

Narrative Maps: A Computational Model to Support Analysts in Narrative Sensemaking

Brian Felipe Keith Norambuena

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Applications

Christopher L. North, Co-chair
Tanushree Mitra, Co-chair
Edward A. Fox
Naren Ramakrishnan
Michael A. Horning

July 6, 2023
Blacksburg, Virginia

Keywords: Narratives, Visual Analytics, Semantic Interaction, Interactive AI
Copyright 2023, Brian Felipe Keith Norambuena

Narrative Maps: A Computational Model to Support Analysts in Narrative Sensemaking

Brian Felipe Keith Norambuena

(ABSTRACT)

Narratives are fundamental to our understanding of the world, and they are pervasive in all activities that involve representing events in time. Narrative analysis has a series of applications in computational journalism, intelligence analysis, and misinformation modeling. In particular, narratives are a key element of the sensemaking process of analysts.

In this work, we propose a narrative model and visualization method to aid analysts with this process. In particular, we propose the *narrative maps* framework—an event-based representation that uses a directed acyclic graph to represent the narrative structure—and a series of empirically defined design guidelines for map construction obtained from a user study.

Furthermore, our narrative extraction pipeline is based on maximizing coherence—modeled as a function of surface text similarity and topical similarity—subject to coverage—modeled through topical clusters—and structural constraints through the use of linear programming optimization. For the purposes of our evaluation, we focus on the news narrative domain and showcase the capabilities of our model through several case studies and user evaluations.

Moreover, we augment the narrative maps framework with interactive AI techniques—using semantic interaction and explainable AI—to create an interactive narrative model that is capable of learning from user interactions to customize the narrative model based on the user’s needs and providing explanations for each core component of the narrative model. Throughout this process, we propose a *general framework for interactive AI* that can handle similar models to narrative maps—that is, models that mix continuous low-level representations (e.g., dimensionality reduction) with more abstract high-level discrete structures (e.g., graphs).

Finally, we evaluate our proposed framework through an *insight-based user study*. In particular, we perform a quantitative and qualitative assessment of the behavior of users and explore their cognitive strategies, including how they use the explainable AI and semantic interaction capabilities of our system. Our evaluation shows that our proposed interactive AI framework for narrative maps is capable of aiding users in finding more insights from data when compared to the baseline.

Narrative Maps: A Computational Model to Support Analysts in Narrative Sensemaking

Brian Felipe Keith Norambuena

(GENERAL AUDIENCE ABSTRACT)

Narratives are essential to how we understand the world. They help us make sense of events that happen over time. This research focuses on developing a method to assist people, like journalists and analysts, in understanding complex information.

To do this, we introduce a new approach called *narrative maps*. This model allows us to extract and visualize stories from text data. To improve our model, we use *interactive artificial intelligence* techniques. These techniques allow our model to learn from user feedback and be customized to fit different needs. We also use these methods to explain how the model works, so users can understand it better.

We evaluate our approach by studying how users interact with it when doing a task with news stories. We consider how useful the system is in helping users gain *insights*. Our results show that our method aids users in finding important insights compared to traditional methods.

Dedication

To my wife, Carolina.

Acknowledgments

The past four years have been an intense journey, going through a doctoral program with a raging global pandemic in the background has sometimes been a strange experience, but overall, I would say it has been one of the most enriching experiences of my life. I am incredibly lucky to have support from a fantastic community of mentors, colleagues, friends, and family. Without the assistance of the following individuals, as well as many others who are not listed, I would not have been able to complete the work described in this dissertation.

First, I would like to express my deepest appreciation to my advisors, Dr. Chris North and Dr. Tanu Mitra, for their great support and guidance throughout this whole process. Thank you for always being receptive and lending an ear to my ideas and worries. From research methodologies to paper writing, they enthusiastically helped me become a better scholar. I have thoroughly enjoyed working with you these past years.

I would like to express my deepest appreciation to my committee members Dr. Edward Fox, Dr. Naren Ramakrishnan, and Dr. Mike Horning. Their feedback and insights have been tremendously useful for guiding my research, improving my dissertation, and discovering potential future directions for my work. Thank you for being readers of my dissertation and providing valuable feedback.

I'm also very grateful to Dr. Eric Krokos and Dr. Sarah Joseph, our collaborators that helped us evaluate and improve our interactive AI models. Your helpful feedback and our discussions have been very useful to my work and I appreciate your support.

I greatly thank those who funded and supported this research. This work was supported in part by NSF I/UCRC CNS-1822080 via the NSF Center for Space, High-performance, and Resilient Computing (SHREC), the NSF grants CNS-1915755 and DMS-1830501, by ANID/Doctorado Becas Chile/2019 - 72200105, and by the Fulbright Faculty Development Program with the support of Universidad Católica del Norte.

I also wish to thank my colleagues Prerna Juneja, Shruti Phadke, Momen Bhuiyan, Kristen Engel, Luis Felipe, Jesse Harden, Kylie Davidson, Wei Liu, and Huimin Han who helped me throughout different stages of the Ph.D. program, either through direct feedback, suggestions, or random conversations and discussion. I would also like to thank all my lab mates from both the InfoVis Lab at VT and the SCALE Lab at UW. I also wish to thank all my colleagues back at Universidad Católica del Norte, who helped me get started in research and be here today. Thank you all for your help.

I would also like to thank my family. First, my parents, Hugo and Rossana, for their steadfast support; without their love and encouragement, I would never have had the chance to get where I am today. They imparted a love for education and curiosity in my formative years and always supported my journey throughout my career. Thanks also to my brother and sister, Hugo and Leslie, for their support. I would also like to thank my in-laws, Ana María and David, for their support throughout these years and welcoming me into their family. I would also like to extend my gratitude to Tabi for their solid support.

Finally, I would also like to deeply thank my wife, Carolina, for her unwavering support throughout these years. Thank you for always being by my side and joining me throughout this quest.

Attributions

Chapter 3 is based on the paper titled “A Survey on Event-based News Narrative Extraction” published in ACM Computing Surveys [190]. Co-authors of this work include Tanushree Mitra and Chris North.

Chapter 4 is based on the paper titled “Evaluating the inverted pyramid structure through automatic 5w1h extraction and summarization” which was presented at the Computational + Journalism Symposium 2020 (C+J 2020) [187]. Co-authors of this work include Tanushree Mitra and Michael Horning.

Chapter 5 is based on the paper titled “Narrative Maps: An Algorithmic Approach to Represent and Extract Information Narratives” which was presented at the 23rd ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2020) and published through the Proceedings of the ACM on Human-Computer Interaction [186]. This paper was co-authored by Tanushree Mitra.

Chapter 6 is based on the paper titled “Narrative Sensemaking: Strategies for Narrative Maps Construction” which was presented at the 2021 IEEE Visualization Conference (VIS 2021) [194] and the paper “Design guidelines for narrative maps in sensemaking tasks” which was published in the SAGE Information Visualization journal [192]. Co-authors of both of these works include Tanushree Mitra and Chris North.

Chapter 7 is based on the paper titled “Characterizing Social Movement Narratives in Online Communities: The 2021 Cuban Protests on Reddit” which was presented at the Computational + Journalism Symposium 2022 (C+J 2022) [195]. Co-authors of this work include Tanushree Mitra and Chris North.

Chapter 8 is based on the paper titled “Mixed Multi-Model Semantic Interaction for Graph-based Narrative Visualizations” which was published through the Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI 2023) [189]. Co-authors of this work include Tanushree Mitra and Chris North.

Chapter 9 is original work that is currently under review [188]. Co-authors of this work include Eric Krokos, Sarah Joseph, and Chris North.

Contents

List of Figures	xvi
List of Tables	xxiii
1 Introduction	1
1.1 Overview and Motivation	1
1.2 Background	2
1.3 Challenges and Goals	3
1.4 Research Questions	3
1.4.1 Milestone 1: Narrative Map Representation and Extraction	4
1.4.2 Milestone 2: Design Guidelines for Narrative Maps	5
1.4.3 Milestone 3: Interactive Artificial Intelligence and Narrative Maps	5
1.5 Structure of This Dissertation	6
2 Review of Literature	9
2.1 Computational Narratives	9
2.1.1 Online Narratives	9
2.1.2 Narrative Theory	10
2.1.3 Narrative Extraction Approaches	10
2.2 Sensemaking Process	12
2.2.1 Sensemaking Loop	12
2.2.2 Cognitive Strategies in the Sensemaking Process	13
2.3 Narratives in Visualization	14
2.3.1 Storytelling with Visualization	14
2.3.2 Narrative Structure and Visual Design	15
2.4 Interactive Artificial Intelligence	15

2.4.1	Semantic Interaction	16
3	A Survey on Computational Narrative Extraction	18
3.1	Introduction	18
3.1.1	Scope of this Survey and Definitions	19
3.1.2	Related Surveys	23
3.2	News Narrative Extraction	24
3.2.1	Overview	24
3.2.2	Events as Sentences	26
3.2.3	Events as Documents	35
3.2.4	Events as Clusters: Event Evolution and Threading	40
3.3	Narrative Extraction Criteria	45
3.4	Evaluation Metrics	49
3.4.1	Computational Metrics	50
3.4.2	User Evaluation Metrics	52
3.5	Discussion	54
3.5.1	Narrative Structure	54
3.5.2	Extraction Methods	55
3.5.3	Evaluation Methods	55
3.5.4	Practical Applications	57
3.5.5	Recent Trends and Open Challenges	58
3.6	Chapter Conclusions	60
4	Main Event Descriptors and the Inverted Pyramid Structure	61
4.1	Introduction	61
4.2	Background and Related Work	63
4.2.1	Structure of News Articles	64
4.2.2	Main Event Descriptor Extraction	64
4.2.3	News Summarization	65

4.3	Data Collection and Annotation	65
4.3.1	AP News Data	65
4.3.2	Annotations	66
4.4	Methods	66
4.4.1	Data Preprocessing	67
4.4.2	Inverted Pyramid Score	67
4.4.3	Main Descriptor Locations	68
4.4.4	Summarization	72
4.5	Results and Discussion	73
4.5.1	5W1H Extraction	73
4.5.2	Inverted Pyramid Score	74
4.6	Chapter Conclusions	75
5	Narrative Maps Representation and Extraction Method	76
5.1	Introduction	76
5.1.1	Route Map Metaphor	79
5.2	An Algorithmic Approach to Represent and Extract Narratives	82
5.2.1	RQ1: Narrative Representation	82
5.2.2	RQ2: Narrative Map Extraction	86
5.3	Case Study: Coronavirus Narrative	94
5.3.1	Overview of the Case Study	95
5.3.2	Analysis: Narratives underlying the Coronavirus Outbreak	95
5.4	User Evaluation	97
5.4.1	Evaluation Guidelines	97
5.4.2	Evaluation Procedure	99
5.4.3	Evaluation Results	100
5.5	Discussion	104
5.5.1	RQ1: Narrative Representation	104
5.5.2	RQ2: Algorithmically Extracting Narratives	104

5.5.3	RQ3: Evaluation	105
5.5.4	Implementation Parameters and Constraints	106
5.5.5	Limitations	106
5.5.6	Implications and Future Work	107
5.6	Chapter Conclusions	109
6	Design Guidelines for Narrative Maps	110
6.1	Introduction	110
6.2	Narrative Maps	112
6.2.1	Motivating Example	112
6.2.2	Research Questions	114
6.2.3	Narrative Extraction	115
6.3	RQ1: Narrative Map Construction Strategies	116
6.3.1	Study Description	116
6.3.2	How do analysts select events?	119
6.3.3	How do analysts connect events?	120
6.3.4	What are analysts' map construction strategies?	124
6.3.5	What are the properties of the created maps?	128
6.3.6	Suggestions and Additional Features	130
6.4	RQ2: Effects of Size and Transitivity	131
6.4.1	Study Description	131
6.4.2	User Performance	134
6.4.3	User Evaluation Results	135
6.5	Design Guidelines	137
6.5.1	What makes a good narrative map?	137
6.6	Discussion	141
6.6.1	Visual Storytelling and Narrative Maps	141
6.6.2	Influence of Analyst Background and Experience	142
6.6.3	Influence of the Data Set and Task Choice	142

6.6.4	Sensemaking Process	143
6.6.5	Limitations	144
6.7	Chapter Conclusions	144
7	Modeling Accepted Narratives in Online Communities	146
7.1	Introduction	146
7.2	Related Work	148
7.2.1	Narrative Extraction	148
7.2.2	Social Movement Narratives	148
7.3	Data Set Description	148
7.4	RQ1: Extracted Narratives	149
7.4.1	Narrative Extraction Method	149
7.4.2	Results	151
7.5	RQ2: Narrative Comparison	155
7.5.1	Narrative Comparison Metric	156
7.5.2	Results	158
7.6	Discussion	160
7.6.1	Social Movement Narratives	160
7.6.2	Extraction and Comparison Methods	160
7.6.3	Implications	161
7.7	Chapter Conclusions	161
8	Semantic Interaction for Narrative Maps	163
8.1	Introduction	163
8.2	Background and Related Work	166
8.2.1	Narrative Extraction and Visualization	166
8.2.2	Semantic Interaction	167
8.3	Narrative Map Extraction	168
8.3.1	Coherence Computation	170

8.3.2	Linear Program Formulation	171
8.3.3	Post-processing	172
8.4	Semantic Interaction Model	174
8.4.1	Mixed Multi-Model Semantic Interaction	174
8.4.2	Semantic Interaction for Narrative Maps	177
8.4.3	Integrating SI into Narrative Extraction	179
8.5	Evaluation Methodology	180
8.5.1	Data Set	180
8.5.2	Simulation Tasks	181
8.5.3	Simulated Analyst Implementation	184
8.5.4	Qualitative Evaluation	188
8.6	Results	190
8.6.1	Regularization Experiment	190
8.6.2	Simulation Results	192
8.6.3	Case Studies	194
8.6.4	Expert Feedback	198
8.7	Discussion	199
8.7.1	Semantic Interaction Model	199
8.7.2	Semantic Interactions and Narrative Sensemaking	200
8.7.3	Model Transparency and Performance	200
8.7.4	Interaction Ambiguity	201
8.7.5	Limitations	202
8.8	Chapter Conclusions	202
9	An Insight-based Evaluation of Interactive AI for Narrative Maps	204
9.1	Background and Related Work	206
9.1.1	Narrative Extraction and Visualization	206
9.1.2	Interactive AI	208
9.1.3	Explainable AI	209

9.1.4	Semantic Interaction	209
9.2	Interactive AI Framework	211
9.2.1	Overview	211
9.2.2	Challenges of Interactive AI	212
9.3	Interactive AI for Narrative Maps	216
9.3.1	Narrative Map Extraction	216
9.3.2	Semantic Interaction Model	217
9.3.3	Explainable AI Components	217
9.4	User Study Description	223
9.4.1	Evaluation Methodology	223
9.4.2	Evaluation Data	224
9.4.3	Participants	224
9.4.4	Evaluation Tasks	224
9.4.5	Insight Coding	225
9.5	Results	226
9.5.1	Insight-based evaluation	226
9.5.2	Quantitative Analysis	229
9.5.3	Qualitative Analysis	233
9.5.4	User Feedback	237
9.6	Discussion	241
9.6.1	Narrative Sensemaking	241
9.6.2	Interactive AI Framework	242
9.6.3	Explainable AI Model	243
9.6.4	Semantic Interaction Model	243
9.6.5	Scalability Issues	244
9.6.6	Limitations	245
9.7	Chapter Conclusions	245

10 Conclusions

247

10.1	Research Questions and System Summary	247
10.1.1	RQ1: How can we computationally represent and extract information narratives?	247
10.1.2	RQ2: How should we design narrative map visualizations to aid analysts in their sensemaking process?	249
10.1.3	RQ3: How can we create an interactive artificial intelligence model for narrative maps?	249
10.1.4	System Summary	250
10.2	Future Work	252
10.2.1	Narrative Representation Issues	252
10.2.2	Narrative Extraction Alternatives	256
10.2.3	Narrative Evaluation Issues	257
10.2.4	Semantic Interaction Challenges	259
10.2.5	Explainable AI Challenges	260
10.2.6	General Challenges	263
10.2.7	Model Extensions and Integration	269
	Bibliography	276
	Appendices	314
	Appendix A Full Insight Tables	315

List of Figures

1.1	Example of the concept of Narrative Maps in addition to semantic interaction and explainable AI techniques. Part (a) shows the base narrative map with annotations produced through explainable methods. Part (b) shows the use of semantic interaction to modify the narrative based on user interaction (e.g., adding events).	3
1.2	Outline of this thesis: analysts interact with a visual narrative representation for sensemaking tasks. The system extracts a narrative graph structure using optimization methods. The semantic interaction system captures users' precise intent and provides assistance to improve the narrative representation.	4
1.3	Structure of this dissertation and dependencies between chapters and components of this dissertation.	7
2.1	Sensemaking loop [282], image taken from [81].	13
2.2	Interactive AI description adapted from [381]. The human cognition component and the AI component are connected through a semantic interaction system. The explainable cognitive intelligence (XCI) component transmits information from the human to the AI. The explainable artificial intelligence (XAI) component transmits information from the AI to the human.	16
2.3	Semantic interaction pipeline adapted from [39]. The pipeline shows how the analyst and the visual analytics system work together to update the visualization and underlying model parameters through an iterative loop. The interactive dimensionality reduction component captures the intent of the analyst based on the modified projection (i.e., the human spatialization) and updates the projection in response (i.e., the model spatialization).	17
3.1	Overview of the article collection process and the inclusion/exclusion criteria used to construct the final article set.	22
3.2	Overview of the different methods used in news narrative extraction categorized by event resolution.	25
3.3	Resolution level and narrative structure. Examples adapted from several works in this survey.	25

4.1	Diagram describing the inverted pyramid structure. The information is shown in descending order of importance, with the key details at the top.	62
4.2	News article [242] showing the Opening Paragraphs (OP) and the rest of the article. The highlighted phrases show the answers to the 5W1H questions that define a news event: <i>What</i> happened? <i>Who</i> is involved? <i>When</i> and <i>Where</i> did it happen? <i>Why</i> and <i>How</i> did it happen?	63
4.3	Example of a question from the evaluation questionnaire. The questionnaire showed the full article with highlighted answers, some additional information about the article on the top left, and the answers below that. The students had to select one answer for each 5W1H question after reading the article and the proposed answers.	67
4.4	Diagram showing the computation of the Inverted Pyramid Score (IPS) of a news article. The first step is preprocessing, then we compute the <i>main event descriptor locations</i> score and the <i>summary similarity</i> score. For the main descriptor locations score we extract the position of the main event descriptors (5W1H answers), then we get the weighted average of these scores. For the summary similarity score, we get the similarity between an extractive summary of the article and the key details. We get the final IPS using the weighted average of the previous scores.	68
4.5	Plot of the IPS distribution and basic statistics for breaking and non-breaking news. On average, breaking news has a higher IPS than non-breaking news.	74
5.1	Illustration of the Route Map Metaphor. (a) A route map showing how to get from Start Stadium to Classroom Hall. (b) A narrative route map showing how to get from the Mysterious Virus in Wuhan event (starting event ) to the US banning flights to Europe (ending event ). We highlight some representative landmarks  for each route in the narrative map.	80
5.2	Figure illustrating the steps in our narrative extraction algorithm that generates a narrative map from data (RQ2). Each phase is annotated with their respective outputs.	87
5.3	Overview of the edge membership computation used in the coverage constraint. The example shows three events from the Coronavirus narrative and two clusters. The first edge does not contribute much to the coverage of either cluster, because it is joining nodes from different clusters. In contrast, the second edge joins nodes from the same cluster, and thus it strongly contributes to the economic impacts cluster.	91

5.4	Antichain examples. The nodes highlighted in green form the first maximum antichain of the graph with 3 elements. The nodes highlighted in purple also form a maximum antichain, but it is not the earliest antichain. The nodes 2 and 3 on their own also form an antichain, but it is not a maximum antichain. Note how each element of the antichain is associated with a chain (left, center, and right)	93
5.5	Narrative Map for the Coronavirus Outbreak during January. Headlines in bold denote the starting event  and ending event  . The blue dashed line corresponds to the main storyline  (the maximum likelihood path). The width of each edge depends on its weight (coherence). Events highlighted in green are representative landmarks  (members of the first maximum antichain). The markers  ,  , and  highlight some storylines that we reference throughout our discussion, we display them in distinctive colors. The  markers show events that share the theme of questioning or criticizing government responses.	94
5.6	Analysis summary of the main route and representative landmarks for the January narrative map. We extract the major themes out of these elements.	96
5.7	Examples of substories from the Narrative Map for the Coronavirus Outbreak during January. For brevity the start event  and ending event  are not shown in detail. Events highlighted in green are representative landmarks  . We use markers  ,  , and  to refer to the storylines as in Figure 5.5. . . .	96
5.8	User evaluation results for the representation dimension. The bars show the percentage of users per response for each question.	101
5.9	User evaluation results for the metaphor dimension. The bars show the percentage of users per response for each question.	102
5.10	User evaluation results for the visualization dimension. The bars show the percentage of users per response for each question.	103
5.11	Potential extensions to narrative maps including different elements. The top-left map considers adding credibility annotations to each of the events, representing how credible or questionable the sources and their claims are. The bottom-left map considers adding political bias annotations, thus helping users understand how bias affects the storylines of the map. The top-right map considers additional information in the form of a news summary and data visualization. The bottom-right map considers adding the answers to the 5W1H questions (who, what, where, when, why, and how), which are the main descriptors of an event.	108

6.1	Example of a Narrative Map showing the COVID-19 narrative during January 2020 from news articles. The highlighted panel shows some of the important outcomes of the outbreak (the lockdown in China, social effects, and economic effects).	113
6.2	Overview of the experiments. The map construction experiment was used to answer RQ1. The user evaluation for size and transitivity was used to answer RQ2.	115
6.3	Narrative map examples created by participants for the two tasks Directed Task (DT) and Open-ended Task (OT). Annotations highlight key elements in these maps. Note how in the DT map all storylines converge into a single ending event, emphasizing the focused nature of this task. In contrast, the OT map has a series of storylines that interact with one another, representing the different outcomes found by the participant.	117
6.4	Narrative map construction strategies. (a) Clustering strategies : no clustering, clustering during preprocessing, or clustering in the middle of construction. b) Construction focus : whether participants created the main story (in blue) first, the side stories first, or a mixed strategy. (c) Algorithm type : the order in which nodes were added, following either a breadth-first, depth-first, or clustering approach. (d) Inter-story connections : some participants checked for inter-story connections (in red) when adding events, others checked when completing a storyline.	125
6.5	Examples of different clustering strategies. The top example shows the creation of a cluster through an intermediate clustering step during map construction (the highlighted events are about "Chinese Media Criticism"). The bottom example shows clustering as preprocessing , where cluster labels are assigned before constructing the map (the highlighted events were initially classified with respect to the questions asked in the instructions).	127
6.6	Example narrative map structure from a participant of the Directed Task. Note that the map has a diagonal layout—the only map that uses this type of layout—with its main story (1) on its center. Moreover, this map was constructed following a depth-first strategy, starting with the main story and then branching into the side stories (2). Some events that were considered too similar or the same were grouped together into a single block (3). Inter-story connections (4) were added following a by storyline pass strategy.	129
6.7	Overview of all the maps used in the evaluation procedure of RQ2. Normal maps (N) have more connections, allowing them to show more details at the cost of more complex layouts compared to their reduced counterparts (T).	133
6.8	Heat map showing the average recall for the main story and the side stories averaged over the issues.	135

6.9	Percentage of favorable responses for each question and each <i>size</i> and <i>transitivity</i> combination. The best result was obtained by Long-T, followed by Medium-T, and then Long-N.	136
7.1	Linear program formulation for the narrative extraction problem with community acceptance.	150
7.2	Narrative map for <i>r/cuba</i> . The numbers in parenthesis beside each submission represent the upvote ratio and score percentile. The main storyline is represented by the blue dashed lines. Highlighted nodes correspond to the representative landmarks of the narrative map. The edge weights represent the coherence normalized as a probability. Manual annotations describing specific storylines have been added with markers 1, 2, and 3.	152
7.3	Narrative map for <i>r/Conservative</i>	154
7.4	Narrative map for <i>r/GenZedong</i>	155
7.5	Narrative map for <i>r/worldnews</i>	156
7.6	Narrative map for <i>r/politics</i>	157
7.7	Examples showing the differences captured by (normalized) GED with unitary costs.	157
8.1	Overview of the semantic interaction model for narrative maps and its core interactions. The analyst modifies the original map on the left by: (1) adding an edge between events <i>a</i> and <i>e</i> , (2) adding event <i>c</i> , (3) removing event <i>d</i> , (4) removing the edge between events <i>e</i> and <i>g</i> , (5) grouping events <i>b</i> , <i>f</i> , and <i>h</i> by assigning them to the same cluster. After this, the system makes changes to the extraction process based on the interactions and generates a new map that integrates the analyst's interactions into the model.	165
8.2	The traditional semantic interaction pipeline has the analyst interact with a dimensionality reduction model by making changes to the projection space. These changes are captured by the machine learning model and used to generate a new projection space. The process repeats as the analyst works through their sensemaking process.	167
8.3	Narrative maps extraction pipeline. The pipeline consists of two main phases (extraction and post-processing).	169
8.4	Linear program used to obtain the optimal narrative map (i.e., the basic structure space).	170
8.5	Generalized Mixed Multi-Model Semantic Interaction (3MSI) visualization pipeline.	175

8.6	(a) Constraints induced by basic interactions (addition and removal). (b) Additional constraints to ensure that user-defined clusters are connected. (b) Examples of weakly connected subgraphs produced by the previous constraints in the narrative map.	181
8.7	Example of the simulation steps for task T3 (Clean Up Storylines). (a) The initial storyline. (b) Removing the inconsistent edges from the storyline. (c) Reconnecting the storyline with consistent edges.	187
8.8	Prototype interface showing an example extracted map in the main canvas, the data set table in the right tab, and the main menu and extraction parameters on the top.	189
8.9	Overfitting analysis for the basic model vs. the regularized model. The left column shows the error rate plots. The middle column shows plots for the number of nodes throughout the simulation. The right column shows plots for the number of edges throughout the simulation. Both models converge towards zero error rate, but the basic model has overfitting issues in tasks T4 and T5 based on its drastic increases in graph complexity metrics (number of nodes and edges).	191
8.10	Error rate vs. the number of iterations. (a, b, c) T1: Remove Irrelevant Events. (d) T2: Remove Inconsistent Edges. (e) T3: Clean Up Storylines. (f, g) T4: Cluster Events. (h, i) T5: Add Events.	193
8.11	(a) Initial map and analyst interactions (add events and cluster events). (b) Second map and analyst interactions (remove events). (c) Third map and analyst interactions (add edge and remove edge). (d) Final map after all iterations. Note that a higher-resolution version is available as part of the supplementary materials.	195
8.12	(a) Initial map and interactions performed by the analyst (add events and cluster events). (b) Final map after a single iteration. Note that a higher-resolution version is available as part of the supplementary materials.	197
9.1	Illustration of the interactive AI framework applied on narrative maps for sensemaking. The core components of this pipeline are the <i>semantic interaction model</i> and the <i>explainable AI components</i> . The user interacts with these components during sensemaking by changing the visual narrative model and leveraging the explanations of the explainable AI components to understand how the narrative was constructed.	205
9.2	Narrative map example based on a COVID-19 news data set. The model includes explanations such as storyline names, connection labels, and high-lighted important events.	207

9.3	Mixed Multi-Model Semantic Interaction pipeline (source: [189]).	210
9.4	Mixed Multi-Model Interactive AI framework. Red arrows represent the “inputs” of the XAI components and what they seek to explain. Blue arrows represent the “outputs” of the semantic interaction model and point toward what they seek to influence.	212
9.5	Example of topic modeling explanations for the low-level space alongside a scatter plot of the low-level space and an overlay of the main storyline going through the low-level space. We also include tooltips that appear when the user hovers over a specific event.	218
9.6	Example of an explanation for an edge between two events in the 2021 Cuban Protests data set.	221
9.7	Coding tree showing our hierarchical coding scheme for insights. The leaves of the coding tree represent the high-level insights and include a histogram showing the frequencies of each high-level insight by group.	227
9.8	(a) High-level (topic) insights results. (b) Low-level (detailed) insights results. (N.S.) denotes that there is no significant difference between the groups ($p \geq 0.1$). (.) Denotes a marginally significant difference ($p \in [0.05, 0.1)$). (*) denotes a significant difference ($p < 0.05$).	228
9.9	(a) Perceived effectiveness. (b) Number of events read. (c) Number of executions. (N.S.) denotes that there is no significant difference between the groups ($p \geq 0.1$). (.) Denotes a marginally significant difference ($p \in [0.05, 0.1)$). (*) denotes a significant difference ($p < 0.05$).	230
9.10	A tally of the answers to our survey questions on <i>usability</i> and <i>representation</i> of the narrative across all three prototypes.	231
9.11	A tally of the answers to our survey questions on the <i>narrative map elements</i> of the BM prototype and the interactive AI prototype.	232
9.12	A tally of the answers to our survey questions on the <i>semantic interaction</i> components and <i>explainable AI</i> components.	233
9.13	Histograms representing the number of users that read each event for each condition: (a) Timeline with Search (b) Basic Map (c) Interactive Map. The red line represents what an ideal uniform distribution would look like.	236
9.14	Lorenz curves describing the diversity of the parameter combinations used by the participants for the BM and IM prototypes. (a) Size (b) Coverage (c) Temporal Sensitivity (d) Starting Event choice.	237

List of Tables

3.1	Summary of the surveyed articles.	27
3.2	Summary of the extraction criteria and evaluation metrics used in the reviewed articles.	46
3.3	A sample of different formulations of coherence from the reviewed articles. .	48
3.4	Benchmark data in the timeline summarization works.	56
4.1	5W1H evaluation results for the breaking news in the AP data set by subject category ($\pm$ standard errors).	73
5.1	Mapping of narrative theory elements and computational elements alongside their route map metaphor equivalent. The concepts shown here are used interchangeably throughout the paper to describe our algorithmic approach. Narratives have at least one storyline, and a storyline has at least two events. Thus, we have the hierarchy: event < storyline $\leq$ narrative.	83
5.2	Table containing the evaluation dimensions and their elements, as well as the questionnaire items and their source (if applicable). Some questions taken from a source have been adapted to our metaphor. The questionnaire has a total of 22 elements. Subjects were asked to evaluate each statement according to a 5-point Likert scale, ranging from “Strongly Disagree” to “Strongly Agree.”	98
6.1	Selection criteria for important events.	119
6.2	Events that were selected by a majority of the participants in at least one task. The first column shows the event, the second and third column show the frequency for that particular task, and the last column shows the global frequency. Note that only one event is common to both tasks (human-to-human transmission event).	121
6.3	Connection types for each participant in our user study.	122
6.4	Connection types for each participant in our user study.	122
6.5	Construction strategies for each participant in our user study.	126
6.6	Graph and Layout Properties for each participant in our study.	130

6.7	Average percentage of favorable responses in our evaluation questionnaire for each <i>size</i> and <i>transitivity</i> combination. The best result was obtained by Long-T, followed by Medium-T, and then Long-N.	135
6.8	Summary of the design guidelines found in our analysis of results from RQ1 and RQ2.	138
7.1	Summary of the analyzed Reddit communities, including the number of subscribers and submissions retrieved over the 10-day period (July 7 - 17). . . .	149
7.2	Description of the variables and parameters of the proposed extraction algorithm. All elements take values between 0 and 1, except K and num-clusters , which are positive integers. Most elements of the model are based on the extraction algorithm of Keith Norambuena and Mitra [186]. The (*) symbol represents new additions to the linear problem formulation designed to incorporate <i>community acceptance</i> in social media into the model.	151
7.3	Costs used in the Graph Edit Distance algorithm and a description of each parameter.	158
7.4	Relevant storylines and events that appear in each narrative map. We note that the presence of storylines is not correlated with the number of submissions. r/Con^* refers to $r/Conservative$, r/GZ^* to $r/GenZedong$, r/wn^* to $r/worldnews$, and r/pol^* to $r/politics$	159
7.5	Pairwise Graph Edit Distance (GED) between the different communities. . .	159
8.1	Evaluation task description and evaluation metrics.	183
9.1	General strategies used by participants with map-based prototypes.	234
9.2	Semantic interaction strategies used by participants. Column codes: C for <i>Corrective SI</i> , A for <i>Additive SI</i> , and P for <i>Parametric Exploration</i>	235
9.3	Keywords used by participants in the TL prototype and the map-based prototypes.	238
9.4	Suggested functionalities by the participants of the user study.	239
9.5	Issues and complaints obtained through user feedback and observation in during the user study.	241
A.1	Table of insights for causes.	316
A.2	Table of insights for effects in Cuba.	317
A.3	Table of insights for effects in the U.S.	318

A.4	Table of insights for international effects and context information.	319

Chapter 1

Introduction

People don't see the world before their eyes until it's put in a narrative mode.

- Filmmaker Brian De Palma [1]

Narratives are fundamental to our understanding of the world [1] and are central to human relations [257]. They are the frameworks that enable humans to associate otherwise unconnected events [55] and play a key role in collaborative sensemaking in society [22, 386].

For decades, communication science scholars have emphasized the importance of narratives in human communication [1]. Moreover, narratives can be used as a sensemaking tool in different tasks, such as news analysis or intelligence analysis [335]. In particular, by constructing appropriate representations from news data, intelligence reports, or other kinds of events, we can obtain a greater understanding of the underlying storylines and the big picture of the narrative. Extracting storylines from data is a task that falls within the scope of *Information Cartography* [324], which seeks to create methods that help us understand the information landscape.

In this introduction, we first provide an overview of our research and its motivation. Then, we provide a brief background in narrative extraction and related works. We next specify our main challenges and goals, which lead to the research questions. Finally, we end this chapter with a description of the structure of the rest of the dissertation.

1.1 Overview and Motivation

In this dissertation, we propose a narrative model and visual analytics method to aid analysts with the sensemaking process [282]. In particular, we propose the *narrative maps* framework—an event-based representation that uses a directed acyclic graph to represent the narrative structure—and a series of empirically defined design guidelines for map construction obtained from a user study.

In addition, we augment the narrative maps framework with interactive AI techniques—using semantic interaction [104] and explainable AI [41]—to create an interactive narrative model capable of learning from user interactions to tailor the narrative model to the user's needs

and providing explanations for each core component of the narrative model. Throughout this process, we propose a *general framework for interactive AI* that can be used for similar models to narrative maps—models that combine continuous low-level representations (e.g., dimensionality reduction) with more abstract high-level discrete structures (e.g., graphs).

Research on narrative maps, and information cartography in general, has widespread implications. At its core, these representations can help users understand the big picture of a narrative. Moreover, expert users, such as journalists, fact-checkers, and intelligence analysts could leverage our methods to explore the narrative landscape in the context of their domain of work. Furthermore, by providing the general public with easy-to-use narrative analysis tools, we hope to allow them to gain a better understanding of how storylines evolve and interact. Moreover, given the issue of pervasive misinformation [214] and information overload, this will also help users understand the underlying biases and potential credibility issues of different storylines.

1.2 Background

Given a large collection of documents from multiple sources (e.g., news articles, tweets, social media posts, intelligence reports) containing events, we seek to extract and visualize the underlying narrative that represents the information landscape. Thus, the overall aim of this investigation is to develop methods to understand and characterize the information landscapes extracted from data through the use of narrative maps.

The problem of extracting storylines from data and “connecting the dots” is a key task in intelligence analysis and similar sensemaking tasks [51, 161]. However, even though there exist visualization tools and storyline extraction methods to approach this problem, most of these initiatives do not make use of the extensive work on formal narrative theory [186].

Previous research initiatives, such as the work on Information Cartography [324], provide a solid foundation for our work. In particular, we highlight previous methods to represent stories such as the Metro Maps method [321]. Building upon similar work and grounded on the scholarly work of formal narrative theory, we developed a computational representation for narratives, an initial extraction method, and a static visualization method [186]. However, our extraction method and the visualization still have some limitations. This dissertation seeks to build upon this research, improving the extraction method and the visualization approach.

In addition to this, recent developments in semantic interaction [107] and explainable AI methods [157] have shown the capabilities of these techniques in aiding the sensemaking process [41]. Thus, we seek to exploit such methods in the context of narrative visualization.

These techniques could prove useful to expert users that seek to extract and understand evolving narratives, such as intelligence analysts trying to uncover the underlying connec-

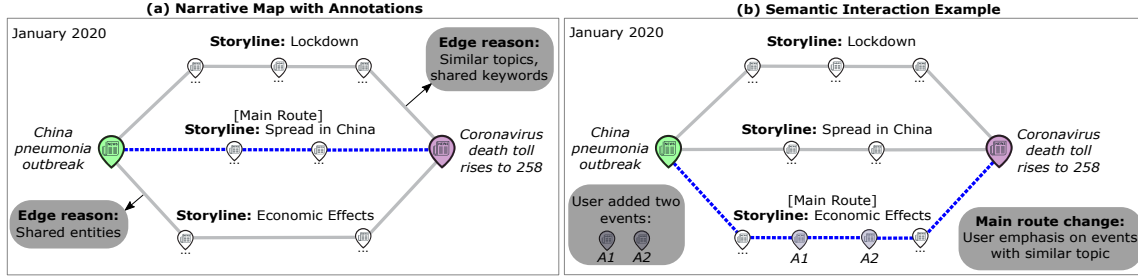


Figure 1.1: Example of the concept of Narrative Maps in addition to semantic interaction and explainable AI techniques. Part (a) shows the base narrative map with annotations produced through explainable methods. Part (b) shows the use of semantic interaction to modify the narrative based on user interaction (e.g., adding events).

tions between documents, journalists attempting to understand the rapidly-evolving news narratives, or fact-checkers trying to understand how misinformation narratives emerge in the information landscape, and other expert users interested in extracting and understanding narratives. Figure 1.1 illustrates the high-level pipeline of our work.

To date, no previous research has sought to use semantic interaction or explainable AI methods in a narrative or story visualization setting. Thus, this research would provide a first step towards creating a human-AI interaction system that aids in the narrative sensemaking process.

1.3 Challenges and Goals

The main objective of this thesis is to develop a computational narrative framework to support the sensemaking process of analysts and model the information landscape. Specifically, this thesis aims to integrate narrative theory, interactive visualization, and artificial intelligence to model and extract narratives computationally, allowing analysts to use the extracted narratives as support in their sensemaking process. Moreover, our framework will use semantic interactions to capture analysts' precise intentions behind their interactions and explainable AI to help users understand the underlying AI model. Furthermore, we focus our studies on exploring news narratives. Figure 1.2 outlines the key process and elements of this thesis.

1.4 Research Questions

To tackle these challenges, we focus this dissertation on exploring the following four research questions:

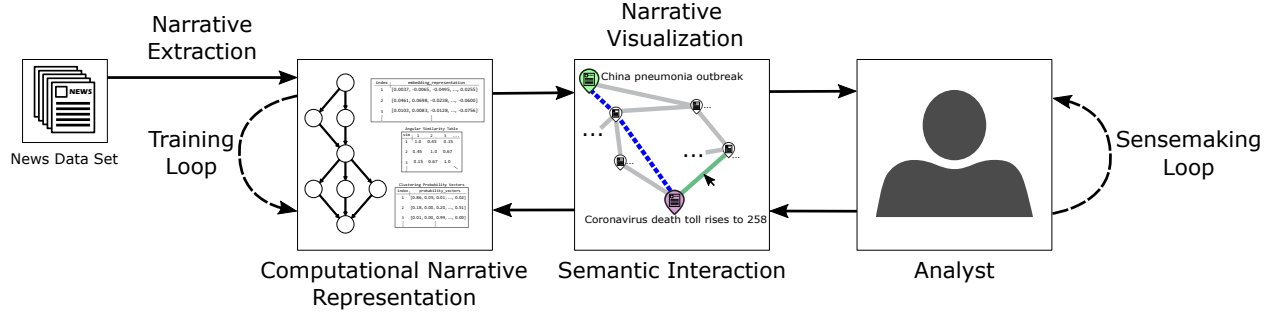


Figure 1.2: Outline of this thesis: analysts interact with a visual narrative representation for sensemaking tasks. The system extracts a narrative graph structure using optimization methods. The semantic interaction system captures users’ precise intent and provides assistance to improve the narrative representation.

- **RQ1 - Narrative Extraction:** How can we computationally represent and extract information narratives?
- **RQ2 - Design Guidelines:** How should we design narrative map visualizations to aid analysts in their sensemaking process?
- **RQ3 - Interactive AI:** How can we create an interactive artificial intelligence model for narrative maps?

We associate each one of these research questions with one of the four milestones of this project. We discuss them in detail in the following subsections.

1.4.1 Milestone 1: Narrative Map Representation and Extraction

The first milestone of this project is focused on developing a computational representation model and an extraction approach to obtain narratives from data.

In Chapter 3, we performed a systematic literature review on narrative extraction. In particular, we screened over 750 articles, which yielded 54 articles focusing on extracting event-centric news narratives. We identified common representational structures, extraction criteria, and evaluation metrics. Based on the studies reviewed, we analyzed a number of recent tendencies, open issues, and prospective research directions.

In Chapter 4, we created a computational approach using natural language processing to extract the main event descriptors—the answers to the 5W1H questions (Who? What? Where? When? Why? How?)—from breaking news articles. The main event descriptors could be used in a narrative representation to capture the details of the events in a more precise manner. In addition, we compute an evaluation metric—the Inverted Pyramid Score (IPS)—that is able to evaluate whether or not an article matches the inverted pyramid

structure [150] used in traditional journalism. Low values of the IPS can be utilized as a potential indicator to determine whether or not an article is credible, which can provide extra information for our narrative representation.

In Chapter 5, we developed a narrative representation model called *narrative maps*, which forms the core of our work throughout this dissertation. In particular, we made three key contributions: a *computational representation* of narratives based on narratological theory, a novel *extraction algorithm* based on linear programming optimization to obtain these representations from data, and a *user evaluation* of our approach. We showcased the effectiveness of our approach by performing a user evaluation to assess the quality of the representation, metaphor, and visualization.

In Chapter 7 we modeled the accepted narratives of online communities related to social movements. In particular, we analyze the 2021 Cuban Protests in five Reddit communities. To perform this analysis, we develop a new narrative extraction algorithm that incorporates *community acceptance* into the model. Moreover, we develop a narrative comparison metric based on graph edit distance to quantitatively analyze the differences between the accepted narratives of different communities.

1.4.2 Milestone 2: Design Guidelines for Narrative Maps

The second milestone of this project is focused on designing an interactive visualization tool to aid in the narrative sensemaking process. To achieve this objective, we first needed to understand how users create and apply narrative maps to solve sensemaking tasks and what makes a good narrative map.

In Chapter 6, our goal was to identify design principles for a narrative map visualization toolkit that can assist analysts in narrative sensemaking. Thus, we set out to gain an understanding of how analysts make and use narrative maps. We ran two experiments using a data collection consisting of 40 news articles. In particular, our contributions consist of the following: a *user evaluation* of specific narrative map elements; an examination of *how analysts construct narrative maps*; and *design guidelines* for narrative maps. Our findings provide insights into useful user interactions as well as strategies to create narrative map representations and extraction algorithms.

1.4.3 Milestone 3: Interactive Artificial Intelligence and Narrative Maps

The third milestone of this project is focused on incorporating higher-level sensemaking capabilities through an interactive AI model [381] into narrative maps representation. Interactive AI models are comprised of two core components: a semantic interaction [104] component

to capture user feedback and an explainable AI [157] component to provide feedback to the user about the underlying AI model.

In Chapter 8 we addressed the first component of the interactive AI model by designing a general semantic interaction framework and applying it to the narrative maps extraction pipeline. In particular, we identify the key challenges associated with the semantic interaction model and show how to apply the framework in the case of narrative maps to create customized narratives based on the user’s needs. We present a quantitative simulation-based evaluation and a qualitative expert-based evaluation.

In Chapter 9 we extended our previous semantic interaction framework into a general interactive AI model that incorporates both semantic interaction and explainable AI. We further identified the key design challenges of this approach and show how we solved them in the case of narrative maps. Finally, we present a user study to evaluate our model following an insight-based evaluation methodology.

1.5 Structure of This Dissertation

To summarize all the work in this dissertation, we present the structure and the contents of each chapter, alongside associated publications.

Figure 1.3 summarizes the structure and dependencies between chapters. We note that there are two core components to this dissertation: the narrative maps framework—which includes Chapters 5 and 6—and the interactive AI framework—which includes Chapters 8 and 9. Furthermore, there are related side projects that deal with specific elements of event extraction and representation (Chapter 4) and narrative extraction in online communities (Chapter 7) that are included in this body of work. Finally, we also include the results of a systematic literature review (Chapter 3) that provides additional context to the narrative maps framework and subsequent parts of this work, but this chapter is not strictly required to understand the contents of the latter chapters.

We now present a description of each chapter and the associated papers (if appropriate):

- Chapter 1 introduces the background and motivation of this work, and lists the key challenges, goals, and research questions.
- Chapter 2 presents a literature review on computational narratives, the sensemaking loop, semantic interaction, and misinformation modeling.
- Chapter 3 presents an extensive literature review on narrative extraction using event-based representations and is part of our basis to address Research Question 1 [190].
- Chapter 4 addresses Research Question 1 by presenting an extraction model for main event descriptors from news articles—which can be used as part of a full narrative

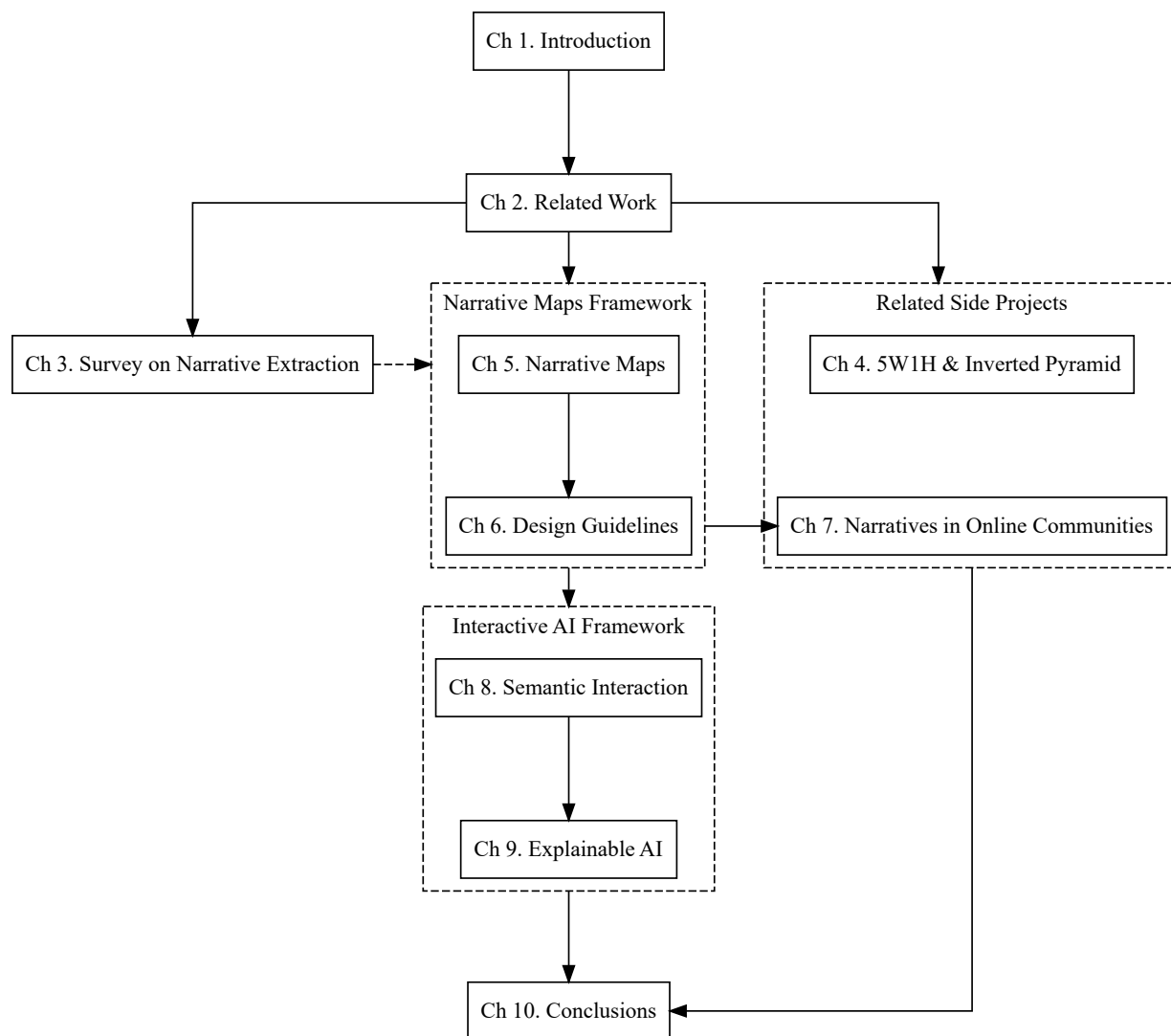


Figure 1.3: Structure of this dissertation and dependencies between chapters and components of this dissertation.

model—and a model of how well the articles follow the inverted pyramid structure—which can be used as an indicator of credibility [187].

- Chapter 5 addresses Research Question 1 by integrating narrative theory, a visual metaphor, and a computational model to represent and extract narratives using narrative maps [186].
- Chapter 6 addresses Research Question 2 by providing design guidelines for narrative maps obtained by evaluating how analysts manually construct and use narrative maps [192, 194]. This chapter contains results from a human subjects study approved by IRB 20-794.

- Chapter 7 addresses Research Question 1 in the context of social media communities by designing and implementing a narrative extraction model incorporating the concept of community acceptance. The model used here also incorporates the design guidelines obtained in the previous chapter [195].
- Chapter 8 addresses Research Question 3 by designing a general semantic interaction framework. In particular, we identify the key challenges associated with this approach and show how to apply the framework in the case of narrative maps to create customized narratives based on the user’s needs. We present a quantitative simulation-based evaluation and a qualitative expert-based evaluation [189].
- Chapter 9 addresses Research Question 3 by extending our previous semantic interaction framework into a general interactive AI model that incorporates both semantic interaction and explainable AI. We further identify the key design challenges of this approach and show how we solved them in the case of narrative maps. Finally, we present a user study to evaluate our model following an insight-based evaluation methodology [188]. This chapter contains results from a human subjects study approved by IRB 21-929.
- Chapter 10 concludes the dissertation, summarizes our core contributions, describes potential future work, and addresses remaining challenges.

Chapter 2

Review of Literature

2.1 Computational Narratives

In this section, we introduce important background concepts from narrative theory and present research related to our main contributions. We start with online narratives, then we move on to narrative theory definitions, and finally, we describe existing approaches to extract and represent narratives and stories.

2.1.1 Online Narratives

The exponential growth of the World Wide Web and the improved accessibility of information has exacerbated the problem of information overload in our daily lives [154]. In particular, this becomes an issue when trying to process and understand the barrage of news events [319]. Furthermore, the rise of social media and its subsequent place as a staging ground for information operations [300, 376] has given place to the creation of alternative narratives [270] and the spread of fake news [78].

Moreover, narratives play a key role as one of the basic building blocks of societal sensemaking [386] and political actors attempt to use strategic narratives to create a shared understanding of the world [258]. Thus, given the importance of narratives in society, it is critical to understand how they emerge and evolve over time in this environment of information overload and fake news.

Existing work has focused on algorithms to extract timelines [319], trees [18], or other variants of graphs to represent stories [401]. However, they have mostly neglected the theoretical foundations of narratives in their models. The present work is an attempt to bridge this theoretical gap by creating a computational representation model for narratives grounded in formal narrative theory. In particular, the proposed model is based on a route map metaphor and it links elements from narrative theory and graph theory in an intuitive way.

2.1.2 Narrative Theory

The design of our narrative representation is informed by concepts from general narrative theory and strategic narrative theory. Both approaches contain elements that help us define and analyze our narrative representation.

General narrative theory focuses explicitly on understanding the general rules of narrative and its different arrangements that make it meaningful [1, 285]. The key intuition in formal narrative theory is that there is a distinction between the story itself and its representation. Narrative theory tries to understand the relationships between stories and their many possible representations [285].

In contrast, instead of focusing on the narrative itself, strategic narrative theory studies the construction and application of these narratives by political actors. In particular, strategic narrative theory focuses on studying how political actors communicate narratives in a strategic way to construct a shared meaning of events with the purpose of influencing the behavior of domestic and international actors [257, 303].

Another approach to narratives by Halverson et al. [145] defines narratives not just as one story, but rather as a system of stories. That is, narratives are a systematic collection of interrelated stories with coherent themes. We rely on this definition of narrative as a system of interrelated stories to build our narrative map extraction method. In particular, our method also ensures a certain degree of thematic coherence between the extracted storylines.

2.1.3 Narrative Extraction Approaches

To inform the design of our extraction method, we refer to the existing story and narrative representations found in the literature. First, we discuss related works based on the level of detail (*representation resolution*), for example, events or topics in the narrative. Afterward, we analyze related works based on the internal structure of the representation (*representation complexity*), for example, a tree or a timeline. We expand on this topic in Chapter 3 by providing a systematic literature review on narrative extraction.

Representation Resolution

Multiple resolution levels could be studied in the context of story and narrative visualization. In particular, most of the story visualization research has focused on either the event level [232, 319] or the topic level [198, 265, 412]. Event-level approaches focus on concrete sequences of events. In contrast, topic-level approaches are more abstract. Thus, instead of focusing on the specific events that happen throughout the story, they focus on the overarching topics and how they relate to each other.

In particular, previous studies have used graph representations based on event clustering

to represent topics [265]. Our method builds upon the idea of using clusters of events to represent topics. In particular, we use them to extract coherent storylines aligned with these clusters.

Some authors have also proposed more fine-grained resolution levels to represent stories, such as individual named entities [114], claims and attributions made in an article [334], or mixed resolution approaches that allow zooming between the different levels, such as the information cartography approach from Shahaf et al. [323]. In contrast to these fine-grained approaches, we focus on the event level to build our narrative map representation. By choosing this resolution level, we are able to easily develop a direct mapping between narrative and computational elements.

Representation Complexity

There exist many approaches to visualize the evolution of news events. The simplest cases correspond to a linear representation of the events (i.e., simple sequences of events), such as timelines [369, 399] and coherent story chains [319, 320]. All these representations consist of a single starting and ending event with a sequence of events joining them.

Previous research on timeline and story extraction relies on optimizing different criteria to select events. Furthermore, these criteria could depend on the type of data used by the extraction method. For example, in the socially informed timelines approach by Wang et al. [369], their optimization method seeks to balance topical cohesion between articles and comments on social media. In contrast, in the news timeline summarization method of Yan et al. [399], the optimization algorithm tries to balance relevance, coverage, coherence, and cross-date diversity. Another approach corresponds to the **Connect-the-Dots** algorithm by Shahaf and Guestrin [319], which creates story chains using a linear programming approach to maximize the coherence of the timeline. An improved version of their approach [320] used integer programming to generate the story chain. Building upon the intuition of these previous works, our approach also relies on an optimization problem to select the events that appear in our narrative representation.

A logical extension to the single timeline approach is to consider more than one timeline. These timelines can be shown in parallel in the story map, such as in the work done by Xu et al. [395], where they extract multiple storylines surrounding a news event. Due to the complexity of generating such maps by brute force, they develop a near-optimal algorithm based on locality-sensitive hashing to perform dimensionality reduction. They display all the storylines in the same story map as parallel timelines. In contrast to this approach, our method allows for the different storylines in the narrative to interact with one another, allowing them to merge or separate into new storylines. This allows a more appropriate representation of narratives as a system of interrelated stories, rather than disjointed timelines.

More advanced cases use directed acyclic graph structures, such as event evolution graphs

[401, 413], and metro maps [321, 323]. In these cases, the maps can have multiple starting and ending events. Furthermore, these structures allow for more advanced local substructures, such as convergence (merging stories) and divergence (separating stories). Scholars have studied different models and features to represent and extract the relationships between events in these graphs. For example, the event evolution graph of Yang et al. [401] uses event timestamps, event content similarity, temporal proximity, and document distributional proximity to model relationships. In contrast, Zhou et al. [413] use a model based on term and event frequency, alongside event sequences, content similarity, and temporal distance costs. Another approach corresponds to optimizing the connectivity of the graph, subject to coherence and coverage constraints [321]. Following these approaches, we model our event relationships using a similarity-based approach in conjunction with clustering information to ensure topical coherence and a certain degree of coverage.

Finally, we also have tree-based approaches to represent stories. For example, the tree-based representation of Ansah et al. [18] leverages information propagation among users in a community, temporal proximity, and semantic context to construct coherent paths and build a timeline summary. In addition, the news content organization system of Liu et al. [232] builds a forest representation using keyword extraction, event clustering, story clustering, coherence, and time-based penalties.

2.2 Sensemaking Process

In this section, we present the general sensemaking process. In particular, we discuss the sensemaking loop and cognitive strategies used in sensemaking.

2.2.1 Sensemaking Loop

Sensemaking is the process through which analysts look for a representation and encode data to answer task-specific questions [305]. This process has been modeled as an iterative loop with a series of steps involving information foraging and synthesizing hypotheses [282].

We now present an overview of the sensemaking process as modeled by Pirolli et al. [282]. In particular, we briefly describe the steps shown in Figure 2.1. First, analysts have to forage for relevant source data by searching and filtering. Then, they extract useful information by reading and extracting the relevant elements from data. Afterward, analysts must organize and represent the information through a cognitive schema. Using these schemas, the analysts then build a case by developing a series of different hypotheses. Then, they choose the best explanation and tell the corresponding story. As part of the iterative process, each one of these steps can be retraced and repeated as necessary. We note that there are two distinct parts in the sensemaking process: the foraging loop and the sensemaking loop. Throughout this dissertation, our focus is on the second loop, as we assume that data has been previously

foraged.

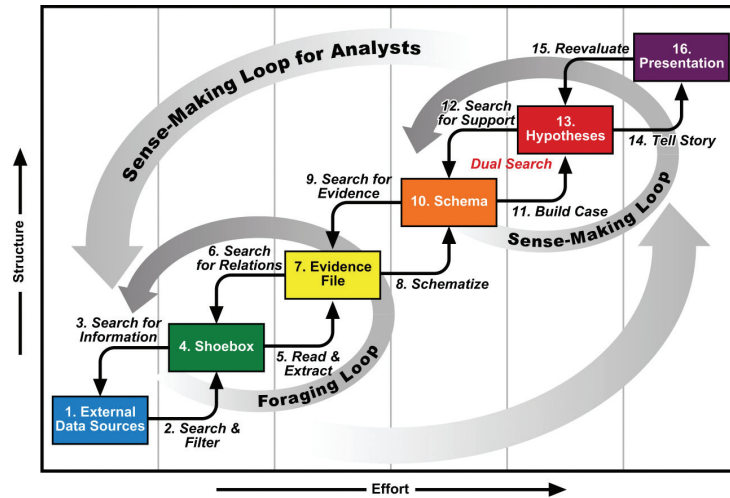


Figure 2.1: Sensemaking loop [282], image taken from [81].

Making sense of data is a cognitively demanding task even for experts. Moreover, in the context of information overload, sensemaking capabilities are hindered by the massive amounts of available data [210]. There are several areas of research that attempt to support the sensemaking process. For example, visual analytics tools have been developed to support specific steps of the sensemaking process [104, 138]. Moreover, machine learning can be used to process data and provide starting points for analysts [389]. However, deciphering the information encoded in text data is still an AI-hard problem [397].

2.2.2 Cognitive Strategies in the Sensemaking Process

Previous research has explored how analysts make cognitive connections between documents in the context of intelligence analysis tasks. For example, Bradel et al. [51] studied how analysts structure information in the context of intelligence analysis tasks, where they found layouts based on linear structures with branching and web-like structures. Our study also shares similarities with the work of Robinson [301], which focuses on analyzing the strategies and organizational methods used during collaborative synthesis, with the purpose of proposing a series of design guidelines for collaborative sensemaking systems.

Other similar work includes Andrews et al. [15, 16], who explore the workspace organization used by analysts in large displays to arrange documents, where most strategies consisted of clustering, although some analysts used timelines. In addition, Wenskovitch and North [380] study how analysts perform grouping and dimensionality reduction, where strategies included divide-and-conquer, incremental layouts, and bottom-up construction. Our work follows a similar approach, but focuses exclusively on the use of narrative maps as a sensemaking

tool, analyzing the different map construction strategies and the underlying graph structures generated during the process.

Previous studies have also found that analysts use strategies such as identifying co-occurrence relationships and aggregating common elements [144], using topical and temporal orderings for document clustering, and evaluating content overlap and similarity for document summarization [61, 106]. However, previous research has not focused on specific narrative sense-making tasks. In narratives, there is an underlying temporal ordering as well as a focus on cause-effect relationships, which leads to a specific description of cognitive connections and construction strategies for narrative sensemaking.

Finally, prior works have shown that graph-based narrative representations [186, 232, 322] are useful as a sensemaking tool. Thus, with the purpose of improving such narrative representations and their associated extraction algorithms, we seek to understand how analysts create such models from scratch by analyzing the narrative mapping process and its strategies.

2.3 Narratives in Visualization

In this section, we discuss the existing literature on narrative visualization. We start by giving a brief introduction to the basic concepts in narrative theory. Then, we present some related work on narrative extraction, representation, and visualization.

2.3.1 Storytelling with Visualization

Narratives are systems of stories interrelated with coherent themes [145]. These stories can be told in different ways, leading to a distinction between the story itself and how it is represented. Narrative studies attempt to understand the relationships between the underlying stories and their representations [1, 285]. In the context of information visualization, we explore how information narratives and storylines can be visualized. The narrative visualization pipeline requires extracting the narrative from data into an internal narrative representation, which is then used to generate the final visualization.

Storytelling and narratives are common metaphors in visual analytics [297, 316, 348]. In our application context, we are interested in extracting and representing narratives taken specifically from data sets of text documents, rather than from numerical or other non-text types of data. Thus, not all of the visual storytelling concepts apply to our work, as they are designed with other types of visualizations in mind. Nevertheless, the visual storytelling framework provides a series of useful definitions [316] as well as techniques and design patterns [297] that could prove useful toward our goal of designing better narrative maps.

There are multiple genres of narrative visualizations. Narrative maps—and other graph-based narrative structures—provide paths that the users can follow to understand the story, similar to how flow charts work. Thus, they fall into the flow chart genre of narrative visualization, as defined by Segel et al. [316]. Next, we consider the concept of messaging [316] in visual storytelling, which refers to the use of text to provide explanations and observations about the visualization. In terms of messaging, narrative maps make heavy use of text, as the events in the maps are described entirely by text (e.g., the article’s headlines) and annotations can be used to provide additional context for each part of the map. Finally, we note that interactivity [316, 348] is another important element of visual storytelling. However, for the purposes of this paper, we did not consider interactive narrative maps in the evaluation. The study of interactivity in the context of narrative maps is left as future work.

Narrative maps usually show multiple storylines that can be visualized at the same time. Therefore, according to the storytelling taxonomy of Tong et al. [348], narrative maps fall between the *narrative visualization for storytelling in parallel* category or the *narrative visualization overview* category. In this context, storytelling systems enable users to detect patterns, structures, or relationships in data, which can help users confirm hypotheses or gain additional knowledge about a specific topic [8, 348]. We note that it would be possible to construct a map as a single timeline, leading to linear storytelling. However, this would be a pathological case and not the typical use case of narrative maps.

2.3.2 Narrative Structure and Visual Design

Finally, we note that the underlying representation of the narrative guides the visual design. For example, timeline approaches visually present the resulting narrative in a linear fashion, and most do not require advanced visualization techniques. Structured approaches using trees or DAGs, in contrast, need more complex visualizations, such as information metro maps [322] or story forests [232]. Moreover, the different structures present trade-offs in terms of expressive power and complexity. For example, DAGs allow us to show divergent and convergent substructures, while trees only allow us to show divergent substructures. However, we still do not have a systematic evaluation of these different underlying structures. Thus, our work seeks to bridge this gap by exploring which one of these structures performs better in the context of narrative sensemaking.

2.4 Interactive Artificial Intelligence

In this section, we present the concept of interactive artificial intelligence [381] for human-AI collaboration. Interactive AI systems are based on two key components: an explainable artificial intelligence (XAI) component and an explainable cognitive intelligence (XCI) com-

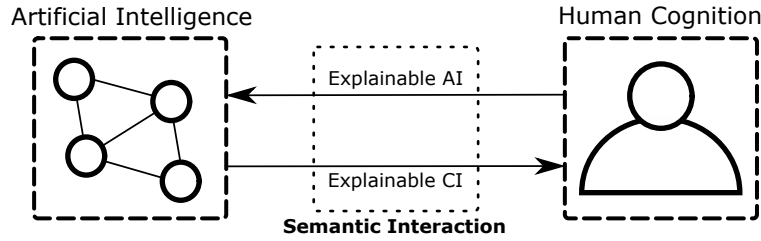


Figure 2.2: Interactive AI description adapted from [381]. The human cognition component and the AI component are connected through a semantic interaction system. The explainable cognitive intelligence (XCI) component transmits information from the human to the AI. The explainable artificial intelligence (XAI) component transmits information from the AI to the human.

ponent.

Explainable AI is an area that leverages machine learning and visualization techniques to convey information about the state of the AI systems to a human user with the goals of building trust in their output, demonstrating the uncertainty in the results, and contextualizing the results. Explainable CI focuses on modeling the cognitive state of the human user and converting the externalized cognition into a format that a machine can understand. For example, XCI can turn user interactions into a set of parameter or value updates that reflect the user’s intent. Based on these interactions, the AI component of the system can learn what the user needs and what is interesting to them [381].

Figure 2.2 shows the core elements of interactive AI systems. These systems enable humans to understand the results and analyses provided by the machine (XAI) and also enable the machine to adapt and learn based on human cognition (XCI). In the context of our work, we use *semantic interaction* to connect the human and machine components.

2.4.1 Semantic Interaction

Semantic interactions exploit the natural interactions within a visualization—usually a projection of the data in a lower dimension space—to learn the intent of the analyst. Then, the internal models and parameters of the system are updated based on these interactions [104]. Thus, instead of trying to manually modify parameters to model a specific concept, semantic interactions learn the parameters associated with that concept from user interactions and changes in the visualization. Semantic interaction has seen applications in text analytics [104], images [155], and other quantitative data with high dimensionality [317]. We now describe the specific elements of the semantic interaction pipeline.

First, we note that semantic interaction systems usually work using an interactive dimensionality reduction model. These models provide analysts with an interactive medium in the form of a spatial layout. Analysts can then modify this spatialization to express their pref-

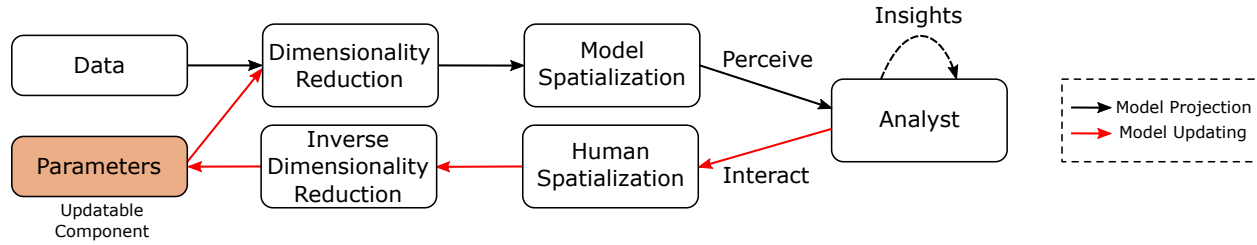


Figure 2.3: Semantic interaction pipeline adapted from [39]. The pipeline shows how the analyst and the visual analytics system work together to update the visualization and underlying model parameters through an iterative loop. The interactive dimensionality reduction component captures the intent of the analyst based on the modified projection (i.e., the human spatialization) and updates the projection in response (i.e., the model spatialization).

ferences, which the interactive dimensionality reduction attempts to learn. To do this, it is necessary to invert the dimensionality reduction and return to the original high-dimensional space. There are multiple machine learning models that attempt to solve the bi-directional transforms required to implement semantic interactions, such as Observation-Level Interaction [103], Bayesian Visual Analytics [162], and Visual to Parametric Interaction [217]. These works show how dimensionality reduction models can be inverted and used to capture the different interactions and modifications made by analysts. In particular, Weighted Multidimensional Scaling (WMDS) [313] is a popular model for dimensionality reduction in the semantic interaction literature.

Figure 2.3 shows the semantic interaction pipeline based on a dimensionality reduction model. Analysts use the visualized projection to get insights. Then, they make changes to this projection by changing the position of data points. By making these changes, analysts are expressing their preferences. Thus, based on these changes, the interactive dimensionality reduction model is able to learn the intent of the analyst behind the changes in the projection. Using this information, the changes are converted back into a high-dimensional space. Afterward, the modified dimensionality reduction model updates the projection based on the new high-dimensional representations.

Chapter 3

A Survey on Computational Narrative Extraction

Narratives are fundamental to our understanding of the world, providing us with a natural structure for knowledge representation over time. Computational narrative extraction is a subfield of artificial intelligence that makes heavy use of information retrieval and natural language processing techniques. Despite the importance of computational narrative extraction, relatively little scholarly work exists on synthesizing previous research and strategizing future research in the area. In particular, this chapter focuses on extracting news narratives from an event-centric perspective. Extracting narratives from news data has multiple applications in understanding the evolving information landscape. This survey presents an extensive study of research in the area of event-based news narrative extraction. In particular, we screened over 900 articles that yielded 54 relevant articles. These articles are synthesized and organized by representation model, extraction criteria, and evaluation approaches. Based on the reviewed studies, we identify recent trends, open challenges, and potential research lines.

3.1 Introduction

Narratives are fundamental to our understanding of the world [1] and they provide a framework that enables humans to associate and represent events over time [55]. Moreover, narratives are a core element of collaborative sensemaking in society [22, 386]. In this context, narratives are defined as a *coherent system of interrelated stories* [145], where stories themselves are defined as sequences of *events* [365]. These systems of stories help humans produce a shared understanding of the world [342]. In particular, extracting narratives from data is a fundamental task in our efforts to achieve this goal of common understanding [186].

In this survey, we focus on a specific type of narrative: news narratives. In particular, we analyze works that extract computational narrative representations from news articles. Work on general computational narratives started as early as the 1960s [306]. However, these early works focused mostly on narrative generation—usually through rule-based methods and grammars [10]—rather than extracting narratives from data. In contrast, the narrative extraction works reviewed in this survey start around the 2000s (e.g., [73, 265, 354]).

From an information retrieval standpoint, extracting narratives from data relies on several techniques from this field, including event [265] and entity extraction methods [46], as well as elements from search and ranking [226] and summarization techniques [207]. Furthermore, narrative extraction is supported by several artificial intelligence techniques, such as machine learning [343] and search and optimization [319].

Despite the importance of news narrative extraction, relatively little work has focused on clarifying the past trajectory and future agenda of news narrative extraction. Our goal with this survey is to fill this gap. This chapter presents a literature review of narrative extraction screening over 900 papers from a variety of journals, conferences, and workshops. In particular, by thematically analyzing 54 articles we identify a taxonomy of representations, extractions methods, and evaluation methods, that helps organize prior work and chart the path forward for future research. Taken together, all these elements provide a detailed account of the core elements of event-based news narrative extraction.

3.1.1 Scope of this Survey and Definitions

Narrative Definition

There are many potential definitions of narrative in the literature. General narrative theory focuses explicitly on understanding the general rules of narrative and its different arrangements that make it meaningful [1, 285]. The key intuition in formal narrative theory is that there is a distinction between the story itself and its representation. Narrative theory tries to understand the relationships between stories and their many possible representations [285]. Other definitions consider narratives as communication tools to construct a shared meaning of events with the purpose of influencing the behaviors [257].

Halverson et al. [145] define narratives not just as one story, but rather as a *system of stories*. That is, narratives are a systematic collection of interrelated stories with coherent themes. Stories are defined as sequences of *events* tied together in a coherent fashion. In this definition, events are the fundamental units of narrative action, they are either an act involving characters and entities or a happening where no entities are causally involved [1]. We leverage this definition to model news narratives. Thus, we have a series of hierarchical definitions starting from the *narrative*, then going into *stories*, and finally into the fundamental units of the narrative: the *event* and its related *entities*. Furthermore, these definitions require an underlying *order* for the events, as they have to be linked sequentially in the stories.

There are two fundamental units in our previous discussion: events and entities. These units provide different perspectives of the narrative, one is focused on the actions and happenings of the narrative, while the other is focused on the characters and other entities that participate in the events. However, to provide a more focused review we will focus exclusively on **event-based narrative representations**. Thus, we define computational narrative repre-

sentations as an *event structure* that represents different stories. We note that these event structures are discrete in nature (e.g., a graph or a timeline of events). Nevertheless, we note that some of the extraction methods that we review will leverage entity-based information, but they are not the focus of their representation.

Finally, we note that the simplest way to computationally represent a narrative is through a linear structure representing sequences of events (i.e., a timeline). In fact, this is the most common approach in our survey. However, we also find that there are more complex representations, based on event graph structures.

News Narrative Extraction

The main focus of this survey is on **narrative extraction** from *news data* (“How do we extract a news narrative from data?”). In particular, we focus exclusively on textual narratives extracted from a set of news articles published in traditional news sources—we exclude works that focus on mixed types of data (e.g., images and text, or videos and text). Thus, all of the surveyed works fall under the umbrella of Natural Language Processing.

Moreover, we note that extraction can be performed at a document level (i.e., extracting a narrative from a single document) or at a corpus level (i.e., extracting a narrative from multiple news articles). As part of our scope definition, we focus on corpus-level extraction methods, where the goal is to obtain a narrative representation from a set of articles, rather than on document-level extraction (e.g., extracting the narrative of a single document).

Throughout this work, we work under the assumption that most news articles focus on a single main event. This is a common assumption in story and narrative extraction methods [186] and a natural assumption when dealing with breaking news articles, as they are likely to present a single event [187]. We note that news articles may sometimes refer to previous or secondary events in their body, which can be used to link articles together. However, for the purposes of our definition, these references are not considered the *main event* of that news article. Following this assumption, we deal with three levels of resolution in our works: events as sentences, events as documents, and events as clusters. Events may be represented by relevant sentences extracted from a news article, usually, a single sentence is used for these purposes. Events may also be represented by an entire document (i.e., a news article). We note that there is some overlap between these two representations when documents are associated with headlines. Finally, events may also be represented as sets of documents that refer to the same main event.

We note that there are more granular views of events in the literature—for example, the notion of event from TimeML [255, 286], where events are a much more specific action (e.g., a perception or state) compared to a news event that may comprise multiple of these events [164]. Contrasting with the granular specifications of TimeML, there are also works that view events as sets of terms (e.g., keywords or entities) [340], akin to how topics are

sometimes characterized in traditional topic modeling works [43], and construct timelines representing them as such. However, this view of events is too broad and lacks the specificity expected from news events. Thus, we do not consider narrative representations that use such approaches. Following these exclusion criteria, we removed approximately 10 articles from the final data set.

Leveraging our previous discussion of narratives as a structured system of interrelated stories, we define the (event-based) narrative extraction task as follows:

News Narrative Extraction: Given a set of news articles, the news narrative extraction task generates a *discrete structure* comprised of *events* to represent the narrative.

We note that the structure is left deliberately ambiguous to allow for different types of representations, such as event timelines or event graphs. However, we note that all these overarching narrative representations are discrete in nature (e.g., event graphs), even if the underlying event representations could be continuous (e.g., text embeddings). Furthermore, the representation of the event itself can be defined in different ways depending on the *resolution* level (sentences, documents, or clusters) of the narrative representation. Furthermore, this definition excludes entity-based representations (e.g., character networks).

Exclusions: Related Tasks

We exclude works that focus on narrative generation, narrative forecasting, and narrative analysis. We also exclude works that only focus on representational issues without an associated method.

Narrative generation is a fundamentally different task from extraction that seeks to create new fictional narratives, rather than extract a narrative that already exists (either fictional or non-fictional) [130, 133]. Furthermore, the focus of narrative generation is usually fictional narratives, not news narratives. Narrative forecasting (i.e., predicting the next events in the narrative) is a task that lies between extraction and generation, but its focus is on generating new events rather than on extracting the complete narrative [411]. Narrative analysis methods use existing extraction approaches to obtain a computational representation of the narrative and then use it to analyze the narrative [272, 314]. However, they do not provide new insight into the extraction task itself, unless they include a novel extraction method as well.

Moreover, we exclude interactive narratives, as these are a fundamentally different type of narratives where the story can be changed through user feedback and actions [63], which would not make sense in the context of news narratives. However, while the underlying story cannot be changed, it might still be possible to interact with the narrative model. In fact, several works rely on interactivity at a presentation level.

Finally, we exclude works that focus on news narratives extracted from social media [29, 227],

as social media narratives follow a different approach that requires not only analyzing content but also the users spreading it, leading to unique challenges that are left beyond the scope of this survey.

Inclusion and Exclusion Criteria

Having defined our scope, we provide details of our collection methodology and inclusion/exclusion criteria. We describe the query and steps used to generate the final article set in Figure 3.1.

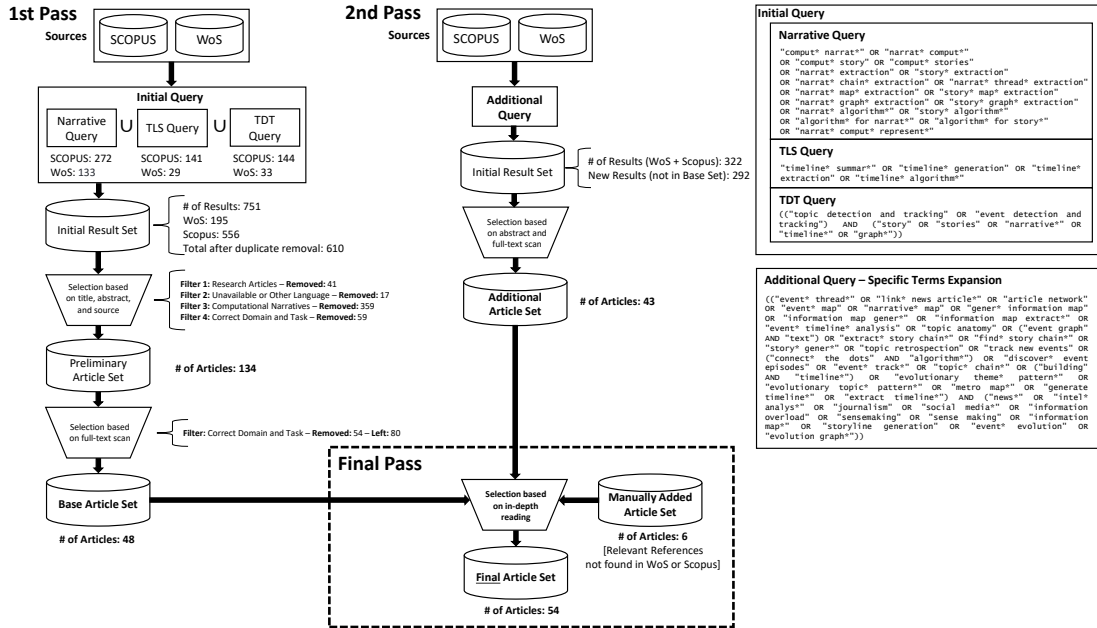


Figure 3.1: Overview of the article collection process and the inclusion/exclusion criteria used to construct the final article set.

We performed two article searches on SCOPUS and Web of Science. The first search was based on three queries that covered broad areas related to the news narrative extraction task: narrative extraction and computational narratives in general, topic detection and tracking (TDT) [13], and timeline summarization (TLS) [134]. These latter two fields are highly related to our task and provide a series of relevant works that we have examined in our survey. In particular, we note that most TDT works view news as flat collections [265] of events without an underlying narrative structure. Instead, we view news data as an interconnected *structure of events*. Nevertheless, some TDT works fit with our view of narratives and thus we include them in the review. In contrast, we consider most of the TLS line of works as a subset of the narrative extraction task and include many works from that field as part of the “event as sentences” resolution level. However, we exclude works that do not generate a full timeline and only focus on identifying relevant dates, as that is a different

sub-task. Next, we performed a second query based on a series of keywords obtained from the initial results. We applied the same inclusion and exclusion criteria for this second set of articles. After this, some additional articles that were not caught by our two main searches were added based on references from some reviewed articles. Finally, we performed a final pass on all the articles based on an in-depth reading of each article.

The rest of this chapter is structured as follows. The rest of Section 1 discusses related surveys and reviews. Section 2 presents an overview and summary of each one of the reviewed articles. Section 3 discusses the different extraction criteria. Section 4 presents a discussion of the evaluation approaches and metrics. Section 5 presents a discussion of our findings and future research directions. This survey concludes with a brief summary and key takeaways in Section 6.

3.1.2 Related Surveys

Most surveys regarding computational narratology focus on the task of narrative generation rather than extraction. In fact, there is an extensive series of survey papers and literature reviews on generation in conferences [355] and journals [133, 206] that cover narrative generation and its different approaches in-depth. Moreover, there is even a book [241] on computational narrative representations for narrative generation and an extensive and in-depth book chapter on different cognitive approaches to narrative generation [273]. Narrative generation is also covered as a specific subtask of the more general field of natural language generation [130]. In contrast, general narrative extraction is not covered by any published survey. More specifically, our domain of interest—news narrative extraction—is also not covered in the literature. However, there are some surveys that touch on related topics. In the rest of this subsection, we provide a general description of these works and how they relate to our own survey.

First, we note a survey on the evaluation of summarization methods by Ermakova et al. [110] as a related approach to narrative extraction. In particular, this survey provides a comprehensive overview of existing metrics for the evaluation of narrative summarization methods. Narrative summarization is related to both narrative extraction and generation, as it requires extracting an internal narrative representation from data and then generating the summary. In comparison, our survey presents evaluation metrics for narrative extraction methods, some of which overlap with the evaluation metrics discussed in the aforementioned survey.

Second, we note the work of Richards et al. [296] which discusses representation models for narratives. Most of the discussion is specific to narrative generation, but there are general models that could be applied in both generation and extraction contexts. Nevertheless, the discussion is focused on what constitutes a narrative in general rather than being directly useful for the news narrative extraction task as defined here.

Third, we note the survey on extracting character networks from fictional narratives by Labatut and Bost [209]. Their work is related to ours as it focuses on the narrative extraction task, but with a much more specific scope focused on character-based models (i.e., entity-based narrative extraction). In contrast, our work has a different scope that considers event-based models. Moreover, their scope focuses on extracting networks from fictional narratives, while we consider extraction methods for non-fictional narratives in news data.

Finally, we note the recent survey on timeline summarization approaches by Ghalandari and Ifrim [134]. While there is plenty of overlap between this survey and our own, the news narrative extraction task that we cover is more general than just timeline summarization, as we include methods that treat events as documents and clusters, rather than at a sentence level. However, we highlight the empirical component of that survey, which includes an experimental section comparing the state-of-the-art methods in timeline summarization.

3.2 News Narrative Extraction

3.2.1 Overview

We found a total of 54 articles focusing on event-based news narrative extraction in our review. We present the articles based on the resolution level that they use: events as sentences, events as documents, and events as clusters. Figure 3.2 summarizes the identified approaches categorized by event resolution and some relevant subsets of these categories. In the sentence-level resolution, *query-based approaches* include an information retrieval step in addition to the narrative extraction itself. For example, these approaches require the user of the method to define a search query (e.g., “COVID” or “Terrorism”) to find related documents in the data set through similarity-based techniques or other methods before extracting the narrative from the queried subset. In contrast, *pre-filtered approaches* assume that the data set has been already filtered and do not require an explicit query. In the document-level resolution, *Connect the Dots* approaches refer to the line of works derived from Shahaf and Guestrin’s seminal method of the same name on storyline extraction [319]. In the cluster-level resolution, event threading and evolution methods refer to a series of works based on Nallapati et al.’s *event threading* concept [265] or Yang et al.’s *event evolution* concept [400]. Works that fall under the “Others” do not fit in any of the defined subsets.

Table 3.1 summarizes the reviewed articles. In particular, we include the following columns in this table: *event resolution*, *number of stories*, *structure*, *type of approach*, and *event representation*. We now provide a brief description of these elements and their possible values.

Event resolution refers to the abstraction level at which the events are extracted. As mentioned in the scope definition, we consider three levels: sentences, documents, and clusters. Sentence-level works represent events as either a single sentence (e.g., the most important

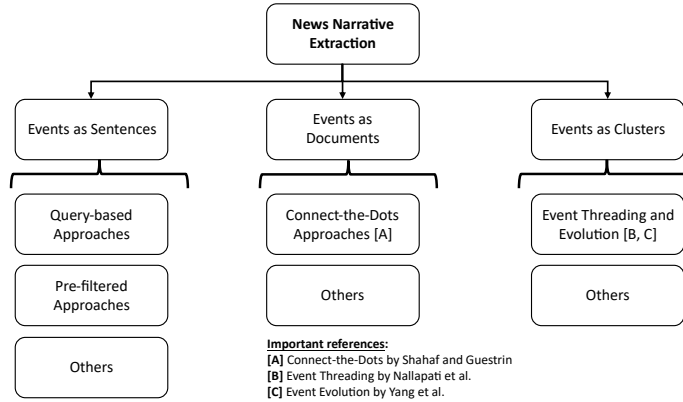


Figure 3.2: Overview of the different methods used in news narrative extraction categorized by event resolution.

sentence or a headline) or a set of sentences (e.g., a sample of representative sentences). Document-level works represent events directly as a single document (e.g., a full news article). Cluster-level works represent events as sets of documents (e.g., multiple news articles that talk about the same basic event). Structure represents whether the extraction method generates a linear structure of events (e.g., a timeline [319, 398]) or a graph-like structure (e.g., a directed acyclic graph [186] or tree [232]). Figure 3.3 exemplifies these concepts.

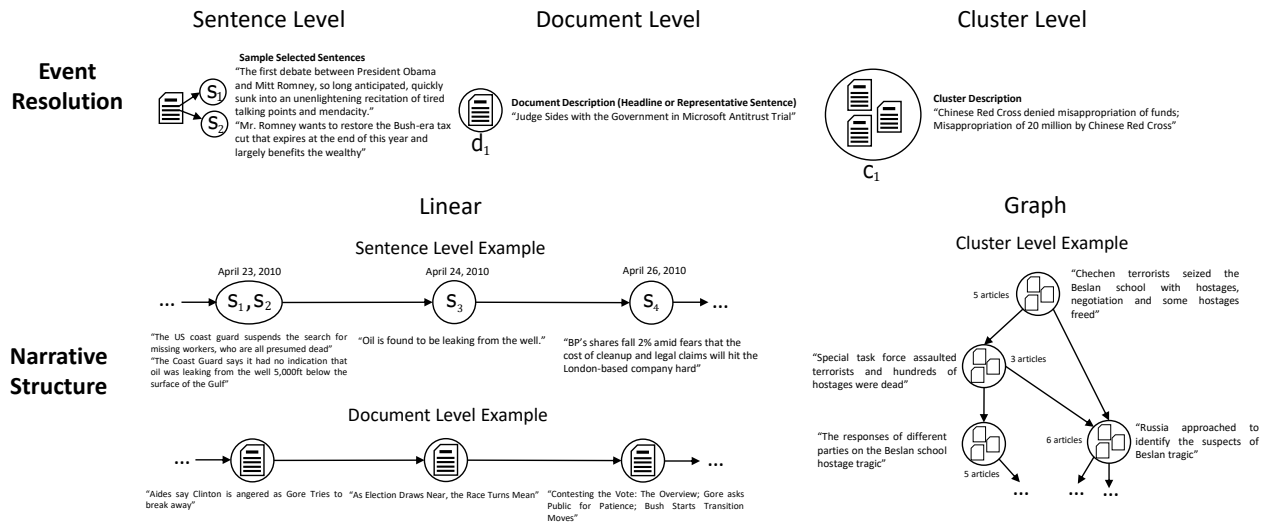


Figure 3.3: Resolution level and narrative structure. Examples adapted from several works in this survey.

Number of stories refers to whether the method is designed to handle a single storyline or multiple storylines. Recall that our definition of a story as a sequence of events. Most time-line extraction methods extract a single story, but some of them extract *parallel* timelines, where each timeline represents a different story from the data [208, 407]. In contrast, most

graph-based works are designed to represent multiple storylines, due to their inherent more complex nature compared to timelines. However, there are some works that represent a single story, but provide extra information by exploiting graph structures. For example, appending additional nodes with related events to the central story [229].

Type of approach represents whether the method is supervised—which requires training data—or unsupervised—which does not require training data. In general, we considered any method where the authors had to train the model with labeled data before using it as supervised. However, some approaches only did this to find the optimal value of a small set of hyperparameters [221, 265, 399] and it could be possible to use them in an unsupervised manner, provided that those hyperparameters were fixed in some other way (e.g., heuristics or previous work information).

Finally, *event representation* provides information about the computational representation of the events. Note that this is separate from the resolution level of the event. In general, we found four types of representations: word frequency models (e.g., TF-IDF and Bag of Words vectors), topic distribution models (e.g., Latent Dirichlet Allocation (LDA) vectors), neural embeddings (e.g., BERT), and entity-based models (e.g., entity frequency vectors). Some works combine these approaches and have a mixed event representation that leverages all these elements in some way to extract the final narrative model. There are some works that did not fit in any of these approaches and were marked as “Other”.

3.2.2 Events as Sentences

We start with works that use a sentence-level resolution. Most of these methods fall under the umbrella of timeline summarization [134]. However, not all of them fit with traditional TLS work. We split the discussion into three parts: query-based approaches, pre-filtered approaches, and others.

Query-based Approaches

These approaches perform an information retrieval step before or during the narrative extraction process based on a user-defined query. In some cases, the query just acts as a simple filter, in others, they explicitly include the query into the narrative extraction model.

Chieu and Lee [73] present a query-based timeline extraction approach where each event is represented as a sentence. This is the earliest form of the “events as sentences” that we could find in the literature. Sentences are first filtered based on the query and then ranked according to two criteria: *interest*—based on the frequency of the reported event in the query—and *burstiness*—based on the idea that important events form clusters around their date of occurrence. To determine whether two sentences are reporting the same event, the authors use cosine similarity. Furthermore, interest is determined based on a time window to

Table 3.1: Summary of the surveyed articles.

Year	Reference	Event Resolution			# of Stories		Structure		Approach		Event Representation				
		Sentences	Documents	Clusters	Single	Multiple	Linear	Graph	Unsupervised	Supervised	Word Vectors	Topic Distribution	Neural Embeddings	Entities	Other
1998	Uramoto and Takeda [354]		x			x		x		x	x				
2004	Nallapati et al. [265]			x		x		x			x				x
2004	Chien and Lee [73]		x		x		x		x	x	x				
2005	Guha et al. [139]		x			x	x		x		x				
2006	Yang et al. [400]			x		x		x			x				
2006	Lin and Liang [228]			x	x			x			x				
2007	Lin et al. [230]			x				x			x				
2008	Chen and Chen [68]			x		x		x			x	x			
2008	Qiu et al. [290]			x		x		x		x	x				
2008	Lin and Liang [229]			x	x			x			x				
2009	Yang et al. [401]			x		x		x			x				
2010	Shahaf and Guestrin [319]		x		x		x		x		x				
2011	Yan et al. [399]	x			x		x			x	x				
2011	Yan et al. [398]	x			x		x			x	x				
2011	Hu et al. [163]	x			x		x		x		x	x			
2011	Khurdiya et al. [197]			x		x		x			x				
2012	Zhu and Oates [416]		x		x		x		x		x				
2012	Chen and Chen [69]			x		x		x			x				
2012	Shahaf and Guestrin [320]		x		x		x		x		x				
2012	Shahaf et al. [321]		x			x		x			x				
2013	Binh Tran et al. [42]	x			x		x			x	x				
2013	Li and Li [218]	x			x		x			x		x			
2013	Tran et al. [350]	x			x		x			x	x				
2013	Huang and Huang [167]	x			x		x		x			x			
2013	Tannier and Moriceau [343]		x			x		x		x	x				
2013	Shahaf et al. [322]		x			x		x			x				
2013	Shahaf et al. [323]			x		x		x			x				
2014	Nguyen et al. [268]	x			x		x		x		x				
2014	Zhu and Oates [417]		x		x		x		x		x				
2014	Huang et al. [166]			x		x		x			x	x			x
2014	Wei et al. [377]			x		x		x			x				
2014	Hu et al. [164]		x			x		x				x			
2014	Zhou et al. [414]	x				x		x			x				
2015	Tran et al. [349]	x					x			x	x				
2015	Li et al. [221]	x					x			x					
2015	Bögel and Gertz [46]		x					x			x			x	
2015	Chen et al. [71]	x					x		x		x	x			x
2015	Shahaf et al. [324]			x		x		x			x				
2017	Wu et al. [391]	x			x		x		x		x				
2017	Liu et al. [332]			x		x		x			x				
2017	Lahan and Heurst [208]		x			x		x		x					
2018	Wang et al. [368]	x				x		x			x				
2018	Tikhomirov and Dobrov [347]	x				x		x			x				
2018	Xu and Tang [394]			x		x		x			x				
2018	Zhou et al. [415]	x				x		x			x				
2019	Camacho Barranco et al. [60]		x			x		x			x	x			
2019	Cai et al. [59]			x		x		x			x				
2019	Yuan et al. [408]	x				x		x					x		
2020	Duan et al. [98]	x				x		x			x		x		
2020	Liu et al. [233]			x		x		x		x	x				
2021	La Quatra et al. [207]	x				x		x		x					x
2021	Yu et al. [407]	x				x		x		x					
2021	Liao et al. [226]	x				x		x		x					
2021	Keith Norambuena and Mitra [186]		x			x		x		x					

avoid combining events that should be separated due to their temporal distance. To reduce *redundancy*, duplicated sentences are removed based on a time window around an important event that depends on the interest value.

Yan et al. [399] proposed a timeline summarization method based on balanced optimization and iterative substitution of sentences. Their optimization problem is defined in terms of *relevance*, *coverage*, *coherence*, and *diversity*. All these terms are based on the Kullback-Leibler divergence (KLD) [204] of the summary items with a target distribution. Relevance is related to a user-defined query and is defined as the KLD between the summary items and the internal representation of the query. Coverage is based on a global term—KLD between the summary items and the whole corpus—and a local term—KLD between the summary items and the set of sentences from the same date. Coherence is defined locally, based on the KLD between each summary item and its neighboring summaries by using an exponential temporal decay term (i.e., consecutive dates should have relatively similar summaries). Diversity is measured across dates and measures the average KLD of each sentence with respect to all other sentences in a leave-one-out manner. The final utility function is a weighted average of these terms with user-defined weights and can be defined

at a local level (to evaluate individual time periods) and a global level (to evaluate the full timeline). To find the sentences, this utility function is optimized in an iterative manner by replacing sentences in the date summaries and improving the utility value in each step using a dynamic programming algorithm that considers both local and global constraints.

Li et al. [218] propose a topic modeling approach for timeline extraction from news called Evolutionary Hierarchical Dirichlet Process (EHDP) to capture the evolution pattern of news topics. This model extends Hierarchical Dirichlet Process models [344] by incorporating time dependencies and background information. In particular, it adds a new dynamic Dirichlet mixture model. Using this proposed topic model, a series of sentences are selected to represent each time period in the timeline based on the weighted average of three criteria: *relevance* (the summary should be related to the overall query), *coverage* (the summary should generalize the important topics in each time period), and *coherence* (each summary should be coherent with neighboring time periods). To score these criteria, the authors propose a topic scoring algorithm based on KLD that leverages their new topic model. The selected sentences are used to represent the relevant events in each time period.

