Spatio-Temporal Storytelling on Twitter

RAY DOS SANTOS and SUMIT SHAH, Virginia Tech
FENG CHEN, Carnegie Mellon University
ARNOLD BOEDIHARDJO, U.S. Army Corps of Engineers
PATRICK BUTLER, Virginia Tech
CHANG-TIEN LU, Virginia Tech
NAREN RAMAKRISHNAN, Virginia Tech

Social media such as *Twitter* have ushered in alternative modalities to propagate news and developments rapidly. Just as traditional IR and web research matured to modeling storylines from search results and linking documents into stories, we are now at a point to study how stories organize and evolve in mediums such as *Twitter*. While there has been a plethora of research for mining and connecting entities through text, the task of deriving a story from short, ill-formed text, is a difficult one. Using spatio-temporal analysis on induced concept graphs, we introduce a new set of methods to automatically derive stories over linked entities in tweets. Our approach models a story as a graph of entities propagating through spatial regions in a temporal sequence, and controls search space complexity by suggesting regions of exploration. Experiments on large scale *Twitter* datasets demonstrate how storylines can be flexibly generated under various syntactic constraints, even without specific endpoints.

1. INTRODUCTION

Social media, e.g., *Twitter*, have provided us an unprecedented opportunity to observe events unfolding in real-time. The rapid pace at which situations play out on social media necessitates new tools for capturing and summarizing the spatio-temporal progression of events.

Take for instance the *Boston Marathon* bombings of April 15, 2013. In the immediate days afterward, law enforcement officers collected a significant number of eyewitness accounts, photo and video footage, and background information on several suspects who were spatially and temporally tagged. What followed was a succession of outcomes: several people were detained near the blast spots; the residence of a Saudi national was searched; MIT police officer S. Collier was killed; the Tsarnaev brothers were identified as two suspects. All these developments could be observed on *Twitter*, but to the best of our knowledge there exists no automated tool that can provide comprehensive and picturesque summaries from tweets. Fast forward to Table V for an illustration of the types of summaries we envisage creating in our project.

The underlying problem is one of *storytelling*, the process of connecting entities through their characteristics, actions, and events [Turner 1994]. *Information retrieval* and web research have studied this problem, i.e., modeling storylines from search results, and linking documents into stories [Kumar et al. 2008][Hossain et al. 2011][Hossain et al. 2012b] (the terms *stories* and *storylines* are used interchangeably). Traditional *storytelling* attempts to link disparate entities that are known ahead of time, such as the connections between two individuals. In this study, however, our focus is not traditional text analysis. Rather, we explore spatio-temporal entity analysis, which can fill some of the gaps left by traditional approaches. Our goal is to not only find meaningful connections, but also to derive new stories for which we do not know the endpoint, if one exists. For example, we

Author's addresses: R. Dos Santos, Sumit Shah, Patrick Butler, Chang-Tien Lu, Naren Ramakrishnan, Computer Science Department, Virginia Tech; Feng Chen, Computer Science Department, Carnegie Mellon University; Arnold Boedihardjo, U.S. Army Corps of Engineers.

DOI:http://dx.doi.org/10.1145/0000000.0000000

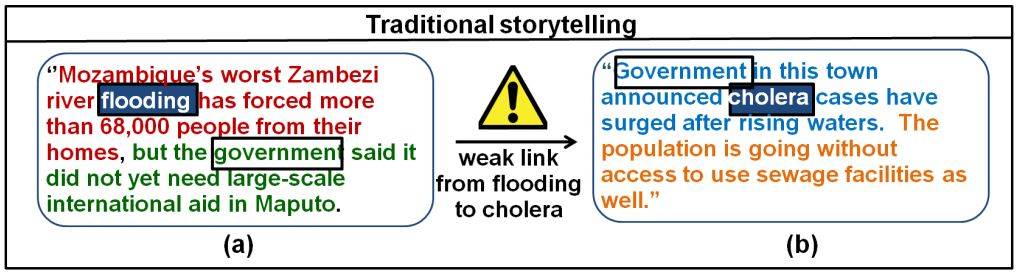

Fig. 1. Under *traditional storytelling*, (a) and (b) represent two partial NY Times articles [2007]. The two documents are weakly connected because no patterns other than two "government" entities relate the two documents, making the flooding-cholera link of difficult identification.

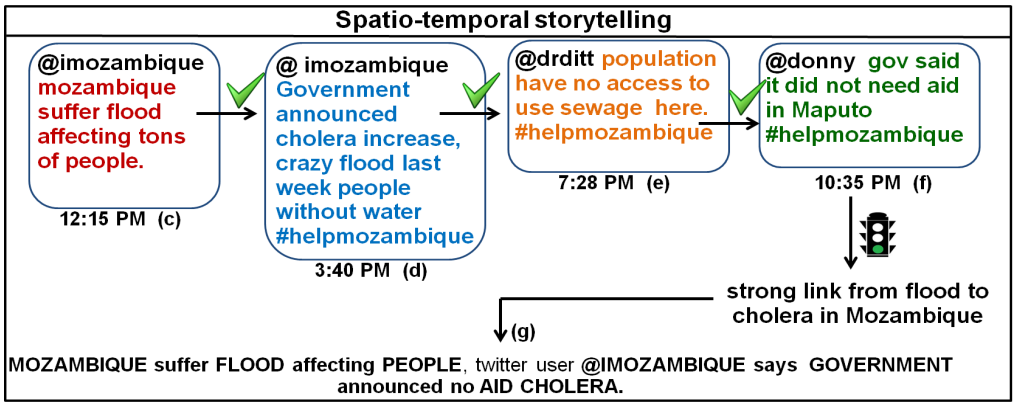

Fig. 2. Under *spatio-temporal storytelling*, (c)(d)(e) and (f) show four tweets similar to the NY Times articles of Fig 1. The four tweets are strongly connected through spatio-temporal features, such as the timestamps and the locations (Mozambique, Maputo), as well as through *Twitter* features (users, hashtags, concepts, and relationships). The Twitter storyline (g) demonstrates how a progression of disparate tweets can be used to make the flood-cholera link in Mozambique.

would be interested in examining the passing of a new law and the reactions it provokes, such as protests in nearby areas. This falls in the field of exploratory analysis where the main focus is discovering new patterns of knowledge. We target spatio-temporal techniques on short, ill-formed text of *Twitter* data for which deriving stories has proven to be a difficult task.