RaRE (Rank and RErank) [268] is a system for building timelines of events from news articles based on a user query. In particular, it extracts timelines in three steps: temporal clustering based on salient dates, event relevance, salience scoring, and sentence re-ranking using an iterative algorithm that seeks to reduce redundancy. The method has an underlying assumption that each document represents a single event that can be described by a single sentence. The temporal clustering step identifies salient dates based on the *number of occurrences* of the date in the documents. The sets of events linked to a specific salient date are called *temporal clusters*. Furthermore, as a preprocessing step, events are clustered into *thematic clusters* inside each date using hierarchical clustering based on normalized Manhattan distance and a user-specified threshold. The event relevance and salience scoring steps use these criteria to rank events (i.e., documents) inside each temporal cluster. In particular, it uses four metrics: *event relevance*, *thematic cluster relevance*, *event salience*, and *date salience*. Event relevance is based on cosine similarity with the initial query. Thematic cluster relevance is based on the similarity of its thematic cluster with the initial query based on the average relevance of each event in the cluster. Event salience is based on the frequency of terms on a specific date. Date salience is based on the (normalized) total relevance of all events happening on that date. Finally, the sentence re-ranking step measures the frequency of unused terms on each date for a specific event to reduce redundancy.

Another topic modeling approach uses a time-dependent Hierarchical Dirichlet Tree Model [221] to capture the evolution of news topics using the Dirichlet Tree distribution—a generalization of the Dirichlet distribution [90]. In particular, the model represents topic distributions in sentences using a tree of fixed depth. Each sentence is associated with a path and with a topic vector and each node has its own topic distribution over words. Using the proposed topic model, sentences are selected by first locating candidate words on the nodes of the tree based on the Jensen-Shannon (JS) divergence of sentences and KLD between word collections. Next, the candidate sentences are scored based on the weighted average

of the following criteria: *focus* (the timeline should be relevant to a given query), *coherence* (the sentences should be correlated), and *coverage* (the sentences and documents should be representative).

Wu et al. [391] propose a sentence-based approach to generate timelines. In particular, all the sentences that contain a user-defined query word are split by date and used to generate a *date vector* representing that specific date. Sentences that do not include parseable dates are grouped based on *similarity* with the date vector. All sentences are then ranked based on similarity with their corresponding date vector and unrelated sentences are filtered out based on a user-defined threshold. The highest-ranking sentence is used to summarize each date.

Tikhomirov and Dobrov [347] propose a news timeline generation approach from a query based on three steps: query extension, inter-document graph extraction, and intra-document sentence ranking. Query extension is based on *pseudo-relevance feedback* and consists of three query levels, which are constructed using the most significant terms based on TF-IDF weights. Next, as a preprocessing step, dates that have a *frequency* below a statistically determined threshold are discarded. The next two steps use an *inverted pyramid* [187] heuristic, which assumes that the upper part of the article contains the most important information and the lower part of the article may contain *references* to important events from the past. In particular, the inter-document graph extraction step constructs a *similarity matrix* between the upper and lower parts of the documents. If the similarity is above a specified threshold, then the articles are considered to be linked, creating a similarity graph. Next, a ranking algorithm—LexRank [109]—is used to determine the importance of each document. Documents that are above a specified importance threshold are used to further expand the original query one more time. Finally, to rank the final selected sentences for the summary, a ranking metric is defined by taking into account content similarity (using cosine similarity) with the extended query (i.e., maximizing relevance) and subtracting similarity with already extracted sentences (i.e., minimizing redundancy).

WILSON (neWs tImeLine SummarizatiON) [226] is a query-based timeline summarization method for news based on a divide-and-conquer approach consisting of two major components: date selection and text summarization for each selected date. For date selection, the method first tags temporal expressions in sentences and constructs a date reference graph based on these annotations. Next, the method assigns weights to the edges of the date reference graph by taking the product of the *number of references* and *temporal distances* with the references. Then, it uses the PageRank algorithm [275] on the extracted graph to find the most salient dates. However, this approach leads to a bias towards older dates, as they have had more time to get references. Thus, the model is augmented with an exponential *recency adjustment* weight, which is used to initialize the Personalized PageRank algorithm [24], which allows for non-uniform initial distributions. Next, the daily summarization can be done using any multi-document summarization approach. Specifically, the authors use TextRank [252] based on BERT [91] representations to generate the summaries.

Pre-filtered Approaches

These approaches assume that the data set has already been filtered as part of a pre-processing step. Thus, they do not explicitly model the query in their extraction model.

Yan et al. [398] propose a system to generate news timelines using a trans-temporal summarization approach, where the summary for each time period depends on its context—that is, nearby time periods. Before generating the timeline, the system chooses the important time periods (e.g., specific days) to be summarized based on *burstiness*. The timeline extraction approach is based on two components: a *global component*—which defines the structure of the overall summary and the inter-temporal relationships between each period of the timeline—and a *local component*—which defines the summary in each time period. The global component is based on a global graph that uses inter-date dependency, which is computed using *temporal proximity* and a global *affinity* model for each sentence based on PageRank. Furthermore, to ensure a diverse set of sentences in the global component, the system incorporates DivRank into the affinity model [247] to penalize the lack of *diversity* in the sentence selection. Next, the local component is based on a local sentence graph for each time period following a similar approach to the global graph. To generate the final sentence selection in each time period, the system optimizes a weighted ranking generated by both components.

Hu et al. [163] propose a timeline overview method for news based on the concept of breakpoints—points in time where a significant development or change occurs (i.e., important events). Their extraction approach consists of three steps. First, they analyze *topic activity* using a Topic-Activeness Hidden Markov Model (HMM) and discard inactive periods. In practice, this is done by measuring whether there is new information using KLD and document frequency. Next, the breakpoints are identified by detecting topic variations in each time period using a topic mixture model—in particular, a generative probabilistic mixture model [246]—and a Theme-Transition HMM model to model topic evolution. Specifically, breakpoints are identified by using JS divergence to measure *topic variation* between two consecutive time points. Then, a summary for each breakpoint is generated by selected representative sentences—based on Jaccard *similarity with topic keywords* and *relevant entities*.

Tran et al. [42] present a supervised learning method to extract timelines from news articles based on Linear Regression. Their model first identifies salient dates based on *burstiness* (i.e., high-frequency periods), and then selects the most representative sentences from the news articles on each of these dates. In particular, the model uses *surface-level features* (e.g., length and position of the sentences), *coherence features* (e.g., causal and temporal signals), *topic features* (e.g., TF-IDF information and cross-entropy), and *time-related features* (e.g., popularity over time and use of temporal expressions) to determine the key sentences of each date. Subsequent work by Tran et al. [350] used SVM-Rank instead of linear regression and expanded upon this supervised framework. In particular, they leverage three metrics to evaluate the event sentences: *relevance*, *novelty*, and *continuity*. Relevance is learned using

the SVM-Rank mentioned earlier. Novelty is evaluated by measuring the non-overlapping n-grams over the total n-grams between a candidate sentence and previously selected sentences. Continuity is a measure of local coherence—there should be smooth transitions in the timeline—that is computed as the average n-gram overlap of all sentences in the current day with the previous summary. The final score is based on a weighted average of these metrics. To learn the relevance function, the authors leverage the same set of features from their previous work [42], but they also added an extra set of features about the event itself. For example, they evaluated whether the sentences properly represent the main event of the article, using the fact that the first sentences should contain the most relevant information (following the *inverted pyramid structure*). Thus, they evaluate the similarity between the sentence summary and the first four sentences. Once the SVM-Rank method was trained, the ranking is fed to a dynamic programming approach to optimize the final score.

Huang and Huang [167] present an event storyline generation method based on a mixture-event-aspect probabilistic model that can detect and distinguish the different types of sub-events in the article data set. Their model is an extension of Probabilistic Latent Semantic Analysis [156] and LDA [44]. In particular, their model detects *global* aspects (i.e., terms that are important throughout the whole story) and *local* aspects (i.e., terms that are important in a specific event inside the story). Based on the extracted aspect model, the *bursty* periods for each aspect are extracted to measure their popularity on a certain date and detect relevant events. Based on these results, it is possible to extract a timeline and select the most representative sentences associated with both global and local aspects to compose the final storyline with adjustable weights for the aspects. Sentences are selected by minimizing the overall *information loss* over each aspect. In particular, the LexPageRank algorithm [108] is used to rank sentences and KLD is used for sentence similarity.

Tran et al. [349] propose a timeline summarization approach based on article headlines. Their approach is based on a random walk model using a topic-sensitive version of PageRank [153] that selects relevant headlines from the data set for each time period. There are three key metrics to evaluate the relevance of a headline: *informing value*, *spread*, and *influence*. The informing value depends on whether the headline provides factual information or an opinion, review, or another non-informing category. It is a binary value computed using a supervised learning approach based on an SVM classifier to separate facts from opinion [406]. Influence tries to measure the impact of an event on future events (e.g., “the president resigns” would lead to a “new election” event) based on references from future events using similarity between future news articles and the headline of the event. Spread is based on the intuitive idea that a relevant event will be reported in multiple news outlets—that is, its reporting will be spread over multiple headlines. Thus, it is a measure of *positive* redundancy and it is formally defined as the probability of a headline being duplicated. To estimate whether two headlines are duplicates, the system uses a supervised logistic regression model trained on semantic similarity measures based on paraphrase detection literature [253]. Having defined these elements, the goal is to maximize all three aspects to select the best headlines. This is done by using PageRank on a graph of headlines, taking

into account both spread (graph edges) and influence (random walk probability), to generate the final rankings. Next, to generate the final timeline for each day, the resulting rankings are selected greedily, subject to redundancy constraints, informativeness constraints, and a maximum number of headlines per day.

Chen et al. [71] present a supervised timeline summarization algorithm based on aging theory for news data sets. Aging theory [70] is a model that tracks the life cycle of events using an energy function, which increases when an event becomes popular and diminishes with time. The method works by extracting sentences (i.e., specific events) and the publication time from news articles and using a classification model built with SVM to determine whether they belong in the output timeline. This approach is based on *surface-level features* (e.g., noun frequencies and stop word frequencies), *importance features* (e.g., latent semantic analysis scores), *topic features* (e.g., topic word frequencies), an *aging score feature* (i.e., changing coverage of an event over time), and a *novelty feature*. The aging score is used to measure the life cycle of each term over time using a recurrence relation with TF-IDF representations. The novelty score is based on the Jaccard similarity of the current summary and the candidate sentence.

Others

Here we present works that use a sentence-level resolution but differ from the majority of the other works that follow the traditional timeline summarization approach. In particular, we consider works on extracting disaster storylines from news and works that present variations on the traditional TLS task.

Disaster Storylines Zhou et al.’s works [414, 415] present a framework to construct spatio-temporal storylines for disaster management from news data based on how the disaster location moves over time (e.g., a typhoon moving through different areas). This approach generates timelines for two levels of representation: a *global level* that follows the progress of the disaster through each location and a *local level* that focuses on a specific location. To extract the storyline, a series of snippets (i.e., event sentences) are extracted from the news articles using named entity recognition methods and grouped together based on a similarity graph. Then, a set of representative sentences is selected by finding the minimum dominating set [326] using a greedy algorithm. Next, an integer linear programming approach is used to select the optimal sequence for the main route of the disaster by maximizing the *coherence* of the story chain, subject to a series of *structural*, *chronological*, and *length* constraints. In this method, coherence is defined based on consecutive content similarity rather than word influence. However, the key difference is that this formulation includes a *smoothness* constraint, which is specifically designed to track the moving location of disasters through time. Smoothness is based on simulating the natural trajectory of a disaster. In particular, the constraints set a maximum distance for consecutive events (i.e., avoiding jumps to locations too far away) and seek to avoid acute angles that could be formed by two

consecutive connections (i.e., avoiding sharp turns in the trajectory of the disaster). Once the main storyline has been constructed, the next step is to analyze the local level storylines. For each main storyline event, a set of similar articles are selected and used to construct a multi-view graph that represents the event relationships based on content similarity. Then, a Steiner tree algorithm is used on the multi-view graph to generate a local storyline for that location.

Yuan et al. [408] propose dTexSL, a disaster storyline extraction approach that extends Zhou et al.’s works [414, 415]. Unlike the previous approach, the news articles are first divided into different subsets based on location and are represented using neural embeddings. Locations are found by measuring the distance of the locations described in each article—using named entity recognition to find location references—and merging locations that are close enough based on a user-defined threshold. Then, an integer linear programming approach is used to select the *key locations* (i.e., document clusters). Instead of choosing events to maximize coherence like before, the goal is to maximize the *number of documents covered* on the map. The model has similar constraints as the original approach: *chronological order*, *length*, and *smoothness*. Once the main storyline has been constructed, a word embedding method is used to construct a multi-view graph that represents the event relationships based on content similarity. Using this graph, a set of representative articles are selected based on two criteria: *uniqueness*—computed using information gain—and *relevance*—computed using a measure of node importance. Then, a dynamic Steiner tree algorithm is used on the multi-view graph to generate a local storyline for that specific location. Finally, a traditional multi-document summarization method [137] is applied to generate a high-level event description for that specific location.

Task Variations Duan et al. [98] introduce another variation on the timeline summarization task called comparative timeline summarization. In this task, the goal is to provide timelines consisting of major contrasting events from two data sets. Their approach is based on three core characteristics: *coverage*, *distinctness*, and *diversity*. Coverage is based on the idea that the timelines should cover most of the important information or topics from each data set. Distinctness is based on the idea that the events in a timeline should be distinct from the events on the other timeline at each time point, to allow for a proper contrast between them. Diversity is based on the idea that each timeline should cover a diverse set of events from its data set. To model these attributes, the authors propose a dynamic Markov model that is built around sentence similarity at a document level for each time step. In particular, sentences are selected from news articles to describe events based on local and global importance measures through the use of an affinity-preserving mutually reinforced Markov random walk model based on the PageRank algorithm. The output is a timeline that contains contrasting events from both data sets.

Yu et al. [407] propose a variation on the basic timeline summarization task, called Multi-TimeLine Summarization (MTLS). In this task, events are represented as sets of sentences and computationally represented by the neural embedding model sentence-BERT [294]. Given a set of time-stamped news articles, MLTS seeks to automatically extract timelines

for important and different stories found in the data set. The authors propose a framework to solve this task called 2SAPS (Two-Stage Affinity Propagation Summarization). There are two key components in their framework: an event generation module and a timeline generation module. The event generation module seeks to extract important events from the document collection. To do so, it uses an *affinity* propagation approach to cluster similar sentences [122] and to identify the event of the article and any other previously *referenced* event. Furthermore, there is a *temporal similarity term* that uses an exponential decay function to penalize similarities of events that are temporally far away. Once the events are identified, a subset of these events is selected based on a weighted average of a *salience metric*—based on event frequency—and a *consistency metric*—based on the intra-event similarity. Next, the timeline generation module has three internal steps: event link, time selection, and timeline summarization itself. Event linking is based on the weighted average between a *co-reference score* (based on entities or terms shared between events) and *semantic similarity* (e.g., cosine similarity). Based on these average scores, the system builds an event graph and uses affinity propagation on it to determine the initial clusters (i.e., timeline sets). Next, there is a timeline selection based on the weighted average of *timeline salience*—the average event salience of the timeline—and *timeline coherence*—the average semantic similarity scores between chronologically adjacent events. The timeline summarizing step selects an exemplar sentence for each event in the timelines, as the most typical and representative member of each event. Finally, there is an add-on timeline tagging step which assigns a label to each timeline, based on the most frequent words of the events.

Summarize Dates First (SDF) [207] is a timeline summarization pipeline that follows a different paradigm for timeline summarization based on generating a summary for each individual date first, and then selecting the most relevant dates using these summaries. This is different from the traditional approach where the relevant dates are selected first. Furthermore, this approach aggregates dates by leveraging higher-level temporal references (i.e., references to previous events in the article). SDF consists of three steps: temporal tagging, per-date summary extraction, and summary-drive date selection. In the temporal tagging stage, the raw text is annotated to identify date-level references (e.g., 31 December 2021) and high-level references (e.g., last December). The per-date summary extraction step uses any traditional sentence-based summarization algorithm from the multi-document summarization literature (e.g., TextRank [252]). Summary-driven date selection is the last step and uses a selection strategy, called Graph-Based Date Selection, which uses graph ranking algorithms (e.g., PageRank, HITS). In particular, a directed date graph model is built using the temporal references of the data set, where the edge weight connecting two dates is influenced by the count of date-level references and the similarity between the date summary and the high-level references to the earlier date.

3.2.3 Events as Documents

Here we present works that use a document-level resolution. We split the discussion into two parts: methods that build upon the Connect the Dots approach by Shahaf and Guestrin [319]—a seminal work in the field of news narrative extraction—and others. We further divide the presentation based on whether the methods are linear or graph-based. We note that the works cataloged as others did not have a discernible pattern beyond using a document-level resolution.

Connect the Dots Approaches

Linear Representations Shahaf and Guestrin [319] proposed the Connect the Dots algorithm to extract temporal chains of documents (i.e., timelines). In particular, they use an optimization approach that seeks to maximize the overall *coherence* of the timeline. Coherence measures the smoothness of a storyline, a coherent story should not have drastic changes in content or topic. To implement this metric, they propose an approach based on *word influence*—a measure of word relevance computed through random walks on a word-document graph—and *word activations*—which measure whether a specific word is active at a given point in the storyline. To extract the story chains, they used linear programming to maximize coherence subject to structural and temporal constraints. However, since linear programming provides non-integer solutions, it required additional heuristics to find the best chain by defining a rounding method. The linear programming approach used in the original Connect the Dots implementation was computationally expensive. Thus, Shahaf and Guestrin [320] proposed a new method to reduce computational costs and avoid the approximate solutions from the linear program. In particular, they used a best-first search algorithm based on an extension heuristic—given a chain of documents, adding a new document to the chain will at most keep the same level of coherence—and the original linear program to individually evaluate each chain.

Expanding upon the Connect the Dots method, Zhu and Oates [416] propose an algorithm to extract story chains from newswire articles that connect two user-defined endpoints based on the following characteristics: *relevance* (the articles on the chain should be relevant to the endpoints), *coherence* (the transition between events should be smooth), low *redundancy* (there should only be one representative article for every event of the chain), and *coverage* (the chain should cover every important event). To compute measures of these characteristics, the article proposes using random walks on bipartite graphs formed by articles and words, where the weights are given by TF-IDF representations. Thus, based on these criteria, their proposed algorithm consists of two iterative stages. The first phase consists of a divide-and-conquer bisecting search problem that adds articles to the story chain. In particular, in this phase the algorithm finds the best article to insert in the middle of each current link of the story chain (i.e., it bisects the current links) based on coherence and relevance criteria. The second phase consists of pruning redundant articles—by removing a certain

percentage based on how much coherence they would add to the current story chain—and irrelevant articles—by removing events that are similar to each other and temporally close with an exponential decay function. These phases are repeated until there are no more articles to add or prune. A subsequent article by the same authors [417] revisits the story chain algorithm and extends this approach by adding an intermediate clustering step that groups documents into document clusters and words into word clusters. These clusters are used to generate a new bipartite *correlation graph* that combines the weight of individual documents and words through a weighted average to assign the edge weights. Furthermore, the model adds a named *entity bias* that assigns a higher weight to named entities compared to other terms. This is modeled through a co-occurrence frequency matrix for entity pairs, which is then used to compute a relevance score for each document in the data set based on the named entities. In turn, these elements are used to modify the cluster and document weights in the correlation graph.

Camacho Barranco et al. [60] propose a storyline extraction algorithm that takes a set of user-defined articles as a seed and generates a timeline of articles based on a series of evaluation metrics. First, the authors propose a temporal criterion to filter candidate documents based on a range between the latest publication date of the seed articles and a maximum threshold away from the earliest publication date of the seed articles (i.e., in the interval $[t_{\min} - t_{\text{threshold}}, t_{\max}]$). Next, there is a topical criterion that measures how much a candidate article can deviate from the seed articles based on KLD and LDA topics. Having defined their basic framework, the authors then formalize an optimization problem to extract the storylines by selecting article connections based on different criteria: *incoherence*, *similarity*, *overlap*, and *uniformity*. Incoherence is based on the average pairwise Soergel distance between documents—measured using TF-IDF information for the entities of the document—with a temporal factor to penalize temporally distant articles. Similarity is used as a penalty factor to enforce diversity in non-adjacent articles of the storyline, implemented as a negative exponential factor based on the Soergel distance. Both of these metrics are weighted by a relevance factor of the documents and are smoothed using modified Gaussian distributions to measure event overlap. Next, an overall overlap factor for the storyline is computed, assigning a penalty based on the difference between publication dates and a user-defined threshold. The overlap factor ensures that the breakpoints occur at sufficiently distinct dates. The uniformity penalty seeks to avoid the case where the optimal solution selects purely irrelevant events as optimal by penalizing uniform weights. The objective function to minimize consists of the sum of the product between incoherence and similarity, multiplied by the overlap and uniformity penalties.

Graph-based Representations Metro Maps [321, 322] are an extension of the Connect the Dots approach that represents more than a single storyline using a directed acyclic graph of events. In particular, the metro maps method is a structured summarization approach that captures the evolution of multiple stories and their interactions. The stories are represented using a metro map metaphor, where each metro line represents a story and stations represent key events. Metro lines intersect in specific stations, representing how storylines

connect with each other. This representation is extracted by solving an optimization problem. In particular, the goal is to maximize *connectivity*, subject to *coverage* and *coherence* constraints. Coverage is computed based on how well specific terms or keywords are represented in the selected events and is defined using a *submodular function* that encourages diversity (e.g., if a term is already covered, adding a document that covers it provides little extra coverage). These keywords depend on the specific corpus or domain of application. Coherence is defined following Shahaf et al.’s previous work [319, 320]. Finally, connectivity is defined as the number of stories that intersect which is used to ensure that the final metro map is connected. The optimization problem is solved in phases. First, a series of coherent candidate metro lines are selected based on a divide-and-conquer approach, which constructs long lines from shorter ones and encodes them in a graph. Then, the method extracts a set of coherent lines that maximize coverage using an approximation algorithm based on the submodularity of the coverage function (otherwise finding these lines is an NP-hard problem). Finally, connectivity is increased using a local search approach that substitutes lines without sacrificing coverage.

Similar to the metro maps metaphor, the Narrative Maps model [186] provides a framework to extract and represent narratives based on a route map metaphor. The narrative and its stories are shown as a series of routes through landmarks, which represent the events. In computational terms, the narrative is modeled through a directed acyclic graph of events. The events are represented through neural embeddings of article headlines. The graph is extracted by solving an optimization problem defined following a linear programming formulation similar to the Connect the Dots approach. The optimization problem is based on maximizing *coherence* subject to *coverage* constraints. Coherence measures how much sense it makes to connect two events together and is defined as the geometric mean of the *content similarity* of events—using cosine or angular similarity—and their *topical similarity*—based on JS similarity of their topic distributions based on clustering. Coverage is measured by the average percentage of topical clusters covered by the selected events based on their topic distributions. Once the optimal map has been found, the main storyline is extracted by normalizing the coherence values of the edges into probabilities and finding the maximum likelihood path. Then, a set of representative landmarks (i.e., important events) of each story by finding the maximum antichain, which corresponds to the point of the maximum width of the graph.

Others

Linear Representations Guha et al. [139] propose an *event threading* approach based on a graph decomposition method that generates document timelines. In particular, they propose decomposing a directed acyclic graph into disjointed node paths that ensure that as many nodes as possible participate in at least one path (i.e., they seek to maximize a notion of *coverage*). The first step is to construct the graph, they propose doing this based on *important terms* (or even entities) in the document collection and their co-occurrence. Fur-

thermore, documents are modeled following a bag of words approach, although the method is also designed to handle TF-IDF representations. Once the graph is constructed, the next step is to solve the event thread extraction problem. To do this, they propose three formulations: an exact algorithm, a maximum approach, and a dynamic programming approach. The first method is an exact algorithm based on minimum cost flow, which has a high computational cost and is impractical. The second is an approximation algorithm based on maximum matching in bipartite graphs that solves the thread extraction problem for a fixed maximum size. The third method is based on an approximation algorithm that uses dynamic programming to solve the thread extraction problem for a range of thread sizes.

Laban and Hearst [208] present newsLens, a system to build and visualize long-ranging news stories. In particular, their system groups news articles based on their *topics*—based on a graph clustering approach—and then selects a sample of headlines from salient dates—based on the *frequency* of publications. In more detail, the first step in their extraction approach is to construct a keyword graph for a starting time period using TF-IDF representations of the articles. Next, a local topic graph is created based on a user-defined threshold for the number of shared keywords between articles. After the initial time period, a sliding window approach with a user-defined length is used to handle the rest of the data. For each time period, a local topic graph is created and compared with the graph from the previous period to check for three types of relationships: linking (connecting a topic from the current graph to a pre-existing topic), splitting (dividing a pre-existing topic into new topics in the current period), or merging (combining separate topics from the previous step into a single one of the current period). However, this approach is not able to handle stories that have *long-time gaps* between publications. To handle these cases, the content similarity of non-overlapping stories is analyzed and merged if above a specific threshold. Afterward, their method assigns a name to the storyline by extracting noun phrases from the news articles and scoring them based on multiple criteria (e.g., length, type of noun, abstractness, and frequency). Finally, salient dates are selected based on local frequency changes, and representative headlines are sampled randomly from these dates to generate the final timeline visualization.

Graph Representations Uramoto and Takeda [354] proposed a graph-based approach to model the relationships between news articles. In particular, they use a directed graph based on temporal ordering and *event similarity*. This is the earliest article that fits with our definitions of event-based narrative representations for news narratives that we found. In particular, the authors use the concepts of *genus* and *differentia* words. For adjacent articles, genus words are computed using the intersection of their word sets and represent already known information in the story. In contrast, differentia words are built from the set difference between the articles (in temporal order) and represent new knowledge in the story. Thus, differentia words are more important when trying to find coherent sequences of articles. The events are represented with a variation of TF-IDF that assigns more weight to *differentia* words.

Tannier and Moriceau [343] propose an approach for building multi-document event threads from news articles. In particular, they use a supervised learning approach with a series of

classifiers to define the type of relationship between news articles: *same-event*, *continuation*, or *reaction*. The output of this method is a temporal event graph, where the nodes correspond to events (represented as news articles) and the edges are labeled with the corresponding relationships. In particular, the first step is to determine whether there is a connection at all between the articles. To do so, an initial classifier is implemented using a series of *content similarity* features (e.g., word overlap, cosine similarity, and similarity of the first sentences) to construct the initial temporal graph. However, this is not enough to find all potential relationships and a second-level classifier is included that takes into account the results from the previous classifier by using *degree-based features* from the temporal graph. Next, after a connection has been established, another classifier determines whether this connection is based on the articles referring to the same news event—same-event connection—or based on a continuation—when an event is a direct continuation or consequence of a previous one. This classifier relies on *date-based features* (e.g., differences in publication time, date references, and references between events themselves) and *keyword-based features* (e.g., usage of temporal words, reaction words, or opinion words). The output is fed into another classifier that leverages degree-based features again to find more relationships. Due to the *transitive* nature of the same-event and continuation relationships, a post-processing step takes the graph and constructs the *transitive closure* for these specific relations. Afterward, a final classifier uses the same features to determine whether a continuation is a reaction—a subset of continuations that relate the reactions of people (or organizations) to an event.

Hu et al. [164] propose a system to model storyline interactions from news events. Their approach generates a series of event timelines focusing on specific entities or topics and their interactions with each other. In particular, this results in a directed graph connecting multiple events. In contrast to other approaches, the underlying representation of events is based on the *main event descriptors* (i.e., the answers to Who, What, When, Where, Why, and How) [187] which are extracted directly from each article and represent the key elements of the event. Based on this information, a coherence graph is constructed and used to identify the storylines through a random walk. Coherence is defined by three factors: *subtopic consistency*, *entity relatedness*, and *time continuity*. To measure subtopic consistency, the first step is to use a generative probabilistic mixture model to discover latent subtopics. Then, JS divergence is used to measure the distance of topic distributions between articles. Next, entity relatedness is measured by the average *affinity* of the entities from each pair of articles using normalized point-wise mutual information. The time continuity factor is simply defined as an exponential penalty term dependent on the temporal distance between events. The coherence graph is built by creating edges between documents that have a coherence score above a given threshold. Based on the coherence graph, a series of *informative events* that connect multiple storylines are identified. Specifically, a topic-sensitive PageRank algorithm [153] is used to discover these events. In turn, these events feed the storyline generation algorithm, an iterative algorithm that selects a single informative event for each story for each day.

Bögel and Gertz [46] present a temporal linking framework based on the concept of article

references. In particular, they exploit the structure of news articles to construct an information network. Instead of comparing articles based on overall content similarity, they exploit the use of *lead paragraphs*, *explanatory paragraphs*, and *additional information paragraphs* in typical news articles. Specifically, they construct the network based on *temporal expressions*, *keywords*, and *entity names*. To select valid event connections, the first step is to filter based on temporal information contained in the text based on a temporal tagger. Next, connections are evaluated based on the *similarity* of the lead paragraph of a news article with all the other paragraphs of another news article (i.e., capturing references to the event). Similarity is computed based on the entities and keywords mentioned in each paragraph based on a weighted average of Jaccard and cosine similarity. Finally, irrelevant edges are pruned based on a user-defined threshold. However, some non-relevant edges are kept if they fulfill the role of a *support path*—paths that have non-relevant edges but share endpoints with fully relevant paths—that provide more evidence of two events being connected. The output is a directed graph based on references, not necessarily acyclic, as there are future temporal references in some articles.

3.2.4 Events as Clusters: Event Evolution and Threading

Now, we present works that use a cluster-level resolution. We divide the discussion into two parts: works related to event threading [265] and evolution [400], and others.

Event Threading and Evolution

Nallapati et al. [265] use a directed graph model to represent to capture the structure and dependencies of events in a news topic. They call this extraction process *event threading*. They represent each event as a cluster of news articles. Event threading is a supervised method that consists of two phases: clustering documents and modeling dependencies. The clustering process starts with a cluster for each document in the data set and merges them iteratively based on similarity until the similarities fall below a predefined threshold. The authors evaluate three types of cluster similarity on the average link, complete link, or single link of the clusters based on document similarities. Document similarities are based on *content similarity* (e.g., cosine similarity), *common locations*, and *common entities*. Furthermore, there is an exponential decay term based on the *temporal distance* to penalize larger temporal distances between documents. Next, dependency modeling uses surface-level features of the document clusters, such as word distributions and time-ordering of the news articles. Based on this information, the authors propose several link extraction criteria (complete-link, simple threshold, nearest parent, best similarity, and maximum spanning tree). These approaches rely on temporal order, similarity information, or structural information.

SToRe (Storyline-based Topic Retrospection) is a topic retrospective system [228, 229, 230] that extracts the main storyline from a given news topic and provides a summary of the

topic based on this storyline. In particular, the extraction process consists of four phases: event identification, topic structure identification, main storyline construction, and storyline-based summarization. In the event identification phase, similar news articles will be clustered together to represent a single event using self-organizing maps. In the topic structure identification step, the events are linked together based on whether their *similarity* exceeds a specific threshold. To compute similarity, the events are represented with a vector of term weights using the concepts of *genus* and *differentia* words [354]. Then, cosine similarity is used to compare the event vectors. Next, in the main storyline construction step, an MST is extracted from the constructed topic structure. The MST is based on the relevance of each event with respect to the topic. The MST is used to generate a timeline of events, and it is further extended with small side branches of other relevant events based on a specific threshold. Finally, in the storyline-based summarization, a summary is generated for each event based on the news articles contained in its cluster using accumulated weight summary [137].

Yang et al. [400, 401] use directed acyclic graphs to represent the evolution of events in online news. They call their approach *event evolution graphs*, which represent temporal and causal relationships between events. Events are defined as sets of news articles and are represented as the average of the TF-IDF vectors of each article they contain. We note that the proposed method assumes that events and their corresponding articles are already computed. In practice, this would require a clustering step before constructing the graph. These events are linked together based on their similarity and a user-specified threshold, which is computed based on *content similarity* (e.g., cosine similarity), *temporal proximity*, and *document distributional proximity* (which penalizes bursty periods with many articles about the same event). The latter two terms are represented through exponential decay factors. Furthermore, users are able to reduce the temporal granularity of the event evolution graph, which merges specific events that occur in short time frames.

Qiu et al. [290] propose another event evolution graph extraction method. Their construction method follows an iterative approach based on *content similarity* and *temporal order*. In particular, documents are first grouped into clusters using the OHC method [288] in the first time period, which gives rise to the initial events. Next, the PRAC method [289] is used to build classifiers and determine whether the documents of the next time period are continuations of a cluster identified in the previous period. If so, a new event node is created using the identified cluster as its parent. This process is repeated until the last time period. Next, *twigs*—paths that die before the end of the timeline—are removed based on a user-set tolerance, and equivalent event nodes are merged to reduce graph complexity.

TSCAN (Topic Summarization and Content ANatomy) [68, 69] is a method to analyze news data that produces a global summary and constructs an event evolution graph. We focus on the event graph component of this method. First, news articles are grouped into themes obtained through a matrix factorization approach with TF-IDF document representations. Next, the news articles of each theme are temporally segmented using an *energy value* threshold based on eigenvalues from the matrix representation. In practice, this generates clusters

of documents based on *frequency*, which are associated with the nodes of the event evolution graph. The evolution graph is a directed acyclic graph, where the edges are constructed using *temporal similarity*—computed using the temporal distance between events, with special cases to consider event overlap—and *content similarity*—based on cosine similarity.

Khurdiya et al. [197] propose a system that extracts directed graphs to represent stories from news data using multi-perspective links. Each node of this graph is associated with multiple news articles. The system uses LDA to extract topics in each time unit (e.g., a day). The extracted topics are associated with sets of articles based on the strength of the topic in each article and form the basis of the story identification model. We note that these topics and their article sets correspond to the notion of event that we use in this survey. Next, article sets are linked chronologically based on *topic correlation* (e.g., Pearson’s correlation coefficient) and a user-defined threshold, generating a directed graph of events.

Wei et al. [377] identify event episodes in news data sets and construct a temporal episode graph (i.e., an event graph under our definitions in the survey). In particular, this article shows a discovery mechanism that organizes news documents into events using novel TF-IDF representations that incorporate a temporal component. Then, the system builds a link structure based on intercluster similarity measures. The first proposed event representation, called TF-IDF_{Tempo}, gives more weight to features with *consecutive occurrences* in a sequence of documents (i.e., it incorporates the surrounding context of the document) by modifying the IDF component of TF-IDF to consider the order of the documents. However, this approach is too strict and is unable to model overlapping events. Moreover, it also has a high bias towards low-frequency articles that are temporally close. Thus, the authors propose a second representation, called TF-Enhanced-IDF_{Tempo} which modifies the IDF component by adopting the *significance factor* proposed by Luhn [236] and a *temporal gap threshold* to allow for short discontinuities in feature appearances. These representations are used with Hierarchical Agglomerative Clustering (HAC) [363] to construct the article clusters that represent the events. For the purposes of clustering, document similarity is defined by content similarity (e.g., cosine similarity) and a negative exponential penalty for temporally distant documents.

Huang et al. [166] propose a different event evolution approach to build and analyze event relationships based on three types of event connections. In particular, they define a *co-occurrence dependence relationship*, an *event reference relationship*, and a *temporal proximity relationship*. The authors define events as a set of news articles and identify them through clustering and topic modeling using a combined similarity measure that leverages LDA and a TF-IDF document model with cosine similarity. Once the events are identified, the method extracts a series of core features (i.e., key entities and terms of the article) by analyzing the *lead* of the articles and evaluating whether their frequency is above a specified threshold. These core features are used to construct a vectorial representation of the events. For the co-occurrence relationship, the method computes the aggregation of all mutual information between all features of the event, generating a symmetric matrix that represents all event-event relationships. For the event reference analysis, the method identifies shared core

features and defines the degree of event reference based on the frequency of references in an event to the core features of a previous event, adjusted by the weight of these terms in the referencing event. Temporal dependency is evaluated using an exponential decay formula.

Event Phase Oriented News Summarization (EPONS) [368] is a timeline summarization approach that assumes that a story summary contains multiple timelines, each one corresponding to a specific event. To model the semantic relations of news articles, EPONS uses a graph model, called Temporal Content Coherence Graph (TCCG), which is an event graph based on two metrics: *content coherence* and *temporal influence*. Content coherence is based on the weighted average of *topic level similarity*—modeled by JS divergence over an LDA topic distribution—and *entity-level similarity*—modeled over a ranking of named entities using the Tanimoto coefficient. Temporal influence is modeled through a Hamming (cosine) kernel to properly separate temporally distinct events. The TCCG is built by selecting edges that are above user-specified thresholds in each metric. Based on this graph, EPONS uses a modified structural clustering approach to group the news articles into different events. Furthermore, small clusters of similar articles are filtered out to ensure that the events are modeled properly. This post-processing is done by using four quality metrics on a pre-trained logistic regression classifier: percentage of new articles, time interval length, pairwise topic similarity, and pairwise entity similarity. Having identified the events, it is now necessary to construct the individual summaries and finalize the timeline. To do so, a vertex-reinforced random walk [247, 279] is used to rank the relevance of news articles inside each event, in a similar manner to PageRank. Next, a supervised model is used to determine whether the headlines are factual (i.e., they are reporting a specific event) or an opinion, as opinion-based headlines are not considered useful for timelines and must be filtered out. Finally, an optimization method is used to maximize the total *relevance*, subject to *non-redundancy* constraints (i.e., disallowing events that are too similar) to select the news articles.

Cai et al. [59] propose a method to extract Temporal Event Maps (TEM) based on the *content dependence* degree and *component event reference* degree for each pair of events. TEMs are directed graphs that have events as nodes, relations as edges, edge weights representing the strength of event relationships, and node weights representing the importance of each event. Events are defined as groups of related documents and identified using a LDA model. After obtaining the events, the next step is to compute the two core metrics that define the temporal event maps. The content dependence degree is defined as the aggregation of all mutual information among the features of each event. The content reference degree is defined by the presence of *core features* of an event—salient terms based on frequency—in other events. Unlike content dependence, this is not a symmetric relationship between events. To construct the temporal event maps, the first step is to order events based on starting time. Then, connections are added for events that surpass a user-specified threshold for the product of content dependence and event reference degrees, which provides the edge weights for the graph. Finally, a ranking procedure based on PageRank is used to generate the event importance values.

Others

Information Cartography Continuing with their work on metro maps, Shahaf et al. [323, 324] propose a new framework called Information Cartography that features *zoomable* metro maps, allowing users of the map to visualize the news at different levels of resolution, allowing the user to zoom in to specific metro stops and generate a new map. Metro stops and events are no longer represented as single documents but as clusters of events. The articles are segmented into time windows and clusters are computed using a community-detection algorithm on word co-occurrence graphs. To extract the maps, an optimization problem is defined based on finding the best *structure* for the map, relying on the idea of minimizing the total number of storylines (to reduce unneeded complexity) and maximizing the *number of covered clusters* (to ensure that the stories are well covered). This approach leads to simple stories being modeled as a single metro line and more complex stories requiring the use of multiple shorter lines. Furthermore, a series of additional constraints for *story coherence*, *cluster quality*, and *map size* is imposed.

Building upon the concept of metro maps and information cartography, Xu and Tang [394] propose a narrative representation in the context of societal risk events (e.g., earthquakes) called Risk Maps. These maps follow the same basic representation of information cartography with events being represented as clusters of documents. However, one key difference is that this approach leverages advances in text representation by using neural word embeddings for news articles before clustering. To obtain the risk map, the authors choose to maximize *coverage* as their primary objective, followed by *connectivity*, subject to a minimal *coherence* constraint. Coverage is defined based on how well each cluster is covered by the different storylines. Connectivity is simply the number of storylines that intersect. Coherence is defined based on the Jaccard similarity of consecutive clusters in the storylines. The optimization problem is solved using a greedy algorithm that finds the best path among clusters at each step.

Story Forests Liu et al. [232, 233] propose the Story Forest approach, where different stories are constructed and represented as a forest of event trees. First, events are clustered using a community detection approach on word co-occurrence graphs using betweenness centrality. Next, documents are associated with each topic through a similarity based on TF-IDF representations. Afterward, a second step groups documents together based on a supervised classifier (SVM) to determine whether pairs of documents refer to the same event based on TF-IDF features and similarities between the contents and titles of articles. The story forest is built iteratively by adding events into its trees by using three operations: merge, extend and insert. Before adding the events it is necessary to determine the correct story tree. This is done based on a measure of *compatibility*, computed as the Jaccard similarity of the keywords of the event and the tree. If no trees are related to the event, a new tree is created with the event as its root. To add the event to an existing tree, the method first tries to *merge* it with any of the existing events into the same node using the previously trained SVM classifier. Otherwise, the method scans all the nodes to identify

which tree to *extend* based on a measure of connection strength determined by three elements: *compatibility*, *coherence*, and *time penalty*. Compatibility is measured by the similarity of their centroids based on cosine similarity. Coherence is a story-level measure that takes into account the path of events from the root of the tree to the newly appended event by measuring the average consecutive compatibility value. Finally, the time penalty is an exponential decay factor that depends on temporal distance. If none of the events are appropriate, the event is *inserted* as a new node connected to the root.

3.3 Narrative Extraction Criteria

In this section, we present a summary of the different construction criteria found in the reviewed articles. These criteria refer to either an evaluation metric or additional information used in the extraction algorithms themselves as part of an objective function (e.g., coherence optimization), selection criteria (e.g., filtering based on content similarity or topic distribution similarity), and other types of extraction heuristics (e.g., leveraging article structure to compute content similarity or evaluating the use of opinionated language). The first part of Table 3.2 provides an overview of the different construction criteria. We note that these criteria are not mutually exclusive and can be combined as needed.

Relevance Relevance metrics evaluate whether the events in the narrative are relevant or significant to a given query or topic [218, 221, 268, 347, 398, 399]. In general, relevance is measured by borrowing techniques from traditional search methods in Information Retrieval, such as PageRank and its variations [349, 368, 416, 417]. However, some approaches use supervised methods to learn a ranking function [42, 350]. The results from such techniques are used to feed other parts of the algorithm or could be directly used to select relevant events, turning this issue into more of a traditional information retrieval problem rather than a narratological one.

Content Similarity Another approach to extracting narratives is based on modeling content similarity between events. Over two-thirds of the methods use some sort of content similarity measure. There are many ways to do this, in particular, we found the following approaches: surface-level similarity comparisons (e.g., Jaccard similarity or cosine similarity) [163, 268, 354], topic similarity based on topic distribution information (e.g., comparing topic vectors extracted from LDA models) [71, 221], and entity-based comparisons (e.g., entity co-occurrence in events) [265, 417].

The exact choice of approach is highly dependent on the event representation. In recent years, researchers have started leveraging advances in text representation with neural embeddings (e.g., BERT) [98, 186, 226, 347, 407], which have several advantages over traditional frequency-based models and are better able to capture semantic similarities.

The use of entity-based information in event-based narrative extraction methods to measure event content similarity remains limited in scope, with sparse usage over the years compared

Table 3.2: Summary of the extraction criteria and evaluation metrics used in the reviewed articles.

Year	Reference	Extraction Criteria														Evaluation Metrics										
		Relevance	Surface Similarity	Topic Distribution	Entities	Coherence	Coverage	Dispersion	Diversity and Redundancy	Output Structure	Article Structure	Content References	Temporal References	Burstiness and Frequency	Temporal Distance	Traditional IR Metrics	Summarization Metrics	Ranking Metrics	Clustering Metrics	Coherence	Coverage	Dispersion	Diversity and Redundancy	User Performance	User Perception	Other or N/A
1998	Uramoto and Takeda [354]		x																							x
2004	Nallapati et al. [265]		x		x										x	x										
2004	Chieu and Lee [73]		x											x	x										x	
2005	Guha et al. [139]		x				x																		x	
2006	Yang et al. [400]		x											x	x		x									
2006	Lin and Liang [228]		x																				x	x		
2007	Lin et al. [230]		x														x									
2008	Chen and Chen [68]		x											x	x		x									
2008	Qiu et al. [290]		x																							x
2008	Lin and Liang [229]		x																					x	x	
2009	Yang et al. [401]		x											x	x		x									
2010	Shahaf and Guestrin [319]						x																	x		
2011	Yan et al. [399]	x					x	x		x							x									
2011	Yan et al. [398]	x												x	x		x									
2011	Hu et al. [163]		x	x	x												x									
2011	Khurdiya et al. [197]				x												x									
2012	Zhu and Oates [416]	x	x			x				x														x		
2012	Chen and Chen [69]		x																							
2012	Shahaf and Guestrin [320]						x																	x		
2012	Shahaf et al. [321]						x	x			x												x			
2013	Binh Tran et al. [42]	x	x	x		x					x		x	x			x	x								
2013	Li and Li [218]	x	x	x		x	x										x									
2013	Tran et al. [350]	x	x	x		x					x		x	x			x									
2013	Huang and Huang [167]			x																						
2013	Tannier and Moriceau [343]		x								x						x									
2013	Shahaf et al. [322]						x	x			x													x		
2013	Shahaf et al. [323]						x	x			x													x		
2014	Nguyen et al. [268]	x	x	x										x			x	x								
2014	Zhu and Oates [417]	x	x		x	x				x															x	
2014	Huang et al. [166]		x	x								x					x									
2014	Wei et al. [377]		x											x	x		x									
2014	Hu et al. [164]				x	x											x	x						x		
2014	Zhou et al. [414]		x			x					x							x								
2015	Tran et al. [349]	x																								
2015	Li et al. [221]	x	x	x		x	x										x							x		
2015	Bögel and Gertz [46]		x		x												x				x					
2015	Chen et al. [71]		x	x																						
2015	Shahaf et al. [324]						x	x			x						x	x						x		
2017	Wu et al. [391]		x																							
2017	Liu et al. [232]		x			x																			x	
2017	Laban and Hearst [208]			x										x	x											x
2018	Wang et al. [368]	x		x	x													x								
2018	Tikhomirov and Dobrov [347]	x	x																							
2018	Xu and Tang [394]						x	x			x															
2018	Zhou et al. [415]		x				x				x							x								
2019	Camacho Barranco et al. [60]		x	x	x	x				x												x			x	
2019	Cai et al. [59]		x																							
2019	Yuan et al. [408]						x				x							x		x						
2020	Duan et al. [98]						x											x	x				x			
2020	Liu et al. [233]		x				x																		x	
2021	La Quatra et al. [207]		x											x				x							x	
2021	Yu et al. [407]		x				x							x	x	x		x								
2021	Liao et al. [226]																	x	x							
2021	Keith Norambuena and Mitra [186]	x	x			x	x																		x	

to other content similarity measures [46, 60, 163, 164, 265, 368, 417]. Combining entity information with other types of similarities would provide a much more holistic view of content similarity. Furthermore, expanding upon this approach, content similarity metrics could exploit the *main event descriptors* [187] to compute a more precise similarity measure.

Coherence Coherence metrics evaluate whether the narrative makes sense. Due to their importance as an extraction metric, we show some mathematical formulations of coherence and coherence-like metrics in Table 3.3.

While coherence has a formal definition in narratological terms [1], it is just as complex and ill-defined as relevance in computational terms. One particular motivation for the definition of coherence that stands out is the idea of *smoothness* from the Connect the Dots [319, 320, 323, 324] series of works. In particular, they use the concept of *word influence* and *word activations* (i.e., the sustained importance of the word in a storyline) to construct stories that have smooth transitions.

Other approaches compute coherence based on content similarity. These works also seek to generate *smooth* stories by avoiding drastic local changes based on content similarity [60, 186, 221, 232, 233, 350, 394, 399, 414, 415, 416, 417], without explicitly defining active words or topics like the original Connect the Dots approach. Finally, one approach also considers coherence around the idea of causality [42, 350] in a supervised setting (e.g., causal signals in text).

Coverage-like Metrics Coverage-like metrics evaluate whether the extracted narrative properly covers the relevant events, stories, or topics. These metrics include coverage itself and related metrics, such as redundancy and diversity. The most basic form of coverage is simply the percentage of topics or relevant events covered by the extracted representation (or some variation of this metric) [98, 139, 186, 408], or a probability estimation [321, 322, 323, 324, 394]. Equation 3.1 shows an example formulation of coverage for a cluster c , where Π represents an extracted narrative with storylines l .

$$\text{Cover}_{\Pi}(c) = 1 - \prod_{l \in \Pi} (1 - \text{Cover}_l(c)) \quad (3.1)$$

Another approach to compute coverage is to do a content similarity comparison between the output and the full data set (or a relevant subset) [218, 221, 399]. In contrast, redundancy and diversity [98, 368, 399, 416, 417] metrics are based on the idea that events should not be covered more than necessary, thus high redundancy can lead to coverage problems.

Structural Information Some works evaluate the structure of the output narrative representation. In particular, these metrics consider aspects such as size (in general) or connectivity (in graph-based narratives).

Size can be used as a proxy for complexity (e.g., length of the timeline) [394]. In most cases, rather than as an evaluation metric, size is used as a constraint (e.g., setting a maximum story length) [321, 322, 323, 324, 408, 414, 415].

Connectivity metrics [321, 322, 323, 324] are used to ensure that narrative graphs avoid isolated stories, as they should be interwoven throughout the narrative. Structure metrics are mostly analyzed at a global level (e.g., the total number of connected stories). However, it is possible to consider local structural features, such as node degrees [343].

Table 3.3: A sample of different formulations of coherence from the reviewed articles.

Formula	Resolution Level	Description	Source
$\frac{1}{ S -1} \sum_{(c_i, c_j) \in S} \text{JaccardSim}(c_i, c_j)$	Events as Clusters	This is a measure of coherence based on average Jaccard Similarity along a story S based on cluster words.	[394]
$\frac{1}{ S -1} \text{CosineSim}(c_i, c_j)$	Events as Clusters	This is a measure of coherence based on average Cosine Similarity along a story S based on cluster centroids.	[232]
$\max_{\text{activations}} \left\{ \min_{(d_i, d_j) \in S} \sum_{w \in \text{Words}} \text{Influence}(d_i, d_j w) \cdot \mathbb{1}(w \text{ active in } d_i, d_j) \right\}$	Events as Documents	This is the full form of the coherence for a storyline S from the original Connect the Dots algorithm. It is based on maximizing the sum of word influences over active words in the storyline. Influence can be changed for any other type of scoring mechanism.	[319]
$\min_{(d_i, d_j) \in S} \text{CosineSim}(d_i, d_j)$	Events as Documents	This is a measure of coherence based on the minimum Cosine Similarity along a story S based on document vectors.	[414]
$\min_{(d_i, d_j) \in N} \sqrt{\text{SurfaceSim}(d_i, d_j) \cdot \text{TopicSim}(d_i, d_j)}$	Events as Documents	This is a measure of coherence for a narrative N based on the minimum geometric mean of Surface-level similarity (e.g., cosine similarity) and Topic-level similarity (e.g., Jensen-Shannon divergence). It is based on document vectors and topic distribution vectors..	[186]
$\frac{\sum_{i,j \in D \times D } w_i \cdot w_j \cdot \Phi \cdot \text{Soergel}(d_i, d_j) \cdot t_i - t_j }{\sum_{i,j \in D \times D } w_i \cdot w_j \cdot \Phi}$	Events as Documents	This is a measure of incoherence rather than coherence. It is based on the average Soergel distance and includes a temporal distance term as well. The events are weighted by their relevance (w_i and w_j) and their temporal distance using a custom kernel Φ .	
$\sum_{s \in E} \frac{\text{Count}_{\text{match}(s, E)}(\text{gram}_n)}{\text{Count}_{\text{previous}(E)}(\text{gram}_n)}$	Events as Sentences	This is a measure of coherence based on the n-gram overlap between the current event sentences and the sentences of the previous summary of the timeline.	[350]
$\frac{1}{1 + \exp(JS(E, \text{Previous}(E)))}$	Events as Sentences	This is a measure of coherence based on the Jensen-Shannon divergence between the current event sentence and the previous event sentence of the timeline.	[221]
$\frac{\sum_{\delta=-\Delta/2}^{\delta=\Delta/2} \exp(-\delta/v) \cdot \text{KLD}(E_t, E_{t-\delta})}{\sum_{\delta=-\Delta/2}^{\delta=\Delta/2} \exp(-\delta/v)}$	Events as Sentences	This is a measure of coherence based on Kullback-Leibler divergence between the current event at time t and all the other local events in a Δ time window surrounding the event. The events are weighted by their temporal distance based on parameter v .	[218]

Exploiting the internal article structure [46, 347, 350] is another piece of structural information used by some methods. Most breaking news articles are written following the *inverted pyramid structure* [187], where the most important information—the main event descriptors—is shown first in the *lead*. Thus, the first few lines of an article describe its main event [350] and subsequent paragraphs may contain more details and reference previous events.

Content References Another criterion to consider in news narrative extraction is the use of content references. As mentioned before, some news articles make explicit references

to previous works in their body. Note that this differs from explicit date-based references discussed before, which rely on explicit temporal information. This approach also differs from general content similarity because of its goal of identifying specific references rather than global similarity.

One way to identify these references is to compare the lead of a news article with the additional information paragraphs of another article [347]. Other approaches identify references based on sentence co-occurrence without considering article structure [407]. Alternatively, a set of core features [59, 166] (e.g., relevant keywords or main event descriptors) could be identified and used to detect references in other articles. Once identified, these references can be used to identify relevant events based on reference-based metrics (e.g., bibliographic coupling).

Temporal Features Temporal information, such as the temporal distance between events or specific date references, has been used. In particular, temporal distance is commonly used to penalize events that would otherwise be similar in content. For example, consider two articles describing separate protests in a city, one during the year 2000 and another in the year 2010. These two articles would likely be very similar in terms of content, including both surface-level features and topic distributions. However, given the temporal separation between them, they would likely refer to different events. Thus, a common strategy is to define an exponentially decreasing term of the form $C_0 \exp\left(\frac{-\Delta t}{\sigma}\right)$ (or similar), where C_0 and σ are pre-defined constants [164, 166, 232, 233, 265, 377, 398, 400, 401, 407], although there are other approaches, such as kernels to perform temporal proximity projections [368, 398] or overlap-based measures [68, 69]. However, we note that the use of a temporal penalty is not always desired. Some events are continuations of stories that did not have anything new to report for a long time. For example, the investigation results of a flight accident might come much after the accident itself has been covered, leading to temporal gaps in story coverage [208, 377]. Thus, it is necessary to distinguish between continuations and completely new storylines when the time gap is high enough.

Burstiness and frequency measures and metrics based on these (e.g., energy values) are other time-based criteria used to identify relevant events and dates [42, 68, 69, 71, 167, 208, 268, 347, 349, 377, 398, 401, 407]. For example, periods with many publications are likely to contain important events. Alternatively, a specific event might be reported several times by different outlets. Finally, other temporal features include the use of specific temporal expressions or date references in the text [42, 46, 207, 226, 343, 349, 350] to identify temporal cross-references between documents.

3.4 Evaluation Metrics

In this section, we discuss the evaluation approaches for the narrative output of the extraction methods. In particular, we show the evaluation metrics used to assess the quality

of the extracted narratives. These output metrics are generally intended to be interpreted by humans, unlike the extraction criteria which may or may not be easy to interpret. In particular, user-based evaluation metrics (e.g., task performance or user perception) are an important subset of output evaluation criteria. The second part of Table 3.2 provides an overview of the different output evaluation criteria.

3.4.1 Computational Metrics

These metrics seek to evaluate the extracted narrative based on computational measures of narrative quality. These metrics are usually supervised, requiring a gold standard data set to be computed.

Supervised

We first discuss supervised approaches. In particular, we identified three broad types of metrics here: traditional information retrieval metrics, summarization metrics, and ranking metrics.

Traditional IR Metrics Several works—about a third of the reviewed articles—rely on classical evaluation metrics such as accuracy, precision, recall, and the F_1 score [46, 59, 71, 98, 163, 164, 166, 197, 230, 265, 343, 349, 368, 377, 400, 401] taken from traditional information retrieval and machine learning literature. In particular, these approaches evaluate the quality of the output by measuring whether *events* or their *connections* were identified correctly. Some methods also use variations of these basic metrics, such as the mean average precision [42, 268] over multiple dates.

Summarization Metrics Specialized metrics from the summarization domain have also been used to evaluate narratives in several works—about a third of the reviewed works use them.

In particular, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics [274] have been used to evaluate the output of narrative extraction methods, mostly in timeline summarization works with a sentence-level event resolution [42, 71, 167, 207, 218, 221, 226, 268, 347, 350, 399, 407], but also in works with a document-level resolution [164, 408, 414, 415]. ROUGE metrics include variations such as ROUGE-N (which measures the overlap of N-grams), ROUGE-L (which measures the longest common subsequence), ROUGE-W (a weighted version of ROUGE-L that favors consecutive subsequences), ROUGE-SU (skip-bigram and unigram-based co-occurrence statistics) and their precision, recall, and F_1 score variants. The most common variant is ROUGE-N, which we show in Equation 3.2.

$$\text{ROUGE-N} = \frac{\sum_{s \in R} \sum_{\text{gram}_n \in s} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{s \in R} \sum_{\text{gram}_n \in s} \text{Count}(\text{gram}_n)} \quad (3.2)$$

In Equation 3.2, n represents the length of the n -gram, and R represents the reference summaries (i.e., the ground truth). $\text{Count}_{\text{match}}(\text{gram}_n)$ represents the maximum number of n -grams that co-occur in a candidate summary and the reference summaries. $\text{Count}_R(\text{gram}_n)$ represents the number of n -grams in the reference summaries.

An alternative is to measure the average *summary-to-document content similarity* where the summary is compared against the documents in the data set using a text similarity measure (e.g., cosine similarity) [68, 69].

We note that these metrics are mostly used with linear representations rather than graph-based models—only three of the reviewed works that extract graphs use summarization metrics [68, 69, 164]. This contrasts heavily with the case of traditional information retrieval metrics, where the split was much more balanced between linear (~40%) and graph representations (~60%). This might be due to the inability of these metrics to handle complex structures.

Ranking Metrics Other works rely on ranking-based metrics, like those used in traditional search tasks from information retrieval. For example, Wang et al. [368] use a relevance-based approach to evaluate their event phase summaries. Liao et al. [226] evaluate the ranking performance of WILSON with the mean reciprocal rank and discounted cumulative gain [21]. Cai et al. [59] use the normalized discounted cumulative gain [405] to evaluate all their events.

Clustering Metrics Liu et al. [232, 233] used clustering metrics to evaluate the event nodes—which are represented as clusters of articles—in their Story Forest method. In particular, they use the homogeneity, completeness, and V-measure scores [304]. These metrics require labeled data sets to be computed, thus they are supervised despite being designed to evaluate unsupervised clustering methods. In particular, homogeneity is larger when each extracted cluster only contains members of a single class. In contrast, completeness is maximized when all members of a true class are in the same cluster. Finally, the V-measure takes both of these metrics and computes the harmonic mean between them, similar to how the F_1 score treats precision and recall in traditional classification metrics. We note that none of the other events as clusters methods used these metrics or other similar clustering metrics to evaluate their models. Instead, they relied on traditional information retrieval metrics like accuracy, precision, recall, and the F_1 score.

Unsupervised Metrics

We now discuss unsupervised approaches. In general, there are far fewer works relying on unsupervised metrics to evaluate the final narrative output.

Coherence In general, coherence is not used to evaluate the output narrative despite being a useful metric during the extraction process. One exception is Xu et al. [394], who evaluate their output using a weighted average of story coherence (based on Jaccard similarity) and

story size.

Coverage Xu et al. [394] evaluate their output by treating coverage as a structural measure, making the assumption that good coverage of topics means that the structure of their metro map representation is good. However, due to the formulation of coverage based on whether the topical clusters of the data set are covered, it does not explicitly consider the structure of the output, which makes it inappropriate to evaluate the structure. In contrast, Bögel et al. [46] use a notion of coverage based on event connections in a graph (i.e., an article is covered if there is at least one edge connecting it) that could be treated more as a structural measure than the topical concept of coverage.

Dispersion Camacho et al. [60] use the *dispersion coefficient*—originally proposed as an evaluation metric for storytelling in the intelligence analysis domain [161]—to evaluate their storyline. In particular, the dispersion coefficient is based on the Soergel distance, although other distance metrics could be used [298]. In particular, dispersion is based on Swanson’s complementary but disjoint hypothesis [341]—where articles that have no explicit common elements yield important inferences or insights when combined. These insights are not apparent from the separate documents. Furthermore, the authors propose a new evaluation metric to measure story flow based on Swanson’s hypothesis called the dispersion coefficient, shown in Equation 3.3. We note that this particular version is based on the Soergel distance (S), but any other distance metric between documents could be used in practice.

$$\text{Dispersion}(d_1, \dots, d_n) = 1 - \frac{1}{n-2} \sum_{i=1}^{n-2} \sum_{j=i+2}^n D(d_i, d_j) \quad (3.3)$$

$$\text{with } D(d_i, d_j) = \begin{cases} \frac{1}{n+i-j}, & \text{if } S(d_i, d_j) < \theta \\ 0, & \text{otherwise.} \end{cases} \quad (3.4)$$

Diversity and Redundancy Finally, another alternative is a diversity metric to ensure proper coverage or low redundancy. In particular, Duan et al. [98] used diversity—based on the average pairwise similarity of sentences (see Equation 3.5—to evaluate the performance of their comparative timeline extraction method.

$$\text{Diversity} = 1 - \frac{1}{|S|^2} \sum_{s_i \in S} \sum_{s_j \in S} 1 - \text{CosineSim}(s_i, s_j) \quad (3.5)$$

3.4.2 User Evaluation Metrics

These metrics seek to evaluate the extracted narrative based on subjective user measures or task performance measures.

Task-oriented Evaluation Task-oriented metrics require designing a series of benchmark tasks to measure the number of correct answers, accuracy, how much time the users take

to complete the task or some other measure of correctness or quality. Few works use task-oriented evaluation metrics: Metro Maps [321, 322], Information Cartography [323, 324], and the SToRe system [228, 229]. These works rely on event-based representations and all of them evaluate extraction methods as retrieval tools following a similar approach. In particular, there are *micro-knowledge* tasks that measure how the extracted narratives help users *retrieve information* faster and *macro-knowledge* tasks that measure how the extracted narratives help users *understand the big picture*.

For micro-knowledge tasks, all works create a series of simple retrieval questions such that the answers can be easily classified as right or wrong. For example, retrieving dates, facts, relevant entities or the main event descriptors. Users are evaluated by measuring how many correct answers (i.e., accuracy) they can get in a fixed amount of time and the rate at which they answer these questions [228, 229, 321]. Another metric used at the micro-knowledge level is the ease of navigation, estimated by the number of documents that users clicked per correct answer [228, 229, 321].

For macro-knowledge, some form of summarization is used to evaluate the narratives. Shahaf et al. [321] asked users to create summaries based on different narrative representations and then used crowdsourcing to evaluate user preference over those summaries. However, these benchmark tasks do not go beyond basic retrieval and summarization. Tasks that require higher levels of knowledge and cognitive work (e.g., analysis tasks) are not covered by these evaluations. In general, the inherent difficulty of designing benchmark tasks that can be easily evaluated might be one of the reasons why user-based evaluations of extraction methods usually rely on subjective ratings rather than task-oriented metrics.

Subjective Evaluation Most of the works that rely on user evaluations use subjective measures (i.e., user perception metrics). These subjective metrics include concepts from usability, including criteria such as user preference [164, 186, 323], visual presentation [186], and ease of use [186, 228]. Other metrics include effectiveness as perceived by the users (e.g., perceived helpfulness or usefulness), satisfaction, and comprehensibility [228, 229, 349]. Alternatively perceived familiarity before and after using the extracted narrative can be a useful measure of usefulness [319].

Lastly, user-perceived quality is another widely used approach to evaluate extracted narratives. The user-perceived quality criteria mostly correspond to the quality criteria metrics defined before [60, 186, 319, 320, 349, 416, 417], including coherence, coverage, redundancy, relevance, dispersion, and similar variations (e.g., broadness). We note that these user perception metrics suffer a similar problem as their computational counterparts—they are fuzzy concepts that could be defined differently. This is further compounded by the subjective nature of these evaluations.

Other works rely on asking users whether they consider specific elements of the narrative as correct. For example, asking whether a specific connection is correct, whether the selected documents are relevant, whether a specific storyline is logically coherent, or about the number of coherent and relevant documents [60, 232, 233]. This is similar to traditional

information retrieval metrics that rely on ground truth information. However, in this case, rather than using a previously defined gold standard, the accuracy measures are defined purely on subjective perceptions. Finally, another approach is to ask users to compare the ground truth with the output narrative—from potentially multiple methods—and rank them according to their preference based on their knowledge of the topic [226].

3.5 Discussion

We now discuss our findings. We start by addressing the structural choices in narrative representation. Next, we address some of the challenges of extraction methods. Then, we turn our attention toward evaluation methods, including benchmark data sets, computational metrics, and user-based evaluations. Afterward, we discuss practical applications of news narrative extraction. Finally, we discuss recent trends, open challenges, and potential research directions.

3.5.1 Narrative Structure

The choice of the core structure is an important aspect of narrative representation. Using a linear structure provides a simple approach to represent a narrative with a single storyline, but it does not appropriately model the nuances of narratives with multiple stories. In contrast, graph-based structures allow the modeling of different interactions between storylines (e.g., convergent and divergent stories) [186]. Linear representations are implicitly directed, but graph-based representations may or may not be directed. Directed graphs usually exploit the underlying temporal relationships to determine the direction of the connections between elements. When the connection between basic units is guided by temporal constraints it naturally gives rise to directed acyclic graphs. Directed acyclic graphs provide the most flexibility while also accounting for the temporal nature of a narrative. However, not all directed graph models are acyclic, as some use specific types of relationships that allow the creation of cycles (e.g., same-event relations).

A representation that falls between linear and fully graph-based representations is the tree-based representation [232, 233, 408, 414, 415]. Such models allow for more flexible structures than linear representations. In particular, they are able to model story divergence (i.e., multiple storylines splitting off from the root or other nodes). Unlike graph-based models, they are not able to model story convergence (e.g., two stories joining into a final event), as that would break the tree structure. Tree-based structures have not been deeply explored in the literature and could provide an intermediate approach between linear and graph-based representations in terms of complexity, allowing easier understanding by users while retaining some flexibility. However, the inability to model story convergence might limit their applications.

3.5.2 Extraction Methods

Scalability and Computational Cost Most extraction methods discussed in this survey suffer from issues when dealing with big data, as the processing pipelines are quite expensive in terms of computational power and they might not be easily parallelizable. One of the simplest methods to reduce computational cost is to filter the data beforehand. This turns the computational cost problem into an information retrieval problem, where the most relevant documents must be retrieved before extracting the narrative itself. Many methods assume that the data has been pre-filtered to a relevant set of news articles. Including a filtering step adds an additional element to the pipeline, thus increasing the risk of errors. Moreover, defining an adequate concept of relevance for this method might prove problematic in itself. Nevertheless, this provides a simple approach to mitigate the ever-increasing available amount of data.

Another approach is to deal with extraction in an online manner [233]. Most news narrative extraction methods are offline methods that analyze an entire set of news articles. However, extracting the stories in an online manner without disrupting the pre-existing structure would offer a computationally cheaper alternative. This is similar to the approach used by traditional TDT systems that sought to track the events of a topic in an online manner [13]. However, it would also require handling the structure of events associated with the narrative, which is not considered by traditional TDT.

Unified Metrics One of the limitations of current approaches is that there are multiple versions of coherence and similar metrics. Coherence itself is an ill-defined term in practice and formalizing it in a computational or mathematical definition is a difficult task. The different definitions of coherence-like metrics focus on measuring different aspects of the narrative. Moreover, additional constraints can be considered to enforce coherence beyond numerical metrics (e.g., events sharing common entities). In general, a hybrid extraction approach that mixes multiple metrics (e.g., through a linear combination) and also includes such constraints might provide better results.

3.5.3 Evaluation Methods

Benchmark Data Sets In general, most works collect their own data or use a subset of a pre-existing news data repository. For example, some use data sets from TDT literature [68, 69, 265, 377], DUC/TAC conferences [71], or other general news repositories [73]. Most works do not publish their data sets. However, there is a subset of timeline summarization works that have provided evaluation data sets that have been adopted in several works as benchmark data. We present these data sets in Table 3.4. These data sets are appropriate for the “events as sentences” resolution level that timeline summarization uses. However, they do not provide a direct way to evaluate methods that use other resolution levels. Furthermore, we note that there are no such benchmark data sets for the other resolution levels of the

narrative extraction tasks considered in this survey. The lack of appropriate benchmark data for the document-level and cluster-level resolutions makes comparing methods harder and makes replicability harder.

Table 3.4: Benchmark data in the timeline summarization works.

Data Set	Source	URL
Timeline 17	[350]	https://github.com/complementizer/news-tls
Crisis	[351]	https://github.com/complementizer/news-tls
COVID-TLS	[207]	https://github.com/MorenoLaQuatra/SDF-TLS
TLS-COVID19	[278]	https://github.com/LIAAD/tls-covid19
Entities	[134]	https://github.com/complementizer/news-tls
MTLS Data	[407]	https://yiyualt.github.io/mtlsdata/

Computational Metrics Limitations We note that most of the narratives discussed in this survey only consider the content (e.g., traditional information retrieval metrics) without accounting for the order nor the structure of the narrative. Some metrics consider ordering information, but only at a linear structure level. For example, story-level measures of coherence consider the connections between consecutive documents [319] or the dispersion coefficient which models story flow [60]. The ranking evaluation metrics also include some underlying notion of order, however, this notion is limited to a linear structure at best. The structural version of coverage based on event connections used by Bögel [46] is based on local connections only, but it does not account for the full structure of the graph. Thus, current metrics for the narrative extraction task are unable to deal with complex narrative structures.

Given this limitation, it would be ideal to consider metrics that account for both order and structure to provide a proper evaluation of a narrative. For linear narratives, it would be sufficient to consider content and order, as the structure itself is fixed. An approach to solve this would be to consider a metric based on weighted edit distance [378] as it considers both the order of the elements and their contents (by defining weights according to event similarities or an adaptation of the previously discussed metrics). For non-linear narratives, a similar approach could use graph edit distance [4] with custom costs, as this metric would consider structure, order, and content.

However, the previous proposal would be supervised, as we would need a narrative against which to compare the output. Devising an unsupervised approach is a more challenging issue, particularly for graph-based narrative representations. One alternative is to attempt to extend coherence and dispersion measures to such graphs. For example, given a directed graph with a single starting event and a single ending event, it could be possible to compute all routes from start to end and obtain a weighted average of the coherence or dispersion of these routes. However, for more complex graph structures it might be too costly in computational terms to do such computation. Designing an unsupervised evaluation metric

remains an open challenge.

Benchmark Tasks and User Evaluations User-based evaluations usually focus on subjective measurements rather than objective task performance. This is due to the lack of properly defined evaluation frameworks and benchmark tasks. Current approaches rely on micro-knowledge tasks—tasks focused on information retrieval—that evaluate the number of correct answers (i.e., user accuracy) over time, and macro-knowledge tasks—tasks focused on summarization—that indirectly evaluate the quality of the extracted narrative by measuring the quality of a user-generated summary. These approaches are limited and do not capture all the nuances associated with narrative sensemaking. Moreover, they do not cover more complex tasks beyond retrieval and summarization.

One possible solution would be to design more holistic evaluations based on different types of benchmark tasks. In particular, the use of Bloom’s taxonomy [45] could provide a useful framework to define such tasks as seen in other sensemaking applications [57] or cognitive tasks in general [80]. Another possible solution would be to borrow the concept of insight-based evaluations [271]. Rather than focusing on benchmark tasks with specifically defined tasks and correct answers, the evaluation would be open-ended and would focus on analyzing the insights generated by the users.

3.5.4 Practical Applications

Event-based news narrative extraction has several practical applications beyond journalistic analysis tasks. Most of these applications seek to help with the issue of *information overload* in different contexts [323]. We briefly discuss some potential applications explored or mentioned in some of the reviewed works.

Disaster Management Disaster management [394, 408, 414, 415] could benefit from using extraction approaches to keep track of disasters or other similarly negative incidents. In particular, to minimize losses caused by a disaster, one of the critical tasks in disaster management is to efficiently analyze and understand situation updates. Doing this requires effective methods to navigate a multitude of documents such as news or reports related to the disaster. Domain experts need to obtain condensed information about the disaster and its evolution [219]. Thus, news narrative extraction could help experts to understand the evolving situation and devise a proper strategy.

Open Source Intelligence Open source intelligence (OSINT) is intelligence that is synthesized using publicly available data [111]. While OSINT data sources leverage more than just traditional news articles [135], OSINT could still benefit from news narrative extraction techniques. In particular, news narrative extraction methods could help intelligence analysts explore the information landscape and find key events [186]. Furthermore, these techniques could help analysts in prediction tasks by providing support and evidence [60].

Misinformation and Fact-Checking News narrative extraction methods could aid fact-

checkers in their tasks by providing them with an overview of the current narrative and highlighting key relevant events [186]. However, current methods do not include explicit ways to model misleading or outright false information.

Financial Markets News narrative extraction could aid financial analysts to understand the information landscape [59]. For example, market news is regarded as an important data source in the context of financial analysis [59]. In particular, being able to understand and exploit the hidden information in the raw news data could help analysts adapt their strategies and reduce their financial risk.

3.5.5 Recent Trends and Open Challenges

Timeline Summarization Variations Recent works have proposed some variations on the traditional timeline summarization task. In particular, Duan et al. [98] proposed the *comparative* TLS task and Yu et al. [407] proposed the Multi-TLS task. These two works highlight the fact that simple linear representations of narratives are naturally limiting unless applied to the most simple of narratives. Thus, the creation of similar tasks to address some of the shortcomings of these representations is a natural progression. However, it raises the question of whether these extensions would benefit from borrowing elements from the methods that use more complex representations discussed in this survey. A natural extension would be to consider a graph-based representation that allows for multiple storylines and comparisons without further modifications. This approach would address both the comparative TLS and MLTS tasks.

In this context, we note that most of the reviewed articles with a sentence-level event resolution used a linear structure (see Table 3.1). The only exceptions were the disaster storyline extraction systems [408, 414, 415] with their local tree representation. However, these methods are designed with a specific news topic in mind—disaster news—and are able to leverage specific characteristics of the topic (e.g., the disaster moves over time). Thus, it would not be possible to directly adapt it to other types of news without addressing this issue.

Furthermore, we note that there are no inherent limitations to sentence-level representations that prevent them from being extended beyond linear narratives, which makes the lack of graph-based approaches an opportunity for future research. Finally, while we did not find such a suitable graph-based approach in the traditional news domain, there is one example from the social media domain—which has its own set of challenges in terms of narrative extraction—that can be found in Ansah et al. [18]. This work proposes a tree-based narrative representation with sentence-level event representation using tweets. This approach extends the traditional TLS by allowing divergent storylines to emerge instead of just a single timeline. Such an approach could be adapted to traditional news narrative extraction.

Multi-resolution Methods Currently, all the narrative extraction approaches that we

reviewed work on a singular resolution level (sentences, documents, or clusters). Existing attempts at multiple resolution levels only change the scope of the data [323, 324] (i.e., applying the method again on a new subset of the data), they do not seek to change the underlying event resolution. Another perspective corresponds to the multi-level presentations of disaster storylines by Zhou et al. and Yuan et al. [408, 414, 415], which use global and local levels to represent the narrative. However, the underlying event representation remains the same and no efforts have been made to make a model that handles multiple levels of event resolution. Developing models that provide a multi-resolution approach remains an open challenge.

Interactivity Most works on news narrative extraction provide surface-level interactions [323, 324, 343] such as re-arranging elements and changing the layout, showing details on demand (e.g., all details about a news article), zooming, or performing basic filtering, highlighting, and searching. However, there is still a need for better interaction models that give users more control and feedback when exploring and manipulating the narrative. Some models [319, 320] allow more in-depth refinement by letting the user specify elements that need to be changed and then evaluating all possible replacement and insertion actions. Building upon this feature-based feedback, Shahaf et al. [321] designed a method to learn a *personalized coverage function* that can be optimized to find a personalized narrative.

Another approach by Bögel et al. [46] allows parametric interaction to modify the extracted graph in real time, helping the user understand how the narrative changes based on the parameters. However, this approach requires the users to understand the underlying model parameters. In this context, *semantic interactions* could be useful to aid users modify the model without deep understanding of the underlying parameters. Semantic interactions [382] are used in sensemaking applications to directly reflect the analytical thought process of analysts about data (e.g., by using information about how analysts organize documents or highlight text), as opposed to parametric interaction that manipulates model parameters (e.g., sliders and keyword weights). Thus, capturing a user model through semantic interaction could lead to a better narrative model.

Misinformation in News Recent works have highlighted the need for future work to model source bias, information validity, transparency, and credibility as an effort to model and counter misinformation [186, 208]. Existing narrative representations could be enhanced by including additional attributes in their representations and extraction algorithms.

Works on misinformation detection focus on the propagation structure and content to determine whether a certain article or publication contains misinformation [140, 390]. Other methods rely on crowdsourcing [140] to detect misinformative content early. However, these methods do not model misinformation as part of an overarching narrative. Instead, they focus on local elements (e.g., a specific event). Thus, a holistic narrative approach could be useful in this context.

The issue of misinformation is also highly relevant for a series of recent works on disaster tracking by using news narrative extraction [394, 408, 415]. However, none of these methods

address this issue and rely on the underlying assumption that the set does not contain false or misleading information. Thus, creating a narrative extraction model that accounts for misinformation would be of vital importance in the context of disaster tracking.

3.6 Chapter Conclusions

This literature review focused on narrative extraction and its related tasks of representation and analysis, synthesizing findings from 54 studies and identifying recurring types of representational structures, extraction criteria, and evaluation metrics. We further analyzed the articles and identified a series of recent trends, open challenges, and potential research directions. In terms of limitations, we highlight the lack of benchmark data sets, the need for better evaluation metrics that are capable of handling complex narratives properly, the high computational costs of most methods, and the lack of standardized benchmark tasks for user-based evaluations. In terms of open challenges, we note the need for better interaction models that allow users to explore the narrative with more control. Finally, we note that current models do not handle misleading or false content, a rising challenge as misinformation compounds with information overload to make understanding the information landscape even harder.

As with other literature reviews, this work has some limitations related to the inclusion and exclusion of relevant pieces of work. In particular, we used the Scopus and Web of Science databases as our initial sources. Previous studies have shown that Scopus and Web of Science are inclusive and extensive sources for literature reviews [113]. Regardless, multiple studies were not included in our initial results and thus we had to include them through other means, such as extracting relevant citations from reviewed works. Moreover, the choice of keywords might have caused some studies that use different terminology to not show up in our searches.

Chapter 4

Main Event Descriptors and the Inverted Pyramid Structure

The inverted pyramid is a basic structure of news reporting used by journalists to convey information and it is considered a key element of objectivity in news reporting. A core part of this structure is the position of the *main event descriptors* in the article—the answers to the 5W1H questions: Who? What? Where? When? Why? and How?—which allow us to represent events in detail. Our goal in this chapter is two-fold. First, we design an extract method to extract the main event descriptors from a news article, as that will allow us to create a representation of the event reported by the article. Next, we propose the Inverted Pyramid Score to evaluate how well a news article follows the inverted pyramid structure based on the position of the main event descriptors and news summarization. We evaluate our proposed methods on a proprietary data set of Associated Press news articles across breaking and non-breaking news spanning two topics—political and business. Our results show that the method works at distinguishing the structural differences between breaking and non-breaking news. In particular, our results confirm that breaking news articles are more likely to follow the inverted pyramid structure.

4.1 Introduction

The inverted pyramid structure—a system of news writing that arranges facts in descending order of importance—has been a cornerstone of journalism since the late 19th century [150]. This style of structuring information emphasizes fact-based reporting and neutrality—two of the key components asserting objectivity in journalistic writing [256]. These elements are also particularly important for hard news reports that require timely reporting and are characterized by high news value (e.g., breaking news stories on political topics). Moreover, scholars have found that an inverted pyramid information structure is a distinctive feature of real journalistic reports; whereas fake news stories often rely on opinion-based reporting and at times are written in structurally ambiguous ways [309]. Thus, determining how well a news story fits the inverted pyramid arrangement could be useful in determining whether the report follows journalistic standards. This chapter introduces one such measure for structural analysis of news writing—the Inverted Pyramid Score (IPS).

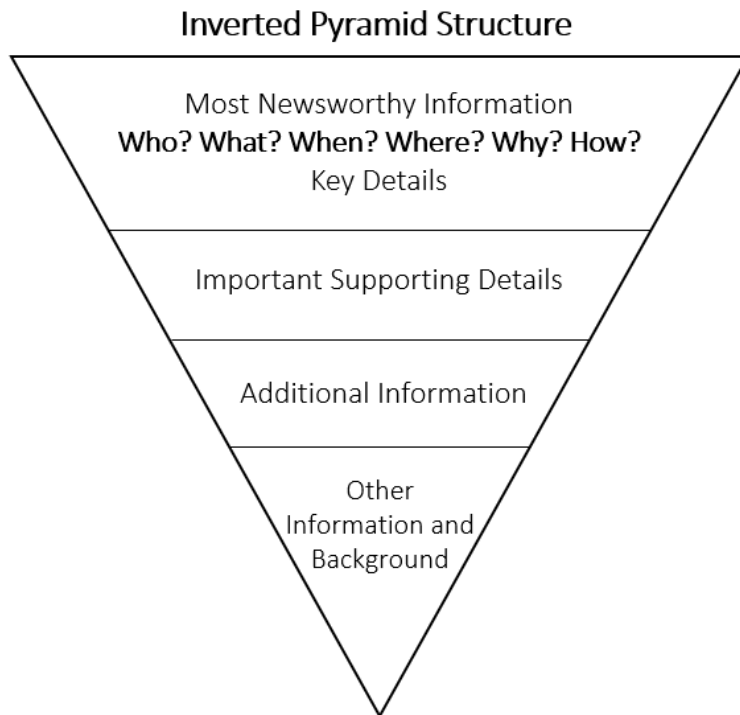


Figure 4.1: Diagram describing the inverted pyramid structure. The information is shown in descending order of importance, with the key details at the top.

Prior work on computationally analyzing the inverted pyramid structure includes Zhang and Liu’s visual and statistical exploration using rhetorical structure theory [410] and Dai et al.’s classifier-based approach to detect structure in news articles using various lexical, syntactic and semantic features [84]. While these approaches exploit the rhetorical and syntactic structures in news reporting, they do not leverage the two distinctive elements of the inverted pyramid reporting style—summarizing and compressing the most newsworthy aspects of the story at the very start [150]. Thus, we design our IPS scoring method by leveraging these two key elements and comprises of the following two components:

1. **Main event descriptor locations:** The most newsworthy aspects of a story (i.e., the 5W1H questions: who, what, when, where, why and how) are compressed in the opening paragraphs (see Figure 4.2). The answers to the 5W1H questions describe the main event of the news article, providing us with the ability to represent the event reported by a news article.
2. **Summary similarity:** The opening paragraphs summarize the story in such a way that it is possible to cut out the last paragraphs without losing key information [248]. Thus, any summary of an inverted pyramid news article should be similar to the Opening Paragraphs (OP), composed of the headline, lead, and second paragraph (see

Figure 4.2). We consider the 2nd paragraph in addition to the headline and the lead because it may include some key details, such as the answers to 5W1H questions [150].

We validate and show the effectiveness of our methods on a proprietary data set of Associated Press’ News articles, comprising news from December 2016 to December 2017. Our experiments demonstrate that our proposed method is capable of evaluating the inverted pyramid structure and, on average, can distinguish the structural differences between breaking and non-breaking news. This is the first step towards evaluating objectivity by the use of the inverted pyramid structure.



Figure 4.2: News article [242] showing the Opening Paragraphs (OP) and the rest of the article. The highlighted phrases show the answers to the 5W1H questions that define a news event: *What* happened? *Who* is involved? *When* and *Where* did it happen? *Why* and *How* did it happen?

4.2 Background and Related Work

We divide this section into three parts. First, we delve into the different news structures used in written journalism. Next, we focus on the two components that form the core of our IPS calculation, namely 5W1H extraction and text summarization.

4.2.1 Structure of News Articles

There are many ways to structure a news article. We describe the four most common structures found in the literature [84, 150, 287]. In the *Inverted Pyramid* structure, the article presents content in descending order of importance with key events placed first and additional details discussed later. In the *Kabob*, the article starts with an anecdote to capture the reader’s attention, then introduces the key events and main story, followed by a general discussion with more details. In the *Martini Glass*, the article relies on narrative chronology beginning with a summary of the main event. The article starts with the inverted pyramid structure but then transitions into a narrative story following a chronological order. Lastly, in the *Narrative* structure, the article presents a chronological sequence of events with more details than usual news articles.

Each one of these structures has distinctive features that distinguish them from one another, in particular, writing style and presentation order [84]. While the last three structures are useful for some news, none enjoy the popularity of the inverted pyramid, the most common structure in written journalism. Furthermore, the inverted pyramid is considered a fundamental pillar of objectivity [256] as well as a key feature of professional journalistic news reporting as opposed to fake news [330]. Hence, for the purposes of this study, we focus on analyzing the inverted pyramid structure, instead of attempting to classify among all possible journalistic structures.

4.2.2 Main Event Descriptor Extraction

There are many methods to extract main event descriptors (i.e., the 5W1H answers). We present some approaches in this subsection.

Most works are purely based on rules that leverage lexical, syntactic and semantic information to obtain answer candidates. *Verb-based approaches* work by identifying the main action in a sentence or text, represented by a verb [277]. Once the method has identified the main verb, it extracts the arguments associated with it to find the main event descriptors (e.g., the subject of the verb). *Semantic role labeling* identifies semantic predicates at a sentence level [65]. Then, it identifies syntactic components through shallow parsing and assigns them a semantic role in the predicate. This method leverages syntactic relationships to identify text semantics. *Machine learning classifiers* can be trained and used to extract main event descriptors [396]. In particular, they can predict the arguments of the main predicate of a sentence (i.e., the 5W1H answers). However, due to the scarcity of annotated data sets [147] there is less work on methods that exploit annotations to improve their results.

Finally, we highlight the recent works of Hamborg et al. [146, 147] which present the development of an open-source system for 5W1H extraction (Giveme5W1H) along with a gold-standard data set. We have used their extraction and scoring methods as a guide for our own 5W1H extraction system.

4.2.3 News Summarization

News summarization is an extensively studied application of natural language processing. It comprises of two main approaches: abstractive and extractive. While abstractive methods rephrase and compress the original text to create the summary, extractive methods select key sentences from the text to build the summary [177].

For our work, we consider using an extractive summarization algorithm. In particular, we use TextRank, an algorithm that has been successfully used in multiple applications [31], is domain-independent and does not require deep linguistic knowledge [251].

Finally, news summarization is closely related to 5W1H extraction, since answers to the main event descriptor questions can be used to provide an explicit summary of the main event [146]. In essence, both summaries and main event descriptors are performing the same task: they distill the article into a simpler representation. In the context of the inverted pyramid structure, the results from both tasks should always be related to the beginning of the article.

4.3 Data Collection and Annotation

We start by describing the data set used, present the subset employed for the IPS evaluation and describe the annotated sample used for evaluating the main event descriptor extraction (5W1H).

4.3.1 AP News Data

Our work is based on a proprietary data set from the Associated Press News (AP News) spanning a full year of news articles from December 2016 to December 2017, a total of 65,535 articles. The Associated Press was instrumental in creating the inverted pyramid structure [367] and continues to use it for reporting. Hence, the AP news data is ideal for testing our inverted pyramid scoring method.

Each news article in our data set includes information about the topics of the news, referred to as subject tags or categories, spanning from serious topics like *Business* to lightweight topics, such as *Entertainment*. Following Bakshy et al’s [27] hard-soft news classification scheme, the following AP News categories are likely to be hard news: *Science*, *Politics*, *Business*, *Health*, and *Weather*. Hard news is characterized by a high level of newsworthiness or news value and require timely publication (e.g., politics or business news). Whereas, soft news has a low level of informational value (e.g., entertainment news) and does not need immediate publication [216]. The AP data also contains a label indicating whether a news article is *Breaking* or *Non-Breaking* news. Breaking news articles usually follow the inverted

pyramid structure and we expect them to have a higher IPS compared to non-breaking news articles. Thus, if our scoring method works, it should be able to score breaking news with significantly higher IPS value than non-breaking new articles.

What genre of news should we evaluate with our scoring method? We decided to test our scoring method on “hard breaking news” since the inverted pyramid is seen as the distinctive feature of hard news reports [346]. Furthermore, we are interested in applying IPS as a means of establishing journalistic standards and as a method to contrast it with fake news. Hence, focusing on hard news makes sense as they are often subjected to misinformation.

We first filtered articles with missing fields (e.g., countries, tags), non-English articles, and retained those that were tagged as “United States”. We focused on a single month to reduce variability in the news articles and to minimize computational costs. In particular, we chose November 2017 as it is the most recent month with a high number of articles (5,045 articles, compared to December 2017 which only has 945 articles). For this first phase of our study, we focus on the two most frequent hard news categories in our data set: *Politics* and *Business*. Our final sample comprised 1,529 articles.

4.3.2 Annotations

One of the key steps in building our scoring method is extracting answers to the 5W1H main event descriptor questions from an article. How well does our 5W1H extraction work? To answer, we extracted a random sample of 30 breaking news from our data (15 political and 15 business articles) and obtained annotations for the 5W1H answers from experts trained in writing journalistic articles following the inverted pyramid. Specifically, we asked senior journalism students to assess our 5W1H answers who are extensively trained in using the inverted pyramid and writing 5W1H answers.

Journalism students were provided a questionnaire with a 3-point Likert scale to evaluate the descriptors. Additionally, if a descriptor was not present in the article, students could mark it as N/A. Figure 4.3 shows an example question. We received six evaluations for each article, totaling 180 annotations for our sample of 30 articles. For each descriptor per article, we averaged the expert-assigned scores. Next, we averaged these results again over all articles to get the final evaluation for each question.

4.4 Methods

Here we present our methods, starting with our pre-processing steps, then the definition of the Inverted Pyramid Score, the details of the 5W1H extraction method, and finally the summarization method.

Hawaii group wants to defend licenses for foreign fishermen

HONOLULU (AP) — A group representing **Hawaii** commercial fishermen has **filed a court motion** to defend the state’s practice of giving fishing licenses to foreign workers.

The Hawaii Longline Association sought court permission **last week** to argue its side after **a Maui resident asked a judge to declare that only those lawfully admitted to the United States should receive commercial fishing licenses.**

[...]

	Correct/good	Partially correct	Incorrect/bad	N/A
Who	☒	☐	☐	☐
What	☒	☐	☐	☐
When	☐	☐	☒	☐
Where	☐	☒	☐	☐
Why	☒	☐	☐	☐
How	☐	☒	☐	☐

Publication Date
2017-11-01

Category
Business

Article Metrics
Inverted Pyramid Score: 90.00%
Subjectivity Score: 23.24%
Polarity: 2.3%

Event Summary

Who: Hawaii group (A group representing Hawaii comercial fishermen)

What: wants to defend licenses for foreign fishermen

When: last week

Where: Hawaii

Why: a Maui resident asked a judge to declare that only those lawfully admitted to the United States should receive comercial fishing licenses.

How: filed a court motion

Figure 4.3: Example of a question from the evaluation questionnaire. The questionnaire showed the full article with highlighted answers, some additional information about the article on the top left, and the answers below that. The students had to select one answer for each 5W1H question after reading the article and the proposed answers.

4.4.1 Data Preprocessing

We start by applying standard preprocessing techniques, such as tokenization, sentence splitting, part-of-speech tagging, dependency parsing and named entity recognition. Additionally, we use neuralcoref to handle coreference resolution. As an example of coreference resolution, consider Figure 4.2. “Hawaii group” is the answer to WHO, but this entity is also mentioned in the text as a “group representing Hawaii commercial fishermen.” This is a coreference because they refer to the same entity.

4.4.2 Inverted Pyramid Score

The Inverted Pyramid Score has two main components: 1) main descriptor locations and 2) summary similarity. The IPS answers the following question: how well does a news article follow journalistic standards? And, in particular, how well does an article fit the inverted pyramid structure?

Figure 4.4 presents a general overview of our method. To find the IPS, we first apply

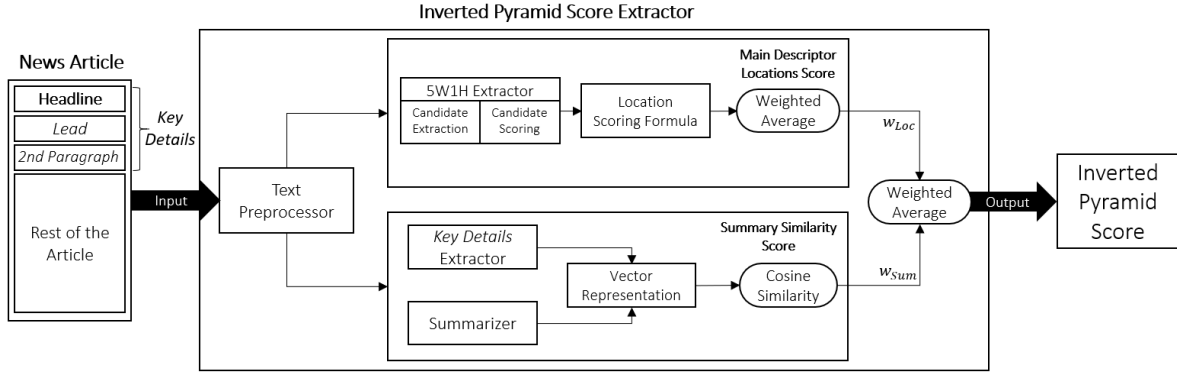


Figure 4.4: Diagram showing the computation of the Inverted Pyramid Score (IPS) of a news article. The first step is preprocessing, then we compute the *main event descriptor locations* score and the *summary similarity* score. For the main descriptor locations score we extract the position of the main event descriptors (5W1H answers), then we get the weighted average of these scores. For the summary similarity score, we get the similarity between an extractive summary of the article and the key details. We get the final IPS using the weighted average of the previous scores.

standard preprocessing steps on our data set. Next, we compute a score for the **main descriptor locations** and the **summary similarity**. Both scores are computed with respect to the opening paragraphs comprising the headline, lead, and 2nd paragraph. Finally, we compute the final IPS as the weighted average of the two component scores. We assigned a higher weight to main descriptor locations since the 5W1H questions guide the writing of the inverted pyramid [131].

4.4.3 Main Descriptor Locations

The first component of the IPS models how well the article captures the main event descriptors. In particular, we first check whether all the main event descriptors (i.e., 5W1H answers) are present in the article. Our 5W1H main event extraction comprises two steps: extracting all possible candidates for 5W1H and scoring candidates to find the best match. Next, we ensure that the main event descriptors appear early in the text. We assign a score based on the descriptor’s location in the article, penalizing those appearing below the headline sentence.

In particular, we build our main event descriptor extraction module by following the generic 5W1H architecture described by Hamborg et al., specifically their implementation of Giveme5W1H . We extend their implementation by including additional rules and refining the candidate scoring mechanism.

Extracting possible candidates for 5W1H

What & Who. We find all sentences of the form NP-VP-NP (e.g., *[The cat]_{NP} [quickly climbed]_{VP} [the apple tree]_{NP}*). Usually, the answer to WHO is contained in the first NP of these structures and the answer to WHAT is contained in the VP-NP part. This strategy works for most journalistic articles since they are usually written in the active voice. We also exclude candidates that contain attribution verbs in the VP (e.g., *said*, *told*). Since attributions usually only offer supporting information, they are unlikely to contain the answer to WHAT and WHO.

When. To extract WHEN candidates, we parse regular dates and relative dates. We also check for dates that escaped automatic date parsing (e.g., “Christmas weekend”). We handle them by adding manual rules. Furthermore, we identify additional time noun phrases using a dictionary of time nouns [240].