Traditional storytelling has been mostly successful on news articles, blogs, as well as structured databases. In general, it makes one strong assumption: the availability of comprehensive data sources, where textual content is robust and ideas are well presented. In this manner, it is able to perform document analysis using several techniques, some of which include vector-space measures such as *cosine similarity*, natural language processing (NLP) for *parts-of-speech tagging*, and keyword matching, among others. A common problem with such methods is that inferences may be missed whenever linkage among documents cannot be strongly asserted. Consider the example of Fig. 1. In **(a)**, a partial *NY Times* news articles describes a flood situation which affected thousands of people in Mozambique in 2007, whereas **(b)** tells about an outbreak of cholera. If the goal is to establish correlation between flooding in document (a) and cholera in document (b), we would first have to link the two documents. Deriving this link is difficult for the following reason: except for government - government, no other terms or patterns are shared between the two documents. A simple *cosine similarity* calculation would yield a low score, and the flooding - cholera link would most likely be missed due to weak connectivity between the two sources.

The above example illustrates why techniques that apply to traditional storytelling tend to perform poorly on social media content, such as *Twitter*, where text lacks proper form and function. For this reason, social media storytelling demands new techniques that can benefit not only from its textual content, however limited, but also from embedded tweet features. These features come in two flavors: (1) spatio-temporal knowledge of the entities described in text; (2) and intrinsic characteristics of social media represented in the form of metadata. An example is given by Fig. 2 **(c)(d)(e)** and **(f)**, which shows four hypothetical tweets modeled after the *NY Times* documents of **(a)** and **(b)**, but written in a more "*Twitter*-like style" (somewhat abbreviated, broken language, little punctuation). Just as in the *NY Times* example, performing *cosine similarity* on any pair of the four tweets would also yield meaningless results, given that very few terms are shared. At closer investigation, however, *Twitter* data allow us to link all four documents through different means. First, tweets **(c)** and **(d)** can be linked because they were issued by the same *Twitter* user (**@imozambique**) and because they share other terms ("flood","people"). Second, tweets **(d)** and **(e)** are connected through the *hashtag* **#helpmozambique**, a strong indication that they address the same general topics. The same link can be made from **(e)** and **(f)**. To close the loop, tweets **(f)** and **(c)** are connected by location: geocoding Maputo and Mozambique allows us to determine that the *latitude*/*longitude* of the former is enclosed in the latter, and thus geospatially related.

As simple as this example may be, it shows that tweets can be linked in many ways, such as by users, locations, and hashtags. In this paper, we strongly emphasize the **spatio-temporal** aspect of the data, considering only tweets that have locations and timestamps. Other features, which we explain later, are also available. Five aspects of this approach are noticeable. First, it allows us to create a short storyline to represent the four tweets. That storyline is shown in Fig. 1(g). It is composed of a sequence of entities identified in the tweets, such as Mozambique and cholera, and relationships, such as $\overset{suffer}{\rightarrow}$ and $\overset{announced}{\rightarrow}$, also from the tweets, which serve to make connections between the entities. In later sections, we explain how to create storylines and discuss other mechanical aspects, such as why some entities are included while others are ignored, and how to use the relationhsips. The attractive aspect of this storyline is that, unlike the *traditional storytelling* example above, it is able to discover a connection between the desired flood - cholera entities. Second, storylines can be made as elastic as necessary by injecting new tweets in an incremental approach. Third, when represented as a graph, a theoretically-unlimited number of tweets can be collapsed into fewer entities and their corresponding relationships. For example, Mozambique or cholera may appear thousands of times in the raw dataset, but in our approach, they are only represented once each, minimizing resource usage. In this manner, the number of generated storylines tends to be several orders of magnitude smaller than the number of tweets that generate them; fourth, they enforce time sequencing, which promotes storyline coherence by preserving the order of facts. In Fig. 1(c), the storyline begins at 12:15 PM when the "flood is announced", and ends with Fig. 1(e) at 10:35 PM when the "government says that aid is not needed". It would not make sense for the government to talk about aid before the flood occurred! Fifth, graph structures are more machine-friendly than file-system documents, allowing efficient searches, spatial operations, and automated data mining.

The importance of location and time: Applying traditional network analysis tools to find and link entities across tweets can lead to 'runaway' stories. Three important problems have to be surmounted. First, to ensure meaningfulness, we must use spatio-temporal coherence as both a desirable aspect of stories and as a way to control computational complexity. It is desirable because entities might be related to one another only under certain circumstances, and modeling spatio-temporal coherence ensures explainable stories. It is a way to control computational complexity because it avoids searching for stories that might not be central to the topic under consideration. For instance, tweets that refer to cases of *cholera* in *South America* are most likely not related to *cholera* cases in *Mozambique*. Thus **spatial** is a fundamental consideration. Second,

time and space must support an approach that can instill some notion of typing to connections inferred from *Twitter*. For instance, a flood $\xrightarrow{increases}$ cholera link can potentially infer causality if these entities are both in **proximal areas** and **close in time**. Otherwise, stating that *flood* causes *cholera* in different places and times is mere speculation. Again such a notion of typing aids in both explainability and scalability objectives. Third, we require algorithms that can operate without specific provision of start and end points as long as entities can be coherently identified **in a location** and **within a timeframe**. The ability to support dynamic storylines as they evolve is critical to modeling fast moving social media streams such as *Twitter*. The goal of this paper is to address the above issues and enhance the current state of storytelling. Our key contributions are:

(1) **Modeling short text over space and time**: We describe arguably the first algorithm to conduct storytelling without specific endpoints (i.e., without supervision) over short text (tweets), represented as an entity graph, and our strategies to enforce coherence, precision, and the influence of spatial entity types on the generated storylines.
(2) **Reasoning over spatio-temporal features**: Key to obtaining coherent stories is to identify regions of spatial propagation where related entities cluster. We demonstrate the use of *Ripley's K* function for this purpose and its use in conjunction with temporal propagation where time windows help keep stories succinct and coherent. In combination, they limit the search space from possibly millions down to the thousands of entities.
(3) **Devising spatio-temporal storylines based on connectivity strength**: We provide a parameter-free relevance measure based on *ConceptRank*, which differentiates relationship types, boosts strongly-connected spatial entities, and helps eliminate large numbers of poorly-connected ones. In addition, storylines are found "on the fly", demonstrating our ability to generate lines of exploration that span across space and time.
(4) **Performing extensive experiments on social media**: To show the effectiveness of spatio-temporal storytelling on *Twitter* data, we evaluate our approach based on current world events that encompass sports corruption in Brazil, social unrest in Mexico, the *Boston Marathon Bombings* developments, and the evolution of the *Syrian Civil War*. Included, we discuss the impact of spatio-temporal parameters, the importance of *Twitter* features, and show the utilization of storytelling in a forecasting application.

Throughout this study, we introduce various components needed for storytelling. Section 2 elaborates on existing work, highlighting differences. In Section 3, we provide background information and definitions. Section 4 explains the spatio-temporal mechanics of entity discovery, ranking, and storyline generation. Section 5 discusses parameter tuning along with challenges that affect this research. Experiments are presented in Section 6 and a conclusion is given in Section 7.

2. RELATED WORK

The work proposed in this paper spans many areas of expertise, from ontological analysis to geographic networks. Our research best lines up with the approaches described below.

2.1. Storytelling

The phrase 'storytelling' has been introduced in an algorithmic context by Kumar et al. [2008] who proposed it as a generalization of *redescription mining*. At a high level, *redescription mining* takes as input a set of objects and a collection of subsets defined over those objects with the goal of identifying objects described in two or more different ways. Such objects are interesting because they may signal shared characteristics and similar behavior, which can be a powerful tool in the context of *storytelling*. One such algorithm is *CARTWheels* [Ramakrishnan et al. 2004] which utilizes induced classification trees to model redescriptions along with the *A* Algorithm* for least-cost path traversal. Hossain et al. [2012b] develop this idea to connect two unrelated PubMed documents where connectivity is defined based on a graph structure, using the notions of hammocks (similarity) and cliques (neighborhoods). This work was generalized to entity networks in Hossain et al. [2012a]

Table I. Comparison of storytelling-related approaches

	Data Source	Link Type	Connecting Elements	Modeling	Edge Strength
traditional storytelling	text / reports	entity relationships	entities	lattice and clique chains	Soergel distance and TF-IDF
connecting the dots	news articles	words / phrases	documents	general graphs	TF-IDF
this work ★	Twitter	entity relationships	entities	spatio-temporal concept graphs	ConceptRank

and specifically targeted for use in intelligence analysis. Their motivation is that current technology lacks better support for entity linkage, explanation of relationships, exploration of user-specified entities, and automated reasoning in general. The tools used in this work include concept lattices as a network where candidate entities are identified with three nearest neighbor approaches (Cover Tree, *k*-Clique, and NN Approximation). The *Soergel Distance* measures the strength between entities, while *coreferencing* serves to identify entities mentioned in various parts of the text using differing terms. All these works require specific start and endpoints, and link entities according to a desired neighborhood size and distance threshold. In many of these works, edge weight has been based on a variation of term frequency × inverse-document-frequency (*TF-IDF*). This class of works represent *traditional storytelling* approaches even though neighborhood distances are considered, albeit not from a geospatial perspective.

2.2. Connecting the Dots

The primary focus of these works is on document linkage rather than entity connectivity. For this reason, textual reasoning is a strong facet of the targeted methods, which departs from a spatio-temporal view of events. Endpoints must (again) be specified and link strength utilizes the notion of *coherence* across documents, which is proposed by Shahaf and Guestrin [2010]. In this work, stories are modeled as chains of articles, where the appearance of shared words across documents help establish their relatedness. Another important aspect is the determination of *influence* between documents based on the presence of a given word. For this purpose, a bipartite graph is built using documents and words as nodes, where edge strength among them can be obtained by third-party tools or with *TF-IDF* scores. Extending that work, they also propose related methods to generate document summaries, i.e. *Metro Maps*, in [Shahaf et al. 2012b] and [Shahaf et al. 2012a], which target scientific literature. Some of the goals are to measure the importance of a paper in relation to the corpus, find the probability that two papers originate from the same source, and identify research lines. The basic data structure is also a directed graph, where for each map that has been generated, its *coverage* is calculated using each document as a vector of word features. The *coverage* is then defined for a set of words as *TF-IDF* values, which can be extended to sets of documents. Connectivity between maps is measured by the number of paths that intersect two maps. Overall, *connecting the dots* methods rely heavily on the abundance of robust content such that the aforementioned calculations (coherence, influence, coverage, etc.) can be calculated acceptably. Social media, however, breaks the assumption of robust content, limiting the amount of textual reasoning that can be performed. Thus, *connecting the dots* is less than ideal for environments that utilize *Twitter* data feeds.

2.3. Link Analysis

This class of work often relies on graphs as a modeling abstraction, such as the evolution of entities in space and time ([Mondo et al. 2013], [Chan et al. 2009]) and the identification of patterns ([George et al. 2009], [Chan et al. 2008]). The spatio-temporal aspects observe how entities propagate. The goal of link analysis, however, is not to explore *stories* per se. Rather, it is an attempt to quantify changes in entities and manage relationships, which leads us to the notion of ranking.

Ranking in terms of link analysis has been popularly applied to web pages since the seminal works of Brin and Page [1998] and Kleinberg [1998]. The former computes the value of the importance

of a web page based on its links and an initial damping factor. The latter also consider the page's links, but is dependent on an initial query that generates a *root set*, and is augmented by other pages that point to the *root set*. In our approach, we propose a variation of *PageRank* for spatio-temporal entities, which we denote as *ConceptRank*, and which introspects relationship types, differentiating them to influence lower or higher rankings.