Where. For WHERE candidates, we find all named entities that are tagged as location. We geolocate them using the OpenCage API.

Why. We search for three elements to extract WHY candidates: adverbs that express causal relationships, causal conjunctions, NP-VP-NP structures with causal verbs and auxiliary verbs that can be used for showing causes but aren’t specific enough (e.g., “to be,” as in “the airplane failure was a mechanical issue.”).

How. We extract sentences that use one of the copulative conjunctions (usually the phrase after the conjunction is the HOW). We also find NP-VP-NP phrases that have adverbs and adjectives (since these modifiers can reflect the answer to HOW). Sometimes a candidate is found in both categories, in this case we mark it as a double candidate from both sources and we assign it a higher score.

Candidate scoring to find the best match

After extracting the potential candidates, we score and rank them to get the final answer for each 5W1H. We designed all scores to be between 0 and 1. We score candidates based on a combination of the following criteria: position, type, frequency, precision, length, and other candidate-specific scoring criteria.

Position Score. For all 5W1H questions, we assign a high score when candidates are found early in the text. For occurrences in the first sentence of the document (or headline), we assign a score of 1. For occurrences in subsequent sentences, the score follows an exponential decay, decreasing with an increase in position, p . Specifically, $S_p(C) = e^{(-dp)}$ with decay coefficient $d > 0$. To illustrate, let’s refer to Figure 4.2. Considering logarithmic decay, $d = \log(2)$, we divide the score by half whenever we move farther away from the headline. Thus, the WHO candidate ($p = 0$) will be scored with 1 and the WHEN candidate with 0.25 ($p = 2$).

Type Score. Scoring based on candidate type, such as proper or common noun, date or time, etc., depends on the 5W1H question being answered. For WHO, it refers to whether the

candidate is a named entity (i.e., a proper noun). For example, if the extracted candidate for WHO is a named entity, we score it as 1, otherwise it is scored as 0. For WHEN, *type* refers to whether the candidate is a proper date or a vague expression. For WHERE, it refers to the type of location (e.g., geopolitical entities, geographical locations, man-made structures, or organizations which can be used to refer to places in some cases). For WHY and HOW, we score candidates based on whether it is expressed through an NP-VP-NP pattern or conjunction or a combination of both.

Frequency Score. For all questions, except WHY and HOW, we rank-score candidates by their frequency of occurrence in the article. The highest frequency candidate is scored as 1. If the candidate is a named entity, we count all its coreferences, otherwise, we simply count the raw occurrences. For example, consider “Hawaii” and “United States” as WHERE candidates for the article in Figure 4.2. If we only consider the parts shown in Figure 4.2, then the article mentions the first candidate four times and the second candidate only once. We normalize the counts by the highest frequency and assign a score of 1 to “Hawaii” and 1/4 to “United States.” Formally, let $\mathcal{S}$ be the set of all scored candidates, then the score is defined as

$$S_f(C) = \frac{\text{Freq}(C)}{\max_{s \in \mathcal{S}} (\text{Freq}(s))}.$$

Precision Score. For WHERE and WHEN, we consider the **Precision** of the candidate. For example, a date with an exact time is ranked higher than a vague phrase like “election time” and “London” is ranked higher than “UK” because it is a more precise location.

For WHEN, this is implemented as a exact date times having a higher score (1), dates without specific times have a lower score (0.6), and abstract dates that cannot be parsed to a proper date have the lowest score (0).

For WHERE, the computation of precision is more involved. In particular, we will use the area size as a proxy for precision. This area can be computed using the bounding box given by the geolocation API (which returns coordinates for the opposite corners of the bounding box). As per Hamborg et al. [147], we use a logarithmic scale to compute this score. For simplicity, we approximate areas as if they were squares (i.e., we assume the Earth is locally flat). For small bounding boxes this works well, for bigger boxes this approximation is not as effective, but it should still provide a good enough approximation. Mathematically, let $d(C)$ be the distance between the two opposite corners of the bounding box of candidate C . Then the area is given by $A(C) = d(C)^2/2$ so in log-scale we get $\log A(C) = 2 \log(d(C)) - \log(2)$.

Moreover, if the type of the candidate is ORG we will make this score a zero to penalize it for not being a “proper location”, they are then excluded from all calculations that follow. We rank all non-ORG using the area and get their strict percentile ($<$ not $\leq$) as a number from 0 to 1. The final score is given by

$$S_s(C) = \begin{cases} 0 & \text{if } C \text{ is of type ORG} \\ 1 - \text{StrictPercentile}(\log A(C)) & \text{otherwise.} \end{cases}$$

Note that this score is designed so the smallest areas have the highest score and the biggest areas have the lowest non-zero score (only ORGs have a zero score).

Length Score. For WHY and HOW, we consider the **Length** of the candidate. We prefer longer explanations for the cause and method. To implement this, we count the number of words in the candidate and divide by the maximum count in all candidates. More formally, let $\mathcal{S}$ be the set of all WHY candidates, we compute the length score as

$$S_t(C) = \frac{\text{Length}(C)}{\max_{s \in \mathcal{S}}(\text{Length}(s))}.$$

Moreover, we add a redundancy penalty if the candidate repeats the answer to WHAT or if we get the same answer for WHY and HOW. More specifically, if the answer to WHY is part of the WHAT (or vice versa) then this score is penalized with a 75% reduction. This is to prevent redundant answers of the form $p \rightarrow p$.

Other Scoring Criteria. For WHEN, we also score candidates by *distance to publication date*, preferring dates closer to the publication date. This score is based on a quadratic formula, with the date farthest away having a score of 0 and the closest date having a score of 1 point. We only use this score for dates that can be properly parsed. For unparseable candidates we assign a score of 0. We note that the distance to the publication dates are taken in absolute value so we can handle both future and past dates. The score is defined as

$$S_d(C) = \begin{cases} \left(\frac{\text{MaxDistance} - \text{TimeDistance}(C)}{\text{MaxDistance} - \text{MinDistance}} \right)^2 & \text{if parseable} \\ 0 & \text{otherwise.} \end{cases}$$

For WHERE, we score candidates by *clustering*. We assign a higher score if a candidate is close to the other candidates. For example, if most locations are in Germany, then we would assign less score to a random location in Japan. To implement this score, we compute the pairwise distances of all locations detected in the text that are not labeled as ORG. If the type of the candidate is ORG we assign this score as 0. The coordinates are obtained using the geolocation API and then we compute the geodesic distance between coordinates. Then, we compute the average distance from each location to all the others. We use the complement of their strict percentile to score the candidates. Let $\text{AvgDistance}(C)$ be the average distance computed for a candidate C , the score will be given by

$$S_d(C) = \begin{cases} 0 & \text{if } C \text{ is of type ORG} \\ 1 - \text{StrictPercentile}(\text{AvgDistance}(C)) & \text{otherwise.} \end{cases}$$

For HOW, we score candidates by *modifier frequency*, which counts the number of adverbs and adjectives used by the candidate. In particular, if our candidate was obtained based on adverb or adjective keywords, we count how many modifiers (adverbs and adjectives) are

contained in the candidate. Then, we take this frequency and normalize it by the maximum count of modifiers from all candidates. If the candidate was obtained based on a conjunction keyword, then this score gets assigned to 0. Let $\mathcal{S}$ be the set of HOW candidates, then we can compute this as

$$S_f(C) = \begin{cases} 0 & \text{if Source is (only) Conjunction} \\ \frac{\text{Freq}_{\text{ADJ}+\text{ADV}}(C)}{\max_{s \in \mathcal{S}}(\text{Freq}_{\text{ADJ}+\text{ADV}}(s))} & \text{otherwise.} \end{cases}$$

Final Location Scores. Finally, we assign the location scores for each main event descriptor using the following criteria: if an article follows an inverted pyramid structure, it should provide answers to the 5W1H questions in the OP (see Figure 4.2). Thus, if we find the answers there, we assign a high IPS. While the headline and lead are usually one sentence long each [150], the 2nd paragraph can have at most three sentences. We found this maximum length by analyzing breaking news articles in the data set. Hence, for the purposes of our estimation, we consider the OP to be the first 5 sentences of an article. We give a full score if all 5W1H descriptors are contained in the OP. Otherwise, we apply an exponential penalty by location of each descriptor. More formally, considering the headline index to be 0, for each descriptor D ,

$$\text{LocScore}(D) = \begin{cases} 2^{4-\max(4, \text{Location}(D))} & \text{if answer found} \\ 0 & \text{if answer not found.} \end{cases}$$

Finally, we obtain a weighted average of all the location scores. Since HOW and WHY are not necessarily present, and even humans may have problems extracting them, we assign them a lower weight than the other descriptors.

4.4.4 Summarization

The second component of the IPS models how well an article is summarized by the OP. By definition, an article following an inverted pyramid structure must be summarizable by removing everything except the OP—the headline, lead, and 2nd paragraph. Note how in Figure 4.2 the OP contains all relevant information about the news story. Hence, our generated summary should be similar to the OP. Thus, we implement our summary similarity module by comparing the summary of the full article with the OP. First, we summarize the full article using an extractive summarization algorithm—TextRank. TextRank ranks an article by the most important sentences and then uses those to build the summary. Next, we compare the full article summary and the OP by comparing the language representations of the two. In particular, we do this using Spacy and their pre-trained `en_core_web_lg` model. This model uses GloVe vectors and it was trained with a multi-task CNN on blogs, news, and comments [158]. We average all the word vectors contained in a text to get its final

representation. Finally, we compute the summarization score using the Cosine similarity distance between the vector representations of the OP and the summary.

4.5 Results and Discussion

Here we show our main findings and discussions. We begin by presenting the evaluation of our main event descriptors extractor. Next, we report the results on the November 2017 AP News articles, showing the IPS distributions for breaking and non-breaking news.

4.5.1 5W1H Extraction

Table 4.1 shows the evaluation results of our 5W1H method. We find that our extractor is capable of obtaining the right answers for the basic 4W with 78% accuracy on average. Out of the four basic descriptors, our method systematically extracted better results for WHERE in this data set. This could be attributed in turn to the date-line being explicitly included in AP News articles.

Table 4.1: 5W1H evaluation results for the breaking news in the AP data set by subject category ($\pm$ standard errors).

Question	Business	Politics	<i>Total</i>
Who	0.74 ± 0.06	0.77 ± 0.07	0.76 ± 0.04
What	0.79 ± 0.05	0.73 ± 0.06	0.76 ± 0.04
When	0.71 ± 0.05	0.83 ± 0.05	0.77 ± 0.04
Where	0.87 ± 0.04	0.81 ± 0.04	0.84 ± 0.03
Why	0.42 ± 0.08	0.51 ± 0.06	0.46 ± 0.05
How	0.46 ± 0.07	0.42 ± 0.07	0.44 ± 0.05
<i>Avg (Total)</i>	0.66 ± 0.08	0.68 ± 0.07	0.67 ± 0.07
<i>Avg (4W)</i>	0.78 ± 0.04	0.78 ± 0.02	0.78 ± 0.02

However, for the full main event descriptors we only achieve 67% average accuracy. This reduction in accuracy makes sense considering the inherent difficulty of extracting the causes and methods from news articles. Even though the accuracy for WHY and HOW is still low compared to the other questions, our method is on par with the state-of-the-art.

As a baseline for comparison, Giveme5W1H gets 0.73 accuracy for all descriptors and 0.82 for the basic 4W on a BBC news data set [147]. However, it is hard to draw a direct comparison because of differences in the background of the annotators (journalism students vs IT students) and of data sets (AP News vs BBC).

4.5.2 Inverted Pyramid Score

After testing the main event descriptor extractor on the previous sample we turn to the main task. Using our full data set, we compute the IPS of each article and show its distribution and basic statistics in Figure 4.5. In general, a higher IPS means that the articles adhere more to the inverted pyramid structure. Thus, these results match our intuition that breaking news usually follows the inverted pyramid structure. Non-breaking news shows more structural variety, as evidenced by their higher standard deviation and lower IPS.

It should be noted that both types of news have instances that with an almost perfect IPS. This makes sense because the inverted pyramid style can be used in both breaking and non-breaking news. However, it is typically used more for breaking news. Hence, the IPS of a non-breaking news article will be lower on average. Thus, we could leverage this difference in the distributions to use the IPS as a predictor to classify breaking and non-breaking news.

While our IPS method gets intuitively correct results on breaking and non-breaking news, there might be other factors that affect whether a news article is written using an inverted pyramid structure or something else. In particular, an important element is the writing style used in the article. Inverted pyramid news will likely follow an expository writing style rather than a narrative writing style [84]. Consequently, we could add a new component to our scoring method that accounted for writing style differences.

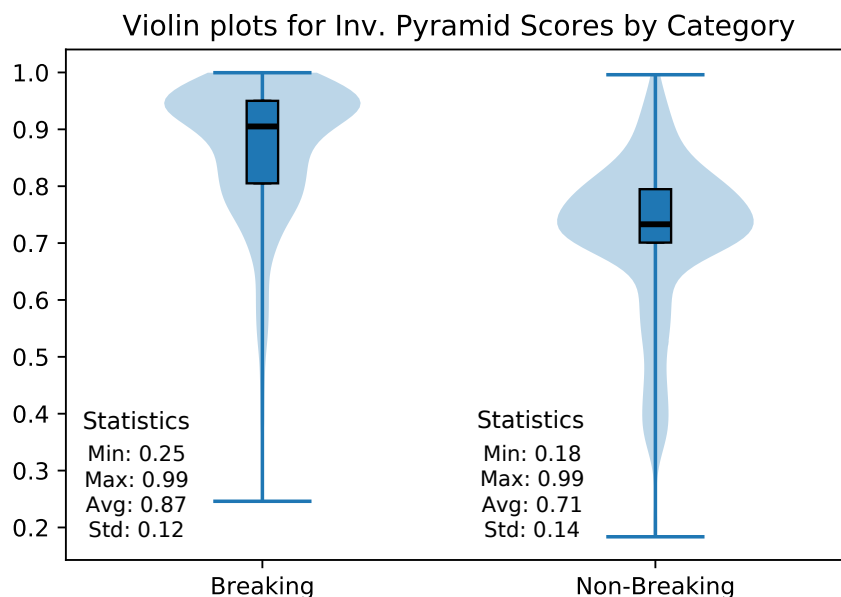


Figure 4.5: Plot of the IPS distribution and basic statistics for breaking and non-breaking news. On average, breaking news has a higher IPS than non-breaking news.

4.6 Chapter Conclusions

We have presented our work on evaluating the inverted pyramid structure using 5W1H extraction and summarization. Our analyses of results show that the methods works well, allowing us to distinguish between breaking and non-breaking news articles. In terms of improving our method, future work could include making improvements to the 5W1H extractor and using state-of-the-art summarization schemes tailored for news articles. In terms of potential applications, future work could extend the IPS to evaluate the different structures of news articles, not only restricting ourselves to the inverted pyramid. The current implementation could also be used, with additional features and descriptors, to provide a classification tool for breaking and non-breaking news.

Our long-term goal is to evaluate how fake news sources structure their articles, as well as comparing them to mainstream outlets. By finding these structural differences we hope to elucidate how fake news articles differ from regular news. In particular, the IPS is used as an indicator for credibility in later chapters to model misinformation. Moreover, our 5W1H method allows us to extract a representation of events based on the main event descriptors, which can be used in narrative representation methods.

Chapter 5

Narrative Maps Representation and Extraction Method

Narratives are fundamental to our perception of the world and are pervasive in all activities that involve the representation of events in time. Yet, modern online information systems do not incorporate narratives in their representation of events occurring over time. This chapter aims to bridge this gap, combining the theory of narrative representations with the data from modern online systems. We make three key contributions: a theory-driven *computational representation* of narratives, a novel *extraction algorithm* to obtain these representations from data, and an *evaluation* of our approach. In particular, given the effectiveness of visual metaphors, we employ a route map metaphor to design a narrative map representation. The narrative map representation illustrates the events and stories in the narrative as a series of landmarks and routes on the map. Each element of our representation is backed by a corresponding element from formal narrative theory, thus providing a solid theoretical background to our method. Our approach extracts the underlying graph structure of the narrative map using a novel optimization technique focused on maximizing coherence while respecting structural and coverage constraints. We showcase the effectiveness of our approach by performing a user evaluation to assess the quality of the representation, metaphor, and visualization. Evaluation results indicate that the Narrative Map representation is a powerful method to communicate complex narratives to individuals. Our findings have implications for intelligence analysts, computational journalists, and misinformation researchers.

5.1 Introduction

Narratives are fundamental to our understanding of the world [1] and are central to human relations [257]. They are the frameworks that enable humans to associate otherwise unconnected events [55] and play a key role in collaborative sensemaking in society [22, 386]. For decades, communication science scholars have emphasized the importance of narratives in human communication [1]. Narrative theorists have gone so far to claim that “narrative is the principled way in which our species organizes its understanding of time.”

In this line of research, narratives are defined as a *coherent system of interrelated stories* [145] and are considered to be omnipresent in all activities that involve the representation of

events in time. Through these systems of stories, humans produce a shared understanding of the world [342].

Despite the compelling case of narratives in human communication, modern-day online information systems do not incorporate narratives in their representation of events happening over time. Instead, most representations are still chronological order of events from the earliest to the latest; for example, consider Google News or Twitter’s news feed.

This chapter aims to bridge this gap, merging the theory behind narrative representations with data left in modern online systems (specifically news systems) while leveraging graph-driven approaches to identify such narrative structures. With theory as a guide, our central question is: Can we computationally represent and extract narratives from large-scale online data?

Our contribution is not just methodological; narrative representation has the potential to significantly impact online information consumers. Consider the unrelenting barrage of online news articles that users get exposed to or the ever-increasing problem of information overload in online systems [154, 319]. Not just regular users, even expert news investigators find it daunting to keep track of the evolving storylines, often losing track of the big picture. Given narratives are the primary mode in which humans apprehend time-based information [1], our paper offers just such a representation and the algorithmic approach to surface them directly from data.

In particular, we use a route map metaphor to provide an intuitive basis for our representation. This metaphor helps connect the elements of narrative theory and graph theory. Building upon this representation (RQ1), we propose a novel extraction algorithm to generate a narrative map representation automatically from data (RQ2). Our algorithmic approach leverages optimization techniques and graph-theoretical concepts to find all the elements in our narrative representation.

Specifically, this chapter addresses three research questions. First, we formalize the computational representation of narratives and ask:

RQ1 Representation: How can we computationally represent a narrative?

Based on a route map metaphor, we define a computational representation of a narrative grounded in formal narrative theory. Our representation maps concepts from narrative theory into graph theory elements. In particular, our proposal considers the use of a directed acyclic graph since this structure naturally reflects the temporal structure of stories. We also consider edge weights that allow us to evaluate the coherence of the storylines in the narrative map. Note that a narrative is composed of at least one storyline. After identifying the elements of our representation, we focus on how to extract these elements from data.

RQ2 Extraction: How can we extract a narrative representation from data?

To extract the narrative representation, we design an optimization approach that seeks to maximize the coherence of the narrative map subject to structural and topic coverage constraints. After generating the basic structure of the narrative map, we extract the main route (the most coherent path in the narrative) and the representative landmarks (sets of representative events from each alternative storyline). Using this information we can infer the themes—the recurrent or prominent ideas—of the narrative. Finally, we evaluate how well our proposed approach works.

RQ3 Evaluation: How well does our approach work to represent narratives?

We present a detailed evaluation of our proposed approach. In particular, we study the news narrative of the Coronavirus outbreak with three maps from the first three months of 2020, each generated after processing hundreds of news articles. We find that our representation provides a coherent narrative and that our metaphors are appropriate to model the narrative. We also show how our representation provides an easy way for users to answer questions such as: What is the main storyline—the main route—of this narrative? Apart from the main storyline, what other alternative storylines are there in the narrative? What are the major themes in the narrative?

Thus, by answering these research questions, our key contributions are:

- A **computational representation** for narratives, based on a mapping between narrative theory concepts and computational elements grounded in the route map metaphor. Our representation also includes a visualization based on the route map metaphor.
- An **extraction algorithm** to construct a narrative representation from data while using an optimization method and graph-theoretical concepts.
- An **evaluation** of our approach through a case study and a user study. In particular, we test our methods on the news narrative of the first three months of the Coronavirus outbreak.

Furthermore, we note that our work builds upon the foundation created by existing techniques. However, our proposed representation extends them by providing a theoretical framework based on formal narrative theory and the route map metaphor. The inclusion of this framework distinguishes our work from similar methods.

We believe that our approach has widespread implications. At its core, our narrative representation and our extraction method can help users understand the big picture of a narrative. Moreover, expert users, such as journalists, fact-checkers, and intelligence analysts can leverage our extraction method to explore the narrative landscape in the context of their domain of work. Furthermore, our narrative representation could be enhanced by considering additional elements, such as contextual information, credibility and political bias annotations,

or more detailed event descriptors. In particular, adding credibility and political bias information can help understand how misinformation spreads in news narratives.

In the rest of this section, we introduce the route map metaphor and provide a motivating example, explaining the reasoning behind the metaphor and detailing how it can be applied to analyze a narrative. In Section 2, we introduce important background concepts and describe related works. In Section 3, we show our algorithmic approach, including both the computational representation of narratives and the extraction algorithm itself. In Section 4, we present a case study of the Coronavirus narrative. Section 5 presents a user evaluation of our approach followed by a discussion of our results in Section 6. Finally, we end with the main conclusions of our work.

5.1.1 Route Map Metaphor

Humanity has relied on maps and cartography to understand its physical surroundings since antiquity. In particular, route maps and similar cartographic representations have been used to describe routes, that is, how to get from one physical place to another. These routes are usually accompanied by information of places and important locations along the way, as well as additional information about relatively less important elements of its surroundings. There can also be many routes joining two places, with differing characteristics and importance.

In general, the idea of using maps to represent stories is called information cartography [323]. In this chapter, we propose a way to represent a narrative (i.e., a system of stories) using a narrative route map (or simply narrative map). Instead of joining two physical places, we join two narrative places (a starting and ending event). The locations shown along the routes in our map should correspond to narrative elements.

Consider a route map that shows routes between two physical places as shown in Figure 5.1(a). The routes connect the “Start Stadium” and the “Classroom Hall” landmarks. Suppose a student needs to go to his engineering class at Classroom Hall after watching a practice session at Start Stadium. However, they are not sure which route to take, so they would take out their phone, enter their current location, and where they want to go into the route map application. The application then provides them with a route map as shown in Figure 5.1. We can identify multiple elements of interest in this map. In particular, it contains multiple interconnected *routes* between the *starting place* (Start Stadium) and the *ending place* (Classroom Hall).

One of these routes is highlighted as the *main route* (in blue) because it is the shortest path. All routes contain information about its distance and the time it takes to go through them, this allows users to decide which one to take (i.e., it is an *evaluation metric*), although in this particular example there is not much difference between the three routes.

We can also see information about different kinds of *landmarks* throughout each route. We note that there are several types of landmarks (e.g., restaurants, shops, parks, and

educational buildings). For example, the main route passes close to the South Market and goes through the Central Park, and then North Court before getting to the Classroom Hall.

In fact, we could identify each route by mentioning a *representative landmark*, such as the Chill Pond, the Central Park, and the Campus Shop for each route in our case. By identifying these representative landmarks, we can differentiate among these routes. Someone using the route map could make choices informed by these representative landmarks. For example, someone seeking to walk through a more scenic view might choose to go through the Chill Pond route, while someone interested in buying a snack or some materials could go through the Campus Shop route. Even though neither of these routes is the main one they still offer a potentially useful alternative.

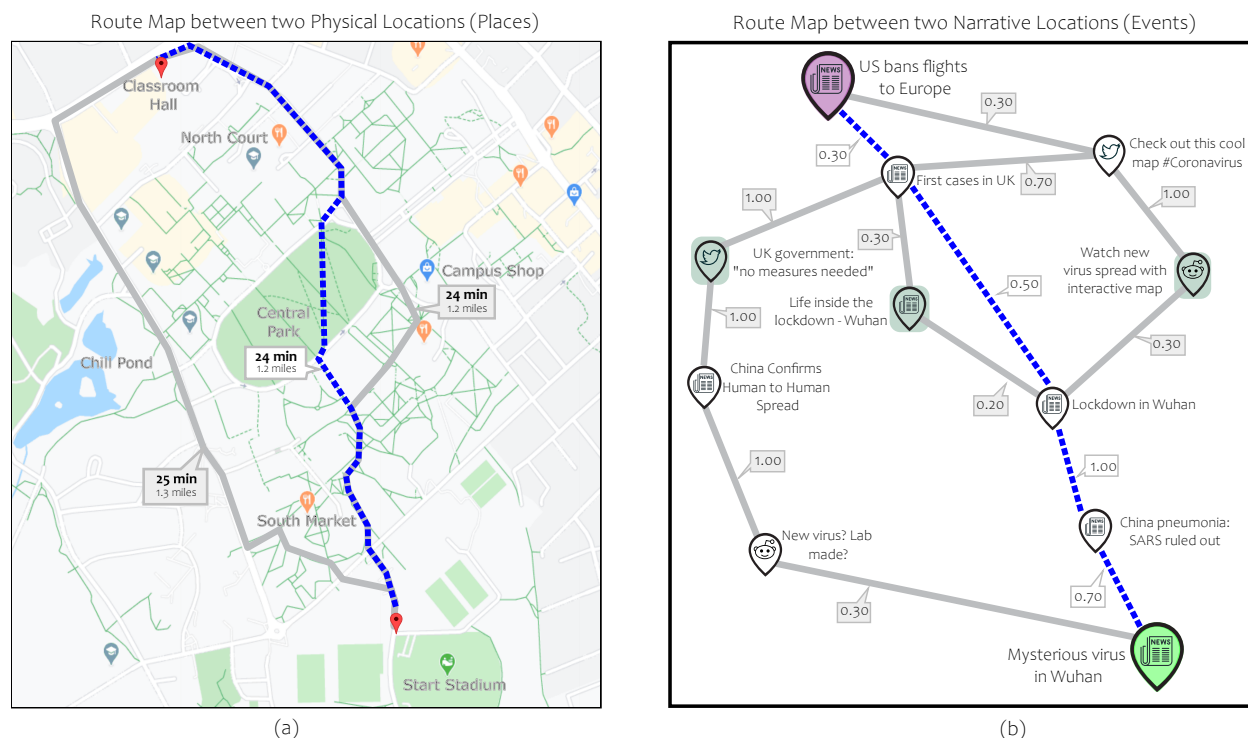


Figure 5.1: Illustration of the Route Map Metaphor. (a) A route map showing how to get from Start Stadium to Classroom Hall. (b) A narrative route map showing how to get from the Mysterious Virus in Wuhan event (starting event ) to the US banning flights to Europe (ending event ). We highlight some representative landmarks  for each route in the narrative map.

From the previous analysis we identify the following key elements of a route map: a starting place, an ending place, routes joining these places, landmarks along those routes, representative landmarks for parallel routes, an evaluation metric (e.g., distance or time) for those routes, and the main route obtained using that metric. Using these elements, we construct a metaphorical representation of narratives with a route map. Just as physical landscapes

can be represented using these route maps, we use the same elements to construct a route map representation of the narrative landscape surrounding a certain event.

According to some authors [1], narratives must contain at least two events: a starting event  and ending event . These events work similarly to the starting and ending place on the map and we need to choose them before creating the narrative map. We are interested in understanding how these two events are joined together by a story (i.e., a series of events and discourse elements). Thus, events would be akin to the landmarks of each route , and discourse elements would be the additional information associated with them. Furthermore, narratives can be seen as a system of stories, which in our metaphor correspond to the routes on the map . By doing this we gather new insights about the original events and the surrounding themes of the underlying stories. In addition, we can identify and characterize these storylines by finding a representative landmark for them . Finally, we can evaluate the stories  (e.g., by looking at its coherence) to find the main storyline of the narrative  (i.e., the main route).

See for example Figure 5.1(b), which shows a metaphoric rendition of a narrative route map representing the first month of the Coronavirus outbreak timeline. This map contains several sources of information (news and social media) and provides an overview of the evolving narrative landscape during this period. The dashed line  represents the main route (i.e., storyline). It reflects the development of the outbreak from its first cases as a mysterious virus to the US banning flights.

Various other elements are intertwined with this route, representing alternative paths towards the ending event. Representative landmarks , in this case, could be the tweets about the UK government indicating that no measures are needed, the news article about the life inside the lockdown in Wuhan, and the submission on Reddit about watching how the virus spreads.

As in the physical route map, these landmarks can inform the user of the map on where to go. For example, a data analyst could be interested in reading about visualizations related to the virus as examples of data visualizations. This analyst would presumably want to go through the interactive map route. A social science researcher might be interested in looking at how news sources presented life during lockdown at the start of the crisis. Therefore, they would likely be interested in reading the lockdown route. Finally, a political activist from the UK could be interested in reading about the UK response to the virus to build a case against the government's handling of the outbreak. Thus, they would probably find the route on the UK response more interesting.

In one of the paths we see conspiracy talks about the virus being artificial, then news about human to human spreading, leading into the spread into other countries. In the continuation of these paths, we see how the lack of appropriate government measures leads to the spread of the virus. Some paths are simply smaller deviations with interesting tidbits of information, such as the article on life during lockdown and people sharing an interactive map in social media.

This route map metaphor forms the intuitive basis of our computational representation and corresponding visualization. However, to give our intuition a stronger foundation, we associate each of the elements of our representation with a formal concept from narrative theory.

5.2 An Algorithmic Approach to Represent and Extract Narratives

We offer a graph-driven approach to represent and find narratives from large troves of unstructured data corresponding to real-world events. We start by showing our computational representation of narratives (RQ1). Next, we present our algorithmic approach to extract these narratives automatically from data (RQ2).

5.2.1 RQ1: Narrative Representation

We offer a conceptual representation of narratives while synthesizing elements from the narrative theory literature and following the route map metaphor. Table 5.1 lays out the key elements in a narrative, their metaphoric equivalents, and their respective computational representations.

What are the key elements in a narrative?

Synthesizing across theories of narratives, we compile eight elements fundamental toward forming a narrative. Most of these elements are based on or related to the concept of events. For example, stories are sequences of events, coherence represents how much sense it makes to join events, and constituent events are a special kind of event that is fundamental to the progress of the narrative. All these elements rely on the basic concept of event for their definition, but they represent a specific narratological concept [1]. The narrative itself is computationally represented as a graph, with each narrative element corresponding to a building block of this graph.

Story, Events, and Discourse Elements According to Abbott’s general narrative theory [1], narratives are composed of two primary elements: *story* and *discourse*. The story is a sequence of *events* involving entities and follows a chronological order. Thus *events* are one of the fundamental building blocks of a *story*. Meanwhile, the narrative *discourse* is the story as narrated, that is, the rendering of the story in a particular narrative (in other words, a representation of the story). While *events* correspond to actions or happenings in the story,

Table 5.1: Mapping of narrative theory elements and computational elements alongside their route map metaphor equivalent. The concepts shown here are used interchangeably throughout the paper to describe our algorithmic approach. Narratives have at least one storyline, and a storyline has at least two events. Thus, we have the hierarchy: event < storyline $\leq$ narrative.

Elements			Description
Narratological	Computational	Metaphorical	
Events	Nodes	Landmarks	Events are actions or happenings that are part of the narrative.
Discourse Elements	Node Attributes	Landmark Attributes	Discourse Elements represent additional or complementary information about the events that are not an event of the story (e.g., the news source that published the event).
Starting Event	Source	Starting Place	The first event of a story. All stories must begin at one of these events and there must be at least one such event on the map.
Ending Event	Sink	Ending Place	The last event of a story. All storylines terminate in an end event and there must be at least one such event on the map.
Story	Chain (Path)	Route	A story is a sequence of events and discourse elements. Under the definition of narrative as a system of stories, there can be multiple stories in the same narrative.
Coherence	Edge Weights	Evaluation Metric	Coherence represents how much sense it makes to join two events and serves as an evaluation metric for the route.
Constituent Events	Maximum Likelihood Chain	Main Route	Constituent events form the core events of the narrative and are part of the main storyline.
Supplementary Events	Maximum Antichain	Representative Landmarks	Supplementary events correspond to events that are not the core part of the narrative but that provide further information and are part of alternative storylines.
Narrative	st-graph	Route Map	The narrative is a system of interrelated stories, we represent the narrative as an st-graph.

Visual Representation							
							
Events	Discourse Elements	Starting Event	Ending Event	Stories	Coherence	Constituent Events (Main Storyline)	Supplementary Events

discourse elements are part of the representation of the story. Discourse elements provide additional framing or complementary information about the events that influence the way the narrative is told but are not events themselves. Examples include public reactions to an event on social media or attributes of an event, such as the credibility of the source reporting the event or bias in reporting.

Thus, in formalizing the computational equivalent of these elements, we represent events as nodes in a graph, discourse elements as node attributes (e.g., an attribute containing the event source), and the graph itself corresponds to the narrative map. We unpack the computational representation of a narrative by introducing two other elements—starting and ending events.

Starting and Ending Events General narrative theory [1] further states that a narratological story is defined by a single starting and ending event. Some authors argue that only one event is needed to have a story [1, 132, 331], but many concur that at least two connected events are essential [33, 299, 353, 387]. We align with the popular view and require our narrative map to contain at least two events: *starting event* and *ending event*. We represent the starting event as a source in the graph—a node with no incoming edges. Likewise, we represent the ending event as a sink—a node with no outgoing edges. As per the route map metaphor, these two events correspond to the starting and ending places in a physical map, respectively. As in the physical map, the starting and ending events are selected by the user.

Story Coherence and System of Stories What makes a story coherent? A narrative story comprising of events and discourse elements needs to flow together, i.e., it needs to be coherent [1]. For example, a narrative story that focuses on a certain topic should not drastically change to a different topic without appropriate events joining them. In other words, *coherence* represents how much sense it makes to join two events. Thus, we can computationally represent story coherence through the weight of the edge joining the two events, where events are represented as nodes in a graph; higher edge weight implies higher coherence. Starting with this representation, next in RQ2, we offer computational techniques to automatically find the edge weight and determine whether to join the two events. Story coherence, thus, corresponds to the evaluation metric of our route map metaphor.

Moreover, prior theoretical work posits “narrative as a system of stories” [145]. Following this definition, our narrative map should not only contain a single story but should represent a narrative as a system of stories. These *stories* must be interrelated and form coherent sequences of events.

Constituent and Supplementary Events An important part of analyzing a narrative is distinguishing which events are the core and which ones simply provide complementary information to the narrative [1]. According to theories of narrative, *constituent events* represent the core. These are the main events that drive the story forward and cannot be omitted without fundamentally altering the story. In contrast, *supplementary events* do not drive the main story forward and if removed, the main story would still remain intact. However, this does not imply that supplementary incidents are not necessary to understand the context and general impact of the narrative. Hence, we allow both constituent and supplementary events as key elements for our narrative representation.

In computationally formalizing these two narratological elements, we want the overall sequence of constituent events to make as much sense as possible. Thus, we represent constituent events as nodes along the most coherent path (i.e., the maximum likelihood path) in the narrative map that joins the starting event with the ending event. The most coherent

path connecting the sequence of constituent events also provides us with the main storyline. In terms of the route map metaphor, the sequence of constituent events corresponds to the “main route” that has the best value for our evaluation metric (i.e., coherence) out of all the routes. For example, recall the blue route in Figure 5.1(b), which reflects the development of the outbreak from its first case as a mysterious virus to the US banning flights to Europe. Note in particular, that this route has a likelihood of $0.7 \times 1.0 \times 0.5 \times 0.3 = 0.105$, the highest out of all routes in the map.

For supplementary events, we focus on finding a representative subset of these events. To do this, we focus on all potential alternative storylines that join the starting event with the ending event. However, enumerating all potential alternative storylines is computationally impractical, except for the smallest of the narrative maps. Instead, we extract meaningful representative events from each alternative storyline in the narrative map. In terms of the route map metaphor, each alternative route can be associated with a representative landmark (i.e., a supplementary event). While these routes do not have the highest likelihood, they can still be viable alternatives and can offer different perspectives. For instance, recall the example from our route map metaphor (Figure 5.1(a)) where we elaborated alternative routes, such as a longer scenic route through the Chill Pond versus the preferred main route through the South Market.

How can we define a structured way to navigate across the narrative elements?

Now that we have formalized the basic building blocks of the narrative representation, we define the underlying computational structure that relates them. Recall that a narrative map is computationally represented as a graph. However, to ensure that our representation aligns with the theoretical structure of narratives, we must specify additional constraints to this graph structure. According to narrative theory, narratives can arrange events in many ways [132], but the underlying stories are made up of *sequences* of events that follow *chronological order* [1]. To capture these properties, we use a directed acyclic graph (DAG) because the event nodes must be joined in sequences (directed) and they must follow chronological order (acyclic). Moreover, this structure also fits the requirements of the route map metaphor, which does not contain cycles and provides a sequential order for the landmarks in a route.

However, there is one caveat of having a general DAG with no additional constraints. A DAG can have multiple sources (i.e., nodes with no incoming edges) and sinks (i.e., nodes with no outgoing edges). Yet, our route map metaphor only contains one source and one sink (the starting and ending place). Thus, we need to find a graph structure that respects these constraints. Specifically, we impose two additional restrictions on our DAG representation: the graph must have exactly one source node and one sink node. This is a special kind of graph called an *st-graph* [172]—a directed acyclic graph with a single source and a single target (sink). See for example the graph of Figure 5.1(b), where the direction of the edges is implicitly given by the chronological order of the events. This representation allows

us to encode a single system of stories connecting the ending and starting events (i.e., a single narrative map). By using an *st-graph* rather than a general directed acyclic graph, the representation allows us to understand how the events are connected without considering unnecessary information about other starting and ending events. Furthermore, we incorporate the coherence measure as edge weights in our graph. Finally, this structure allows us to navigate throughout the elements of the narrative map, offering us a chronological view of all the events and storylines contained in it.

Formalizing the narrative map definition

Given our previous discussion, we propose the following formal definition for a narrative map:

Definition 5.1. Narrative map: We formally define a narrative map as a weighted directed acyclic graph $G = (D, E)$ with a single source $s \in D$ (the starting event) and a single sink $t \in D$ (the ending event), such that weights represent the probability of going from one node to another based on coherence and the paths from s to t represent the different storylines in the narrative.

Figure 5.1(b) from the motivational example section shows an example of a weighted *st-graph* for a sample narrative, where the direction of the edges is implicitly given by the chronological order of the events. It should be noted that our representation is general enough, such that, if we modify some of the constraints, we can obtain other representations. For instance, a simple timeline (one that only allows sequential patterns, such as a single story chain [319]) or an event tree (one that allows at most one parent for each node) can be constructed from our generic representation. As a case in point, to convert our narrative map in Figure 5.1(b) to a narrative tree, we could require that nodes have at most one parent, relax the restriction of only one ending event, and then drop the less important edges. More complex representations, such as parallel timelines [395] or story forests [232], can also be obtained by combining the simpler representations.

5.2.2 RQ2: Narrative Map Extraction

Here we introduce our narrative map extraction method. We first define the narrative map problem formally before presenting the narrative extraction process. Figure 5.2 gives an overview.

Formal Problem Definition

Based on our previous formalization of a narrative map as a directed acyclic graph with a single source (the starting event) and a single sink (the ending event), we provide a formal

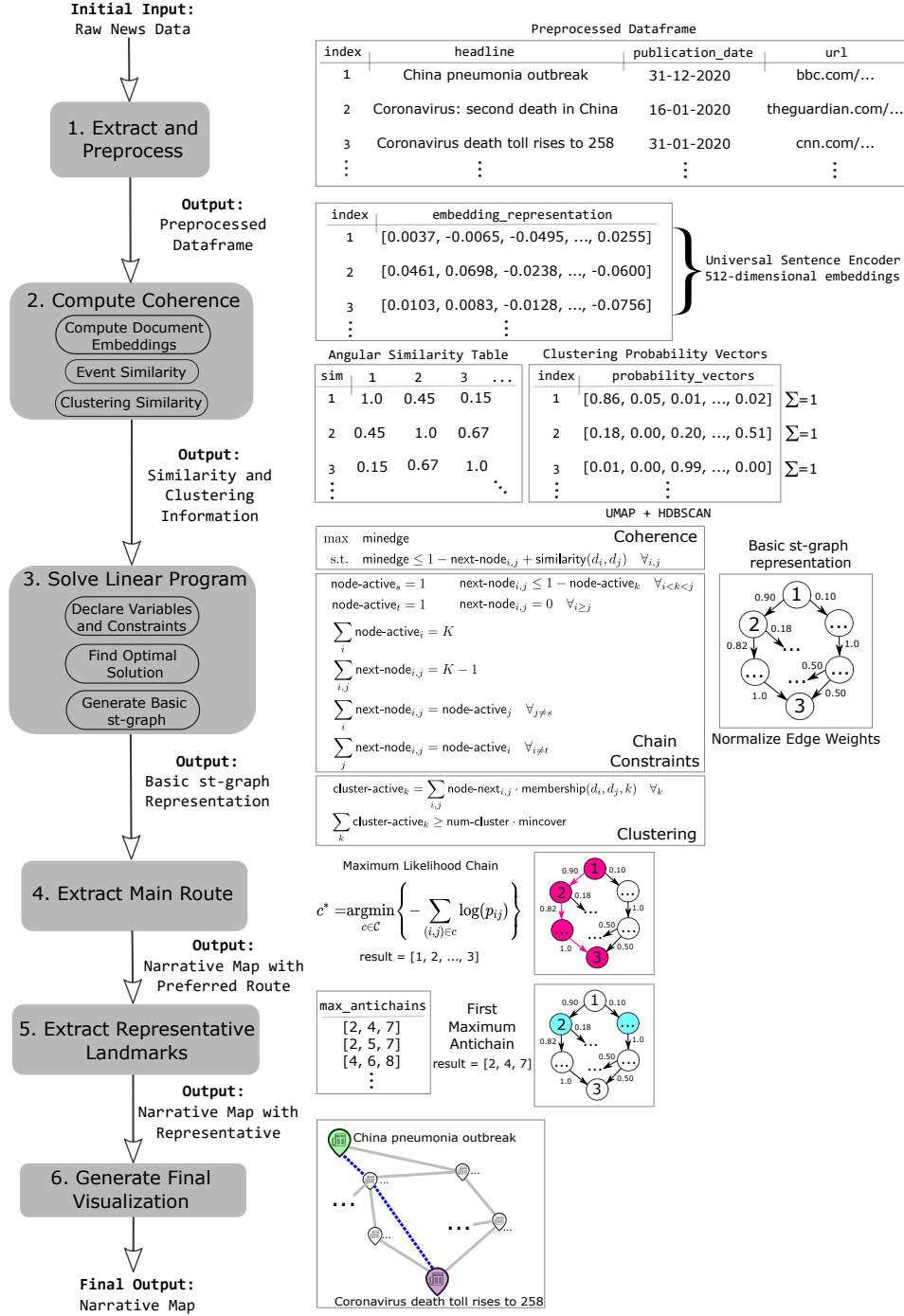


Figure 5.2: Figure illustrating the steps in our narrative extraction algorithm that generates a narrative map from data (RQ2). Each phase is annotated with their respective outputs.

definition of the narrative map problem.

Definition 5.2. Narrative map problem: Given a set of time-stamped documents D

representing events for a certain issue of interest (e.g., news articles for the Coronavirus outbreak), a starting event $s \in D$ and an ending event $t \in D$, our objective is to find a graph in the form of a narrative map $G = (D, E)$, where source s and sink t are connected (by potentially many paths).

Since the narrative map problem has multiple potential solutions, we need to define the criteria to find an appropriate solution—a good narrative map.

What makes a good narrative map?

To answer, we rely on the notion of narrative coherence [319, 321, 399]—how much sense it makes to join two events—and coverage [321, 323, 399]—how well the map covers the different topics in the narrative.

Inspired by previous works [232, 323, 369, 413], we base our measure of coherence on the following intuition: coherence should be high if the events are similar and if they share common topics. That is, in a narrative map when a user follows a chain of events along edges joining them, they should get a clear understanding of the underlying theme in the storyline. Therefore, to determine whether it makes sense to join two events with an edge on the narrative map graph, we need to compute both event similarity [413] and topic similarity [232, 323, 369].

Based on previous works [321, 322], we base our measure of coverage on the following intuition: coverage should be high if the map covers as many topics of the narrative as possible. Prior work has relied on simple word frequency-based models and pre-defined keyword lists to compute coverage and give priority to important words [321]. However, a pre-defined keyword-based approach makes assumptions about how events will look a priori, potentially increasing the risk of selection bias [259]. Thus, to avoid this issue we use a more abstract approach by defining coverage with respect to clusters of events (representing topics) that are automatically extracted from data.

In general, we want narrative maps to be as coherent as possible while also covering the topics in an appropriate way. Thus, we regard a narrative map as good if it is maximally coherent for a given coverage requirement. This leads naturally to an optimization approach, which we cover in more detail in the next subsection.

Narrative Extraction Process

Following our previous discussion on what makes a good narrative map, we seek to maximize the coherence of the narrative map, subject to coverage and structural constraints. First, we give a high-level overview of our optimization method and the extraction process. Then, we provide additional details for each phase of our algorithmic approach starting from data extraction and ending with the final visualization.

Extraction Process Overview We develop an optimization-based approach informed by prior scholarly work [319]. In particular, we employ a linear programming technique which automatically generates a narrative map by selecting a subset of events and connections between them, such that it maximizes coherence subject to two constraints: 1) coverage constraints to ensure that our solution has at least a certain level of coverage and 2) structural constraints to ensure that our solution has an st-graph structure.

Based on this optimization approach, our extraction process consists of 6 phases. The input corresponds to a data table with headlines, publication dates and URLs. The output corresponds to a visualization representing the narrative through the route map metaphor. In phase 1 we extract the data and perform basic preprocessing. In phase 2, we compute coherence information for each pair of events. Coherence values are then used as input to the linear program. In phase 3, we solve the linear program to obtain the final st-graph representation. Note that during this phase we compute coverage and ensure that our solution meets the required threshold. In phase 4, we extract the main route from the st-graph representation by finding the route with the highest coherence. In phase 5, we extract the representative landmarks from the st-graph representation using graph antichains—sets of pairwise incomparable nodes that can be used to represent the storylines. Finally, in phase 6, we generate the final narrative map and its visualization based on the route map metaphor.

Phase 1: Extracting Data and Preprocessing Our data set of choice is a collection of news articles. News articles capture real-world information and are structured to provide the key event details at the very beginning in the news headline [187]. Thus, we retain headlines of the news articles as our unit of analysis. Each headline corresponds to an event in the narrative. Next, we perform basic data preprocessing steps, such as fixing character encoding issues.

Phase 2: Computing Coherence Coherence comprises of two components—event similarity and topic similarity. First, to compute event similarity, we represent the events (corresponding news headlines in our case) using neural embeddings of the headlines. In particular, for each event, we used the pre-trained models from the Universal Sentence Encoder v4.0 from TensorFlow [64] to extract the document embeddings from their headlines. This representation is appropriate for our purpose because it has been trained on news data and is designed for short text [64]. After mapping news headlines on the document embedding space, we find semantically similar headlines by comparing the similarities between pairs of embeddings. Instead of traditional raw cosine similarity, we use angular distance-based similarity measure which distinguishes near parallel vectors better and is well suited when used with the Universal Sentence Encoder [64]. Angular similarity ranges between 0 and 1; values close to 1 denote similar events.

Next, to compute the topic similarity component, we obtain potential topics through clus-

tering, where each cluster corresponds to an underlying topic. In particular, we cluster the documents using HDBSCAN [243] with UMAP [244]. HDBSCAN is an advanced density-based technique, but it scales poorly with the dimensionality (number of features) of the data. Thus, to improve the speed and performance of HDBSCAN, we use UMAP on the document embeddings for dimensionality reduction as a preprocessing step. In addition, an important advantage of HDBSCAN is that it does not require specifying the number of clusters (or topics) a priori, allowing us to automatically surface the most important topics directly from data. The step results in cluster probability vectors, containing the probability of whether a certain document belongs to each cluster.

Next, we need to compare the topic similarity of two events using their cluster probability distributions. We compare the cluster distribution vectors using the Jensen-Shannon similarity (JS-score)—a preferred method to compare probability distributions [142]. If the events have perfectly equal topic or cluster distribution vectors, we obtain a JS similarity score of 1 (e.g., both events have $[0.6, 0.3, 0.1]$ as its cluster distribution vector). If the events focus on completely different storylines or topics, the JS similarity score will be 0 (e.g. $[0.4, 0.6, 0]$ vs. $[0.0, 0.0, 1.0]$).

How can we combine the two components—event similarity and topic similarity—to obtain the final coherence score? A simple approach would be to compute the arithmetic mean. However, this approach is not strict enough for our purposes. We want a scoring mechanism that has strong penalties if any of the components are low and forces a strict balance them. Thus, we rely on the geometric mean (square root of their product), which imposes stricter penalties when a score is low. In particular, by formulating it through a geometric mean rather than a simple arithmetic mean, we ensure that the coherence will be zero if and only if any of the components are zero.

Phase 3: Solving the Linear Program Using our coherence information from the previous step, we can now implement the linear program that maximizes coherence. The solution that we generate from this linear program is an *st-graph* that represents interrelated stories, thus fitting with the definition of narrative as a system of stories from Halverson et al. [145]. The mathematical formulation of the linear program is shown in the corresponding phase of Figure 5.2. Here we explain each component of this formulation in an intuitive way. The linear program seeks to optimize the strength of the weakest link in the narrative map in terms of coherence. Intuitively, this is based on the idea that a story is as coherent as its weakest link [319]. Furthermore, recall that narratives have a single starting event, a single ending event, and are formed by sequences of events ordered chronologically, represented as an *st-graph*. Thus, our optimization problem needs to include a series of structural constraints to ensure that our solution is an *st-graph*. Namely, it needs to set up a single source (starting event) and a single sink (ending event) and then force the events to be joined chronologically. These are all characteristics of a directed acyclic graph (DAG for short).

Finally, the last constraint—coverage constraints—ensures that a certain percent of the

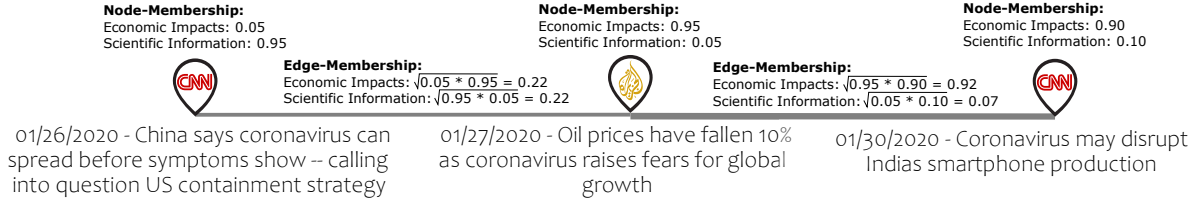


Figure 5.3: Overview of the edge membership computation used in the coverage constraint. The example shows three events from the Coronavirus narrative and two clusters. The first edge does not contribute much to the coverage of either cluster, because it is joining nodes from different clusters. In contrast, the second edge joins nodes from the same cluster, and thus it strongly contributes to the economic impacts cluster.

extracted topic clusters are covered. We found empirically that a minimum coverage of 80% worked well for our data in preliminary qualitative evaluations. Using this parametrization, we found that in general, the extracted storylines made sense and the covered topics were relevant. Furthermore, we note that coverage, unlike coherence, can not be pre-computed because it can change with each iteration of the linear program optimization process (different potential solutions of the linear program can yield different coverage values).

To compute coverage, we rely on the following intuition: a topic is covered if it appears on coherent sequences of events. Following this intuition, nodes on their own are not enough to consider a topic covered, because they might be isolated events sprinkled throughout the map. Instead, we rely on edges to determine if a topic is properly covered. Consider for example the storyline in Figure 5.3 comprised of three events. This storyline focuses mostly on the economic impact of the virus. However, the first event [*China says Coronavirus can spread before symptoms show (...)*] is mostly unrelated to economic impacts, instead, it talks about scientific information. Since this event is isolated in a storyline about a completely different topic (economic impacts) and does not connect with other events talking about its own topic (scientific information), it would not meaningfully contribute to the scientific information topic coverage.

Computationally, we define the coverage of the topic cluster k as the sum of all edge weights that belong to that cluster weighted by edge membership to cluster k (see for example Figure 5.3). The membership of an edge to a cluster is defined as the geometric mean of the membership of each endpoint. We use the geometric average because its properties work intuitively well with the concept of membership. Similar to what we did with the coherence computation, we want a scoring mechanism that has strong penalties if any of the endpoints does not belong to the cluster.

For example, if both endpoints have a non-zero probability of being in that cluster, their edge gets a positive membership. Otherwise, if any of them has a zero probability, the edge membership is also zero. Consider the storyline in Figure 5.3 based on the Coronavirus narrative, with only two clusters we can clearly see the impact of joining two nodes from

distinct clusters to the values of edge membership. Furthermore, while the first edge does not contribute much [0.22] to the coverage of either of the clusters, the second edge strongly contributes [0.92] to the economic impacts cluster. We normalize the outgoing edge weights for each node to ensure that they sum up to 1. This turns our coherence weights into probabilities, which helps to interpret them.

Phase 4: Extracting the Main Storyline Thus far, our algorithmic approach has constructed the basic narrative map structure. Now we use this structure to extract the main route or storyline (i.e., the constituent events). To do this, we find the most coherent storyline from the starting event to the ending event in the narrative map. In other words, we find the path with the highest probability from the source to the sink, or equivalently, the maximum likelihood path. In general, we can find the probability of each path by multiplying the edge weights. However, instead of maximizing the multiplication of probabilities, in practice, it is more convenient to minimize the sum of negative $\log$ -probabilities [121]. By turning this into an additive minimization problem, we can use shortest path algorithms on our graph with $\log$ -probability weights to find the maximum likelihood path.

In practice, there is usually only one such route, but it is theoretically possible to have more. If there are multiple routes, this would imply that the algorithm has produced two potential sets of constituent events to describe the narrative and is incapable of choosing the best one.

Phase 5: Finding Representative Landmarks with Antichains Now that we have found the main route, we next find the representative landmarks in the narrative map. Before delving into the method, we introduce some important concepts from graph theory that informed our technique.

To extract the representative landmarks, we leverage the concept of maximum antichains in directed acyclic graphs. Formally, antichains are sets of pairwise incomparable nodes in the DAG (i.e., nodes that are not connected by a path) [281]. For example nodes 2 and 3 form an antichain in Figure 5.4. In particular, an antichain is called *maximum* antichain if no other antichain has more nodes. For example, nodes 2, 3, and 6 form one possible maximum antichain, no other antichain that we select can have more nodes.

Previous works have shown that maximum antichains formalize the concepts of abstract graph views, slices [125], or abridgments (i.e., a small part of the graph that is of interest) [101]. Furthermore, researchers have used them to define a partitioning of the nodes of the graph [3]. In particular, according to Dilworth’s theorem [93], the size of the maximum antichain—called the width of the DAG—is equal to the minimum number of chains (paths) into which we can partition the DAG. Intuitively, we can partition the graph into chains (i.e., storylines) and interpret each node of the maximum antichain as a representative element for one of the chains. This intuitive interpretation is backed by Dilworth’s formal theorem. Figure 5.4 demonstrates this with an example. Take the chains of Figure 5.4, $A : 1 \rightarrow 2 \rightarrow$

$4 \rightarrow 7 \rightarrow 10$, $B : 1 \rightarrow 3 \rightarrow 5 \rightarrow 8 \rightarrow 9$, and $C : 1 \rightarrow 6 \rightarrow 9 \rightarrow 10$, we can associate one element from an antichain to each of these chains, such as $(A, 2)$, $(B, 3)$, $(C, 6)$.

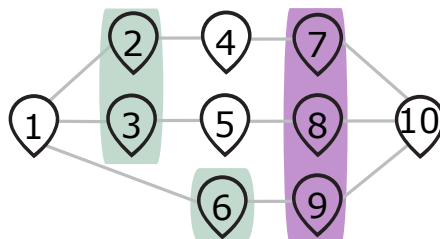


Figure 5.4: Antichain examples. The nodes highlighted in green form the first maximum antichain of the graph with 3 elements. The nodes highlighted in purple also form a maximum antichain, but it is not the earliest antichain. The nodes 2 and 3 on their own also form an antichain, but it is not a maximum antichain. Note how each element of the antichain is associated with a chain (left, center, and right)

To find the maximum antichains, we use the antichain extraction method implemented in the **NetworkX** library [143] in Python. This gives us a list with all the antichains (i.e., list of node indices). We filter the list, ensuring that it only contains maximum antichains. Note that there are potentially multiple maximum length antichains. In this case, our tiebreaker strategy is to select the earliest antichain according to their timestamps, as this provides us with representative landmarks close to the starting points of each storyline. For example, consider the two maximum antichains in Figure 5.4. In that case, we would choose the green antichain for our representative landmarks.

Phase 6: Generating the Final Visualization

Finally, we turn our computational representation into a visual representation, which is used for the case study and user evaluation. In particular, we use GraphViz [102] and the DOT language [124] to generate the layout of our visualization in an SVG file. Then, we embed the SVG file into an HTML page and add basic interactive elements using CSS.

Our visual representation follows the route map metaphor that we described in the introduction. The metaphor elements in Table 5.1 guide our design and we try to faithfully follow our original concept in the illustrations of Figure 5.1. The visualization shows the date and the headline accompanied by a landmark icon with the news source logo. We note that our visualization omits the edge weight information. We removed it because preliminary evaluations proved that adding explicit weights confused the users. Instead, we make edges wider to represent higher weights.

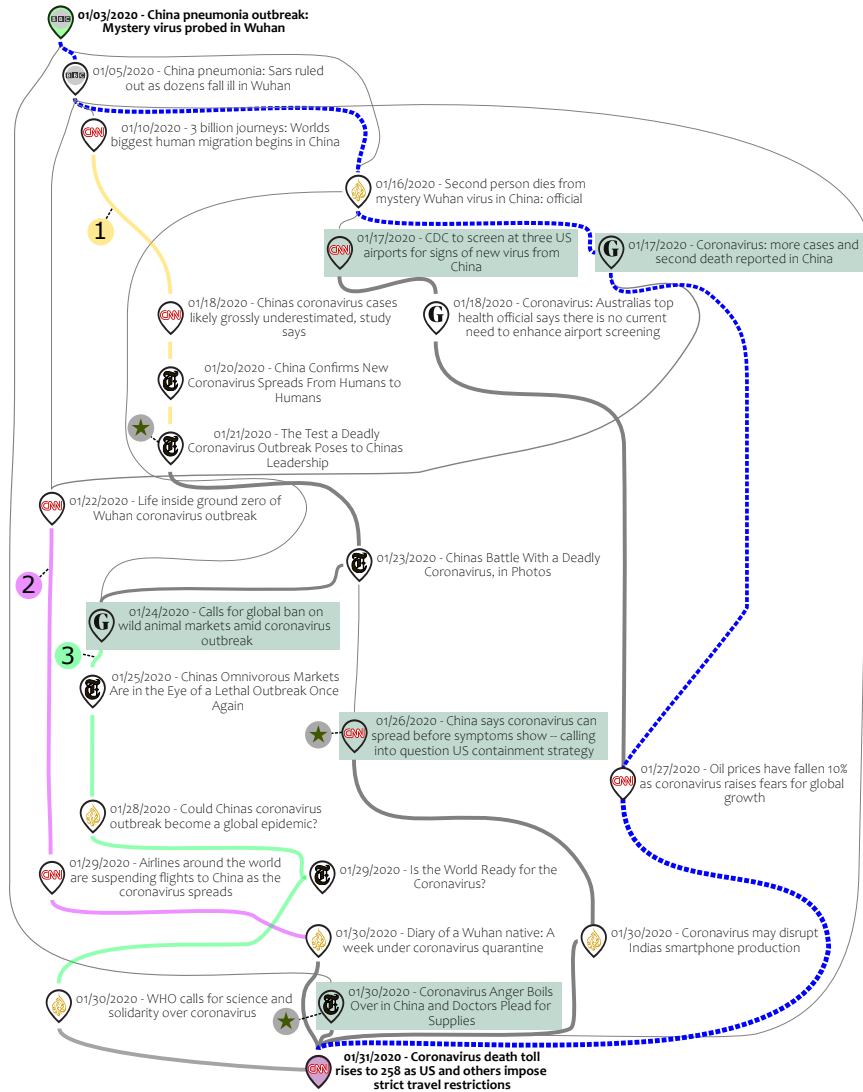


Figure 5.5: Narrative Map for the Coronavirus Outbreak during January. Headlines in bold denote the starting event  and ending event . The blue dashed line corresponds to the main storyline  (the maximum likelihood path). The width of each edge depends on its weight (coherence). Events highlighted in green are representative landmarks  (members of the first maximum antichain). The markers , , and  highlight some storylines that we reference throughout our discussion, we display them in distinctive colors. The  markers show events that share the theme of questioning or criticizing government responses.

5.3 Case Study: Coronavirus Narrative

In this section, we showcase our narrative extraction algorithm in action. We present a case study to demonstrate how we can leverage the narrative route map representation and our

computational extraction method to understand the narrative as it evolves and changes. First, we present the data used to generate the narrative map. Then, we show our results and analysis.

5.3.1 Overview of the Case Study

We focus on the Coronavirus outbreak due to its high impact and how clearly defined the major events are (e.g., first case, lockdown in Wuhan, the closing of borders, stay at home orders). Note that it is not a requirement to have such clearly defined major events to create a narrative map, but it facilitates the analysis and thus it makes for an informative example.

We obtained data related to the Coronavirus outbreak from the COVID-19 archive¹. This archive maintains a list of news articles related to the outbreak and comprised 1576 articles from 373 unique sources spanning three months—Jan’20 to Mar’20. For the purposes of this case study, we focused on the first month of the outbreak and retained articles from December 31st to January 31st, resulting in a data set of 607 news articles.

To reduce the search space of the optimization algorithm, we selected articles from the top 5 news domains in terms of publications (BBC, Al Jazeera, The Guardian, CNN, and New York Times), resulting in 102 news articles from the first month of the outbreak. Focusing on the major sources allows us to obtain an overview of the most prominent storylines and the major themes in the narrative while ignoring less important stories from more obscure sources. For the starting and ending events, we simply took the chronological first and last elements in the corpus. Out of the 102 articles, our narrative extraction algorithm returned an optimal solution with 22 events, in addition to the fixed starting and ending events. The rest were not selected because they lowered the overall coherence when added to any of the extracted storylines. Figure 5.5 shows our results.

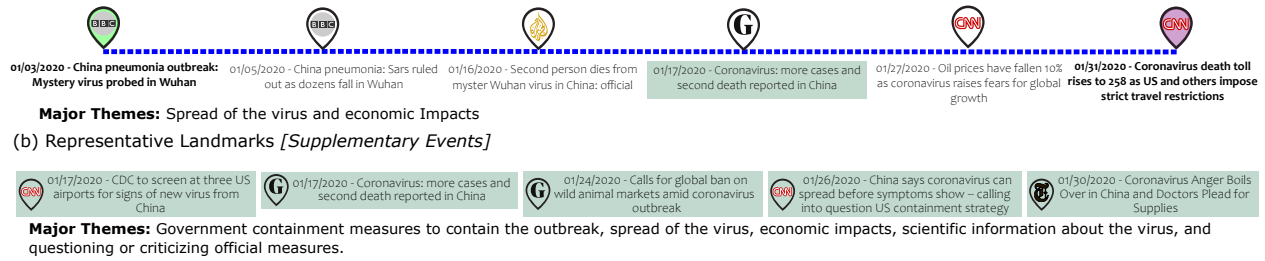
5.3.2 Analysis: Narratives underlying the Coronavirus Outbreak

What are the narratives present in the first month of the Coronavirus outbreak? Figure 5.5 shows the full narrative map generated by our algorithm. Figure 5.6 shows a summary of the major themes of the narrative according to the main route  and the representative landmarks .

We begin by analyzing the **main route** in Figure 5.6(a), which starts by referring to the new “mysterious pneumonia virus” in Wuhan. At this point in the outbreak, information about the virus is scarce, but at the very least SARS was ruled out. Then, less than a week later, there are already two deaths from this mystery virus. The next day news outlets report the second death and the virus has now been called “Coronavirus”. After this point, the story

¹<https://www.covid19-archive.com/>

(a) Main Route [Constituent Events]



(b) Representative Landmarks [Supplementary Events]

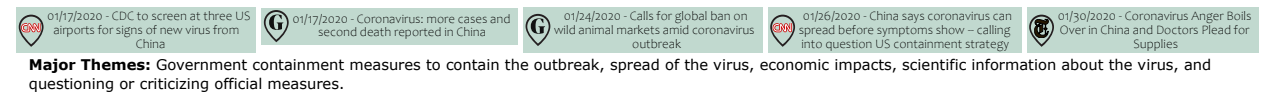


Figure 5.6: Analysis summary of the main route and representative landmarks for the January narrative map. We extract the major themes out of these elements.

changes its focus and talks about the drop in oil prices as the Coronavirus raises fears for global growth. Finally, this leads to the final event with a death toll of 258 and the US and other countries imposing restrictions. Thus, the major themes of the main route are the general spread of the virus and its economic impacts.

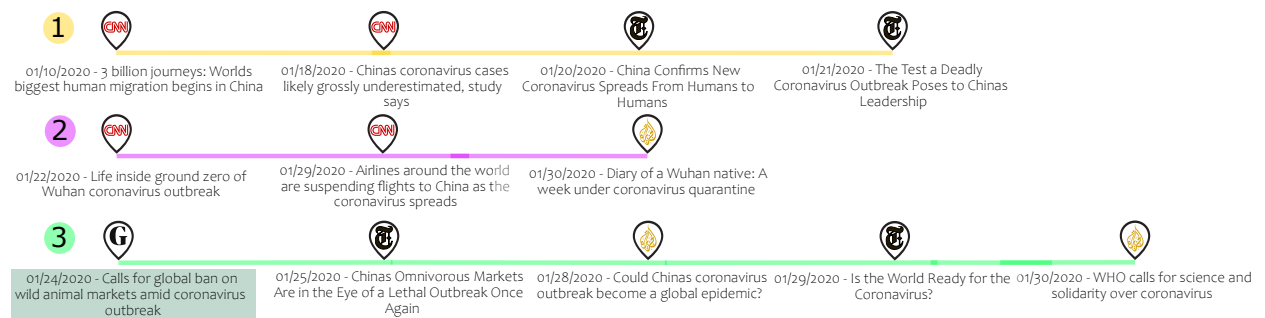


Figure 5.7: Examples of substories from the Narrative Map for the Coronavirus Outbreak during January. For brevity the start event  and ending event  are not shown in detail. Events highlighted in green are representative landmarks . We use markers **1**, **2**, and **3** to refer to the storylines as in Figure 5.5.

The **representative landmarks** in Figure 5.6(b) reveal a thematic overlap with the main route as well as surface new elements, where common topics overlapping with the main route include the “spread of the virus” and its “economic impact.” New themes in the representative landmarks refer to “scientific information about the virus” [*China says Coronavirus can spread before symptoms show (...)*], “government measures to contain the outbreak” [*CDC to screen at three US airports (...)*], and “criticisms and questioning of government responses” [*Coronavirus anger boils over in China (...)*].

Next, we turn our attention to some of the **stories** on the map. For instance, see **1** in Figures 5.5 and 5.7. Here we see references to mass migration as people try to escape from the virus. Then we find out that the number of cases might be grossly underestimated and that humans can transmit the virus to other humans. Thus, this storyline implies that the virus is likely to spread from human mass migration.

Another storyline deals with the lockdown in Wuhan and the surrounding provinces (see 2 in Figures 5.5 and 5.7). At this point, we see multiple articles about how life goes on during the quarantine. This storyline focuses on a more human aspect of the outbreak, seeing how it affects the daily lives of people. This is a theme that we did not identify directly by looking at the main route and the representative landmarks. This implies that some themes are abstracted away when we look at the narrative exclusively from the main route and representative landmarks perspective.

Amidst the battle with the virus, we see a route dealing with wild animal markets (see 3 in Figures 5.5 and 5.7), one of the potential causes of the outbreak according to the information available at that time. Then, this turns into a discussion about whether the virus could become a global pandemic and whether the world is ready to handle it. This is followed by the World Health Organization (WHO) asking for science and solidarity over the virus.

There are also events dealing with anger over how the Chinese government is handling the situation, talks about the challenges the outbreak could pose to the Chinese government, and questioning the United States containment strategy. These events are part of different storylines (see the ★ markers in Figure 5.5), yet they share a common theme in terms of questioning governments and how they are handling the outbreak.

5.4 User Evaluation

Finally, to provide additional support for the effectiveness of our narrative map representation and answer RQ3, we perform a user evaluation. Here, we present the guidelines driving our evaluation followed by the results of our user study.

5.4.1 Evaluation Guidelines

The narrative map representation should be an effective method to communicate complex narratives to individuals. Hence, to measure the effectiveness of the representation, the underlying metaphor, and the overall algorithmically generated narrative-map visual, we evaluate three dimensions:

1. **Representation:** How well does the map represent the narrative?
2. **Metaphor:** How well does the route map metaphor work for narratives?
3. **Visualization:** How well does the visualization work in general?

For each of these dimensions, we adapt evaluation questionnaires from a range of prior scholarly work. Table 5.2 shows an overview of the questionnaire. For example, to evaluate along

Table 5.2: Table containing the evaluation dimensions and their elements, as well as the questionnaire items and their source (if applicable). Some questions taken from a source have been adapted to our metaphor. The questionnaire has a total of 22 elements. Subjects were asked to evaluate each statement according to a 5-point Likert scale, ranging from “Strongly Disagree” to “Strongly Agree.”

Dimension	Element	Questionnaire Item	Source
Representation	Coherence	The map presents a coherent overview of the narrative.	Shahaf (2010)
	Relevance	The map presents relevant information about the narrative.	Shahaf (2010)
	Redundancy	The map has too much redundant information.	Shahaf (2010)
	Effectiveness	I feel more familiar with the narrative after using this map.	Shahaf (2010)
	Completeness	The amount of information on the map is appropriate to represent the narrative.	Burkhard (2005)
	Size	The size of the map is appropriate to represent the narrative.	Burkhard (2005)
Metaphor	Nature	The metaphor is natural (i.e., it is similar to the real-world).	García (2015)
	Understanding	The metaphor is understandable (i.e., easy to understand and extract information).	García (2015)
	Interpretability	It is easy to draw conclusions and get new insights from the metaphor.	García (2015)
	Landmarks	Presenting narrative events as landmarks makes sense.	Burkhard (2005)
	Main Route	The main storyline serves well as an overview of the most important events in the narrative.	New
	Representative Landmarks	The representative landmarks serve as an overview of all the stories in the narrative.	New
Visualization	Visual Impact	The visualization is eye-pleasing.	Shamim (2015)
	Overall Performance	The visualization is easy to understand.	Shamim (2015)
	Overall Design Style	The visualization is user-friendly.	Shamim (2015)
		The visualization is informative.	Shamim (2015)
	Information Quality	The visualization is intuitive.	Shamim (2015)
		The visualization is useful.	Shamim (2015)
		The comprehensiveness of the data is good.	Shamim (2015)
	Visual Representation Model	The visualization allows us to easily compare storylines.	Shamim (2015)
		The representation style of the data is good.	Shamim (2015)
	Information Presentation Model	Previous knowledge is required to understand the visualization.	Shamim (2015)

the representation dimension, we adopt Shahaf et al’s [319] evaluation scheme from their Connect-the-Dots algorithm—an algorithm that heavily informed our extraction approach. We also take elements from Burkhard and Meier’s work [56], where they present a framework for evaluating knowledge visualization tools that are based on visual metaphors. To evaluate

the effectiveness of our metaphor, we adopted elements from García et al.’s [128] work, where they focused on visualization metaphors to process complex information and interpret data. Finally, we also take elements from the visualization evaluation questionnaire of Shamim et al. [325]—a survey questionnaire designed to assess the usability of visualization systems pertaining to news articles. To measure the internal consistency of our final 22-item evaluation questionnaire, we carried out a reliability analysis. Cronbach’s alpha with $\alpha = 0.9560$ demonstrates the high internal consistency of our evaluation questionnaire.

5.4.2 Evaluation Procedure

Can users with no training in narrative or news analysis use and interpret our algorithmically generated narrative map? To answer, we asked Amazon Mechanical Turk workers (MTurkers) to do short tasks. Recent work by Cheng et al. [72] has shown the promise in employing MTurkers in the context of explaining complex AI algorithms, where they leverage the results from an MTurk evaluation with non-expert users to guide the design of their systems. Moreover, prior scholarly works have used MTurk in intelligence analysis tasks and have obtained high-quality results, demonstrating the capabilities of MTurk [223]. For example, MTurkers have performed at par with intelligence analysts in connecting entities in a visualization map [222] and in identifying relationships in large sets of data [267]. Therefore, even if MTurkers are not a replacement for expert feedback, they could still act as good proxies and provide valuable insights towards how our narrative maps proposal representation, metaphor, and visualization work in the context of sensemaking and intelligence analysis tasks.

Prior research [152, 276] indicates that it is possible to obtain good quality data from MTurk. Specifically, as long as we follow certain guidelines, the data quality is similar to traditional sources (e.g., student recruitment pools). Furthermore, research using MTurk faces many ethical challenges [136, 385], in particular with regards to payment. These challenges have given rise to various initiatives to tackle them, such as Dynamo [308]. We have followed the general Dynamo guidelines for academic requesters in MTurk. We have ensured that our MTurk workers receive appropriate compensation, above the federal minimum wage according to the estimated response time. Furthermore, we provided workers with an explanation in case they were rejected due to failing attention check questions.

The task required that MTurkers spent time to interact with the visualization, understand the structured data it surfaced, infer one of the major themes of the main route and two themes of the whole narrative map (different from the main route one), followed by responding to our 22-item evaluation questionnaire (shown in Table 5.2). Subjects were required to answer how much they agreed with the statements shown in Table 5.2 on a 5-point Likert scale with the following options: Strongly Disagree, Disagree, Undecided, Agree, and Strongly Agree. In addition to the 22 questions, we added one attention check question. Three workers responded incorrectly to the attention check question. We discarded their answers and requested additional workers to keep our total to 20 valid responses for each nar-

rative map task evaluation. The entire duration of the task along with reading instructions was approximately 30 minutes. Workers were compensated for \$4 for each task, adhering to federal minimum wage standards.

To warrant high-quality responses, we required MTurkers to be Master Turkers, have at least a 95% approval rating, and to have completed at least 50 tasks on MTurk. To ensure that the workers correctly understood the task, we first trained them by providing an explanation of the basic elements of the narrative route map metaphor. We also showed them a simple example narrative map with just three nodes, including a short discussion of the common themes that could be extracted from the example map. Prior work in crowdsourcing suggests that training workers by showing them examples results in high-quality responses [213, 260].

We performed our MTurk evaluation with three different narrative maps, each requiring 20 responses. All the maps were focused on the Coronavirus outbreak but spread across three different months to facilitate the evaluation: January, February, and March 2020.

5.4.3 Evaluation Results

Here, we present the results of our user study across each dimension (representation, metaphor, and visualization). Overall, all dimensions were evaluated well by the MTurkers as shown in Figures 5.8, 5.9, and 5.10. These results are mostly consistent throughout the three months. For brevity, we focus on a few interesting results from our analysis.

Representation

First, we analyze the results of the representation dimension. Overall, the maps are considered as a *coherent* overview of the narrative for each month. Next, we note that there is a slight drop in *relevance* during March, which might be because the March map has 40 events and is significantly larger than the other maps. Thus, there might be too many irrelevant events compared to the other maps. In contrast, we see that most users did not consider our representation *redundant*. This result is encouraging considering that one of the potential issues of using a similarity-based approach was that our representation could be redundant since it would naturally keep trying to connect highly similar items. In terms of *size*, we see that March had better results than February, despite being the larger map. This difference in results could be due to layout differences rather than the number of nodes itself. Moreover, we see that there is a trade-off between an appropriate map size and *completeness* of the data. In this context, and considering our results, our narrative extraction method balances both of these requirements adequately. Finally, we see that at least half of the users felt more familiar with the topic after using the visualization, thus showing the *effectiveness* of our representation. While we expected a more positive response to this question, it should be noted that due to how pervasive Coronavirus news was at the time we did our user study, it is possible that Turkers were already very well acquainted with the information at hand.

With a more obscure topic, the visualization could have had a greater impact on the users by providing new knowledge.

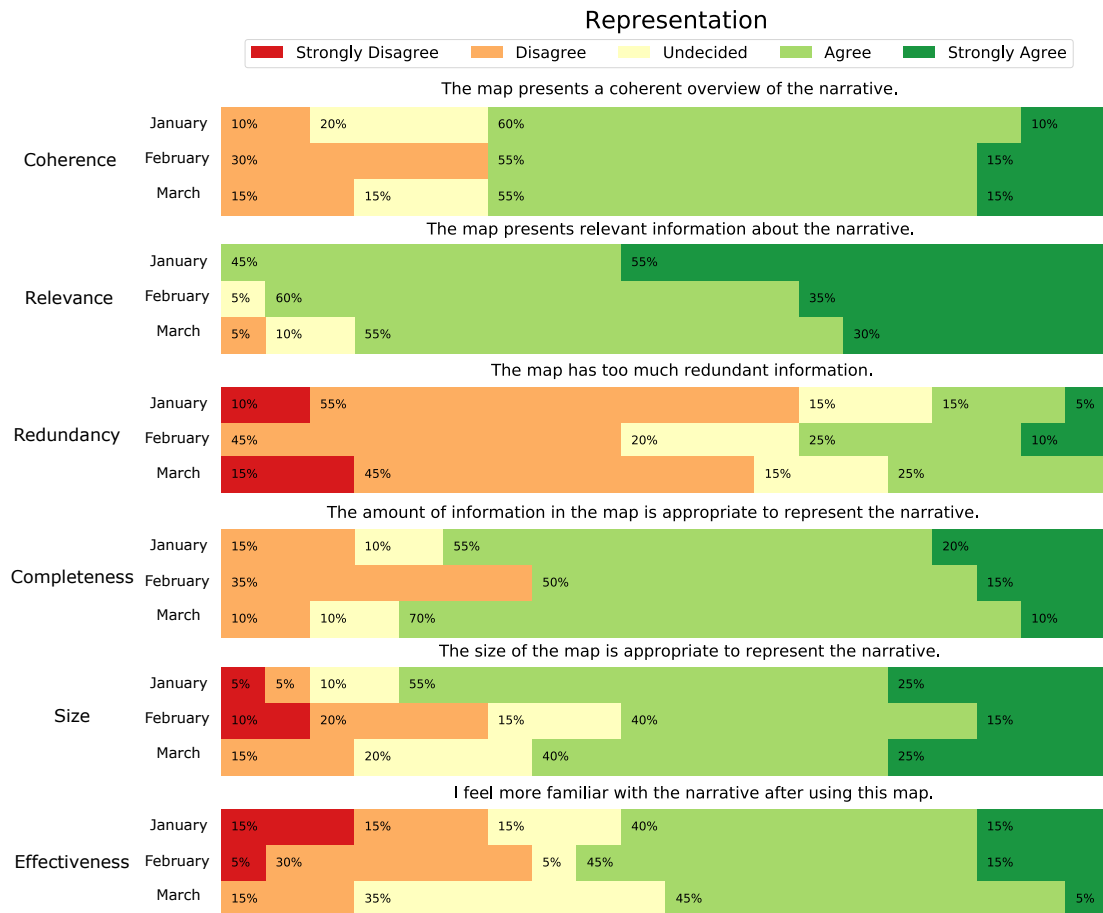


Figure 5.8: User evaluation results for the representation dimension. The bars show the percentage of users per response for each question.

Metaphor

Second, we analyze the results of the metaphor dimension. In general, the results for this dimension were positive. Users agree that our route map metaphor as well as the more specific metaphors (*landmarks*, the *main route*, and *representative landmarks*) are indeed an intuitive and *natural* way to represent a narrative. Furthermore, the users agree that it is easy to draw conclusions from the metaphor, thus it is highly *interpretable*. Moreover, most users found the metaphor to be *understandable* for January and March. However, fewer users agreed to this for the February map. We believe that this is due to the complex layout of the February map compared to the other two.

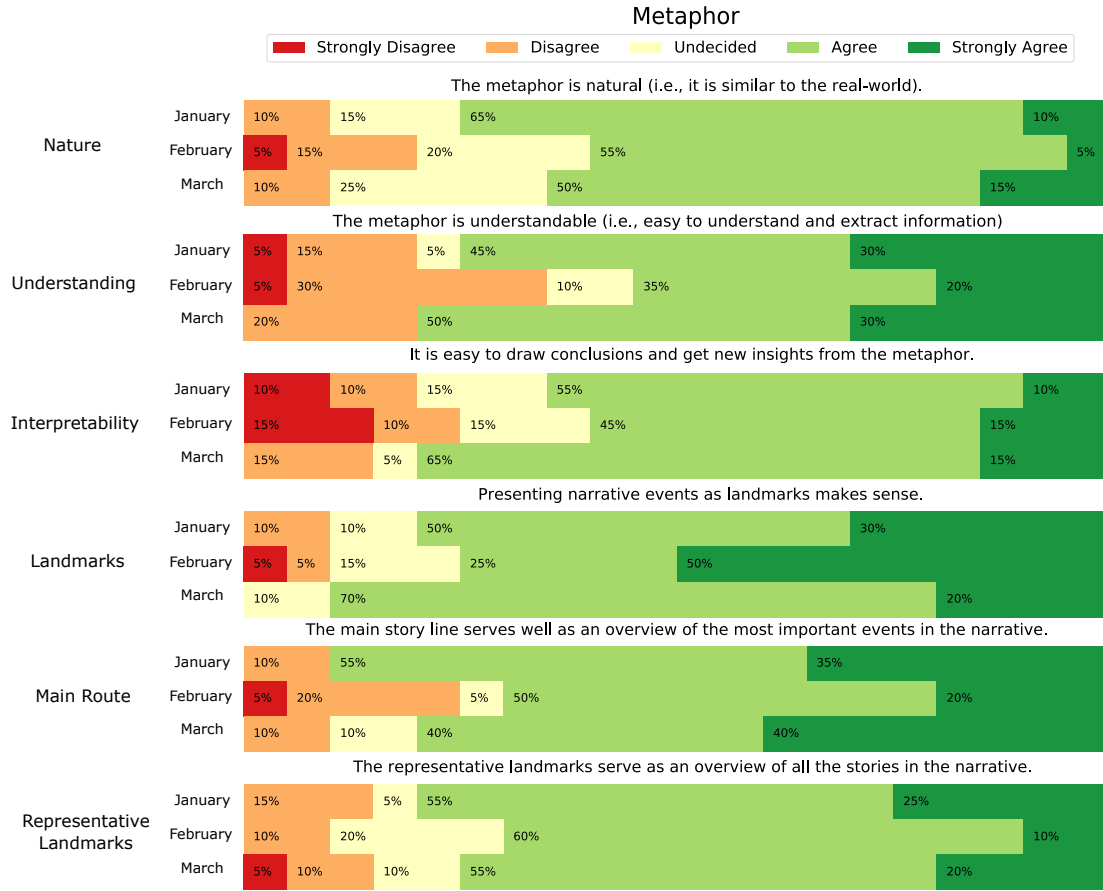


Figure 5.9: User evaluation results for the metaphor dimension. The bars show the percentage of users per response for each question.

Visualization

Finally, we turn to our analysis of the visualization dimension. Overall, the evaluation results for this dimension were positive as well. In terms of *visual impact*, most users found the visualization to be eye-pleasing, except for March. This difference could be due to the size of the map. In terms of its *overall performance*, the visualization had good results in both ease of understanding and user-friendliness. Regarding the *overall design style*, users mostly agreed that the current design is informative and intuitive. Likewise, in terms of *information quality*, the visualization had good results regarding its usefulness and the comprehensiveness of the data. The *visual representation model* was also evaluated well by the users in the context of comparing storylines. Finally, with respect to the *information presentation model*, around half of the Turkers considered that the narrative representation did not require previous knowledge. The rest of the users believed that previous knowledge would be required to understand the visualization. This split in the user evaluation indicates that additional measures might be needed to facilitate using the visualization without prior

knowledge. However, it might also reflect the fact that most people likely had previous knowledge about Coronavirus due to the constant barrage of news surrounding this topic. We leave additional evaluations, with a more obscure topic, for future work.

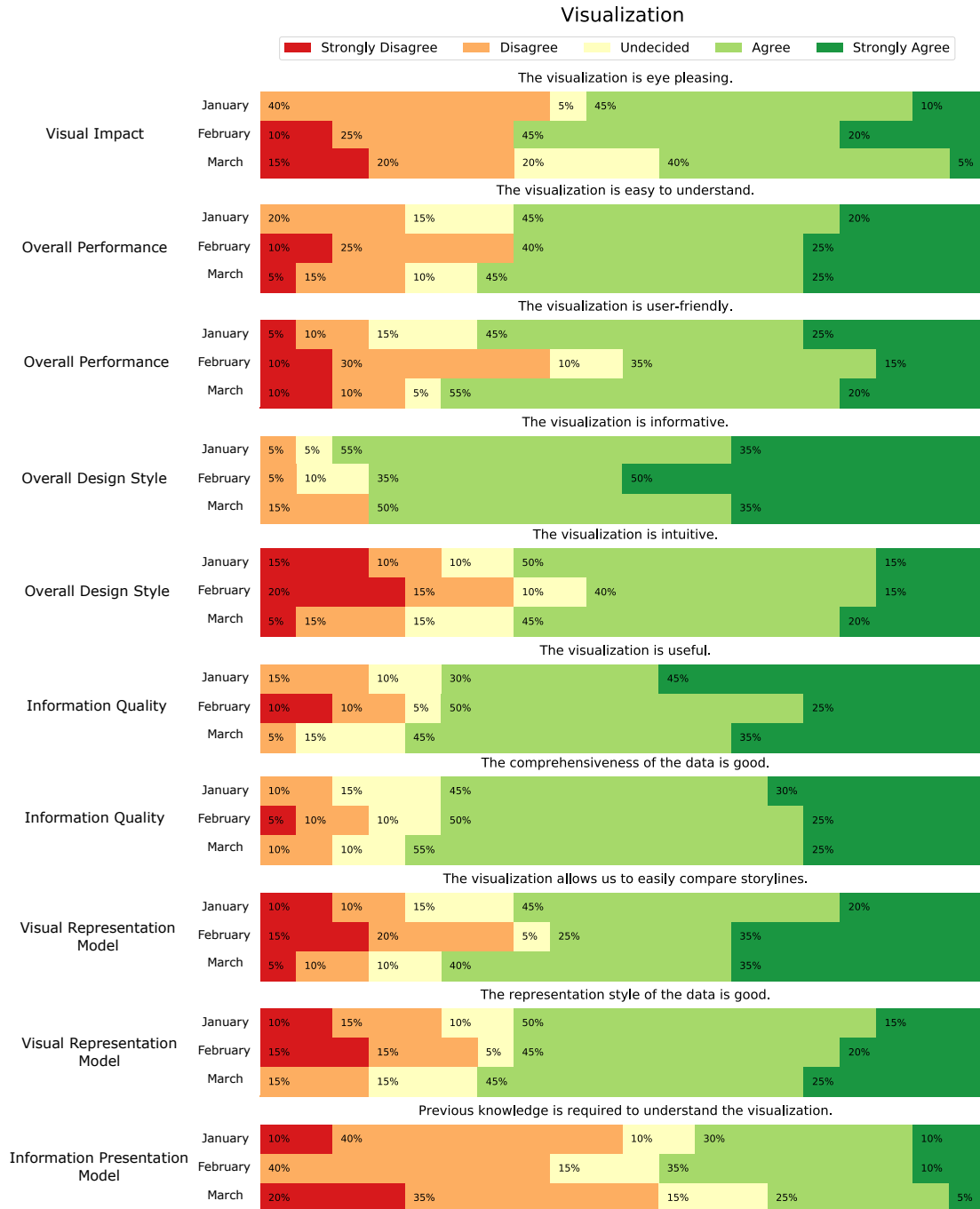


Figure 5.10: User evaluation results for the visualization dimension. The bars show the percentage of users per response for each question.

5.5 Discussion

Our work provides a computational representation of narratives grounded in formal narrative theory. We also offer an algorithmic approach to extract these representations from data using a linear programming technique. We summarize our approach and outline the key implications.

5.5.1 RQ1: Narrative Representation

Unlike previous attempts at extracting storylines or narratives from data [319, 323, 369, 399, 413], our work is based on a comprehensive survey of narrative theory. Hence our narrative representation provides a general framework to represent narratives and can serve as a starting point for understanding narratives through a computational lens.

Our representational framework emulates the real-world metaphor of a route map, making it more usable and understandable while displaying abstract unstructured knowledge (also demonstrated by our user evaluation results). Moreover, the computational representation of our framework—graph-based—can be effortlessly extended to encompass other narrative phenomena. For example, we can augment our data set to contain social media information such as comments, tweets, or posts related to our news events. By feeding this additional data to our narrative building model, we can incorporate the reactions of people towards certain articles. In turn, this provides us with insights into the issue of narrative reception [303] (i.e., how the public perceives narratives) and how it relates to the projected narrative by the media.

Additionally, we could also add new discourse elements in the form of contextual information through text summaries and data plots. For example, in the case of the Coronavirus narrative, our narrative map could be augmented by including a text summary of the outbreak during that month and traditional data plots to show the growth of the virus as the narrative advances (see Figure 5.11).

5.5.2 RQ2: Algorithmically Extracting Narratives

The optimization-driven approach for generating narrative maps uses an objective function that seeks to maximize the coherence of the narrative based on similarity and clustering information. The resulting solution of this optimization problem selects real-world events that are displayed in the narrative map along with their connections. Our user evaluation study shows that our approach constructs a primarily coherent and relevant narrative from hundreds of news articles.

Moreover, we include a coverage constraint to force the narrative map to cover many topics. This constraint could be modified and expanded to consider different kinds of coverages,

similar to the idea of user-personalized coverage from previous scholarly work [322]. For example, by adding a term to consider the political bias of the news sources, we could require a certain level of coverage for each side of the political spectrum. Thus, ensuring that all sides of the story are covered.

To help users interpret the storylines, major themes, and the connection between events, we provide them with the main route and representative landmarks of our map. These elements offer a way to characterize narratives by helping identify the storylines and major themes of the narrative. Our user study confirmed that both the main route and the representative landmarks served well in providing an overview and highlighting key elements of the narrative. While the *main route* supplies with a *longitudinal view* of the narrative, *representative landmarks* created by the maximum antichain offer a *transverse view*. In essence, the representative landmarks and the main route provide us with an understanding of both the breadth and the depth of the narrative, respectively.

5.5.3 RQ3: Evaluation

The evaluation proves that the Narrative Map representation, alongside the extraction method, is an effective mechanism to help users understand complex news narratives. The route map metaphor was considered natural by the users, providing a coherent overview of the narrative. Moreover, the main route and the representative landmarks are able to convey the key elements of the narrative.

Although the user evaluation demonstrated the effectiveness of our method, there is room for improvement. We focus in particular on the possibilities for enhancing the visualization method. As noted in the user evaluation, it is necessary to refine its visual appeal. Moreover, reducing the need for previous knowledge in the visualization is an important consideration. This could be achieved by including auxiliary discourse elements to provide additional context (e.g., news summaries and complementary data visualization).

Further improvements could be made by adding more interactivity to the visualization. For example, a simple tweak would be to highlight all edges of a node when an edge is selected. More complex upgrades could allow users to move the nodes along with its edges or allow them to hide or merge certain nodes while preserving structure. For instance, an approach similar to the recursive summarization technique for discussion forums [409] could be adopted for narratives, enabling a user to explore different narrative storylines based on their interests, with varying levels of details. Investigating these extensions could be fruitful avenues for future research. Nevertheless, in its current state, the visualization correctly represents the narrative and is able to provide users with new insights, as shown by our user evaluation study with MTurkers.

5.5.4 Implementation Parameters and Constraints

Our linear programming problem has a total number of variables in the order of $O(|D|^2 \cdot k)$, where $|D|$ is the number of documents and k is the number of clusters. In practical terms, the number of clusters is small enough to be considered constant. Thus, our main computational bottleneck lies in generating the pairwise combinations of nodes required by the optimization problem. Future work could deal with improvements to the linear programming approach or even applying a completely different optimization technique, similar to how the Connect-the-Dots method [319] evolved from its basic linear programming formulation into a divide-and-conquer approximation approach [324].

We also note that there are two parameters of interest in our linear program proposal: the expected length of the main route (K , based on the Connect-the-Dots formulation [324]), and the minimum average coverage of clusters (**mincover**). Tuning these parameters is required to obtain a good narrative map representation. Moreover, since our method depends on document embeddings and clustering, the choice of the vector embedding space representation, the clustering algorithm, and all the corresponding parameters for clustering can also influence the final representation. Thus, special care is required to choose an appropriate vector representation for the events, as this can influence both the similarity and the clustering results.

5.5.5 Limitations

One limitation of our study is that we constructed the event representation based on the headlines of the articles. Although headlines generally contain key event information, it would be useful to consider either the full text of the article or an appropriate summary. In particular, considering the headline and the lead of the article might be enough to capture the main event descriptors [187].

In terms of the user evaluation of our narrative maps, we only focused on the Coronavirus narrative with only 20 users for each map. In order to properly generalize our results, future work should focus on evaluating different narrative maps with more users and with different topics, ranging from well-known issues (e.g., Coronavirus) to more obscure ones. Nevertheless, our user evaluation provides an important baseline for future works involving this narrative representation.

5.5.6 Implications and Future Work

A General Framework to Extract and Represent Narratives

Our method is grounded in the scholarly definitions of narrative theory. We start with abstract theoretical definitions and move toward concrete representations of the essential elements in a narrative. Hence, our representation should be able to capture any kind of narrative as defined in the narrative theory literature. As long as data is in the form of (event, timestamp) tuples, we can use our algorithm to construct a map representation. Moreover, we can generalize this approach to any kind of data provided it has a temporal ordering and that we can define a measure (i.e., the evaluation metric of our metaphor) that allows us to compute coherence through similarity and clustering. At a technical level, the generalizability of our approach stems from the abstract mathematical formulation used to extract the map, which could, in turn, be used for other types of data, such as image data (e.g., creating a narrative map out of pictures from an embedding representation and timestamps).

Design Implications

Grounded in theoretical constructs of narratology, our approach can automatically extract structured information from multiple data points and represent it as a narrative map—a visual metaphor similar to physical maps. We believe that this capability offered by our approach can be used in the design of a wide range of systems. For example, imagine a news reading tool that accompanies articles with a narrative map visualization to help readers understand the big picture of the narrative, or consider a social media monitoring system that shows how the online narrative surrounding a certain issue evolves over time. Indeed, literary theorists, while outlining the importance of narratives noted [1]: “*we do not have any mental record of who we are until narrative is present as a kind of armature, giving shape to that record.*”

Furthermore, narrative maps could be used in the context of computational journalism, which focuses on providing tools for various computational needs in the journalistic context [79], such as topic detection, visualization, and sensemaking. Moreover, the theoretical basis of our representation, along with the general approach of extracting narratives from data, ensures that a designer can easily adapt our proposal to a different domain. For example, our model could be adopted by psychologists to help analyze personal narratives [199] or by narratologists to study literary works.

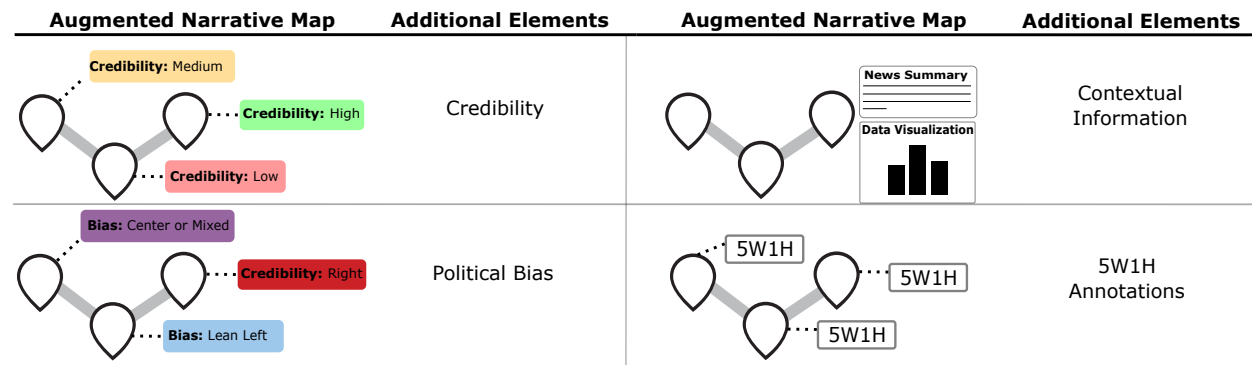


Figure 5.11: Potential extensions to narrative maps including different elements. The top-left map considers adding credibility annotations to each of the events, representing how credible or questionable the sources and their claims are. The bottom-left map considers adding political bias annotations, thus helping users understand how bias affects the storylines of the map. The top-right map considers additional information in the form of a news summary and data visualization. The bottom-right map considers adding the answers to the 5W1H questions (who, what, where, when, why, and how), which are the main descriptors of an event.

Extending the Narrative Map Representation

Using the foundations and methods presented in this chapter, future work can create an enhanced representation, for example, by considering the bias and credibility of the sources of each event. See for instance how we added this information to the landmarks of the narrative maps in Figure 5.11. With these enhanced maps, intelligence analysts would have an additional tool at their disposal to help detect potential online misinformation campaigns by analyzing emergent narratives. The rapidly shifting media ecology has made strategic narratives more difficult to control [257], thus analysts would benefit from being able to detect counter-narratives as they emerge. Additionally, fact-checkers could use this enhanced representation to obtain an overview of how facts, both credible and questionable, spread throughout the narrative. This could help in their fact assessment process. Moreover, they could augment our representation by manually modifying it to add credibility annotations.

Furthermore, narrative maps can also be used as a summarization tool to identify high-level patterns in narratives. For example, for the Coronavirus narrative shown in our case study, we identified “economic impacts” and the “spread of the virus” as major themes. In conjunction with more advanced graph-theoretical elements, it would also be possible to identify more complex patterns in the narrative. For instance, we could use network motifs—frequent patterns in the structure of a graph [254]—to encode interesting properties and relationships of its elements [37].

5.6 Chapter Conclusions

This chapter presents an algorithmic approach to represent and extract narratives using *narrative maps*. Based on a comprehensive survey of formal narrative theory, we outline a computational representation of narratives comprising eight elements. We used an intuitive route map metaphor to connect the elements, their metaphoric renditions, and their computational counterparts. Building upon existing approaches of story extraction, we designed a novel optimization-based approach to extract the underlying graph representing a narrative map. In particular, we seek to maximize the coherence of the map, subject to structural and coverage constraints. Using this graph, we then extract the main storylines and the representative landmarks of the narrative map. These help us characterize the major themes in a narrative.

We presented a detailed case study following the Coronavirus outbreak during its first month. We explored the major events and identified the major themes surrounding the outbreak. This extended example helped us demonstrate how an analyst could leverage our representation to understand the evolution of a narrative. Next, we presented a user evaluation of our method considering three aspects: the narrative representation itself, the effectiveness of the metaphor, and the usability of the visualization. While there are still some aspects that could be improved, our evaluation results demonstrate that our narrative map proposal fulfills its intended purpose.

This work opens up multiple avenues for future research, such as adapting the current approach to another type of data (e.g., images) or to a different field (e.g., helping psychologists analyze personal narratives) or extending the narrative map representation by adding contextual information. In particular, incorporating measures of source bias and credibility into our representation could pave the way towards modeling how misinformation propagates in news narratives.

Chapter 6

Design Guidelines for Narrative Maps

Narrative sensemaking is a fundamental process to understand sequential information. Narrative maps are a visual representation framework that can aid analysts in their narrative sensemaking process. Narrative maps allow analysts to understand the big picture of a narrative, uncover new relationships between events, and model the connection between storylines. We seek to understand how analysts create and use narrative maps in order to obtain design guidelines for a narrative map visualization toolkit that can aid analysts in narrative sensemaking. We perform two experiments with a data set of news articles. The insights extracted from our studies can be used to design narrative map visualizations, extraction algorithms, and visual analytics tools to support the narrative sensemaking process. The contributions of this chapter are three-fold: (1) an analysis of how analysts construct narrative maps; (2) a user evaluation of specific narrative map features; and (3) design guidelines for narrative maps. Our findings suggest ways for designing the narrative map representations and extraction algorithms, as well as providing insights towards useful interactions. We discuss these insights and design guidelines and reflect on the potential challenges involved. As key highlights, we find that narrative maps should avoid redundant connections that can be inferred by using the transitive property of event connections, reducing the overall complexity of the map. Moreover, narrative maps should use multiple types of cognitive connections between events such as topical and causal connections, as this emulates the strategies that analysts use in the narrative sensemaking process. We note that this chapter contains results from a human subjects study approved by IRB 20-794.

6.1 Introduction

Narratives are systems of stories [145]—sequences of events tied together in a coherent fashion. Events are the fundamental units of narrative action, they are either an act involving characters and entities or a happening where no entities are causally involved [1]. Narratives are fundamental to our understanding of the world and provide a natural way to capture relationships between sequences of events, as well as the goals, motivations, and plans of actors [119]. Narratives are used in the process of “connecting the dots” between apparently unrelated pieces of information [160, 161] and modeling causal relationships [75]

Storytelling in general is an accepted metaphor used in visual analytics and analytical reason-

ing [297, 316, 348]. However, unlike general visual storytelling, our work focuses specifically on visualizing textual narratives, such as those created by news. In this context, narratives provide a way to understand the information landscape, a key part of several narrative sense-making tasks [186]. Example narrative sensemaking tasks range from a journalistic analysis of news narratives [53], where the goal might be to understand the big picture, to intelligence analysis [107], where the goal is to uncover hidden or implicit relations between events.

To aid analysts with sensemaking tasks, scholars have created visual analytics software, which allow analysts to process and understand greater quantities of data and information [81]. These tools focus on different parts of the sensemaking loop [282]. For example, while some tools focus on the foraging loop [178], others focus on the synthesis loop [388] to generate hypotheses. However, there is still a lack of support towards building tools that use narrative representations to aid in narrative sensemaking tasks, such as connecting events, extracting storylines, and constructing narratives [186].

In this work, we focus on a specific type of graph-based visual narrative representation—narrative maps [186]. Narrative maps provide a generic foundation to encode different types of narratives extracted from data, requiring only the existence of a total ordering (e.g., in the form of timestamps) and text representation of the event (e.g., a news headline). Narrative maps are a useful visualization framework to understand the information landscape. In particular, they offer a way to keep track of the big picture of a narrative in the context of the ever-increasing problem of information overload [154, 319]. Moreover, they allow for uncovering connections between events in the narrative, which helps analysts connect the dots and understand events as well as their context. Furthermore, narrative maps could be used to explore how narratives and counter-narratives emerge over time, thus providing a way to model how misinformation spreads [186].

However, from a visualization standpoint, the optimal design of narrative maps for the sense-making process remains unexplored. We attempt to remedy this gap by defining a series of *design guidelines* for narrative maps. In particular, we explore how analysts create, structure, and use narrative maps to determine the characteristics of *good* narrative maps. Through our exploration, we develop design guidelines that provide the basis for the creation of a narrative map visualization toolkit; this toolkit can aid analysts in their narrative sensemaking process. Thus, the contributions of our paper are the following: (1) an analysis of how analysts construct narrative maps, including the types of cognitive connections and structures; (2) a user evaluation of specific narrative map features, namely size and transitivity; (3) a series of design guidelines for narrative maps and extraction algorithms.

Finally, the overarching goal in this work is to improve on the narrative map representation and extraction algorithm proposed by Keith and Mitra [186]. While this representation made heavy use of narrative theory in its inception, it did not include analyst feedback in the context of the narrative sensemaking process. Thus, the main findings and design guidelines proposed in this chapter provide empirical scaffolding in the context of sensemaking that can be used to improve the narrative map representation and extraction algorithm.

6.2 Narrative Maps

6.2.1 Motivating Example

To show how narrative map visualizations work, consider the narrative surrounding the Coronavirus outbreak at the start of 2020 using real data extracted from news articles. Bob, an analyst working in investigative journalism, wants to explore how the start of the outbreak led to the US travel restrictions. Moreover, he is interested in exploring other outcomes of the outbreak during this time. These two tasks are examples of narrative sensemaking. In particular, finding out how two events are connected is a *directed task*, because the analysis is focused on understanding the connection between the two events. In contrast, exploring all the outcomes of the outbreak is an *open-ended task*, as it does not focus on any particular outcome, leaving the analyst with more room to explore the branching system of stories. Thus, Bob decides to use a narrative map with a data set of articles on the Coronavirus outbreak from the top five news sources at the time. We will show how these two questions can be answered using a narrative map.

In general, narrative maps can be used to answer the directed and open-ended tasks [186]. That is, their main purpose is to aid analysts in connecting the dots between events, such as those represented by news articles or intelligence reports, and understanding the different storylines that emerge from these events. Thus, narrative maps provide a generic sensemaking framework for analysts. In particular, intelligence analysts could also use it as a graphical representation of their mental model, similar to other narrative-based models [38, 310].

However, we note that extracting narrative information manually from data and then analyzing it to generate a narrative map is a labor-intensive task. Thus, it is necessary to use an automatic extraction algorithm to deal with the data. In general, users interested in generating narrative maps from data to aid them in the “connecting the dots” task often have a specific list of requirements or constraints to impose on their extracted narratives (e.g., they are interested in specific events or outcomes). For example, users might impose requirements such as the starting and ending event of the narrative, as well as the parameters of the extraction algorithm. Further constraints could be added to the narrative map extraction method, such as making the presence of certain connections or events mandatory. For the purposes of this example, we limit ourselves to basic constraints.

In this context, Bob uses a narrative extraction algorithm and selects two points of interest based on his tasks: a starting and ending point for the narrative. In particular, he starts the narrative with the mysterious pneumonia outbreaks in Wuhan at the start of the month and ends it with the US imposing travel restrictions. The narrative map extraction algorithm selects a coherent subset of these articles to build a visual representation of the underlying narrative. We show the output visualization in Figure 6.1. The map extraction algorithm has selected a total of 25 events from the data set to display on the map.

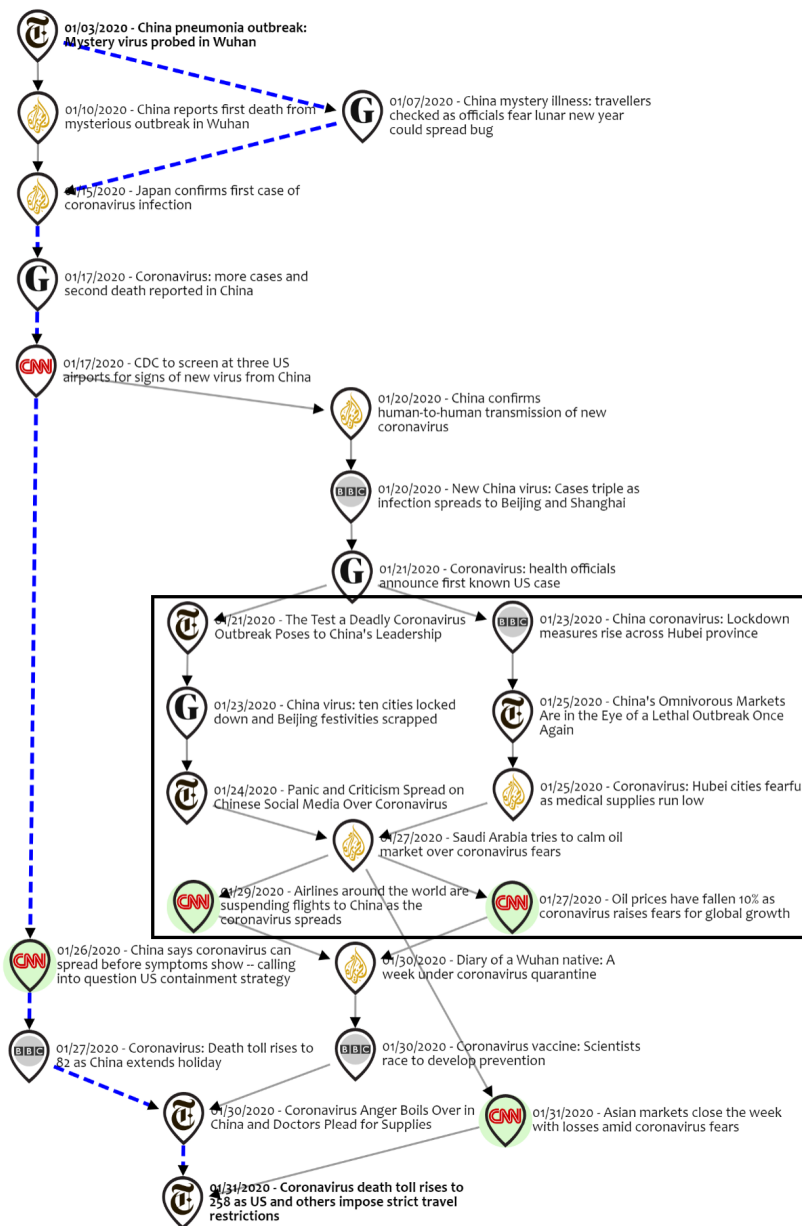


Figure 6.1: Example of a Narrative Map showing the COVID-19 narrative during January 2020 from news articles. The highlighted panel shows some of the important outcomes of the outbreak (the lockdown in China, social effects, and economic effects).

The nodes in the graph represent the events, which we call *landmarks* based on the original route map metaphor. The edges represent narrative connections between the events. After generating the map, we extract the main storyline—the most coherent path in the graph—which we represent with dashed blue edges. This storyline provides us with the most coherent

way to get from the starting event to the ending event. Next, we find the representative landmarks—a set of representative events from each storyline—which we highlight with green nodes. These events give us an overview of the side storylines of the narrative and focus on issues not covered by the main storyline.

To complete the directed task, Bob looks at the main storyline, which begins with the mysterious outbreak. Then, the main storyline starts reporting the spread of the virus. Next, it mentions how the CDC is screening airports for signs of the virus, the first sign of the US responding to the outbreak. After that, China confirms that the virus can spread asymptotically, which makes the US screening strategy insufficient. Taking into account this issue and the subsequent rise in cases, deaths, and medical supply problems in China, the US decides to impose travel restrictions. Thus, based on the main storyline of the narrative map, Bob is able to identify the core causes of imposing travel restrictions: rising cases and deaths, medical supply issues, and asymptomatic spread.

To complete the open-ended task, Bob looks at the side storylines. In particular, he focuses on the zoomed-in section of the map. This area shows some key side storylines. We see the lockdowns starting in China, as well as the socio-political effects that the outbreak had there. Next, it moves into the economic effects of the virus and how it affected oil prices. Thus, Bob is able to identify three important outcomes from the narrative map: lockdowns, social impacts, and economic impacts.

6.2.2 Research Questions

The motivating example shows how an analyst could apply a narrative map to extract important information from the data. Studying the narrative map allowed the analyst to answer the questions defined by the directed and open-ended tasks. Having shown this narrative map example, we now present our research questions. As mentioned previously, our goal is to determine the characteristics of a *good* narrative map. We do this by understanding how analysts construct narrative maps, as this gives us an insight into the structures and types of connections they would use, and we also explore how specific characteristics affect the utility of narrative maps from a consumer perspective. Thus, we sought to answer the following research questions:

- **RQ1: How do analysts manually construct narrative maps?** We focus on the strategies, cognitive connections, and structures used during the construction process of the narrative map.
- **RQ2: How do map size and transitivity affect the utility or effectiveness of the map?** We explore the effects of different combinations of map size—as defined by the length of the main story—and transitivity—whether we should allow redundant connections that can be induced by transitivity or not.

Figure 6.2 shows an overview of the experiments and research questions, which provides an overview of our experiments. We note that each of these research questions is also associated with a different type of user of narrative maps, while RQ1 is focused on users who create the maps, RQ2 is focused on users who only consume the maps without creating them. These users might have different needs, for example, map creators might want tools that make it easier to find new connections, and map consumers might prefer having additional interactivity to navigate the map. However, in both use cases, the narrative maps aid analysts in the connecting the dots task.

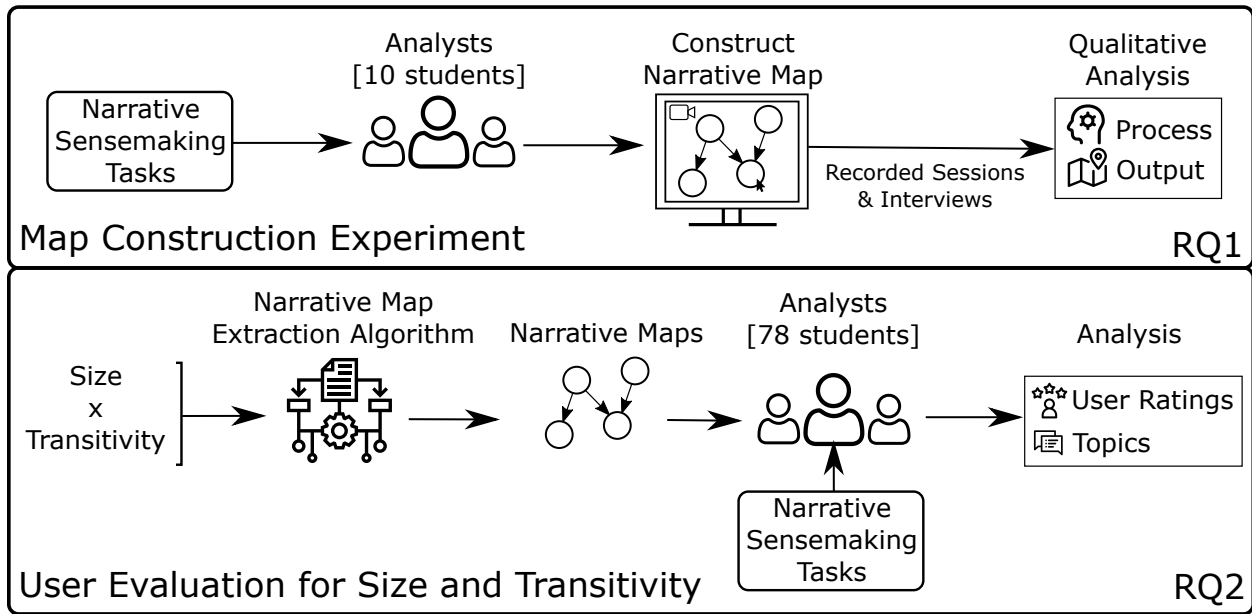


Figure 6.2: Overview of the experiments. The map construction experiment was used to answer RQ1. The user evaluation for size and transitivity was used to answer RQ2.

6.2.3 Narrative Extraction

Regardless of the underlying structure or representation used to model narratives, extraction and construction usually rely on optimizing different criteria to extract narratives from data, such as topical cohesion (whether connected events focus on the same topic) [369], coherence (how much sense it makes to join two events) [319], or coverage (the proportion of the events covered by the narrative) [321]. In this work, we use a narrative extraction algorithm based on the criteria of coherence maximization through linear programming [186]. However, none of these narrative extraction methods are backed by an evaluation of how analysts construct narratives from data. Thus, in order to create better extraction methods, we seek to understand the narrative sensemaking process of analysts.

6.3 RQ1: Narrative Map Construction Strategies

In this section, we focus on answering **RQ1**. We first describe our user study. Then, we explain the event selection criteria used by the participants. After that, we study the types of cognitive connections used to construct the map. Next, we describe the different map construction strategies. Furthermore, we describe the graph and layout properties of the narrative map. Finally, we list some of the suggestions and features of narrative maps discussed by the participants. Throughout this process, we used *open coding* [196] to discover the different types of cognitive strategies and analyze the results.

6.3.1 Study Description

Data Set

We used a data set comprised of 40 COVID-19 news articles from January 2020 that cover the start of the Coronavirus outbreak in all our experiments. This data set is a subset of the COVID-19 archive data used in previous works on narrative maps [186]. The events were carefully curated in order to have a sufficiently small data set for our manual map construction experiment while covering a series of different topics and issues regarding the COVID-19 narrative. In particular, the articles cover topics such as the economic consequences of the pandemic, the sociopolitical effects in China, the worldwide response, and others. As our data set was made up of breaking news, the main event is usually described explicitly in the headlines [187]. Thus, we focused on the headlines rather than the full article. We also included the publication dates and sources.

Task Definitions

As in our motivational example, we defined two tasks to explore how analysts constructed narrative maps, a directed task that required participants to join two events and an open-ended task that required participants to expand on the outcomes of an initial event (see examples in Figure 6.3). In both tasks, participants were given a list of events (i.e., nodes) and asked to construct a narrative map by designing its overall structure, layout, and specific connections. The participants were also asked to label their main storyline—the core events of the narrative—and their side stories—stories relevant to the overall narrative but not directly related to the main storyline. The focus of this experiment was to glean insights on the construction process, rather than comparing how the tasks themselves influence the construction. By considering two tasks rather than a single one, we expected to gather additional insights regarding the construction of narrative maps.

The directed task required participants to construct a narrative map to answer the following question: *"How did the Wuhan outbreak lead to the US travel restrictions?"*, which referred

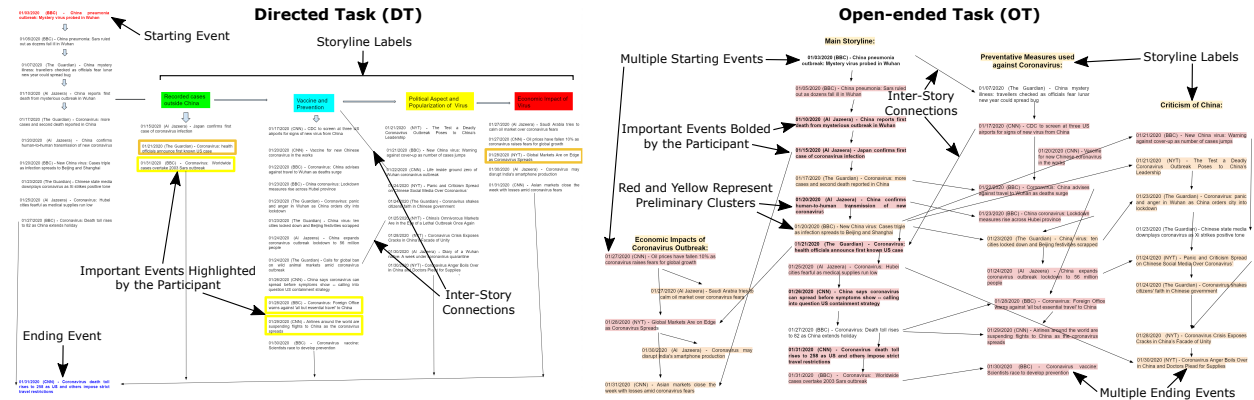


Figure 6.3: Narrative map examples created by participants for the two tasks **Directed Task (DT)** and **Open-ended Task (OT)**. Annotations highlight key elements in these maps. Note how in the DT map all storylines converge into a single ending event, emphasizing the focused nature of this task. In contrast, the OT map has a series of storylines that interact with one another, representing the different outcomes found by the participant.

to two specific events in the data set. This task is also known as “connecting the dots” and it is a fundamental task in narrative sensemaking [186]. Previous research has attempted to understand how analysts perform this process [51] and sought to automate this process through algorithmic approaches [319]. Note that while users are allowed to create side stories, the focus is on finding the connections between the two events rather than on finding other outcomes.

In contrast, the open-ended required participants to construct a narrative map to answer the following question: “*What outcomes occurred as a result of the Wuhan outbreak?*”. This task is a variation of the basic “connect the dots” task [320] that only provides the starting event as a fixed point, requiring the participants to explore the storylines that emerge because of this event. The focus is on finding storylines and outcomes in the narrative, rather than connecting two specific events. We designed this task to allow participants more degrees of freedom in their analysis, letting them define what they consider to be an important outcome. More specifically,

Both tasks required participants to label their storylines and to answer a follow-up question with their map: “*What are the key events (i.e., the most important events or turning points)?*”. All other instructions and examples were the same for both tasks. The only difference being the basic question that guides the map construction process.

Finally, we note that the tasks defined for this experiment represent simplified and constrained versions of what analysts would do in a real-world setup. In particular, they exclude the foraging loop from the sensemaking process, as we provide a pre-selected and curated data set. Moreover, they all use the same document as a starting point. These constraints were imposed in order to make analysis simpler by eliminating the additional complexity

and variables that foraging and unguided analysis could create. Thus, the created maps are easier to compare and analyze. Regardless of these constraints, the tasks still provide valuable insights into narrative sensemaking, and more specifically into the synthesis loop of the sensemaking process.

Evaluation Procedure

We recruited ten participants, five performed a directed task and five an open-ended task. All participants were advanced undergraduate students part of a national security program and hence, had a background in intelligence analysis. They also had previous knowledge on the topic which they could leverage while conducting the tasks. Prior knowledge ranged from general knowledge about COVID-19 to stronger backgrounds since some participants were ardent followers of the pandemic news right from its start. Figure 6.3 shows examples of the maps created in each task.

To provide initial training and to avoid inducing biases in subsequent task performance, participants were provided with a short example narrative map on a different topic. We engaged with our participants in an hour-long semi-structured session in a video call where they completed their assigned narrative sensemaking task. Participants were provided with a short example of a narrative map to guide them. The example map was on a different topic to avoid inducing biases in potential connections. We encouraged the participants to think aloud and ask questions and share any observations as they worked. We explained that there were no correct or incorrect answers; as our goal was to understand the cognitive strategies used by the analysts to complete the tasks. However, the quality and conceptual cohesion varied among the solutions. All participants were recorded and the videos were analyzed to understand their construction strategies. In particular, we used *open coding* [196] to perform a qualitative analysis of the created maps and the sessions themselves.

To construct the map, we gave participants a canvas on Google Drawings with the instructions and the list of articles chronologically ordered. The participants had to drag and drop the articles into the available space. Then, they had to add connections with arrows. The participants were instructed to design the map with other analysts as potential users in mind. The participants were familiar with Google Drawings and similar editing tools, thus they did not require additional training in its use, even if it might not have been their preferred tool for such an exercise. Moreover, they had full access to this tool through their institutional accounts.

We opted for Google Drawings in our study for several reasons. First, it provided a closer approximation of what a computational narrative map tool would look like compared to an approach using hand-drawn notes. Thus, even though it might influence the kinds of strategies used by the participants, these strategies should be closer to what we would expect with a computational tool. Second, given the limitations caused by the pandemic, using Google Drawings allowed us to do virtual sessions, thus minimizing the risks for the

participants. Finally, it also provided a detailed editing history which, in conjunction with the recorded sessions, was useful to precisely analyze the steps taken by the participants.

6.3.2 How do analysts select events?

We asked participants to explain their event selection process during the creation of the map. Table 6.1 shows the results for each analyst.

Table 6.1: Selection criteria for important events.

Property	Code	D1	D2	D3	D4	D5	O1	O2	O3	O4	O5
Important Events	Hard Facts	×						×	×	×	
	Perceived Impacts		×	×		×					×
	Map Structure				×	×	×				

First, regarding the selection of important events, participants either focused on “hard facts” (e.g., number of deaths and scientific information), the “perceived impacts” of an event (e.g., panic and social issues), or the map structure (e.g., number of connections or how an event summarizes the surrounding events). Four participants focused on hard facts and avoided referring to opinions or speculations in their selection of important events, as they wanted their narrative to be as objective as possible. This included reporting events such as the number of deaths, statistics, scientific information, and government responses. In contrast, the four participants who focused on “impacts” did not shy away from opinion-based or speculative headlines, since these events might provide insight into the actual perceived impacts of the outbreak. The directed-task participants that focused on impacts explicitly mentioned that they were concerned with the impact concerning the travel restrictions, as the directed task made them focus on this issue. The open-ended participant that focused on impacts used their own experiences with the virus to determine impacts. Finally, those who focused on the map structure selected the events based on their context in the underlying graph, considering whether the event acted as a hub node or whether it served as a summary of its surrounding articles or storyline.

Next, we explored how participants used the information regarding the news source of each event during the event selection process. Most participants did not use the sources, with some of them outright ignoring them. Reasons vary from “the sources are filtered and reliable enough” to simply “I was focused on the dates and headlines”. Most participants found that the sources were reliable enough and as they were relatively mainstream sources, they did not question their content. In this context, some participants commented that a narrative map should have more sources and that the sources should be balanced to prevent introducing biases in the narrative (e.g., having all sources come from one side of the political spectrum). In particular, some participants suggested limiting the sources to mainstream media.

The actual usage of the news source information varies. For example, a participant used his knowledge about the BBC to determine that one of the articles referred to a governmental office in the UK. Someone found the early Al Jazeera articles on the virus as an important sign indicating the spread and impact of the virus. Thus, we found that the news sources did not influence the selection of events or their connections, at least with this data set.

We note that in a real-world application the quality of the sources would be a very important consideration for analysts, which might affect the results of such experiments. However, in this experiment, the data was pre-selected and curated, as our work did not focus on the foraging loop of the sensemaking process. Instead, our goal was to understand the narrative structures that analysts would create, rather than how they would filter and collect the data and sources. Thus, for the purpose of this experiment, only mainstream and reputable sources were selected, in order to avoid the additional layer of complexity of dealing with biased or untrustworthy news sources.

We then turned our attention towards the events that were selected by the participants. We present the most common ones in Table 6.2 (i.e., those selected by a majority of participants in at least one task). The directed task had fewer common events than the open-ended task, which could be due to the nature of the directed task requiring to focus specifically on how the outbreak led to the US travel restrictions. However, the event regarding human-to-human transmission was considered by all participants in their narrative map.

Finally, we studied the alignment between participants in terms of included and excluded events. For each event, we measured the number of times they were included and excluded in the maps. Then, we took the maximum value among these two and averaged over all events. This gave us the average alignment among all participants. The best possible value of alignment would be 1.0, which means that either all participants agreed that it should be included or excluded. The worst possible value of alignment would be 0.5, as that would mean that the event is equally included and excluded by the participants. Following this approach, we find that the directed task has an alignment of 76.32% (excluding the pre-defined starting and ending events). In contrast, the open-ended task only has an alignment of 56.41% and much higher variability in terms of event inclusion and exclusion. This makes sense as the directed task gives a specific guiding question to the participants.

6.3.3 How do analysts connect events?

To answer this question, we asked participants to explain their connection strategies as they constructed the map as well as in the follow-up interviews. We identified seven types of connections, which we further divided into low-level, high-level, and supporting connections. Low-level connections are those that can be made directly from the content of the document (e.g., dates, keywords, entities present) without an in-depth analysis. In contrast, high-level connections involve applying cognitive schemas to synthesize information between events [51]. Supporting connections are used in conjunction with high-level connections as an auxiliary

Table 6.2: Events that were selected by a majority of the participants in at least one task. The first column shows the event, the second and third column show the frequency for that particular task, and the last column shows the global frequency. Note that only one event is common to both tasks (human-to-human transmission event).

Events	Directed Task	Open-Ended Task	Global
China pneumonia: Sars ruled out as dozens fall ill in Wuhan	0.4	0.6	0.5
China reports first death from mysterious outbreak in Wuhan	0.8	1	0.9
Japan confirms first case of coronavirus infection	0.2	0.6	0.4
Coronavirus: more cases and second death reported in China	0.8	0.8	0.8
CDC to screen at three US airports for signs of new virus from China	0.6	0.2	0.4
Vaccine for new Chinese coronavirus in the works	0.2	0.6	0.4
China confirms human-to-human transmission of new coronavirus	1	1	1
New China virus: Cases triple as infection spreads to Beijing and Shanghai	0.8	0.8	0.8
Coronavirus: health officials announce first known US case	0.8	0.6	0.7
China virus: ten cities locked down and Beijing festivities scrapped	0.2	0.6	0.4
China says coronavirus can spread before symptoms show – calling into question US containment strategy	0.6	0.6	0.6
Coronavirus: Death toll rises to 82 as China extends holiday	0.4	0.6	0.5
Coronavirus: Worldwide cases overtake 2003 Sars outbreak	0.8	0.6	0.7
Total events with at least 0.5 frequency	8	12	7

strategy to help connect events. For example, a connection could be based on cause-effect relationships between events (high-level connection) and analyst speculation (supporting connection). Table 6.3 summarizes the different connection types and Table 6.4 shows the results for the different connection types for each analyst.

Table 6.3: Connection types for each participant in our user study.

Type	Code	Description	Number of Maps
Low-level	Temporal	<i>A</i> happened before <i>B</i>	10
	Similarity	<i>A</i> is similar to <i>B</i>	7
	Entity	<i>A</i> is about the same entity as <i>B</i>	4
High-level	Topical	<i>A</i> shares a common theme or topic with <i>B</i>	10
	Causal	<i>A</i> leads to <i>B</i>	7
Supporting	Domain Knowledge	<i>A</i> is related to <i>B</i> because of external knowledge <i>X</i>	4
	Speculative	<i>A</i> is connected to <i>B</i> because of inferred <i>X</i>	3

Table 6.4: Connection types for each participant in our user study.

Type	Code	D1	D2	D3	D4	D5	O1	O2	O3	O4	O5
Low-level	Temporal	×	×	×	×	×	×	×	×	×	×
	Similarity	×	×	×	×		×		×	×	
	Entity	×	×		×		×				
High-level	Topical	×	×	×	×	×	×	×	×	×	×
	Causal	×		×	×		×		×	×	×
Supporting	Domain Knowledge					×		×		×	×
	Speculative			×					×	×	

Low-level Connections

We identified 3 low-level connections: temporal, similarity, and entity.

1. **Temporal connections** (*A* happened before *B*): Most participants used temporal connections as their default connection strategy. In particular, this type of connection was used when there were no other explicit relationships between events and participants wanted to maintain the temporal sequence of the events (e.g., “I just followed chronological order”). All participants used temporal connections in their maps, because of the inherent chronology of narratives [25]. However, some included event connections that broke the explicit chronological order. These non-chronological connections were a result of inferred causality relations—discussed later in the high-level connections.

2. **Similarity connections** (*A* is similar to *B*): Users determined two events as similar primarily based on keyword matching and a superficial similarity evaluation (e.g., “All these events mentioned markets or production”). These connections can be seen as a low-level counterpart of the topical connections discussed later. In computational terms, these connections can be easily implemented through text similarity functions or keyword matching techniques. Moreover, we note a special case of similarity connections, where documents that explicitly refer to the same event or that are too similar are grouped together as a single

document (e.g., “Calls for a global ban on wild animal markets amid coronavirus outbreak” and “China’s Omnivorous Markets Are in the Eye of a Lethal Outbreak Once Again” could refer to the same event).

3. Entity connections (A is about the same entity as B): These connections are based on named entity co-occurrences in events. For example, some participants focused on whether the events referred to specific entities; in this case, the entities mostly referred to the countries being affected (e.g., “These events talk about China”). Such connections could be extracted computationally by using named entity recognition techniques. We note that entity-based connections have been identified before as one of the common low-level techniques that analysts use to “connect the dots” between documents [51].

High-level Connections

We classified connections as high-level if they involved the use of a cognitive schema to connect information between documents. In particular, these connections arise usually from inferences made by the users rather than a superficial characteristic of the document. We identified two high-level connections in the user-generated maps.

1. Topical connections (A shares a common theme or topic with B): These connections are a more abstract version of the similarity connection. They focus on the overarching topic or theme of the articles (e.g., “These events are about the Chinese government response”). These connections are distinct from the low-level similarity ones because they are based on a semantic viewpoint rather than superficial similarity. From a computational perspective, topical connections can be implemented through a topic modeling or clustering approach. In particular, we could determine whether the events fall under the same topic or cluster. A special type of topical connection is based on how information is presented, rather than on the topic itself. For example, events that share a specific media frame [88] could be related by this connection (e.g., “These headlines criticize the Chinese government”). Based on the idea of media frames, we call this subtype of connection a framing connection. Moreover, follow-up questions suggest that this connection might be more useful for specific kinds of articles (e.g., op-eds).

2. Causal connections (A leads to B): These are high-level relationships that join events if one is caused (or could be caused) by another (e.g., “The number of cases surpassing SARS led to stricter travel restrictions”). Causal connections also cover events that could be reactions to another event, even if they are not explicitly caused by it. A special type of causal connection used by one of the participants corresponds to a “supporting argument” connection that joins events if one of them provides a supporting argument for the conclusion or occurrence of another event, even if they do not directly cause it (e.g., “The foreign office warning against travel supports the decision of airlines to suspend flights”). Finally, some causal relationships might defy temporal ordering because the reporting date of an event is

not the same as the date when the event happened. In those cases, the participants changed the order of the events to respect the cause-effect relation.

Supporting Connections

We classified connections as supporting if they are auxiliary strategies used in conjunction with a high-level connection. In particular, we identified two supporting connections in the user-generated maps.

1. **Speculative connections** (A is connected to B because of inferred X): These connections are based on implications based on the participant's beliefs (e.g., "The call for a global ban on animal markets made the global markets panic", a paraphrased causal speculative connection from one of the participants). Thus, speculative connections relate events that do not share any explicit relationship but could be connected based on speculative reasons. Note that these reasons might be right or wrong, but what matters is their speculative nature.
2. **Domain Knowledge connections** (A is related to B because of external knowledge X): These connections are a special type of connection where documents that do not share any explicit relationship are connected based on external domain knowledge (e.g., "Air travel and oil demand are related"). Note that these reasons might be right or wrong, but what matters is the dependence on external knowledge.

6.3.4 What are analysts' map construction strategies?

We studied the construction process by following the individual steps taken by the participants as they built their narrative maps. We also asked follow-up questions about the process during the interviews. We identified a series of construction strategies for each analyst, that we display in Table 6.5. We also display a diagrammatic overview of the different strategies in Figure 6.4. Note that these strategies are abstract versions of the actual strategies that were obtained after analyzing the narrative map construction process step by step. Thus, these models provide a general idea of the construction strategy followed by participants, although there might be minor differences in some steps.

1. **Clustering Strategy:** Clustering allows analysts to group documents based on specific characteristics (e.g., topic, type of document, source). Half of the participants had an explicit clustering step during the creation of the map. The use of clustering in sensemaking tasks has also been reported in previous research, either as a story construction strategy [51] or as the final product [106]. However, in the context of narrative maps, clustering is just a means to an end. That is, their main purpose is to aid in storyline construction and they are not explicitly part of the final narrative map in most cases. Clustering was either done

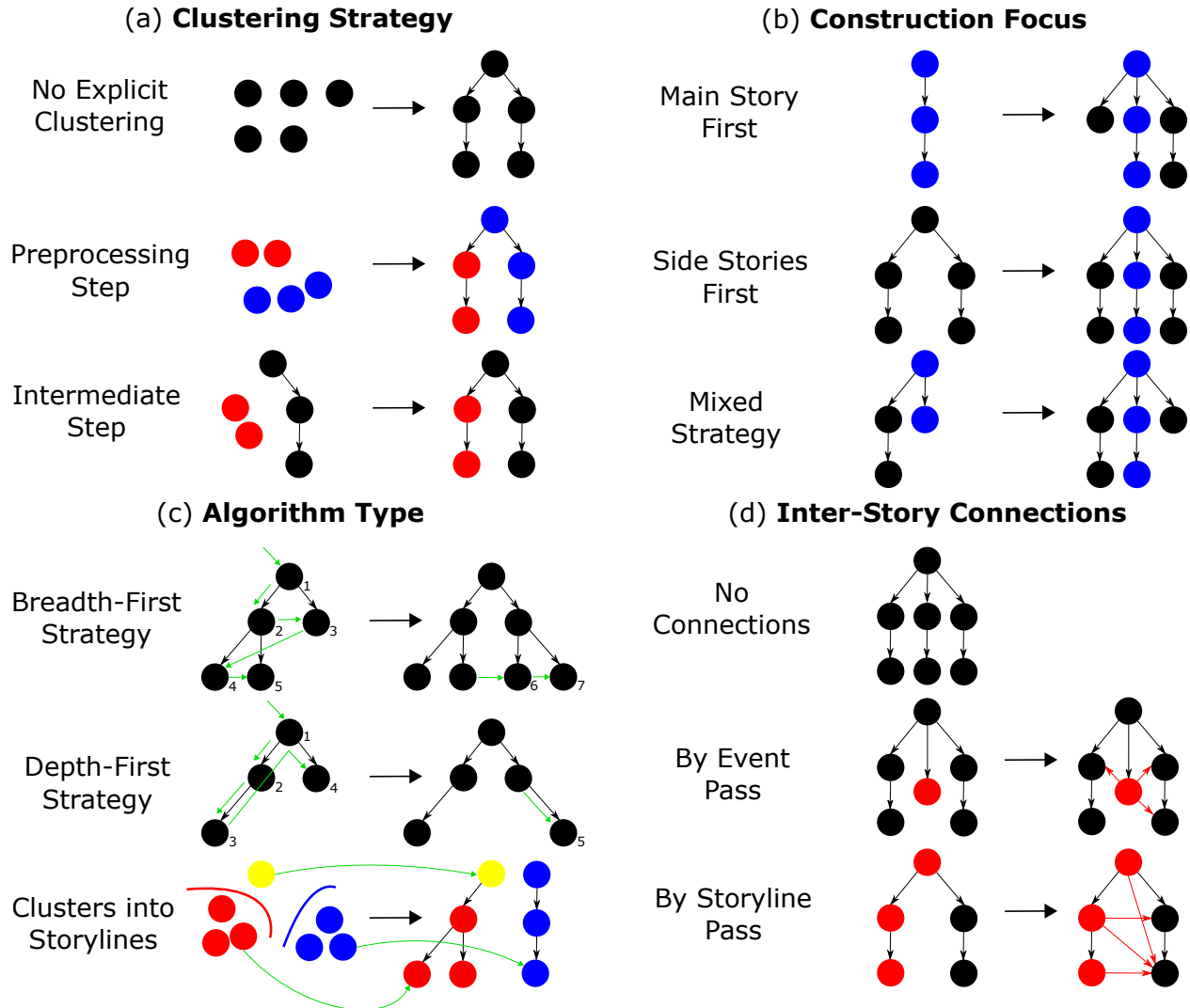


Figure 6.4: Narrative map construction strategies. (a) **Clustering strategies**: no clustering, clustering during preprocessing, or clustering in the middle of construction. (b) **Construction focus**: whether participants created the main story (in blue) first, the side stories first, or a mixed strategy. (c) **Algorithm type**: the order in which nodes were added, following either a breadth-first, depth-first, or clustering approach. (d) **Inter-story connections**: some participants checked for inter-story connections (in red) when adding events, others checked when completing a storyline.

as a preprocessing step (i.e., before starting with the connections) or as an intermediate step (i.e., after starting with the connections) to help identify storylines. Not all documents were clustered by people, although one of the participants did cluster all documents before creating the map. See Figure 6.5 for examples of both clustering strategies.

Moreover, some clusters changed over time (e.g., adding new documents to an existing cluster

Table 6.5: Construction strategies for each participant in our user study.

Property	Code	D1	D2	D3	D4	D5	O1	O2	O3	O4	O5
<i>Clustering Strategy</i>	No Explicit Clustering		×			×		×		×	×
	Preprocessing Step			×	×				×		
	Intermediate Step	×					×				
<i>Construction Focus</i>	Main Story First	×				×	×	×			
	Side Stories First			×							×
	Mixed Strategy		×		×				×	×	
<i>Algorithm Type</i>	Breadth-First Strategy		×							×	
	Depth-First Strategy	×				×	×	×			×
	Clusters into Storylines			×	×				×		
<i>Inter-story Connections</i>	No Connections			×	×	×	×	×			×
	By Event Pass		×							×	
	By Storyline Pass	×							×		

as the map was created) while others remained static (i.e., the clusters did not change after creation). Finally, there are also cases where the participants did not perform any explicit clustering step. For participants performing the directed tasks, there was a slight preference for clustering in comparison to the open-ended task participants. This could be due to the directed nature of the task, which could have allowed participants to define clusters more easily, as the guiding question could be answered by grouping events that focused on travel or the US. In contrast, the open-ended task did not provide any explicit guidelines for cluster formation.

2. Construction Focus: This strategy refers to the part of the narrative map that was created first. Participants either focused on the main story, the side stories, or followed no particular order (i.e., a mixed strategy going back and forth). The main story refers to the sequences of core events in the narrative, those that move the narrative forward [1]. In contrast, the side stories do not form part of the narrative core. Instead, they provide further information and useful context to the narrative. Note that there is an even split between focusing on the main story and following a mixed strategy. Only two participants decided to focus on the side stories first.

3. Algorithm Type: This strategy refers to the general algorithm that participants followed to construct the map. By analyzing the order in which participants constructed the maps, we found three types of strategies. The first two strategies are conceptually similar to basic graph searching algorithms—constructing the map in a depth-first or breadth-first fashion—while the third strategy is based on clustering—turning clusters into storylines. Note that depth-first approaches either focused on side stories first or on the main story first. In contrast, breadth-first approaches followed a mixed strategy by definition. The strategy of turning clusters into storylines either focused on side stories first or followed a mixed strategy. We note that it would be technically possible to focus on the main story when using the

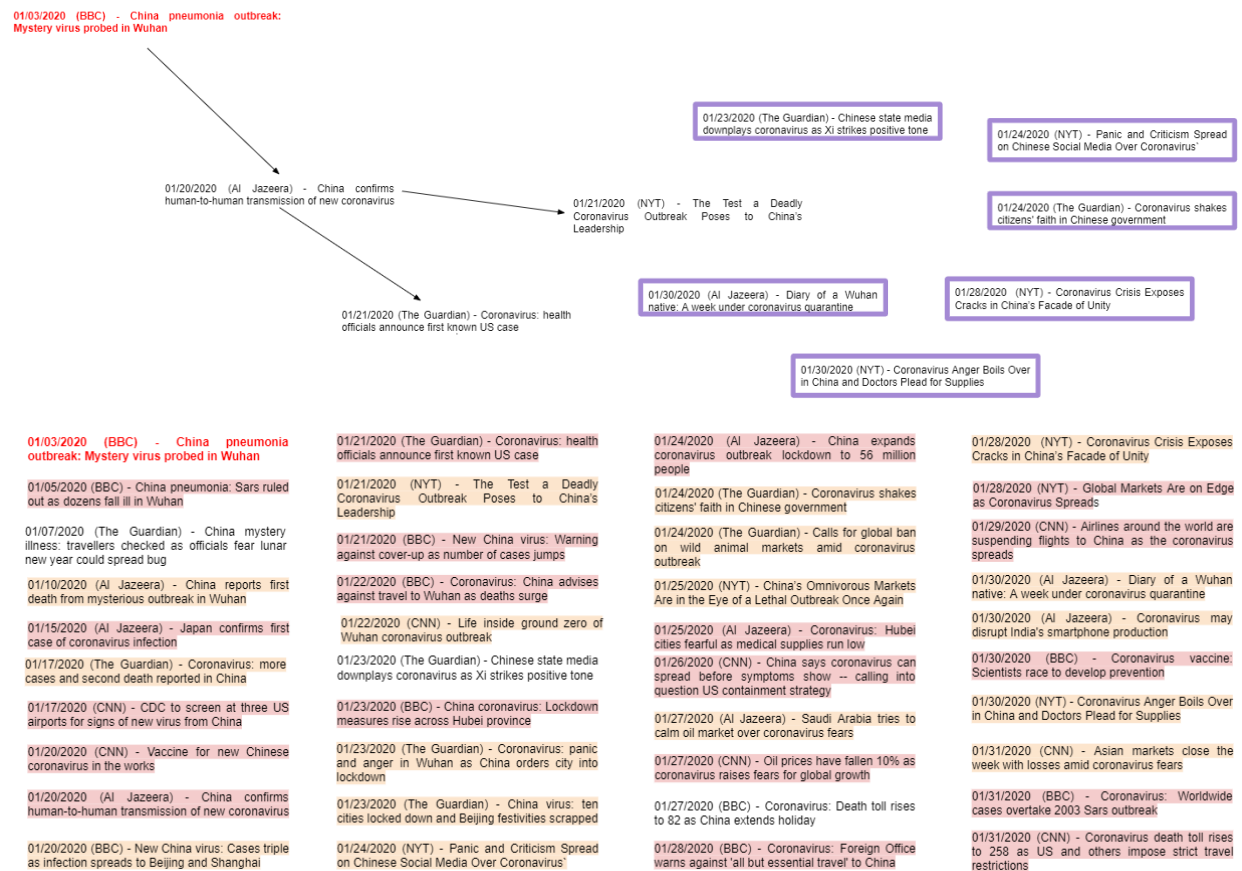


Figure 6.5: Examples of different clustering strategies. The top example shows the creation of a cluster through an **intermediate clustering step** during map construction (the highlighted events are about "Chinese Media Criticism"). The bottom example shows **clustering as preprocessing**, where cluster labels are assigned before constructing the map (the highlighted events were initially classified with respect to the questions asked in the instructions).

clusters into storylines approach. However, we did not observe this behavior during our experiments.

4. Inter-story Connections: This strategy refers to how the participants connected storylines. In most cases, participants did not add inter-story connections; making their storylines independent from the rest of the graph, except for the initial connection where they split off. In other cases, they added connections on a by-event basis, checking whether an event should be connected to other stories as they add it. Alternatively, they added connections on a by-storyline basis, checking whether to connect the storyline with others only after completing the whole storyline. For example, the inter-story connections in the open-ended map of Figure 6.3 were added after the storylines were completed. We note that, in gen-

eral, there were few inter-story connections, as each storyline was clearly defined and mostly self-contained.

6.3.5 What are the properties of the created maps?

We answer by focusing on multiple structural aspects of the underlying graph and the layout considerations made by participants (see Table 6.6).

1. **Graph Structure:** We found that participants used three types of underlying graph structures: lists , trees , and directed acyclic graphs (DAGs) . These results are in line with prior work on story and narrative representations, which has focused on similar types of structures to represent stories [186, 321], such as timelines [319], trees [18], or other graph variants [401]. For our study, structures were evenly split between trees and DAGs, with only two list-like graphs, where one of them was a single timeline and the other comprised three parallel timelines . The person who used the single list structure explained that they were trying to create a timeline that covered the important events, rather than expanding on side stories.

2. **Layout and Main Story Position:** Most participants went for a vertical (top-down) approach  with storylines presented as parallel columns and the main story placed first  (i.e., the left-most story in a vertical layout or the top story in a horizontal layout). Horizontal layouts  were not preferred; as noted by participants, computer displays seem to favor vertical layouts due to how scrolling works. Finally, one participant used a unique diagonal layout  (shown in Figure 6.6; we did not observe this behavior in any of the other participants).

3. **Number of Source and Sink Nodes:** Most people had multiple storyline endings (i.e., sink nodes) . In particular, all five open-ended task maps had multiple endings. In contrast, the directed task had two participants constrain themselves to a single ending as defined by the task , while the others added endings or dead-end events for some of the other storylines. For source nodes, participants that had the directed task were more likely to have a single source  than those that had the open-ended task . The tendency of open-ended maps to have multiple sources and sinks intuitively makes sense given the unrestricted nature of the task. In contrast, the directed task maps are naturally more focused on just answering the main question (“How did the Wuhan outbreak lead to the US travel restrictions?”), thus leading to structures that did not have as many loose ends.

4. **Connectivity:** Most participants created connected graphs . In graph-theoretical terms, we classify the map as connected if its underlying graph is weakly connected (i.e., we disregard the direction of the arrows). However, there were two cases where the graphs had separate components . The first had a separate component for the “social response and effects of COVID” that was not connected to any other story. The second had three parallel timeline structures (the main story, economic effects, and preventive measures) without any explicit connection between them.

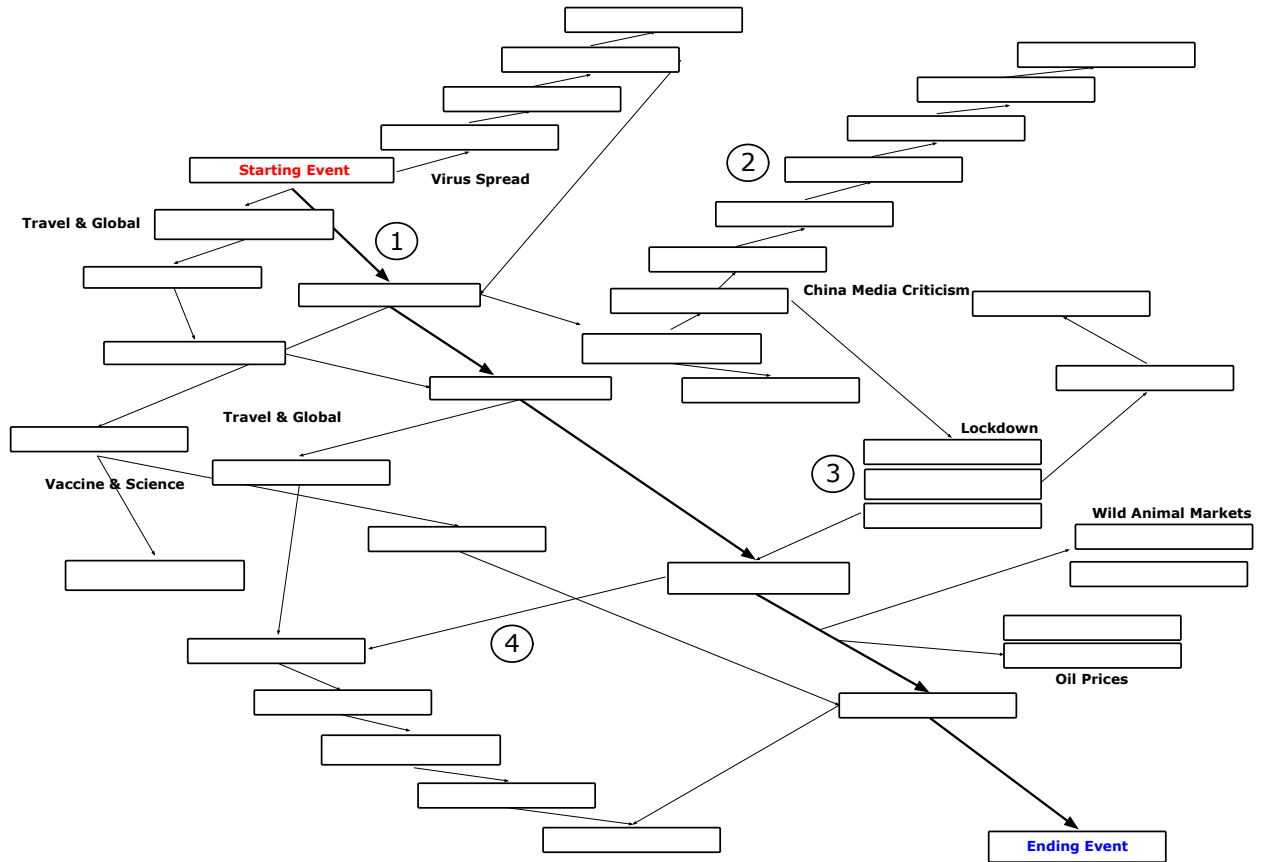


Figure 6.6: Example narrative map structure from a participant of the Directed Task. Note that the map has a diagonal layout—the only map that uses this type of layout—with its main story (1) on its center. Moreover, this map was constructed following a depth-first strategy, starting with the main story and then branching into the side stories (2). Some events that were considered too similar or the same were grouped together into a single block (3). Inter-story connections (4) were added following a by storyline pass strategy.

5. **Transitivity:** We considered whether participants explicitly included connections that are implied by transitivity (i.e., $\rightarrow$ vs. $\rightsquigarrow$). We observed that most people do not include these transitive connections. In particular, only two participants who worked on the open-ended task used explicit transitive connections to emphasize the relationship between events. However, even in the maps where they were used, they were scarce. Thus, in general, transitive connections were either not needed or participants had difficulty finding such connections in the first place. In contrast, the computer-generated maps from the original narrative map extraction algorithm [186] were able to easily extract explicit transitive connections and were well-evaluated by users. Therefore, we considered exploring whether including such explicit connections is useful. If so, using algorithms that can extract explicit transitive connections to emphasize specific relationships in narrative maps could help analysts in their narrative sensemaking process.

6. **Size:** The size of the map is inherently related to the length of the main story since longer main stories lead to bigger maps in general. Thus, the length of the main story acts as a proxy for map size. The number of events in the main story ranged from 6 to 25 events out of a total of 40 events in our data set. The mean number of events in the main story is 14.3 events and the median is 12.5 events.

Table 6.6: Graph and Layout Properties for each participant in our study.

Property	Code	D1	D2	D3	D4	D5	O1	O2	O3	O4	O5
<i>Graph Structure</i>	List 					×					×
	Tree 			×	×		×	×			
	DAG 	×	×						×	×	
<i>Layout</i>	Vertical (top-down) 		×	×	×		×	×	×	×	
	Diagonal (left to right) 	×									
	Horizontal (left to right) 					×					×
<i>Main Story Position</i>	Main Story First 		×	×	×	×	×			×	×
	Main Story Center 	×						×	×		
<i>Source Nodes</i>	Single 	×	×		×	×		×		×	
	Multiple 			×			×		×		×
<i>Sink Nodes</i>	Single 		×			×					
	Multiple 	×		×	×		×	×	×	×	×
<i>Connectivity</i>	Connected 	×	×		×	×	×	×	×	×	
	Disconnected 			×							×
<i>Transitivity</i>	Implicit 	×	×	×	×	×	×	×			×
	Explicit 								×	×	

6.3.6 Suggestions and Additional Features

From the follow-up interviews, we also gathered a series of recommendations and suggestions for additional narrative map features. These suggestions were mostly oriented towards providing further support to the construction process and the subsequent use of the map. Setting aside the addition of basic functionalities, such as searching, highlighting, color-coding, or modifying the graph, as well as including more data, we summarize some of the key takeaways. First, participants mentioned the necessity of explanations in event connection (i.e., why are they connected?) and important events (i.e., why are they important?). Participants did not include any edge labels in their constructed maps, but they explained that they would prefer if maps created by other analysts included edge labels with explanations. Next, the participants mentioned the idea of getting automated recommendations on how to complete the map or expand it during the construction process, as this would make the construction process easier. Furthermore, maps should provide directions regarding the

general topics or storylines in a specific part of the map (e.g., similar to section titles) and a way to focus on specific topics by zooming in with more details. Finally, events should be able to be merged if they are the same or above a certain similarity threshold, in order to reduce redundancy in the map.

6.4 RQ2: Effects of Size and Transitivity

Based on our previous findings, we sought to explore the effects of size and transitivity on narrative maps. In particular, in RQ1 we found that the length of the main story in the analyst-generated maps had high variability, ranging from only 6 events to 25 events. Thus, we explored the effect of size on the utility of narrative maps. Moreover, in RQ1 we also found that most participants did not include explicit transitive connections. However, previous research has found that narrative maps that included these transitive connections were successful in terms of user evaluations [186]. Thus, we sought to compare maps with and without explicit transitive connections.

6.4.1 Study Description

To explore these characteristics, we performed a new experiment evaluating multiple combinations of sizes and the use of transitive connections. We opted to generate the maps computationally because this allows for easier scalability compared to manually generating maps for all the factor combinations in the experiment. Specifically, as our goal is to improve on the narrative map representations created by Keith and Mitra [186], we generated a series of maps using their narrative maps extraction algorithm. For the events, we used the same data set from RQ1.

Map Size

We extracted maps of different **sizes** based on the K parameter of the extraction algorithm, which represents the expected length of the main story. We tested several levels: *Small* ($K = 4$), *Medium* ($K = 8$), *Long* ($K = 12$), and *Longer* ($K = 16$). A high value of K leads to long and narrow maps, while a low value of K results in shorter and wider maps. We generated all maps with a required minimum coverage of at least 50% of the clusters found in the data (i.e., at least half of the relevant topics in the data should be covered by the map).

Transitive Connections

To study the effect of explicit transitive connections, we created maps with all their connections (normal maps) and maps with all explicit transitive connections removed (transitive reduced maps). To remove the extra connections from one of the base narrative maps we used *transitive reduction*, an operation that removes edges on directed graphs while preserving its structure and important properties [99]. This operation is a way to reduce the complexity of large and dense graphs, which makes their layouts easier to read [123]. Thus, we would expect it to have a similar effect on narrative maps. We labeled maps using their *Size* followed by a dash and *N* for regular maps or *T* for transitive reduced maps (e.g., Short-N).

Evaluation Procedure

For evaluation purposes, we provided participants with a single map and asked them to complete narrative sensemaking tasks. We show a zoomed out overview of all the maps used in this evaluation in Figure 6.7.

In particular, we used the directed and open-ended tasks from our first experiment. The directed question could be answered by finding the main storyline in the extracted maps, while the second one could be answered by exploring the side storylines. Thus, these two tasks ensured that the participants had to make full use of the narrative map. We also included an evaluation questionnaire with ten 5-point Likert-scale questions. Then, we considered the percentage of favorable answers to evaluate the effectiveness of the narrative maps. We adapted the evaluation questionnaire used by Keith and Mitra [186]. This questionnaire considered multiple dimensions for the evaluation of narrative maps and adapted elements from similar procedures to evaluate the representation [56, 319], the metaphor [56, 128], and the visualization [325]. We used a simplified version due to the stricter time constraints in this experiment. Nevertheless, this version covers all the main points of the original questionnaire (evaluating the underlying representation, the visualization itself, and the map metaphor). The relevant questions are listed below:

- **Usefulness:** The map was helpful to answer the questions.
- **Coherence:** The map presents a coherent overview of the narrative.
- **Relevance:** The map presents relevant information about the narrative.
- **Redundancy:** The map has redundant information. [Note the result for this question is reversed, as redundancy is a negative characteristic].
- **User-friendliness:** The map is easy to understand.
- **Comparability:** The map allows us to easily compare storylines.

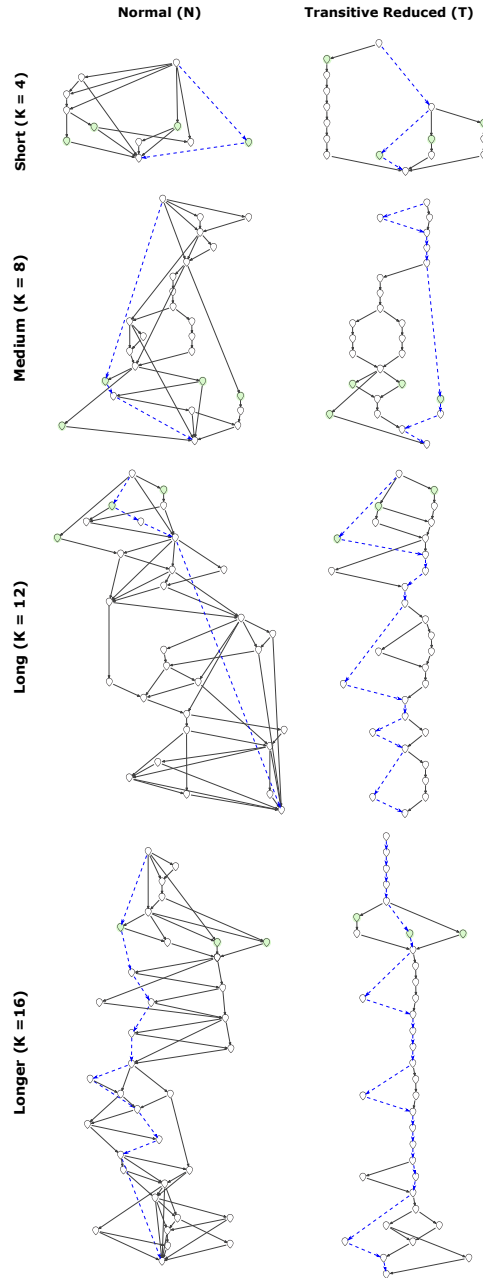


Figure 6.7: Overview of all the maps used in the evaluation procedure of RQ2. Normal maps (N) have more connections, allowing them to show more details at the cost of more complex layouts compared to their reduced counterparts (T).

- **Completeness:** The amount of information on the map is appropriate to represent the narrative.
- **Size:** The size of the map is appropriate to represent the narrative.

- **Landmarks Metaphor:** The representative landmarks (green events) serve as an overview of all the stories in the narrative.
- **Main Route Metaphor:** The main storyline (blue path) serves well as an overview of the most important events in the narrative.

Study Participants

Our design considered 91 potential subjects, which we randomly distributed among the factor combinations, ensuring that every combination had at least 11 subjects. The original sample consisted of 68 males and 29 females. The students were undergraduate students in a data analytics program. The participants had a lower level of experience compared to the participants of our first experiment, as their knowledge base consisted mostly of basic data analytics techniques. Nevertheless, most participants were able to complete the tasks. After filtering through blank and invalid responses, we had a total of 78 responses. Table 6.7 shows the number of valid responses for each factor combination and the average effectiveness results.

6.4.2 User Performance

How well do users perform narrative sensemaking tasks with these narrative maps? To evaluate user performance, we identified a series of important *high-level events* in the main story and the side stories. These high-level events are abstract representations of relevant events throughout the narrative. These high-level events were identified based on the narrative maps created for RQ1 as well as the follow-up interviews with participants. We evaluate user performance based on *recall* (fraction of the high-level events that are successfully retrieved).

In particular, the following high-level events that contributed to the US travel restrictions (i.e., the main story): the geographic spread of the virus, the reports on the virus's contagiousness, the death toll, and the worldwide responses. Moreover, we have the following notable high-level events for the side stories: the lockdown in China, the economic impacts, and the social impacts. We present the percentages of users that correctly identified these high-level events in the main story and the side stories are shown in Figure 6.8.

Medium-T (clean version shown in Figure 6.1) had the highest recall of high-level events in both the main story and the side stories. Performing an ANOVA we find that the difference in main storylines is significant with respect to both map size and use of transitive connections ($p < 0.05$). In particular, Medium-T has the best performance in terms of recall. For the side storylines, the difference was not significant. The performance difference between the main story and the side stories shows that attempting to construct a single narrative map that covers both tasks is sub-optimal. Thus, using a series of task-specific maps rather than

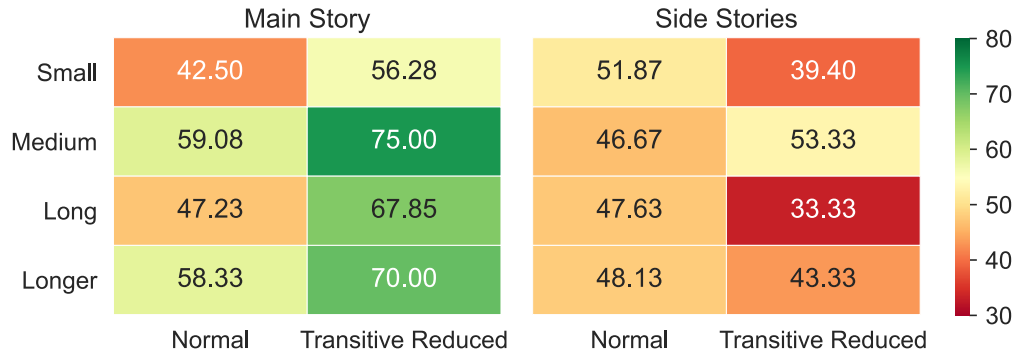


Figure 6.8: Heat map showing the average recall for the main story and the side stories averaged over the issues.

a single general map could lead to better results.

6.4.3 User Evaluation Results

How well do users evaluate these narrative maps in terms of effectiveness or utility? We show the evaluation questionnaire results in Figure 6.9 and in Table 6.7.

Table 6.7: Average percentage of favorable responses in our evaluation questionnaire for each *size* and *transitivity* combination. The best result was obtained by Long-T, followed by Medium-T, and then Long-N.

Size	Normal (N)		Transitive Reduced (T)	
	Responses	% Agree	Responses	% Agree
Short	10	67.00	12	63.33
Medium	11	66.36	10	77.00
Long	9	73.33	7	91.43
Longer	9	60.00	10	65.00

First, our best performing map is Long-T on most evaluation metrics, except for the metaphor-related metrics. On average, the second-best performing map is Medium-T and then Long-N. In particular, Long maps have the best performing results for all metrics.

The user preference for Long maps could be caused by their resemblance to timelines, which makes them more intuitive to use, while at the same time providing enough additional complexity to be useful as a narrative map. The tendency of users to prefer timeline-like structures could be related to the fact that timelines are the most basic and natural representation for narratives. Thus, users tend to prefer structures that are most familiar

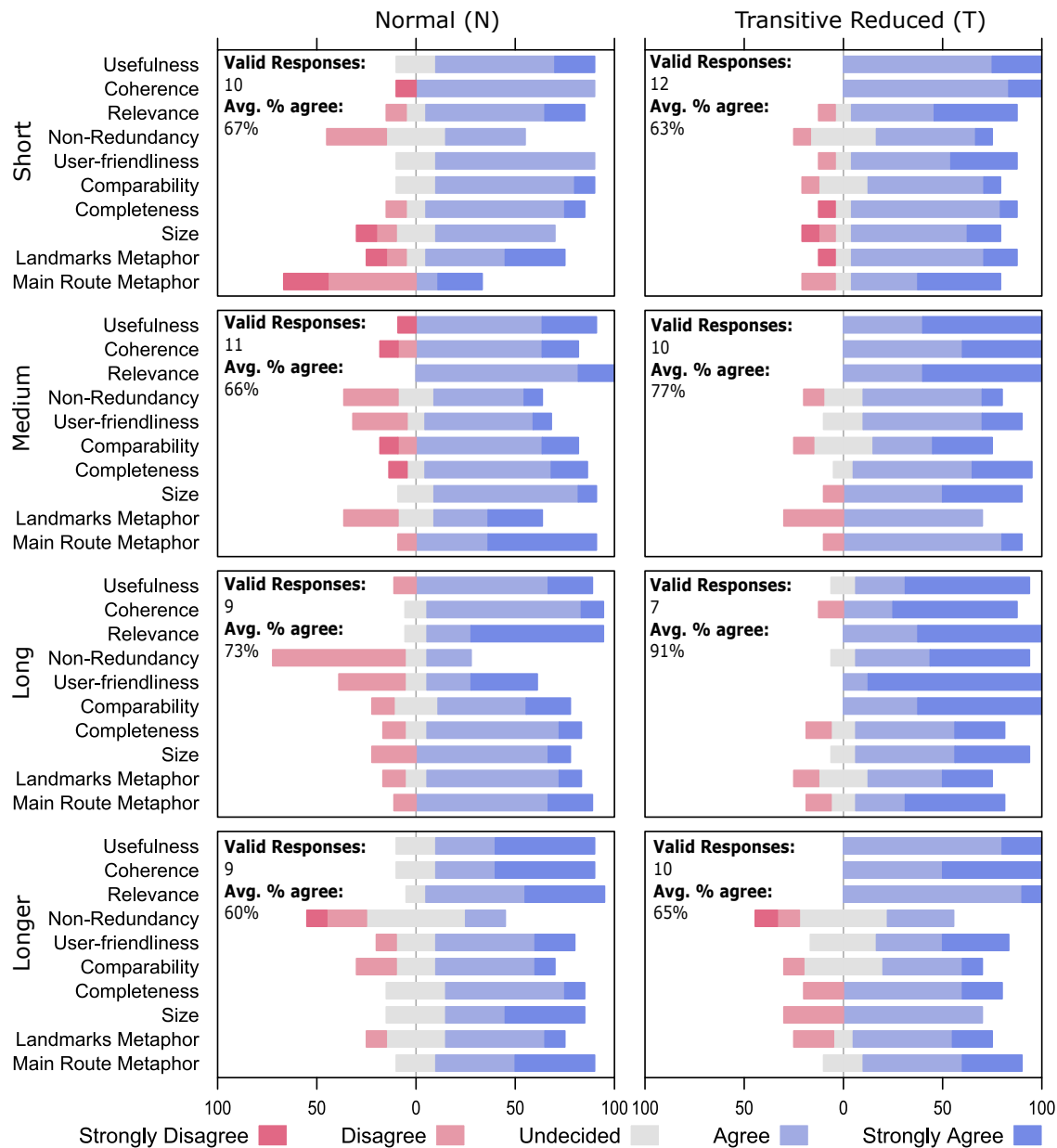


Figure 6.9: Percentage of favorable responses for each question and each *size* and *transitivity* combination. The best result was obtained by Long-T, followed by Medium-T, and then Long-N.

to them. Moreover, bigger maps are naturally able to contain more information than their smaller counterparts. However, Longer maps were not as well-received as Long maps, likely due to their unwieldy size and amount of content which made them impractical to use.

Transitive reduced maps were better in all metrics except comparability (i.e., the ability to compare storylines), with an overall average of 72.8% compared to 67.7% for normal maps.

For comparability, normal maps had 71.1% favorable responses compared to only 56.4% for transitive reduced maps. We hypothesize that this difference could be due to transitive reduction removing too many connections between storylines. Thus, by simplifying the map we lost important connections, making storyline comparison more difficult. Next, if we aggregate maps of the same size disregarding the effects of transitivity, Long maps have the advantage with an average of 81.3% favorable responses, followed by Medium maps with 71.4%.

Finally, we note that using transitive reduction on Short maps actually hurt the overall effectiveness (63% compared to 67%). This could mean that the extra connections present in the Small-N map were useful. Since smaller maps have fewer events, users benefited from knowing the connections between them. In contrast, bigger maps benefited from the use of transitive reduction to minimize complexity, at least up to a certain point, as Long-T performed better than Longer-T. Transitive reduction on Long maps had the highest positive effect on average user evaluation (from 73% to 91%). However, the benefits started to reduce as the map got bigger (from 60% to only 65% for Longer maps). Thus, this indicates that there is a sweet spot for the size of the map where transitive reduction has its greatest impact on effectiveness.

6.5 Design Guidelines

6.5.1 What makes a good narrative map?

Based on our analysis of the results from all experiments, we present our narrative map design guidelines. These guidelines try to encapsulate the optimal design of narrative maps in the context of visual analytics and narrative sensemaking tasks. These recommendations seek to provide a general overview of what makes a “good” narrative map. Table 6.8 summarizes the design guidelines.

Map Layout and Main Story Position: Narrative maps should have a vertical layout. Participants commented that vertical maps worked better since scrolling on a screen lends naturally to vertical layouts. In visual storytelling terms, narrative maps should follow the *scrollytelling* article design [318]. Moreover, as our event headlines were horizontal, placing them vertically simulates natural reading in English. The influence of reading order is a known factor in how people perceive visual narratives [316]. Furthermore, we note that all our participants used standard computer screens. However, for smaller displays, such as cellphones or tablet screens, displaying maps horizontally might make more sense (i.e., closer to a *slideshow* [316] in visual storytelling terms). Besides the general layout, we found that the main story should be the first story in the layout (assuming a top-bottom and left-to-right reading order).

Starting and Ending Events: Narrative maps should have a single starting event and

Table 6.8: Summary of the design guidelines found in our analysis of results from RQ1 and RQ2.

Design Guidelines	Summary
Map Layout and Main Story Position	According to our analysis, maps should be vertical  with the main story displayed first  (assuming a top-bottom and left-to-right order).
Starting and Ending Events	From our results, we conclude that maps should have a single starting event  (i.e., one source node) and potentially multiple ending events  (i.e., many sink nodes).
Connectivity and Transitivity	Narrative maps should be connected  (i.e., there should be no islands). Moreover, users prefer maps without explicit transitive connections  .
Map Size and Main Story Length	In the context of our data set, narrative maps should have about 12 events in the main story.
Cognitive Connections	Narrative maps should include multiple types of cognitive connections; ideally, they would leverage all types of connections.
Event Selection	According to our analysis, maps should focus on hard facts and impactful events .
Graph Structure	Structured approaches (trees  or DAGs ) are preferred over simple timelines. DAGs provide a more flexible and general approach over trees.
Map Labeling	Narrative map edges should include connection explanations and express quantitative information as edge widths . Important events should include a justification for their importance.
Storyline Presentation	Events should be partitioned into multiple parallel storylines , labeled and represented as columns . There should be a small number of important inter-story connections .
Interactions	Relevant interactive features include adding a recommendation system for events in the narrative and the ability to emphasize relevant keywords based on user feedback.
Task-specific Maps	Narrative maps should be designed with a specific target task (e.g., directed and open-ended task). Users should have the option to create different maps based on their intended task.

potentially multiple ending events. However, our results might be biased because we gave participants a starting event in all tasks, but only the directed task had a predefined ending. However, most participants did not add their own starting events, even though many of them added their own ending events in addition to the predefined one when applicable.

Connectivity and Transitivity: Narrative maps should be connected and avoid using

connections that can be induced by transitivity. During the creation process, almost no participants created disconnected components for their maps. This makes intuitive sense, since if something is so separate from the map that it must be its own disjointed component, then it is likely irrelevant to begin with. Regarding transitivity, both experiments showed that users generally prefer simpler maps where explicit transitive connections are omitted.

Map Size and Main Story Length: Based on our analyses, we found narrative maps should have main stories of length 12, as this size worked better in our evaluation context. In RQ1 the mean number of events in the main storyline was 14.3 events. However, these maps had high variability and there were many outliers. Thus, the median provides a better measure of the actual central point. In particular, the median was 12.5, which aligns with Long maps ($K = 12$) having the highest user evaluation in RQ2. Nevertheless, there are some caveats with this main story size. In particular, these results might depend on the size of the data set, the length of the timeline itself, and our resolution level. For example, if we are constructing a narrative that lasts a few years and we have daily updates, then 12 events might not be enough. Therefore, the results of this design guideline are specific to this data set. However, it might be applicable to similarly sized data sets, such as those used in other studies on sensemaking [51, 301]. Finally, we note that based on the recall performance, an argument could be made towards favoring medium-sized maps. However, for the purposes of our design guidelines, we are interested in what users consider good in maps, rather than maximizing performance on this specific experiment.

Event Selection: Regarding the event selection criteria, narrative maps should be focused on hard facts and impactful events. Some events could be highlighted based on structural criteria, such as events that act as a hub node in the graph. Moreover, events should be selected from a variety of sources to create an unbiased narrative (e.g., including articles from left-leaning and right-leaning outlets to provide a politically balanced view of a narrative).

Cognitive Connections: Regarding connection types, rather than focusing on creating maps with a specific type of preferred connection, narrative map tools should allow the creation of different types of connections and provide an explanation of the type of connection (e.g., “Common entity: China”, “Cause-effect relationship”, or “Same topic”). In general, based on the results and the use of different connection strategies by participants in our first experiment, a good narrative map would use a mix of different types of cognitive connections, rather than focusing on a specific one.

In this respect, no existing tool in the literature handles all these cases. For example, the narrative maps extraction algorithm uses similarity and topical connections [186], but it does not include any cause-effect relationship or entity-based connections. In contrast, consider the Analyst’s Workspace designed by Hossain et al. [161], which uses entity-based connections to generate storylines, but does not leverage topical information. As another example, consider the causal storytelling visualization technique developed by Choudhry et al. [75], which explicitly models causal relationships, but does not exploit other types of cognitive connections. Thus, we posit that there is a need to develop a narrative representation and

extraction model that can leverage all these types of connections.

Finally, we note the absence of *citation-based connections* in the constructed narrative maps (i.e., A references B). This is a consequence of only considering headlines rather than the full articles, which could theoretically include links to previous articles in their body. However, even if we had the full text of the articles, we do not have HTML versions with hyperlinks available. Thus it would not be possible to find such type of connections with this data set. We note that this is a low-level type of connection, as it only requires analysts to detect the reference in the document, without necessarily analyzing it in more detail. However, it could turn into a higher level connection if the analysts detect why the reference was made in the first place.

Graph Structure: Regarding the map structure, the map construction experiment showed that users preferred structured approaches (trees and DAGs) over simple timelines. However, there was no apparent preference between the tree-based approaches and the DAG-based approaches for the participants of RQ1. Thus, narrative maps should either use trees or DAGs for their underlying structure. However, we find that in theoretical terms, DAGs provide the most flexible representation. DAGs can be used to show divergent and convergent storylines, allowing for greater representation capabilities compared to a tree. In particular, DAGs are also able to model timelines and trees. Thus, these representations could be interpreted as special cases of DAGs.

Map Labeling: The map should include labels that explain key components (i.e., connections and important events). Regarding edges, we note that the use of edge labels with numerical information can hurt usability, as discussed by Keith and Mitra, a better alternative is to instead show such information via edge widths [186]. In our experiments, we found that most participants did not include explicit weight information for the connections nor any other labeling. Moreover, they did not use different width arrows to denote connection strength. However, in the follow-up interviews, some participants were interested in having an explanation of why two events were joined together in the connections. This would aid other analysts in understanding the reasoning behind map connections. Thus, edge labels should provide qualitative information about the connection (e.g., “Causal Connection”) and could also be color-coded for user convenience. Quantitative information, such as connection strength (e.g., coherence), could be displayed using edge width. Similarly, participants also showed interest in explanations or justifications regarding why a specific event was deemed as important. Thus, maps should include additional labels providing information regarding connections and important events.

Storyline Presentation: Events should be partitioned into parallel storylines. The storylines should be labeled and represented as columns. Moreover, there should be a small number of important inter-story connections joining them. We note that all participants separated their events into clearly defined parallel storylines of varying lengths and labeled their storylines. The labels were based on the general topic or issue presented in the storyline (e.g., prevention or economic impacts). In most cases, these storylines had few inter-story

connections between them. Participants also labeled their storylines.

Interactions: In addition to all the structural and layout information obtained from our first experiment, we also obtained some insight towards potential useful interactions for a narrative maps tool. Beyond basic surface-level interactions, such as zooming, moving elements, changing layouts, and adding/deleting elements, participants were interested in obtaining recommendations on which events to add next as they constructed the map. In general, users expressed interest in a recommendation system to enrich the narrative map. Moreover, this system could be extended to also detect missing events in existing storylines. We note that there are other interactions or approaches that could be beneficial when confronted with a larger corpus of related articles. For example, using similarity of the articles to find related story pieces or citations and cross-references to detect other relevant articles. Such dependencies could be offered so that analysts could drag and drop new events into existing paths.

A close analogue in the literature to the narrative maps method is the metro maps approach developed by Shahaf et al. [321]. This visualization tool incorporates its own narrative representation and extraction algorithm. In particular, it incorporates user feedback through the selection of important tags (i.e., selecting relevant words according to the user’s interests). A similar approach could be used to incorporate user feedback into narrative maps. Lastly, our participants mentioned the idea of incorporating user feedback through keywords as a way to obtain a more relevant map. This could be implemented by emphasizing events based on input keywords on a search bar or highlighted words by the user.

Task-specific Maps: Narrative maps should be designed with a specific target task. We found that using a single narrative map to attempt answering both the directed and open-ended tasks was sub-optimal. The current extraction algorithm focuses on the directed task, thus leading to lower performance in recognizing important side stories. While a combined map for both tasks can provide an appropriate overview, it would be better to create task-specific maps. For example, for the open-ended task, we could generate a map without a fixed ending event and with high topic coverage, as this would likely lead to a narrative map that explored multiple outcomes.

6.6 Discussion

6.6.1 Visual Storytelling and Narrative Maps

Regarding the use of visual storytelling techniques with narrative maps, we discuss some concepts, taken from the work of Segel and Heer [316], that could be potentially useful in the design of a narrative maps visualization tool.

Regarding *visual narrative* elements. We note that narrative maps should guide viewers to

explore paths in the visualization through the use of *visual highlighting* (e.g., color, size, boldness). In practice, this would correspond to highlighting the main storyline, but there should also be clear indications for the side stories. The ability to perform *close-ups* or *zooming* into relevant narrative map sections is also important.

Regarding *messaging*, narrative maps already include the *headlines* of the events as core elements of the narrative. Nevertheless, there are other messaging tools from storytelling that could be used. For example, *annotations*, such as edge labels, storyline names, or other macro-structures names (e.g., clusters) to the narrative map could prove useful as well. The inclusion of a *summary* could also be a useful feature for narrative maps, as it would be able to provide additional context and a brief overview of the content of the map.

Regarding *interactivity* elements, narrative map visualizations should also consider including a *details-on-demand* feature, either by mousing-over an event on the graph or by clicking on them. Such a feature could open a special details tab, containing information such as the full article, a snapshot of the original publication, or even a list of related articles. It could also be useful to include a *timeline slider* element, as it could allow users to change the scope of the visualized narrative to a different time window. Moreover, it should be possible to perform *filtering*, *selection*, and *searching* over the events of the narrative.

6.6.2 Influence of Analyst Background and Experience

First, we note that the analysts were not working professional analysts, but were student analysts-in-training. Thus, their specific sensemaking strategies might be influenced by their lower level of experience compared to real analysts. Moreover, if analysts were familiar with structured analytic techniques [280], such as the *generic narrative space model* [310] or other methods, they might affect their sensemaking process, as these techniques provide ways to develop compelling narrative rationales [72, 280]. Nevertheless, previous work has shown that studies with real analysts and with students have similar findings and implications [179]. Exploring the influence of specific analyst experience and is left as future work. More specifically, future work should include the study of more cases with professional analysts.

6.6.3 Influence of the Data Set and Task Choice

Regarding the data set, we note that the use of a current topic such as COVID-19 might have influenced the results, as participants could have been influenced heavily by their own experiences with the pandemic. Moreover, the data set was relatively small, a limitation imposed due to time constraints. The data set size could make it difficult to scale the detected strategies or results to larger data sets, which, for example, could place more emphasis on the foraging steps of the sensemaking loop or require more complex narrative map structures. Regardless of these issues, the COVID-19 data set should still provide valuable insight into

the synthesis loop part of the sensemaking process. Moreover, the data set size is in line with related works [51, 301] that use intelligence analysis data sets [169], such as *The Sign of the Crescent* data set (41 documents) or the *Atlantic Storm* data set (47 documents) to understand the analyst sensemaking process.

It should also be noted that experience and prior knowledge might heavily influence the work done by participants, especially due to the use of a recent and high-profile topic such as COVID-19. In particular, participants had different levels of expertise on the topic and were able to bring insights from their own knowledge and experiences. Specifically, in the RQ1 experiment, we note that only four analysts made explicit remarks on how they used domain knowledge in their construction process. However, the other six analysts might have drawn on this knowledge implicitly without properly acknowledging it.

In addition to this, we note that the specific choice of starting and ending events in the directed task also influences the construction of the map and what is considered part of the main storyline or a side story. For example, when trying to find the connection between the initial outbreak and travel restrictions, it is unlikely that documents relating to oil prices are directly part of the main story. However, if the question required connecting the dots between the initial outbreak and the economic impacts it would make more sense as part of the main storyline.

Finally, we note that both of the tasks used in this study represent simplified and constrained versions of what analysts would do in a real-world setup, but they still provide valuable insights into the general narrative sensemaking process. Nevertheless, as these tasks do not capture the full sensemaking process, caution should be exercised when attempting to generalize these conclusions, especially as higher complexity tasks might yield different kinds of strategies or structures.

6.6.4 Sensemaking Process

We note that much of the evaluation process rediscovers parts of the larger sensemaking process. However, in this chapter, we focus exclusively on how the synthesis loop of the sensemaking process applies to narrative maps. Thus, the results are only applicable to this scope. Future work could address how other types of sensemaking strategies or tools compare against narrative maps in the context of sensemaking.

Furthermore, it would be useful for future work to do multiple evaluations with data sets with different characteristics and analysts with different levels of experience. Such work could ask analysts to create a narrative map based on their own analytical work that they have previously completed as part of their regular practice, as opposed to using a specific toy data set, although such an approach would have several more variables to account for, requiring careful experimental design. However, there would be value in drawing lessons and guidelines from a more diverse set of analytic problems, as this would also provide

information on where and how narrative maps could be best applied.

6.6.5 Limitations

Our work is not without limitations. First, there is an unbalanced number of participants in the two experiments, the first one has 10 subjects, based on the methodology of Bradel et al. [51], while the second had 78 valid responses. Due to the qualitative nature of the first experiment and the need to understand the construction strategies in depth, it was necessary to use a much smaller sample size. In contrast, the second experiment did not require such a level of detail, making it much simpler to scale up. However, we note that the difference in sample size makes comparing results between these experiments more complex.

Regarding the limitations of the RQ1 experiment, we note that we conducted interviews with only a handful of analysts (10). While the number was small, all participants had a background in intelligence analysis. They also spanned a variety of majors and had reasonable gender representation (6 females and 4 males). Nevertheless, even with 10 participants, we were able to observe diverse strategies and structures for narrative map construction.

Regarding the limitations of the RQ2 experiment, we first note that each factor combination had a different response rate, as not all participants completed the assigned tasks. Nevertheless, the general trend still provided useful insight towards how to design narrative maps. Another issue was the lack of experience of the participants; however, the data set was small enough and the questions were designed to be simple so even non-expert users could answer them. Finally, we note that this experiment lacks an explicit baseline, such as a basic timeline or similar representation.

6.7 Chapter Conclusions

We studied how analysts construct narrative maps and the characteristics of these maps. In particular, our user study detected 7 types of cognitive connections. In particular, we have shown the importance of topical and causal relationships in the construction of narrative maps, as these were the most common high-level connections in the user-generated maps.

In terms of strategies, we found three major ways to construct maps. Each one of these strategies can be the basis of a new narrative extraction algorithm. Furthermore, in terms of the structure of the map, we saw an even distribution between tree-like maps and DAG-like maps. Regarding layout, we found that most users preferred a vertical top-down layout (i.e., *scrollytelling*), with the main story shown first. We also evaluated the effect of map size and transitivity, finding that users preferred long maps without transitive connections.

All these results led to a series of design guidelines for narrative maps. These guidelines can be used in the design of new narrative map extraction algorithms and interactive visualization

tools. Future work will deal with the implementation of such algorithms and tools, as well as their evaluation based on the insights gathered in this work.

Future work could explore how strategies differ when applied to different domains, data set sizes, and analyst experience. In particular, it would be useful to consider how previous analyst training (e.g., experience with structured analytic techniques) could influence the construction strategies or the narrative map structures.

Finally, as mentioned before, the overarching goal of our study was to improve on the narrative map representation proposed by Keith and Mitra [186]. Thus, by extracting these design guidelines and understanding the narrative sensemaking process, we have provided the basis for future improvements of the narrative map model. Thus, future work should focus on using these findings to improve the narrative map representation and the associated extraction algorithms.

Chapter 7

Modeling Accepted Narratives in Online Communities

In our goal to model complex narratives (RQ4), we tackle the problem of modeling the narratives of social movements. Social movements are dominated by storytelling, as narratives play a key role in how communities involved in these movements shape their identities. Thus, recognizing the accepted narratives of different communities is central to understanding social movements. In this work, we incorporate the concept of community acceptance into a narrative extraction algorithm to find the accepted narratives in social media communities discussing a social movement. Furthermore, we develop a narrative comparison metric based on graph edit distance to compare the narratives of different communities. In particular, we study the 2021 Cuban protests and characterize the top five communities in terms of submissions related to this movement. The extracted narratives differ in structure—ranging from pseudo-linear narratives to tree-like structures—and content—ranging from pro-communist propaganda to right-leaning narratives using the protests to criticize their political opponents in the US. Our work has implications in the study of social movements, intelligence analysis, computational journalism, and misinformation research.

7.1 Introduction

Social movements are dominated by storytelling [86]. In fact, narratives play a key role in social change and social movements [36], as people use them to learn and exercise agency, shape identity, and motivate actions [94, 126]. Social movements are made possible by the construction of collective narratives within free spaces [82]. In particular, online communities in social media provide free spaces for people to develop such narratives [203]. Thus, studying how narratives evolve in these online communities is key to understanding social movements.

Narratives provide a *cognitive roadmap* by which community members are to live and accepted narratives define the collective identity of the group and their relations with other groups [183, 184, 212]. In the context of social movements, the narratives promoted by members of the movement and opposition forces can be leveraged to mobilize, channel, and legitimate collective action [250]. Thus, modeling the *community acceptance* of a narrative can offer insights into their motivations and goals.

In this chapter, we study the *2021 Cuban protests*—the biggest protests in decades in Cuba [302]—and the narratives surrounding this movement that have emerged in online communities. Cuba has a unique sociopolitical, economic, and cultural landscape [100]. Its long-standing communist regime faces unique challenges [12]. In particular, the continued US embargo [384] has undermined the Cuban economy [180, 332] and its health care system [9, 32]. The protests were triggered by food and medicine shortages [120], as well as by the inefficient handling of the COVID-19 pandemic by the Cuban government [112].

In order to understand the narratives that emerged in online communities surrounding the 2021 Cuban Protests, we sought to answer the following research questions:

- **RQ1:** What are the accepted narratives in online communities surrounding the social movement of the 2021 Cuban Protests?
- **RQ2:** How do the accepted narratives differ across online communities?

We focus on a specific social media platform—Reddit—to explore how narratives surrounding the 2021 Cuban protests change across communities. Specifically, we characterize the accepted narratives in a number of communities that show a high level of activity related to the protests. We perform our analysis by extracting narrative maps [186], an event-based directed acyclic graph narrative representation. To find the accepted narratives of a community, we further propose a novel extraction algorithm that models *community acceptance*. Additionally, we offer a comparison metric based on graph edit distance to quantitatively evaluate how these narratives differ. We make the following contributions:

1. An **analysis of the accepted narratives** espoused by online communities surrounding the social movement of the 2021 Cuban protests.
2. A narrative **extraction algorithm** that models *community acceptance* in online communities.
3. A narrative **comparison metric** based on graph edit distance to quantitatively evaluate differences in narratives.

We believe our approach has widespread implications, both for the study of social movement narratives online and narratives in general. At its core, our methods allow users to understand the accepted narratives of online communities. Furthermore, expert users, such as intelligence analysts, journalists, and misinformation researchers could use our methods to explore the narrative landscape of social movements emerging in online communities. Beyond the social movement context and the use of community acceptance, our methods could be adapted to include other attributes, such as bias, credibility, or sentiment to model complex narratives in social media.

7.2 Related Work

7.2.1 Narrative Extraction

Narratives are systems of stories with coherent themes [145], where each story is represented as a sequence of events [1]. Most narrative extraction methods use an event-based representation, where each event is associated with a single document in a data set [186, 232, 319]. We follow this approach, as event-based representations have strong theoretical foundations in narratology [1].

Narrative extraction approaches usually rely on optimizing different criteria to extract narratives from data [319, 321, 369]. We use a narrative extraction algorithm based on the criteria of coherence maximization through linear programming, expanding upon the narrative map extraction algorithm proposed by [186]. However, unlike previous models, we integrate the concept of *community acceptance* in our optimization model to select important events and storylines.

7.2.2 Social Movement Narratives

Narratives are key to the formation of social movements [36, 87]. In this context, social media plays a key role in shaping the narrative [176, 203] and mobilizing people [215]. Thus, several works have sought to model these narratives. Computational methods to study social movement narratives in social media usually rely on keyword and hashtag analysis [370, 393], topic modeling [333], and user network analysis [359]. Other approaches include qualitative research, focusing on interviews and the experiences of people [83, 94, 202]. However, none of these approaches consider the underlying structure of the narrative, which depends on temporal and causal relationships between events. In contrast, our approach extracts the narrative structure, allowing for a detailed model of the social movement narrative via narrative maps—an event-based directed acyclic graph narrative representation.