Within the same family of the above approaches, there have been other proposed methods. The *Indegree Algorithm* is a simple heuristic that considers the *popularity* factor as a ranking measure [Marchiori 1997]. For social media, *popularity* is a gray area: works well for high-visibility events, but may fail miserably for events that are important, but that do not get much exposure. For *storytelling*, this type of applicability is possible, but too subjective in terms of ranking. The *HITS Algorithm* of Kleinberg [1998] introduced the notion of *hub and authority*, where authorities are the pages that hold "legitimate" information, and hubs are the pages directing the user to the authorities. For example, users seeking to buy a television would favor pages such as *sony.com* or *panasonic.com*, since they are authorities in the subject. They may, however, give different preference to a directory page such as *Yahoo! Shopping* (i.e., the hub) that points to the authorities. In terms of *storytelling*, this type of ranking would be challenging since there is no clear-cut way to determine which entities would be authorities and which would be hubs. It represents an open line of research, but outside the scope of this document.

2.4. Visualization Tools

Even though the goal of this study is not visualization (we do include some in the experiments), we point out some third-party tools available for this purpose. Most of these tools perform information extraction into a graph, linking them according to different criteria for subsequent use. In general, some of them leave reasoning to the end user, while others provide more robust knowledge exploration. *NetLens* [Kang et al. 2007], for example, operate on user queries, permitting them to be refined interactively. For more advanced graph analysis, Sentinel Visualizer [2013] and Palantir [2013] provide capabilities such as geospatial mapping, shortest-path calculus, and a variety of filtering options. Collaboration among users and across data repositories are powerful aspects of these solutions. Entity extraction in the form of persons, locations, and organizations is also accomplished by [Jigsaw 2013], with views based on lists, graphs, and connections among entities. *Siren* takes a different direction, using redescription mining as a means to identify and visualize geospatial entities that may have equivalent descriptions in disparate sources [Galbrun and Miettinen 2012]. Our *storytelling* approch differs from the above tools in the following manners: (1) our foundation resides on a spatio-temporal reasoning process, establishing firm decisions on what a story should consist of, and justifying the why and why nots of connections. Although important, the visualization aspect is a secondary priority that can be achieved using some of the above tools; (2) most of these tools work on a supervised manner, i.e., an analyst must be available to introduce modifications or provide feedback. Our approach is a standalone framework that requires at most a few initialization parameters; (3) none of these tools can be considered an end-to-end *storytelling* solution. Rather, they are extremely valuable in integrating sources, presenting data, and generating different views of analysis. Little guidance, however, is provided in terms of reasoning and inference.

Each of the above approaches have different goals, apply different techniques, and target vastly different datasets. As a result, hard comparisons to our proposal are inadequate, if not misleading. Table I constrasts our approach to the first two classes of approaches. While traditional *storytelling* can handle text and reports in general, the *connecting the dots* approach uses standard news articles. Both link pre-defined endpoints, the former linking entities, and the latter documents. Our approach operates on ill-formed *Twitter* data, builds stories dynamically without specific endpoints, and makes heavy use of entity relationships as opposed to word frequencies. Since *Twitter* data have poor textual content, our approach refrains from *TF-IDF* measures, calculating instead a *ConceptRank* to link entities, which, in practice, works as a form of edge strength. Because we rely on a

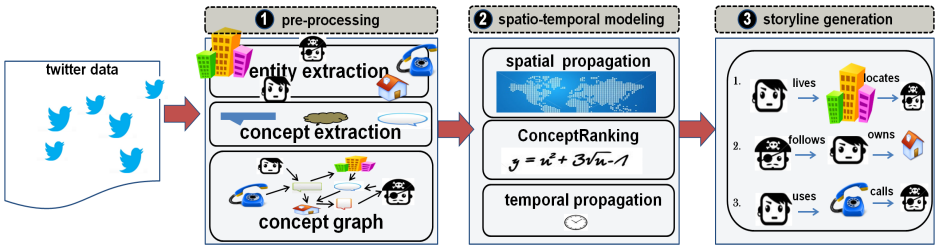

Fig. 3. Three-step process for spatio-temporal storyline generation using *Twitter* data. In the pre-processing stage, entities and concepts (relationships) are extracted and used to build the concept graph. Under spatio-temporal modeling, spatial propagation first discovers entities in nearby locations. For each entity, *ConceptRanking* determines its relevance in the graph, and the entities are subsequently time-ordered for proper temporal propagation. Storylines are then generated by linking the *top-k* ranked entities in time order.

spatio-temporal model, both geographical proximity and time ordering are favored, as we coalesce many data points into *bounding rectangles* (BR) and time windows. Given the many differences, we do not claim to have a competing approach. Rather, we present complementary techniques that cover a spatio-temporal niche, which to the best of our knowledge, has not been explored before.

3. PRELIMINARIES

In this section, we provide a visual representation of our methods and introduce the definitions and nomenclature used throughout the remainder of this paper. Fig. 3 shows the three steps taken by our approach: (1) in the pre-processing stage, entities such as people and events, as well as concepts (i.e., relationships), are extracted from *Twitter* data. Combining the extracted entities and their relationships allows a concept graph to be constructed; (2) in the spatio-temporal modeling stage, we find entities located in regions through which a storyline is most likely to propagate, using the concept graph to further rank those entities, and temporally order them; (3) Storylines are then generated using the highest-ranked entities and their observed relationships.

3.1. Definitions

In the scope of our study, an entity network is a graph $G(E,R)$ where entities $E=\{e_1, ..., e_n\}$ can be linked to one another through relationships $R=\{r_1, ..., r_n\}$ defined by conceptual interactions, and thus called a *concept graph*. Given a set of documents $D=\{d_1, d_2, ..., d_n\}$, the following definitions apply:

Definition 3.1. An entity e represents a person, location, organization, event, or object described in at least one document $d_i \in D$. Only entities for which a location and a timestamp can be obtained are considered in this study.

Definition 3.2. A semantic constraint is a user-defined data delimiter similar to a query parameter. For example, if one seeks stories related to "explosion" and other related terms (e.g., "bombing" or "blast"), he/she may use these terms as semantic constraints to guide the storytelling process toward those concepts.