7.3 Data Set Description

We used Pushshift [34] to extract Reddit data. In particular, we extracted all submissions that contained the word “Cuba” from 07/07/2021 to 07/17/2021, covering a few days before the protests started on July 11th, to capture any useful signals on social media prior to the start of protests. We retrieved 4810 submissions. We focused on the top five communities according to their number of posts (see Table 7.1). We stopped at five because the sixth community (r/FreeKarma4U) was a *karma farming* group [11] unrelated to the protests that did not have a cohesive ideology or group identity. Moreover, the number of posts rapidly declines after this point. We note that the keyword “Cuba” could lead to results that are

not immediately relevant to the protests. However, during this short time period, most of the retrieved submissions were indeed related to the protests or their causes. To confirm, we manually analyzed all the posts in the top five communities by the number of submissions and found that roughly 95% of the submissions out of 550 were relevant.

Table 7.1: Summary of the analyzed Reddit communities, including the number of subscribers and submissions retrieved over the 10-day period (July 7 - 17).

Community	Description	Posts	Subscribers
r/cuba	The biggest Cuban community in Reddit.	168	36k
r/Conservative	US-centric right-leaning community focused on political issues.	126	833k
r/GenZedong	Pro-communist community focused on humor and world news.	99	33k
r/worldnews	Community focused on world news (excluding US news).	79	26.6M
r/politics	US-centric left-leaning community focused on political issues.	78	7.6M

7.4 RQ1: Extracted Narratives

In this section, we present our narrative extraction method and the narratives from each community.

7.4.1 Narrative Extraction Method

We use narrative maps [186] as our basic representation. Narrative maps provide a generic foundation to encode different types of narratives extracted from data, requiring only the existence of a total ordering (e.g., in the form of timestamps) and text representation of the event (e.g., a news headline or a post title). Based on a *route map* metaphor, they consist of a directed acyclic graph of event nodes, where each path represents a route connecting events. These nodes are called *landmarks*, based on the idea of route maps having a series of landmarks to highlight important elements. Certain events are highlighted—called *representative landmarks*—at the point of highest width in the graph to capture all parallel storylines. Finally, by finding the path of maximum coherence we find the main route that contains the core events of the narrative. Narrative maps are extracted by optimizing coherence relations between events—based on event similarity and topical consistency—subject to coverage constraints.

In this work, we propose a new optimization model for narrative maps that accounts for *community acceptance* in addition to coherence and story coverage. First, we incorporate

additional parameters to account for the **submission score** (i.e., the difference between upvotes and downvotes) and **upvote ratio** as a way to measure **community acceptance** of a submission. In particular, posts that have a high score and upvote ratio in a subreddit are considered accepted by that community [58, 224] and therefore should appear in the corresponding narrative. We model this by imposing a new constraint that requires that the average score percentile of all the submissions in the narrative map to be above a certain threshold. Note that this does not necessarily discard all the low score submissions, as their presence might be necessary to construct a coherent narrative.

Furthermore, we incorporate **community acceptance** into the optimization objective itself, extracting maps by maximizing both coherence and acceptance. We show the proposed linear programming formulation in Figure 7.1. Our goal is to maximize the minimum edge strength as defined by **minedge**, which depends on **coherence**—a measure of how much sense it makes to join two events together, based on event similarity and being part of the same storyline or cluster—and **acceptance**—a measure of community acceptance that depends on the upvote ratios and submission score in the community. Table 7.2 shows all the elements of our model.

$$\begin{aligned}
& \max \quad \text{minedge} \\
& \text{s.t.} \quad \text{minedge} \leq \text{coherence}_{ij} + \text{acceptance}_{ij} - \text{edge}_{ij} \quad \forall_{ij} \\
& \quad \sum_i \text{event}_i = K \\
& \quad \sum_{ij} \text{edge}_{ij} = K - 1 \\
& \quad \sum_{i < j} \text{edge}_{ij} \geq \text{event}_j \quad \forall_j \\
& \quad \sum_{j > i} \text{edge}_{ij} \leq \text{event}_i \quad \forall_i \\
& \quad \text{edge}_{ij} = 0 \quad \forall_{i \geq j} \\
& \quad \text{cluster}_k = \sum_{ij} \text{edge}_{ij} \cdot \text{membership}_{ijk} \quad \forall_k \\
& \quad \text{mincover} \leq \frac{1}{\text{num-clusters}} \cdot \sum_k \text{cluster}_k \\
& \quad \text{minscore} \leq \frac{1}{K} \cdot \sum_i \text{event}_i \cdot \text{score-percentile}_i.
\end{aligned}$$

Figure 7.1: Linear program formulation for the narrative extraction problem with community acceptance.

Regarding the parameters, we set **mincover** = 0.5, ensuring that at least half of the stories are covered without excessively restricting the narrative. We set $K = 8$ to get main storylines with around 8 events, based on estimates from similar approaches [186, 319]. We

Table 7.2: Description of the variables and parameters of the proposed extraction algorithm. All elements take values between 0 and 1, except K and **num-clusters**, which are positive integers. Most elements of the model are based on the extraction algorithm of Keith Norambuena and Mitra [186]. The (*) symbol represents new additions to the linear problem formulation designed to incorporate *community acceptance* in social media into the model.

Type	Name	Description
Variables	minedge	Represents the lowest coherence value over all active edges.
	event _{<i>i</i>}	Represents the presence of event <i>i</i> in the narrative map.
	edge _{<i>ij</i>}	Represents the strength of the connection between events <i>i</i> and <i>j</i> in the narrative map.
	cluster _{<i>k</i>}	Represents the average presence of cluster <i>k</i> in the narrative map.
Data-based Parameters	num-clusters	Represents the number of clusters of events.
	membership _{<i>ijk</i>}	Represents whether the connection of events <i>i</i> and <i>j</i> belongs to cluster <i>k</i> . Defined as $\sqrt{\text{membership}_{ik} \cdot \text{membership}_{jk}}$ —the geometric mean between the probabilities of the events belonging to cluster <i>k</i> .
	coherence _{<i>ij</i>}	Represents the coherence score of the connection between events <i>i</i> and <i>j</i> . Defined as $\sqrt{\text{angular-similarity}_{ij} \cdot \text{cluster-similarity}_{ij}}$ —the geometric mean between the angular similarity of the underlying event embeddings and the clustering similarity based on comparing the cluster membership probability distributions of each event using Jensen-Shannon divergence.
	score-percentile _{<i>i</i>} (*)	Represents the score percentile of event (submission) <i>i</i> in the community.
	upvote-ratio _{<i>i</i>} (*)	Represents the upvote ratio of event (submission) <i>i</i> in the community.
	acceptance _{<i>ij</i>} (*)	Represents the acceptance score of the connection between events <i>i</i> and <i>j</i> . Defined as $\sqrt{\text{score-percentile}_i \cdot \text{score-percentile}_j} \cdot \sqrt{\text{upvote-ratio}_i \cdot \text{upvote-ratio}_j}$ —the product between the geometric means of submission score percentiles and upvote ratios.
User-defined Parameters	K	Defines the expected length of the main storyline in the narrative map.
	mincover	Defines the minimum average coverage over all clusters in the narrative map.
	minscore (*)	Defines the minimum average acceptance over all events (submissions) in the narrative map.

set $\text{minscore} = 0.85$, ensuring that the average community acceptance of at least 85%—intuitively, this means that roughly 85% of the community regards the selected submissions positively. In particular, lower thresholds allowed too many events with a low acceptance level to appear on the map. We encode submissions with Multilingual USE embeddings [403], as some communities have submissions in English and Spanish. Finally, we use HDBSCAN [243] with UMAP [244] for clustering as used in previous extraction methods [186].

7.4.2 Results

Next we present the accepted narratives extracted from each community. We focus on the community with the highest number of submissions (*r/cuba*), although we also present the most relevant parts of the other communities.

r/cuba: Figure 7.2 shows the narrative map extracted from *r/cuba*. The main storyline starts with a submission criticizing the standing of Cuba as a leader in health care. Moreover, it highlights economical issues and political repression. Next, a submission discusses a Cuban

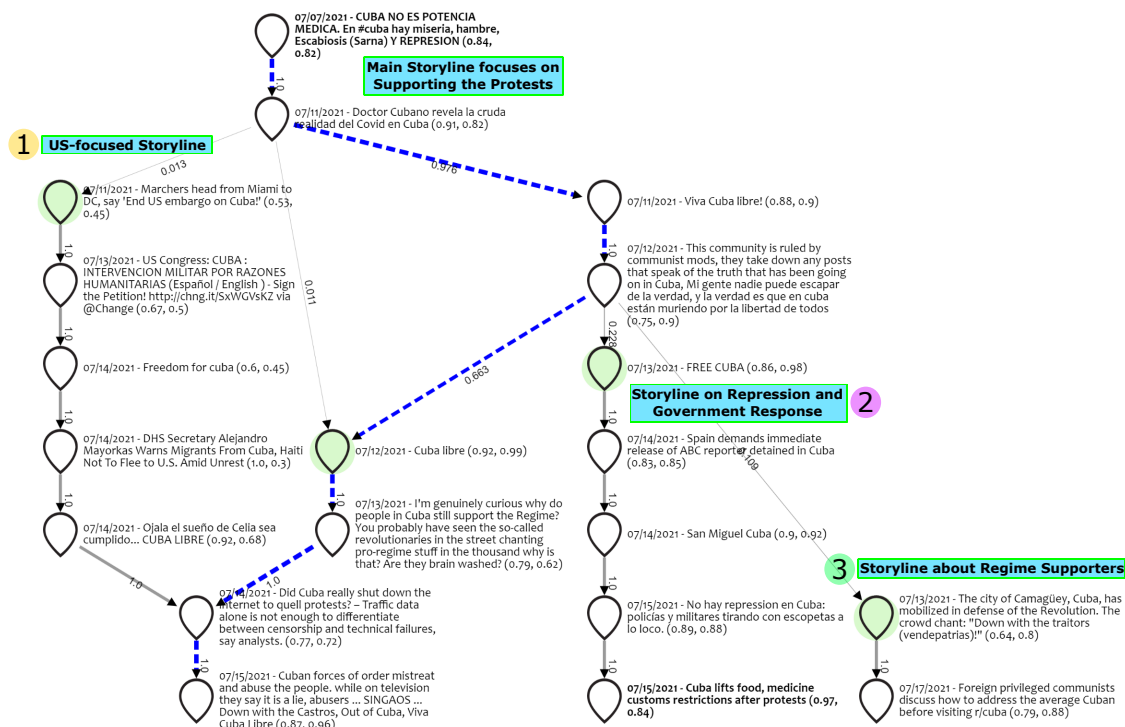


Figure 7.2: Narrative map for *r/cuba*. The numbers in parenthesis beside each submission represent the upvote ratio and score percentile. The main storyline is represented by the blue dashed lines. Highlighted nodes correspond to the representative landmarks of the narrative map. The edge weights represent the coherence normalized as a probability. Manual annotations describing specific storylines have been added with markers 1, 2, and 3.

doctor revealing the reality of the pandemic in Cuba on the same day that the protests start. Overall, the main storyline gives the impression that the protests started due to health care issues, alongside other catalysts.

Afterward, we see a simple submission with the rallying call “Viva Cuba libre!” (translated as “long live free Cuba”) with an image about tyranny in its internal contents. We note that this submission was locked by the moderators of the community. The next day there is a post criticizing the moderators for taking up anything about the protests. Then, we see a rallying call with a video showing the protests. After this, there are two submissions seeking information about the protests, asking about counter-protests and whether the internet has been shut down. Finally, the storyline ends with reports of abuse by Cuban police and a call against the regime.

Second, we analyze the representative landmarks highlighted in green and the storylines they correspond to. Storyline 1 is focused on the US, discussing the end of the embargo, a potential military intervention, and a potential refugee and migratory crisis. We note that the storyline is sprinkled with more rallying calls for a free Cuba, but these are not

immediately relevant to this specific storyline. The second representative landmark is the rallying call with a video showing protests in the main storyline, which makes sense as a key narrative event that shows the actual protests. The third representative landmark is another rallying call with a photo of protesters carrying a Cuban flag. The rest of the storyline [2](#) focuses mostly on repression in the protests, with Spain demanding the release of a reporter, and the police and military shooting protesters. However, it ends on a better note with changes in customs restrictions by the Cuban government due to the protests. The final representative landmark and side storyline [3](#) focus on the counter-protests and support for the regime, with people mobilizing in defense of the communist revolution.

Overall, the side storylines provide three key points that complement the main storyline and point towards potential future events. First, the US is likely to be a key player in this narrative, although it is too early at the time of this writing how it might respond. Second, the protests have at least some limited success so far, such as the lifting of customs restrictions. Finally, there is a strong response by the government and its forces, with police and military forces being used, as well as a counter-movement that backs the government.

r/Conservative: We present the narrative of r/Conservative—a right-leaning community centered on US politics [\[356\]](#)—in Figure [7.3\(a\)](#). The main story starts with the protests and how Cubans seek to end the regime. However, the rest of the narrative focuses mostly on US political actors. The narrative of r/Conservative uses the protests as a tool to criticize its political opponents (left-leaning media, Democrats, Black Lives Matter, etc.), rather than focusing on the protests themselves.

Regarding the representative landmarks, the first one is a satirical article criticizing communist supporters and their views on the regime after the protests. The second one is part of the main story and highlights how celebrities are also stepping in to support the protests against communism.

r/GenZedong: We present the narrative of r/GenZedong—a pro-communist community—in Figure [7.4](#). Due to the linear nature of this narrative, extracting representative landmarks is not useful, as they rely on the existence of parallel stories [\[186\]](#). The storyline starts with an unrelated submission criticizing the presence of an anti-communist Cuban exile in a documentary about Cuba. Although not related to the protests, it aligns with the pro-communist stance of the community. The rest of the timeline supports the regime in Cuba with mostly satirical renditions of anti-communist and pro-US submissions. Overall, this narrative supports the Cuban regime and criticizes its opponents.

r/worldnews: Figure [7.5](#) shows the narrative of r/worldnews—a neutral community focused on global news [\[23, 159\]](#). The main storyline starts with an unrelated event on Cuba developing effective vaccines. This is an interesting parallel with the r/cuba narrative that also started

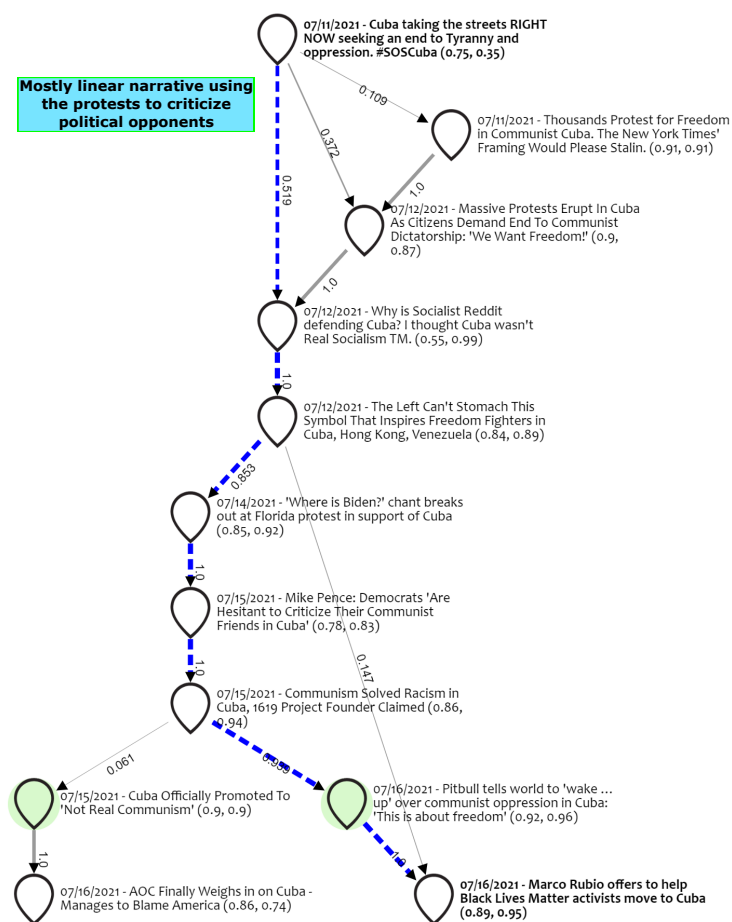


Figure 7.3: Narrative map for r/Conservative.

with a health care-related post but in a negative light. The next event on the chain is the Mexican president calling for an end to the US embargo—a representative landmark of this map. Next, the storyline focuses on how the protests developed in Cuba. Then, the Miami mayor is discussing bombing Cuba. Immediately after this, we see Cuba lifting the customs restrictions, heavily implying that fear of US intervention led to removing the restrictions, unlike the r/cuba narrative that presented this as a result of the protests.

The side story on the left focuses mostly on the international response, including Russia, China, Iran, and the US. Overall, this narrative has two main themes: the protests themselves and the responses by international actors. This makes sense as the purpose of the community is to discuss world news. Moreover, this external perspective is an interesting complement to the r/cuba narrative, as it provides a bigger picture view rather than an emotional one. For example, note the lack of rallying call submissions in r/worldnews. This makes sense as this community is neutral and has a more objective look towards world events.

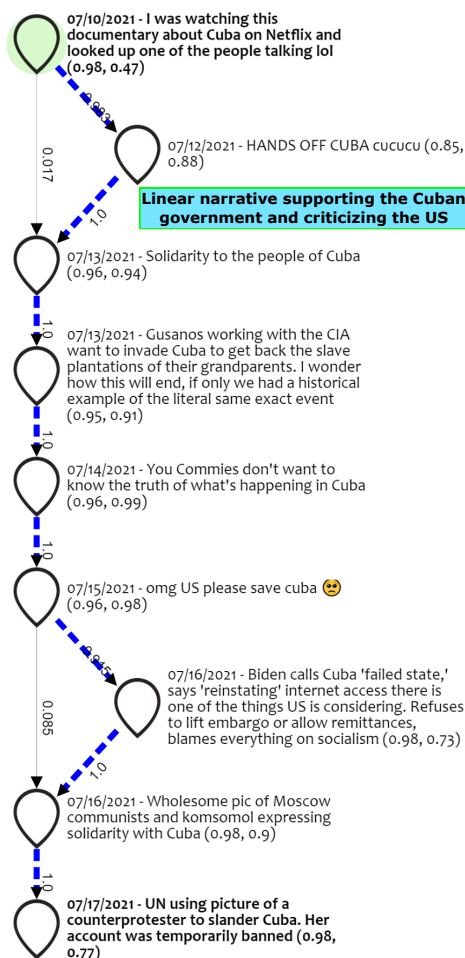


Figure 7.4: Narrative map for r/GenZedong.

r/politics: We present the narrative of r/politics—a left-leaning community centered on US politics [328, 356]—in Figure 7.6. The narrative starts with a report on the protests, then it mostly focuses on calls to lift the embargo. The storyline then shifts towards the protests in Miami, with a focus on the anti-riot law, as marked by the representative landmark of the main storyline. The narrative uses this to criticize the political opponents of the community, in a similar way to how r/Conservative used the Cuban protests to criticize their own opponents.

7.5 RQ2: Narrative Comparison

In this section, we present our narrative comparison metric and then analyze how the narratives espoused by each community differ from each other.

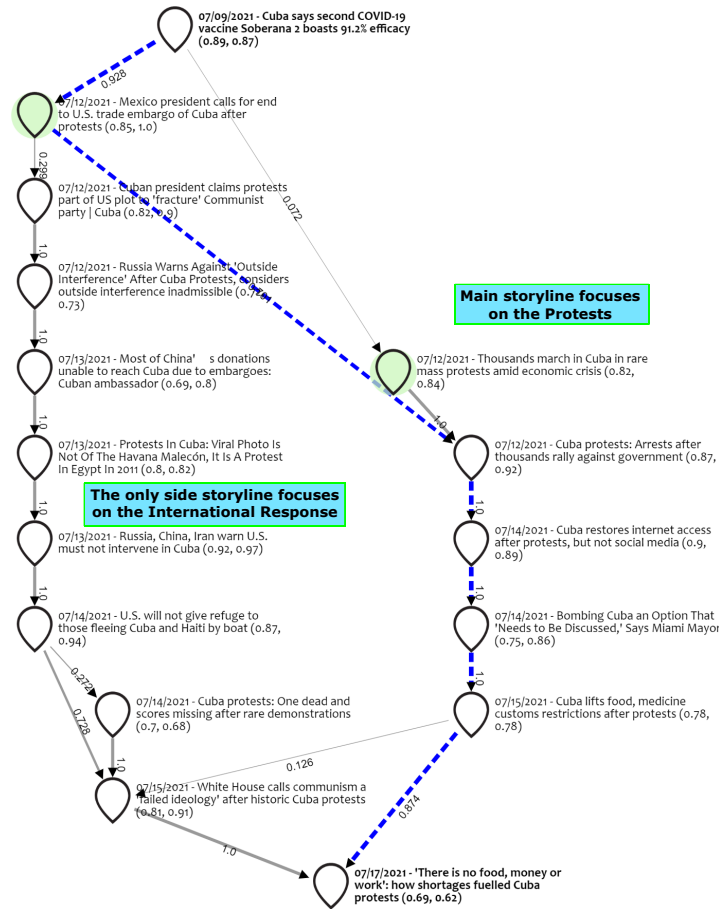


Figure 7.5: Narrative map for r/worldnews.

7.5.1 Narrative Comparison Metric

We quantitatively compare narratives using Graph Edit Distance (GED) [4]. GED is defined by the total cost of the operations (e.g., node and edge substitution) needed to transform one graph to another. In our case, we define the costs based on the similarity of the affected events (nodes) and their connections (edges). For example, replacing a node about an US intervention with a node about the Cuban health care crisis would be costly. However, replacing it with a node about the US embargo would be cheaper, because they share a common entity (the US). GED can detect three types of differences: structure, order, and content (see Figure 7.7).

Table 7.3 shows the costs used to compute Graph Edit Distance. Intuitively, these costs are designed to prioritize substitutions of similar events or connections over deletions or insertions. In practice, due to the use of an approximation algorithm to find the graph edit distance, this is not always achieved and some suboptimal operations are used when others

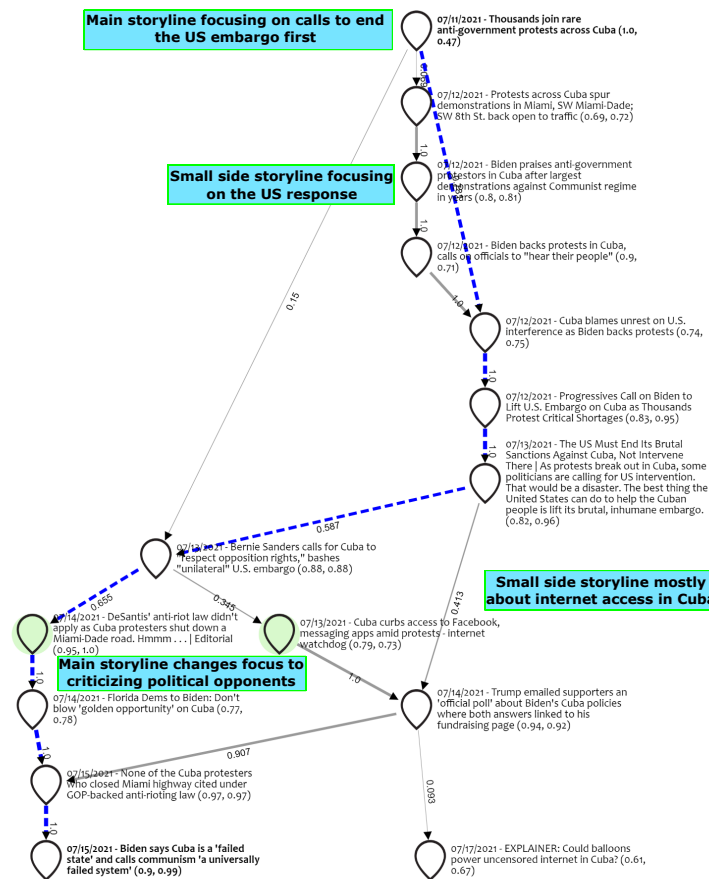


Figure 7.6: Narrative map for r/politics.

would be preferred.

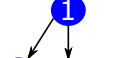
Narrative Maps			
Content Structure Order	Different Same N/A	Same Different Same	Same Same Different
Steps GED	SUB 4 nodes ~0.67	DEL 3 edges ~0.27	SUB 1 edge ~0.25

Figure 7.7: Examples showing the differences captured by (normalized) GED with unitary costs.

Table 7.3: Costs used in the Graph Edit Distance algorithm and a description of each parameter.

Costs	Formula	Description
Node Substitution	$1 - s_{ij}$	The cost of substituting event i with event j depends on the similarity between these events.
Node Deletion	1	The cost of deleting an event is 1. This value ensures that deleting events is always at least as costly as substituting events, the optimal edit distance will prefer substituting events unless necessary.
Node Insertion	1	The cost of adding an event is 1, following the same reasoning as the node deletion cost.
Edge Substitution	$(\sqrt{s_{ij} \cdot w_{ij}} - \sqrt{s_{kl} \cdot w_{kl}})^2$	The cost of substituting edge ij with edge kl depends on how similar these edges are based on their weights and the similarity between their endpoints. The geometric mean is used to combine weights and similarity, as it strictly penalizes low values on either part. The overall difference between the two edges is squared to ensure a non-negative cost.
Edge Deletion	$\sqrt{s_{ij} \cdot w_{ij}}$	The cost of deleting edge ij is based on the geometric mean of its weight and the similarity between its endpoints.
Edge Insertion	$1 - \sqrt{s_{ij} \cdot w_{ij}}$	The cost of adding edge ij is based on the geometric mean of its weight and the similarity between its endpoints.

Due to the high computational cost of computing graph edit distance—an NP-complete problem [47]—we rely on an iterative approximation algorithm implemented in the NetworkX package [4, 143]. In particular, we let the approximation algorithm run until a timer runs out or it naturally completes, whichever occurs first.

Since the costs are defined asymmetrically, as the costs for deletion and addition are not the same for edges, the graph edit distance is not necessarily symmetric—the cost of converting one narrative map into another might depend on the direction in which we do it. Moreover, due to the nature of the approximation algorithm, the order in which we do the computations affects the results. To account for both of these issues, we compute the graph edit distance in both directions and average the results to get the final approximation of GED. Furthermore, we normalized the values over the total number of edges in the two graphs, as this number is a proxy for graph complexity and it provides easier to interpret results.

7.5.2 Results

Qualitative Comparison We manually categorize the relevant storylines and events highlighted by each narrative. We show the results in Table 7.4.

We note that *r/cuba* focuses on all aspects of the protest except the international response, save for the US, given its relevance in the historical and geopolitical context of Cuba. *r/-Conservative* supports the protests, but mostly uses them to criticize their political opponents. *r/GenZedong* supports the protests and also criticizes its opponents, particularly the US. In

Table 7.4: Relevant storylines and events that appear in each narrative map. We note that the presence of storylines is not correlated with the number of submissions. *r/Con** refers to *r/Conservative*, *r/GZ** to *r/GenZedong*, *r/wn** to *r/worldnews*, and *r/pol** to *r/politics*.

Relevant Storylines	<i>r/cuba</i>	<i>r/Con*</i>	<i>r/GZ*</i>	<i>r/wn*</i>	<i>r/pol*</i>
Causes of the Protest	×			×	
Cuban Government Response	×			×	×
Police Brutality and Repression	×			×	
Supporting the Protests	×	×			×
Supporting the Regime (counter-protests)	×		×		
International Response				×	
US Response and Potential Intervention	×		×	×	×
Ending the US Trade Embargo	×		×	×	×
Criticizing Political Opponents	×	×	×		×

contrast, *r/worldnews* focuses on the protests, without explicitly giving support for either side. Finally, *r/politics* shows support for the protests and uses them to criticize their political opponents, like *r/Conservative*. However, it also emphasizes the need to end the US trade embargo.

Quantitative Comparison Our quantitative comparisons comprise the pairwise graph edit distance (GED) comparisons between the communities (see Table 7.5).

Table 7.5: Pairwise Graph Edit Distance (GED) between the different communities.

GED	<i>r/Conservative</i>	<i>r/GenZedong</i>	<i>r/worldnews</i>	<i>r/politics</i>
<i>r/cuba</i>	0.7097	0.7719	0.5641	0.6224
<i>r/Conservative</i>	—	0.4164	0.4988	0.4890
<i>r/GenZedong</i>	—	—	0.6275	0.5938
<i>r/worldnews</i>	—	—	—	0.4498

We find that *r/cuba* differs the most from *r/GenZedong* compared to the other communities. This makes sense intuitively, as the pro-communist linear narrative is quite different both in terms of content and structure compared to the complex tree-like structure of the pro-protest Cuban narrative. We note that the Cuban narrative also differs from *r/Conservative*, as the structure of their narrative is relatively linear. Moreover, that narrative focuses on criticizing political opponents rather than the protest themselves. In contrast, the most similar narrative to *r/cuba*—*r/worldnews*—focuses on the protests and the international response, rather than political aspects.

In contrast, consider the narrative from *r/Conservative*. Despite being on the opposite side

of the political spectrum to *r/GenZedong*, they share structural and superficial content similarities (e.g., similar words). In particular, both of them are mostly linear narratives, both criticize their political opponents, and both express support towards one side of the protests. Likewise, the narrative is also considered similar to *r/politics*. This makes sense since both communities focus on US politics, although on opposite sides.

The pro-communist community of *r/GenZedong* shares some similarities with the left-leaning community of *r/politics*. Likely, this similarity stems from the relative closeness of the content in the political spectrum and the criticisms of the US embargo and, by extension, the US government.

Finally, the narrative from *r/worldnews* is relatively similar to all the other narratives. This similarity could be attributed to either narrative order, content, or structure. The least similar narrative emerges from *r/GenZedong*, which is a highly politicized community without much focus on the news itself. More importantly, this result aligns with the fact that *r/worldnews* should be a neutral community since it is focused on factual content rather than siding with any ideology.

7.6 Discussion

7.6.1 Social Movement Narratives

Our work unravels the different narratives espoused by online communities pertaining to the 2021 Cuban protest movement. In general, our work provides a computational framework to analyze the accepted narratives that emerge in online communities in the context of social movements, allowing analysts and social scientists to understand the motivations and relations between different groups in the context of a social movement. In particular, we found that *r/cuba* had the most complex narrative with a tree-like structure and diverse content covering multiple storylines. In contrast, *r/GenZedong* had the simplest narrative—a single linear storyline—and had pro-communist content. Meanwhile, both the US-based political communities—*r/politics* and *r/Conservative*—used the protests to criticize their opponents and further their own agendas, rather than focusing on the protests themselves. Finally, *r/worldnews* showed the most neutral presentation, focusing on facts and news articles rather than emotional content, with news discussing the international response and the events of Cuba itself.

7.6.2 Extraction and Comparison Methods

Unlike previous narrative extraction methods [186, 319, 323, 369, 399, 413], our approach incorporates the concept of *community acceptance* into the optimization problem. Specifi-

cally, we leverage submission scores and upvote ratios in Reddit communities to compute an acceptance score.

Our comparison metric provides us with a way to quickly compare narratives in a holistic manner, considering their underlying structure, order, and content. For example, the metric helped us to identify similarities between politically opposed communities, such as *r/Conservative* and *r/GenZedong*, despite their superficially different content. However, the metric is not capable of providing a nuanced distinction regarding how these elements are contributing to the differences between two narratives. Thus, specialized metrics would be required to study these components.

7.6.3 Implications

Our work could be applied by social scientists interested in a narrative analysis of how social movements are evolving, as the accepted narratives provide a window into what different communities are thinking and their views. Our work could also be used in the context of computational journalism [79], which provides tools for various computational needs in journalism, such as topic detection, sensemaking, and visualization. Additionally, we could include other elements into our representation. For example, narrative maps could be part of a narrative exploration suite, which could include maps in addition to text summaries, data plots, and references to trustworthy sources. Such a tool could aid analysts, journalists, and other users in trying to understand the narratives espoused by online communities.

Furthermore, our community acceptance approach could be enhanced with additional attributes. In particular, we could model the credibility and political bias of Reddit submissions (e.g., by analyzing the article and source they are linking). These attributes could aid in modeling how misinformation spreads during social movements. Another possible extension would be using sentiment analysis techniques [291] in conjunction with our narrative model to obtain more nuanced narratives that explicitly model how communities feel with respect to specific storylines and events. Finally, we note that our methods are grounded in theoretical constructs of narratology [186] and could be applied beyond social movement narratives.

7.7 Chapter Conclusions

This chapter presents an algorithmic approach to extract the accepted narratives of online communities in the context of social movements, specifically the 2021 Cuban Protests during its first few days. Building upon existing narrative extraction methods based on linear optimization, we designed a novel approach that integrates the concept of *community acceptance* with pre-existing definitions of coherence and coverage. Our technique allows us to identify the major themes surrounding the protests in five different online communities.

Furthermore, we developed a narrative comparison metric using graph edit distance with costs based on the similarity of the events and connections to measure the differences between narratives. Moreover, our work opens up multiple avenues for future research, such as adapting the optimization method to find rejected narratives instead of accepted ones, improving the graph edit distance metric with different costs, and adapting this approach to other social networks.

Our work is not without limitations. The current embedding model only considers the text of a submission, disregarding images and videos. Thus, future work could consider multimedia embeddings to model these additional elements. Moreover, we note that our data spans a relatively small period of time and that the full effects and repercussions of the protests are still not completely clear at the time of this writing. Thus, the analysis is not able to fully capture all the narratives surrounding this social movement. However, it is able to provide an overview of the narratives at the start of the protests.

Chapter 8

Semantic Interaction for Narrative Maps

Narrative sensemaking is an essential part of understanding sequential data. Narrative maps are a visual representation model that can assist analysts to understand narratives. In this work, we present a semantic interaction (SI) framework for narrative maps that can support analysts through their sensemaking process. In contrast to traditional SI systems which rely on dimensionality reduction and work on a projection space, our approach has an additional abstraction layer—the structure space—that builds upon the projection space and encodes the narrative in a discrete structure. This extra layer introduces additional challenges that must be addressed when integrating SI with the narrative extraction pipeline. We address these challenges by presenting the general concept of Mixed Multi-Model Semantic Interaction (3MSI)—an SI pipeline, where the highest-level model corresponds to an abstract discrete structure and the lower-level models are continuous. To evaluate the performance of our 3MSI models for narrative maps, we present a quantitative simulation-based evaluation and a qualitative evaluation with case studies and expert feedback. We find that our SI system can model the analysts’ intent and support incremental formalism for narrative maps.

8.1 Introduction

Narratives are systems of stories [145]—sequences of events connected in a coherent manner. Narratives are fundamental to the sensemaking process and our understanding of the world [192], as humans use them as a natural way to capture relationships between sequences of events, alongside the goals, motivations, and plans of actors [119]. Narrative sensemaking tasks range from intelligence analysis [107], where analysts try to find hidden or implicit connections between events, to journalistic analysis of news narratives, where analysts seek to understand the big picture across many articles [53].

To help analysts in sensemaking tasks, scholars have developed visual analytics tools, which aid analysts in processing and understanding greater quantities of data [81]. These tools usually focus on different parts of the sensemaking process [282]. For example, some tools focus on the synthesis loop [388] to help analysts generate hypotheses. Others focus on

the foraging loop [178], where the goal is to gather and select appropriate data for further analysis.

In this chapter, we focus on the synthesis loop of the sensemaking process. In particular, we deal with extracting narratives from large sets of documents describing events (e.g., news articles) in the form of a graph structure—a *narrative map*—that represents the system of storylines [186]. These composite structures can then be used to aid analysts in narrative sensemaking tasks by providing a higher-level view and explicit connections based on chronology [192] that traditional visualization systems, such as those based on dimensionality reduction (DR) and clustering of documents, would not well capture.

Recent work has sought to develop computational models to assist in the narrative sensemaking process [192]. However, current approaches are static and lack refinement based on user- or task-specific goals beyond basic interactions such as searching or emphasizing specific keywords [18, 59, 186, 232, 319, 343]. More specifically, we design and evaluate an interactive narrative sensemaking tool that integrates previous work on narrative extraction and representation [186, 192, 193] with the semantic interaction (SI) framework—a framework for visual analytics that enables steering models by inferring data characteristics that are of interest to the user based on their interactions [104, 155, 317]. Our tool’s purpose is to help analysts in narrative sensemaking tasks by building a better narrative model through *incremental formalism* [329]—the ability to learn incrementally over multiple iterations to produce better models. Our main hypothesis is that the proposed SI models effectively support incremental formalism for narrative map models.

In particular, recent developments [107] have shown the capabilities of semantic interaction techniques in aiding the sensemaking process [41] through human-AI collaboration [381]. Thus, we seek to integrate semantic interaction methods with narrative maps in order to create an interactive AI narrative sensemaking framework that is capable of learning from the analyst. These techniques could prove useful to an expert user who needs to extract and understand evolving narratives, such as intelligence analysts trying to uncover the underlying connections between documents or journalists attempting to understand rapidly evolving news narratives [192].

To date, no previous research has sought to integrate semantic interaction with narrative extraction and visualization. Thus, our research provides a first step towards creating a human-AI interaction system that aids in the narrative sensemaking process. Furthermore, our overarching research goal is to explore *how to design an SI model for narratives*. This requires dealing with the issue of how to modify the machine learning pipeline of narrative extraction to support SI (i.e., how do we model user intent in our narrative extraction model?).

As a specific example of how semantic interaction techniques could aid in the narrative sensemaking process, consider an analyst seeking to understand the causes and effects of a newsworthy event (e.g., the COVID-19 pandemic or social movements) from news data. The analysts would generate a visual representation of the narrative from data and provide

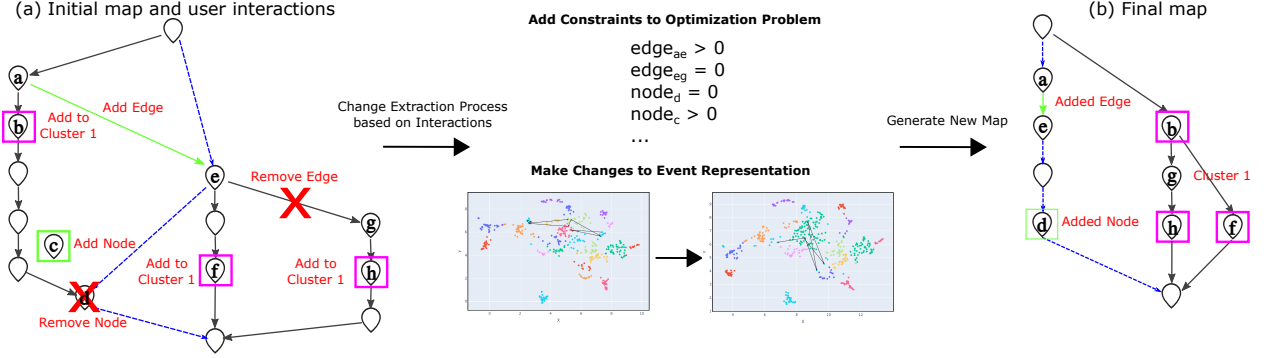


Figure 8.1: Overview of the semantic interaction model for narrative maps and its core interactions. The analyst modifies the original map on the left by: (1) adding an edge between events *a* and *e*, (2) adding event *c*, (3) removing event *d*, (4) removing the edge between events *e* and *g*, (5) grouping events *b*, *f*, and *h* by assigning them to the same cluster. After this, the system makes changes to the extraction process based on the interactions and generates a new map that integrates the analyst’s interactions into the model.

their feedback through interactions (see Figure 8.1), leading to incremental improvements as they work through their sensemaking process (e.g., removing biases, changing the focus of the narrative, or including specific storylines). We show a concrete example in Figure 8.11 as part of a case study.

Furthermore, we note that traditional SI systems usually work directly on projection spaces generated by DR methods [52], leveraging distance changes induced by the user on the elements of the projection space to improve the underlying model. However, our proposed SI model has to deal with an additional challenge in the form of an intermediate abstraction layer: the structure space (i.e., the space of possible narrative graphs). Unlike the projection space, distance is not meaningful in the structure space. Instead, relationships are determined by the edges of the narrative graph. Thus, instead of continuous distance-based interactions, the natural interactions that arise in the structure space are discrete (e.g., removing a specific connection from the graph).

In this context, we need to design a semantic interaction model that interacts with the discrete nature of the higher-level structure layer in a meaningful way. To solve this problem, we propose a general framework that builds upon the Multi-Model Semantic Interaction (MSI) concept of Bradel et al. [52] and addresses the challenges of developing SI pipelines with a mix of higher-level discrete structure space and low-level continuous models. Figure 8.1 shows an overview of the supported semantic interactions and the changes they induce on the extraction model.

In summary, the core contributions of this work are as follows:

- The concept of the **Mixed Multi-Model Semantic Interaction (3MSI)** pipeline—an SI pipeline comprised of a high-level discrete structure and a lower-level continuous

model that helps build the structure.

- An **SI Model for Narrative Extraction** that handles the intermediate abstraction layer defined by the narrative structure space and showcases the challenges of 3MSI.
- An **Evaluation** of the SI model to show whether it supports incremental formalism in the sensemaking process using a quantitative simulation-based approach and a qualitative approach with expert feedback.

8.2 Background and Related Work

8.2.1 Narrative Extraction and Visualization

Narratives are defined as systems of stories interrelated with coherent themes [145]. The same story can be told in countless ways, leading to a distinction between the underlying story itself and its representation. Narratological studies seek to understand the relation between stories and their representations [1, 285].

Most computational narrative representation and extraction methods rely on event-based models [186, 232, 319]. Events are the fundamental unit of narratives [192] and they represent actions involving entities and characters or happenings without casually involved entities [1]. However, event-based representations are not the only approach. For example, some works use a topic-level analysis, abstracting the narrative away from specific events [198, 265, 412]; some scholars propose even more fine-grained approaches, such as claims [334] and named entities [114].

Furthermore, there are three general structures for narrative representations [192]: timelines [320, 369], trees [18, 232], and directed acyclic graphs (DAGs) [186, 323, 413]. Of these structures, DAGs provide the most flexibility by allowing the representation of divergent and convergent story substructures [186].

Independent of the representational structure of the narrative model, extraction methods usually rely on optimizing a notion of narrative quality to generate the narrative representation [192]. There are multiple optimization criteria, such as topical cohesion [369] (measuring whether two events share the same topic), coherence [319] (measuring whether it makes sense to join two events together), and coverage [321] (measuring whether the narrative is properly covering the events). In this work, we use an extraction approach that maximizes coherence subject to coverage constraints [186].

In the context of visual analytics and information visualization, narratives and storytelling are common techniques used to present data [297, 316, 348]. In particular, presenting visualization as stories can be used to aid in the sensemaking process [170, 171]. Visual storytelling systems can help users detect relationships, structures, and other patterns, which can help

in confirming hypotheses or gaining additional knowledge [8, 348]. In this regard, constructing and interacting with visual narratives could be interpreted through the lens of a visual knowledge generation process [307], as users generate new knowledge as they work through their hypotheses and insights with the visual narrative model.

In this work, we develop an interactive visualization model of narratives based on the narrative map as defined by Keith and Mitra [186]—a DAG describing the connections between events (see Figure 8.11 for examples). The original narrative extraction method for narrative maps is grounded in narrative theory and previous work has found that it provides analysts with a useful narrative representation for sensemaking tasks [186, 193]. However, subsequent work showed that this narrative representation could be improved by following a series of design guidelines [192] based on an empirical study of how analysts create and use narrative maps to solve sensemaking tasks. Thus, we implement an improved version of the extraction method following these design guidelines to support our semantic interaction model. However, we note that the core contribution of our work is the semantic interaction model, rather than the incremental improvements to the narrative extraction process.

8.2.2 Semantic Interaction

Semantic interactions [105] exploit the natural interactions within a visualization—usually a projection of the data in a lower dimension space—to learn the intent of the analyst. In particular, instead of trying to manually modify parameters to model a specific concept, semantic interactions learn the parameters associated with that concept from user interactions and changes in the visualization. Semantic interactions have seen applications in text analytics [104], images [155], and other quantitative data with high dimensionality [317].

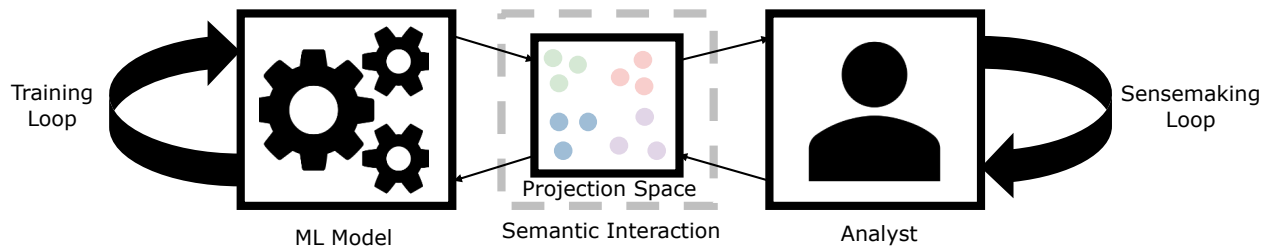


Figure 8.2: The traditional semantic interaction pipeline has the analyst interact with a dimensionality reduction model by making changes to the projection space. These changes are captured by the machine learning model and used to generate a new projection space. The process repeats as the analyst works through their sensemaking process.

Figure 8.2 shows semantic interaction as a bidirectional pipeline [97] based on an interactive DR model [52], the most common approach used when implementing SI systems. In this pipeline, analysts use the visualized projection to get insights, then they make changes to this projection by changing the position of data points. By making these changes, analysts are

expressing their preferences. Thus, based on these changes, the interactive DR model is able to learn the intent of the analyst behind the changes in the projection. Using this information, the changes are converted back into a high-dimensional space. Afterward, the modified DR model updates the projection based on the new high-dimensional representations. Note that while this example is focused on DR, the underlying approach is not limited to just DR models and it can be generalized to other methods, such as force-directed graph layouts [52, 104].

Regardless of the implementation of the pipeline, it is necessary to capture the changes from the visualization and turn them into changes to the model. There are multiple machine learning models that attempt to solve the bidirectional transforms required to implement SI, such as Observation-Level Interaction [103], Bayesian Visual Analytics [162], and Visual to Parametric Interaction [217]. Furthermore, the semantic interaction pipeline can be extended to leverage multiple models by chaining them and providing functionality and interactions at each level through the concept of Multi-Model Semantic Interaction [52]. In general, any number of models can be used in this pipeline. Then, the interaction feedback from the user is interpreted as a change into one (or many) of the models using an appropriate inverse. These works show how DR models can capture the different interactions and modifications made by analysts.

However, unlike the previous examples, narratives have an underlying temporal and causal structure [1], and in particular, graph-based representations of narratives have a discrete structure that has no natural continuous notion of distance that can be leveraged to define the necessary inverse transformation like in DR. Therefore, SI models for narrative or story visualizations must account for the underlying temporal structure and the discrete nature of the narrative representation, introducing an additional layer of complexity to the development of an SI model for narrative sensemaking. To address these issues, we develop the concept of 3MSI, which accounts for the usage of mixed models (i.e., a high-level discrete structure and a low-level continuous space) in the multi-level pipeline. Finally, to date, no previous research has sought to use SI in a narrative or story visualization setting. Thus, this research would provide a first step towards creating a human-AI interaction system that aids in the narrative sensemaking process by allowing the analysts to manipulate a narrative structure.

8.3 Narrative Map Extraction

In this section, we present our narrative extraction pipeline, which builds upon the extraction algorithm for narrative maps proposed by Keith and Mitra [186]. We further introduce optimizations to reduce computational costs and post-processing to align our results with the design guidelines for narrative maps [192]. However, we note that the core contribution of our work is building an SI model for narrative maps, rather than the extraction process itself. We show the narrative extraction pipeline in Figure 8.3. There are 2 phases in this

pipeline: extraction and post-processing.

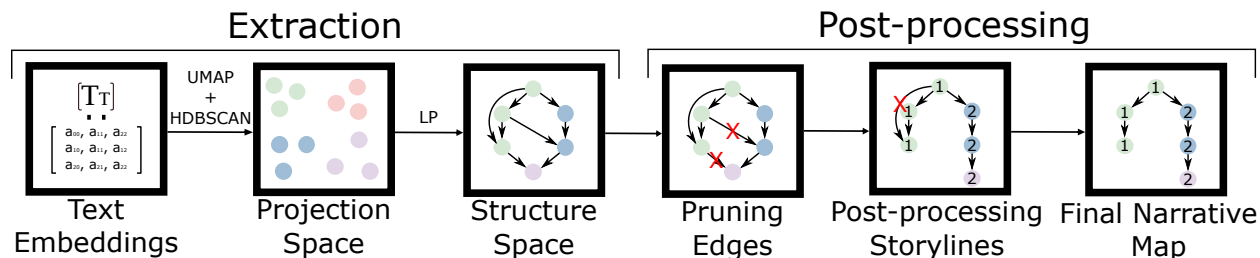


Figure 8.3: Narrative maps extraction pipeline. The pipeline consists of two main phases (extraction and post-processing).

The extraction phase builds upon the original extraction algorithm proposed by Keith and Mitra [186], using UMAP to generate a projection space [244], HDBSCAN to compute topical clusters [243], and linear programming to find the optimal narrative map. The narrative map corresponds to the structure space of our SI model.

Next, the post-processing phase takes the basic map and simplifies it, following the design guidelines defined by Keith et al. [192]. In particular, the post-processing phase seeks to decrease the overall complexity of the base map in order to reduce the cognitive load of the users. Specifically, post-processing involves removing low coherence edges, transitive connections inside each storyline, and redundant inter-story connections.

We note that there are two key steps in the extraction phase: the coherence computation step—which provides a quantitative measure of how much sense it makes to connect specific events—and the linear program (LP) formulation itself—which seeks to find the optimal narrative structure. We focus on these two steps to develop our SI model.

The coherence computation step relies on DR and clustering. We use the UMAP [244] and HDBSCAN [243] algorithms, respectively. The coherence computation step is highly dependent on the underlying projection space. Thus, our SI model intervenes at the projection space level to induce changes in the coherence values. The LP formulation step handles structural constraints and integrates the coherence values from the previous step. Furthermore, the LP formulation can be modified to handle constraints induced by the user interactions directly.

We present the LP used for extraction in Figure 8.4, which seeks to maximize the coherence of the map. We note that compared to the original formulation [186], our approach allows multiple endings and removes unnecessary constraints to reduce computational costs while maintaining the general structure.

Finally, we note that the interactive and iterative nature of SI can lead to overfitting issues [74] as the model attempts to satisfy the incremental requirements imposed by the user through each iteration of semantic interactions. Furthermore, each interaction performed by the user only affects a small subset of the data, which can lead to further overfitting issues,

$$\begin{aligned}
\max \quad & \text{minedge} - \lambda \cdot \sum_{i,j} \text{edge}_{i,j} \\
\text{s.t.} \quad & \text{minedge} \leq 1 - \text{edge}_{i,j} + \text{coherence}_{i,j} \quad \forall_{i,j} && [\text{Objective}] \\
& \text{node}_s = 1 && [\text{Fixed Start}] \\
& \text{node}_i = 0 \quad \forall_{i < s} && [\text{No Previous Starts}] \\
& \text{edge}_{i,j} = 0 \quad \forall_{i \geq j} && [\text{Chronological Order}] \\
& \sum_i \text{node}_i = K && [\text{Map Size}] \\
& \sum_{i < j} \text{edge}_{i,j} = \text{node}_j \quad \forall_{j \neq s} && [\text{Incoming Edges}] \\
& \sum_{j > i} \text{edge}_{i,j} \leq \text{node}_i \quad \forall_i && [\text{Outgoing Edges}] \\
& \text{mincover} \leq \frac{1}{|\mathcal{C}|} \cdot \sum_{k \in \mathcal{C}} \sum_{i,j} \text{edge}_{i,j} \cdot \text{membership}_{i,j,k} && [\text{Coverage}]
\end{aligned}$$

Figure 8.4: Linear program used to obtain the optimal narrative map (i.e., the basic structure space).

similar to how few-shot learning methods are prone to overfitting issues in general [372]. Thus, we include a regularization term that seeks to minimize the sum of all edge weights (i.e., L_1 regularization) and produce a sparse and less complex solution [95]. We discuss regularization and how we handled it after defining our simulation-based evaluation. We now provide more details on the narrative maps extraction pipeline, including the coherence computation step, the linear programming formulation, and the post-processing steps.

8.3.1 Coherence Computation

The coherence computation step relies on the projection space and the associated topical clusters [186]. Thus, it depends on the dimensionality reduction and clustering algorithms used to build this space. In particular, coherence is computed as the geometric mean between content similarity and topical similarity. The content similarity term is computed based on the normalized cosine similarity of the projected document embeddings, which are based on the **all-MiniLM-L6-v2** model of the **sentence-transformers** library [293]. The topical similarity term is computed using the Jensen-Shannon similarity [142] to compare the distribution of the topical cluster probabilities of each document.

However, this definition of coherence leads to issues when dealing with data sets that span longer periods of time, such as generating long-ranging connections between events that are seemingly related but too temporally distant to be relevant. To deal with this, previous approaches have used exponential decay factors based on the temporal distance between

events [60, 232, 233, 407, 416]. Thus, we extend the previous notion of coherence with a temporal component that penalizes events that are temporally distant using an exponential factor. Let $\Delta t_{i,j}$ be the temporal distance between events i and j . The new value for coherence is defined as

$$\text{coherence}_{i,j} = \exp\left(-\frac{\Delta t_{i,j}}{\sigma_t}\right) \cdot \sqrt{\text{CS}_{i,j} \cdot \text{TS}_{i,j}}.$$

where σ_t represents the temporal sensitivity, a user-defined parameter for the extraction algorithm that regulates the rate of decay, $\text{CS}_{i,j}$ represents content-based similarity of the events and $\text{TS}_{i,j}$ represents topical-based similarity of the events. For the purposes of our work, we measure all temporal distances and σ_t in days and we use a default value of 30. However, for other data sets and contexts, it could be necessary to change the units of temporal measurement.

8.3.2 Linear Program Formulation

Now we provide more details about the components of the linear program formulation (see Figure 8.4).

Variables and Parameters. The map is built by finding the optimal weights for its nodes and edges. These weights are represented by the node_i and $\text{edge}_{i,j}$, respectively, where i and j represent the indices of the corresponding events in the data set. These variables take values between 0 and 1. There are three user-given parameters: K , the expected main story length that regulates the size of the map; mincover , the minimum average coverage for the topical clusters; and λ , the strength of the regularization term in the objective function. There are also pre-defined parameters that depend on the results of the projection space. Namely, the coherence values of each connection $\text{coherence}_{i,j}$, and the membership of an edge to a given topical cluster $\text{membership}_{i,j,k}$ [186]. In this context, C represents the set of topical clusters (indexed by k). Finally, the minedge variable represents the minimum coherence value of the narrative map over the edges and is used as the optimization goal.

Objective Function and Constraints. Following the model of Keith and Mitra [186], the linear program seeks to maximize the value of the minimum edge, under the assumption that the best narrative structure should have no weak connections (i.e., the minimum coherence of the map is as high as possible). This is done using the minedge constraints that depend on the coherence values.

In our evaluations, we fixed the regularization strength parameter (λ) to the inverse of the total number of potential edges in a directed acyclic graph of size n . That is, we set $\lambda = \frac{2}{n(n-1)}$. Our empirical evaluations showed that this value worked well for our purposes (see analysis of regularization in the Results section).

Starting Event Constraints. We note that the aforementioned design guidelines study showed that analysts tend to use maps with a single source event and potentially multiple

endings [192]. Thus, we allow users of our extraction algorithm to select a single starting event s when extracting the map. This is handled by the $\text{node}_s = 1$ which sets the corresponding starting node as active. Furthermore, to avoid interfering with the user-defined starting event, we prevent events that occur before the starting event from appearing on the map, which is handled by the constraint that assigns their node variables to 0. We note that it would be possible to adapt this approach to a fixed ending event, or other combinations of starting and ending events, including none of them.

Chronological Order. To ensure that the events in the map follow a chronological order, we impose a series of constraints based on the index of the events. This constraint assumes that the data set has been sorted by time before running the extraction algorithm.

Map Size Constraint. This constraint determines the expected length of the main storyline on the map and thus it regulates the map size. This constraint is determined by the parameter K , which is defined by the user. We note that the original extraction model [186] had another constraint to ensure that the sum of edge weights added up to $K - 1$. However, we removed this constraint and replaced it with the regularization term that seeks to minimize this sum.

Edge Constraints. These constraints seek to generate the connection structures of the map, relating the incoming and outgoing edges of each node with the activation value of that node. Unlike the original formulation by Keith and Mitra [186], we use inequality constraints for the outgoing edges, as this approach allows for maps with multiple endings.

Coverage. To ensure appropriate topical diversity and coverage over the narrative map, the original formulation of the extraction algorithm also included a coverage constraint based on the average coverage of each topical cluster (obtained from the HDBSCAN clusters in the projection space). A topical cluster is considered to be covered if sufficient edges belonging to the cluster (defined by the $\text{membership}_{i,j,k}$ parameter) are on the map. The minimum coverage required (mincover) is a user-defined parameter. We set 20% minimum coverage as the default value of this parameter.

8.3.3 Post-processing

We note that despite the use of regularization, the optimal map in terms of coherence could still be highly complex and have redundant edges or connections that make it difficult for a human to interpret adequately. Thus, as described in the extraction pipeline, we perform a series of post-processing on the optimal map for usability purposes and to keep in line with the design guidelines defined by Keith et al. [192].

Pruning Edges. In particular, after extracting the initial narrative map, we pruned edges to reduce the cognitive load for users. In particular, we only keep a certain number of edges per node, thus, restricting the branching factor of the DAG. Furthermore, after the initial pruning, we remove any leftover edges that have too low of a coherence value in the final

model.

More specifically, we only leave the top $\lceil \sqrt{K} \rceil$ outgoing edges of each node and remove the rest. Next, we removed all edges that had less than $0.1/K$ coherence after re-normalizing edge weights. In both formulas, K is the map size parameter of the linear program described in the previous subsection. These heuristics were found through empirical testing. We note that the cut-off values depend on K because a fixed cut-off value became problematic for larger maps, as it leads to too many edges being filtered. These heuristics helped prevent most cases of excessive edge density without eliminating potentially relevant connections.

Storyline Identification. Next, we also identified the different storylines of the narrative map. This is a necessary step to remove the transitive connections from storylines (i.e., redundant connections) and to minimize inter-story connections (i.e., only keep the first and last of such connections for each pair of stories). Furthermore, identifying these storylines is used for evaluation purposes (e.g., analyzing a specific storyline of a map). However, the original narrative map extraction method does not directly generate a partition of the map into storylines, it only highlights the main storyline [186]. Thus, we need to develop an algorithm to partition the map into storylines.

To do this, we use the recursive algorithm described in Algorithm 1. This algorithm takes a narrative map as an input (i.e., a weighted directed acyclic graph), with potentially many sources (starting events) and sinks (ending events). This algorithm is based on the idea that we can recursively extract the storylines by finding the most coherent paths of the graph in each step. The first storyline is easy to identify, as that would simply be the main storyline, which we identify by finding the maximum likelihood chain of the graph based on coherence. To find the rest of the storylines, we first remove the events from the main storyline and all the corresponding connections. Then, we find the new “main” storyline using whatever events and connections remain in the graph. We recursively perform this process and store all the extracted storylines until only storylines of size 1 remain (i.e., a graph of disconnected single events). As we remove storylines, we might end up separating the graph into multiple connected components. Intuitively, these components correspond to different parts of the narrative (e.g., topics) and should contain different storylines.

Transitive Reduction. We note that the original implementation of the narrative map extraction algorithm had an additional constraint to reduce redundant transitive connections [186]. However, this approach introduced a cubic number of constraints in the worst case, greatly increasing the computational cost. Thus, removing this constraint from the optimization problem and performing post-processing to remove transitive connections reduces the overall computational cost of the extraction process, allowing a more responsive system.

In particular, for each identified storyline in the graph, we use transitive reduction [7] on the corresponding storyline sub-graph making sure to keep the edges with higher coherence values. Afterward, we re-normalize the edges and obtain the final storyline. Note that this process does not affect inter-story connections.

Input: Narrative map $G = (V, E)$ - Weighted directed acyclic graph with weights between 0 and 1.

Output: List of sequences of events (storylines).

Function `GraphStories( $G$ ):`

```

    if  $|E| == 0$  then
        return List(Singletons( $V$ ))          // Base Case - graph with no edges.
    end
     $sp = \text{ShortestPath}(G)$                 // Shortest Path (Negative Log Likelihood).
     $H = G - sp$                              // Delete Nodes in  $sp$  (and edges) from graph  $G$ .
     $H = \text{Normalize}(H)$                      // Normalize outgoing edge weights (sum = 1).
    return List( $sp$ ) + GraphStories( $H$ )       // + means list concatenation.

```

Algorithm 1: Recursive algorithm to extract all the storylines based on their overall coherence.

Inter-story Connections. For inter-story connections, we only keep the first and last of such connections for each pair of storylines on the map. To do this, we find the boundary edges between the sub-graphs defined by each storyline. Then, we keep the first identified connection and the last identified connection, based on the indices of the corresponding nodes. After removing all redundant inter-story connections, we re-normalize the edges of the graph.

8.4 Semantic Interaction Model

In this section, we present our general 3MSI concept and how to address the challenges of using a mix of continuous spaces and discrete structures in the SI pipeline. Then, we present the SI model for narrative maps.

8.4.1 Mixed Multi-Model Semantic Interaction

The concept of 3MSI deals with a specific type of MSI where the top-level model of the SI pipeline—which is associated with the visualization shown to the user—is represented by a discrete structure space. Furthermore, the model must not have a continuous notion of distance that can be leveraged to model user interactions. For example, a graph represented with a force-directed layout can use a continuous notion of distance to model interactions, but a graph represented with a hierarchical layout would not be able to directly translate display distances into model changes (e.g., narrative maps).

The general 3MSI pipeline is shown in Figure 8.5. In this pipeline, the lower-level model uses

a continuous representation and the high-level model corresponds to a discrete structure. This model extends the traditional MSI approach [52]—which only considers continuous spaces as internal models—with an additional layer of abstraction in the form of a discrete structure space, which provides users with further scaffolding to perform sensemaking tasks.

Once the user perceives the visualization associated with the high-level model, they can interact with it. However, the introduction of a discrete structure in the 3MSI framework makes capturing user interactions and feedback for semantic interaction purposes more difficult compared to the traditional pipeline. The interaction feedback needs to be interpreted and fed to the inverse models in an appropriate manner. Defining how to interpret and feed the interactions back to the inverse models is the key step in using the 3MSI framework. Once the feedback has been captured by the models, the updated parameters are stored and used alongside the original data to update the visualization. This process repeats as the user keeps interacting with the updated models.

In the context of our work, the forward models are defined by the narrative extraction pipeline, with the low-level model corresponding to the UMAP projection and the higher-level model corresponding to the narrative structure itself. The inverse models depend on the specific type of interactions that we define and can rely on mathematically rigorous, heuristic, or probabilistic approaches [97]. In our implementation of 3MSI for narrative maps, we consider heuristic approaches, rather than defining formal mathematical inverses. In particular, we rely on adding specific constraints to the LP definition to modify the structure space and using semi-supervised learning to modify the projection space.

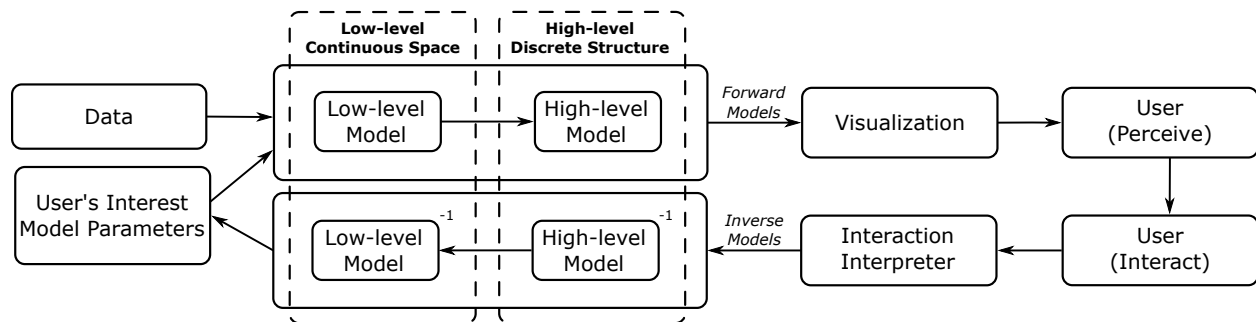


Figure 8.5: Generalized Mixed Multi-Model Semantic Interaction (3MSI) visualization pipeline.

As mentioned before, interpreting the interactions and feeding them into the inverse models is the key step in the 3MSI pipeline. In particular, there are two key challenges that arise at this point. These two challenges are further amplified by the mixed nature of the pipeline:

- The *placement challenge*—how do we choose the levels at which we will handle each one of the relevant semantic interactions?
- The *transformation challenge*—how do we transform changes from the higher-level

discrete structure (e.g., the structure space) back to the low-level continuous space (e.g., the projection space)?

We note that these challenges are not necessarily unique to the mixed model context, but the discrete nature of the high-level model induces extra difficulties (e.g., defining an inverse function between the discrete structure and the low-level model). Furthermore, the solutions to these challenges depend on the specific architecture of the system. In this context, we propose general guidelines to aid in solving these challenges. In particular, we focus on providing ways to solve the placement challenge and circumventing the transformation challenge when possible, as this is the more complex challenge due to the difficulty of defining proper inverse transformations for these interactions, which can rely on mathematically rigorous, heuristic, or probabilistic approaches [97].

The proposed guidelines follow the general design principle of *separation of concerns* and apply a *divide-and-conquer* strategy when required for more complex interactions, adapting the general tenets of algorithmic and software design [249] to the context of our SI framework. More specifically, the goal is to minimize the need of dealing with multiple levels at the same time—avoiding the transformation challenge. Assuming that the relevant semantic interactions have been identified, we present the general guidelines and specific examples pertaining to our narrative extraction case.

Handling Single-level Interactions: Separation of Concerns As our core design guideline is to avoid the transformation challenge and reduce the implementation complexity of the 3MSI pipeline, interactions that have an intuitive and direct effect on a level should be implemented exclusively at that level and avoid changing the other level. This is the simplest case to implement for a semantic interaction, as it only requires interpreting the changes on a single level. For example, removing an event from a narrative map has a direct effect on the structure space (i.e., the model must create a representation that avoids that event). Thus, we implement this by only inducing changes to the structure space, without changing the underlying projection space, avoiding the transformation challenge.

Furthermore, interactions that can be solved at different levels in equally valid ways should be solved at the highest level possible (i.e., closest to the visualization and the user). As the model moves away from the visualization and deeper into the semantic interaction levels, each step makes the resulting changes potentially more opaque to the user. Thus, we recommend choosing the highest level possible, as this should provide a solution that is closer to the changes expected by the user, which would likely be biased by the visualization. In particular, this would be particularly relevant in cases where there are more internal lower-level models. In the narrative extraction context, removing an edge from a narrative map could be solved at the structure space level by simply imposing a specific constraint to avoid this edge in subsequent solutions or at the projection space level by forcing the events of that edge to be sufficiently distant to reduce the likelihood of being connected. However, following the

guideline of using the highest level possible, we should choose the structure space level, as it is closer to the user and thus more likely to produce the intended effect.

Handling Multi-level Interactions: Divide and Conquer Interactions that are more abstract and complex and do not have a direct effect on a single level should be separated into their constituent effects and each one of these should be implemented separately at a different level as appropriate. By separating the interaction into simpler components, we address both the placement challenge and the transformation challenge. Instead of relying on an explicit inverse function to get from the discrete structure space back to the continuous space or vice versa—which might not even be properly defined—we directly induce changes at each separate level following heuristics. This strategy can also be easily scaled to pipelines with many internal lower-level models.

For example, in contrast to the remove event example, grouping events is a more abstract and indirect interaction, as it does not specify how the events should be connected in the structure space. Instead, it simply implies that the events should be “close” and that the events should be somehow connected. Thus, we implement this interaction at both levels, inducing a change in the projection space (closeness) and adding constraints to the structure space (connectedness). We note that if it is not possible to adequately divide the semantic interaction by level, then we would have to deal with finding an appropriate inverse function instead of simpler heuristics.

8.4.2 Semantic Interaction for Narrative Maps

Pipeline Overview

Following the 3MSI concept, our approach has an additional abstraction layer compared to traditional SI systems: the structure space. This space builds upon the projection space, as it uses the projected representations to compute a discrete structure (the narrative map) that has no natural notion of continuous distance. The structure space captures higher-level relationships that might not be apparent in the projection space.

In contrast to the traditional pipeline of Figure 8.2, the proposed pipeline has an additional level of abstraction: the structure space, which builds upon the projection space. The analyst interacts directly with the structure space (i.e., the narrative map), rather than with the continuous projection space. The changes in the structure space are used to update the underlying models, either by defining constraints that affect the narrative extraction process or by inducing changes into the projection space through semi-supervised learning. Thus, the semantic interactions with the narrative map can lead to either direct changes to the structure space or indirect changes to the projection space.

Design Goals for SI in Narrative Maps

Following previous work [192] and preliminary evaluations with users, there are two general ways in which narrative maps are used: **exploratory analysis**—where users are trying to find the different storylines in the data and understand the big picture—and **directed analysis**—where users seek to understand specific connections between events and stories. Thus, we seek to design an interactive model to support these analyses. In particular, there are three core subtasks that we support:

Correcting Storylines: Analysts need to be able to make corrections to existing storylines. This includes modifying intra-story connections (e.g., changing a specific edge, removing an event from the storyline) and inter-story connections (e.g., adding a connection with a different storyline to highlight common events or entities). This is mostly related to the directed analysis task, as users are refining the narrative schema.

Creating Storylines: Analysts need to be able to create new storylines and integrate them into the narrative map. For example, analysts might want to add new events or specify which events should be grouped into a coherent and consistent storyline. This is mostly an exploratory analysis task, as users are uncovering the different stories in the data.

Shifting Focus: Analysts need to be able to change the focus of the narrative map (e.g., changing the main storyline or the overall contents of the data). This interaction is relevant to both tasks, as users perform work through the sensemaking process they might require changing the focus of the narrative.

Relevant Interactions

We capture five relevant user interactions in our model. Examples of these interactions and their effects are shown in Figure 8.1. We note that all the user interactions are done at the structure level, but they might imply a change in the underlying projection space too. Following the 3MSI framework, we need to identify the level at which these interactions must be handled and if necessary divide them into their constituent effects. We briefly describe the interactions and how they influence the underlying projection and structure space.

Add/Remove Event: Adding an event implies that the analyst would prefer for this specific event to appear in all subsequently extracted narrative maps. Likewise, removing an event implies it should not appear in any of the subsequently extracted narrative maps. In our implementation, these interactions only affect the model at the structure space level, without changing the underlying projection space.

Add/Remove Connection: Adding a connection between two event nodes implies that the analyst would prefer for this specific connection to appear in all subsequently extracted narrative maps. Likewise, removing a connection implies that it should not appear. These interactions only affect the structure space.

Group Events (Clustering): This interaction implies that the analyst would like for these events to appear in the same storyline or be close to each other in the narrative map (i.e., in the same route or at least connected in some way). As this interaction is more abstract and there is no clear definition of how a cluster should be handled in the structure space, we will handle this interaction at both levels (projection and structure space).

8.4.3 Integrating SI into Narrative Extraction

There are two steps in the extraction pipeline where we can implement semantic interaction: the coherence computation step (i.e., the projection space) and the linear programming step (i.e., the structure space). For the coherence computation step, we can modify the DR model used to generate the projection based on user interactions (e.g., forcing two events to be close together in the projection). For the LP step, we can modify the problem formulation by adding explicit constraints that force the solution to include the feedback from the user interactions (e.g., removing an event from the map should prevent it from appearing in the solution again).

Single-level Semantic Interactions: We first deal with semantic interactions that can be handled at a single level by adding constraints directly to the LP formulation shown in Figure 8.4. This is a direct and straightforward method to model simple changes to the narrative map. We summarize all induced constraints in Figure 8.6(a). In particular, node and edge removal can be directly incorporated by adding constraints of the form $\text{node}_i = 0$ and $\text{edge}_{ij} = 0$, respectively. These constraints ensure that nodes and edges removed by the user are not included in any of the subsequently extracted narrative maps. In contrast, node and edge addition are not as direct. It is not possible to include constraints of the form $\text{node}_i > 0$ and $\text{edge}_{ij} > 0$, as linear programming does not support constraints with strict inequalities [418]. Instead, we use non-strict inequality constraints with a small positive number ε to ensure that nodes and edges added by the user are included in all subsequently extracted narrative maps. We set ε empirically, testing different threshold values we found that 0.01 for edges and 0.05 for nodes ensured that the elements were added in a mostly coherent way to the map.

We note that this approach only requires working on the structure space, as the constraints only induce changes to the extracted narrative map, rather than the lower-level projection space. Furthermore, as the user interacts with the visualization through their sensemaking process, the LP formulation will accumulate a series of induced constraints, forcing the optimization algorithm to find ways to fit the user feedback into the overall narrative structure it extracts.

Multi-level Semantic Interactions: Now, we turn our attention to the more complex interaction of grouping events. Analysts can group events by manually selecting event nodes and assigning them a cluster number. For example, in Figure 8.1, events b , f , and h are assigned to Cluster 1. Unlike the previous cases, clustering events is a more complex in-

teraction, as it not only implies that events should be connected or added directly in the structure space but also imposes the notion of these events being part of the same group (i.e., relatively similar).

Following the 3MSI guidelines, we divide the interaction into two effects: the events should be close in the projection space (closeness) and the events should be connected in the structure space (connectedness). Thus, we first use semi-supervised DR with the cluster information to keep the events close in the projection space. Next, we impose constraints on the LP formulation, seeking to keep the events connected in the structure space. This way, we induce changes in both levels of the SI pipeline without explicitly computing an inverse function between the structure space and the projection space. Instead, we rely on intuitive notions of closeness and connectedness to create appropriate heuristics.

In more detail, we incorporate these manual clusters as training information for the DR step that uses UMAP. These clusters would be created based on user-defined labels (e.g., the user could assign all nodes with a specific keyword to be part of the same cluster) and fed to the extraction pipeline. UMAP supports semi-supervised learning [244]. Thus, including the label information is direct and does not require further modifications. In particular, we create a label vector that contains the value -1 for all data points, which means that no labels have been assigned yet. Then, for each data point that belongs to a user-defined cluster, we assign the corresponding integer label to their entry in the label vector, following the format defined by the semi-supervised UMAP implementation [244]. Next, we recompute the projection space using the label vector and the original embeddings. Note that when there are no user-defined clusters, we simply use the unsupervised version of the method. Once we obtain the new projection, we use it to recompute the similarity and clustering tables for the extraction algorithm.

Regarding the structure space changes, we add constraints to the LP to ensure that the clustered events appear on the map ($\text{node}_i \geq \varepsilon$). We also impose further constraints—shown in Figure 8.6(b)—by requiring the user-created cluster to be (weakly) connected, as shown in Figure 8.6(c). As with the previous constraints, we set ε empirically. We found that 0.01 for nodes and 0.05 for the edge sum worked well in most test runs. In conjunction with the new projection space, this ensures that the resulting map properly connects the grouped events. With all these changes, we solve the linear program again to find the new optimal narrative map.

8.5 Evaluation Methodology

8.5.1 Data Set

To show the effectiveness of our SI model we use news data covering the 2021 Cuban Protests that occurred in July 2021—the biggest protests in decades in Cuba [302]—which provide a

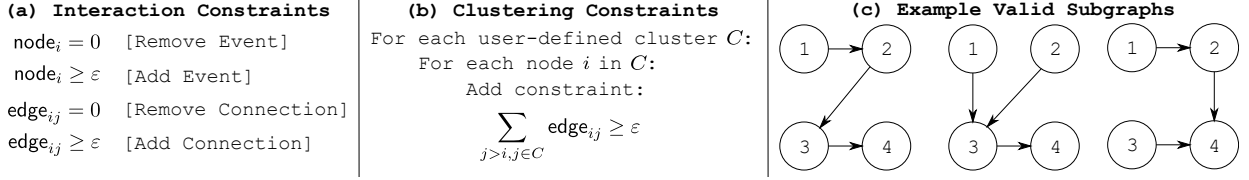


Figure 8.6: (a) Constraints induced by basic interactions (addition and removal). (b) Additional constraints to ensure that user-defined clusters are connected. (c) Examples of weakly connected subgraphs produced by the previous constraints in the narrative map.

sensemaking task of moderate difficulty for our experiments. We focus mostly on breaking news, as each news article represents a single main event [187]. We scraped a data set of 500 online news articles from 20 news sources from different sides of the political spectrum. We categorized the sources based on external bias ratings taken from [AllSides.com](https://allsides.com)¹ and the Media Bias/Fact Check database².

Throughout our evaluation tasks, we rely on keywords and political leanings as labels. Previous research has shown that it is possible to computationally distinguish political leanings from the content of a news article [28, 181]. Particularly, in times of political crises, there is a strong use of partisan content frames, which can be computationally detected [181]. Thus, we expect our SI model to be capable of learning such distinctions.

8.5.2 Simulation Tasks

The human-centered approach [48] is the primary evaluation method for SI systems [40]. However, this method is highly dependent on human feedback and makes it challenging to compare different SI systems. This is due to the inherent difficulty of replicating user interactions in SI systems, as the semantic interactions build upon each other and modify the internal model through incremental formalism [329]. To deal with these issues, simulation-based evaluations [39, 40] work by creating a simulated analyst agent that attempts to replicate human interactions. Since there are no ground truths available for analyst intent and their cognitive process in general, these approaches use keywords or other easy-to-define criteria to label the data. Thus, this ground truth is used to guide the simulated interactions and compute error rates, allowing us to evaluate the ability of these models to perform incremental inference [40].

¹allsides.com/media-bias/media-bias-ratings

²mediabiasfactcheck.com/search

Task Overview and Definitions

We define five simple tasks with basic narrative goals and easy-to-define ground truths. These tasks use a series of labels that can be easily extracted from the data to simulate the creation of user-defined target labels (i.e., relevant events, consistent connections, etc.). These user-defined labels are based on keywords or other metadata that can be easily retrieved and do not require complex computations. In practice, users would define more complex and nuanced labels in their sensemaking process. Nonetheless, simple labels—based on keywords or pre-defined classifications—are sufficient to showcase the learning capabilities of our SI model.

Furthermore, we note that the simulated analyst will only interact with a subset of the potentially relevant nodes or edges from the data set (i.e., those that appear on the map). Moreover, the model is not aware of these labels, it only knows about the embedding representation of the articles and the interaction feedback from the analyst. We expect the model to be able to *learn* these user-defined labels based on the simulated analyst interactions.

Moreover, we make the simulated analysts follow the same rules in each iteration (i.e., selecting events or edges based on a specific label). However, we note that each execution of the simulation is done using a different random seed, which leads to different narrative maps and, in turn, a different sequence of interactions in each task throughout the iterations, even if the simulated analysts always follow the same rules. We note that while it would be possible to create simulated analysts with more diverse interactions, such as interacting with random subsets of map elements or choosing random interactions, we opt to define simple rules for our simulation in order to reduce sources of variability.

We note that the evaluation metrics all measure the *error rate* of the model with respect to a specific element of interest. In particular, each task has a different metric based on its goal and the user-defined labels that we use (e.g., a specific set of relevant nodes, consistent connections, or consistently connected clusters). The tasks and their evaluation metrics are shown in Table 8.1.

Specific Task Definitions

T1 - Remove Irrelevant Events We define three sub-tasks for T1 using different user-defined labels of increasing complexity. We start with a simple label based on keywords, where events are marked as irrelevant if they contain “Florida” or “Miami” in their headline (41 events). The goal is to create a narrative map that does not cover these specific events, as they are not relevant to the protests in Cuba itself or the US response in general. Next, we move to a more complex label based on the publication source of the event, where events are marked as irrelevant if they were published by Breitbart or Fox News (155 events). Both of these news sources are right-leaning and in general present a highly biased version of the events with right-wing framing. Thus, the goal is to create a narrative map that avoids

Table 8.1: Evaluation task description and evaluation metrics.

Task	Description	Interactions	Evaluation Metric
Remove Irrelevant Events (T1)	Generate a narrative that does not include events corresponding to user-defined labels.	Remove Node	Fraction of irrelevant nodes with respect to the total number of nodes in the map.
Remove Inconsistent Connections (T2)	Generate a narrative with consistent edges according to user-defined labels.	Remove Edge	Fraction of inconsistent edges with respect to the total number of edges in the map.
Clean Up Storylines (T3)	Generate a narrative with consistent storylines according to user-defined labels.	Remove Node, Add Edge	Fraction of inconsistent nodes with respect to the total number of nodes in the map.
Cluster Events (T4)	Generate a narrative with consistent clusters according to user-defined labels.	Clustering	Fraction of relevant nodes that are not connected over the total number of relevant nodes in the map.
Add Relevant Events (T5)	Generate a narrative that incorporates additional events in a consistent manner according to user-defined labels.	Add Node, Clustering	Fraction of relevant nodes that are not connected over the total number of relevant nodes in the map.

these highly biased articles, thus providing a more neutral view of the narrative. Unlike the previous keyword-based approach, this requires the model to understand the differences between these highly biased sources and other sources. Finally, the last label is also based on the publication source of the event, where events are marked as irrelevant if they were published by any right-leaning news outlet (207 events). Generating a map that excludes this label is even harder than the previous example, as it considers highly biased sources the same as sources with only mild bias, making the distinction fuzzier.

T2 - Remove Inconsistent Connections T2 is based on our definition of inconsistent edges (connecting right-leaning to left-leaning articles and vice versa). Intuitively, minimizing inconsistent edges in a narrative map leads to a map that does not have abrupt changes in the framing and coverage perspective. Thus, the goal of removing these inconsistent edges is to create a map that avoids drastic changes in the framing and presentation of the facts. We seek to create a map that does not directly connect left-leaning and right-leaning articles. Thus, any change in the political leaning of a storyline would be mediated by an unbiased article, allowing for a slower shift in framing.

T3 - Clean Up Storylines T3 is based on our definition of consistent nodes (i.e., having the same political leaning as their storyline). The goal of cleaning up storylines is to create internally consistent storylines that share their political leaning and avoid storylines that shift their political leaning. This is similar to the goal of removing inconsistent edges, but using a mix of node and edge operations at a storyline level, rather than edge operations

on the whole map. Ideally, this would lead to a map with parallel storylines showing how each side of the political spectrum presents the events, allowing analysts to contrast these different perspectives.

T4 - Cluster Events The goal of this task is to create a narrative map that properly covers the events from the user-defined clusters, by creating a consistent presentation of these events. That is, the nodes of these clusters should be connected according to our definition. The SI model should be capable of inferring other events (modeled by the changes in the projection space) that should be in the cluster and connecting them properly, even if the user did not explicitly label them as part of the cluster. We define two sub-tasks for T4.

For the first sub-task, we define two clusters based on headline keywords. The first cluster (41 articles) is made from articles focusing on Florida and Miami. The second cluster (96 articles) is made from US-focused articles, excluding Florida and Miami. The goal is to create a narrative map that includes coverage of the general US response and the additional protests in Florida. For the second sub-task, we define two clusters based on publication source and political leaning. The first cluster (83 articles) is made exclusively from articles published by Breitbart, a highly biased news source. The second cluster (92 articles) is made from articles published by unbiased sources (AP news and Reuters). The goal in this example is to create a narrative map that includes highly-biased coverage from Breitbart and unbiased coverage. The resulting map would help an analyst compare and contrast these perspectives.

T5 - Add Relevant Events We define two sub-tasks for T5. For the first sub-task, we define relevant events as those that contain “Florida” or “Miami” in their headline (41 articles). The goal is to generate a narrative that includes events about the protests in Florida. To do this, the analyst would mark relevant events in the map as part of a single cluster and add more relevant events. These two interactions tell the narrative extraction model that the relevant events are closely related and that the narrative map should also integrate more relevant elements into the map, respectively. For the second sub-task, we define relevant events as those that contain “Biden” in their headline (92 articles). The goal is to generate a narrative map that includes events about the response of the US president to the Cuban protests.

8.5.3 Simulated Analyst Implementation

We now describe our simulation experiment and the implementation of the simulated analyst in detail.

Simulation Overview

Throughout our experiments, we test the ability of the models to learn incrementally over the course of several interactions with the simulated analyst. Due to the computational costs of executing the narrative map extraction process and the subsequent interactions and refinements, we only take 10 samples of each task. Each sample will generate slightly different narrative maps, as the underlying embedding spaces are projected and clustered with a different seed each time. Our results show that regardless of the starting conditions, after sufficient interactions the models are capable of learning based on the simulated interactions. Furthermore, we fixed the starting event for all maps to reduce variability and make comparisons easier. In particular, we chose a news article from an unbiased source (Reuters) describing the start of the protests: “Cuba sees biggest protests for decades as pandemic adds to woes”³.

In each iteration, we first compute the evaluation metric for the current narrative map using the definitions in Table 8.1. Then, the simulated analyst performs a series of interactions with the model and generates a new map. We rinse and repeat until the error rate achieves the desired target level (0%). Finally, we discarded all samples that started with a 0% error rate, as it would not make sense to use semantic interactions to improve the narrative model in those cases. We kept executing our simulation until we obtained 10 valid samples.

Finally, we note that the SI system has no awareness of the user-defined labels throughout the evaluation process. Only the simulated analyst knows these definitions and uses them to determine its actions in each task. Moreover, the analyst only interacts with a subset of the potentially relevant elements from the data set, but we expect the model to be able to *generalize* based on these interactions. For example, for T1 with the Miami/Florida label (41 events in the data set), the average number of total user interactions throughout the simulation was approximately 7 node removals. Thus, the different metrics seek to capture the model’s generalization ability.

Simulation Definitions

We provide more details on the definitions used for the task definitions and associated evaluation metrics in Table 8.1. In particular, the tasks make use of the following definitions:

- **Relevant and Irrelevant Nodes** We define event nodes as **relevant** if they are associated with a user-defined label, which can be defined by the presence of a keyword, the political leaning of the associated news outlet, or even the news outlet itself. We define **irrelevant** nodes in a similar manner.
- **Inconsistent Edges** We define edges as **inconsistent** if they connect articles from

³<https://www.reuters.com/world/americas/street-protests-break-out-cuba-2021-07-11/>

opposite political leanings. For example, articles from right-leaning outlets with articles from left-leaning outlets, or vice versa. Any other combination is considered consistent.

- **Inconsistent Nodes** We define event nodes as **inconsistent** with respect to their storyline if they conflict with the most frequent political leaning of their storyline, excluding unbiased articles. Ties are broken by selecting the first political leaning in the storyline.
- **Connected Nodes** We define event nodes as **connected** with respect to their cluster if they fulfill any of the following conditions based on general narrative structures [186]: they are directly connected to another relevant node of the same cluster, they are in the same storyline as another relevant node of the same cluster, or they share direct predecessors or successors with another relevant node of the same cluster.

Simulated Analyst Actions

We now describe the actions taken by the simulated analyst in each iteration for each task. We note that each iteration consists of multiple actions (e.g., multiple node removals or additions).

T1 - Remove Irrelevant Events. The simulated analyst removes all nodes marked as irrelevant according to the defined target labels. These interactions should eventually lead to a narrative map that avoids events with this keyword. When there are no more events to remove from the map, the error rate is zero and the simulation ends.

T2 - Remove Inconsistent Connections. The simulated analyst removes all edges marked as inconsistent according to our criteria. These interactions should eventually lead to a narrative map that avoids inconsistent connections. When there are no more connections to remove from the map, the error rate is zero and the simulation ends.

T3 - Clean Up Storylines. For each storyline in the narrative map, the simulated analyst removes all inconsistent edges. To do this, the simulated analyst first computes the political leaning of a storyline by selecting the most frequent political leaning, excluding center articles. The simulated analyst breaks ties by selecting the political leaning of the first biased article in the storyline.

Once the political leaning of a story has been chosen, all nodes inside that have the opposite leaning are marked as inconsistent. To make the storyline consistent, the simulated analyst disconnects the inconsistent nodes from the rest of the storyline by removing the inconsistent edges in the storyline. Then, the simulated analyst reconnects the storyline by adding edges between the events that are missing connections. See Figure 8.7 for an example of this interaction.

These interactions should eventually lead to a narrative map with only consistent storylines. Note that we do not delete the inconsistent nodes, as they could be re-used as part of

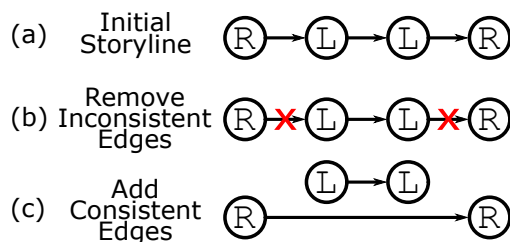


Figure 8.7: Example of the simulation steps for task T3 (Clean Up Storylines). (a) The initial storyline. (b) Removing the inconsistent edges from the storyline. (c) Reconnecting the storyline with consistent edges.

another storyline by the extraction algorithm. When there are no more inconsistent nodes in all storylines of the map, the error rate is 0% and the simulation ends.

T4 - Cluster Events. For each pre-defined cluster label, the simulated analyst marks the relevant nodes in the narrative map as part of their respective clusters. These interactions should eventually lead to a narrative map that integrates the relevant clusters as part of the map by properly connecting them in the same storyline or through common predecessors or successors. When there are no more relevant isolated nodes on the map, the error rate is zero and the simulation ends. The simulation also ends if there are no more relevant nodes to add to the cluster (note that this termination condition did not occur in practice during our experiments).

We note that our clustering task only considers disjoint clusters. The current implementation of our SI model is not able to handle the multi-label case, where a single event belongs to more than one cluster.

T5 - Add Relevant Events. The simulated analyst marks the relevant nodes present in the narrative map as part of a single cluster. Then, they add new nodes to the map by taking the first 10% (rounded up) of the relevant events that did not appear on the map from the data set.

In this task, we cannot measure success based on the simple frequency of relevant events associated with the user-defined label, because our goal is not to generate a map filled with only the relevant events. Instead, we care about the added events being properly integrated into the narrative map, which is why we need to use both clustering and node addition in this task. Thus, we use the same metric from the clustering task (T4), but with only a single “cluster”. Thus, like in T4, when there are no more relevant isolated nodes on the map, the error rate is zero and the simulation ends. The simulation also ends if there are no more events to add (although this ending condition did not occur in practice). Preliminary evaluations showed that using node addition without clustering does not lead to convergence, because the model does not understand that the added nodes are necessarily related.

8.5.4 Qualitative Evaluation

To perform a qualitative evaluation of our model, we first implemented an interactive prototype that allows analysts to extract narrative maps and use semantic interactions on the extracted narrative maps. Using this prototype, we studied two different usage scenarios for SI and narrative maps (COVID-19 and Cuban protests). Then, to evaluate whether our proposed SI model provides value to actual users, we performed an evaluation and review with three experts in visual analytics working in the intelligence analysis domain. Furthermore, the prototype was shown to two analysts working in the same domain. In particular, we discussed the general capabilities of the model and the prototype, the resulting maps of each case study, and the effects of the semantic interactions as we navigated the different examples.

Case Studies. For our case studies, we considered two data sets: a smaller version of the Cuban data set with only 160 documents and the COVID-19 data set used in the design guidelines study [192] which contains 40 documents. We briefly describe these case studies.

1. **Case 1 - The First Month of COVID-19:** The goal of the analyst in this case study is to understand the global spread of COVID-19 and its effects during its first month in January 2020.
2. **Case 2 - The Causes and Effects of the 2021 Cuban Protests:** The goal of the analyst in this task is to gain an understanding of the causes and effects of the Cuban protests. For the effects, the analyst considers both Cuba itself and the US, due to its geopolitical importance in this context.

We note that, unlike the simulated analysts, the case studies do not attempt to correct the map until its error rate is zero. Instead, they show a more natural approach that combines different types of semantic interactions to evaluate their quality as an analysis tool in a more realistic context. Furthermore, they consist of only a few iterations, showing that the model provides valuable insights even with a few iterations of semantic interactions.

Interactive Prototype. We implemented an interactive prototype using the Dash Cytoscape library ⁴. The graph layout of the narrative map is generated using GraphViz [102] and the DOT language [124]. We show the interface in Figure 8.8. This interface supports all the interactions and tasks defined in our simulation methodology and case studies, as well as basic navigation and providing details-on-demand about the elements of the map.

The prototype shows the narrative map on the main canvas on the left. Events are shown on the map with the landmark icon $\bigcirc$ and connected with edges $\nearrow$. Nodes include the logo of the news outlet that published the corresponding article. The main storyline is shown with the dashed blue line $\dashrightarrow$. Edge width is based on the coherence value of the connection.

⁴<https://dash.plot.ly/cytoscape>

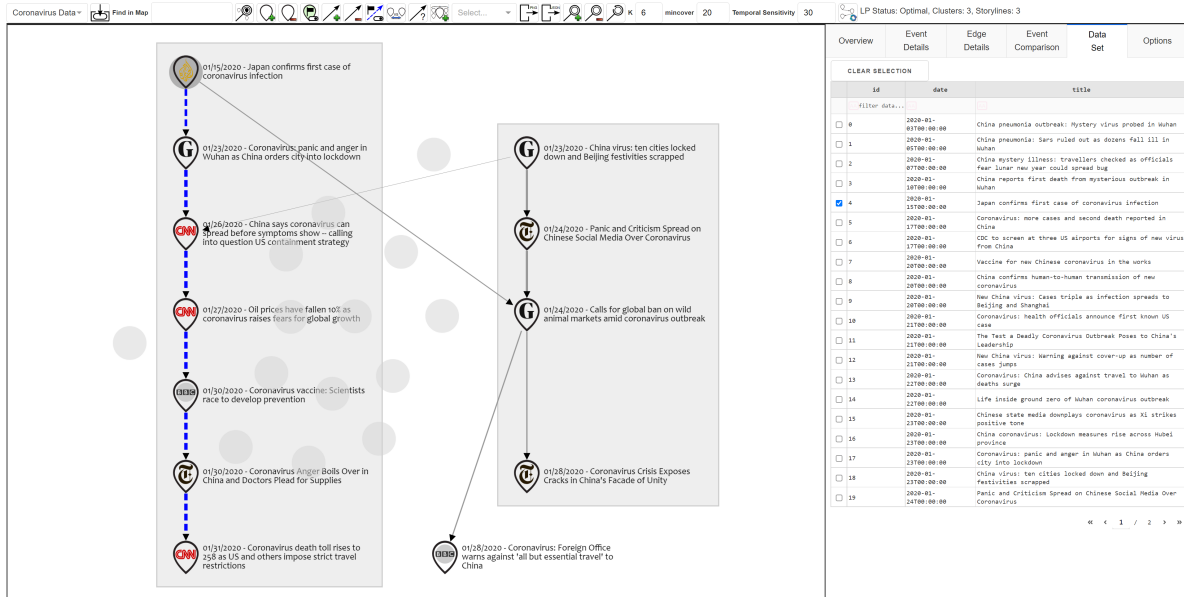


Figure 8.8: Prototype interface showing an example extracted map in the main canvas, the data set table in the right tab, and the main menu and extraction parameters on the top.

We note the addition of background gray nodes  that show similar events to those in the map but were not chosen by the optimization algorithm (i.e., they do not improve the coherence of the map). These extra nodes are useful to guide the “Add Event” interaction, although that interaction can also be done directly from the data table. The gray nodes are organized using a force-directed layout based on similarity. We note that the original narrative maps implementation did not show related events [186], focusing only on the extracted structure. However, we chose to display them to provide analysts with potentially relevant events and help with the node addition interaction.

The top menu allows selecting the data set and loading it into the system . The top menu also contains the core interactions that allow users to manipulate the narrative map, such as “Add Event” , “Remove Event” , “Add Connection” , “Remove Connection” , and “Cluster Events” . The parameters of the extraction algorithm (expected length K , minimum average coverage, and temporal sensitivity σ_t) can also be modified by the user. To create a map, the user must click the “Generate Map” button .

The right panel has six tabs containing additional information and options (overview, event details, edge details, event comparison, data set details, and additional options). For the purposes of the case study, we only used the event details tab (i.e., the contents of the article) and the data set details tab (i.e., a list of all the articles), as the rest are used for development purposes. Finally, we note that the prototype contains other supporting interactions that can aid in navigation and analysis (e.g., searching), but these interactions are not used in the current SI model.

8.6 Results

We now present our results. In particular, we start with our regularization analysis using simulation. Then, we present the main simulation results for all the defined simulation tasks, showcasing the capabilities of the SI model. Next, we present our qualitative analysis of two case studies, followed by our discussion with expert users and the analysis of their feedback.

8.6.1 Regularization Experiment

In this section, we present an additional simulation evaluation to study overfitting issues and the effect of regularization in our extraction model.

We addressed potential overfitting issues by adding a regularization term to the objective of the linear program. However, we have not shown the actual effects of overfitting when this term is not considered nor how this addition properly curtails overfitting issues. Preliminary qualitative evaluations showed that attempting to solve certain tasks led to overly complex maps after performing semantic interactions. Thus, we added a regularization term into the extraction model to minimize the risks of overfitting. In particular, the regularization term added to Figure 8.4 acts like L_1 regularization. This term seeks to minimize the sum of all edge weights, leading to a sparse solution in the number of edges [95]. This provides two-fold benefits: first, it simplifies the resulting map, making it easier to understand, and second, it reduces the likelihood of overfitting (i.e., generating an overly complex map after using SI).

Following these qualitative evaluations, we ran the simulation experiments again to compare both the basic model without regularization and the regularized model. For the purposes of this simulation-based evaluation, we use the simplified version of the Cuban protests data set with only 160 documents instead of the full data set with 500 documents (i.e., the same data used in the the Cuban Protests case study), as running the full-scale simulation is a much more computationally expensive endeavor.

In this context, we first need to define a way to measure overfitting in the context of narrative map extraction. We do this by measuring the *change in complexity* of the map after using SI. A drastically higher complexity value is associated with a higher cognitive load when analyzing the map (e.g., consider the difficulty of using a map with triple the number of edges compared to the baseline). To measure the complexity of the map we use the number of nodes and edges as our main metrics. We note that these complexity values do not have an inherent target or optimal value and their values depend on the map size parameter (K) used in the extraction. However, changes in these complexity measures after interacting with the narrative map are related to our intuitive understanding of changes in narrative complexity. Hence, their values must be evaluated relative to a baseline—the initial map before using semantic interactions.

We summarize our results in Figure 8.9. Due to space constraints, we only show one example

per task for those tasks that had multiple examples. The first column compares the error rate for the basic model and the regularized model. The second and third columns show the average number of nodes and edges, respectively, for each model.

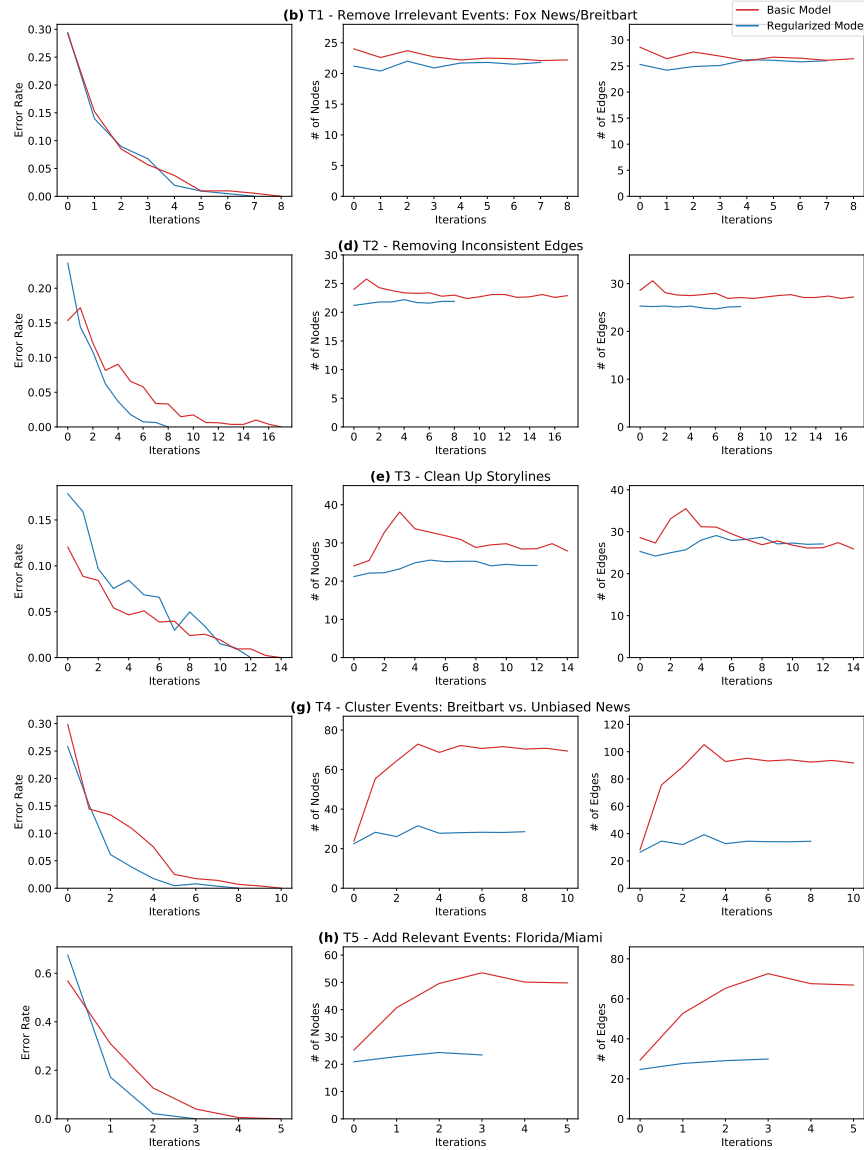


Figure 8.9: Overfitting analysis for the basic model vs. the regularized model. The left column shows the error rate plots. The middle column shows plots for the number of nodes throughout the simulation. The right column shows plots for the number of edges throughout the simulation. Both models converge towards zero error rate, but the basic model has overfitting issues in tasks T4 and T5 based on its drastic increases in graph complexity metrics (number of nodes and edges).

We note that the error rate converges to zero as expected in both models. There are no

apparent differences in the rate of convergence between the two models. Thus, both models are able to capture user intent, just like in the full-scale simulation with the larger version of the data set. Regarding graph complexity, we note that, in general, the basic model usually has a slightly higher base complexity. This is expected, as regularization should reduce the model complexity in general. However, the difference is not particularly significant.

Next, we note that in tasks T1 and T2 both models behave similarly in terms of complexity—there are no significant changes in the number of edges or nodes throughout the simulation. However, we note that for T2 the regularized model converges significantly faster compared to the basic model. This faster rate of convergence could be due to the edge-based nature of task T2. Since the regularized model tries to construct sparser maps, the maps are less likely to contain inconsistent edges.

Regarding T3, we note that the basic model has overfitting issues. The number of nodes increases compared to the initial map and remains higher throughout the simulation. The number of edges increases initially but then returns to lower levels. In contrast, the regularized model only shows a mild increase compared to the basic model.

For T4 and T5, we note that the basic model has drastic increases in the number of nodes and edges even after a single iteration. The regularized model also has an increase in these metrics, but the changes are much milder. Thus, following our interpretation of complexity, the basic extraction model with no regularization has overfitting issues for the clustering-based tasks, while the regularized model is able to reduce the overfitting impacts of the interactions.

In general, we note that semantic interactions that only work on the structure space level do not seem to lead to significant overfitting issues when used in simple tasks (T1 and T2) and only mild overfitting issues when used in more complex tasks (T3). However, semantic interactions that work on the projection space can lead to significant overfitting issues (T4 and T5). Nevertheless, the addition of regularization minimizes the impact of semantic interactions on narrative map complexity and solves most overfitting issues under our complexity-based interpretation for all tasks, without significantly increasing the convergence of the error rate.

8.6.2 Simulation Results

We now present our simulation-based evaluation results for each task in Figure 8.10. In particular, we present the average error rate in each iteration and how many iterations are needed in each task to achieve the target error rate. We note that for all tasks, the error rate tends to zero after a sufficient number of iterations. Moreover, in most cases, it only takes a few iterations to reach a low error rate (e.g. around 5% error rate). However, some tasks are harder than others, taking more iterations to converge. In some cases, even inside the same task, we see some variability in the convergence rate.

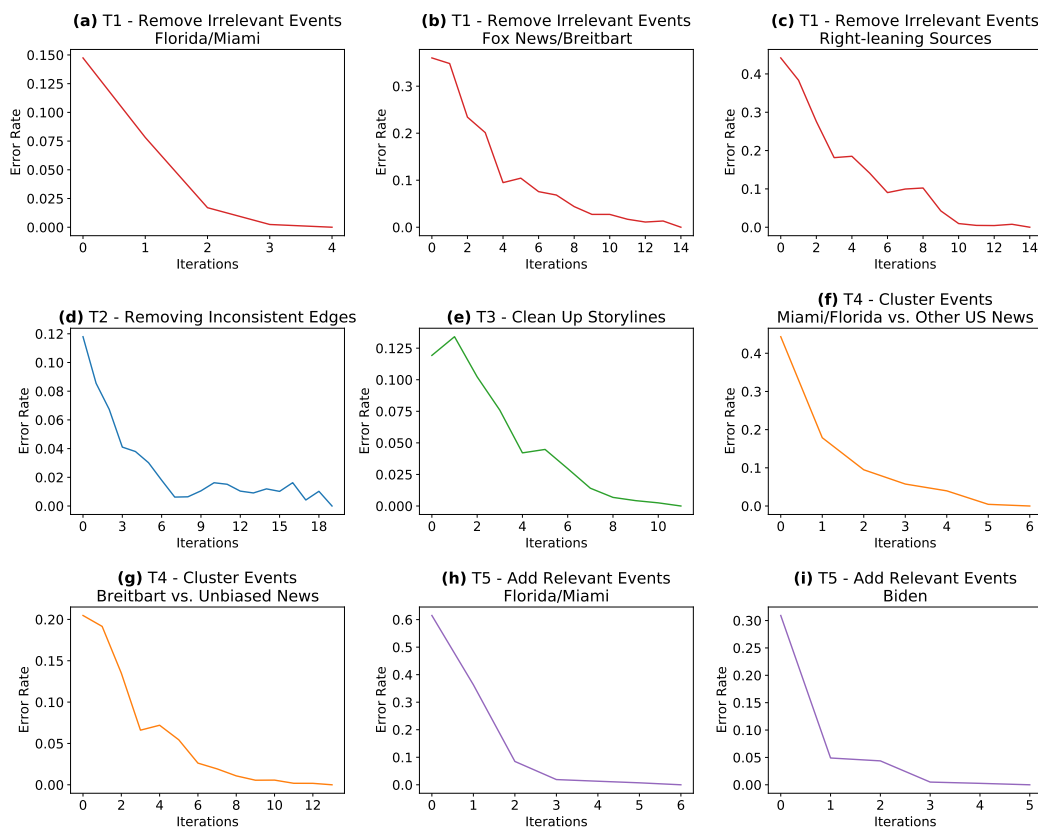


Figure 8.10: Error rate vs. the number of iterations. **(a, b, c)** T1: Remove Irrelevant Events. **(d)** T2: Remove Inconsistent Edges. **(e)** T3: Clean Up Storylines. **(f, g)** T4: Cluster Events. **(h, i)** T5: Add Events.

The hardest task—in terms of iterations needed—corresponds to the edge-based task T2. In particular, after the initial drop in error rate, the number of inconsistent edges stabilizes below 2%, but it takes up to 18 iterations to completely remove them. We note that this is an inherently harder task compared to event-based tasks due to the number of possible combinations of edges that can arise in a graph, which rise quadratically with the number of nodes. Nevertheless, if our goal is to remove most of the inconsistent edges and achieve a mostly consistent narrative map, we could stop between 7 to 9 iterations, which results in around 2% inconsistent edges.

As an example of intra-task variability, consider T1, where the second and third cases take longer than the simpler case of removing Florida/Miami events. The slower convergence is expected, as these two specific sub-tasks are more complex, as the user-defined label does not rely on simple keywords, but rather the abstract concept of political bias. Likewise, T4 shows similar behavior, where the keyword-based clustering creates an easier task compared to the more abstract clustering based on bias vs. unbiased news.

8.6.3 Case Studies

COVID-19

We now present the qualitative analysis of the COVID-19 case study as an extended example. The goal of the analyst in this case study is to understand the global spread of COVID-19 and its effects during its first month in January 2020. However, during this time period, most of the news articles reported events from China, leading to the data set being mostly focused on Chinese news. Thus, the analyst needs to set an appropriate starting point that is more likely to lead to a map detailing the global spread of the virus and its effects. In this context, the analyst sets the 5th event of the data set—which reports Japan’s first COVID-19 case—as the starting point, as it corresponds to the first news article that is not about China in the data set. In terms of parameters, the analyst uses $K = 6$ for the map size and default values for the rest. Figure 8.11 shows all the maps created throughout the analysis process and the interactions done by the analyst.

Figure 8.11(a) shows the initial map and the interactions of the analyst. The map has two major storylines—shown in the map as linear subgraphs in gray boxes. The main storyline represents the most important path in the narrative map and is computed using maximum likelihood [186] and is shown with blue edges. All other connections use gray edges. Finally, there is also a singleton storyline—a storyline with only one element—shown without a gray box.

On a more detailed inspection, the main storyline has some relevant events about the effects of the virus and its spread (e.g., comments about the US containment strategy, oil prices, vaccines, and travel restrictions). However, there are also some events about the effects in China that are not necessarily useful for the analyst’s goal. The other storylines are also all focused on China. Thus, to shift the focus of the map towards the global spread and its effects, the analyst performs the sequence of interactions shown in Figure 8.11(a).

In particular, the analyst adds two potentially relevant events to the map: global markets being on edge and the CDC screening at airports. The first event relates to the global economic effects and the second to the international prevention strategies. Afterward, to ensure that the added events are integrated consistently into the map, the analyst clusters them together with other relevant events from the main storyline (see the highlighted events in Figure 8.11(a)).

Then, after completing the interactions, the analyst generates the new map shown in Figure 8.11(b). This map still has two storylines, and the events highlighted in blue that we assigned to the same cluster are mostly connected (some indirectly). However, there are still some issues, such as irrelevant events and the fact that the map still focuses on China. To fix the first issue, the analyst removes the irrelevant events on animal or omnivorous markets—which are potentially important as a cause of the start of the pandemic, but not relevant to its ongoing spread and effects—and the event about daily life under quarantine—which is

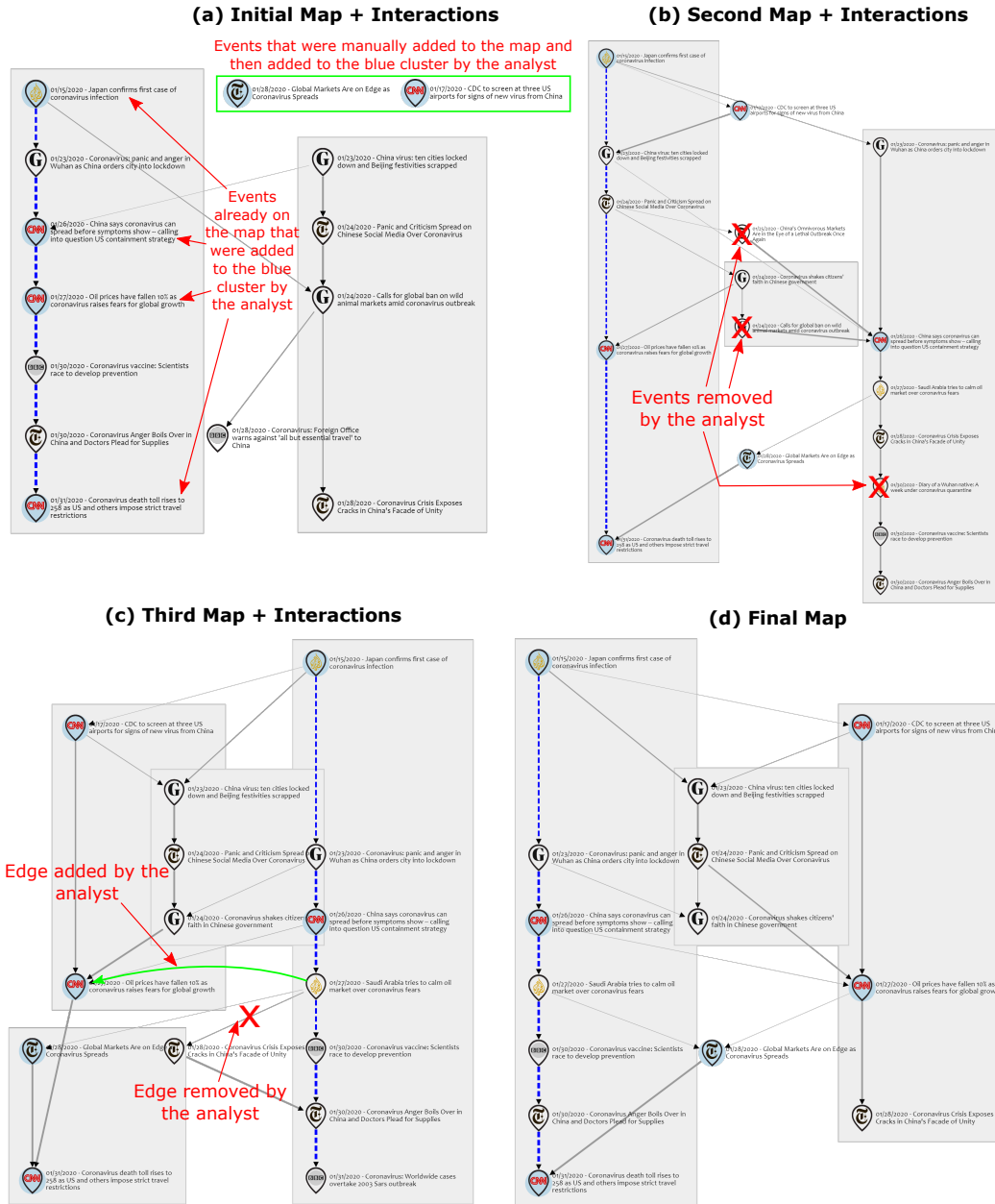


Figure 8.11: (a) Initial map and analyst interactions (add events and cluster events). (b) Second map and analyst interactions (remove events). (c) Third map and analyst interactions (add edge and remove edge). (d) Final map after all iterations. Note that a higher-resolution version is available as part of the supplementary materials.

also irrelevant for the analyst's goal.

These interactions lead to the map shown in Figure 8.11(c). This map has four major storylines and all highlighted events are connected (technically, their induced subgraph is

weakly connected). The main storyline now covers more relevant events, and although it still contains events about China, they fit into the overall context (e.g., lockdowns and lack of supplies). Two of the storylines consist of only highlighted events, thus they focus exclusively on the global spread and effects on a worldwide scale. Furthermore, these storylines are connected through relevant inter-story connections with the main storyline. Furthermore, the map now has distilled an extra storyline about Chinese socio-political effects into its own storyline. Thus, providing a better structure to represent the narrative.

While the current map provides a good overview, it could be improved further. In particular, the analyst could add a direct edge between the two events about the oil market. Furthermore, the analyst could disconnect the event about Saudi Arabia from the Chinese facade of unity event, as they are unrelated. These operations lead to the map shown in Figure 8.11(d). This time the interactions did not introduce major changes. In fact, two storylines are preserved from the previous map (main story and Chinese socio-political effects). The third story is about relevant side effects, such as the US screening for the virus leading to oil prices dropping due to fear. However, the last event of this storyline does not seem to fit and could be removed to further improve the representation. The only singleton storyline of this map is placed in an interesting position, its connections imply that the fear in oil markets has extended to global markets.

Overall, the final structure showcases some of the key storylines about the early impact of COVID-19, including the global spread and the ensuing economic issues. Furthermore, the map still contains relevant information about the socio-political effects in China, which were not prevalent in other countries at this point (e.g., lockdowns and unrest). All these storylines provide valuable insights into the pandemic and even foreshadow future effects of the pandemic.

Cuban Protests

The goal of the analyst in this task is to gain an understanding of the causes and effects of the Cuban protests using a smaller version of the Cuban data set with only 160 documents. For the effects, the analyst considers both what happened in Cuba itself and any relevant developments in the US, due to its geopolitical importance in this context. Relevant effects range from the general response by the administration to more specific socio-political effects. The analyst generates a map using the same event we used for our simulation experiments, namely: “Cuba sees biggest protests for decades as pandemic adds to woes”. In terms of parameters, the analyst uses $K = 6$ for the map size and default values for the rest. Figure 8.12 shows the resulting maps and interactions of the analyst.

Using these parameters, the system generates the map shown in Figure 8.12(a). This map has two major storylines, as well as four minor singleton storylines. The main storyline starts focuses on the Cuban protests at first, but it slowly starts to drift toward US-related issues. The side storyline that comes afterward is focused mostly on socio-political effects

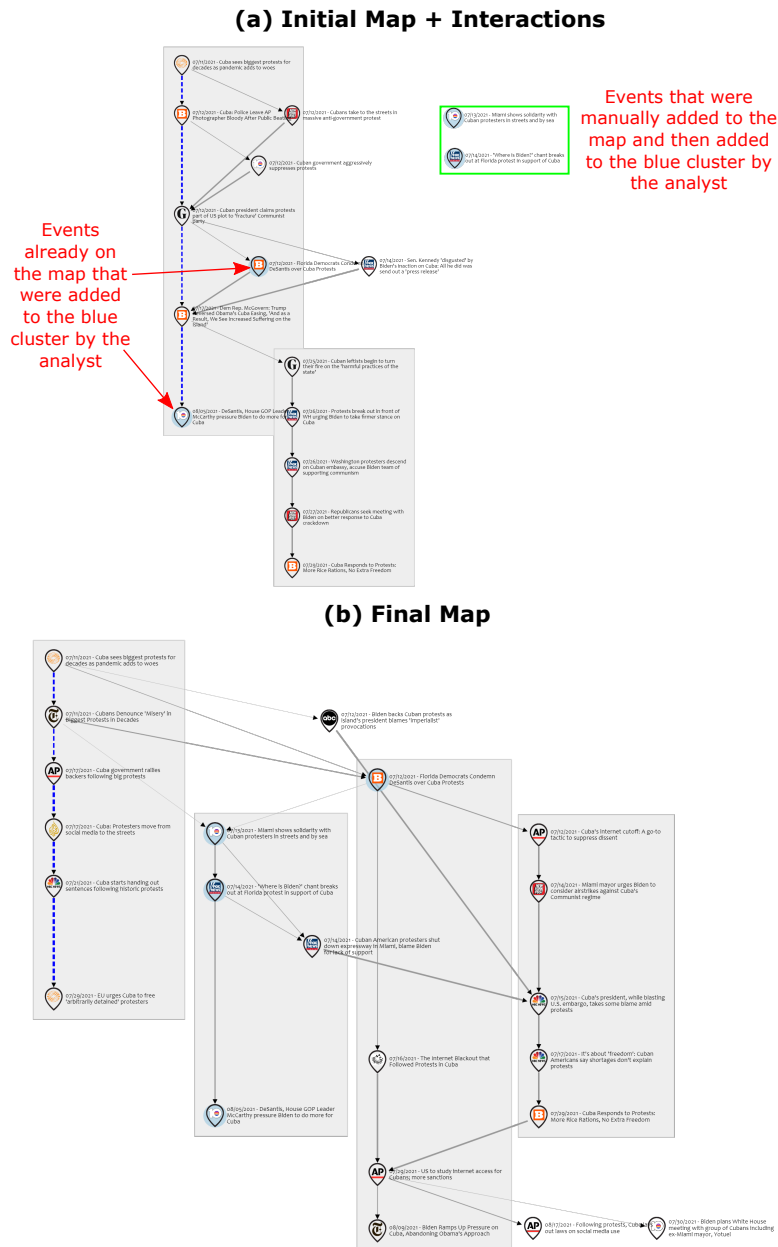


Figure 8.12: (a) Initial map and interactions performed by the analyst (add events and cluster events). (b) Final map after a single iteration. Note that a higher-resolution version is available as part of the supplementary materials.

in the US, including protests and political struggles. However, there are two events that are relevant to the Cuban protests themselves. Furthermore, the first two singleton storylines simply reinforce some aspects of the protests (e.g., aggressive suppression of the protests). The second two singleton storylines, which happen after the Cuban president claims that the protests are a US plot, provide some insights into how the US is responding to these

events, particularly, there is political infighting in Florida and calls to action from the senate. Despite some useful insights, the current structure is not that practical as it has a disjointed coverage of different aspects of relevant events. Thus, the analyst will perform semantic interactions to generate a better structure to represent the narrative.

In particular, the analyst wants to find out more about the effects on Florida, as several events mention this state and related entities or keywords (e.g., DeSantis and Miami). The analyst adds two relevant protest events, as seen in Figure 8.12(a). Next, the analyst groups these events with the DeSantis events that were already present the clustering interaction. The expected result of this interaction is a map that focuses more on the effects on Florida and the US.

The actual results are shown in Figure 8.12(b), where we see four major storylines and four singleton storylines. The main storyline is now exclusively focused on the Cuban protests, including counter-protests and government repression. which is a nice side effect of the interactions. Second, we see a storyline that has only highlighted events that the analyst clustered, corresponding to a storyline focused on Miami. Next, the third storyline starts with a DeSantis event but then turns its attention to a series of events about trying to provide internet access to Cubans. This connection makes sense when delving deeper into the article's contents, which mentions that DeSantis is pushing for the administration to provide internet access to the island. One of the singleton storylines right after this also provides further insights, as Cuba outlaws social media usage after the protests. The last storyline also mentions the internet cutoff on the first event, but then shifts focus toward calls for military actions against the Cuban regime. Afterward, the storyline gives us some insights into the potential causes of the protests from different points of view. From the Cuban perspective, the US embargo was the key cause, although the government also admits some blame for the economic shortages and lack of supplies. From the Cuban Americans' perspective, it is more about freedom than supply shortages. Finally, the last event highlights how the final response of the government was to attempt to fix the supply issues by giving more rations without any fundamental change to the regime. While further improvements would be possible, the current layout and contents provide a decent overview of the main developments of the Cuban protests, and the protests in Florida, as well as some insight into the potential causes of the protests (e.g., supply shortages), as well as some relevant effects (e.g., banning social media and censorship).

8.6.4 Expert Feedback

First, the experts provided feedback on the general narrative maps framework, highlighting how the prototype could provide analytic value by generating a structured representation of the documents organized over time into different storylines. Second, regarding the SI model, the experts found that it was able to properly capture user feedback in most cases and examples, leading to improved maps or clearer storylines. Furthermore, we discussed the

potential value of the model for exploratory tasks and directed tasks. For exploratory tasks, the value of the model was clearer, as they provide analysts with a series of potentially relevant storylines, which can be then explored in-depth via SI. In contrast, for directed tasks, the value would depend on the type of objective, as not all directed tasks would need a structured approach (e.g., they could be solved with simpler search-based models). However, if the task required finding complex connections, then the narrative model provided more value.

In general, the experts found the general narrative maps framework useful and the SI model capable of properly capturing user feedback. However, they highlighted two potential issues: *transparency* and *scalability*. In particular, the current SI model acts as a black box, and understanding its effects is not straightforward. Thus, it could be complemented with an explainable AI approach to provide clear explanations of the changes induced by the model. Regarding scalability and performance issues, we note that this is one of the main drawbacks of the current implementation, as the optimization approach and the layout algorithm have issues scaling to larger data sets. Nevertheless, the experts noted that scalability and transparency are generally considered open problems in complex sensemaking systems. Thus, regardless of these issues, the experts found the 3MSI narrative model to be a welcome advancement to help during the sensemaking process. Finally, the experts noted that they had not seen any previous work that integrated SI capabilities with narrative extraction, which provided additional value and it would be “extremely useful” for certain sensemaking tasks.

8.7 Discussion

8.7.1 Semantic Interaction Model

Our results show that the proposed 3MSI system is capable of integrating analyst feedback into the narrative model. In all tasks, the average error rate tends to zero as the analyst refines the model, although we note that the convergence rate depends on the specific task and its inherent difficulty. Thus, our experimental results support the hypothesis that our 3MSI model for narrative maps is effective at supporting incremental formalism [329] based on iterative interactions.

In general, our 3MSI concept provides a flexible approach that can enable researchers to use higher-level discrete structures as outputs of a semantic interaction pipeline in conjunction with a lower-level continuous model (or even multiple lower-level models). Furthermore, our proposed guidelines for handling single-level and multi-level interactions can aid researchers and practitioners in the construction of complex SI models. However, while the 3MSI approach is flexible, it could become increasingly complex depending on the number of levels we include in the pipeline.

Furthermore, in the context of interactive AI [381], if we seek to add an explainable AI

component to complement our SI model, each additional layer in the system’s architecture increases the complexity of generating explanations, as each transformation between layers requires some way to “invert” it and extract meaningful information. This increasing complexity makes dealing with the transparency issues highlighted in our expert review a more complicated endeavor.

8.7.2 Semantic Interactions and Narrative Sensemaking

Our simulation-based analysis provides a replicable and scalable evaluation for our narrative SI models without requiring human analysts. However, we note that due to the relative simplicity of our test tasks compared to real sensemaking tasks, these results might overestimate the incremental formalism capabilities of the SI model, yet it provides a good baseline. In practice, we would expect real sensemaking tasks to require more iterations to refine the model. Nevertheless, the fact that, on average, we were able to learn from different types of analyst interactions is promising.

Regarding our qualitative evaluation, our case studies and subsequent expert evaluation show that SI is capable of providing value to analysts and aiding in the sensemaking process. In particular, both the final structure provided by the narrative map after SI and the intermediate maps generated in each step are useful for gathering insights from the data. In this context, despite the main goal of semantic interaction being to generate a “good” final map by trying to capture user intent through successive iterations, the iterative process itself also has value to the analyst. Even if each individual iteration is not a perfect map, the exploratory nature of this process can provide valuable insights, which can be compounded as the analyst explores different configurations of the structure space. Even in cases where SI fails to generate a “good” map, there is value in the exploratory analysis that it allows.

Overall, our findings suggest that combining semantic interactions with narrative visualizations can assist in the sensemaking process and allow analysts to change the story according to their own preferences. Intelligence analysis [107], computational journalism [53], and other fields that require analyzing and understanding narratives could benefit from our proposed SI model for narrative maps.

8.7.3 Model Transparency and Performance

Two of the key issues of the current SI model as identified by our analysis with experts are *transparency* and *scalability*.

Regarding transparency, the system should provide appropriate explanations of the changes it makes to the narrative to the users. Thus, it would be useful to complement SI with an explainable AI model that provides such feedback. In the context of creating a general interactive AI model [381] for narrative maps, adding an explainable AI component to our

pipeline would be the natural next step. Furthermore, we also note that the transparency issues are likely accentuated due to the usage of non-incremental layout algorithms for the output graphs, which can lead to additional cognitive load due to drastically different positioning of events.

Regarding performance and scalability, we note that it would be possible to reduce the computational cost by applying techniques to group documents together and treat them as a single event, rather than keeping the assumption of a one-to-one relationship between documents and events. Such an approach could be helpful to reduce the computational cost of the map extraction process and the cost of the subsequent layout generation. However, it would require extensive pre-processing and appropriate machine learning models to detect which documents refer to the same event.

In this context, referring back to topic detection and tracking literature [197, 265] to identify events or other structured narrative extraction approaches [18, 232, 343] that represent events as clusters of documents could be useful. In particular, adding an event extraction model in the middle of the proposed 3MSI pipeline for narrative maps could help reduce the overall computational cost and make the model more scalable, as the last step to generate the structure space is the most computationally expensive part of the pipeline.

8.7.4 Interaction Ambiguity

While our SI model was able to capture knowledge from the simulated analyst, there is ambiguity in how to interpret the user interactions properly. The challenge of defining user intent from interactions has been called the “With Respect to What” problem [165, 383] in SI literature. Possible solutions to this issue include providing additional interactions [383] so that the user can specify with more precision what they actually want to obtain (e.g., by using additional interactions) or providing the ability to control the impact [165] of the interactions (e.g., by designing a weighting scheme that integrates elements from original representation and the post-interaction representation). In our particular example, analysts could specify that the current stories should be kept together by using the clustering interaction. By doing so, the model would be forced to keep the original storylines in the representation while also integrating the new story. Thus, ensuring that the model properly captures the incremental formalization process, rather than throwing away the previous narrative map. However, further exploration is required to define a robust model of user interaction for narrative maps that captures user intent correctly when there is ambiguity.

Furthermore, based on the feedback from the experts, an important step that can aid in understanding how the interactions are working is explicitly showing the changes introduced by the semantic interactions. Adding such explanations would help users determine whether the interactions are capturing their intentions properly when there is potential ambiguity. Furthermore, in case the semantic interactions have misinterpreted the intent of the analyst, they could apply corrective interactions to the highlighted changes and guide the extraction

algorithm to generate a new map that deviates less from the original intent.

8.7.5 Limitations

Our work is not without limitations. While the general 3MSI framework provides a general approach to semantic interaction, our specific SI approach for narrative maps is not model agnostic. The current model is built around the linear programming extraction model and the use of a DR method for the embedding representation of documents. Thus, any changes to the extraction pipeline would require figuring out how to integrate constraints into methods.

Regarding the evaluation of the SI model, we note that the simulation-based experiments only used simple tasks that do not represent the full complexity range of narrative sense-making tasks that analysts could perform with a narrative map. To ameliorate this issue, we presented in-depth case studies that attempt to replicate real sensemaking tasks and gathered feedback from experts in visual analytics working in the field of intelligence analysis.

Finally, we note that there are other potential interactions that could be used to capture the intent of analysts. For example, manually changing the main story of the narrative map or highlighting a specific event to mark it as more important. However, the current model does not handle such interactions in its current form. Despite these limitations, our quantitative and qualitative results show that our 3MSI model has the potential to help analysts modify computational narrative models through simple interactions with narrative visualizations, without having to understand the underlying extraction or representation models used to generate the narrative visualization.

8.8 Chapter Conclusions

In this chapter, we proposed the concept of 3MSI—Mixed Multi-Model Semantic Interaction—an SI pipeline defined by a higher-level discrete structure and lower-level continuous models. Furthermore, we showcased the capabilities of this approach by developing an SI model for narrative maps, which had a lower-level model in the form of a document embedding projection space and a higher-level narrative structure space.

We evaluated our SI model through a quantitative simulation-based approach. The evaluation showed that the SI model is capable of integrating the simulated analyst’s interactions into the narrative model. Thus, supporting the hypothesis that the SI models are effective at supporting incremental formalism for narrative maps.

Furthermore, our review of case studies and validation with experts led to valuable feedback regarding the SI model in particular and the interactive narrative maps framework in general. The expert feedback showed that the SI model for narrative maps could provide value to analysts in their sensemaking process. In terms of broader impact, integrating SI models

with narrative maps allows us to better support the sensemaking loop for narrative sense-making tasks, such as journalistic analysis of news narratives [53] and intelligence analysis [107]. However, we note that there are two key issues in the current SI model for narrative maps: *scalability* and *transparency*. Thus, future work should seek to improve the extraction approach to allow for the analysis of larger data sets and explainable AI techniques should be used to improve the transparency of the SI model and the changes it makes to the narrative maps.

Finally, in terms of broader impact, the 3MSI concept provides a flexible framework that allows researchers to employ higher-level discrete structures as outputs of an SI pipeline in combination with different continuous internal models and representations. Furthermore, our proposed guidelines for dealing with single-level and multi-level interactions can help academics and practitioners build complex SI models.

Chapter 9

An Insight-based Evaluation of Interactive AI for Narrative Maps

Human-AI collaboration for sensemaking is an emerging topic with multiple challenges in the context of visual analytics (VA). In this work, we present a mixed multi-model interactive AI framework that seeks to address some of the issues that arise when building VA systems for human-AI collaboration. In particular, the proposed framework utilizes semantic interaction and explainable AI to enable human-AI collaboration in the context of a mixed multi-model pipeline. In particular, this framework addresses some of the specific challenges that arise when combining high-level discrete structures with low-level continuous spaces in AI models. We showcase the capabilities of the proposed framework in the context of narrative sensemaking with an interactive narrative maps prototype. We evaluate the effectiveness of our prototype and models through a user study following an insight-based evaluation methodology. We further analyze the behavior of users and explore their cognitive strategies. We find that our proposed framework supports analysts in their sensemaking process and aids them in the process of generating insight when compared to the baseline. We note that this chapter contains results from a human subjects study approved by IRB 21-929.

Visual analytics (VA) seeks to support the analysis of large and complex data sets through interactive exploration of data [81]. In particular, VA systems assist users in their sensemaking process by allowing them to interact with statistical and artificial intelligence (AI) models through visualizations to explore connections and hypotheses in the data, with the goal of obtaining insights [104].

The sensemaking process of users is formalized by the *sensemaking loop*, which is comprised of two main sub-loops: foraging and synthesis [282]. The foraging loop consists of filtering and gathering collections of relevant information, which naturally lends to computational support. The synthesis loop consists of creating and testing hypotheses about the foraged information and the underlying relations, which relies on human intuition to find relationships in the data.

In this context, recent research has sought to better integrate human cognition with the underlying models to support the sensemaking process and enable further collaboration between humans and AI [381]. This integration can be approached in two ways: by helping the human understand the underlying model better through the use of *explainable AI* (XAI) [5, 30] and by helping the computational model understand better the users' underlying

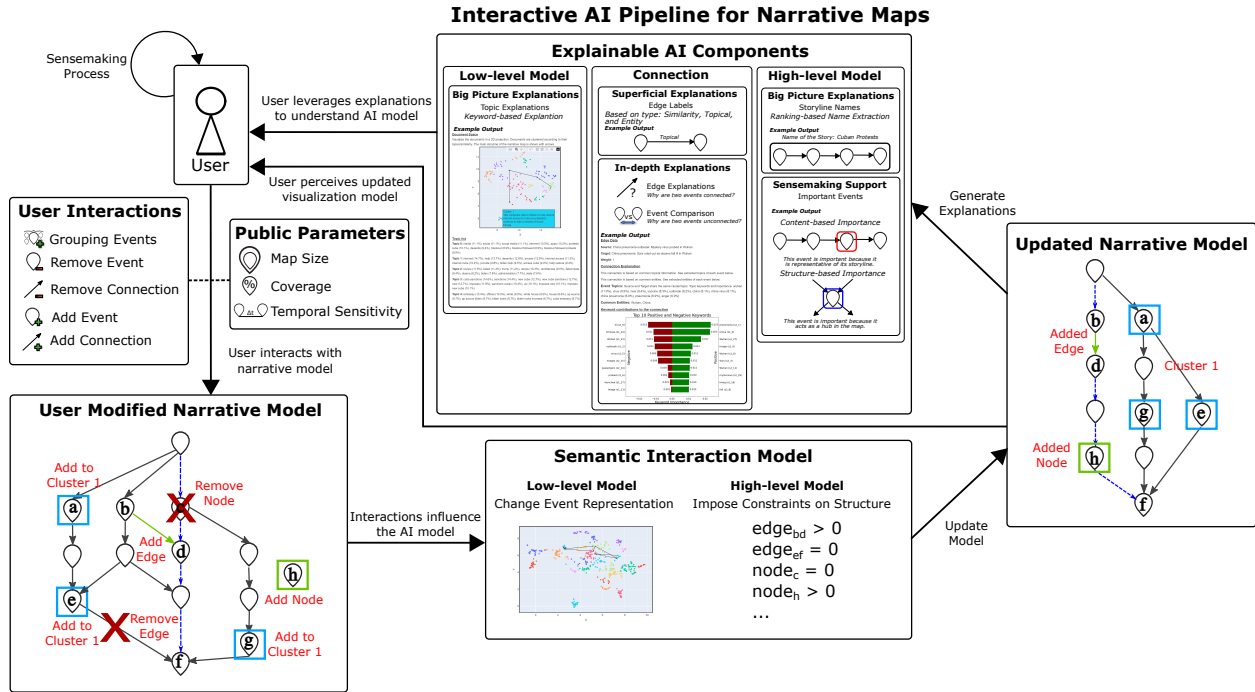


Figure 9.1: Illustration of the interactive AI framework applied on narrative maps for sensemaking. The core components of this pipeline are the *semantic interaction model* and the *explainable AI components*. The user interacts with these components during sensemaking by changing the visual narrative model and leveraging the explanations of the explainable AI components to understand how the narrative was constructed.

cognition by capturing their intent through interactions through the use of *semantic interaction* (SI) [52, 189]. This bidirectional pipeline of human-AI collaboration is referred to as *interactive AI* [381].

However, the integration of interactive AI with VA systems raises several challenges that need to be addressed, such as how to externalize cognition and convert it into a machine-understandable form, and how to provide contextual outputs from the underlying AI models [381].

In this work, we present an interactive AI [381] framework for mixed multi-model [189] VA systems—which relies on a high-level discrete structure for visualization (e.g., a graph) and low-level continuous models to construct this structure. This proposal seeks to address the aforementioned challenges and provide a general framework to develop complex interactive AI systems. Furthermore, we identify challenges that are specific to the mixed multi-model case and provide general guidelines for how to approach them.

To demonstrate the capabilities of our interactive AI framework, we develop a VA prototype for narrative sensemaking [186, 189] implemented following our proposed framework. We showcase our implementation and describe the application of the framework and specific

challenges that arise in the context of narrative sensemaking. Finally, we present a quantitative and qualitative evaluation of our prototype against a non-interactive map-based prototype and a timeline-based prototype with search functionalities through a user study with an insight-based evaluation methodology. Thus, we present three core contributions:

1. The concept of the **Mixed Multi-Model Interactive AI** framework, an interactive AI pipeline that integrates SI and XAI using a high-level discrete structure built from lower-level continuous models.
2. A prototype of a VA system for narrative visualization to demonstrate the capabilities of our interactive AI framework.
3. A user study following an insight-based evaluation methodology to better understand the affordances of our framework and the capabilities of our VA prototype for narrative sensemaking.

Our results show that our interactive AI prototype provides advantages compared to the baseline, enabling analysts to generally find more insights and develop complex sensemaking strategies in the process. We show an overview of our proposed framework and our interactive narrative sensemaking prototype in Figure 9.1.

9.1 Background and Related Work

9.1.1 Narrative Extraction and Visualization

Narratives can be defined as a system of interrelated storylines that share coherent themes [145, 186]. Moreover, there are many ways of telling the same story, which leads to one of the core distinctions of narratological studies that seek to understand the relationship between the underlying story and its representation [1, 285].

In the context of computational narrative representation and extraction, most approaches [189] rely on event-based models [190]. Events are the fundamental unit of analysis for narratives [192] and they either represent actions performed by characters (or entities in general) or happenings that do not directly involve entities [1]. Event-based representations rely on discrete structures—such as timelines, trees, or directed acyclic graphs—to represent the narratives. The choice of structure is associated with a trade-off between complexity and representational capabilities (e.g., the ability to model parallel storylines) [186]. Regardless of the underlying structure, computational extraction of narratives typically relies on an optimization approach that seeks to maximize a notion of quality (e.g., coherence) [190]. In this work, we rely on an extraction approach that maximizes coherence subject to coverage constraints [189].

In the context of VA and information visualization, narratives and storytelling are common techniques used to present data [297, 316, 348]. In particular, presenting visualization as stories can be used to aid in the sensemaking process [170, 171]. Visual storytelling systems can help users detect relationships, structures, and other patterns, which can help in confirming hypotheses or gaining additional knowledge [8, 348]. In this regard, constructing and interacting with visual narratives could be interpreted through the lens of a visual knowledge generation process [307], as users generate new knowledge as they work through their hypotheses and insights with the visual narrative model.

In this work, we develop an interactive AI model for narrative sensemaking based on the concept of narrative maps as defined by Keith and Mitra [186]—a DAG describing the connections between events. We show an example of a narrative map generated by our interactive AI prototype in Figure 9.2.

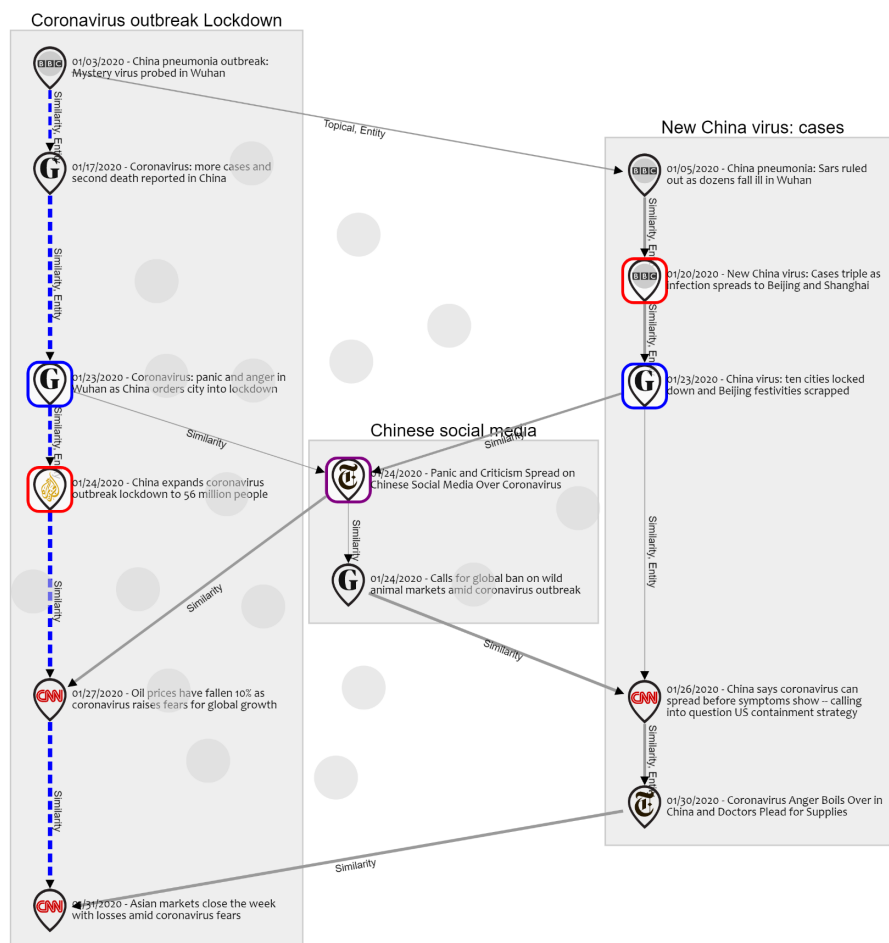


Figure 9.2: Narrative map example based on a COVID-19 news data set. The model includes explanations such as storyline names, connection labels, and highlighted important events.

Previous work has found that this approach provides analysts with a useful representation

for narrative sensemaking tasks [186, 193]. Furthermore, we build upon recent work on an SI framework for narrative maps [189] to develop our interactive AI framework. This pre-existing SI framework has been tested through a simulation-based evaluation [40] and expert feedback, which has shown that it is capable of supporting the sensemaking process of analysts. We note that while this aforementioned framework [189] partially addresses the issue of interactive AI for narrative sensemaking, it only covers the SI component. Our work builds upon this foundation by extending the SI model with XAI components and creating a full interactive AI pipeline. Thus, this research provides a novel approach to creating an interactive AI system that aids in the narrative sensemaking process by integrating XAI and SI to help the analysts understand and improve the narrative structures generated by the system.

9.1.2 Interactive AI

Interactive AI [381] is a symmetric model for human–artificial intelligence collaboration built upon the design philosophy of SI [105]. Interactive AI seeks to solve the “two black boxes” problem, which refers to the communication challenges that exist between the AI system and human cognition. Interactive AI systems address these challenges through two key components: an explainable artificial intelligence (XAI) component and an explainable cognitive intelligence (XCI) component. These systems enable humans to understand the results and analyses provided by the machine—the XAI component—and also enable the machine to adapt and learn based on human cognition—the XCI component. This symmetric loop of understanding is enabled through SI.

Explainable AI [157] is an area that leverages machine learning and visualization techniques to convey information about the state of the AI systems to a human user with the goals of building trust in their output, demonstrating the uncertainty in the results, and contextualizing the results. In contrast, explainable CI focuses on modeling the cognitive state of the human user and converting the externalized cognition into a format that a machine can understand. For example, XCI can turn user interactions into a set of parameter or value updates that reflect the user’s intent. Based on these interactions, the AI model can learn what the user needs and finds interesting [381].

In this context, we highlight recent work on a human-AI interactive framework for continuous sensemaking in the domain of image processing and classification [327]. In particular, the aforementioned framework is implemented for a classification task with deep learning models. The framework leverages expert feedback to continuously improve the underlying model, through a process similar to SI. However, while the proposed framework is applicable to many relevant sensemaking processes, it is designed to support pattern recognition tasks where training data is available (i.e., supervised tasks). In contrast, our proposed framework does not require training data, but instead focuses on a *multi-model* pipeline involving a *mixed model* approach (i.e., it mixes continuous models with a high-level discrete structure),

without paying attention to the supervised or unsupervised nature of these models.

9.1.3 Explainable AI

One of the key challenges of effective human-AI interaction is overcoming the black-box nature of AI models [327]. In this context, the field of XAI arises as an approach to solving this issue. In particular, XAI tries to overcome this challenge by opening the black boxes to humans and uncovering the underlying reasoning and processes of the AI models [381].

Furthermore, recent human-computer interaction research has sought to address XAI in the context of visualizations and sensemaking [361]. In particular, under the assumption that transparency of the AI model would aid in building trust between humans and AI [338], XAI provides a way to make the models more transparent [5], and thus enhance users' trust in the model and foster human-AI collaboration [373].

In general, XAI methods can be classified into two broad categories [30]: model-agnostic techniques—which can be applied to any machine learning model without consideration of its internal structure—and model-dependent techniques—which are specifically tailored to a model and sometimes overlap with certain agnostic techniques. In general, model-agnostic approaches include simplification-based explanations [200], relevance-based explanations [238], local explanations [295], and visual explanations [238, 295]. In this work, we use a mix of model-agnostic visual explanations and relevance-based explanations through Shapley plots [238]. Furthermore, we develop model-dependent explanation methods for our narrative model which try to explain specific elements of the extracted maps.

We note that, to the best of our knowledge, there are no works that address XAI in the context of narrative extraction and visualization for sensemaking. Thus, in this work, we address this gap in the literature by using our proposed interactive AI framework to create an XAI model for the narrative extraction process based on the narrative maps model [186]. In particular, we implement a mix of model-agnostic and model-specific approaches.

9.1.4 Semantic Interaction

Semantic interaction models leverage the natural interactions of a visualization model to learn the intent of the user and modify the underlying model [106]. More specifically, SI models allow the users to indirectly modify the parameters of the AI model to fit a user-defined concept through interactions with the visualization. SI has seen applications in the context of text analytics [104] and high-dimensional data in general [317].

In this context, SI systems can also be related to knowledge-assisted visualization systems [117], which rely on traditional knowledge bases to augment their capabilities. These systems can capture user interactions to acquire tacit knowledge from the user and improve their own

model [235]. However, unlike these systems, SI does not explicitly rely on external knowledge bases, but rather on machine learning models (e.g., dimensionality reduction).

Recent research has sought to develop SI models as a multi-model pipeline [52, 97], which are usually implemented using a continuous dimensionality reduction (DR) method. In this context, users rely on the visualized DR projection to obtain insights about the data and make changes based on these insights by interacting with the visualization (e.g., changing the position of specific data points). The SI model captures these changes and interprets them to update the underlying AI model and its parameters through a series of *inverse transformations*, which can be implemented through probabilistic, heuristic, or mathematically rigorous methods [189]. Then, these changes are used to create a new projection that should capture the user’s intent. Regardless of the specific AI model used, the general approach remains the same: users interact with the visualization to indirectly modify the underlying model and its parameters [52, 104].

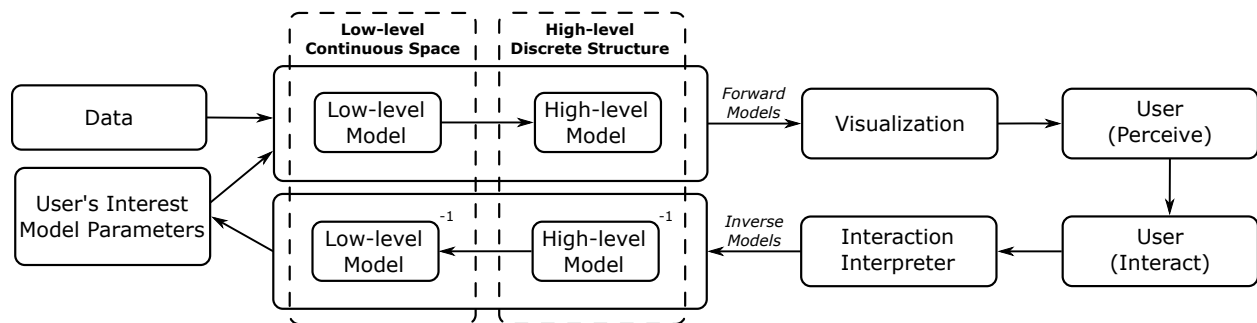


Figure 9.3: Mixed Multi-Model Semantic Interaction pipeline (source: [189]).

In the context of narrative sensemaking, we face an additional challenge because narratives have an underlying temporal and causal structure [1], unlike DR models or other continuous models. In particular, the narrative map model uses a graph-based representation that does not have a natural notion of continuous distance that could be used to define inverse transformations in the SI model [186]. Thus, it is necessary to deal with the discrete nature of the model. Recent work [189] has sought to develop a framework to handle *mixed* multi-model SI that combines a series of continuous low-level models that are used to build a high-level discrete structure. We show this pipeline in Figure 9.3 and we note how it comprises multiple models of mixed nature. Our work builds upon this SI framework and extends it into a full interactive AI framework that can handle the mixed nature of the narrative extraction pipeline.

9.2 Interactive AI Framework

9.2.1 Overview

In the proposed interactive AI framework—shown in Figure 9.4—the *AI model* takes some data and model parameters and generates a visualization. Internally, the AI model consists of a low-level continuous space—or potentially multiple of these continuous spaces—and a high-level discrete structure that is used to generate the visualization.

Our framework is *mixed*—in the sense that it is designed to address the specific challenges of using a discrete structure built on top of continuous spaces—and *multi-model*—in the sense that it relies on multiple internal AI and statistical models. The framework further assumes that the high-level model does not have a continuous notion of distance that can be leveraged to model user interactions. Otherwise, all models could be treated as continuous and there would be no need for relying on the mixed nature of the framework.

Consider for instance a force-directed layout of a graph. This representation is discrete as it uses a graph, but it is possible to use a continuous notion of distance to model interactions. In contrast, representing a graph with a hierarchical layout does not have such a notion and it would not be possible to directly translate display distances into changes in the internal model.

On top of the AI model, a series of *XAI components* help provide additional contextual information and insight into how the model has produced its output, representing the first half of the interactive AI loop. In particular, the XAI components involve explanations of each internal model. Thus, we need to provide explanations for the low-level continuous space and the high-level discrete structure. Furthermore, the model also seeks to explain how the internal models connect with one another. Therefore, we need to provide explanations of how the internal models are connected.

The visualization is perceived by the user on the *system interface*. The user then performs their internal cognitive process following the sensemaking model. Based on this internal process, the user interacts with the VA system interface by making changes to the visualization or by directly manipulating the *public* parameters of the model. The parameter manipulations are fed back into the AI model and the visualization interactions are interpreted and fed back to the SI model.

The SI model captures user cognition by influencing the underlying model parameters and the internal processes of the AI model. The SI model closes the interactive AI loop by communicating human cognition back to the AI model. We note that this part of our interactive AI framework builds upon a SI framework for mixed multi-level models [189], which addresses specific challenges that arise in this context. To implement this feedback process, it is necessary to define appropriate *inverse models* that can properly model the user’s cognitive process. These inverses can rely on mathematically rigorous, heuristic, or

probabilistic approaches depending on the type of interaction that we are using [97]. Finally, the changes induced by the SI model should be explained to the user, so that they can properly understand the effects of their interactions. This is handled by the interaction explanation module of the XAI components.

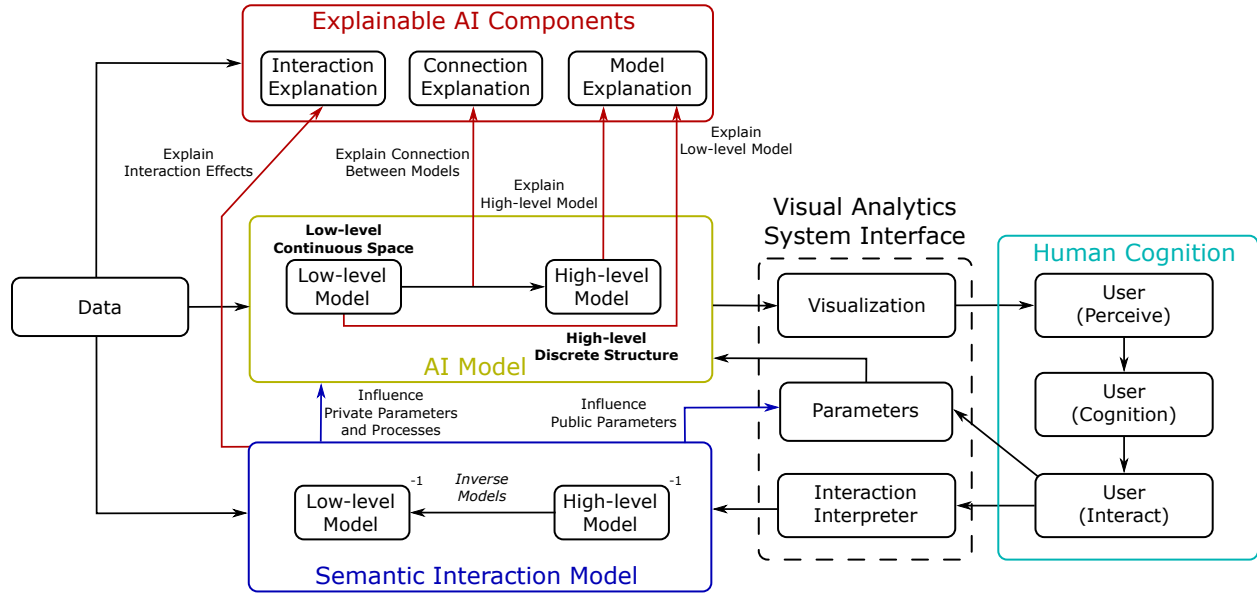


Figure 9.4: Mixed Multi-Model Interactive AI framework. Red arrows represent the “inputs” of the XAI components and what they seek to explain. Blue arrows represent the “outputs” of the semantic interaction model and point toward what they seek to influence.

9.2.2 Challenges of Interactive AI

Based on this general framework for mixed multi-model interactive AI, we identify some general challenges that arise in each key component and how we can address them.

Explainable AI Challenges

Scope Challenge. The *Scope* challenge in the XAI part of the interactive AI loop refers to determining which forward models should be explained by the XAI components, as not all models might be useful for the end user. While an argument can be made that all models should be explained for the purposes of transparency and increasing trust, in practice, the usefulness of some of these explanations depends on the technical background of the users.

For example, many internal models are too far away from the final representation to be useful for an end user. Depending on the level of technical background, a user could find it useful to dig deeper into the model, but for general use, it is unlikely that all components

are needed. Thus, in general, if it is not possible to implement explanations for all levels, then levels that are closer to the visualization should be prioritized for explanations.

In the case of our narrative sensemaking tool, we provided explanations for both the low-level projection space and high-level narrative structure space. However, during our user study, we found that most participants did not bother with the low-level representation and instead just focused on the high-level discrete structure.

Identification Challenge. The *Identification* challenge in the XAI part of the interactive AI loop refers to the difficulty of identifying the specific elements that should be explained in each model. This challenge assumes that the scope challenge has been resolved. In general, given our goal of supporting the sensemaking process, there are two strategies that can be followed to address this challenge: providing the big picture or highlighting relevant details. The first approach helps users understand the model in general, without overwhelming them with information. The second approach can provide pointers that can guide the exploration process that occurs during sensemaking.

In the context of our narrative sensemaking tool, we provided users with *big picture* explanations for both spaces. In particular, we explained the underlying clusters of events in the low-level space by identifying their relevant keywords and we provided an overview of the narrative structure by extracting storyline names that give a general glimpse of each storyline. Furthermore, we highlighted specific relevant events by selecting important events based on different metrics and generated an explanation of their importance based on these metrics.

Connection Challenge. The *Connection* challenge in the XAI part of the interactive AI loop refers to the need of addressing the explanations of the link between the low-level spaces and the high-level discrete structure. Defining how to address this challenge depends on how the internal models are connected with one another. For this particular challenge, if there are multiple low-level models, we do not recommend addressing it at all levels, but instead focusing on the last low-level space and how it transitions into the high-level discrete structure. Furthermore, depending on the case, it might be better to provide a global explanation or focus on local explanations of core discrete elements of the structure space.

In the case of our narrative sensemaking tool, we followed a local approach and explained the edges that connect the events of the narrative. These edges are built based on the topical clusters and the similarity of the embeddings in the low-level space. Thus, the first step in our explanation is to identify the type of connection (e.g., shared topics, common entities, and similar contents), which generates a specific label for that edge. Next, if the user seeks more details, they can generate an in-depth explanation that shows the topical cluster associated with each endpoint of the connection, the keywords that make the contents similar or different, and a list of named entities in common, if any. We note that these explanations are all generated in a post-hoc manner, rather than during the extraction process.

Finally, we note that an alternative approach in our narrative sensemaking case would be to attempt to explain the connection between the models by generating a global explanation of the linear programming optimization, but this would require that the user has knowledge of linear programming optimization, and it would likely be much more computationally complex.

Interaction Challenge. The *Interaction* challenge is the last one that we identified in the XAI part of the interactive AI loop. This challenge refers to the need of explaining the effects of SI on the AI model back to the user. Assuming that the SI model properly captures human cognition, this should not be explicitly necessary. However, user interactions can be interpreted in different ways, and it is not always clear what the effect should be, an issue known as *interaction ambiguity* or the “With Respect to What” problem [189, 381]. While this is a known issue in the SI context, it has not been addressed through the lens of XAI.

In this context, it would be useful to provide feedback on the actual effects of the SI so that the user can confirm that it is working as they intended or detect if they need to correct the SI model. This challenge was identified through the user study, as participants sometimes found unexpected effects from their SI, as such, our current implementation of the narrative sensemaking prototype does not address this issue beyond showing the changes in the final model. In this context, we note that simply showing the difference between the previous model and the model after SI is somewhat helpful, but this does not suffice as an explanation, as it would still be hard to understand why the changes happened. Addressing this challenge also depends on the specific SI used, as some might affect the low-level space, others might affect the structure space, and others might affect both. Thus, finding the exact way the interactions influenced the AI model is a hard task.

Semantic Interaction Challenges

Relevant Interactions Challenge. The *Relevant Interactions* challenge for the SI part of the interactive AI loop focuses on finding which user interactions should be captured and used to provide feedback to the AI model.

In general, we should focus on natural interactions with the visualization that can lead to relatively simple interpretations in the internal models. In some cases, the relevant interactions could be natural with respect to the sensemaking task, rather than the visualization itself. In this context, we suggest limiting interactions that are not natural to the visualization itself, as they likely will be more complex for the end user and require additional explanations.

For the purposes of our narrative sensemaking tool, we focused on five interactions that we found relevant through pilot studies. The first four are natural interactions for a graph-based structure: adding event nodes, removing event nodes, adding edges, and removing edges between events. The fifth interaction is not a natural interaction for the graph-based

model, but it is natural from a sensemaking perspective: grouping or clustering events that are related.

Private vs. Public Influence. The *Private vs. Public Influence* challenge for the SI part of the interactive AI loop is focused on the issue of defining how a specific SI will influence the AI model. In particular, the model can be affected by turning the user’s interaction into *public* parameter changes—that is, changing any parameter that the user can directly manipulate (e.g., updating weights in a DR model or the map size in the narrative model)—or by inducing changes in the *private* parameters or internal processes of the AI, that the user cannot directly influence in any way without SI (e.g., imposing constraints into an optimization algorithm). Both approaches might be viable, but there are some trade-offs in terms of flexibility with complexity and interpretability. Furthermore, depending on the specific AI models, only one approach might be practical.

Inducing changes to the public parameters is conceptually simpler and easier to explain to users, as long as the parameters are not overly complex (e.g., weights in a DR model). Furthermore, by allowing indirect manipulation of these public parameters through SI, the user does not have to worry about manipulating those parameters directly, providing extra cognitive focus to the sensemaking task itself. Thus, in cases where such an approach is possible, we would recommend implementing it.

However, not all models lend themselves to such manipulations and in these cases, we would need to rely on changes into the *private* parameters and inner processes of the AI models, rather than public parametric changes. This approach has the advantage of providing more flexibility, as we are not limited to changes to the public parametric interface of the model. However, this leads to higher complexity and is harder to explain to the user (which in turn leads back to the *Interaction* challenge of the XAI part of the interactive AI loop). This is the approach we took in our narrative sensemaking prototype, where we directly induced changes into the low-level space—through the use of semi-supervised learning to induce changes based on user feedback—and into the high-level structure—by imposing direct constraints into the optimization algorithm that finds the optimal narrative map.

Placement Challenge. The *Placement* challenge for the SI part of the interactive AI loop addresses the following question: How do we choose the levels at which we will handle each one of the relevant SI? Following guidelines from previous work [189], one approach is to select the level closest to the visualization whenever possible, as this will lead to a more natural interpretation of the interaction.

In our narrative sensemaking prototype, we implement most of our interactions on the high-level space by directly influencing the constraints of the extraction model. Only our clustering interaction requires the use of both spaces.

Transformation Challenge. The *Transformation* challenge refers to the SI part of the interactive AI loop and addresses the following question: How do we transform changes from the higher-level discrete structure (e.g., the structure space) back to the low-level

continuous space (e.g., the projection space)? Following guidelines from previous work [189], one approach is to minimize the instances that require explicit transformations by following a divide-and-conquer approach that separates the interactions into independent effects on each level.

In our narrative sensemaking prototype, only the clustering interaction was handled at multiple levels. In this case, we addressed this issue by separating the clustering interaction into two distinct effects. First, in the low-level space, the events should be projected relatively close to each other to represent the clustering effect. This was achieved using a semi-supervised projection algorithm. Next, in the high-level space, we imposed constraints that required the clustered events to be (weakly) connected to each other to ensure that they stayed relatively close to each other.

9.3 Interactive AI for Narrative Maps

9.3.1 Narrative Map Extraction

In this section, we present the narrative extraction pipeline, which is replicated from the implementation of the mixed multi-model SI framework for narrative maps [189]. This implementation is itself built upon the extraction algorithm for narrative maps proposed by Keith and Mitra [186], with changes to account for narrative map design guidelines [192] as well as general optimizations and simplification measures, such as regularization and post-processing to remove unneeded complexity [189].

There are two sub-phases in the narrative extraction process: extraction itself and post-processing. We focus our description on the extraction itself and omit the post-processing, which only makes simplifications and removes unnecessary clutter [189]. We note that the extraction algorithm is not a core contribution of this work, but we need to expose the key elements of this process to properly explain the components of the interactive AI framework.

The extraction sub-phase takes the event data in the form of text embeddings (e.g., BERT) and generates a 2D projection using UMAP [244], which corresponds to the low-level continuous space. Based on this projection, a series of topical clusters are obtained through hierarchical clustering with HDBSCAN [243]. Afterward, linear programming is used to find the optimal map based on content similarity (e.g., sharing keywords) and topical similarity (i.e., belonging to the same or similar clusters). The narrative map corresponds to the high-level discrete structure space [189].

Following the previous description, we note that there are two steps in the extraction sub-phase that form the core of the method: coherence computation and linear program optimization. The coherence computation step is used to measure how much sense it makes to narratively connect two events, based on content similarity and topical similarity, which

are the basic building blocks of how analysts construct narrative maps [192]. The linear programming step is focused on finding the optimal narrative structure, which relies on optimizing the previous notion of coherence over the whole graph, by maximizing the coherence of the weakest link [186, 319].

In this context, our interactive AI model allows analysts to indirectly influence these two steps by performing interactions on the visualization. Furthermore, the interactive AI model seeks to explain how these steps are performed, by explaining different elements of the visualization and its contents (e.g., explaining why two events are connected).

9.3.2 Semantic Interaction Model

Our interactive AI relies on the mixed multi-model SI framework [189] to define the SI components of the system. Following this framework, the analyst interacts with a high-level structure space—the narrative map—and these interactions are used to modify the underlying AI model, as shown in Figure 9.1. In particular, the effects of the SI model include adding constraints that affect the extraction process or changing the projection space based on semi-supervised learning.

In particular, based on previous research [192], the SI model for narrative maps must support analysts in exploratory tasks—which have an open-ended goal—and directed analysis tasks—which have a specific goal. Supporting these tasks requires the ability to correct pre-existing stories, create new stories, and influence the focus of the extracted narrative [189]. Following these requirements, five relevant user interactions are captured by the SI model: event addition, event removal, connection addition, connection removal, and clustering events (grouping).

Since this part of the interactive AI model has been evaluated in depth in previous works, we refer the interested reader to the original evaluation of this approach using a simulation-based methodology and expert feedback [189].

9.3.3 Explainable AI Components

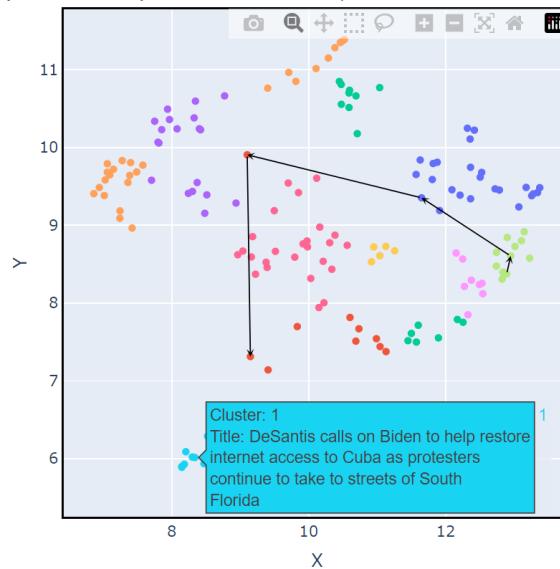
We now provide a description of the XAI components used in our system to help users understand the extracted narrative map. In particular, we add explanations to the low-level model through clustering and topic modeling. Next, we add explanations to the connection between the low-level model and the high-level model by explaining the *edges* that connect events in the high-level space, which are directly dependent on the low-level model. Finally, to explain the high-level model, we extract storyline names and important events.

Low-level Space: Topic Modeling

To explain the low-level space we use the HDBSCAN clusters found during the extraction process. In particular, we generate a keyword-based explanation to represent each cluster. To obtain these keywords, we rely on a TF-IDF representation [89] of the clusters. We show an example of the explanation generated by the system in Figure 9.5.

Document Space

Visualize the documents in a 2D projection. Documents are clustered according to their topics/similarity. The main storyline of the narrative map is shown with arrows.



Topic list

Topic 0: media (11.1%), social (11.1%), social media (11.1%), internet (10.5%), apps (10.2%), protests cuba (10.1%), desantis (9.4%), blackout (8.9%), blackout followed (8.9%), blackout followed protests (8.9%)

Topic 1: internet (14.7%), help (13.7%), desantis (12.8%), access (12.8%), internet access (11.2%), internet cuba (10.4%), provide (9.8%), biden help (9.5%), access cuba (9.4%), help restore (9.4%)

Topic 2: review (11.5%), failed (11.4%), trump (11.2%), ramps (10.3%), remittances (9.0%), failed state (8.4%), obama (8.2%), biden (7.8%), administration (7.7%), state (7.6%)

Topic 3: cuba sanctions (14.6%), sanctions (14.4%), new cuba (12.7%), new cuba sanctions (12.7%), new (12.7%), imposes (11.9%), sanctions cuban (10.4%), us (10.1%), imposes new (10.1%), imposes new cuba (10.1%)

Topic 4: embassy (13.4%), official (10.0%), white (9.0%), white house (9.0%), house (8.8%), ap source (8.7%), ap source biden (8.7%), biden looks (8.7%), biden looks increase (8.7%), cuba embassy (8.7%)

Figure 9.5: Example of topic modeling explanations for the low-level space alongside a scatter plot of the low-level space and an overlay of the main storyline going through the low-level space. We also include tooltips that appear when the user hovers over a specific event.

In particular, we first obtain a global TF-IDF representation of all documents using n -grams with n from 1 to 3 (with stop words removed). Then, we obtain the average representation for each extracted cluster, which gives us a first approximation of the relevant words for

each cluster. However, this approach considers an IDF component over the whole data set, which dilutes the importance of cluster-specific words. Thus, we further increase the weight of cluster-specific words by obtaining a local IDF term for the specific cluster [173].

Connection Explanations

To generate explanations of the connection between the low-level and high-level models, we develop three XAI models: connection label generation (i.e., classification), connection explanation generation (i.e., keywords and topics), and event comparison (i.e., explaining unconnected events).

Connection Labels. Our current narrative map model supports three types of connections: *similarity*, *entity*, and *topical*. These types of connections are based on the taxonomy defined by the narrative maps design guidelines of Keith et al. [192]. We show examples of these labels in Figure 9.2.

The first two connections are based on low-level information, namely, surface content similarity and sharing common entities, respectively. In contrast, topical connections represent a higher-level type of connection that refers to events that should be connected because they share an overarching theme or belong to the same topical clusters. We note that there are also *temporal* connections, but all connections are inherently temporal in the narrative maps model [186], thus we do not explicitly label them.

To categorize the different connections into these elements, we follow a heuristic-based approach using the components of the coherence measure. First, we evaluate whether the connection is based on topical similarity or superficial similarity. To do this, we leverage the coherence matrix used in the extraction process. The values of this matrix depend on clustering similarity (i.e., events belonging to the same topic category) and basic text similarity (i.e., the angular similarity between event embeddings).

In particular, we determine whether the main contributor to the coherence score of a specific connection is the clustering component or the basic similarity component. If more than 50% of the contribution comes from the clustering component, we assign the label “Topical” to the connection. Otherwise, the main contributor is just basic similarity and we use the label “Similarity” for the connection.

In addition to the previous labels, we consider an additional “Entity” label that can be used in conjunction with the previous labels. In particular, we use this label to represent that the connected events share some entities (e.g., names or places). To do this, we use named entity recognition [158] and check whether there are any common entities on a headline level. If there is at least one match, we assign an additional “Entity” label to the connection. Furthermore, to address partial matching cases, we consider an overlap score based on Jaccard similarity that considers the individual tokens of an entity (e.g., “President Barack Obama” and “President Obama” would have a Jaccard similarity of approximately 0.67). We use a

threshold of 0.5 for partial matches (i.e., if half of the entity tokens match we consider it a full match).

Finally, we note that there are more types of connections, such as *causal*, *domain knowledge*, and *speculative* [192] that cannot be captured by the current narrative model, as these types of connections would require an external knowledge base to store domain knowledge and an inference engine or probabilistic models to allow for causal inferences about events and their cause-effect relationships.

Connection Explanations. While the labels provide a general view of why two specific events were connected, they do not provide sufficient details. Thus, we expand upon each relevant element of the different types of connections, including a breakdown of the *topic* to which each event belongs, the common *entities* they have, and the specific contributions of each *keyword* towards similarity. The information shown depends on the specific label assigned to the connection (e.g., a connection with an “Entity” label will display the common entities).

For the topical information, we specify whether the events share common topics and the keywords that define these topics, following the same format from the explanations of the low-level space. For the common entities, we simply intersect the sets of identified entities and display them accordingly. To find the keywords with the highest contribution (in positive or negative terms), we use the SHAP library [238] with a permutation-based approach.

In particular, we adapt an existing implementation of SHAP-based explanations for text similarity ¹. This implementation allows us to find the most relevant keywords that contribute to (or reduce) similarity. We modify this implementation by generating a tornado plot instead of the default cascade plot from the SHAP library [238]. We further remove stop words from the explanations and ensure that the baseline for the model similarity is set to zero so that increasing and decreasing values are easily interpretable. Due to computational power limitations, we only consider the headlines and the first thirty words of the event to build the explanation, which should contain the most relevant information, assuming that the data we use follows the usual breaking news structure [187].

We show an example of the generated explanations in Figure 9.6 with a connection based on common topics and entities. We note that the plot displays keywords from both events in the same plot, using the notation S1-*i* to denote the *i*-th word of the first event and S2-*j* to denote the *j*-th word of the second event. Moreover, some keywords are repeated as they appear in different positions through the text (denoted by the index beside each code).

Event Comparison. Finally, the last element we consider for the XAI component focused on the connection between the low-level space and the high-level space is an explanation of why two specific events were *not* connected. We implement this through a functionality we call *event comparison*, which allows users to select any two events and generate a similar explanation to the one used with the connections. The explanation only focuses on the

¹<https://github.com/yg211/explainable-metrics>

Edge Data

Source: China pneumonia outbreak: Mystery virus probed in Wuhan

Target: China pneumonia: Sars ruled out as dozens fall ill in Wuhan

Weight: 1

Connection Explanation

This connection is based on common topical information. See extracted topics of each event below.

This connection is based on common entities. See extracted entities of each event below.

Event Topics: Source and Target share the same cluster/topic. Topic keywords and importance: wuhan (11.0%), virus (8.9%), new (8.4%), vaccine (8.3%), outbreak (8.2%), china (8.1%), china virus (8.1%), china pneumonia (6.9%), pneumonia (6.9%), anger (6.9%)

Common Entities: Wuhan, China

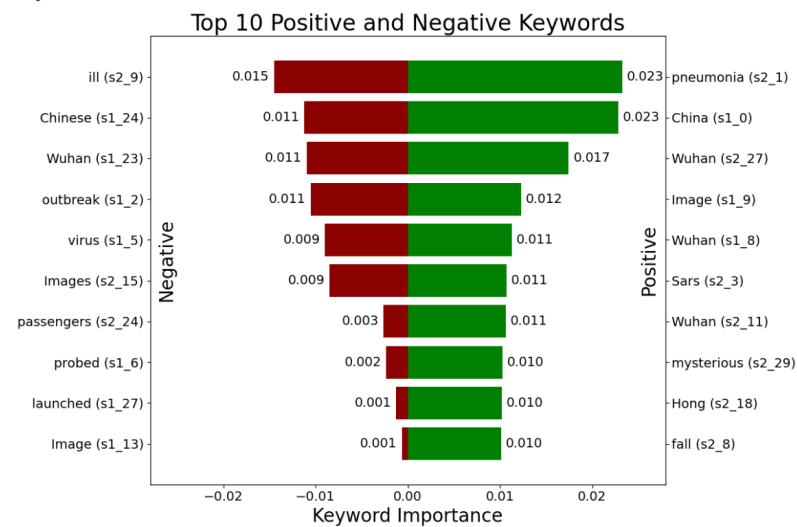
Keyword contributions to the connection

Figure 9.6: Example of an explanation for an edge between two events in the 2021 Cuban Protests data set.

keyword contributions, which provides a general idea of why two specific events were not connected.

High-level Space

To explain the elements of the high-level space, we generate two explanations. First, we extract storyline names that help the user understand the big picture of the narrative model. Next, we find the important events of the narrative model based on content and structural concerns.

Naming Storylines. We identify the different storylines of the narrative map and assign them a name. This will help users of the narrative map understand the general topic of each storyline by providing a big picture overview of the narrative model. We show examples of

automatically extracted story names in Figure 9.2.

The first step is to identify each storyline inside the narrative map. To do this, we use the recursive algorithm described in the mixed multi-model SI framework implementation [189]. The algorithm works by extracting the most coherent storyline from the narrative map and removing it. This process is repeated recursively on the left-over connected components until the graph is empty or full of isolated singleton storylines (i.e., stories with only one event).

Once the storylines have been identified, we proceed to extract potential candidates and score them. We name the stories in the order that they were identified (i.e., the main story first). We also seek to minimize overlap and redundancy in names, by penalizing names that have been selected for other storylines.

In particular, we implement a modified version of the naming method proposed by Laban and Hearst to find story names in news timelines [208]. Their approach is based on the following principles: a story name should be a maximal noun phrase that contains a proper noun (i.e., an entity), a common noun, and an abstract word (e.g., crisis, outbreak). Based on these basic rules, they proposed a series of scoring functions. We use a modified version of these rules adapted to our smaller storylines. Furthermore, we also penalize redundancy by adding an overlap criterion to avoid redundant names. Using a linear combination of these criteria, we rank the candidates and select the top candidate for each storyline, which provides an overview of potential key points and entities of the storylines. The individual weights were selected empirically to ensure storyline names that made sense.

Important Events. To provide users with additional information, we select important events based on two measures of relevance: node degree centrality (i.e., structural importance) and representativeness (i.e., content importance). We note that this approach differs from the purely structural approach of Keith and Mitra [186], which relied on the concept of antichains. Instead, this approach considers the contents of the events and the coherence of their connections.

The first step to identifying important events is to identify the different storylines of the map. Having identified the storylines, we can extract one important event from each one of them. To do this, we find the event that is closest to the *centroid* of the storyline in the embedding space. That is, we compare each event embedding in the storyline with the average embedding of the whole storyline and select the one with the highest similarity to be the important event of that story. Thus, if there are n stories, we identify n relevant events based on content. Next, we seek to identify n events that are relevant based on structural concerns. To do this, we use degree centrality as a measure of event importance based on the normalized coherence weights of the narrative map [186] and select the top n events based on centrality. We note that some events might fulfill both conditions, being both relevant because of their content and being structurally relevant due to their connections.

Thus, this extraction approach based on storyline identification and event ranking based on the centroid (content importance) and degree centrality (structural importance) criteria

allows us to model important events based on the content of the events and the general structure of the narrative map. We note that the extraction algorithm does not directly account for these notions of content and structural importance in its process, but they act as intuitive proxies for the underlying coherence values. The important events selected by this approach help provide more background for the narrative and can aid the user in their sensemaking process by providing a quick overview of the relevant information in the data. We mark these events in different colors depending on their origin: red for content-based important events, blue for structure-based important events, and purple for events that fulfill both conditions. We show examples of these elements in Figure 9.2. Moreover, we add explanations to the selected events, based on how they were selected (e.g., “this event was marked as important because it acts as a hub in the map” or “this event was marked as important because it represents its storyline”), with a special case to handle when an event fulfills both conditions.

9.4 User Study Description

9.4.1 Evaluation Methodology

Insight-based evaluation [271, 311] is a methodological approach to evaluate and compare visualization systems and techniques, which has proven useful to evaluate VA systems and visualization techniques [14, 67, 141, 271]. In contrast to traditional evaluations based on benchmark tasks (e.g., accuracy and completion time), insight-based evaluations seek to capture how analysts use a VA system and how they arrive at insights. Furthermore, insight-based evaluations allow comparing different systems in a way that reflects the practical implications of their usage as a VA system [141] and does not hinder the evaluation of exploratory strategies [14] that arise in the narrative sensemaking process [192].

To evaluate our interactive AI model, we rely on an insight-based evaluation methodology and ask users to identify as many insights as possible regarding the causes and effects of the protests and how they relate to each other, alongside any other relevant insights. In this evaluation approach, analysts identify and report the insights that they find in the data, which we then categorize and count. This open-ended task gives analysts freedom to explore the data set and provides an approximation of a realistic narrative sensemaking task. Based on this evaluation, we consider our interactive AI system to be effective if the users are able to gather more insights with it compared to the baseline.

However, the previous evaluation on its own does not give us the full picture, as we still need to consider the users’ perception of the model and its usefulness. Thus, we evaluate whether users perceived the explanations as useful and whether they believe that the SI model is capturing their intent. In particular, we ask participants to complete a series of follow-up questionnaires regarding their perception of the different components of the

system. Furthermore, at the end of the session, we ask participants to give us feedback on the system and the underlying models. This second evaluation helps us gauge whether the model is properly capturing the user’s cognitive process and learning what the user expects it to learn. Finally, we also perform an in-depth quantitative and qualitative evaluation of the user strategies and behavior, beyond their insights and responses to the follow-up surveys (e.g., reading behavior and parametric diversity). This evaluation lets us understand how users are applying the different elements of the system and how the affordances of the models influence their behavior.

9.4.2 Evaluation Data

To show the effectiveness of our interactive AI model, we need to use a data set with the appropriate complexity level for a sensemaking task. In particular, we use news data covering the 2021 Cuban Protests that occurred in July 2021—the biggest protests in decades in Cuba [302]—to design our evaluation tasks. We collected a data set of 160 online news articles scraped from multiple news sources from all sides of the political spectrum. The 2021 Cuban Protests provide an appropriate level of difficulty to evaluate the capabilities of our interactive AI model due to their complex nature [189].

9.4.3 Participants

We recruited a total of 36 participants from a pool of graduate and undergraduate students registered in a Human-Computer Interaction email list. In particular, our recruits were comprised of mostly undergraduate Computer Science students, but we were also able to recruit Communications, National Security, and Engineering majors. Most participants were male students (64%). The participants were distributed throughout three conditions: a timeline-based prototype with search functionalities used as a baseline (TL), a basic narrative map prototype without interactive AI features (BM), and a narrative map prototype with interactive AI capabilities (IM). One participant dropped out from the study, another participant was discarded due to failing attention check questions, and another participant was discarded due to not following instructions and refusing to use the narrative map, instead brute-forcing the task. The final distribution of participants was 10 for the TL condition, 13 for the BM condition, and 10 for the IM condition. Participants were anonymized and identified with codes (S_i , B_i , and I_i for the i -th participant of the TL, BM, and IM condition, respectively).

9.4.4 Evaluation Tasks

All participants worked on the same tasks. We provided participants with a brief training session with pre-recorded videos before doing the task. The pre-recorded videos showed the

functionalities of the prototype with a data set about COVID-19 news. The training session lasted 15 minutes. We also allowed participants to ask as many questions as needed about the task and prototype throughout the session. After the tutorial, participants were asked to fill out a pre-task survey asking about familiarity with the topic.

After completing the tutorial, participants were given 1 hour to complete the task. The task consisted of two parts: two warm-up questions with multiple choices and an open-ended analysis question. The warm-up questions required finding two specific pieces of information: the exact date the protests started and who was the president of the country at the time. The analysis question asked participants to find the causes and effects of the protests and to report as many insights as possible. Furthermore, to provide participants with the proper context before doing the tasks, we also included a descriptive prompt that contextualized their analysis as part of a journalistic analysis assignment. Participants were given as many sheets of paper as they needed to write down their insights. Participants worked on the task until they ran out of time or they decided that they were done and could not find any more insights.

Furthermore, we used a think-aloud protocol to help us understand the thought process of the participants and all sessions were recorded (screen and voice). We also recorded the number of executions (search queries for the TL condition and map generations for the map-based conditions), and the number of events read (events clicked with the event details tab open, regardless of time spent on them). We also recorded specific search queries in the TL condition and the keywords used in the *Find in Map* function for the map-based conditions. Moreover, we kept track of the parameter combinations and start events used by participants in the map-based conditions.

After completing the task, we had a feedback session of 15 minutes, where we asked questions about the elements of the prototype. For participants with the TL condition, we also gave them a quick tour of the map-based prototypes to gather additional feedback. Finally, participants were asked to fill out a post-task survey, which included an assessment of their familiarity with the topic after completing the task and a series of system evaluation questions. For participants with map-based conditions, we also asked them to complete a series of questions about the map elements. Finally, participants with the IM condition had to complete a series of questions about the interactive AI components.

9.4.5 Insight Coding

Finally, as part of our insight-based evaluation [271], we tagged all the insights provided by the participants, including spoken insights (taken from the transcripts of the sessions), written insights (taken from their reports), and implicit insights (taken from their interactions, such as creating a cluster). Throughout this process, we used *open coding* [196] to categorize these insights. We reviewed our tags with two experts in VA working in the intelligence analysis domain.

In total, we found 180 individual insights. Generally, these are *low-level insights* that represent specific facts or details identified by the participants (e.g., the government lifted import restrictions after the protests), but also include inferences or comparisons (e.g., comparing the response of the Cuban government to the U.S. response to BLM protests). Moreover, the identified insights sometimes overlap, with some insights providing a deeper level of detail for the same specific element (e.g., the U.S. trying to provide internet to Cubans vs. the U.S. working with the private sector to provide internet to Cubans). While these could be grouped together into a single insight, we opt to count them separately due to them providing different levels of detail. For the purposes of our analysis, we refer to all these insights as low-level or detailed insights, as they represent the full range of insights provided by the participants.

We note that there are varying levels of abstraction and detail in the low-level insights provided by the participants (e.g., there is an economic crisis vs. people turning to black markets). These two insights are part of the same topic (e.g., people turning to the black market is a symptom of the economic crisis). Thus, once we had collected all the insights, we grouped them together into *high-level insights* that represented broad topics or themes identified by the participants (e.g., economic issues or political discontent). This analysis led to 26 different high-level insights that can be roughly grouped into causes, effects, and context of the 2021 Cuban Protests. We present the high-level insights in the coding tree of Figure 9.7, which provides an overview of our hierarchical coding scheme. We do not present the low-level insights in this tree due to space limitations, but they would be the next level of the coding tree after the high-level insights. The full table of insights is provided as supplementary material in Appendix A.

9.5 Results

9.5.1 Insight-based evaluation

We start with the insight-based evaluation at two levels: high-level (or topical) insights and low-level (or detailed) insights. We show our results in Figure 9.8.

High-level insights

We show the distribution of high-level insights in Figure 9.8(a). We performed a one-way ANOVA on the number of high-level insights obtained by each participant. According to the ANOVA, there is a significant difference in the number of high-level insights by group ($p = 0.00117$). Tukey’s test shows that a significant difference occurs between the IM condition with the TL condition ($p = 0.00078$). There is no significant difference between the BM condition and the other conditions. However, we note that the difference between

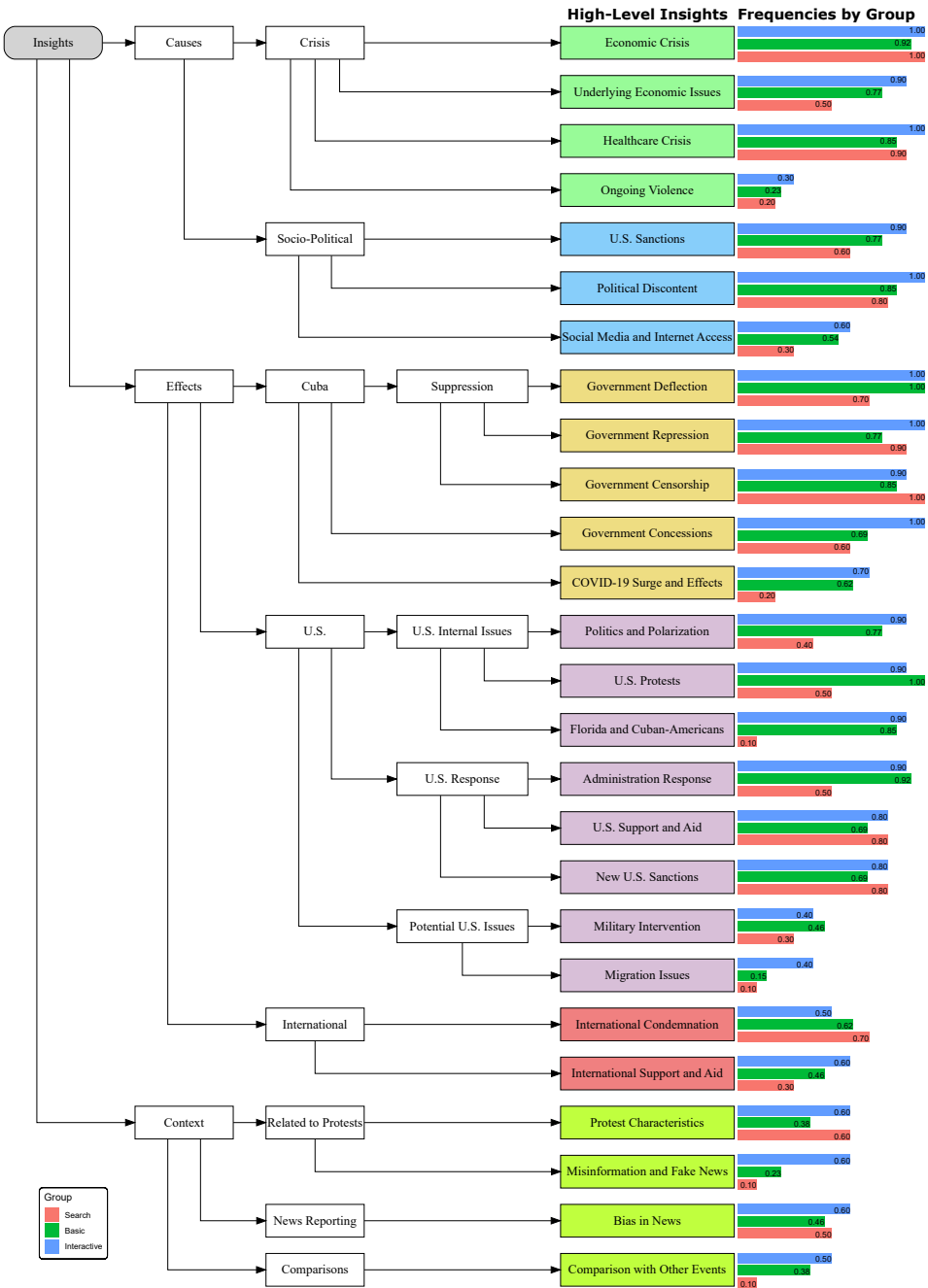


Figure 9.7: Coding tree showing our hierarchical coding scheme for insights. The leaves of the coding tree represent the high-level insights and include a histogram showing the frequencies of each high-level insight by group.

the number of elements in the BM condition and the TL condition could be considered marginally significant, with a p -value of 0.055. This analysis suggests that, in the context of our narrative sensemaking setting, the IM condition provides better sensemaking support to

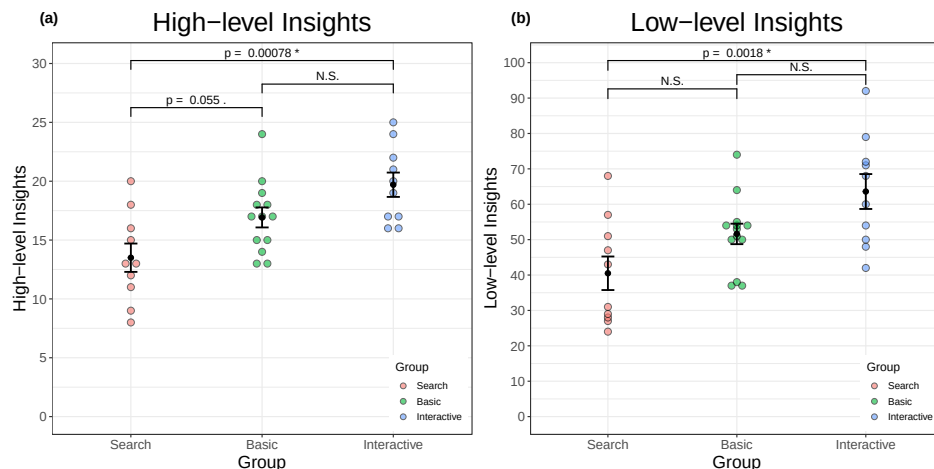


Figure 9.8: **(a)** High-level (topic) insights results. **(b)** Low-level (detailed) insights results. (N.S.) denotes that there is no significant difference between the groups ($p \geq 0.1$). (.) Denotes a marginally significant difference ($p \in [0.05, 0.1)$). (*) denotes a significant difference ($p < 0.05$).

analysts in terms of high-level or topical insights when compared to the baseline.