Definition 3.3. A relationship, connection, or link defines a unit of interaction between two entities and is denoted by $e_i \xrightarrow{interaction} e_j$. It is deemed *explicit* if it is extracted from tweet text, such as in $D.Tsarnaev \overset{talks-to}{\rightarrow} T.Tsarnaev$. A relationship is *implicit* if it comes from metadata, as in the *Twitter* case of "follows". Note that all relationships $e_i \xrightarrow{interaction} e_j$ are intended to be directional.

Definition 3.4. An entrypoint is any entity e in the dataset and the point from where the story evolves. It is application-dependent. For instance, in the *Boston Marathon Bombings* scenario, the

Fig. 4. Concept graph example. The solid lines between entities represent *explicit* relationships extracted from tweet textual content. The dashed lines denote *implicit* relationships from tweet metadata.

entrypoint can be the blast site (i.e., a location), an individual seen in the vicinity (i.e., a person), or any other entity of interest.

Definition 3.5. A storyline is a time-ordered sequence of n entities $\{e_1, ..., e_n\}$ where consecutive pairs (e_i, e_j) are linked by one relationship. The number of entities n is the length of the storyline.

3.2. Twitter Features

In order to capture the importance of entities, we use both tweet metadata and textual content in the following manner:

(1) **users** are person entities and the subject and objects of **mentions**, **reply-to**, **following**, and **follower** relationships. They help establish implicit relationships, as defined above in 3.3.
(2) **countries**, **states/provinces**, **cities**, and **addresses** are geocoded and become location entities, both coming from metadata and text. Tweets without location are not considered.
(3) **hashtags** implicitly link entities either in the same or across tweets.
(4) **created At** (from tweet metadata) and **dates** (when available from tweet text) are both used for temporal analysis. Whenever an entity is extracted from text, a timestamp is associated to it. If the tweet text has an inline timestamp that can be associated to the entity, this timestamp will be used. Otherwise, the timestamp of the tweet metadata is used instead. Dates extracted from text are always given preference, if available.
(5) **organizations** are extracted from text (i.e., not metadata).

Fig. 4 shows a simple concept graph where the entities were extracted from several tweets. Solid lines represent explicit relationships, while dashed lines denote implicit ones. We have the following: D. Tsarnaev (D.T.) and T. Tsarnaev (T.T.) are connected through a "talk" relationship, which was extracted from *Twitter* text (not *Twitter* metadata), and is thus defined as explicit. The same is true for the "meets" link between T.T and S. Collier (S.C.), the "works" link from S.C. to MIT, and the "drives" link from D.T. to MIT. The various links to other unknown entities (small triangles) come from *Twitter* metadata ("follows", "following"), and therefore are implicit.

The reason to differentiate the above relationships comes from a simple notion: in entity networks such as *Twitter*, semantic closeness in the form of social interactions is probabilistically correlated to spatio-temporal proximity [Groh et al. 2012] akin to Tobler's first law of Geography, in which similar things tend to be near one another. Intuitively, this notion has several implications to storytelling:

— Explicit connections are more helpful than implicit ones. Knowing that "D.Tsarnaev spoke to T. Tsarnaev" is more powerful than simply learning that "D.Tsarnaev mentions (in the *Twitter* sense) T.Tsarnaev". We explore this idea in Subsection 4.3 in our Concept Ranking.
— A story can be modeled as a graph of entities and semantic relationships propagating through spatially-close regions in a temporal sequence. Consider Fig. 5(a) which depicts several loca-

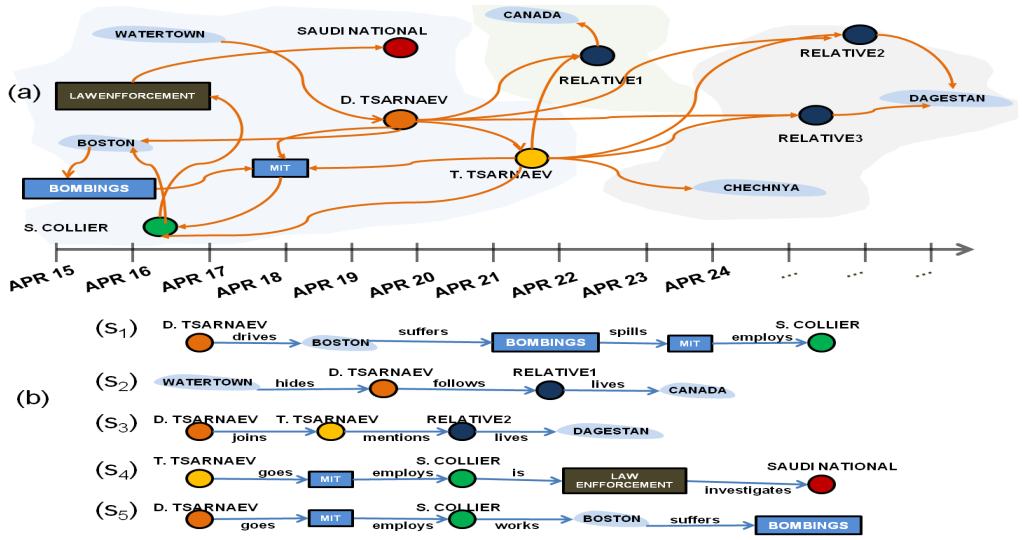

Fig. 5. Boston Marathon Bombings spatio-temporal sequence. In (a), each shape represents an entity observed in a tweet. The edges denote relationships between the entities. In (b), S_1 through S_5 represent five storylines connecting different entities. The english verbs define their relationships and correspond to the edges of the concept graph in (a).

tions related to the *Boston Marathon Bombings*. Most of its developments took place in an 8-day interval (Apr 15-22, 2013) and in proximal areas: Boston - MIT Campus - Watertown. Developments in Canada or Chechnya are an evolving part of the story, but do not necessarily play a major role. Based on these ideas, we use spatio-temporal propagation in Section 4 to define and explore constrained regions of entity connectivity where stories can evolve from.

— Stories do not necessarily have endpoints. Entities come and go, relationships develop, and locations vary. In the *Boston Marathon Bombings*, the entry point could be any one of thousands of persons. The end could propagate through Canada, Russia, and other places. We use this idea to further justify the use of evolving stories in our experiments.