Next, we consider the individual high-level insights and their distribution across the different groups, as shown in Figure 9.7. Performing a one-way ANOVA we find that there is a significant difference ($p = 0.00472$) in the average frequencies of each group, with the IM condition having the highest average frequency (0.76) and a statistically significant difference to the TL condition (0.52), confirmed through a Tukey’s test ($p = 0.0032$). As before, the BM condition lies in the middle (0.65) with no significant difference with either group.

The general trend is for the map-based conditions to have a higher frequency in each high-level insight. This pattern holds for most insights, with some having drastic differences between the map-based groups and the TL baseline (e.g., Florida and Cuban-Americans) and others having a relatively uniform distribution (e.g., Economic Crisis). However, there is one outlier that presents a different behavior: the insight related to “International Condemnation”, which sees a reverse trend, with the TL condition having the highest frequency (0.70) compared to the BM condition (0.62 and IM conditions (0.50).

We credit this difference to one specific event, which discusses the protests in Cuba from a “Diplomatic Standpoint” and is found on the second page of the default timeline generated by the TL prototype when no query is given. 90% of the participants with the TL prototype accessed this event, while only 15% did so for the BM condition and 30% for the IM condition. We note that there are other events that contain information about international condemnation, which participants in the map-based prototype accessed, but this specific event contained a detailed breakdown of the international effects that could lead to insights related to this topic. Thus, as this event was relatively easy to find in the TL prototype

compared to the map-based prototypes, it makes sense for this specific insight to be higher for the TL prototype. In contrast, we note that the similar “International Support and Aid” insight followed the regular trend of being higher for the map-based conditions, which makes sense in the light of the contents of the aforementioned article focusing mostly on the negative sides of the international response, rather than the support from allied countries or general humanitarian aid from abroad.

Low-level insights

We show the distribution of the low-level or detailed insights in Figure 9.8(b). We performed a one-way ANOVA on the number of low-level insights obtained by each participant.

According to the ANOVA, there is a significant difference in the number of low-level insights by group ($p = 0.00267$). Tukey’s test shows that a significant difference occurs between the IM condition with the TL condition ($p = 0.0018$). These results concur with the previous analysis of high-level insights, with only slight differences in the statistical significance values. Furthermore, as before, there is no significant difference between the BM condition and the other conditions. Unlike before, the difference between the number of insights in the BM condition and the TL condition is not close to significance ($p = 0.1064$). This analysis suggests that the IM condition provides better sensemaking support at a detailed level and further confirms the capabilities of our interactive AI model to support sensemaking.

9.5.2 Quantitative Analysis

Effectiveness

We define effectiveness as the fraction of gap closed by using the system (measured as a percentage), following the methodology of Shahaf et al. [319, 321] for metro maps and story chains. In particular, let F_b be the familiarity of the participant before doing the task and F_a the familiarity after completing the task, both defined on a scale from 1 to 5. Perceived effectiveness is then defined as

$$E = \frac{F_a - F_b}{5 - F_b}.$$

For example, someone that has $F_b = 2$ at the start and goes to $F_a = 4$ has a 67% effectiveness, as they only closed 2 points of their 3-point gap with the perfect familiarity score.

The average effectiveness of the TL prototype is 0.68, for the BM condition it is 0.79, and for the IM condition, it is 0.83. The results suggest that there might be an advantage in effectiveness in favor of the map-based prototypes. The results are shown in Figure 9.9(a). However, an ANOVA shows that there is no significant difference ($p = 0.0703$). Nevertheless,

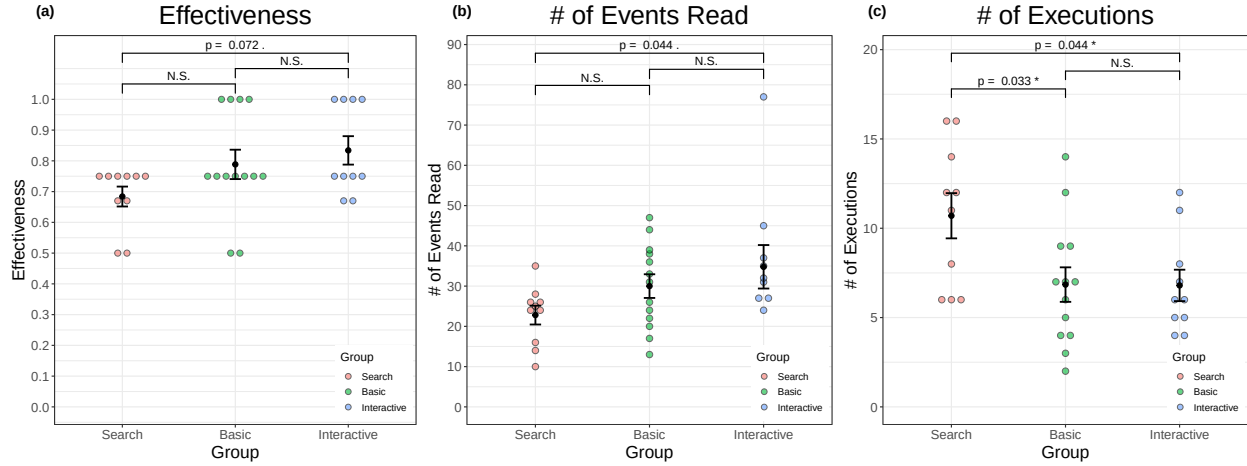


Figure 9.9: **(a)** Perceived effectiveness. **(b)** Number of events read. **(c)** Number of executions. (N.S.) denotes that there is no significant difference between the groups ($p \geq 0.1$). (.) Denotes a marginally significant difference ($p \in [0.05, 0.1)$). (*) denotes a significant difference ($p < 0.05$).

we note that the difference between the IM prototype and the TL prototype is marginally significant ($p = 0.072$).

Number of Events Read

We also check the number of events read in each group. The results are shown in Figure 9.9(b). The average for the TL prototype is 22.80, for the BM condition it is 30.00, and for the IM condition, it is 34.80. This would suggest that the map-based prototypes may lead to participants reading more events in detail (or at least skimming them). However, an ANOVA shows that there is no statistically significant difference between groups, although it could be considered to be trending toward significance ($p = 0.103$).

Number of Executions

The difference in the number of (valid) executions is statistically significant between groups. In particular, both map-based prototypes have a lower number of executions compared to the TL prototype. We show the results in Figure 9.9(c).

An ANOVA shows that there is a significant difference between the groups ($p = 0.0217$). The follow-up Tukey's tests show that the TL prototype is different from both map-based prototypes. Thus, on average, participants with the TL prototype required more executions (i.e., more timeline generations) to solve the task compared to the participants that used the map-based prototypes. There is no difference between the map-based prototypes.

Usability and Narrative Representation

We show the tallies of the answers to our survey questions on usability and narrative representation in Figure 9.10. We find that there is a significant difference according to a MANOVA on the average Likert scores by group ($p = 0.01774$). A follow-up analysis shows that the differences are in perceived *Usefulness*, *Ease Of Use*, and *Reuse*. There is no significant difference for the narrative representation questions. All models are perceived the same in terms of their capability to represent the narrative.

Items	Search						Responses Basic						Interactive					
	Strongly Disagree					M	Strongly Disagree					M	Strongly Disagree					M
	1	2	3	4	5		1	2	3	4	5		1	2	3	4	5	
	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree		Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree		Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	
Usefulness	0	0	0	9	1	4.10	0	0	0	9	4	4.31	0	0	0	2	8	4.80
Ease of Use (*)	0	0	1	3	6	4.50	0	0	0	13	0	4.00	0	1	2	7	0	3.60
Reuse	0	0	2	5	3	4.10	0	0	2	4	7	4.38	0	0	0	1	9	4.90
Representation	0	0	2	5	3	4.10	0	0	3	8	2	3.92	0	0	0	8	2	4.20
Redundancy (*)	0	5	2	2	1	2.90	1	3	6	3	0	2.85	0	1	2	7	0	3.60
Relevance (*)	0	0	2	5	3	4.10	0	1	0	7	5	4.23	0	0	0	5	5	4.50
Depth	0	0	1	7	2	4.10	0	0	2	6	5	4.23	0	0	0	8	2	4.20

Figure 9.10: A tally of the answers to our survey questions on *usability* and *representation* of the narrative across all three prototypes.

We note that if we had aggregated the usability scores into a single metric there would be no significant difference, as the scores cancel themselves out when averaged. However, there are some interesting takeaways when looking at the specific questions. The IM prototype has the highest *Usefulness* and *Reuse* ratings compared to the other prototypes and the TL prototype had the lowest scores. Furthermore, map-based prototypes are harder to use than the TL prototype, and, in particular, the IM condition has the worst rating in terms of *Ease of Use*.

Narrative map elements

Next, we evaluate the questions about map elements for the two map-based conditions. We show the results in Figure 9.11. In particular, we use a MANOVA on the average Likert scores. However, we find that there are no statistically significant differences ($p = 0.5297$).

		Responses											
		Basic						Interactive					
		Strongly Disagree					M	Strongly Disagree					M
		Disagree	Neutral	Agree	Strongly Agree	Disagree		Neutral	Agree	Strongly Agree			
		1	2	3	4	5		1	2	3	4	5	
Items	Coverage	0	0	3	7	3	4.00	0	0	1	7	2	4.10
	Layout	0	0	1	6	6	4.38	0	0	2	6	2	4.00
	Important Events Relevance	0	1	2	7	3	3.92	0	0	1	5	4	4.30
	Important Events Usefulness	0	1	1	8	3	4.00	0	0	2	4	4	4.20
	Main Story Relevance	0	0	2	5	6	4.31	0	1	1	4	4	4.10
	Main Story Usefulness	0	0	1	7	5	4.31	0	0	3	5	2	3.90
	Storylines	0	0	1	8	4	4.23	0	0	1	4	5	4.40

Figure 9.11: A tally of the answers to our survey questions on the *narrative map elements* of the BM prototype and the interactive AI prototype.

Interactive AI elements

We show the results of the user evaluation of XAI and SI with Likert scales in Figure 9.12. Note that these do not have comparable results in the BM or TL results.

Regarding SI, we note that most participants found the interactions and the changes made by the map useful ($M = 4.05$). Most participants found that the system learned from their interactions ($M = 4.2$). However, the ability to learn only slightly increased trust in the system ($M = 3.8$).

Regarding the XAI components, we note that explanations were considered useful in general ($M = 4.3$) and that they increased trust in the system (4.5), even more than the ability to learn from user feedback. This difference is statistically significant (paired t-test, $p = 0.02485$).

Storyline names were the most controversial element ($M = 3.5, SD = 1.10$). Opinions ranged from them being the most useful element of the map, to being useful despite their issues, to being distracting and useless. Paraphrased from a participant: “Storyline names are not always correct, but they are relevant and useful, they provide pointers.”

Regarding connection labels (e.g., similarity and topical connections), they were mostly considered correct, but not necessarily useful. In fact, a participant mentioned that they were distracting and would make more sense to show if they were seeking to explain a specific edge.

Regarding detailed explanations of the connections, they were considered useful, even if they were not used in depth. Some participants used them to confirm that a specific connection

was superficial and could be deleted (e.g., only based on the keyword Cuba). Thus, rather than explaining why something was connected, they were more useful for finding superficial connections. The event comparison function was mostly unused or used in unintended ways (e.g., to generate edge explanations), although participants considered it potentially useful (e.g., comparing bias between news outlets).

		Responses					
		Strongly Disagree Disagree Neutral Agree Strongly Agree					
		1	2	3	4	5	M
SI Items	Interaction Relevance	0	0	1	9	0	3.90
	Learning Model	0	0	0	8	2	4.20
	Usefulness	0	0	1	6	3	4.20
	Model Trust	0	0	4	4	2	3.80
XAI Items	Storyline Names Correctness	1	4	1	2	2	3.00
	Storyline Names Relevance	1	0	4	1	4	3.70
	Explanations Trust	0	0	1	3	6	4.50
	Explanations Usefulness	0	0	1	5	4	4.30
	Connection Labels Correctness	0	0	2	6	2	4.00
	Connection Labels Usefulness	0	1	3	5	1	3.60
	Storyline Names Usefulness	0	1	2	5	2	3.80
	Connection Explanations	0	0	1	7	2	4.10
	Event Comparison	0	0	4	4	2	3.80

Figure 9.12: A tally of the answers to our survey questions on the *semantic interaction* components and *explainable AI* components.

9.5.3 Qualitative Analysis

Strategies

For the TL prototype, there was not much variety, which makes sense given the limited functionalities of the TL prototype. In particular, there were two strategies: normal search and brute force. Participants who used brute force (S_1 and S_7) had limited usage of the search functionality, opting to just analyze the timeline without search queries. S_1 generated a timeline with as many events as possible. S_7 did not increase the number of events displayed, but opened a second window and used them to access information more quickly than with a single window. Participants who used normal search relied on keywords, although the quality of these keywords varied across users (e.g. S_2 had very low-quality queries compared to S_5 , S_9 or S_{10}).

For the map-based prototypes, we identified several different strategies. We identified strategies in two phases, as strategies usually shifted as participants learned how to use the system (e.g., starting with a focus on a single map and then shifting to parametric exploration, or vice versa). Notable strategies are shown in Table 9.1.

Table 9.1: General strategies used by participants with map-based prototypes.

Strategy	Description
Single Map	Participants focused on a single map trying to extract as much information as possible from this singular representation. Depending on the participant, the exploration of the map focused on specific elements (e.g., highlights or main story).
Comprehensive Timeline	A subtype of single map focus, where the participants generated a large timeline-like map (i.e., a long main storyline with only a few side stories).
Parametric Exploration	Participants generated several maps with different parameters (or starting events). Instead of the single map strategy, the participants did not stay too long on each map and tried to generate many maps to extract as much information as possible.
Limited Parametric Exploration	This is a subtype of parametric exploration where the participants only generated a few maps. This is an intermediate type between full-blown parametric exploration and the single map strategy.
Parametric Exploration with Corrections	This is a subtype of parametric exploration where the participant tried to correct the map. However, for the BM prototype, this was not really useful because it does not have any capabilities to learn from the interactions. Nevertheless, interacting more in-depth with the events and map could lead to a better understanding of the narrative.
Incremental Maps	This is a subtype of parametric exploration focused on iteratively increasing map size, starting from a small map that only provides a brief overview and slowly turning it into a more complex map with many details.
Sliding Window	This is a subtype of parametric exploration focused on iteratively changing the starting event (and other parameters as needed) in the manner of a sliding window (e.g., start with event 0, then 10, then 20, and so on).
Increasing Coverage	This is a subtype of parametric exploration focused on iteratively increasing map coverage, starting with a map with low coverage and then increasing it to cover more topics.
Parametric Exploration with Keyword Search	This is a subtype of parametric exploration that complements the strategy with keyword search to explore the unselected events of the map. The narrative maps generated through parametric exploration are used as a tool to generate keywords to search the data set.
Headline Exploration	This is an even more specific type of parametric exploration with keyword search, where the focus was solely on headlines, without actually reading the events until after the big picture was figured out.

For the IM prototype, in addition to single map focus and parametric exploration, we found some unique strategies. We describe these strategies and also categorize them based on how they use SI and parametric exploration. In particular, we found two approaches for SI: corrective and additive. We show these strategies in Table 9.2.

Corrective SI focused on the existing map, making changes to pre-existing elements of the map (e.g., clustering existing events, adding edges between existing events, or removing events). Additive SI focused on attaching new elements to the map (e.g., adding events, connecting those events, or clustering them). Finally, we note that both of these strategies can be used in a single-use or cumulative manner. In cumulative strategies, the changes are stacked through multiple iterations. In contrast, single-use strategies perform a single SI and keep its effects until either the end of the task or a full reset, without building upon them. We note that single-use and cumulative strategies can be mixed throughout different “resets”.

Table 9.2: Semantic interaction strategies used by participants. Column codes: **C** for *Corrective SI*, **A** for *Additive SI*, and **P** for *Parametric Exploration*.

Strategy	C	A	P	Type	Description
Removing Bias	×			Cumulative	The participant started with corrective SI, then decided to remove right-wing bias and clickbait from the map (e.g., Breitbart). This is similar to our simulation example.
Topical Clusters		×		Single-use	The participant tried to add information to the map by defining topical clusters before using the map. This was a single-use semantic interaction, as it was not built upon but kept throughout the map.
Iterative Refinement	×	×		Cumulative	The participant tried to refine the map through different interactions in a cumulative manner. The refinements continued until the system failed and were re-started with different interactions after that. This strategy did not include parametric exploration.
Resetting and Small Clusters	×	×	×	Mixed	The participant tried to improve the map through different interactions, mostly clustering. The participant reset the map after testing a cluster (usually after a single set of semantic interactions, rather than in a cumulative manner). This strategy included parametric exploration.
System as a Playground	×	×	×	Cumulative	The participant tried to use all possible interactions to push the limits of the system, including all types of semantic interactions in a cumulative manner and parametric exploration. We note that the participant only did this after solving the task through traditional parametric exploration.

Reading space

We now compare the reading distributions of the groups (number of users that read each event). We show the event histogram in Figure 9.13 for each group. It can be observed that the distribution of the TL group is skewed towards events later in the timeline, while the other groups are mostly uniform. To confirm this, we perform pairwise comparisons between reading distributions. Both map-based prototypes ($p = 0.996$) have a similar reading distribution (which could be described as roughly uniform) but differ ($p < 0.005$) from the TL reading distribution. This difference in distribution could be caused due to the tendency of participants with the TL prototype to focus on the first two pages of results, which usually contain events from later in the timeline by design. In contrast, participants that used the maps tended to cover the data set in a more uniform manner in temporal terms. Thus, these results imply that timeline methods with search lead to limited exploration by the users compared to the map-based methods which provide a more holistic view of events.

Parameter space

From a qualitative standpoint, participants that had the IM prototype had less parametric diversity compared to the BM prototype in terms of its public parameters: *size*, *coverage*, and *temporal sensitivity*. These results are expected, as participants with the IM prototype had access to SI capabilities and may not have needed to make many changes to the parameters to find a good map. To confirm our observation of less diversity in parameters in a quantitative

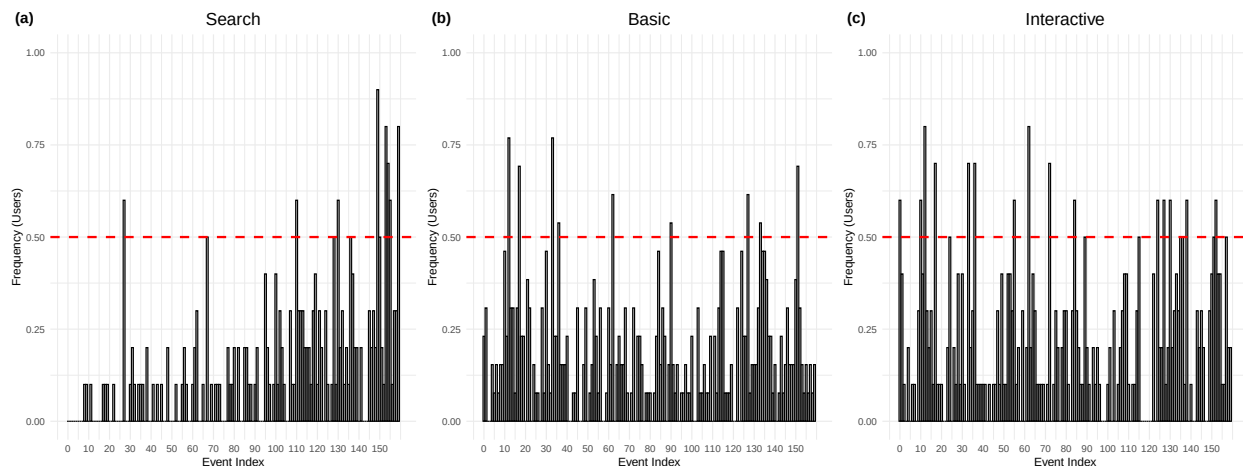


Figure 9.13: Histograms representing the number of users that read each event for each condition: (a) Timeline with Search (b) Basic Map (c) Interactive Map. The red line represents what an ideal uniform distribution would look like.

way, we use the Gini Coefficient to measure the diversity of the data [262]. In particular, for size, the BM condition had a Gini of 0.25, and the IM condition had a Gini of 0.19. For coverage, it was 0.29 for the BM condition and 0.25 for the IM condition. Finally, for temporal sensitivity, it was 0.22 for the BM condition and 0.19 for the IM condition. All these differences in diversity are significant based on Gini confidence intervals (with $\alpha = 0.01$). We show a visualization of these results through Lorenz curves [129] in Figure 9.14.

We note that despite the fact that users with the IM prototype had less *direct interactions* with the public parameters, they still generated a similar number of maps compared to their BM condition counterparts—as shown by the number of valid executions in Figure 9.9(c). This behavior implies that while they did not explicitly change public parameters, they *indirectly* modified the underlying model and its private parameters through interactions, as otherwise, they would not have needed to generate a similar number of maps.

Search queries

We briefly discuss the different search queries (i.e., keywords) used by the participants. For the TL prototype, these queries represent the main interaction they have with the system. For the map-based prototypes, these queries are used to explore the map and background events and function akin to a typical “find” function in a document.

Queries could be roughly categorized into eight types: general (e.g., effects, causes, crisis), economic (e.g., economy, market), infrastructure (e.g., power outages, internet), media (e.g., Fox News, social media), medical (e.g., COVID), political (e.g., Trump, Obama, communism), geographical (e.g., worldwide, United States), temporal (e.g., start, July), or

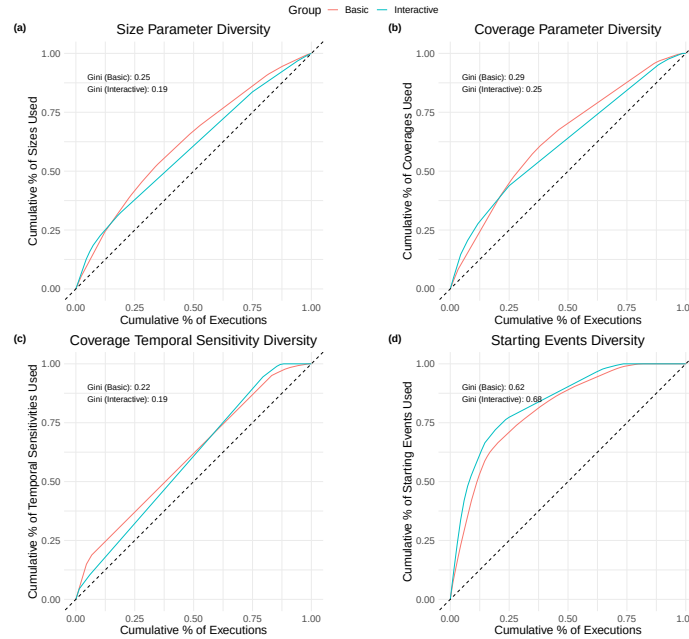


Figure 9.14: Lorenz curves describing the diversity of the parameter combinations used by the participants for the BM and IM prototypes. (a) Size (b) Coverage (c) Temporal Sensitivity (d) Starting Event choice.

repression-related (e.g., human rights, mercenaries).

We note that participants who worked with the TL prototype presented a higher diversity of keywords on average. This makes sense, as their tool required them to come up with queries to properly solve the tasks. In most cases, participants with the map-based prototypes did not rely on search to perform their tasks. In particular, we highlight the usage of B_9 and B_{13} to complement their parametric exploration with keywords. The rest of the participants with map-based prototypes had limited usage of the *Find in Map* function.

In general, we can identify some classes of keywords (e.g., economic, medical, political). We note that only participants with the map-based prototypes searched for Medical keywords, usually related to the context of the COVID-19 pandemic, one of the causes of the protests.

9.5.4 User Feedback

Finally, we present our analysis of user feedback.

Table 9.3: Keywords used by participants in the TL prototype and the map-based prototypes.

Class	TL Prototype	Map-based Prototypes
General	effects, protest, causes, demonstration, results, crisis, aid, help	effects, protest, causes, history, conflict, spread, flooded
Economic	economy, market, stock, business, inflation, economic, sanctions, embargo, restrictions, import	sanctions, embargo, trade, economy
Infrastructure	power outages, internet, internet access	internet
Media	Fox News	social, media
Medical	N/A	coronavirus, COVID, medicine, vaccine
Political	president, Miguel Díaz-Canel, communism, socialism, leftist, Biden, Obama, China, Díaz-Canel, government change, Cuban government, Trump sanctions	Biden, Trump, DeSantis, president, Raul, Castro, freedom, Castro, political
Geographical	geography, worldwide, United States	Florida, Miami
Temporal	start, July, date, began, April	start, July
Repression	rights, human right, arrests, peaceful, violent, Cuban response, dead	death, mercenaries

Functionalities

We now discuss functionalities proposed by the participants, which are categorized into five types: extraction, support, explanation, model guidance, and specific interactions.

Extraction functionalities refer to suggestions that would require changing either the internal narrative map structure through extra pre-processing (e.g., grouping events), the extraction algorithm (e.g., online map generation), or simplifying the output (e.g., turning the map into a timeline). Furthermore, participants also required support for extraction based on other forms of *seeding*, such as providing multiple starts, multiple endings, or a fixed event in the middle to generate the map.

Support functionalities refer to changes that would provide additional information or help to the users without modifying the narrative map or its contents. Examples of these functionalities include providing summaries of individual articles and summaries of specific days (or any relevant time unit). Furthermore, one participant (B_6) suggested providing an automatic choice for a starting event or recommendations for this choice would be useful. Finally, the last two support functionalities suggested by users focus on providing extra information to help the users make decisions, such as fact-checking information for an event or source information (e.g., credibility and bias).

Explanation functionalities are a subtype of supportive functionalities that would require the usage of XAI to implement. Some of the suggested functionalities by participants with the BM condition are covered by our XAI model (e.g., providing connection explanations and a topic modeling overview). Other suggested functionalities include storyline-level explana-

Table 9.4: Suggested functionalities by the participants of the user study.

Type of Functionality	Functionalities
Extraction	Grouping Events into Higher-level Structures Simplified Timeline View Online Map Generation Different Seeding Support
Support	Article Summarization Temporal Summarization (e.g., day summary) Automatic Choice of Starting Event Recommendations for Starting Events Fact Checking Information Additional Source Information (credibility and bias)
Explanation	Storyline Explanations Connection Explanations Event Comparison to Evaluate Bias Topic Modeling Overview
Model Guidance	Interaction-based Guidance Query-based Guidance Relevance-based Guidance (event weights) Spatialization-based Guidance Negative Clustering (group and delete) Storyline Name Changes Keep Element Interaction
Specific Interactions	Note Taking Highlighting Quote Extraction Mass Removal from Map (e.g., by source) Add Event to Multiple Clusters Remove Storyline

tions (e.g., explaining why all the events in a storyline are connected, rather than local-level explanations), using event comparison as a tool to evaluate the bias of different articles (e.g., right-leaning vs left-leaning news sources).

Model guidance functionalities refer to suggestions on SI or similar capabilities to guide the internal model. For example, one participant (B_{14}) from the BM condition suggested the idea of influencing the map from basic interactions, such as removing or adding events, which corresponds to our SI model. Another participant (B_{13}) suggested a query-based approach, where events would be emphasized based on a user-given query. Another user (I_2) suggested relevance-based guidance using custom event weights on top of the pre-existing SI model to emphasize specific elements of the narrative. Two participants (I_2 and I_8) tried to use a spatialization-based approach [52] to guide the model by placing groups of related events close together. However, the current SI model does not support such interactions

(due to the distances not having a direct meaning in the discrete narrative space). Finally, other participants suggested specific SI, such as *negative clustering* (grouping events and ensuring that they and related events did not appear in the map), *storyline name changes* that influence the contents of the story, and a *keep element* interaction that instructs the model to preserve a specific element.

Beyond model guidance suggestions, participants also suggested basic interactions. We note that these interactions could be turned into SI by defining an appropriate way to induce changes in the AI model. In particular, participants suggested the ability to *take notes* and save them in the system. Furthermore, participants also wanted to *highlight text*, either in the map or in the event details tab. A similar suggestion was the ability to *extract quotes* from the text and save them for later use. Other suggestions include the ability to perform *mass removals* based on specific criteria (e.g., by source), adding events to *multiple clusters* at the same time (unlike the current model that only supports one cluster), and *removing storylines* from the map, rather than individual events or edges.

Issues

We also identified issues with the current system, either by observing specific interactions that participants attempted to do or by recording verbalized observations by the participants. We discuss the most important ones here.

The most common complaint or issue noted by participants was the map having layout problems (e.g., long temporal jumps and overlapping storylines). This could be solved by tuning the parameters of the layout engine, but it would require intensive testing to optimize properly.

The second most common complaint was unrelated to the prototype itself but rather was caused by the constraints of the user study and our task. In particular, participants noted that there was not enough time to use everything, while this could be explained by the complexity of the task and time constraints, it also points towards the general complexity of the system.

Another relatively frequent issue was that the main storyline was shorter in maps with isolated sub-graphs or large maps with lots of storylines. This was caused by small storylines having a higher likelihood (due to less multiplications to dilute their weight). It could be possible to fix this by including additional constraints in the algorithm that finds the main storyline.

Furthermore, some participants had some issues understanding the priorities of the extraction algorithm, which could be helped by creating better XAI components that should help clarify the priorities of the extraction algorithm and reduce general confusion about results.

Finally, some participants requested better error reporting for the extraction process, as

the current prototype does not provide in-depth reports about the failures in the linear programming optimization (e.g., no solution due to conflicting user-imposed constraints). We note that some of these errors arose due to unexpected usage patterns, such as using events at the end of the data set as starting events or intentionally testing the limits of the system.

We show all the issues in Table 9.5 and discuss the most important ones below.

Table 9.5: Issues and complaints obtained through user feedback and observation in during the user study.

Type of Issue	Issues
System and Task Complexity	Complex and Confusing Parameters
	Time Constraints
	Confusion about Results without XAI
	Confusion about Starting Events as Data Sets
	Confusion about Starting Event and Constraints
AI Model and XAI Issues	Scared or Uncomfortable to Modify Map
	Unclear Priorities of Algorithm
Interactive Visualization Issues	Explanations Contaminated by Frequent Words
	Layout Issues (temporal jumps and overlap)
	Main Story Issues in Large and Isolated Maps
System Failure and Unexpected Behavior	Interactions are Meaningless without SI
	Deleting or Changing Seed Events and Semantic Interaction
	Lack of Support for Ending Events
	Testing the Limits of the System

9.6 Discussion

We now discuss our results and the limitations of our methods.

9.6.1 Narrative Sensemaking

Our insight-based analysis provides an in-depth look into how narrative maps can aid in the sensemaking process, especially when imbued with interactive AI capabilities. In particular, the map-based prototypes enable analysts to explore the data set through a mix of diverse and unique strategies. In terms of the contributions of interactive AI, XAI increased the trust that users had in the model and SI enabled further exploration strategies based on corrective or additive approaches.

We note that our evaluation approach has some limitations. In particular, our evaluation task is relatively simple compared to a real sensemaking task. Thus, these results could overestimate the capabilities of the interactive AI method in the context of narrative sensemaking. However, we also note that our experiments had a strict time limit that only allowed for limited training. Performing a longer study that allowed users to familiarize themselves with the system over an extended period of time would likely produce better results.

In this context, a few participants commented on the issues of developing more structured strategies in a short period of time. One participant commented that “*it can take maybe like few days or weeks working with this before you, like, understand what’s the best strategy*” (B₅). Another participant commented that “*if somebody used this software more than just like, once in a study, they [would] definitely develop different strategies*” (I₄).

Nevertheless, despite these issues, the results of our insight-based analysis are promising and point toward the capabilities of interactive AI to aid in sensemaking. In fact, our results confirm the capabilities of the narrative maps model for sensemaking, which had only been tested in smaller-scale settings with static pre-generated maps [186] or with manually generated maps [192]. Even in the cases where the model fails to generate a good map (e.g., due to conflicting SI or bad parameterization), there is value in the exploratory analysis it enables.

Furthermore, our evaluation data set and the associated detailed and high-level insights collected in our analysis could be used in future research to evaluate other narrative sensemaking systems. In particular, the catalog of insights that we provide could be used as a *gold standard* data set in subsequent studies.

Finally, in a broader context, our work on narrative sensemaking could benefit fields such as intelligence analysis [107], computational journalism [53], or any field that requires understanding and analyzing complex narratives [189].

9.6.2 Interactive AI Framework

The mixed multi-model interactive AI framework provides a general approach that can enable the development of VA systems with complex pipelines that mix higher-level discrete structures as outputs with potentially many lower-level continuous models. Moreover, the identified challenges and their respective design guidelines discussed in this article can aid practitioners and researchers that seek to construct their own interactive AI models. However, we note that this flexibility comes at the cost of increasing complexity with the number of levels. The deeper the pipeline, the more difficult it would be to address the discussed challenges, as each level would potentially require its own interaction and explanation model.

We note that the implementation of the specific pipeline depends on the specific structure space and internal models used in the system. Researchers working on such systems would need to address the challenges of interactive AI according to their needs. The general

guidelines proposed in our work could aid them in addressing these challenges. However, developing a model-agnostic approach that does not require redesigning interactions for each specific case remains an open challenge.

9.6.3 Explainable AI Model

Our proposed interactive AI framework showcases how to approach the development of XAI components in the context of mixed multi-model pipelines. Our results show that the proposed approach provides support to analysts in the context of narrative sensemaking, helping them obtain additional insights from data. Furthermore, the participants of our user study considered the addition of XAI beneficial and it increased their trust in the underlying models, even if the explanations themselves were not always particularly useful for the assigned task. However, our user evaluation also shows that there is some contention on the usefulness of each individual component, such as storyline names, which were sometimes considered very useful and sometimes discarded as mostly useless, depending on the participant.

In general, XAI provides useful scaffolding and enables users to potentially find more insights. However, developing appropriate XAI methods that capture the intricacies of the underlying models and are useful from a user's perspective is a complex task. In general, based on our results, we recommend designing such explanations with the goal of helping users understand the big picture by providing general pointers and providing support for specific confirmatory tasks, such as checking whether an element of the model is valid (e.g., event connections). In our evaluation, methods that provided too much information or were too technical in nature were generally regarded negatively, as participants were likely overwhelmed by the information overload of the explanations [118].

9.6.4 Semantic Interaction Model

Our proposed interactive AI framework integrates user feedback into the underlying AI model through SI. Our qualitative analysis shows that the model is generally capable of properly capturing user feedback and turning it into something useful. Thus, our experimental results show that our framework is effective at supporting *incremental formalism* [329] through SI. However, some challenges remain and need to be addressed to fully enable proper human-AI collaboration for sensemaking.

In particular, while our interactive AI model was able to capture knowledge from the analysts during our user study, we note that in some cases, there were issues regarding how specific interactions should be interpreted. More specifically, these issues typically arose when the participants performed unexpected interactions which were harder to interpret or outright led to internal model failures. The general challenge of discerning user intent from interactions is called the “With Respect to What” problem in SI literature [165, 383].

In the context of the challenges in our interactive AI framework, this issue relates to the *Interaction* challenge of the XAI component. In particular, interactive AI models should strive to explain the effects of SI on the underlying AI model. Doing so will help ameliorate the issue of *interaction ambiguity* [189], allowing the user to perform corrective interactions to redirect the model as needed.

Addressing the challenge of interaction ambiguity depends heavily on the underlying SI model and how it has been implemented. In general, determining the exact manner in which the SI model has influenced the underlying model is a complex task. A general approach towards dealing with SI explainability that can handle diverse models is an open-ended challenge that requires more research.

9.6.5 Scalability Issues

Previous research identified two key drawbacks of using SI in the context of narrative sense-making [189]: *transparency*—how can we explain why the model acts in the way it does?—and *scalability*—how can we handle bigger data sets?

Our interactive AI framework addresses the *transparency* issue through the use of XAI techniques, providing explanations of the low-level space through topic modeling, explanations of the high-level space through storyline names, and explanations of the connection between those two spaces by explaining the edges that connect two events. Thus, our approach overcomes the issue of transparency by exposing the inner workings of the algorithm—at least partially, as some components still require more in-depth explanations based on our evaluation and there are other challenges, such as layout issues, that hinder explainability. However, our approach does not solve the other key drawback: *scalability*.

In terms of scalability and performance, our framework intensifies some of the issues found in previous research, as the XAI components require even further computations. Thus, using a full interactive AI pipeline compared to only an SI model comes with a trade-off between *scalability* and *transparency*.

In this context, we could consider some of the user feedback to develop new narrative representation and extraction models that offer a better deal in terms of scalability. However, this would require revising the extraction process accordingly. In particular, we focus on the proposal of generating a map based on *grouping events into higher-level structures*. This approach breaks the assumption of the 1-to-1 relationship between events and documents that forms the basis of the current model [189].

Following such an approach would reduce the computational cost of the extraction process and layout generation by including an extra pre-processing step. This step would attempt to group documents that represent the same event into a single group and treat them as a singular entity. However, this requires the development and training of appropriate machine learning models, as well as their corresponding inverse models and explanations to con-

struct the full interactive AI pipeline. In this regard, we refer to other narrative extraction approaches that group documents to represent higher-level events [18, 232].

9.6.6 Limitations

Despite the generally satisfactory results of our framework in the user study, we note that our work is not without limitations. The proposed interactive AI framework provides a general approach to dealing with the challenges of XAI and SI in a mixed multi-model setting. However, we note that our specific implementation for narrative maps is not model-agnostic and would require extensive adaptation to integrate with other models.

Regarding the evaluation of our model, we note that the insight-based evaluation used a relatively simple task that only deals with 160 documents. However, we note that the size and complexity of this evaluation are in line with similar research in sensemaking [51, 169, 301]. Furthermore, the time constraints imposed on the participants do not necessarily reflect the real conditions that analysts would face when performing sensemaking tasks. Nevertheless, despite the limited time available to learn how to use the prototypes, the results showed that our interactive AI model provided the highest number of insights. We also note that another issue was the general lack of experience of the participants in narrative sensemaking. However, the task and the data set were designed to be sufficiently simple so that even non-expert analysts could answer them.

Finally, we note that there are other approaches that could be used to implement the XAI and SI components. In particular, as can be seen from the user-suggested functionalities, there are several missing elements that our model does not handle in its current form. Despite these limitations, our quantitative and qualitative results display the capabilities of our interactive AI model in helping analysts perform narrative sensemaking.

9.7 Chapter Conclusions

In this chapter, we proposed a mixed multi-model framework for interactive AI, which relies on a SI component—to model the user’s cognition and guide the underlying AI model—and a series of XAI components—to provide insight into the inner workings of the AI model to the user and help them understand its extraction process. Furthermore, we showcased the capabilities of this approach by developing an interactive AI model for narrative maps and performing a quantitative and qualitative evaluation through a user study with an insight-based methodology.

Our results show that our interactive AI prototype provided advantages compared to a basic map-based prototype and a timeline-based prototype with search. In particular, participants with the interactive AI prototype were able to provide more relevant insights. Thus, sup-

porting the hypothesis that the use of the interactive AI framework is an effective way to support the sensemaking process of analysts. We further identified specific usage patterns and strategies.

In terms of broader impact, using interactive AI models with narrative maps allows us to better support the sensemaking loop for narrative sensemaking tasks, such as journalistic analysis of news narratives [53] and intelligence analysis [107]. Furthermore, our proposed framework provides a general approach that can be used in other mixed-model contexts and the identified challenges and design guidelines can provide a starting point for researchers and practitioners working with their own interactive AI models.

Finally, while our work addresses the issue of *transparency* in the context of narrative extraction methods through XAI, we still have not addressed the issue of *scalability*. Thus, future work should seek to improve the extraction approach to allow for the analysis of larger data sets, which raises the additional challenge of reducing the computational cost of the interactive AI loop without diminishing its capabilities.

Chapter 10

Conclusions

In this chapter, we present the conclusions for this dissertation and discuss future work. We begin by returning to the research questions posted in Chapter 1 and summarize the key elements of our system. We show how the research presented in this dissertation addresses each one of the research questions. Finally, we discuss potential avenues for future work.

10.1 Research Questions and System Summary

For this dissertation, we posed three research questions focusing on challenges and goals in information narrative representation, extraction, and visualization. We now discuss some of the implications of our results for each research question. Then, we summarize the key elements of our system in terms

10.1.1 RQ1: How can we computationally represent and extract information narratives?

The first research question of this project was focused on developing a computational representation model and an extraction approach to obtain narratives from data.

In Chapter 3, we performed a systematic literature review on narrative extraction. Particularly, we screened over 750 articles, which yielded 54 articles concentrating on extracting event-centric news narratives. We identified common representational structures, extraction criteria, and evaluation metrics. In addition, we analyzed a number of recent tendencies, open issues, and prospective research directions. We highlight the lack of benchmark data sets, the need for improved evaluation metrics capable of handling complex narratives correctly, the high computational costs of most methods, and the absence of standard benchmark tasks for user-based evaluations as limitations of this field. Concerning open challenges, we observe the need for improved interaction models that give users more control over their exploration of the narrative. Lastly, we observe that current models do not handle misleading or false content, a growing problem as misinformation combines with information overload to make it more difficult to comprehend the information landscape. This literature review is the basis for our methods and future work in the field of narrative extraction.

In Chapter 4, we developed a natural language processing model to extract the *main event descriptors*—the answers to the 5W1H questions (Who? What? When? Where? Why? How?)—from breaking news articles. The main event descriptors can be used as part of a narrative representation, providing more details on the events. Moreover, we compute an evaluation metric, called the Inverted Pyramid Score (IPS), which can determine whether a news article follows the inverted pyramid structure. Our analyses of results show that the method works well, allowing us to distinguish between breaking and non-breaking news articles. However, this approach could be improved by using state-of-the-art summarization schemes tailored for news articles. Moreover, in terms of potential applications, we highlight how this work could be used to evaluate other structures of news articles, not only restricting ourselves to the inverted pyramid. Finally, the IPS provides a useful method to evaluate how fake news sources structure their articles, as well as compare them to mainstream outlets. Low values of IPS can be used as a potential indicator of whether an article is credible [187].

In Chapter 5, we developed a narrative representation model called narrative maps. In particular, we made three key contributions: a theory-driven *computational representation* of narratives, a novel *extraction algorithm* to obtain these representations from data, and an *evaluation* of our approach. We showcased the effectiveness of our approach by performing a user evaluation to assess the quality of the representation, metaphor, and visualization. Unlike previous attempts at extracting storylines or narratives from data [319, 323, 369, 399, 413], our work is based on a comprehensive survey of narrative theory. Hence our narrative representation provides a general framework to represent narratives and can serve as a starting point for understanding narratives through a computational lens [191]. The evaluation proves that the narrative map representation, alongside the extraction method, is an effective mechanism to help users understand complex news narratives. The route map metaphor was considered natural by the users, providing a coherent overview of the narrative. Moreover, the main route and the representative landmarks are able to convey the key elements of the narrative.

In Chapter 7 we modeled the accepted narratives of online communities related to social movements. In particular, we analyzed the 2021 Cuban Protests in five Reddit communities. To perform this analysis, we designed a new narrative extraction algorithm that incorporates *community acceptance* into the model. Moreover, we developed a narrative comparison metric based on graph edit distance to quantitatively analyze the differences between the accepted narratives of different communities. In terms of implications, our work could be applied by social scientists interested in a narrative analysis of how social movements are evolving, as the accepted narratives provide a window into what different communities are thinking and their views. In general, narratives are considered key to social movements [36, 87], as they shape the discourse and perception of society. Furthermore, we note that the narratives pushed by social media in this context could lead to further developments in the social movement, which in turn could lead to further narratives being pushed, resulting in a potential feedback loop that could be of interest to analysts. Thus, the ability to model the narratives that arise in social media could provide useful insights into how these narratives are evolving

and shaping social movements. Furthermore, our work could also be used in the context of computational journalism [79], which provides tools for various computational needs in journalism, such as topic detection, sensemaking, and visualization. Additionally, we could include other elements in our representation. For example, narrative maps could be part of a narrative exploration suite, which could include maps in addition to text summaries, data plots, and references to trustworthy sources. Such a tool could aid analysts, journalists, and other users in trying to understand the narratives espoused by online communities.

10.1.2 RQ2: How should we design narrative map visualizations to aid analysts in their sensemaking process?

The second research question of this dissertation was focused on designing an interactive visualization tool to aid in the narrative sensemaking process. To achieve this objective, we first needed to understand how users create and apply narrative maps to solve sensemaking tasks and what makes a good narrative map.

In Chapter 6, we sought to understand how analysts create and use narrative maps in order to obtain design guidelines for a narrative map visualization toolkit that can aid analysts in narrative sensemaking. We performed two experiments with a data set of news articles. In particular, our contributions are three-fold: an analysis of *how analysts construct narrative maps*; a *user evaluation* of specific narrative map features; and *design guidelines* for narrative maps. Our findings suggest ways for designing narrative map representations and extraction algorithms, as well as providing insights into useful user interactions.

In particular, we found three major ways to construct maps. Each one of these strategies can be the basis of a new narrative extraction algorithm. Furthermore, in terms of the structure of the map, we saw an even distribution between tree-like maps and DAG-like maps. Regarding layout, we found that most users preferred a vertical top-down layout, with the main story shown first. We also evaluated the effect of map size and transitivity, finding that users preferred long maps without transitive connections. All these results led to a series of design guidelines for narrative maps. These guidelines can be used in the design of new narrative map extraction algorithms and interactive visualization tools. Future work could deal with the implementation of such algorithms and tools, as well as their evaluation based on the insights gathered in this work.

10.1.3 RQ3: How can we create an interactive artificial intelligence model for narrative maps?

The third research question of this dissertation is focused on incorporating higher-level sense-making capabilities into the narrative map representation.

In Chapter 8 we augmented the functionality of narrative maps through semantic interaction [107] to better support the sensemaking loop. Incorporating semantic interaction into the visualization tool required developing a proper user-interaction model that understands the intent of users with each interaction. In particular, we proposed the concept of 3MSI—Mixed Multi-Model Semantic Interaction—an SI pipeline defined by a higher-level discrete structure and lower-level continuous models. We showed the effectiveness of our pipeline by implementing it for narrative maps and evaluating it through a quantitative simulation-based approach and a qualitative expert-based approach. Our results support the hypothesis that the SI models are effective at supporting incremental formalism for narrative maps. However, despite the model’s capabilities. We note that there are two key issues with the proposed 3MSI approach: *scalability* and *transparency*.

In Chapter 9 we extend our SI pipeline into a full interactive AI pipeline through the addition of explainable AI techniques [157]. In particular, we proposed a mixed multi-model framework for interactive AI with two key components: an SI component that models the user’s cognition to guide the internal AI model and an XAI component to provide insight into the internal AI model and the narrative extraction process. We implemented interactive AI for narrative maps following this model and performed a quantitative and qualitative evaluation through an insight-based user study. Our results showed that our interactive AI model provided advantages compared to the baseline regarding the number of insights extracted, supporting the hypothesis that the use of the interactive AI framework is an effective way to support the sensemaking process of analysts. We note that while our work addresses the issue of *transparency* through XAI, our model still has the issue of *scalability*. Thus, future work should seek to design more efficient methods of narrative extraction and interactive AI techniques. Finally, in general, integrating SI and XAI models with narrative maps allows us to provide better support to analysts performing narrative sensemaking, including tasks such as journalistic analysis of news narratives [53] and intelligence analysis [107].

10.1.4 System Summary

We summarize our work in terms of mappings between different aspects, including documents, events, the system itself, and the potential users.

Documents and Events: The current system can extract and visualize narrative maps from a collection of documents under the assumption that there is a one-to-one mapping between documents and events. We note that this is a strong assumption, as in general, documents may be associated with one or many events and vice versa. However, most of the data we used throughout our evaluation consisted of breaking news articles, which tend to focus on a single event [187]. Thus, the assumption of one document being associated with one event can be considered appropriate in that context.

Furthermore, we note that the documents in our data set consist of text (e.g., the headline and body of a news article) and a timestamp. The text content is represented through neural

embeddings and these embeddings are used to determine coherence scores in our narrative model. The timestamp information is used to join these events in chronological order and also influences the coherence score. Furthermore, we also extract entities from the text content, to aid in the process of connection explanation of our AI model. We do not take into account other dimensions of the documents (e.g., location, bias, accuracy, or validity).

System Description: The system itself contains several visualizations, screens, and interactions. The main screen of the system is the visualization of the extracted narrative map, where events are mapped into the narrative structure through our extraction pipeline. This visualization is always present during the interaction with the system. The system contains a main menu at the top that allows the users to load different data sets, create new maps based on user-defined parameters, and interact with the map itself by removing or adding elements to the map, among other options.

Next, the system contains a side panel with multiple tabs: an overview tab that shows the projection space with a visualization of the topical clusters found in the data set; a document view tab that presents the contents of a selected document from the map; a connection explanation tab view that provides an in-depth report of why two events are connected alongside a visualization of how each keyword contributes to the connection; a data set tab that shows a table that shows all the documents and allows users to select the starting and ending event of the narrative; and an advanced settings tab that allows the user to configure advanced elements of the model.

User Personas, Goals, and Tasks: The system can support various user personas with different goals and tasks, such as researchers, journalists, intelligence analysts, and fact-checkers. In particular, we briefly discuss how the system may help these users.

Researchers. The system can be used to help in the literature review process of researchers. For example, the narrative map can provide a big-picture view of the different trends in the literature. In turn, this may help researchers see relationships, spot trends, and unearth fresh insights. Additionally, they can use the interactive AI capabilities of the system to adjust and improve the narrative representation, which may help them obtain more insights for their research. However, one limitation of this model for researchers is that it does not handle reference-based connections, which would be particularly useful to better understand how scientific articles connect to each other.

Journalists. The system can be used by journalists, particularly those that work in investigative journalism, to research stories, explore the information landscape, and effectively convey information by using the narrative map itself as a potentially useful output. The system can help them gather facts and organize them into a logical narrative structure. Furthermore, the customization capabilities of the interactive AI model allow journalists to view the information landscape from a different perspective.

Intelligence Analysts. The narrative maps system can help intelligence analysts in their sensemaking process, particularly in the process of schematizing the data, generating hy-

potheses, and storytelling. The map itself could be used to effectively convey information in intelligence reports. The system can help analysts connect pertinent pieces of information, such as posts on social media, news stories, and classified reports. Based on the narrative structure, intelligence analysts can spot possible dangers, follow trends, and produce useful intelligence by visualizing how events are connected. Moreover, the interactive AI capabilities allow intelligence analysts to tailor the narrative model to their specific needs. In particular, the connection explanations generated by our system could be particularly useful for these users. However, we note that the current model lacks support for cause-effect relationships, which would be particularly useful for intelligence analysts.

Fact-checkers. The system can help fact-checkers to confirm and refute false or misleading narratives. The maps can provide fact-checkers with an overview of the information landscape and help them organize and cross-reference multiple assertions, sources, and pieces of evidence. The need for tools that can assist in fact-checking is increasing due to the large scale of online misinformation diffusion [175]. In this context, fact-checkers could use the narrative structure to help determine the truthfulness of a claim by identifying how a specific piece of information fits in the narrative and locating discrepancies or inconsistencies. Furthermore, by providing a visual picture of the debunking process, narrative maps can also aid fact-checkers in effectively communicating their results.

10.2 Future Work

We now discuss potential future work and provide guidance on potential avenues of research that arise from this body of work. In particular, we address several of the issues and challenges faced in our research and discuss alternative ways to solve these problems, extensions to the proposed models in this dissertation, and integration of these models with pre-existing elements of the literature.

10.2.1 Narrative Representation Issues

We start with some of the challenges that arise in terms of narrative representation.

Narrative Definition

The literature contains numerous potential definitions of narrative. General narrative theory is explicitly concerned with comprehending the general principles of narrative and its various meaningful arrangements [1, 285]. Formal narrative theory assumes that there is a distinction between the actual story and its representation. The objective of narrative theory is to comprehend the connections between stories and their numerous potential representations

[285]. Other definitions view narratives as communication tools for constructing a shared understanding of events in order to influence behavior [257].

Halverson et al. [145] define narratives as a system of stories as opposed to a single story. In other words, narratives are a collection of interconnected *stories* with consistent themes. The definition of a story is a series of events strung together in a logical manner. According to this definition, events are the basic elements of narrative action; they are either an act involving characters and entities or an occurrence in which no entities are causally involved [1]. This definition is used to model news narratives. Thus, we have a series of hierarchical definitions beginning with the narrative, proceeding to stories, and ending with the narrative's fundamental units: the *event* and its associated *entities*. In addition, these definitions require a fundamental *order* for the events, as they must be linked sequentially.

Events and Entities

Based on our previous discussion, there are two fundamental units of narratives: events and entities. These elements offer unique perspectives on the narrative. One unit focuses on the happenings and actions of the story, while the other focuses on the characters and other entities involved. In our work, we developed an *event-based narrative representation*. While we include the concept of entities for the purposes of explainable AI, they are not the central focus of our representation. Thus, one natural extension of the current narrative model is to include an explicit entity model in the representation and properly integrate it with the extraction pipeline.

Furthermore, we highlight the fuzzy nature of the definition of events. For example, an event may refer to an abstract representation of a single action (e.g., a single verb or a predicate) [255, 286], a full sentence inside a text document [134], a document in the data set describing something that happened (e.g., a breaking news article) [319], or a group of documents all related to the same event [265, 400]. In this context, the narrative map model could be modified to handle such cases (e.g., generating a map of predicates inside a document or a map of clusters of documents), provided that the new representation properly accounts for the chronological order of the events, which may require additional pre-processing.

Finally, an alternative definition of events is based on the notion of *changes of state* in the narrative's world [163, 315]. In this context, extending our model with a representation of the current state of the world could help us model the narrative more accurately (e.g., using a knowledge graph [266]).

Structure

Choosing the representational structure is one of the core aspects of narrative representation. Our model relies on a directed acyclic graph representation that provides the most flexibility

to model convergent and divergent stories while also accounting for the temporal nature of a narrative. However, it is not the only event-based structure that can be used to represent narratives. In particular, there are also linear representations [60, 319, 416] (e.g., timelines), directed graphs [46, 343] that allow for non-chronological relationships (e.g., same-event relations), and tree-based models [232, 408, 415] that fall between linear and fully graph-based representations. Our narrative extraction pipeline could be adapted to extract such structures rather than a directed acyclic graph. In particular, sacrificing some flexibility in representation for reduced computational cost could be a decent trade-off considering the high computational cost of our current extraction model.

Main Storyline and Important Events

In order to help users interpret the storylines, major themes, and relationships between events captured by our narrative, the extraction pipeline finds the main storyline and representative landmarks (i.e., important events) on our map. These elements serve to characterize the extracted narratives, helping users in their sensemaking tasks. Regardless of the extraction method used to identify these elements, our user evaluations revealed that both the primary route and the significant events were effective at providing an overview and emphasizing the narrative’s essential components. While the main storyline provides a *longitudinal* perspective of the narrative, the representative landmarks provide a *transverse* perspective. In essence, the representative landmarks and the main storyline provide us with a sense of the narrative’s breadth and depth, respectively.

Regarding the main storyline, we define it as the *most coherent* storyline in the narrative map by finding the route with maximum likelihood after normalizing the edges of the graph. However, this is not necessarily the only way to find the most important storyline in a narrative map. One possible alternative would be to take an *entity-based perspective*, where a storyline could be selected based on whether the events share a common protagonist or entities in general.

Regarding the important events, our initial implementation of the narrative maps extraction method leveraged the concept of maximum antichains to find the point where the graph has its maximum width and selected those elements as the important events. This is a *structure-based* approach, as it relies on the graph structure to select the important events. However, in our subsequent models, we used different methods to select these elements: another structure-based method using centrality measures and a *content-based* approach by finding the document closest to the centroid of a storyline.

Narrativity and Eventfulness

In the context of finding the main storyline and important events, the concept of *narrativity* from narratological literature could be useful. One of the possible definitions of narrativity

provides a scalar sense of the *narrativeness* of a narrative [2]. In general, narrativity [185, 284] is a measure of the set of properties that characterize a narrative and what distinguishes it from non-narrative texts. Narrativity can be related to the concept of *eventfulness*, where events are defined as a non-trivial change of state in the narrative world and narrativity depends on the non-triviality of these changes [315].

This notion of non-triviality could be used to find the most relevant parts of the narrative (e.g., the main storyline or important events), as points where there is high narrativity (i.e., a significant change in the world state) could be important to identify. Recent work has delved into the problem of detecting narrativity in texts through linear regression [336] and classification [358]. However, these works focus on single documents and do not address the problem from the perspective of a structured narrative over multiple documents.

In this context, our work assumes that narrativeness is measured by a coherence function that relies on surface text similarity and topical information, but this is not the only possible way to define this notion. For example, we could rely on any of the different measures of coherence showcased in our survey from Chapter 3 or any of the other metrics discussed there that could work as a proxy for non-triviality and eventfulness (e.g., relevance metrics and event frequency).

Discourse vs. Actual Story

From a theoretical standpoint, the study of narratives distinguishes the story itself from how it is being told—the *discourse*. In this sense, our narrative maps model is able to generate different *discourses* from the same data, depending on its parameters and any user interactions captured by the semantic interaction model. From this point of view, the multiple narrative maps extracted represent multiple discourses, which in conjunction, provide a bigger view of the narrative.

However, at this point, we come to the issue of what even is the *real* narrative, because our data is limited by the media narrative and the associated discourses of mass media organizations. Thus, even if the system captured the narrative in full, it would still be biased by the data it comes from. This difference in presentation by different sources is represented by the concept of *framing* [362]—which refers to how these sources highlight specific issues and make them more salient, while other aspects are omitted. In general, framing plays a key role in shaping societal perception and accepted narratives as certain frames dominate the discourse [182]. Moreover, the framing and public perception of narratives might evolve as new events occur, previously unknown information is revealed, or as new interpretations arise. These changes could even be abrupt depending on the specific type of narrative (e.g., crisis events might induce drastic changes in the narrative and its framing). In general, without the In this context, being aware of the specific framing used by the sources and how it changes over time could prove valuable for analysts and users trying to understand the narrative.

Thus, to further help users in understanding the narrative, we could incorporate framing information into the narrative maps model. For example, framing information could be integrated into the narrative extraction algorithm, enabling the model to capture articles with similar framing in the same stories (e.g., by increasing coherence scores based on *frame similarity*). Alternatively, we could attempt to model framing information using semantic interaction techniques. In this context, we highlight that previous works [62, 234, 264] have shown that journalistic frames can be learned by classifiers based on deep learning representations (i.e., embeddings). Thus, we would expect our semantic interaction model to be able to capture the signals associated with framing.

Finally, to mitigate the impact of dominant frames in the discourse, it would be useful to include different types of news sources (e.g., Western vs. non-Western sources). For example, our user study in Chapter 9 had mostly Western sources and lacked any significant presence of non-Western sources. Thus, the insights obtained by the participants were influenced by the dominant frames of this data set, shaping their perception of the Cuban protest narrative. To find a more unbiased narrative and expose users to different perspectives, the narrative map could explicitly consider framing in its extraction process and warn the user about potential framing issues. In this context, it could also be useful to include data from other types of sources beyond news to include other perspectives (e.g., complementing news data with intelligence data).

Alternative Representations and Comparison

We note that there are other types of conceptually similar representations that could be used in the context of narrative sensemaking. For example, mind maps [35], concept maps [357], and knowledge graphs [115]. Beyond these three general examples, we also note that there are specialized tools available to the various user personas. In this context, it would be useful to perform a comparison against these types of representations and evaluate how well they help users in performing their tasks. Furthermore, another possible line of research would be to develop an integrated model of narrative maps with these other representations, which could provide users with a holistic view of the narrative.

10.2.2 Narrative Extraction Alternatives

We now discuss the computational cost issues of our extraction pipeline and some potential alternatives to our current approach.

First, we note that the total number of variables in our linear programming problem is $O(|D|^2 \cdot k)$, where $|D|$ is the number of documents and k is the number of clusters. The number of clusters is sufficiently small to be considered constant in practice. Thus, the generation of the pairwise combinations of nodes required by the optimization problem is our primary computational bottleneck.

Thus, in consideration of the fact that our linear programming framework has a high computational cost, we could evaluate alternative optimization strategies with a lower computational cost. In particular, we could follow a similar approach to the Connect-the-Dots method [319], which evolved from its basic linear programming formulation into a divide-and-conquer approximation approach [324]. In particular, we could endeavor to improve the current linear programming approach or switch to a completely different optimization technique to address this issue. In this context, we discuss potential methods that we could use to extract the narrative maps, based on similar narrative or storytelling extraction tasks explored by other researchers and general optimization literature.

For example, we could use a search-based approach (e.g., greedy or A* algorithm) [320, 323]. In this case, the process of narrative map construction can be viewed as a search problem in which we try to find the optimal paths from Event A to Event B (assuming that we are providing a starting and ending event). Using a suitable heuristic (e.g., approximating global narrative coherence by local coherence), we can employ search algorithms such as A* to determine the optimal path. By repeating this procedure in a systematic way to generate multiple paths, we can construct a narrative map without using linear programming. Alternatively, we can follow a greedy search procedure to reduce computational costs even more. If this approach is properly defined, it could yield a reasonable approximation of the optimal solution.

Another alternative would be to use relaxed linear programming [6]. Relaxation simplifies an optimization problem by transforming it into an approximation with looser constraints. Consequently, we obtain solutions to this new problem that approximate the original difficult problem. This strategy can make solving the problem much simpler and faster, at the expense of not obtaining the precise solution.

Furthermore, we could use reinforcement learning [404] techniques to optimize the coherence of our narrative representation. In particular, this would require defining an appropriate agent that can add events to a blank map to start building the narrative. Based on their actions, the agent would receive feedback based on the relevance and coherence (or any other appropriate narrative evaluation metrics) of the elements they add and how they structure them.

10.2.3 Narrative Evaluation Issues

There are several challenges when evaluating narratives for the purposes of narrative extraction. In general, defining what makes a narrative better than another is not easily defined. Thus, designing an appropriate metric could help extract better narratives to explore the information landscape. Our work relies on a notion of coherence, but as shown by our survey in Chapter 3, there are potentially many definitions of coherence that could have been used. So in this section, we discuss the issues of lack of standardized metrics and the fact that most metrics are purely based on content, without accounting for ordering or structure.

First, we note that there are multiple versions of coherence and comparable metrics, which is one of the drawbacks of current methods. In practice, coherence is ill-defined, and formalizing a computational or mathematical definition of it is challenging. The various definitions of coherence-like metrics focus on measuring distinct narrative components. In addition, extra constraints beyond numerical metrics can be considered to enforce coherence (e.g., events sharing common entities). In general, a hybrid extraction method that combines multiple metrics (e.g., via linear combination) and incorporates such constraints may produce superior results and could be one way to improve our current narrative extraction pipeline.

The majority of narrative extraction techniques only consider the content (e.g., traditional information retrieval metrics), without taking into consideration the narrative's order or structure. Some metrics take ordering information into account, but only at the level of linear structure. For instance, story-level coherence measures evaluate the relationships between consecutive documents [319] or the dispersion coefficient [60], which models story flow. We note that some works rely on ranking evaluation metrics, which incorporate an underlying notion of order; however, this notion is at best restricted to a linear structure. Another example is Bögel's [46] structural variant of coverage, which is based on local event connections, but does not account for the entire graph structure. Therefore, the prevalent metrics for the task of narrative extraction are incapable of handling complex narrative structures.

Given this limitation, it would be preferable to consider metrics that take both order and structure into consideration in order to provide an accurate evaluation of a narrative. For linear narratives, it is sufficient to consider the content and order because the structure is fixed. Consider a metric based on weighted edit distance [378], which takes into account both the order of the elements and their contents (by designating weights based on event similarities or an adaptation of the previously discussed metrics). Graph edit distance [4] with custom costs could be used in a similar manner for non-linear narratives, as this metric would incorporate structure, order, and content.

Nevertheless, the previous proposal would be supervised, as we would need a narrative against which to evaluate the output. Developing an unsupervised method is more difficult, especially for graph-based narrative representations. Attempting to extend coherence and dispersion measures to such graphs is an alternative. For instance, given a directed graph with a single starting event and a single ending event, it may be possible to compute all routes from start to end and derive a weighted average of these routes' coherence or dispersion. However, for more complex graph structures, it may be computationally prohibitive to perform this operation. Thus, designing a metric for unsupervised evaluation remains a challenge.

10.2.4 Semantic Interaction Challenges

We now discuss some of the challenges that arise in the context of semantic interaction with our narrative maps model.

Semantic Interaction Model

Our proposed interactive AI framework incorporates user feedback into the underlying AI model through the use of semantic interaction. Our qualitative analysis demonstrates that the model is generally capable of capturing and transforming user feedback into something useful. In turn, our experimental results demonstrate that our framework effectively supports *incremental formalism* [329] via semantic interaction. Nevertheless, to truly enable proper human-AI collaboration for sensemaking, a number of obstacles must be addressed.

Specifically, despite the fact that our interactive AI model was able to extract knowledge from the analysts during our user study, there were ambiguities in how specific interactions should be interpreted in some cases. These issues typically arose when participants engaged in unexpected interactions that were difficult to interpret or directly caused internal model failures. In SI literature, the difficulty of determining user intent from interactions is referred to as the “With Respect to What” problem [165, 383].

Semantic Interaction and Narrative Evaluation

Another issue is that there is no guarantee that the models generated after semantic interaction are actually “better” than the original narrative extracted by the system. In fact, one might argue that the new narrative is objectively worse according to our narrative evaluation metrics, otherwise, the optimization problem would have found it instead (assuming that there are no multiple optimal solutions).

In this context, considering our previous analysis on multiple presentations of the same underlying narrative (i.e., different discourses), semantic interactions on narrative maps are about finding a different perspective—another discourse—of the underlying real narrative, rather than trying to extract the *correct* or *better* narrative. In this context, semantic interactions enable navigating the *discourse space* [85] of a narrative.

Following this line of thought, an analysis that could be useful would be to qualitatively evaluate the quality of the narratives generated with semantic interaction against non-interactive narratives generated only using parametric interaction. While our user study did not show significant differences in the number of insights obtained by the participants with non-interactive maps vs. those with interactive maps, we did not evaluate the quality of the final narrative maps used by the participants. Thus, it would be useful to qualitatively evaluate these results.

Other Interaction Models

We note that there are other potential interactions that could be used to capture the intent of analysts. For example, an analyst might manually change the main story of the narrative map or highlight a specific event to mark it as more important. Other kinds of model guidance could include a query-based approach, where events would be emphasized based on a user-given query. This could be implemented by including an additional factor in our mathematical definition of coherence.

Another approach would be relevance-based guidance using custom event weights on top of the pre-existing semantic interaction model to emphasize specific elements of the narrative. Alternatively, an analyst could interact with the relevant keywords of the narrative (or one of its components) to assign a higher weight to it. Another strategy could be a spatialization-based approach [52] to guide the model by placing groups of related events close together. However, the current SI model does not support such interactions (due to the distances not having a direct meaning in the discrete narrative space).

Most of these interactions are event-based, as they either address an event or their connections. However, entities are another key element of narratives that could be used to guide the model. For example, an analyst could assign a higher weight to specific entities and the narrative would adjust accordingly.

Furthermore, there are other specific interactions that were discussed by our participants in our user study, such as *negative clustering* (grouping events and ensuring that they and related events did not appear in the map) and a *keep element* interaction that instructs the model to preserve a specific element.

Finally, we note that we could also develop semantic interactions on the explanations generated by the XAI components. For example, we could implement *storyline name changes* that influence the contents of the story. We could also allow direct manipulation of the topic explanations for the low-level model, indirectly influencing the projection space.

10.2.5 Explainable AI Challenges

We now discuss some of the challenges that arise in the context of explainable AI with our narrative maps model.

Explainable AI Model

Our interactive AI framework demonstrates how to develop XAI components within the context of mixed multi-model pipelines. Our findings indicate that the proposed method aids analysts in the context of narrative sensemaking, enabling them to derive additional insights from data. Moreover, the participants in our user study viewed the addition of

XAI as advantageous, and it increased their trust in the underlying models, despite the fact that the explanations themselves were not always particularly helpful for the prescribed task. However, our user evaluation reveals that there is a degree of disagreement regarding the utility of each individual component, such as storyline names, which, depending on the participant, were deemed either extremely useful or largely useless.

In general, XAI provides valuable scaffolding and enables users to potentially discover additional insights. Developing user-friendly XAI methods that capture the complexities of the underlying models is a difficult undertaking. In general, based on our findings, we recommend designing such explanations to aid users in grasping the big picture by providing general pointers and support for specific confirmatory tasks, such as determining whether a model element is valid (e.g., event connections). Methods that provided too much information or were too technical were viewed negatively in our evaluation, as participants were likely overwhelmed by the explanations' information overflow [118].

Low-level Model Explanations

Our explanations of the low-level model (i.e., projection space) were created based on clustering and a topic model using TF-IDF. However, TF-IDF is a basic representation and it would be possible to use more advanced topic modeling techniques to generate better explanations. In particular, dynamic and temporal topic models [43, 371] are particularly appropriate for modeling the events in the low-level space, as they take into account the temporal nature of narratives (or any other underlying total order). Such models would be particularly important for domains where the rate of change in the topics is high (e.g., social media data) [77].

High-level Model Explanations

Our explanations of the high-level model (i.e., the narrative space) consist of storyline name generation and identification of the important events and their associated explanations (e.g., hub event or representative event). Both of these approaches seek to provide a big-picture view of the narrative. However, they are still limited as shown by our user study.

Storyline name generation can be seen from the perspective of summarization, as a storyline name is simply a highly compressed summary of the contents of the story. In this context, our method is an *extractive* approach [261] that simply selects an appropriate name from parts of the headlines available in the storyline. However, it could be better to follow an *abstractive* approach [231] that generates a storyline name that does not necessarily appear in the events but represents them accurately. This could be accomplished by using advanced generative neural models [127] to abstract away the key elements of the storyline.

Regarding the extraction and explanation of important events, we highlight again the pos-

sibility of using the concept of narrativity or narrativeness and the idea of identifying non-trivial changes of state. Using this notion, we could explain the important events by virtue of their impact on the overall measure of narrativity. In this context, another potential approach is to identify the *breakpoints* [163] in the narrative—points in time where a decisive change occurs—which could provide us with better candidates for events.

Connection Explanations

Our explanations of the connection between the high-level model and the low-level model were constructed based on a post-hoc approach that generated explanations for each edge of the narrative map using edge labels and a SHAP-based tornado plot. This approach provides a general idea of why something was connected and can help identify superficial connections, as shown by our user study. However, these explanations are still limited and do not provide in-depth insights into how the extraction algorithm actually generated the map. In this context, there are some avenues that we could pursue in future work.

First, we could try to create white-box explanations based on the linear programming framework. In general, linear programs are considered white boxes [337] for the purposes of explainability, and some techniques exist to explore linear programs’ internal components in detail, such as sensitivity analysis [374] and dual analysis [201]. However, unlike traditional linear programs, our work relies on text similarities that come from neural embeddings. Thus, even if we understood which similarities are key to the solution, we would still have issues uncovering what they actually mean in practical terms due to the generally obscure nature of neural embeddings [352].

Second, we could try a post-hoc approach that relies on conversational generative neural models [151] to answer the question: “Why are these two events (not) connected?” In turn, these models could provide the user with an in-depth *narrative-based* answer that explains the connection (or lack of) between two events rather than a keyword-based or numerical answer like the current approach. In general, we could develop a conversational AI assistant that supports analysts in their narrative sensemaking process by leveraging information from the narrative map, explaining the contents of the map, and helping users draw information from the models.

Semantic Interaction Explanations

As identified in our interactive AI framework, explaining the effects of SI on the AI model back to the user is an important component of a fully realized interactive AI system. However, our current model is not capable of providing such feedback. Thus, explaining the effects of SI remains an open challenge that should be addressed.

Assuming that the SI model properly captures human cognition, this should not be explicitly

necessary. However, user interactions can be interpreted in different ways, and it is not always clear what the effect should be, an issue known as *interaction ambiguity* or the “With Respect to What” problem [381]. While this is a known issue in the SI context, it has not been addressed through the lens of XAI.

In this context, it would be useful to provide feedback on the actual effects of the SI so that the user can confirm that it is working as they intended or detect if they need to correct the SI model. This challenge was identified through the user study, as participants sometimes found unexpected effects from their SI, as such, our current implementation of the narrative sensemaking prototype does not address this issue beyond showing the changes in the model.

In this context, we note that simply showing the difference between the previous model and the model after SI is somewhat helpful, but this does not suffice as an explanation, as it would still be hard to understand why the changes happened. Addressing this challenge also depends on the specific SI used, as some might affect the low-level space, others might affect the structure space, and others might affect both. Thus, finding the exact way the interactions influenced the AI model is a hard task.

One potential approach to do this would be to leverage existing interpretation techniques for linear programs, such as sensitivity analysis [374], to find the key connections in the extracted narrative. Based on how these results change through each iteration, the analyst could get an idea of how their interactions are affecting the underlying extraction algorithm.

10.2.6 General Challenges

Now, we present some general challenges and issues that arise in our research.

Scalability, Performance, and Big Data

Most narrative extraction methods suffer from issues when dealing with big data, as the processing pipelines are quite expensive in terms of computational power and they might not be easily parallelizable. One of the simplest methods to reduce computational cost is to filter the data beforehand. This turns the computational cost problem into an information retrieval problem, where the most relevant documents must be retrieved before extracting the narrative itself. Many methods assume that the data has been pre-filtered to a relevant set of news articles. Including a filtering step adds an additional element to the pipeline, thus increasing the risk of errors. Moreover, defining an adequate concept of relevance for this method might prove problematic in itself. Nevertheless, this provides a simple approach to mitigate the ever-increasing available amount of data.

Another approach is to deal with extraction in an online manner [233]. Most news narrative extraction methods are offline methods that analyze an entire set of news articles. However, extracting the stories in an online manner without disrupting the pre-existing structure

would offer a computationally cheaper alternative. This is similar to the approach used by traditional TDT systems that sought to track the events of a topic in an online manner [13]. However, it would also require handling the structure of events associated with the narrative, which is not considered by traditional TDT.

Another strategy would be a divide-and-conquer approach, which could provide further reductions in computational costs. This method would work by judiciously splitting the data into smaller sets (e.g., through hierarchical clustering [17]) that can be easily handled by the more costly optimization algorithms. Then, we generate separate narratives for each part and merge them.

Another approach could be to merge documents that represent the same event in the narrative [343]. Very important events are likely to be reported several times by different sources. Therefore, merging these documents into a single one can be a way to reduce the computational cost of the main extraction algorithm. However, we should keep in mind that only events that are reported similarly should be merged, as differences in how the event is reported could provide important contextual information for the narrative (i.e., framing).

Moreover, in a similar way to how boosting techniques [312] work in machine learning, we could use sampling to create several weaker narrative models and use them to create a better aggregate model. This would reduce the computational cost by reducing the data set by only working on a sample rather than the full data set. However, this process would have to be repeated enough times to ensure that the aggregate final model is appropriate.

Users and Tasks

Regarding the potential users of our models, we note that most of our work has focused on news narrative extraction. Thus, a natural user of this model could be a journalist trying to understand the information landscape. In particular, investigative journalists and editors could benefit from the models developed in this dissertation. However, they are not the only type of user that could leverage our system for their sensemaking needs. In particular, our work could be of use to intelligence analysts seeking to connect documents and uncover hidden storylines; to historians trying to piece together a narrative from multiple historical documents; or to fact-checkers trying to determine whether a given piece of information is true or false.

In general, any user that seeks to perform narrative sensemaking tasks on a set of documents stands to benefit from our models (i.e., anyone seeking to understand the stories that arise in the data). More specifically, our models support tasks of the form “*How did event A lead to event B?*”—connecting a starting and ending event—and “*What are the effects of event A?*”—exploring the information landscape from a starting event. While we explicitly supported these two tasks in our system, we note that it would be straightforward to implement variants of these tasks based on a fixed ending event, a fixed middle event, or

even a combination of multiple starting, middle, and ending events. All these tasks can be implemented by directly modifying the linear programming formulation in our pipeline.

Furthermore, we note that the current model would need some adaptations to fully support the needs of certain domains. For example, intelligence analysts might require further support for entity-based connections and causal connections. Additionally, fact-checkers would benefit from a model that explicitly incorporates notions of credibility and bias in its extraction process. Thus, further extensions of the model would be required to fully support other types of users in their sensemaking needs.

Finally, we note that the design guidelines for narrative maps obtained in this work might not be applicable in all domains of applications. In general, some fields and contexts may require further changes to improve the narrative map model for their specific needs (e.g., intelligence analysts and their need for modeling causal relationships). In this context, following the practices of user-centered design [364] could prove useful when addressing the goals, tasks, and specific workflows of each user persona associated with the narrative maps system. More specifically, performing long-term studies involving different types of experts using the system in various settings for extended periods of time would be useful to improve the design of our models, including both the narrative maps themselves and the interactive AI components.

Data Set Availability

The vast majority of news narrative extraction works collect their own data or utilize a subset of an extant news data repository. For example, some works use data sets derived from TDT literature [68, 69, 265, 377], DUC/TAC conferences [71], or other general news archives [73]. A subset of timeline summarization works [134, 207, 278, 350, 351, 407] have supplied evaluation data sets that have been adopted as benchmark data in multiple works of timeline summarization with a sentence-level resolution for events. They do not, however, provide a direct method for evaluating methodologies that employ other resolution levels.

Furthermore, we note that such benchmark data sets do not exist for the other resolution levels of the narrative extraction tasks. The lack of relevant benchmark data for document-level and cluster-level resolutions makes it more difficult to compare methods and replicate results. Thus, the evaluation of such methods has to rely on user-based evaluations, which are relatively costly and time-consuming.

User-based Evaluations

Due to the aforementioned availability issues of benchmark data sets, most narrative extraction works rely on user-based evaluations. In particular, user-based evaluations usually emphasize subjective measurements over objective task performance. This is because eval-

uation frameworks and benchmark assignments are not well-defined. Current approaches rely on micro-knowledge tasks—tasks focused on information retrieval—that evaluate the number of correct answers (i.e., user accuracy) over time and macro-knowledge tasks—tasks focused on summarization—that evaluate the quality of the extracted narrative indirectly by measuring the quality of a user-generated summary. These methods are limited and do not account for all the nuances of narrative sensemaking. In addition, they do not address more complicated duties than retrieval and summarization.

A possible solution would be to devise more comprehensive assessments based on a variety of benchmark tasks. Specifically, the application of Bloom’s taxonomy [45] could provide a useful framework for defining such tasks as they appear in other sensemaking applications [57] or cognitive tasks in general [80].

Another option is to rely on the concept of insight-based evaluations [271], which is the approach we followed in our work. Instead of focusing on benchmark tasks with clearly defined tasks and correct answers, this type of evaluation is open-ended and requires extensive analysis of the user-generated insights. In this context, our insight-based evaluation can be further used as a baseline in future studies and as a benchmark task (e.g., measuring how many insights out of the baseline users identify or if they identify any different insights from the baseline). However, we note that insight-based evaluations are time intensive and require significant qualitative analysis.

Finally, we note that another alternative is to deploy the current system and test the effectiveness of our system in a real application context. This would require a long-term user study, where we would recruit expert users (e.g., investigative journalists or intelligence analysts) and ask them to use the system for their sensemaking tasks.

Trust in Computational Models

In our user evaluation, we found that participants trusted the explainable AI model significantly more than the semantic interaction model, despite some of the flaws of the explainable AI components (e.g., storyline names). This difference in trust could be caused by the more obscure nature of semantic interactions, as our current model does not provide any explanation of how they actually work for the users. Thus, this issue also ties in with our previous discussion of the explanation of semantic interactions.

For the purposes of increasing trust in the system in general, it could be useful to leverage concepts from the Computers As Social Actors (CASA) theory [292], which states that humans treat computers as social actors, in a similar way to how they treat other sentient beings. In particular, by providing more human-like explanations based on the CASA theory, going beyond relatively simple numerical information or text-based explanations. Thus, the narrative map system could be redesigned in a way that increases trust in the system following the CASA theory.

Furthermore, additional emphasis could be placed on the three core technology-related trust beliefs [211] when designing the system. These beliefs include *functionality*—the system helps with the task; *reliability*—the system operates in a consistent manner; *helpfulness*—the system provides adequate and responsive help. Our experiments prove that the system is functional, but there are some issues with reliability (e.g., inconsistent maps) and helpfulness (e.g., no adequate help when SI fails to produce good results).

Ethical Issues

There are some potential ethical issues that arise with the use of our proposed narrative map model, its extraction process, and interactive AI.

First, we note that the model does not distinguish between truth and misinformation, which could allow users to explore how misinformation spreads throughout the information landscape, but also lead to incorrect interpretations of the narratives. Furthermore, it could contribute to the spread of false narratives if maps with misinformation were released into the wild. We could reduce the amount of potential misinformation in maps by giving control to the analysts over the information sources used. However, in the same way, it would be possible for a user to create a narrative purely with misinformation and biased sources, which could then be used to spread.

Second, we note that our model works under the assumption that the user is always right for semantic interaction purposes. This can lead to the issue of augmenting any pre-existing biases that the user may have. In general, users are subject to their cognitive biases when they interact with text documents [366]. For example, users may suffer from confirmation bias, focusing on specific documents that confirm their pre-existing hypotheses [269, 375]. In turn, the semantic interactions performed by the user lead the model toward a biased narrative that reflects the user’s worldview.

Third, consider our use of narrative maps with social media data to investigate the accepted narratives of communities. There are two stakeholder groups in this instance: members of the studied communities and potential system users (such as analysts, journalists, and researchers). As with any system that could be used to monitor social movements, there is a risk that malicious actors exploit the system for nefarious purposes, such as tracking activists involved in organizing or participating in these movements or designing information operations to counter or support particular narratives. Concerning tracking issues, the system operates at an aggregate level, without user-level details, and utilizes non-confidential and publicly available data without requiring user-specific information. Therefore, it would be difficult to re-identify a user.

In this context, the release of the system and the distribution of its source code raises some ethical concerns. The system could be used by malicious actors to create and spread specific narratives to promote their harmful agendas. However, releasing the model and its code as

an open-source project will help aid in the reproducibility of our techniques and generally help toward transparency in research. Appropriate publishing of the source code could help ensure that the potential risks are outweighed by the benefits of sharing our narrative maps model in full.

Misinformation and Bias

Recent works have highlighted the need for future work to model source bias, information validity, transparency, and credibility as an effort to model and counter misinformation [140, 208, 390]. Our narrative model could be enhanced by including additional attributes in the representation itself or in the extraction algorithm.

Works on misinformation detection generally focus on the propagation structure and content to determine whether a certain article or publication contains misinformation [140, 390]. Other methods rely on crowdsourcing [140] to detect misinformative content early. However, these methods do not model misinformation as part of an overarching narrative. Instead, they focus on local elements (e.g., a specific event). Thus, a holistic narrative approach could be useful in this context.

The issue of misinformation is also highly relevant for a series of recent works on disaster tracking by using news narrative extraction [394, 408, 415]. However, none of these methods address this issue and rely on the underlying assumption that the set does not contain false or misleading information. Thus, creating a narrative extraction model that accounts for misinformation would be of vital importance in the context of disaster tracking.

One possible way to approach modeling this in our narrative maps approach is through the incorporation of *bias* and *credibility* attributes in the narrative maps. We could rate these attributes for each article using information about the source of the news article (i.e., source cues). We could get credibility information from resources such as Media Bias/Fact Check, which considers where news sources fall in the political spectrum (bias), and the veracity of their claims across news stories (factualness) [49, 379]. Furthermore, we can leverage other potential credibility cues based on the contents of the news article (i.e., message cues), such as the proper use of the *inverted pyramid structure* in breaking news and the appropriate definition of the *main event descriptors* (5W1H).

In the context of modeling bias in news narratives, it could be useful to compare the narratives generated by news sources with different biases (e.g., left- vs. right-leaning articles). One possible approach to this comparison problem would be to perform a qualitative evaluation of the different elements of the narrative (e.g., the main storyline, side stories, and important events). Another approach would be to attempt to compare these elements in a computational manner by extracting relevant keywords from each narrative or what elements they share.

10.2.7 Model Extensions and Integration

In this section, we present some possible extensions of the model, applications to other domains, and potential integration with other tasks.

Data Types and Multimodality

Throughout this work, we tested the narrative map model and the associated interactive AI techniques with mostly news-based data and one instance of social media data. However, we note that our method is grounded in the scholarly definitions of narratology. More specifically, we began with abstract theoretical definitions and moved toward concrete representations of the narrative’s essential elements. Therefore, our representation should be capable of capturing any type of narrative as defined in the narrative theory literature. So long as data is in the form of (event, timestamps) tuples, we should be able to construct a map representation using our algorithm. In accordance with this model, it would be possible to use our approach with synthetic data (e.g., synthetically generated documents or articles) assuming that it conforms to the aforementioned format requirements.

Moreover, we can generalize this approach to any kind of data provided it has a *total order* (e.g., temporal information) and that we can define a measure (i.e., the evaluation metric of our metaphor) that allows us to compute coherence through similarity and clustering. On a technical level, the generalizability of our method derives from the abstract mathematical formulation used to extract the map, which could be applied to other types of data, such as image data (e.g., constructing a narrative map from images based on an embedding representation and timestamps).

Another potential extension of our model would be the ability to handle *multimodal* data. In particular, being able to handle both text and image data at the same time would provide useful to model narratives in social media spaces, which rely on multimodal data in general [149]. In this context, we note that one of the limitations of our social media study in Chapter 7 was our inability to handle image and video data, as we simply focused on the textual content of the data. However, many of the posts included in the data set had embedded images, videos, or links to these elements. Thus, being able to handle multimodal data in a coherent way would provide additional value in understanding how narratives spread in social media. We note that such a model would require appropriate multimodal embeddings [392] to handle the different types of data.

Domains of Application

Focusing on applications with text data, there are several potential domains of applications beyond journalistic analysis tasks. Most of these applications seek to help with the issue of *information overload* in different contexts [323].

Disaster Management: To keep track of disasters or other similar adverse events, disaster management could benefit from utilizing extraction methods [394, 408, 414, 415]. In order to minimize losses caused by a disaster, one of the most important duties in disaster management is to analyze and comprehend situation updates effectively. This necessitates efficient methods for navigating a multitude of disaster-related documents, such as news articles and reports. Domain experts require condensed information on the catastrophe and its progression [219]. Thus, narrative maps with interactive AI could assist specialists in comprehending the evolving situation and formulating an appropriate strategy.

Financial Markets: The extraction of news narratives could aid financial analysts in comprehending the information landscape. In the context of financial analysis, market news is regarded as an essential data source [59]. In particular, the ability to comprehend and exploit the concealed information in raw news data could assist analysts in adapting their strategies and minimizing financial risk. Thus, narrative maps could help such analysts by providing an overview of the information landscape.

Intelligence Analysis and Open Source Intelligence: Another relevant field that could stand to benefit from narrative maps is intelligence analysis. In particular, intelligence analysts could use narrative maps to explore the underlying storylines in the data and find interesting connections between documents.

Particularly, we emphasize the potential application to open source intelligence (OSINT), which is intelligence derived from publicly accessible data [111]. OSINT data sources contain news articles alongside other types of documents. Thus, narrative maps with interactive AI could assist intelligence analysts in exploring the information landscape and locating significant events and connections in these data sets. In addition, these techniques could aid analysts in prediction tasks by supplying evidence and support [60]. In this context, comparing the narrative based on intelligence data to the narrative based on news articles could be an interesting application.

Temporal and Geographical Dimensions

We now discuss the use of two relevant dimensions in event and narrative modeling. In general, events can be defined using the main event descriptors (5W1H) [187]. This definition includes an explicit notion of temporal (When?) and geographic dimensions (Where?). These dimensions could be exploited to detect relevant connections between events. In fact, the notion of chronological order is inherent to our directed acyclic graph model [186] and our final coherence model includes a *temporal component* in the form of a penalty factor for temporally distant events [189].

However, our model does not account for geographical information in its formulation of coherence. This dimension is particularly useful for certain domains, such as disaster management [408, 415]. In fact, recent work has developed narrative extraction models that

incorporate geographical information in their notion of coherence [408] or through *location smoothness* constraints [415] (i.e., avoiding sudden jumps in location).

The geographical dimension allows us to track how narratives spread not only in temporal terms but how it spreads physically (e.g., tracking a disaster moving through a country or a social movement spreading from city to city). Thus, extending the model to consider the geographical dimension could prove valuable. In this context, we could add an additional factor for *geographical closeness* in our coherence definition or appropriate constraints in our linear program definition. We note that particular care would need to be applied to ensure that these additions do not affect the performance of the semantic interaction components of our model. Furthermore, the addition of the geographical dimension could help construct better explanations with our XAI components.

Collaborative Sensemaking

In our current approach, the narrative map model has been designed with a single user in mind. However, in practice, many analysis tasks require analysts to work in a team to perform collaborative sensemaking [54, 76]. In this context, our narrative maps model could be extended to facilitate this type of sensemaking.

The first possible extension would be to extend the model from a single graph into a multi-graph approach, where there could be multiple disconnected graphs. Each one of these graphs would be shared between multiple users and could be manipulated in a collaborative way (e.g., allowing all users to add and remove events). These graphs could represent different parts of the narrative or overlap, depending on the specific objectives of the system. Furthermore, we could link each graph through a higher-level structure (or even another narrative map) to show the relationship between them.

In this context, it could also be useful to be able to compare the narratives created by different users. Thus, the system should facilitate the comparison between narratives by identifying common elements and highlighting any differences. These elements could be identified through direct set differences or more complex techniques, such as graph edit distance [4].

Another possible extension that could be useful for collaborative sensemaking would be to implement multiple *views* of the narrative—which would act like database views [174], where instead of a virtual table showing a subset of data, we would have a “virtual” map showing a subset of events in a structured manner. Different users would use different views of the narrative, thus enabling better collaboration by only showing what is relevant to each user in each case. Finally, we note that these map views relate back to the idea of different discourses for the same underlying story from narrative theory.

Other Connection Types

Our model is based on three types of connections: similarity, entity, and topical. Moreover, the model relies on the underlying assumption that the data is chronologically ordered. But, there are other types of connections that could be handled by the model, such as connections based on causality relationships, logical inference, or cross-referencing between documents. Thus, natural extensions to the narrative map model include being able to handle these different types of connections. Furthermore, we could drop the chronological order assumption of our narrative model and allow for non-chronological connections.

These connections are hard to model in general. For example, determining whether one event was caused by another is a complex task. Work has been done in this area by connecting events through causality in a *single document*—where specific cues can be leveraged to obtain causality relationships (e.g., the use of the word “because”) [402]. However, this becomes harder when we try to account for causality relationships across *multiple documents*.

In this context, we can leverage resources from the multi-document summarization literature [231]. Some of the approaches that have been used to capture more complex relationships rely on graph-based structures, such as knowledge graphs [115], semantic link networks [225], and entailment graphs [245]. In this context, all of these approaches could be used to augment the capabilities of the narrative map model and capture more complex relationships between events (or entities) by incorporating additional knowledge into the extraction process.

Finally, there is the possible extension of including an *uncertainty* measure for edges. While our current model does have weights based on normalized coherence, these weights are not direct representations of uncertainty in the edges, but rather of how strong the connection is. In particular, one possible way of doing this could be to use fuzzy graphs [339] which have a natural notion of uncertainty in them. Furthermore, going back to the original route map metaphor that we used to design the maps, we note that fuzzy graphs have been applied in the context of real route maps [50]. Thus, we could leverage this approach and our original metaphor to develop a concept of fuzzy narrative maps.

Adjacent Narrative Tasks

We also consider adjacent tasks to computational narrative extraction, such as narrative forecasting, narrative analysis, and narrative generation.

Narrative generation is a fundamentally different task from extraction that seeks to create new fictional narratives, rather than extract a narrative that already exists (either fictional or non-fictional) [130, 133]. Moreover, these methods usually generate a single document. In this context, we could adapt our representational structure to provide the basis for the generation process, but it would not be a direct approach. Alternatively, assuming that we can split the generated narrative into multiple “documents” (e.g., event sentences with

timestamps), our extraction model could be used to validate whether the generated narrative has the desired structure and content.

Narrative forecasting (i.e., predicting the next events in the narrative) is a task that lies between extraction and generation, but its focus is on generating new events rather than on extracting the complete narrative [411]. This would be a particularly important task for the domain of intelligence analysis, which would benefit from being able to predict future events [96]. In this context, our narrative maps framework could be leveraged to predict future events (e.g., as an input for the generation method). However, as the current model does not account for the state of the narrative’s world, it would not be enough on its own for this task and would require support from another method that could represent the world’s state (e.g., knowledge graphs [266]).

Narrative analysis methods use existing extraction approaches to obtain a computational representation of the narrative and then use it to analyze the narrative [272, 314]. In this context, our narrative maps method could be directly plugged into the analysis pipeline and provide additional input for the analysis task.

Personal Narratives

Our narrative model and visualization method could be applied to personal narratives [237] where events represent past, present, or future occurrences in a person’s life. In particular, personal narratives have applications in healthcare [345] and psychology [263].

In the context of our work, personal narratives could be represented using the narrative maps framework. In particular, using our graph-based representation would allow users to visualize a coherent and structured representation of their life story. The model could be used by psychologists or medical personnel to analyze the personal narratives of their patients [263], providing them with an overview of their patients’ life events and allowing for a more customized treatment if required. However, the system would require extensive changes to properly handle this type of data. For example, the current internal representation relies on general text embeddings, while specialized representations would be required to handle the data and events of personal narratives [360].

As another potential application based on personal narrative maps, we could develop an AI assistant that could converse with the user and help them navigate through this representation. The personal narrative map could help users find new insights and influence their perception of their own life. Moreover, the insights they gather from this representation could help them navigate their current circumstances. Furthermore, our semantic interaction capabilities could allow users to customize their narrative by providing feedback to the underlying AI model. Moreover, the explainable AI component could help users understand why specific life events are connected in the narrative representation.

Graph Analysis and Graph Motifs

Another possible application of our model is to leverage more advanced graph-theoretical elements, such as *graph motifs*—frequent patterns in the structure of a graph [254]—to encode interesting properties and relationships of its elements [37]. These motifs could help us identify more complex patterns in the narrative.

Graph motif analysis has been successfully used in the context of determining text quality using a sentence-level graph that represents shared nouns [19]. Graph motifs have also been applied in the context of analyzing the language used by politicians [20]. Specifically, this analysis used graph motifs on a graph comprised of semantic frames [26].

Although there are works applying graph motifs on news events at the topic level, such as in the context of media and public attention [205], to date no work has applied graph motif analysis to event-level graphs, narrative maps, or similarly defined representations. Thus, it could be useful to perform graph motif analysis on a narrative map. In particular, we could extract framing patterns using graph motifs, which could allow us to discover statistically significant *frame motifs*, thus showing us how these narratives frame the events.

Script Learning and the Cloze Task

To further extend our narrative model, it could be helpful to borrow elements from the *script learning* literature [148], which is highly related to our notion of narrative extraction. However, script learning relies on a lower-level notion of events (i.e., verbs or predicates rather than documents). Scripts are a structured representation of events in a narrative, following a predicate form with a subject and an object [148]. In particular, script learning has a heavy focus on the idea of performing inferences, beyond just capturing the narrative.

In this context, we discuss the *Narrative Cloze Task* proposed by Chambers and Jurafsky [66], which evaluates narrative knowledge by removing an event from a narrative and attempting to reconstruct it from its context and surrounding events. This evaluation task is based on the basic Cloze Task [92] which requires completing a sentence with a missing word in the middle. This evaluation approach is suited for linear narrative representations, such as timelines, but it could be adapted to fit non-linear narrative representations, although it would require considering all the surrounding structure of the event, not just the preceding and subsequent events.