4. SPATIO-TEMPORAL MODELING AND STORYTELLING

In this section, we explore the ideas discussed previously, applying them in mathematical form, which closely reflect the steps of Fig. 3. In brief, our methods take as **input** a set of *Twitter* feeds, and generate as the **output** a set of storylines. The intermediate steps are explained below at a deeper level of detail.

4.1. Spatial Modeling

In the process of telling a story, the entrypoint can be any entity such as a person or event, as in the "bombing" scenario. Given an entrypoint, the goal is to delimit a region where the "most amount of information" can be found, and grow that region until seemingly relevant information becomes sparse. To find this region, several techniques could be explored, but not all of them fit spatio-temporal *storytelling* adequately. One of them would be to perform a simple *Nearest Neighbor*(*NN*) search on the area of study and collect the found entities. *NN* searches, however, are "blind" to the dataspace, i.e, they find entities without relaying information about how they disperse, and thus not used here. Another alternative method is *Pair Correlation Function* (*PCR*) [Reed and Gubbins 1973], which divides the data space into spatial segments, allowing each segment to be weighted higher (lower) for closer (farther) entities. In spatio-temporal *storytelling*, however, we only need close-by regions, thus segmenting them does not serve a useful purpose. *PCR*, therefore, is not an ideal choice. Other possible methods are the variations of partitional clustering, such as *K-means*,

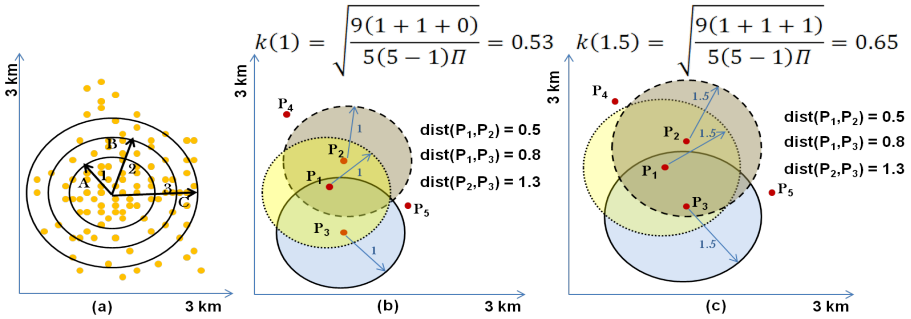

Fig. 6. Spatial scaling for different radii. (a) Circle A depicts high entity density, becoming more sparse in circles B and C. (b) and (c) shows the calculation of *Ripley's K* function for a 1 and 1.5-km radius respectively.

which could serve to group related entities before linking them. While feasible, this type of clustering demands several initialization centroids, which *stortytelling* does not provide (in our approach, only one entrypoint is initially given). In addition, this early in the process, performing any type of clustering adds complexity that can be avoided by other approaches. Below, we explain our preferred method.

Consider Fig. 6(a) where each point represents a person who tweeted during the *Boston Marathon Bombings* near the blast sites. Circle A designates an area of 1 km around the entrypoint (i.e., blast site) with a high concentration of person entities. If we consider 2 km, as in circle B, the density decreases, while circle C becomes even more sparse. Intuitively, the investigation should focus on the 1 or 2-km radii where most of the information resides. In theory, this is the modeling of a point process (i.e., a collection of persons who sent tweets) in terms of a randomly chosen event E (i.e., bombing) with an estimator distance function for a given density λ, which is given by *Ripley's K-coefficient* $K(r) = \lambda^{-1}E$. Mathematically, $K(r)$ can be stated as:

$$K(r) = \sqrt{\frac{A \sum_{i=1}^{n} \sum_{j=1}^{n} w(i,j)}{\pi n(n-1)}}, i \neq j \tag{1}$$

where r is a desired radius originating at a chosen entrypoint, n is the total number of entities in the data space, A is the entire area of study, and $k(i, j)$ represents a weight. $K(i, j) = 1$ if distance(e_i, e_j) $< r$, and 0 otherwise. In effect, $K(r)$ performs a nearest-neighbor search and can be viewed as a clustering coefficient for a desired type of entity (e.g., persons sending tweets) within a limited radius. The coefficient can be evaluated at different scales, such as $r = 1$ km or $r = 1.5$ km. Fig(s) 6(b) and (c) show two simple calculations of the *K-coefficient* for 3 persons {P_1,P_2, and P_3} located in a (3 km x 3 km) area A. In 6(b), the chosen radius is 1 km. The calculation follows: using each entity P_i as the center of a 1 km circle, count the number of other entities P_j within that radius, adding 1 if their distance is less than the radius, zero otherwise. In that range, P_1 "can see" 2 others (P_2 and P_3), since their respective distances ($dist(P_1, P_2)$ and $dist(P_1, P_3)$) are both less than $r = 1$. Using P_2 as the center of a 1km-radius, P_2 "sees" only P_1. The same is true for P_3, which yields $K(1) = 0.53$. In Fig. 6(b), the radius is increased to 1.5 km, and the calculations are repeated, yielding a $K(1.5) = 0.65$.

Comparing the two calculations indicates that the larger radius picked up more points and resulted in more clustering, with the same density. Increasing the radius can potentially find more empty space, which is undesirable. *Ripley's K-coefficient* is an elegant method of discovering related nearby things, but does not tell what a good radius should be or whether lower/higher density is better or worse. *Ripley's* gives us an opportunity to present a set of heuristics that calculates a feasible $K(r)$ in the discussion below.

4.1.1. Finding a feasible K(r). In our analysis, we need a systematic way to determine if the 1-km radius is better than 1.5 km, or vice-versa, to avoid guessing. The most adequate region is where the storytelling process will initiate. Given that a real-world dataset may contain millions of entities, a feasible region is one that includes enough data points, but not all of them. Looking at Fig. 6(a), Circle B covers most of the entities in that dataset, which may be excessive for many applications. The problem is that Circle B has a 2-km radius, which corresponds to most of the length of the entire study area of $9km^2$. A better approach in this case can be done according to Algorithm 1, which we explain below.