Summarization Tasks and Report Generation

The narrative map model can be seen from the perspective of *multi-document summarization* [239]. The graph representation could be interpreted as a structured summary of the narrative formed by all the documents in the data set. Thus, narrative maps can be used as a

summarization tool to identify high-level patterns in narratives. For example, by leveraging the main storyline and the important events we could generate a summary of the narrative based on these elements.

In particular, the graph representation used by our narrative maps could be understood as an extension of *timeline summarization* (TLS) [98, 407] models. In this context, we note that recent research has proposed variations on the traditional TLS task. In particular, Duan et al. [98] and Yu et al. [407] proposed the comparative TLS task and the Multi-TLS task, respectively. These two works demonstrate that simple linear representations of narratives are inherently restrictive unless they are applied to the simplest narratives. By leveraging our narrative map representation and adapting our extraction method, we could provide an alternative to both of these approaches. Thus, instead of generating a series of timeline summaries, we would use a graph-based structured summary to compare and contrast the different stories contained in the data set.

Furthermore, narrative map extraction could be used to generate reports, through template-based methods [283] or neural generative models [168]. In particular, these reports could be arranged based on the narrative structure and have sections based on the different storylines extracted by our method and a highlights section based on the important events. Furthermore, our model could be combined with pre-existing summarization methods to obtain better reports. Furthermore, through the application of our interactive AI framework, these reports could be customized by modifying the narrative map. In this context, generating text reports from text data based on our narrative map models would be *closing the loop* in terms of sensemaking.

End-to-End Narrative Analysis

Finally, continuing with our previous idea of *closing the loop*, our narrative map models can be considered part of an interactive *end-to-end narrative analysis suite* that considers the whole sensemaking loop. In particular, such a system would first extract events (e.g., by crawling the web and finding relevant elements [116]). Then, it would extract a world representation (e.g., in the form of a knowledge graph [266]). Afterward, the system would extract a narrative based on these events and knowledge, which is where the narrative maps would come in. Next, the system would identify potential hypotheses based on the extracted narrative map and the knowledge schema. Moreover, such a system could be complemented by other integrated approaches that analyze event collections [220] to provide a holistic view of the narrative. Finally, the system would generate a report based on the extracted narrative and the previously generated hypotheses.

Bibliography

- [1] H Porter Abbott. *The Cambridge introduction to narrative*. Cambridge University Press, One Liberty Plaza, New York, NY, USA, 2008.
- [2] H. Porter Abbott. Narrativity. In Peter Hühn, Jan Christoph Meister, John Pier, and Wolf Schmid, editors, *Handbook of Narratology*, pages 587–607. De Gruyter, Berlin, München, Boston, 2014. ISBN 9783110316469.
- [3] James Abello, Frank Van Ham, and Neeraj Krishnan. ASK-GraphView: A large scale graph visualization system. *IEEE transactions on visualization and computer graphics*, 12(5):669–676, 2006.
- [4] Zeina Abu-Aisheh, Romain Raveaux, Jean-Yves Ramel, and Patrick Martineau. An Exact Graph Edit Distance Algorithm for Solving Pattern Recognition Problems. In *4th Intl. Conf. on Pattern Recognition Applications and Methods 2015*, pages 1–9, Lisbon, Portugal, January 2015. HAL Open Science.
- [5] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.
- [6] Shmuel Agmon. The relaxation method for linear inequalities. *Canadian Journal of Mathematics*, 6:382–392, 1954.
- [7] Alfred V. Aho, Michael R Garey, and Jeffrey D. Ullman. The transitive reduction of a directed graph. *SIAM Journal on Computing*, 1(2):131–137, 1972.
- [8] Mina Akaishi, Kato Yoshikiyo, Ken Satoh, and Hori Koichi. Narrative based topic visualization for chronological data. In *2007 11th International Conference Information Visualization (IV’07)*, pages 139–144. IEEE, 2007.
- [9] Narges Akbarpour Roshan and Mohsen Abbasi. The impact of the US economic sanctions on health in Cuba. *International Journal of Resistive Economics*, 6(2):17–36, 2018.
- [10] Arwa I. Alhussain and Aqil M. Azmi. Automatic story generation: A survey of approaches. *ACM Comput. Surv.*, 54(5), May 2021. ISSN 0360-0300.
- [11] Meysam Alizadeh, Jacob N Shapiro, Cody Buntain, and Joshua A Tucker. Content-based features predict social media influence operations. *Science advances*, 6(30): eabb5824, 2020.

- [12] Antón L Allahar and Nelson P Valdés. The Bureaucratic Imperative: Economic and Political Challenges to Cuban Socialism in the Early 21st Century. *The CLR James Journal*, 19(1/2):392–422, 2013.
- [13] James Allan. *Topic detection and tracking: event-based information organization*, volume 12. Springer Science & Business Media, New York, 2012.
- [14] Tomás Alves, Carlota Dias, Daniel Gonçalves, Sandra Gama, and Joana Henriques-Calado. How personality and visual channels affect insight generation. In *2022 IEEE Evaluation and Beyond-Methodological Approaches for Visualization (BELIV)*, pages 47–55. IEEE, 2022.
- [15] Christopher Andrews and Chris North. Analyst’s workspace: An embodied sense-making environment for large, high-resolution displays. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 123–131. IEEE, 2012.
- [16] Christopher Andrews, Alex Endert, and Chris North. Space to think: large high-resolution displays for sensemaking. In *Proc. of CHI 2010*, pages 55–64, 2010.
- [17] Nicholas O Andrews and Edward A Fox. Recent developments in document clustering. Technical report, Department of Computer Science, Virginia Polytechnic Institute & State University, 2007.
- [18] Jeffery Ansah, Lin Liu, Wei Kang, Selasie Kwashie, Jixue Li, and Jiuyong Li. A graph is worth a thousand words: Telling event stories using timeline summarization graphs. In *The World Wide Web Conf., WWW ’19*, page 2565–2571, NY, USA, 2019. ACM. ISBN 9781450366748.
- [19] Thomas Arnold and Karsten Weihe. Network motifs may improve quality assessment of text documents. In *Proc. of TextGraphs-10: the Workshop on Graph-based Methods for Natural Language Processing*, pages 20–28, 2016.
- [20] Thomas Otmar Arnold. *Advanced Motif Analysis on Text Induced Graphs*. PhD thesis, Technische Universität Berlin, Berlin, Germany, 2018.
- [21] Eleftherios Avramidis. Rankeval: Open tool for evaluation of machine-learned ranking. *Prague Bull. Math. Linguistics*, 100:63–72, 2013.
- [22] Chris Baber, Dan Andrews, Tom Duffy, and Richard McMaster. Sensemaking as narrative: Visualization for collaboration. *VAW2011, University London College, VAW2011*: 7–8, 2011.
- [23] Mohamed Bahgat, Steve Wilson, and Walid Magdy. Towards using word embedding vector space for better cohort analysis. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 919–923, 2020.

- [24] Bahman Bahmani, Abdur Chowdhury, and Ashish Goel. Fast Incremental and Personalized PageRank. *Proc. VLDB Endow.*, 4(3):173–184, dec 2010. ISSN 2150-8097. doi: 10.14778/1929861.1929864. URL <https://doi.org/10.14778/1929861.1929864>.
- [25] Alok Baikadi and Rogelio E Cardona-Rivera. Towards finding the fundamental unit of narrative: A Proposal for the Narreme. In *The Third Workshop on Computational Models of Narrative*, pages 44–46, Boston, MA, USA, 2012. MIT CSAIL, MIT.
- [26] Collin F Baker, Charles J Fillmore, and John B Lowe. The Berkeley FrameNet project. In *Proc. of the 17th Int. Conf. on Computational linguistics-Volume 1*, pages 86–90. ACL, 1998.
- [27] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.
- [28] Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. We can detect your bias: Predicting the political ideology of news articles. In *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online, November 2020. ACL.
- [29] Kiran Kumar Bandeli, Muhammad Nihal Hussain, and Nitin Agarwal. A framework towards computational narrative analysis on blogs. In *Text2Story@ ECIR*, pages 63–69, Lisbon, Portugal, 2020. CEUR-WS.
- [30] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020. ISSN 1566-2535.
- [31] Federico Barrios, Federico López, Luis Argerich, and Rosita Wachenchauzer. Variations of the Similarity Function of TextRank for Automated Summarization. In *Argentine Symposium on Artificial Intelligence (ASAI 2015)*, 2015.
- [32] Michele Barry. Effect of the US embargo and economic decline on health in Cuba. *Annals of internal medicine*, 132(2):151–154, 2000.
- [33] Roland Barthes and Lionel Duisit. An introduction to the structural analysis of narrative. *New Literary History*, 6(2):237–272, 1975. ISSN 00286087, 1080661X.
- [34] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The Pushshift Reddit dataset. *CoRR*, abs/2001.08435, 2020. URL <https://arxiv.org/abs/2001.08435>.

- [35] Joeran Beel and Stefan Langer. An exploratory analysis of mind maps. In *Proceedings of the 11th ACM Symposium on Document Engineering, DocEng '11*, page 81–84, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450308632. doi: 10.1145/2034691.2034709. URL <https://doi.org/10.1145/2034691.2034709>.
- [36] Lee Anne Bell. *Storytelling for social justice: Connecting narrative and the arts in antiracist teaching*. Routledge, 2010.
- [37] Austin R Benson, David F Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.
- [38] Floris Bex, Henry Prakken, and Bart Verhey. Anchored narratives in reasoning about evidence. *Frontiers in Artificial Intelligence and Applications*, 152:11, 2006.
- [39] Yali Bian and Chris North. *DeepSI: Interactive Deep Learning for Semantic Interaction*, page 197–207. ACM, New York, NY, USA, 2021. ISBN 9781450380171.
- [40] Yali Bian, Michelle Dowling, and Chris North. Evaluating semantic interaction on word embeddings via simulation. *EEvaluation of Interactive Visual Machine Learning systems Workshop, VIS 2019*, 2019.
- [41] Yali Bian, John Wenskovitch, and Chris North. Deepva: Bridging cognition and computation through semantic interaction and deep learning. In *Proc. of the IEEE VIS Workshop MLUI*, 2020.
- [42] Giang Binh Tran, Mohammad Alrifai, and Dat Quoc Nguyen. Predicting relevant news events for timeline summaries. In *Proceedings of the 22nd International Conference on World Wide Web, WWW '13 Companion*, page 91–92, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320382.
- [43] David M. Blei and John D. Lafferty. Dynamic topic models. In *Proc. of the 23rd Intl. Conf. on Machine Learning*, page 113–120, NY, USA, 2006. ACM. ISBN 1595933832.
- [44] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- [45] Benjamin Samuel Bloom. *Taxonomy of educational objectives: The classification of educational goals*. Longman, Ann Arbor, Michigan, 1956.
- [46] Thomas Bögel and Michael Gertz. Time will tell: Temporal linking of news stories. In *Proc. of the 15th ACM/IEEE-CS Joint Conf. on Digital Libraries*, page 195–204, NY, USA, 2015. ACM. ISBN 9781450335942.
- [47] Sébastien Bogleux, Benoit Gaüzere, and Luc Brun. Graph edit distance as a quadratic program. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 1701–1706. IEEE, 2016.

- [48] Nadia Boukhelifa, Anastasia Bezerianos, and Evelyne Lutton. Evaluation of interactive machine learning systems. In *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*, pages 341–360. Springer International Publishing, Cham, 2018. ISBN 978-3-319-90403-0.
- [49] Lia Bozarth, Aparajita Saraf, and Ceren Budak. Higher ground? How groundtruth labeling impacts our understanding of fake news about the 2016 US presidential nominees. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 48–59, 2020.
- [50] Alexander Bozhenyuk, Stanislav Belyakov, Margarita Knyazeva, and Vitalii Bozheniuk. Temporal reachability in fuzzy graphs for geographic information systems. In Sergey Kovalev, Valery Tarassov, Vaclav Snasel, and Andrey Sukhanov, editors, *Proceedings of the Fourth International Scientific Conference “Intelligent Information Technologies for Industry” (IITI’19)*, pages 507–516, Cham, 2020. Springer International Publishing. ISBN 978-3-030-50097-9.
- [51] L. Bradel, J. Z. Self, A. Endert, M. S. Hossain, C. North, and N. Ramakrishnan. How analysts cognitively “connect the dots”. In *2013 IEEE International Conference on Intelligence and Security Informatics*, pages 24–26, New York, NY, USA, 2013. IEEE.
- [52] Lauren Bradel, Chris North, and Leanna House. Multi-model semantic interaction for text analytics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 163–172. IEEE, 2014.
- [53] Lauren Bradel, Nathan Wycoff, Leanna House, and Chris North. Big text visual analytics in sensemaking. In *2015 Big Data Visual Analytics (BDVA)*, pages 1–8. IEEE, 2015.
- [54] Verica Buchanan, Yafeng Lu, Nathan McNeese, Michael Steptoe, Ross Maciejewski, and Nancy Cooke. The role of teamwork in the analysis of big data: a study of visual analytics and box office prediction. *Big Data*, 5(1):53–66, 2017.
- [55] Kenneth Burke. *A grammar of motives*, volume 177. Univ of California Press, 155 Grand Ave. Suite 400. Oakland, CA, 1969.
- [56] Remo Aslak Burkhard and Michael Meier. Tube map visualization: Evaluation of a novel knowledge visualization application for the transfer of knowledge in long-term projects. *J. UCS*, 11(4):473–494, 2005.
- [57] Alyxander Burns, Cindy Xiong, Steven Franconeri, Alberto Cairo, and Narges Mahyar. How to evaluate data visualizations across different levels of understanding. In *2020 IEEE Workshop on Evaluation and Beyond - Methodological Approaches to Visualization*, pages 19–28, UT, USA, 2020. IEEE.

- [58] Tolga Buz and Gerard de Melo. Should You Take Investment Advice From WallStreet-Bets? A Data-Driven Approach. *arXiv preprint arXiv:2105.02728*, 2021.
- [59] Yi Cai, Haoran Xie, Raymond YK Lau, Qing Li, Tak-Lam Wong, and Fu Lee Wang. Temporal event searches based on event maps and relationships. *Applied soft computing*, 85:105750, 2019.
- [60] Roberto Camacho Barranco, Arnold P Boedihardjo, and M Shahriar Hossain. Analyzing evolving stories in news articles. *Intl. Journal of Data Science and Analytics*, 8(3):241–256, 2019.
- [61] Renata T Camargo, Verônica Agostini, Ariani Di Felippo, and Thiago AS Pardo. Manual typification of source texts and multi-document summaries alignments. *Procedia-Social and Behavioral Sciences*, 95:498–506, 2013.
- [62] Dallas Card, Amber E. Boydston, Justin H. Gross, Philip Resnik, and Noah A. Smith. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2072. URL <https://aclanthology.org/P15-2072>.
- [63] Marc Cavazza and David Pizzi. Narratology for interactive storytelling: A critical introduction. In Stefan Göbel, Rainer Malkewitz, and Ido Iurgel, editors, *Technologies for Interactive Digital Storytelling and Entertainment*, pages 72–83, Berlin, Heidelberg, 2006. Springer. ISBN 978-3-540-49935-0.
- [64] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder. *CoRR*, abs/1803.11175: 1–7, 2018.
- [65] Kunal Chakma and Amitava Das. A 5W1H based annotation scheme for semantic role labeling of English tweets. *Computación y Sistemas*, 22(3):747—755, 2018.
- [66] Nathanael Chambers and Dan Jurafsky. Unsupervised learning of narrative event chains. In *Proc. of ACL-08: HLT*, pages 789–797, 2008.
- [67] Remco Chang, Caroline Ziemkiewicz, Tera Marie Green, and William Ribarsky. Defining insight for visual analytics. *IEEE Computer Graphics and Applications*, 29(2): 14–17, 2009.
- [68] Chien Chin Chen and Meng Chang Chen. TSCAN: A Novel Method for Topic Summarization and Content Anatomy. In *Proc. of the 31st Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, page 579–586, NY, USA, 2008. ACM. ISBN 9781605581644.

- [69] Chien Chin Chen and Meng Chang Chen. TSCAN: A content anatomy approach to temporal topic summarization. *IEEE transactions on Knowledge and Data Engineering*, 24(1):170–183, 2012.
- [70] Chien Chin Chen, Yao-Tsung Chen, Yeali Sun, and Meng Chang Chen. Life cycle modeling of news events using aging theory. In Nada Lavrač, Dragan Gamberger, Hendrik Blockeel, and Ljupčo Todorovski, editors, *Machine Learning: ECML 2003*, pages 47–59, Berlin, Heidelberg, 2003. Springer. ISBN 978-3-540-39857-8.
- [71] Jie Chen, Zhendong Niu, and Hongping Fu. A multi-news timeline summarization algorithm based on aging theory. In Reynold Cheng, Bin Cui, Zhenjie Zhang, Ruichu Cai, and Jia Xu, editors, *Web Technologies and Applications*, pages 449–460, Cham, 2015. Springer. ISBN 978-3-319-25255-1.
- [72] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O’Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proc. of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, page 1–12, New York, NY, USA, 2019. ACM. ISBN 9781450359702.
- [73] Hai Leong Chieu and Yoong Keok Lee. Query based event extraction along a timeline. In *Proc. of the 27th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR ’04, page 425–432, NY, USA, 2004. ACM. ISBN 1581138814.
- [74] In Kwon Choi, Taylor Childers, Nirmal Kumar Raveendranath, Swati Mishra, Kyle Harris, and Khairi Reda. Concept-driven visual analytics: An exploratory study of model- and hypothesis-based reasoning with visualizations. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, page 1–14, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702.
- [75] Arjun Choudhry, Mandar Sharma, Pramod Chundury, Thomas Kapler, Derek WS Gray, Naren Ramakrishnan, and Niklas Elmqvist. Once upon a time in visualization: Understanding the use of textual narratives for causality. *IEEE Transactions on Visualization and Computer Graphics*, 2020.
- [76] Haeyong Chung, Chris North, Jessica Zeitz Self, Sharon Chu, and Francis Quek. Vis-reporter: facilitating information sharing for collaborative sensemaking on multiple displays. *Personal and Ubiquitous Computing*, 18:1169–1186, 2014.
- [77] Rob Churchill and Lisa Singh. The evolution of topic modeling. *ACM Comput. Surv.*, 54(10s), nov 2022. ISSN 0360-0300.
- [78] Giovanni Luca Ciampaglia. Fighting fake news: a role for computational social science in the fight against digital misinformation. *Journal of Computational Social Science*, 1(1):147–153, 2018.

- [79] Sarah Cohen, James T Hamilton, and Fred Turner. Computational journalism. *Communications of the ACM*, 54(10):66–71, 2011.
- [80] Kevyn Collins-Thompson, Soo Young Rieh, Carl C. Haynes, and Rohail Syed. Assessing learning outcomes in web search: A comparison of tasks and query strategies. In *Proc. of the 2016 ACM on Conf. on Human Information Interaction and Retrieval*, page 163–172, NY, USA, 2016. ACM. ISBN 9781450337519.
- [81] Kristin A Cook and James J Thomas. Illuminating the path: The research and development agenda for visual analytics. Technical report, Pacific Northwest National Lab.(PNNL), Richland, WA, 2005.
- [82] Richard A Couto. Narrative, free space, and political leadership in social movements. *The Journal of Politics*, 55(1):57–79, 1993.
- [83] Clara Crivellaro, Rob Comber, John Bowers, Peter C Wright, and Patrick Olivier. A pool of dreams: facebook, politics and the emergence of a social movement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3573–3582, 2014.
- [84] Zeyu Dai, Himanshu Taneja, and Ruihong Huang. Fine-grained structure-based news genre categorization. In *Proc. of the Workshop Events and Stories in the News 2018*, pages 61–67, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.
- [85] Barbara Dancygier and Lieven Vandelanotte. Discourse viewpoint as network. In Barbara Dancygier, Wei lun Lu, and Arie Verhagen, editors, *Viewpoint and the Fabric of Meaning: Form and Use of Viewpoint Tools across Languages and Modalities*, pages 13–40. De Gruyter Mouton, Berlin, Boston, 2016. ISBN 9783110365467.
- [86] Joseph E Davis. Narrative and social movements. *Stories of change: Narrative and social movements*, pages 3–29, 2002.
- [87] Joseph E Davis. *Stories of change: Narrative and social movements*. SUNY Press, 2012.
- [88] Claes H De Vreese. News framing: Theory and typology. *Information design journal & document design*, 13(1):51–62, 2005. ISSN 1871-1138.
- [89] Franca Debole and Fabrizio Sebastiani. Supervised term weighting for automated text categorization. In *Text mining and its applications*, pages 81–97. Springer, 2004.
- [90] Samuel Y Dennis III. On the hyper-Dirichlet type 1 and hyper-Liouville distributions. *Communications in Statistics-Theory and Methods*, 20(12):4069–4081, 1991.
- [91] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018.

- [92] Tony Deyes. Towards an authentic ‘discourse cloze’. *Applied Linguistics*, 5(2):128–137, 1984.
- [93] Robert P Dilworth. *A Decomposition Theorem for Partially Ordered Sets*, pages 139–144. Birkhäuser Boston, Boston, MA, 1987. ISBN 978-0-8176-4842-8.
- [94] Jill P Dimond, Michaelanne Dye, Daphne LaRose, and Amy S Bruckman. Hollaback! the role of storytelling online in a social movement organization. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 477–490, 2013.
- [95] David L Donoho and Jared Tanner. Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proceedings of the national academy of sciences*, 102(27):9446–9451, 2005.
- [96] Raimundo F Dos Santos Jr, Sumit Shah, Arnold Boedihardjo, Feng Chen, Chang-Tien Lu, Patrick Butler, and Naren Ramakrishnan. A framework for intelligence analysis using spatio-temporal storytelling. *GeoInformatica*, 20(2):285–326, 2016.
- [97] Michelle Dowling, John Wenskovitch, Peter Hauck, Adam Binford, Nicholas Polys, and Chris North. A bidirectional pipeline for semantic interaction. In *Proc. Workshop on Machine Learning from User Interaction for Visualization and Analytics (at IEEE VIS 2018)*, volume 11, page 74, 2018.
- [98] Yijun Duan, Adam Jatowt, and Masatoshi Yoshikawa. Comparative timeline summarization via dynamic affinity-preserving random walk. In *ECAI 2020*, pages 1778–1785. IOS Press, Santiago de Compostela, Spain, 2020.
- [99] Vincent Dubois and Cécile Bothorel. Transitive reduction for social network analysis and visualization. In *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI’05)*, pages 128–131. IEEE, 2005.
- [100] Michaelanne Dye, David Nemer, Josiah Mangiameli, Amy S Bruckman, and Neha Kumar. El Paquete Semanal: The Week’s Internet in Havana. In *Proc. of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.
- [101] Peter Eades and Mao Lin Huang. Navigating clustered graphs using force-directed methods. In *Graph Algorithms and Applications 2*, pages 191–215. World Scientific, 5 Toh Tuck Link, Singapore, 596224, Singapore, 2004.
- [102] John Ellson, Emden R. Gansner, Eleftherios Koutsofios, Stephen C. North, and Gordon Woodhull. Graphviz and dynagraph — static and dynamic graph drawing tools. In Michael Jünger and Petra Mutzel, editors, *Graph Drawing Software*, pages 127–148. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-642-18638-7.
- [103] Alex Endert, Chao Han, Dipayan Maiti, Leanna House, and Chris North. Observation-level interaction with statistical models for visual analytics. In *2011 IEEE conference on visual analytics science and technology (VAST)*, pages 121–130. IEEE, 2011.

- [104] Alex Endert, Patrick Fiaux, and Chris North. Semantic interaction for visual text analytics. In *Proc. of CHI 2012*, CHI '12, page 473–482, New York, NY, USA, 2012. ACM. ISBN 9781450310154.
- [105] Alex Endert, Patrick Fiaux, and Chris North. Semantic interaction for sensemaking: inferring analytical reasoning for model steering. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2879–2888, 2012.
- [106] Alex Endert, Seth Fox, Dipayan Maiti, Scotland Leman, and Chris North. The semantics of clustering: Analysis of user-generated spatializations of text documents. In *Proc. of the International Working Conference on Advanced Visual Interfaces*, AVI '12, page 555–562, New York, NY, USA, 2012. ACM. ISBN 9781450312875.
- [107] Alex Endert, M Shahriar Hossain, Naren Ramakrishnan, Chris North, Patrick Fiaux, and Christopher Andrews. The human is the loop: new directions for visual analytics. *Journal of intelligent information systems*, 43(3):411–435, 2014.
- [108] Gunes Erkan and Dragomir Radev. Lexpagerank: Prestige in multi-document text summarization. In *Proc. of the 2004 Conf. on Empirical Methods in Natural Language Processing*, pages 365–371, Barcelona, Spain, 2004. ACL.
- [109] Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479, 2004.
- [110] Liana Ermakova, Jean Valère Cossu, and Josiane Mothe. A survey on evaluation of summarization methods. *Information processing & management*, 56(5):1794–1814, 2019.
- [111] João Rafael Gonçalves Evangelista, Renato José Sassi, Márcio Romero, and Domingos Napolitano. Systematic literature review to investigate the application of open source intelligence (osint) with artificial intelligence. *Journal of Applied Security Research*, 16(3):345–369, 2021.
- [112] Anthony Faiola. Cubans take to the streets for the biggest anti-government protests in decades. *The Washington Post*, Jul 2021. URL https://web.archive.org/web/20210712033912if_/https://www.washingtonpost.com/world/2021/07/11/cuba-protests/.
- [113] Matthew E Falagas, Eleni I Pitsouni, George A Malietzis, and Georgios Pappas. Comparison of pubmed, scopus, web of science, and google scholar: strengths and weaknesses. *The FASEB journal*, 22(2):338–342, 2008.
- [114] Christos Faloutsos, Kevin S. McCurley, and Andrew Tomkins. Fast discovery of connection subgraphs. In *Proc. of the Tenth ACM SIGKDD Int. Conf. on Knowledge*

- Discovery and Data Mining*, KDD '04, page 118–127, New York, NY, USA, 2004. ACM. ISBN 1581138881.
- [115] Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. Using local knowledge graph construction to scale seq2seq models to multi-document inputs. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, 2019.
 - [116] Mohamed MG Farag, Sunshin Lee, and Edward A Fox. Focused crawler for events. *International Journal on Digital Libraries*, 19:3–19, 2018.
 - [117] Paolo Federico, Markus Wagner, Alexander Rind, Albert Amor-Amorós, Silvia Miksch, and Wolfgang Aigner. The role of explicit knowledge: A conceptual model of knowledge-assisted visual analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 92–103, Phoenix, AZ, USA, 2017. IEEE.
 - [118] Aidah Nakakande Ferguson, Matija Franklin, and David Lagnado. Explanations that backfire: Explainable artificial intelligence can cause information overload. In *Proc. of the Annual Meeting of the Cognitive Science Society*, volume 44, 2022.
 - [119] Mark A Finlayson and Steven R Corman. The military interest in narrative. *Sprache und Datenverarbeitung*, 37(1-2):173–191, 2013.
 - [120] Pascal Fletcher. Cuba protests: Frustration at government runs deep. *BBC*, Jul 2021. URL <https://web.archive.org/web/20210714110213/https://www.bbc.com/news/world-latin-america-57823130>.
 - [121] Jonathan P. Forsyth, Rachel M. Bittner, Michael Musick, and Juan P. Bello. Improving and adapting finite state transducer methods for musical accompaniment. In Richard Dudas, editor, *41st Int. Computer Music Conf., ICMC 2015*, pages 290–297, Denton, TX, January 2015. Int. Computer Music Association.
 - [122] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.
 - [123] Emden R Gansner and Stephen C North. An open graph visualization system and its applications to software engineering. *Software: practice and experience*, 30(11):1203–1233, 2000.
 - [124] Emden R Gansner, Eleftherios Koutsofios, Stephen C North, and K-P Vo. A technique for drawing directed graphs. *IEEE Transactions on Software Engineering*, 19(3):214–230, 1993.
 - [125] Emden R Gansner, Yehuda Koren, and Stephen C North. Topological fisheye views for visualizing large graphs. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):457–468, 2005.

- [126] Marshall Louis Ganz. The power of story in social movements. In *Proceedings of the 2001 Annual Meeting of the American Sociological Association*. American Sociological Association, 2001.
- [127] Yang Gao, Yang Wang, Luyang Liu, Yidi Guo, and Heyan Huang. Neural abstractive summarization fusing by global generative topics. *Neural Computing and Applications*, 32:5049–5058, 2020.
- [128] Félix García, M^a Ángeles Moraga, Manuel Serrano, and Mario Piattini. Visualisation environment for global software development management. *IET Software*, 9(2):51–64, 2015.
- [129] Joseph L Gastwirth. The estimation of the lorenz curve and gini index. *The review of economics and statistics*, pages 306–316, 1972.
- [130] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018.
- [131] Delia Gavrilu. From the print press to online press: Constraints and liberties of the journalistic discourse. *Procedia - Social and Behavioral Sciences*, 63:263 – 270, 2012. ISSN 1877-0428. 4th Intl. Conf: Paradigms of the Ideological Discourse 2012.
- [132] Gérard Genette. *Narrative discourse: An essay in method*, volume 3. Cornell University Press, 512 East State St., Ithaca, NY, 1983.
- [133] Pablo Gervás, Eugenio Concepción, Carlos León, Gonzalo Méndez, and Pablo Delatorre. The long path to narrative generation. *IBM Journal of Research and Development*, 63(1):8–1, 2019.
- [134] Demian Gholipour Ghalandari and Georgiana Ifrim. Examining the state-of-the-art in news timeline summarization. *CoRR*, abs/2005.10107, 2020.
- [135] Helen Gibson. Acquisition and preparation of data for osint investigations. In Babak Akhgar, P. Saskia Bayerl, and Fraser Sampson, editors, *Open Source Intelligence Investigation: From Strategy to Implementation*, pages 69–93. Springer, Cham, 2016. ISBN 978-3-319-47671-1.
- [136] Ilka H Gleibs. Are all “research fields” equal? Rethinking practice for the use of data from crowdsourcing market places. *Behavior Research Methods*, 49(4):1333–1342, 2017.
- [137] Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Jamie Callan. Creating and evaluating multi-document sentence extract summaries. In *Proc. of the Ninth Intl. Conf. on Information and Knowledge Management, CIKM '00*, page 165–172, NY, USA, 2000. ACM. ISBN 1581133200.

- [138] David Gotz, Michelle X Zhou, and Vikram Aggarwal. Interactive visual synthesis of analytic knowledge. In *2006 IEEE Symposium On Visual Analytics Science And Technology*, pages 51–58. IEEE, 2006.
- [139] R. Guha, Ravi Kumar, D. Sivakumar, and Ravi Sundaram. Unweaving a web of documents. In *Proc. of the Eleventh ACM SIGKDD Intl. Conf. on Knowledge Discovery in Data Mining*, KDD '05, page 574–579, NY, USA, 2005. ACM. ISBN 159593135X.
- [140] Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. The future of misinformation detection: New perspectives and trends. *CoRR*, abs/1909.03654, 2019.
- [141] Hua Guo, Steven R Gomez, Caroline Ziemkiewicz, and David H Laidlaw. A case study using visualization interaction logs and insight metrics to understand how analysts arrive at insights. *IEEE transactions on visualization and computer graphics*, 22(1): 51–60, 2015.
- [142] Fethallah Hadjila, Amine Belabed, and Mohammed Merzoug. Flexible service discovery based on multiple matching algorithms. *Int. Journal of Web Engineering and Technology*, 14(4):315–340, 2019.
- [143] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab. (LANL), 2008.
- [144] Johanna Doppler Haider, Patrick Seidler, Margit Pohl, Neesha Kodagoda, Rick Adderley, and BL William Wong. How analysts think: sense-making strategies in the analysis of temporal evolution and criminal network structures and activities. In *Proc. of the Human Factors and Ergonomics Society Annual Meeting*, volume 61, pages 193–197. SAGE Publications: Los Angeles, CA, 2017.
- [145] Jeffry Halverson, Steven Corman, and H Lloyd Goodall. *Master narratives of Islamist extremism*. Springer, 175 5th Ave., New York, NY, USA, 2011.
- [146] Felix Hamborg, Soeren Lachnit, Moritz Schubotz, Thomas Hepp, and Bela Gipp. Giveme5W: Main Event Retrieval from News Articles by Extraction of the Five Journalistic W Questions. In *Transforming Digital Worlds*, pages 356–366, Cham, 2018. Springer International Publishing. ISBN 978-3-319-78105-1.
- [147] Felix Hamborg, Corinna Breitingner, and Bela Gipp. Giveme5W1H: A Universal System for Extracting Main Events from News Articles, 2019.
- [148] Yi Han, Linbo Qiao, Jianming Zheng, Hefeng Wu, Dongsheng Li, and Xiangke Liao. A survey of script learning. *Frontiers Inf. Technol. Electron. Eng.*, 22(3):341–373, 2021.

- [149] Haiyan Hao and Yan Wang. Leveraging multimodal social media data for rapid disaster damage assessment. *International Journal of Disaster Risk Reduction*, 51:101760, 2020. ISSN 2212-4209. doi: <https://doi.org/10.1016/j.ijdr.2020.101760>. URL <https://www.sciencedirect.com/science/article/pii/S2212420920312620>.
- [150] Tim Harrower. *Inside reporting*, volume 310. McGraw-Hill Education, 1221 Avenue of the Americas, New York, NY 10020, 2010.
- [151] Hossein Hassani and Emmanuel Sirmal Silva. The role of ChatGPT in data science: how AI-assisted conversational interfaces are revolutionizing the field. *Big data and cognitive computing*, 7(2):62, 2023.
- [152] David Hauser, Gabriele Paolacci, and Jesse J Chandler. Common concerns with mturk as a participant pool: Evidence and solutions, Sep 2018.
- [153] Taher H Haveliwala. Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, 15(4):784–796, 2003.
- [154] Jinwon Ho and Rong Tang. Towards an optimal resolution to information overload: An infomediary approach. In *Proc. of the 2001 Int. ACM SIGGROUP Conference on Supporting Group Work*, GROUP '01, page 91–96, New York, NY, USA, 2001. ACM. ISBN 1581132948.
- [155] Nathan Oken Hodas and Alex Endert. Adding semantic information into data models by learning domain expertise from user interaction. *arXiv preprint arXiv:1604.02935*, 2016.
- [156] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proc. of the 22nd Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR '99, page 50–57, NY, USA, 1999. ACM. ISBN 1581130961.
- [157] Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE trans. on visualization and computer graphics*, 25(8):2674–2693, 2018.
- [158] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings. *Convolutional Neural Networks and Incremental Parsing*, 7(1), 2017.
- [159] Benjamin D Horne, Sibel Adali, and Sujoy Sikdar. Identifying the social signals that drive online discussions: A case study of Reddit communities. In *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–9. IEEE, 2017.

- [160] M Shahriar Hossain, Joseph Gresock, Yvette Edmonds, Richard Helm, Malcolm Potts, and Naren Ramakrishnan. Connecting the dots between pubmed abstracts. *PloS one*, 7(1):e29509, 2012.
- [161] Mahmud Shahriar Hossain, Christopher Andrews, Naren Ramakrishnan, and Chris North. Helping intelligence analysts make connections. In *Workshops at the Twenty-Fifth AAAI Conf. on Artificial Intelligence*, pages 1–10, San Francisco, CA, USA, 2011. AAAI.
- [162] Leanna House, Scotland Leman, and Chao Han. Bayesian visual analytics: Bava. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 8(1):1–13, 2015.
- [163] Po Hu, Minlie Huang, Peng Xu, Weichang Li, Adam K. Usadi, and Xiaoyan Zhu. Generating breakpoint-based timeline overview for news topic retrospection. In *2011 IEEE 11th Intl. Conf. on Data Mining*, pages 260–269, Vancouver, Canada, 2011. IEEE.
- [164] Po Hu, Min-Lie Huang, and Xiao-Yan Zhu. Exploring the interactions of storylines from informative news events. *J. of Computer Science and Technology*, 29(3):502–518, 2014.
- [165] Xinran Hu, Lauren Bradel, Dipayan Maiti, Leanna House, and Chris North. Semantics of directly manipulating spatializations. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2052–2059, 2013.
- [166] Dongping Huang, Shuyu Hu, Yi Cai, and Huaqing Min. Discovering event evolution graphs based on news articles relationships. In *2014 IEEE 11th Intl. Conf. on e-Business Engineering*, pages 246–251, Guangzhou, China, 2014. IEEE, IEEE.
- [167] Lifu Huang and Lian’en Huang. Optimized event storyline generation based on mixture-event-aspect model. In *Proc. of the 2013 Conf. on Empirical Methods in NLP*, pages 726–735, Seattle, WA, USA, 2013. ACL.
- [168] Yongkang Huang and Hongzhi Yu. Research on text generation techniques combining machine learning and deep learning. In *Proceedings of the 3rd Asia-Pacific Conference on Image Processing, Electronics and Computers*, pages 319–326, 2022.
- [169] F Hughes and D Schum. Discovery-proof-choice, the art and science of the process of intelligence analysis-preparing for the future of intelligence analysis. *Washington, DC: Joint Military Intelligence College*, 2003.
- [170] Jessica Hullman, Steven Drucker, Nathalie Henry Riche, Bongshin Lee, Danyel Fisher, and Eytan Adar. A deeper understanding of sequence in narrative visualization. *IEEE Transactions on visualization and computer graphics*, 19(12):2406–2415, 2013.

- [171] Jessica Hullman, Robert Kosara, and Heidi Lam. Finding a clear path: Structuring strategies for visualization sequences. In *Computer Graphics Forum*, volume 36, pages 365–375. Wiley Online Library, 2017.
- [172] Thore Husfeldt. Fully dynamic transitive closure in plane dags with one source and one sink. In Paul Spirakis, editor, *Algorithms — ESA '95*, pages 199–212, Berlin, Heidelberg, 1995. Springer Berlin Heidelberg. ISBN 978-3-540-44913-3.
- [173] Vitavin Ittipanuvat, Katsuhide Fujita, Yuya Kajikawa, Junichiro Mori, and Ichiro Sakata. Measuring relatedness between technology and social issue citation networks. In *ISPIM Conference Proceedings*, page 1. The International Society for Professional Innovation Management, 2012.
- [174] Johannes Jakob, Alexander Königs, and Andy Schürr. Non-materialized model view specification with triple graph grammars. In *ICGT*, pages 321–335. Springer, 2006.
- [175] Prerna Juneja and Tanushree Mitra. Human and technological infrastructures of fact-checking. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), nov 2022. doi: 10.1145/3555143. URL <https://doi.org/10.1145/3555143>.
- [176] Joseph Kahne, Erica Hodgins, and Elyse Eidman-Aadahl. Redesigning civic education for the digital age: Participatory politics and the pursuit of democratic engagement. *Theory & Research in Social Education*, 44(1):1–35, 2016.
- [177] Jagadish S Kallimani, KG Srinivasa, and B Eswara Reddy. Summarizing news paper articles: experiments with ontology-based, customized, extractive text summary and word scoring. *Cybernetics and Information Technologies*, 12(2):34–50, 2012.
- [178] Y. Kang, C. Gorg, and J. Stasko. Evaluating visual analytics systems for investigative analysis: Deriving design principles from a case study. In *2009 IEEE Symposium on Visual Analytics Science and Technology*, pages 139–146, New York, NY, USA, 2009. IEEE.
- [179] Youn-ah Kang, Carsten Görg, and John Stasko. How can visual analytics assist investigative analysis? design implications from an evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 17(5):570–583, 2010.
- [180] Antoni Kapcia. *A short history of revolutionary Cuba: revolution, power, authority and the state from 1959 to the present day*. Bloomsbury Publishing, 2020.
- [181] Dmytro Karamshuk, Tetyana Lokot, Oleksandr Pryymak, and Nishanth Sastry. Identifying partisan slant in news articles and Twitter during political crises. In *8th International Conference on Social Informatics, SocInfo 2016*, pages 257–272. Springer-Verlag Berlin Heidelberg, 2016.

- [182] Peter Kareithi and Nixon Kariithi. Black hawk down and the framing of somalia: Pop culture as news and news as pop fiction. *Africa media review*, 16(1):1–20, 2008.
- [183] Kerry Kartchner and Jennie Johnson. *Strategic culture and weapons of mass destruction: culturally based insights into comparative national security policymaking*. Springer, 2009.
- [184] Stuart J Kaufman. Narratives and symbols in violent mobilization: the palestinian-israeli case. *Security Studies*, 18(3):400–434, 2009.
- [185] Suzanne Keen. *Narrative form*. Springer, 2003.
- [186] Brian Keith Norambuena and Tanushree Mitra. Narrative maps: An algorithmic approach to represent and extract information narratives. In *Proc. ACM Hum.-Comput. Interact.*, volume 4, page 33 pages, New York, NY, USA, December 2020. ACM.
- [187] Brian Keith Norambuena, Michael Horning, and Tanushree Mitra. Evaluating the Inverted Pyramid Structure through Automatic 5W1H Extraction and Summarization. In *Proc. of the 2020 Computation + Journalism Symposium*, pages 1–7, Boston, MA, USA, 2020. C + J 2020.
- [188] Brian Keith Norambuena, Eric Krokos, Sarah Joseph, and Chris North. Mixed Multi-Model Interactive AI Framework for Narrative Sensemaking, 2023. In Review.
- [189] Brian Keith Norambuena, Tanushree Mitra, and Chris North. Mixed multi-model semantic interaction for graph-based narrative visualizations. In *28th International Conference on Intelligent User Interfaces*, New York, NY, USA, 2023. ACM. ISBN 979-8-4007-0106-1/23/03.
- [190] Brian Keith Norambuena, Tanushree Mitra, and Chris North. A survey on event-based news narrative extraction. *CSUR*, 2023.
- [191] Brian Felipe Keith Norambuena and Tanushree Mitra. Narrative maps: An algorithmic approach to represent and extract information narratives. *Proc. of the ACM on Human-Computer Interaction*, 4(CSCW3):1–33, 2021.
- [192] Brian Felipe Keith Norambuena, Tanushree Mitra, and Chris North. Design guidelines for narrative maps in sensemaking tasks. *Information Visualization*, 21(3):220–245, 0.
- [193] Brian Felipe Keith Norambuena, Tanushree Mitra, and Chris North. Narrative sense-making: Strategies for narrative maps construction. In *2021 IEEE Visualization Conference (VIS)*, pages 181–185, 2021.
- [194] Brian Felipe Keith Norambuena, Tanushree Mitra, and Chris North. Narrative sense-making: Strategies for narrative maps construction. In *2021 IEEE Visualization Conference (VIS)*, pages 181–185, 2021.

- [195] Brian Felipe Keith Norambuena, Tanushree Mitra, and Chris North. Characterizing Social Movement Narratives in Online Communities: The 2021 Cuban Protests on Reddit. *Computational Journalism C+J*, 2022.
- [196] Shahedul Huq Khandkar. Open coding. *University of Calgary*, 23:2009, 2009.
- [197] Arpit Khurdiya, Lipika Dey, Nidhi Raj, and Sk Mirajul Haque. Multi-perspective linking of news articles within a repository. In *Twenty-Second Intl. Joint Conf. on Artificial Intelligence*, pages 2281–2286, Barcelona, Spain, 2011. AAAI.
- [198] Dongwoo Kim and Alice Oh. Topic chains for understanding a news corpus. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 163–176, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-19437-5.
- [199] Catherine Kohler Riessman. Analysis of personal narratives. In *Qualitative research in social work*, chapter 7, pages 168–191. Columbia University Press, 61 West 62 St., New York, NY, USA, 2000.
- [200] Rikard Konig, Ulf Johansson, and Lars Niklasson. G-rex: A versatile framework for evolutionary data mining. In *2008 IEEE International Conference on Data Mining Workshops*, pages 971–974, 2008.
- [201] Delia Koo and Delia Koo. Linear programming—duality and sensitivity analysis. *Elements of Optimization: With Applications in Economics and Business*, pages 102–127, 1977.
- [202] Yubo Kou, Yong Ming Kow, Xinning Gui, and Waikuen Cheng. One social movement, two social media sites: A comparative study of public discourses. *Computer Supported Cooperative Work (CSCW)*, 26(4):807–836, 2017.
- [203] Yong Ming Kow, Yubo Kou, Bryan Semaan, and Waikuen Cheng. Mediating the undercurrents: Using social media to sustain a social movement. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3883–3894, 2016.
- [204] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.
- [205] Haewoon Kwak, Jisun An, Joni Salminen, Soon-Gyo Jung, and Bernard J Jansen. What we read, what we search: Media attention and public attention among 193 countries. In *Proc. of the 2018 World Wide Web Conference*, pages 893–902, 2018.
- [206] Ben Kybartas and Rafael Bidarra. A survey on story generation techniques for authoring computational narratives. *IEEE Transactions on Computational Intelligence and AI in Games*, 9(3):239–253, 2016.

- [207] Moreno La Quatra, Luca Cagliero, Elena Baralis, Alberto Messina, and Maurizio Montagnuolo. Summarize dates first: A paradigm shift in timeline summarization. In *Proc. of the 44th Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, page 418–427, NY, USA, 2021. ACM. ISBN 9781450380379.
- [208] Philippe Laban and Marti A Hearst. newsLens: building and visualizing long-ranging news stories. In *Proc. of the Events and Stories in the News Workshop*, pages 1–9, Vancouver, Canada, 2017. ACL.
- [209] Vincent Labatut and Xavier Bost. Extraction and analysis of fictional character networks: A survey. *ACM Comput. Surv.*, 52(5), September 2019. ISSN 0360-0300.
- [210] Daniel Lafond, René Proulx, Alexis Morris, William Ross, Alexandre Bergeron-Guyard, and Mihaela Ulieru. HCI dilemmas for context-aware support in intelligence analysis. In *Adapt. 2014, Sixth Int. Conf. Adapt. Self-Adaptive Syst. Appl*, pages 68–72, 2014.
- [211] Nancy K. Lankton and D. Harrison McKnight. What does it mean to trust facebook? examining technology and interpersonal trust beliefs. *SIGMIS Database*, 42(2):32–54, may 2011. ISSN 0095-0033.
- [212] Jeffrey A Larsen and Jeannie L Johnson. Comparative strategic cultures curriculum project: assessing strategic culture as a methodological approach to understanding wmd decision-making by states and non-state actors. Technical report, Science Applications International Corp McLean VA, 2006.
- [213] Walter S. Lasecki, Mitchell Gordon, Danai Koutra, Malte F. Jung, Steven P. Dow, and Jeffrey P. Bigham. Glance: Rapidly coding behavioral video with the crowd. In *Proc. of the 27th Annual ACM Symposium on User Interface Software and Technology, UIST '14*, page 551–562, New York, NY, USA, 2014. ACM. ISBN 9781450330695.
- [214] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. The science of fake news. *Science*, 359(6380):1094–1096, 2018. ISSN 0036-8075.
- [215] Reinoud Leenders and Steven Heydemann. Popular mobilization in Syria: opportunity and threat, and the social networks of the early risers. *Mediterranean politics*, 17(2): 139–159, 2012.
- [216] Sam N. Lehman-Wilzig and Michal Seletzky. Hard news, soft news, ‘general’ news: The necessity and utility of an intermediate classification. *Journalism*, 11(1):37–56, 2010.

- [217] Scotland C Leman, Leanna House, Dipayan Maiti, Alex Endert, and Chris North. Visual to parametric interaction (V2PI). *PloS one*, 8(3):e50474, 2013.
- [218] Jiwei Li and Sujian Li. Evolutionary hierarchical dirichlet process for timeline summarization. In *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 556–560, Sofia, Bulgaria, 2013. ACL.
- [219] Lei Li and Tao Li. An empirical study of ontology-based multi-document summarization in disaster management. *IEEE Transactions on Systems, Man, and Cybernetics: systems*, 44(2):162–171, 2013.
- [220] Liuqing Li. *Event-related collections understanding and services*. PhD thesis, Doctoral dissertation, Virginia Tech, 2020.
- [221] Rumeng Li, Tao Wang, and Xun Wang. *Tracking Events Using Time-dependent Hierarchical Dirichlet Tree Model*, pages 550–558. SIAM, Vancouver, Canada, 2015.
- [222] Tianyi Li, Asmita Shah, Kurt Luther, and Chris North. Crowdsourcing intelligence analysis with context slices. In *CHI 2018 Sensemaking Workshop*, 2018.
- [223] Tianyi Li, Chandler J Manns, Chris North, and Kurt Luther. Dropping the baton? Understanding errors and bottlenecks in a crowdsourced sensemaking pipeline. *Proc. of the ACM on Human-Computer Interaction*, 3(CSCW):1–26, 2019.
- [224] Vicki Li. The Race Talk on Reddit: “China Virus” and its Effect on Asian-Americans. *Moore Undergraduate Research Apprentice Program (MURAP)*, 2020.
- [225] Wei Li, Lei He, and Hai Zhuge. Abstractive news summarization based on event semantic link network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 236–246, 2016.
- [226] Yiming Liao, Shuguang Wang, and Dongwon Lee. Wilson: A divide and conquer approach for fast and effective news timeline summarization. In Yannis Velegrakis, Yannis Velegrakis, Demetris Zeinalipour, Panos K. Chrysanthis, Panos K. Chrysanthis, and Francesco Guerra, editors, *Advances in Database Technology - EDBT 2021*, Advances in Database Technology - EDBT, pages 635–645, Nicosia, Cyprus, 2021. OpenProceedings.org.
- [227] Chen Lin, Chun Lin, Jingxuan Li, Dingding Wang, Yang Chen, and Tao Li. Generating event storylines from microblogs. In *Proc. of the 21st ACM Intl. Conf. on Information and Knowledge Management*, CIKM ’12, page 175–184, NY, USA, 2012. ACM. ISBN 9781450311564.
- [228] Fu-ren Lin and Chia-hao Liang. Topic retrospection with storyline-based summarization on news reports. In *PACIS 2006 Proc.*, Kuala Lumpur, Malaysia, 2006. AIS.

- [229] Fu-ren Lin and Chia-Hao Liang. Storyline-based summarization for news topic retrospection. *Decision Support Systems*, 45(3):473–490, 2008.
- [230] Fu-ren Lin, Feng-mei Huang, and Chia-hao Liang. Individualized storyline-based news topic retrospection. In *PACIS 2007 Proc.*, page 140, Auckland, New Zealand, 2007. AIS.
- [231] Hui Lin and Vincent Ng. Abstractive summarization: A survey of the state of the art. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9815–9822, 2019.
- [232] Bang Liu, Di Niu, Kunfeng Lai, Linglong Kong, and Yu Xu. Growing story forest online from massive breaking news. In *Proc. of the 2017 ACM on Conf. on Information and Knowledge Management*, page 777–785, NY, USA, 2017. ACM. ISBN 9781450349185.
- [233] Bang Liu, Fred X Han, Di Niu, Linglong Kong, Kunfeng Lai, and Yu Xu. Story forest: Extracting events and telling stories from breaking news. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 14(3):1–28, 2020.
- [234] Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/K19-1047. URL <https://aclanthology.org/K19-1047>.
- [235] Anna-Pia Lohfink, Simon D. Duque Anton, Heike Leitte, and Christoph Garth. Knowledge rocks: Adding knowledge assistance to visualization systems. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1117–1127, 2022.
- [236] Hans Peter Luhn. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165, 1958.
- [237] Stephanie M Lukin, Kevin Bowden, Casey Barackman, and Marilyn A Walker. Personabank: a corpus of personal narratives and their story intention graphs. In *10th conference on International Language Resources and Evaluation (LREC’16)*, pages 1026–1033. European Language Resources Association, 2016.
- [238] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [239] Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z Sheng. Multi-document summarization via deep learning techniques: A survey. *ACM Computing Surveys*, 55(5):1–37, 2022.

- [240] Michaela Mahlberg. *English general nouns: A corpus theoretical approach*, volume 20. John Benjamins Publishing, 2005.
- [241] Inderjeet Mani. Computational modeling of narrative. *Synthesis Lectures on Human Language Technologies*, 5(3):1–142, 2012.
- [242] Audrey McAvoy. Hawaii group wants to defend licenses for foreign fishermen. *Associated Press News*, 2017. URL <https://apnews.com/00962efef9e941ee8563cd24855c1fb8>.
- [243] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017.
- [244] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv preprint arXiv:1802.03426, 2018.
- [245] Yashar Mehdad, Giuseppe Carenini, and Raymond Ng. Abstractive summarization of spoken and written conversations based on phrasal queries. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1220–1230, 2014.
- [246] Qiaozhu Mei and ChengXiang Zhai. Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In *Proc. of the Eleventh ACM SIGKDD Intl. Conf. on Knowledge Discovery in Data Mining*, KDD '05, page 198–207, NY, USA, 2005. ACM. ISBN 159593135X.
- [247] Qiaozhu Mei, Jian Guo, and Dragomir Radev. DivRank: The Interplay of Prestige and Diversity in Information Networks. In *Proc. of the 16th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, KDD '10, page 1009–1018, NY, USA, 2010. ACM. ISBN 9781450300551.
- [248] Melvin Mencher and Wendy P Shilton. *News reporting and writing*. Brown & Benchmark Publishers, 2011.
- [249] T. Mens. On the complexity of software systems. *Computer*, 45(08):79–81, aug 2012. ISSN 1558-0814.
- [250] David Meyer. Claiming credit: Stories of movement influence as outcomes. *Mobilization: An International Quarterly*, 11(3):281–298, 2006.
- [251] Rada Mihalcea. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proc. of the ACL Interactive Poster and Demonstration Sessions*, pages 170–173, 2004.

- [252] Rada Mihalcea and Paul Tarau. TextRank: Bringing order into text. In *Proc. of the 2004 Conf. on empirical methods in natural language processing*, pages 404–411, Barcelona, Spain, 2004. ACL.
- [253] Rada Mihalcea, Courtney Corley, and Carlo Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, pages 775–780, Boston, MA, USA, 2006. AAAI.
- [254] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298 (5594):824–827, 2002.
- [255] Anne-Lyse Myriam Minard, Manuela Speranza, Eneko Agirre, Itziar Aldabe, Marieke van Erp, Bernardo Magnini, German Rigau, and Ruben Urizar. Semeval-2015 task 4: Timeline: Cross-document event ordering. In *9th Intl. workshop on semantic evaluation*, pages 778–786, CO, USA, 2015. ACL.
- [256] David TZ Mindich. *Just the facts: How “objectivity” came to define American journalism*. NYU Press, 2000.
- [257] Alister Miskimmon, Ben O’loughlin, and Laura Roselle. *Strategic narratives: Communication power and the new world order*. Routledge, 711 3rd Ave. #8, New York, NY, USA, 2014.
- [258] Alister Miskimmon, Ben O’Loughlin, and Laura Roselle. *Forging the world: Strategic narratives and international relations*. University of Michigan Press, 839 Greene St., Ann Arbor, MI, 2017.
- [259] Tanushree Mitra and Eric Gilbert. Credbank: A large-scale social media corpus with associated credibility annotations. *Int. AAAI Conf. on Web and Social Media*, 2015: 258–267, 2015.
- [260] Tanushree Mitra, C.J. Hutto, and Eric Gilbert. Comparing person- and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proc. of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI ’15, page 1345–1354, New York, NY, USA, 2015. ACM. ISBN 9781450331456.
- [261] N Moratanch and S Chitrakala. A survey on extractive text summarization. In *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*, pages 1–6. IEEE, 2017.
- [262] Karl Mosler. Multidimensional indices and orders of diversity. *Community ecology*, 2 (2):137–143, 2001.
- [263] Michael Murray. Narrative psychology and narrative analysis. In *Qualitative research in psychology: Expanding perspectives in methodology and design*. American Psychological Association, 2003.

- [264] Nona Naderi and Graeme Hirst. Classifying frames at the sentence level in news articles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 536–542, Varna, Bulgaria, September 2017. INCOMA Ltd. doi: 10.26615/978-954-452-049-6_070. URL https://doi.org/10.26615/978-954-452-049-6_070.
- [265] Ramesh Nallapati, Ao Feng, Fuchun Peng, and James Allan. Event threading within news topics. In *Proc. of the Thirteenth ACM Int. Conf. on Information and Knowledge Management, CIKM '04*, page 446–453, New York, NY, USA, 2004. ACM. ISBN 1581138741.
- [266] Rungsiman Nararatwong, Natthawut Kertkeidkachorn, and Ryutaro Ichise. Knowledge graph visualization: challenges, framework, and implementation. In *2020 IEEE Third International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 174–178. IEEE, 2020.
- [267] Alexis R Neigel, Justine P Caylor, Sue E Kase, Michelle T Vanni, and Jefferson Hoye. The role of trust and automation in an intelligence analyst decisional guidance paradigm. *Journal of Cognitive Engineering and Decision Making*, 12(4):239–247, 2018.
- [268] Kiem-Hieu Nguyen, Xavier Tannier, and Véronique Moriceau. Ranking Multidocument Event Descriptions for Building Thematic Timelines. In *COLING 2014, the 25th Intl. Conf. on Computational Linguistic*, pages 1208 – 1217, Dublin, Ireland, 2014. HAL Open Science.
- [269] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.
- [270] A. Conrad Nied, Leo Stewart, Emma Spiro, and Kate Starbird. Alternative narratives of crisis events: Communities and social botnets engaged on social media. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW '17 Companion*, page 263–266, New York, NY, USA, 2017. ACM. ISBN 9781450346887.
- [271] Chris North, Purvi Saraiya, and Karen Duca. A comparison of benchmark task and insight evaluation methods for information visualization. *Information Visualization*, 10(3):162–181, 2011.
- [272] Nir Ofek, Sándor Darányi, and Lior Rokach. Linking motif sequences with tale types by machine learning. In *2013 Workshop on Computational Models of Narrative*, pages 166–182, Germany, 2013. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, Dagstuhl Publishing.

- [273] Takashi Ogata. Computational and cognitive approaches to narratology from the perspective of narrative generation. In *Computational and cognitive approaches to narratology*, pages 1–74. IGI Global, Hershey, PA, USA, 2016.
- [274] Yasuhiro Ogawa, Masaki Mori, and Katsuhiko Toyama. Recall-oriented evaluation metrics for consistent translation of japanese legal sentences. In Manabu Okumura, Daisuke Bekki, and Ken Satoh, editors, *New Frontiers in Artificial Intelligence*, pages 141–154, Berlin, Heidelberg, 2012. Springer. ISBN 978-3-642-32090-3.
- [275] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.
- [276] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419, 2010.
- [277] Kristen Parton, Kathleen McKeown, Robert Eric Coyne, Mona T Diab, Ralph Grishman, Dilek Hakkani-Tür, Mary Harper, Heng Ji, Wei Yun Ma, Adam Meyers, Rosenthal Sara, Ang Sun, Gokhan Tur, Wei Xu, and Sibel Yamam. Who, What, When, Where, Why? Comparing Multiple Approaches to the Cross-Lingual 5W Task. In *Proc. of the Joint Conf. of the 47th Annual Meeting of the ACL and the 4th Intl. Joint Conf. on NLP of the AFNLP*, pages 423–431, Suntec, Singapore, August 2009. Association for Computational Linguistics.
- [278] Arian Pasquali, Ricardo Campos, Alexandre Ribeiro, Brenda Santana, Alípio Jorge, and Adam Jatowt. TLS-Covid19: A New Annotated Corpus for Timeline Summarization. In Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani, editors, *Advances in Information Retrieval*, pages 497–512, Cham, 2021. Springer International Publishing. ISBN 978-3-030-72113-8.
- [279] Robin Pemantle. Vertex-reinforced random walk. *Probability Theory and Related Fields*, 92(1):117–136, 1992.
- [280] Randolph H Pherson and Richards J Heuer Jr. *Structured analytic techniques for intelligence analysis*. Cq Press, 2020.
- [281] Wim Pijls and Rob Potharst. Another Note on Dilworth’s Decomposition Theorem. *Journal of Discrete Mathematics*, 2013:692–645, 2013. ISSN 2090-9837.
- [282] Peter Pirolli and Stuart Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proc. of international conference on intelligence analysis*, volume 5, pages 2–4, McLean, VA, USA, 2005. International Conference on Intelligence Analysis.

- [283] Colin M. Potts and Arnav Jhala. Narraport: Narrative-based interactions and report generation with large datasets. In Alex Mitchell and Mirjam Vosmeer, editors, *Interactive Storytelling*, pages 118–127, Cham, 2021. Springer International Publishing. ISBN 978-3-030-92300-6.
- [284] Gerald Prince. *A dictionary of narratology*. U of Nebraska Press, 2003.
- [285] Kent Puckett. *Narrative theory*. Cambridge University Press, One Liberty Plaza, New York, NY, USA, 2016.
- [286] James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. TimeML: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3: 28–34, 2003.
- [287] Horst Pöttker. News and its communicative quality: the inverted pyramid—when and why did it appear? *Journalism Studies*, 4(4):501–511, 2003.
- [288] J Qiu and C Tang. Topic oriented semi-supervised document clustering. In *Proc. of SIGMOD 2007 Workshop IDAR*, pages 57–62, Beijing, China, 2007. DISC.
- [289] Jiangtao Qiu, Changjie Tang, Tao Zeng, Shaojie Qiao, Jie Zuo, Peng Chen, and Jun Zhu. A novel text classification approach based on enhanced association rule. In Reda Alhajj, Hong Gao, Jianzhong Li, Xue Li, and Osmar R. Zaiane, editors, *Advanced Data Mining and Applications*, pages 252–263, Berlin, Heidelberg, 2007. Springer. ISBN 978-3-540-73871-8.
- [290] Jiangtao Qiu, Chuan Li, Shaojie Qiao, Taiyong Li, and Jun Zhu. Timeline analysis of web news events. In Changjie Tang, Charles X. Ling, Xiaofang Zhou, Nick J. Cercone, and Xue Li, editors, *Advanced Data Mining and Applications*, pages 123–134, Berlin, Heidelberg, 2008. Springer. ISBN 978-3-540-88192-6.
- [291] Kumar Ravi and Vadlamani Ravi. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems*, 89:14–46, 2015.
- [292] Byron Reeves and Clifford Nass. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK*, 10:236605, 1996.
- [293] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL <https://arxiv.org/abs/1908.10084>.
- [294] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *CoRR*, abs/1908.10084, 2019.

- [295] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. ACM. ISBN 9781450342322.
- [296] Whitman Richards, Patrick Henry Winston, and Mark Alan Finlayson. Advancing computational models of narrative. Technical report, MIT-CSAIL, 2009.
- [297] Nathalie Henry Riche, Christophe Hurter, Nicholas Diakopoulos, and Sheelagh Carpendale. *Data-driven storytelling*. CRC Press, 2018.
- [298] J. T. Rigsby and Daniel Barbará. Storytelling with signal injection: Focusing stories with domain knowledge. In Petra Perner, editor, *Machine Learning and Data Mining in Pattern Recognition*, pages 425–439, Cham, 2018. Springer. ISBN 978-3-319-96133-0.
- [299] Shlomith Rimmon-Kenan. *Narrative fiction: Contemporary poetics*. Routledge, 711 3rd Ave. #8, New York, NY, USA, 2003.
- [300] Corina Robbins. The weaponization of social media. Technical report, Mercy Corps, 2019.
- [301] Anthony C Robinson. Collaborative synthesis of visual analytic results. In *2008 IEEE Symposium on Visual Analytics Science and Technology*, pages 67–74. IEEE, 2008.
- [302] Frances Robles. Cubans Denounce 'Misery' in Biggest Protests in Decades. *The New York Times*, Jul 2021. URL <https://web.archive.org/web/20210711212619/https://www.nytimes.com/2021/07/11/world/americas/cuba-crisis-protests.html>.
- [303] Laura Roselle, Alister Miskimmon, and Ben O'Loughlin. Strategic narrative: A new means to understand soft power. *Media, War & Conflict*, 7(1):70–84, 2014.
- [304] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proc. of the 2007 joint Conf. on empirical methods in natural language processing and computational natural language learning*, pages 410–420, Prague, Czech Republic, 2007. ACL.
- [305] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 269–276, 1993.
- [306] James Ryan. Grimes' fairy tales: A 1960s story generator. In Nuno Nunes, Ian Oakley, and Valentina Nisi, editors, *Interactive Storytelling*, pages 89–103, Cham, 2017. Springer. ISBN 978-3-319-71027-3.

- [307] Dominik Sacha, Andreas Stoffel, Florian Stoffel, Bum Chul Kwon, Geoffrey Ellis, and Daniel A. Keim. Knowledge generation model for visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1604–1613, 2014.
- [308] Niloufar Salehi, Lilly C. Irani, Michael S. Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and Clickhappier. We Are Dynamo: Overcoming Stalling and Friction in Collective Action for Crowd Workers. In *Proc. of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 1621–1630, New York, NY, USA, 2015. ACM. ISBN 9781450331456.
- [309] Jaakko Salo. A genre analytical comparison of real and fake news. Master’s thesis, University of Helsinki, Faculty of Arts, Department of Modern Languages, 2019.
- [310] Laura Sappelsa, H Van Dyke Parunak, and Sven Brueckner. The generic narrative space model as an intelligence analysis tool. *American Intelligence Journal*, 31(2): 69–78, 2013.
- [311] Purvi Saraiya, Chris North, and Karen Duca. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE transactions on visualization and computer graphics*, 11(4):443–456, 2005.
- [312] Robert E Schapire. The boosting approach to machine learning: An overview. *Non-linear estimation and classification*, pages 149–171, 2003.
- [313] Susan S Schiffman, Susan B Schiffman, M Lance Reynolds, and Forrest W Young. *Introduction to multidimensional scaling: Theory, methods and applications*. Academic Press Incorporated, 1981.
- [314] Jason Schlachter, Alicia Ruvinsky, Luis Asencios Reynoso, Sathappan Muthiah, and Naren Ramakrishnan. Leveraging topic models to develop metrics for evaluating the quality of narrative threads extracted from news stories. *Procedia Manufacturing*, 3: 4028–4035, 2015.
- [315] Wolf Schmid. Narrativity and eventfulness. In Tom Kindt and Hans-Harald Müller, editors, *What Is Narratology? Questions and Answers Regarding the Status of a Theory*, pages 17–34. De Gruyter, Berlin, New York, 2003. ISBN 9783110202069.
- [316] Edward Segel and Jeffrey Heer. Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics*, 16(6):1139–1148, 2010.
- [317] Jessica Zeitz Self, Michelle Dowling, John Wenskovitch, Ian Crandell, Ming Wang, Leanna House, Scotland Leman, and Chris North. Observation-level and parametric interaction for high-dimensional data analysis. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–36, 2018.

- [318] Doris Seyser and Michael Zeiller. Scrollytelling—an analysis of visual storytelling in online journalism. In *2018 22nd International Conference Information Visualisation (IV)*, pages 401–406. IEEE, 2018.
- [319] Dafna Shahaf and Carlos Guestrin. Connecting the dots between news articles. In *Proc. of the 16th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD '10*, page 623–632, New York, NY, USA, 2010. ACM. ISBN 9781450300551.
- [320] Dafna Shahaf and Carlos Guestrin. Connecting two (or less) dots: Discovering structure in news articles. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4):1–31, 2012.
- [321] Dafna Shahaf, Carlos Guestrin, and Eric Horvitz. Trains of thought: Generating information maps. In *Proc. of the 21st Int. Conf. on World Wide Web, WWW '12*, page 899–908, New York, NY, USA, 2012. ACM. ISBN 9781450312295.
- [322] Dafna Shahaf, Carlos Guestrin, and Eric Horvitz. Metro maps of information. *SIG-WEB Newsl.*, Spring, 2013. ISSN 1931-1745.
- [323] Dafna Shahaf, Jaewon Yang, Caroline Suen, Jeff Jacobs, Heidi Wang, and Jure Leskovec. Information cartography: Creating zoomable, large-scale maps of information. In *Proc. of the 19th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD '13*, page 1097–1105, New York, NY, USA, 2013. ACM. ISBN 9781450321747.
- [324] Dafna Shahaf, Carlos Guestrin, Eric Horvitz, and Jure Leskovec. Information cartography. *Communications of the ACM*, 58(11):62–73, 2015.
- [325] Azra Shamim, Vimala Balakrishnan, and Muhammad Tahir. Evaluation of opinion visualization techniques. *Information visualization*, 14(4):339–358, 2015.
- [326] Chao Shen and Tao Li. Multi-document summarization via the minimum dominating set. In *Proc. of the 23rd Intl. Conf. on Computational Linguistics (Coling 2010)*, pages 984–992, Beijing, China, 2010. ACL.
- [327] Haifeng Shen, Kewen Liao, Zhibin Liao, Job Doornberg, Maoying Qiao, Anton van den Hengel, and Johan W. Verjans. Human-AI Interactive and Continuous Sensemaking: A Case Study of Image Classification Using Scribble Attention Maps. In *Extended Abstracts of 2021 CHI, CHI EA '21*, New York, NY, USA, 2021. ACM. ISBN 9781450380959.
- [328] Ryan P Shepherd. Gaming Reddit’s algorithm: r/the_donald, amplification, and the rhetoric of sorting. *Computers and Composition*, 56:102572, 2020.

- [329] Frank M. Shipman and Catherine C. Marshall. Formality considered harmful: Experiences, emerging themes, and directions on the use of formal representations in interactive systems. *Computer Supported Cooperative Work (CSCW)*, 8(4):333–352, Dec 1999. ISSN 1573-7551.
- [330] Pamela J. Shoemaker. *News Values: Reciprocal Effects on Journalists and Journalism*, pages 1–9. American Cancer Society, 2017. ISBN 9781118783764.
- [331] Barbara Herrnstein Smith. Narrative versions, narrative theories. *Critical inquiry*, 7(1):213–236, 1980.
- [332] Wayne S Smith. Cuba: Confronting the US Embargo. *Bulletin of the Atomic Scientists*, 55(5):68–68, 1999.
- [333] Anna S Smoliarova, Svetlana S Bodrunova, Alexandr V Yakunin, Ivan Blekanov, and Alexey Maksimov. Detecting pivotal points in social conflicts via topic modeling of twitter content. In *International Conference on Internet Science*, pages 61–71. Springer, 2018.
- [334] Sandeep Soni, Tanushree Mitra, Eric Gilbert, and Jacob Eisenstein. Modeling factuality judgments in social media text. In *Proc. of the 52nd Annual Meeting of the ACL (Volume 2: Short Papers)*, pages 415–420, Baltimore, Maryland, June 2014. ACL.
- [335] John Stasko, Carsten Görg, and Zhicheng Liu. Sensemaking across text documents: human-centered, visual exploration with jigsaw. In *Sensemaking Workshop@ CHI 2008*, 2008.
- [336] Max Steg, Karlo HR Slot, and Federico Pianzola. Computational detection of narrativity: A comparison using textual features and reader response. In *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 105–114, 2022.
- [337] David Steinmann, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Machines explaining linear programs. *arXiv preprint arXiv:2206.07194*, 2022.
- [338] Chairs Constantine Stephanidis, Gavriel Salvendy, Members of the Group Margherita Antona, Jessie Y. C. Chen, Jianming Dong, Vincent G. Duffy, Xiaowen Fang, Cali Fidopias, Gino Fragomeni, Limin Paul Fu, Yinni Guo, Don Harris, Andri Ioannou, Kyeong ah (Kate) Jeong, Shin’ichi Konomi, Heidi Krömker, Masaaki Kurosu, James R. Lewis, Aaron Marcus, Gabriele Meiselwitz, Abbas Moallem, Hirohiko Mori, Fiona Fui-Hoon Nah, Stavroula Ntoa, Pei-Luen Patrick Rau, Dylan Schmorrow, Keng Siau, Norbert Streitz, Wentao Wang, Sakae Yamamoto, Panayiotis Zaphiris, and Jia Zhou. Seven HCI Grand Challenges. *International Journal of Human-Computer Interaction*, 35(14):1229–1269, 2019.

- [339] MS Sunitha and Sunil Mathew. Fuzzy graph theory: a survey. *Annals of Pure and Applied mathematics*, 4(1):92–110, 2013.
- [340] Russell Swan and James Allan. Automatic generation of overview timelines. In *Proc. of the 23rd Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR '00, page 49–56, NY, USA, 2000. ACM. ISBN 1581132263.
- [341] Don R. Swanson. Complementary structures in disjoint science literatures. In *Proc. of the 14th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR '91, page 280–289, NY, USA, 1991. ACM. ISBN 0897914481.
- [342] Joanna Szostek. Defence and Promotion of Desired State Identity in Russia’s Strategic Narrative. *Geopolitics*, 22(3):571–593, 2017.
- [343] Xavier Tannier and Véronique Moriceau. Building event threads out of multiple news articles. In *Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 958–967, Seattle, Washington, USA, October 2013. ACL.
- [344] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet processes. *Journal of the american statistical association*, 101(476):1566–1581, 2006.
- [345] Marie Thompson. Narrative mapping: Participant-generated visual methodology for health communication research and pedagogy. *Health Communication*, 36(5):630–638, 2021. doi: 10.1080/10410236.2020.1733228. PMID: 32160770.
- [346] Elizabeth A. Thomson, Peter R. R. White, and Philip Kitley. “Objectivity” and “Hard News” reporting across cultures. *Journalism Studies*, 9(2):212–228, 2008.
- [347] Mikhail Tikhomirov and Boris Dobrov. News timeline generation: Accounting for structural aspects and temporal nature of news stream. In Leonid Kalinichenko, Yannis Manolopoulos, Oleg Malkov, Nikolay Skvortsov, Sergey Stupnikov, and Vladimir Sukhomlin, editors, *Data Analytics and Management in Data Intensive Domains*, pages 267–280, Cham, 2018. Springer. ISBN 978-3-319-96553-6.
- [348] Chao Tong, Richard Roberts, Rita Borgo, Sean Walton, Robert S. Laramée, Kodzo Wegba, Aidong Lu, Yun Wang, Huamin Qu, Qiong Luo, and Xiaojuan Ma. Storytelling and visualization: An extended survey. *Information*, 9(3):65, 2018.
- [349] Giang Tran, Mohammad Alrifai, and Eelco Herder. Timeline summarization from relevant headlines. In Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr, editors, *Advances in Information Retrieval*, pages 245–256, Cham, 2015. Springer. ISBN 978-3-319-16354-3.

- [350] Giang Binh Tran, Tuan Tran, Nam-Khanh Tran, Mohammad Alrifai, and Nattiya Kanhabua. Leveraging learning to rank in an optimization framework for timeline summarization. In *SIGIR 2013 Workshop on Time-aware Information Access (TAIA'2013)*, Dublin, Ireland, 2013. ACM.
- [351] Tuan A. Tran, Claudia Niederee, Nattiya Kanhabua, Ujwal Gadiraju, and Avishek Anand. Balancing novelty and salience: Adaptive learning to rank entities for timeline summarization of high-impact events. In *Proc. of the 24th ACM Intl. on Conf. on Information and Knowledge Management, CIKM '15*, page 1201–1210, NY, USA, 2015. ACM. ISBN 9781450337946.
- [352] Jacob Turton, Robert Elliott Smith, and David Vinson. Deriving Contextualised Semantic Features from BERT (and Other Transformer Model) Embeddings. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 248–262, 2021.
- [353] Barbara Tversky. Narratives of space, time, and life. *Mind & Language*, 19(4):380–392, 2004.
- [354] Naohiko Uramoto and Koichi Takeda. A method for relating multiple newspaper articles by using graphs, and its application to webcasting. In *Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th Intl. Conf. on Computational Linguistics - Volume 2, ACL '98/COLING '98*, page 1307–1313, USA, 1998. Association for Computational Linguistics.
- [355] Josep Valls-Vargas, Jichen Zhu, and Santiago Ontañón. From computational narrative analysis to generation: A preliminary review. In *Proc. of the 12th Intl. Conf. on the Foundations of Digital Games, FDG '17*, NY, USA, 2017. ACM. ISBN 9781450353199.
- [356] Katherine Van Koevinger, Austin R Benson, and Jon Kleinberg. Frozen binomials on the web: Word ordering and language conventions in online text. In *Proceedings of The Web Conference 2020*, pages 606–616, 2020.
- [357] Jim Vanides, Yue Yin, Miki Tomita, and Maria Araceli Ruiz-Primo. Concept maps. *Science Scope*, 28(8):27–31, 2005.
- [358] Michael Vauth, Hans Ole Hatzel, Evelyn Gius, and Chris Biemann. Automated event annotation in literary texts. In *CHR*, pages 333–345, 2021.
- [359] Stefania Vicari. Twitter and non-elites: Interpreting power dynamics in the life story of the (#)BRCA Twitter stream. *Social Media+ Society*, 3(3):2056305117733224, 2017.
- [360] Prashanth Vijayaraghavan and Deb Roy. Modeling human motives and emotions from personal narratives using external knowledge and entity tracking. In *Proceedings of the Web Conference 2021, WWW '21*, page 529–540, New York, NY, USA, 2021.

- Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3449997. URL <https://doi.org/10.1145/3442381.3449997>.
- [361] Adriana S Vivacqua, Roberto Stelling, Ana Cristina B Garcia, and Livia C Gouvea. Explanations and Sensemaking with AI and HCI. In *Proc. of the IX Latin American Conference on Human Computer Interaction, CLIHC '19*, New York, NY, USA, 2020. ACM. ISBN 9781450376792.
- [362] Christian von Sikorski and Jörg Matthes. Framing and journalism, 02 2020. URL <https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-817>.
- [363] Ellen M Voorhees. Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. *Information Processing & Management*, 22(6):465–476, 1986.
- [364] Karel Vredenburg, Ji-Ye Mao, Paul W. Smith, and Tom Carey. A survey of user-centered design practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '02, page 471–478, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 1581134533. doi: 10.1145/503376.503460. URL <https://doi.org/10.1145/503376.503460>.
- [365] Paul Wake. Narrative and narratology. In *The Routledge Companion to Critical and Cultural Theory*, pages 39–52. Routledge, NY, USA, 2013.
- [366] Emily Wall, Leslie M Blaha, Lyndsey Franklin, and Alex Endert. Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In *2017 ieee conference on visual analytics science and technology (vast)*, pages 104–115. IEEE, 2017.
- [367] Patrick Walters. Beyond the inverted pyramid: Teaching the writing and all-formats coverage of planned and unplanned breaking news. *Teaching Journalism & Mass Communication*, 7(2):9–22, 2017.
- [368] Chengyu Wang, Xiaofeng He, and Aoying Zhou. Event phase oriented news summarization. *World Wide Web*, 21(4):1069–1092, 2018.
- [369] Lu Wang, Claire Cardie, and Galen Marchetti. Socially-informed timeline generation for complex events. In *Proc. of the 2015 Conference of the North American Chapter of the ACL: Human Language Technologies*, pages 1055–1065, Denver, Colorado, 2015. ACL.
- [370] Rong Wang, Wenlin Liu, and Shuyang Gao. Hashtags and information virality in networked social movement: Examining hashtag co-occurrence patterns. *Online Information Review*, 2016.

- [371] Xuerui Wang and Andrew McCallum. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 424–433, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933395.
- [372] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.*, 53(3), jun 2020. ISSN 0360-0300.
- [373] Yunlong Wang, Priyadarshini Venkatesh, and Brian Y Lim. Interpretable directed diversity: Leveraging model explanations for iterative crowd ideation. In *Proc. of CHI 2022*, CHI '22, New York, NY, USA, 2022. ACM. ISBN 9781450391573.
- [374] James E Ward and Richard E Wendell. Approaches to sensitivity analysis in linear programming. *Annals of Operations Research*, 27(1):3–38, 1990.
- [375] Peter C Wason. On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12(3):129–140, 1960.
- [376] Jen Weedon, William Nuland, and Alex Stamos. Information operations and facebook. Technical report, Facebook, 2017.
- [377] Chih-Ping Wei, Yen-Hsien Lee, Yu-Sheng Chiang, Chun-Ta Chen, and Christopher CC Yang. Exploiting temporal characteristics of features for effectively discovering event episodes from news corpora. *Journal of the Association for Information Science and Technology*, 65(3):621–634, 2014.
- [378] Achim Weigel and Frank Fein. Normalizing the weighted edit distance. In *Proc. of the 12th IAPR Intl. Conf. on Pattern Recognition, Vol. 3-Conf. C: Signal Processing (Cat. No. 94CH3440-5)*, volume 2, pages 399–402, Jerusalem, Israel, 1994. IEEE, IEEE.
- [379] Galen Weld, Maria Glenski, and Tim Althoff. Political bias and factualness in news sharing across more than 100,000 online communities. *arXiv preprint arXiv:2102.08537*, 2021.
- [380] John Wenskovitch and Chris North. An examination of grouping and spatial organization tasks for high-dimensional data exploration. *IEEE Transactions on Visualization and Computer Graphics*, 2020.
- [381] John Wenskovitch and Chris North. Interactive artificial intelligence: Designing for the “two black boxes” problem. *Computer*, 53(8):29–39, 2020.
- [382] John Wenskovitch, Lauren Bradel, Michelle Dowling, Leanna House, and Chris North. The effect of semantic interaction on foraging in text analysis. In *2018 IEEE Conf. on Visual Analytics Science and Technology (VAST)*, pages 13–24, Berlin, Germany, 2018. IEEE, IEEE.

- [383] John Wenskovitch, Michelle Dowling, and Chris North. With respect to what? Simultaneous Interaction with Dimension Reduction and Clustering Projections. In *Proc. of the 25th International Conference on Intelligent User Interfaces*, pages 177–188, 2020.
- [384] Nigel D White. Ending the US embargo of Cuba: International law in dispute. *Journal of Latin American Studies*, 51(1):163–186, 2019.
- [385] Vanessa Williamson. On the ethics of crowdsourced research. *PS: Political Science & Politics*, 49(1):77–81, 2016.
- [386] Tom Wilson, Kaitlyn Zhou, and Kate Starbird. Assembling strategic narratives: Information operations as collaborative work within an online community. *Proc. of the ACM on HCI*, 2(CSCW):1–26, 2018.
- [387] Sarah E. Worth. Storytelling and narrative knowing: An examination of the epistemic benefits of well-told stories. *Journal of Aesthetic Education*, 42(3):42–56, 2008. ISSN 00218510, 15437809.
- [388] William Wright, David Schroh, Pascale Proulx, Alex Skaburskis, and Brian Cort. The sandbox for analysis: Concepts and methods. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, page 801–810, New York, NY, USA, 2006. ACL. ISBN 1595933727.
- [389] Hao Wu, Michael Mampaey, Nikolaj Tatti, Jilles Vreeken, M Shahriar Hossain, and Naren Ramakrishnan. Where Do I Start? Algorithmic Strategies to Guide Intelligence Analysts. In *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics*, pages 1–8, 2012.
- [390] Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter*, 21(2):80–90, 2019.
- [391] Yaguang Wu, Haichun Sun, and Chungang Yan. An event timeline extraction method based on news corpus. In *2017 IEEE 2nd Intl. Conf. on Big Data Analysis (ICBDA)*, pages 697–702, Beijing, China, 2017. IEEE.
- [392] Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1038–1046, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413650. URL <https://doi.org/10.1145/3394171.3413650>.
- [393] Ying Xiong, Moonhee Cho, and Brandon Boatwright. Hashtag activism and message frames among social movement organizations: Semantic network analysis and thematic

- analysis of Twitter during the #MeToo movement. *Public relations review*, 45(1):10–23, 2019.
- [394] Nuo Xu and Xijin Tang. Generating risk maps for evolution analysis of societal risk events. In Jian Chen, Yuji Yamada, Mina Ryokey, and Xijin Tang, editors, *Knowledge and Systems Sciences*, pages 115–128, Singapore, 2018. Springer. ISBN 978-981-13-3149-7.
- [395] Shize Xu, Shanshan Wang, and Yan Zhang. Summarizing complex events: a cross-modal solution of storylines extraction and reconstruction. In *Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1281–1291, Seattle, Washington, USA, October 2013. ACL.
- [396] Sibel Yaman, Dilek Hakkani-Tür, Gokhan Tur, Ralph Grishman, Mary Harper, Kathleen R McKeown, Adam Meyers, and Kartavya Sharma. Classification-based strategies for combining multiple 5-w question answering systems. In *Tenth Annual Conf. of the Intl. Speech Communication Association*, pages 2703–2706. INTERSPEECH-2009, 2009.
- [397] Roman V Yampolskiy. Turing test as a defining feature of AI-completeness. In *Artificial intelligence, evolutionary computing and metaheuristics*, pages 3–17. Springer, 2013.
- [398] Rui Yan, Liang Kong, Congrui Huang, Xiaojun Wan, Xiaoming Li, and Yan Zhang. Timeline generation through evolutionary trans-temporal summarization. In *Proc. of the 2011 Conf. on Empirical Methods in Natural Language Processing*, pages 433–443, Edinburgh, Scotland, UK, 2011. ACL.
- [399] Rui Yan, Xiaojun Wan, Jahna Otterbacher, Liang Kong, Xiaoming Li, and Yan Zhang. Evolutionary timeline summarization: A balanced optimization framework via iterative substitution. In *Proc. of the 34th Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, page 745–754, NY, USA, 2011. ACM. ISBN 9781450307574.
- [400] Christopher C. Yang, Xiaodong Shi, and Chih-Ping Wei. Tracing the event evolution of terror attacks from on-line news. In Sharad Mehrotra, Daniel D. Zeng, Hsinchun Chen, Bhavani Thuraisingham, and Fei-Yue Wang, editors, *Intelligence and Security Informatics*, pages 343–354, Berlin, Heidelberg, 2006. Springer. ISBN 978-3-540-34479-7.
- [401] Christopher C Yang, Xiaodong Shi, and Chih-Ping Wei. Discovering event evolution graphs from news corpora. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 39(4):850–863, 2009.
- [402] Jie Yang, Soyeon Caren Han, and Josiah Poon. A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, 64(5): 1161–1186, 2022.

- [403] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernández Ábrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*, 2019.
- [404] Yuxuan Ye and Edwin Simpson. A proposal: Interactively learning to summarise timelines by reinforcement learning. In *Proceedings of the First Workshop on Interactive Learning for Natural Language Processing*, pages 25–31, 2021.
- [405] Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. A simple and efficient sampling method for estimating ap and ndcg. In *Proc. of the 31st Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR '08, page 603–610, NY, USA, 2008. ACM. ISBN 9781605581644.
- [406] Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proc. of the 2003 Conf. on Empirical methods in natural language processing*, pages 129–136, Sapporo, Japan, 2003. ACL.
- [407] Yi Yu, Adam Jatowt, Antoine Doucet, Kazunari Sugiyama, and Masatoshi Yoshikawa. Multi-TimeLine Summarization (MTLS): Improving Timeline Summarization by Generating Multiple Summaries. In *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Intl. Joint Conf. on Natural Language Processing (Volume 1: Long Papers)*, pages 377–387, Online, 2021. ACL.
- [408] Ruifeng Yuan, Qifeng Zhou, and Wubai Zhou. dTexSL: A dynamic disaster textual storyline generating framework. *World Wide Web*, 22(5):1913–1933, 2019.
- [409] Amy X. Zhang, Lea Verou, and David Karger. Wikum: Bridging discussion forums and wikis using recursive summarization. In *Proc. of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 2082–2096, New York, NY, USA, 2017. ACM. ISBN 9781450343350.
- [410] Hongxin Zhang and Haitao Liu. Visualizing structural “inverted pyramids” in english news discourse across levels. *Text & Talk*, 36(1):89–110, 2016.
- [411] Liang Zhao. Event Prediction in the Big Data Era: A Systematic Survey. *ACM Computing Surveys (CSUR)*, 54(5):1–37, 2021.
- [412] Houkui Zhou, Huimin Yu, Roland Hu, and Junguo Hu. A survey on trends of cross-media topic evolution map. *Knowledge-Based Systems*, 124:164–175, 2017.
- [413] P. Zhou, B. Wu, and Z. Cao. EMMBTT: A Novel Event Evolution Model Based on TFXIEF and TDC in Tracking News Streams. In *2017 IEEE Second Int. Conf. on Data Science in Cyberspace (DSC)*, pages 102–107, Shenzhen, China, 2017. IEEE.

- [414] Wubai Zhou, Chao Shen, Tao Li, Shu-Ching Chen, and Ning Xie. Generating textual storyline to improve situation awareness in disaster management. In *Proc. of the 2014 IEEE 15th Intl. Conf. on Information Reuse and Integration*, pages 585–592, Redwood City, CA, USA, 2014. IEEE, IEEE.
- [415] Wubai Zhou, Chao Shen, Tao Li, Shu-Ching Chen, Ning Xie, and SS Iyengar. A new two-layer storyline generation framework for disaster management. *Intl. Journal of Next-Generation Computing*, 9(3), 2018.
- [416] Xianshu Zhu and Tim Oates. Finding story chains in newswire articles. In *2012 IEEE 13th Intl. Conf. on Information Reuse & Integration (IRI)*, pages 93–100, Las Vegas, NV, USA, 2012. IEEE.
- [417] Xianshu Zhu and Tim Oates. Finding story chains in newswire articles using random walks. *Information Systems Frontiers*, 16(5):753–769, 2014.
- [418] Constantin Zopounidis and Michael Doumpos. Multi-group discrimination using multi-criteria analysis: Illustrations from the field of finance. *European Journal of Operational Research*, 139(2):371–389, 2002.

Appendices

Appendix A

Full Insight Tables

We present a series of tables with all the insights provided by the participant in our user study from Chapter 9. Table A.1 has all the insights identified by participants related to the causes of the 2021 Cuban protests. Table A.2 has all the insights related to the effects of the Cuban protests in Cuba itself. Table A.3 has all the insights related to the effects of the Cuban protests in the United States. Finally, Table A.3 has all the insights related to the international effects (non-U.S.) of the Cuban protests and general contextual insights. In all these tables, a marked (×) cell represents that a specific participant (columns) obtained a specific insight (rows). Colored rows represent high-level insights, which are also counted towards the detailed insight total.

Insights		S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B11	B12	B13	B14	I1	I2	I3	I4	I6	I7	I8	I9	I10	I11
Causes: Economic Crisis Worst economic crisis since fall of Soviet Union / 30 years Basic necessities and food shortages Medicine and hygiene products shortages People turning to black market / alternative sources High / rising prices, job shortages, inability to support businesses Power outages and unstable utilities Internet outages Lack of flights available out of the country	Causes: Economic Crisis	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Worst economic crisis since fall of Soviet Union / 30 years	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Basic necessities and food shortages	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Medicine and hygiene products shortages	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	People turning to black market / alternative sources	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	High / rising prices, job shortages, inability to support businesses	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Power outages and unstable utilities	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Internet outages	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Lack of flights available out of the country	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Causes: Underlying Economic Issues	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Causes: Underlying Economic Issues Economic crisis and shortages worsened by the COVID-19 pandemic Tourism industry impacted by COVID-19 Trade / exports impacted by COVID-19 Economic crisis worsened by Cuban policies / mishandling of spending Restrictions on food and medicine imports / customs fees General inefficiencies due to communist economy / Soviet-style system	Causes: Underlying Economic Issues	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Economic crisis and shortages worsened by the COVID-19 pandemic	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Tourism industry impacted by COVID-19	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Trade / exports impacted by COVID-19	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Economic crisis worsened by Cuban policies / mishandling of spending	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Restrictions on food and medicine imports / customs fees	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	General inefficiencies due to communist economy / Soviet-style system	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Causes: Healthcare Crisis	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Ongoing pandemic / general effects of COVID-19 in the healthcare system	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Government mishandling of the COVID-19 pandemic	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Causes: Ongoing Violence Government violence and repression Public riots / gang beatings Human rights abuses Religious persecution	Cuba's universal / government-controlled healthcare system issues	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Recent surge in COVID-19 cases	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Lack of access to vaccines / vaccination pace	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Food scarcity is compounding healthcare issues	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Cuban officials blaming / suppressing overworked doctors	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Causes: Ongoing Violence	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Government violence and repression	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Public riots / gang beatings	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x								

Table A.1: Table of insights for causes.

Insights		S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B11	B12	B13	B14	I1	I2	I3	I4	I6	I7	I8	I9	I10	I11
Effects: Government Deflection		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Government denial and deflecting / rejection of coverage		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Cuban government blames unrest on criminals / rioters		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Cuban government blames social media / online misinformation		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Cuban government blames the US		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Cuban government blames the US sanctions / embargo / economy		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Cuban government blames US-backed protesters / mercenaries		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Cuban government blames the Cuban-American mafia		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Effects: Government Repression		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Crackdown on protests, leading to arrests / disappearances of protesters		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Arrests / brutality on journalists		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Use of excessive force / violence / human rights violations by the government		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Sham trials or no proper judicial procedure / overwhelmed justice system		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Flooding cities with soldiers / heavy police presence		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Use of special forces and spies by the government (black berets)		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Silencing people / fear of speaking up		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Cuban rallying its own backers to counter-protest / order of combat		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Churches calling out government repression		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Increased protests due to repression		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Castro calling to defend revolution after backlash		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Effects: Government Censorship		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Internet access restrictions		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Social media restrictions		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Messaging apps restrictions / jammed communications		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Censorship of "misinformation" by government (social media laws)		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Censorship on TV signals		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Censorship as an example to other authoritarian regimes		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Effects: Government Concessions		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Acknowledgment of issues / president takes some blame		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Internal dissent from party members		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Relaxing of restrictions / customs on food and medicine for travelers		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Additional food rations with no fundamental change		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Improve access to utilities (e.g., electricity / internet)		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Provided medicine / vaccines to the people		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Allowing state-owned enterprises to set wages		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Concessions after protests can lead to further protests later		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Effects: COVID-19 Surge / Dead Generals		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
COVID-19 surge after protests		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Dead generals / soldiers (potentially due to COVID-19 surge)		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
WHO discourages protests over COVID-19		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Table A.2: Table of insights for effects in Cuba.

[illegible]

Table A.3: Table of insights for effects in the U.S.

Insights	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B11	B12	B13	B14	I1	I2	I3	I4	I6	I7	I8	I9	I10	I11
Effects: International Condemnation	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Other countries condemned the Cuban government / support protesters	x																																
EU, UK, and Canada condemned the Cuban government																																	
International condemnation of human rights violations																																	
U.N. general assembly discussion on Cuba / calls for releasing prisoners																																	
Change.org petition demonstrating global backlash																																	
Effects: International Support and Aid	x																																
International community critical of / does not support U.S. embargo																																	
Cuban allies criticize US embargo																																	
Humanitarian aid by the international community																																	
Support and aid from other socialist/communist/allied countries																																	
Effect of COVID-19 on international aid (trade issues)																																	
Context: Protest Characteristics	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Biggest protests since '59																																	
Geographic locations of protests / events (San Antonio, West Cuba, Havana)																																	
Peaceful protests																																	
Mix of peaceful and violent protests / riots																																	
Criminals take advantage of protests / mass looting and vandalism																																	
Protest slogans / "Patria y Vida" / Hashtags																																	
Organizations and groups involved in the protests (Ladies in White)																																	
Context: Misinformation and Fake News																																	
Use of propaganda / counter-narratives by the government																																	
Fake news are proliferating / misinformation surrounding the protests																																	
Fake news examples (e.g., Venezuela, Raul Castro, etc)																																	
Context: Bias in News																																	
Political bias in reporting																																	
Recognition of highly biased / sensationalistic articles and source quality (e.g., Breitbart)																																	
U.S. reaction / U.S. politics centric news sources																																	
Context: Comparison with Other Events (e.g. BLM, historical)																																	
Comparing Cuban response with U.S. response to BLM protests																																	
Comparing Cuban protests with Hong Kong protests																																	
Comparing the Cuban protests with the Sri Lanka protests																																	
Comparing the build-up towards the protests to 9/11 (multifactorial, long time)																																	
Comparing Cuban internet firewall with Berlin Wall																																	
Comparing the U.S. providing internet to the use of Starlink in the Ukraine War																																	
Comparing Cuba's censorship and political system with North Korea or China																																	
Comparison and speculation about protests in Venezuela																																	
Comparison with other communist countries and their "market economies"																																	

Table A.4: Table of insights for international effects and context information.