We begin by picking an initial random radius, which the algorithm specifies as 1/2 the length of the data set (r_i in Line 1). This initial radius can be manipulated higher or lower to comply with application needs or when better knowledge of the dataset is known apriori. Using radius r_i from the story's entrypoint, a list of entities is obtained by performing a range query over the spatially-indexed entities in the database L (Line 2). A simple check is then made: if the ratio of retrieved entities (|Ents|) and total number of entities ($|e_A|$) is equal to or greater than a certain threshold, say 10%, then too many entities have been retrieved (Line 3) and they are discarded (Line 5). The algorithm halves the initial radius (Line 6) and tries again (Line 7). Once the calculation hits a point below the threshold, the algorithm has found r_{Limit}, i.e., a radius that covers an adequate number of entities (Line 8).

ALGORITHM 1: Distance Computation

inputs : spatially-indexed entity database L, area A, entity count threshold T_e, number of entities in A $|e_A|$, story entrypoint e

output: radius r_i

1: initialize: i=1; k = i-1; $r_i = \frac{length(A)}{2}$; // set the initial radius as half of the length of the study area (customizable).
2: List {Ents} ← rangeQuery(L, e, r_i) ; // create a list of entities by performing a range query from the radius.
3: **if** $\left(\frac{|Ents|}{|e_A|} \geq T_e\right)$ // compare the list of entities against a desired threshold
4: **then**
5: discard {Ents} ; // if too many entities are found , discard them all.
6: **set** $r_i = \frac{r_i}{2}$; // shorten the radius by half of the previous size.
7: iterate Line 2 ; // and run a new iteration with the new radius
8: **set** $r_{Limit} = r_i$; // save the newly found radius.
9: $K(r_i) = calculateK(r_{Limit})$; // calculate Ripley's K function for the new radius.
10: initialize $K(r_k) = 0$;
11: **while** $\left(K(r_i) > K(r_k)\ and\ \frac{|Ents|}{|e_A|} < T_e\right)$ // run more iterations until K stops increasing and threshold is not met, then output r_i
12: **do**
13: {Ents} ← rangeQuery(L, entrypoint, r_i) ;
14: $K(r_k) = K(r_i)$;
15: **set** $r_i = r_i + \frac{r_{Limit}}{2}$; // as long as $K(r_1)$ keeps increasing, increase r_i by half its previous value.
16: **set** $r_{Limit} = \frac{r_{Limit}}{2}$; // save the new radius temporarily.
17: $K(r_i) = calculateK(r_i)$; // calculate Ripley's K function for the increased radius.
18: **end**
19: **output** r_i;

On its own, r_{Limit} is possibly good enough, but not necessarily the best radius. For example, it is possible that r_{Limit} corresponds to Circle B of Fig. 6(a). Ideally, however, we would like to find Circle A, or even a smaller circle inside of A, as they seem to concentrate most of the entities. Our goal, then, is to find the highest clustering coefficient beginning with r_{Limit}, which we store as $K(r_i)$, through an iterative process, but one which does not exceed threshold T_e. Using r_{Limit}, $K(r_i)$ is computed (Line 9). In successive steps, r_i is incremented by half the value of r_{Limit} and its K is recomputed (Lines 11-17). As soon as $K(r_i)$ stops growing from its previous value or the number of retrieved entities reaches threshold T_e, the process stops. $K(r_i)$ has reached an adequate coefficient

for this specific radius, which is output in Line 19. In theory, there is no guarantee the "truly best" radius has been found, but since increments of r_{Limit} become smaller and smaller over many iterations, we hit the law of "diminishing returns" and stop the process for the sake of efficiency. We now state that the storytelling process will include all entities located within range r_i of e.

4.2. Spatio-temporal Propagation

Subsection 4.1 finds entities according to a desired radius of observation. It is now necessary to organize those entities such that they are not only spatially correlated, but also time-ordered in a way that makes sense to the human mind. The key concept here is that entities evolve along space and time, and thus we use the notion of *spatio-temporal propagation*, which adds coherence to facts and becomes an integral part of the storytelling process. In the *Boston Marathon Bombings* scenario, for instance, it is clear that the BOMBING event should precede the arrest of suspect D.TSARNAEV, and not the other way around. Temporal propagation over *Twitter* data is challenging for three reasons:

(1) In many instances, entities are spread throughout long periods of time (e.g., a war), while in others the time span can be very short (e.g., a terrorism act). Therefore, varying-length time intervals must be accounted for;
(2) Often, entities display bursty behavior. In an initial time period, for example, millions of tweets can be issued due to a high-visibility event (e.g., Barack Obama's election). But that same event may subside over time when it is no longer considered "news". Thus, distribution becomes important;
(3) Many entities may be observed at the same time, in which case ordering them is not intuitive. Therefore, ties must be somehow dealt with.

One way to get around the above problems is to utilize a *time matrix*, which provides an intuitive way of aggregating spatial entities in flexible time intervals. In a *time matrix*, each column is a *time unit* and each row is a fraction of the *time unit*. Each cell of the matrix holds the entities observed at specific times. Fig. 7(a) shows an example where each column represents one day of the week (i.e., the *time unit*), and each row represents the time of the day. A *time matrix* permits entities to be observed as a sequence of interactions and can be made as short or as long as the situation dictates. One can then perform data analysis on the entire matrix or on a subset of rows and columns, which we denote as a *time window*. In the scope of this study, a *time window* is defined with a simple rule:

Definition 4.1. Given a *time matrix* of interest (TM) composed of n *time units* (TU), a time window (TW) is composed of one or more TU_i where $0 < i \leq n$. In other words, a time window corresponds to a pre-defined time interval or a subset of it.

For example, consider the *Boston Marathon Bombings*, where some of the developments took place over 7 days starting on April 15, 2013. We can establish its time matrix as **TM = *one week***, each *time unit* TU_i = ***one day***, and each column as the hours/minutes of the day, with n =**7**. Fig. 7 shows the corresponding time matrix, where $TU_1 = 15Apr$, $TU_2 = 16Apr$, etc... For more granular applications, the time matrix can be adjusted to one day and each *time unit* can be the minutes/seconds of the day. The point is that the user must determine the time units that make sense for the task at hand. Having established the time units, we must now define the length of each time window. A simple approach is to make each time window the same as a *time unit*. In Fig. 7(a), for instance, each time window *TW* corresponds to one *time unit* (e.g., $TW_1 = APR15$ or $TW_2 = APR18$). Alternatively, a time window can be a combination of several time units, as is shown in Fig. 7(b) where $TW_3 = APR19 - 22$.

The time window parameters above are decided on a per-application basis. Once established, each time window can be populated with the entities found according to the method in Subsection 4.1. This is easily accomplished by allocating each entity to the appropriate TW_i based on the entity's timestamp. On *Twitter* data, the timestamp is ideally extracted from text. Since that is not always

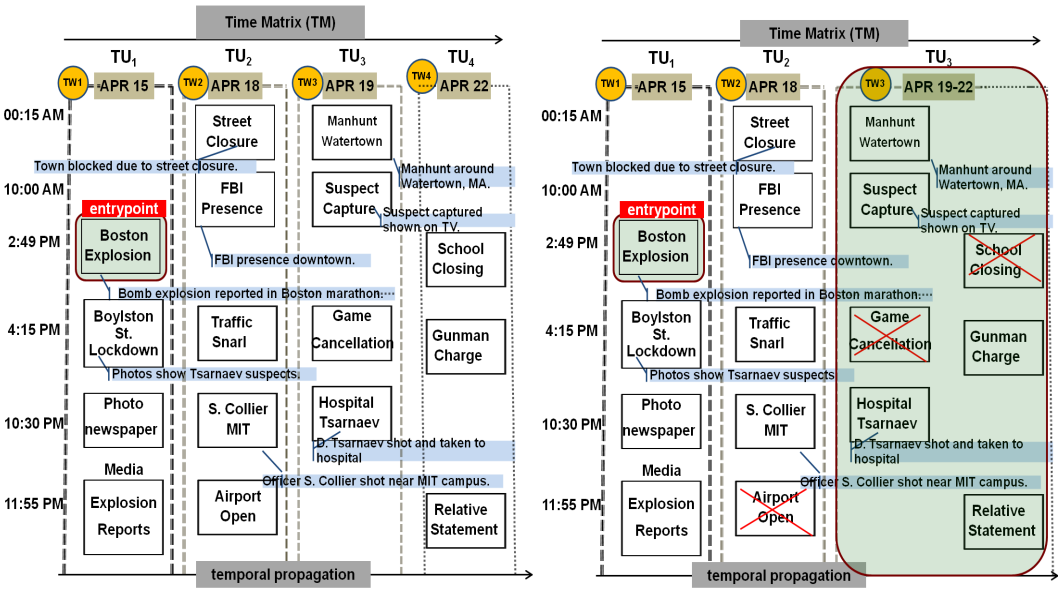

Fig. 7. Visualization of a *Time Matrix*. (a) Temporal propagation of entities in 4 time windows TW1-TW4. Each entity is designated by a box and allocated to a *time unit* TU_i according to the entity's timestamp. (b) The crossed entities indicate that they have been pruned. Time units TU_3 and TU_4 are merged as a new single *time unit* TU_3.

available, the tweet's metadata timestamp can be used as a good-faith approximation. We make one additional caveat: we only insert entities that meet a minimum value of *ConceptRank* (*ConceptRank* is explained in Subsection 4.3). We follow with a visual example.

Fig. 7(a) depicts a partial time interval of four discrete days (Apr 15,18,19,22) related to the *Boston Marathon Bombings*. Some textual description is included for illustration purposes. It is assumed no data is available for the missing days (Apr 16,17,21). Here, we set $TM = 4$ days and set each each $TU_i = 1$ day on an hourly basis. Knowing that the Boston Explosion, which we set as the storyline's entrypoint, occurred on April 15 at 2:49 PM, we place that entity in TU_1. It is followed by the Boylston St. Lockdown at 4:15 PM, and so forth. The same is done for the rest of the days until all entities in the data space have been addressed for that time matrix. This organizational model is not only attractive for its simplicity, but it also serves as a look-up data structure where sequences of developments can be easily found. In Section 4.4, we revisit time windows, putting them to use after computing entity connectivity.

4.2.1. Time Windows Considerations. The model explained above provides an efficient view of time-ordered entities and events, which facilitates reasoning. However, it raises design questions for which decisions must be made and are explained below:

— Entities may span more than one time window. It is possible, for instance, that the Street Closure of April 18 may last several days. In this case, allocation to a time window is done according to the entity's earliest observation time. That entity, therefore, is placed in TW_2 since its earliest occurrence is indeed April 18.
— Entities may have the same timestamp, in which case there is no clear way to order them. In such scenarios, we make the following differentiation: preference is given to the entities that contain either a *semantic constraint* (see *Definition* 3.2) or the most specific location. If the tie still cannot be broken, arbitrary ordering is taken as the last option. For example, if the user seeks semantic constraint "explosion", then entities with such a mention are placed in its time window before another entity that has the same timestamp, but with no such mention. Similarly, an entity located

at *Boylston St.* precedes any simultaneous entity located in *Boston*, since the former location is more specific than the latter.

— Rare entities can be pruned since they provide little connectivity strength (connectivy, an important feature of this approach, is explained in Subsection 4.3). For example, assume that the Airport Open in TW_2, the Game Cancellation in TW_3, and the School Closing in TW_4 appear very few times. In this case, they are removed from the analysis, which is indicated by the red crosses in Fig. 7(b). Pruning removes non-interesting entities, thus saving processing cycles.

— Two time windows TW_j and TW_k can be merged when they are deemed too sparse. For example, in Fig. 7(b), *Apr 19* and *Apr 22* had some entities pruned, leaving them relatively unpopulated as compared to the other TW_i. To save computing resources, they are combined into a single window, namely "*Apr19-22*", denoted by the shaded area. The time sequence of the remaining entities are still preserved.

— In theory, a time window TW can hold any number of entities and can be composed of any number of time units, only limited by the length n of the time matrix. In addition, they do not have to have uniform lengths. However, long time windows, whether uniform or not, may generate excessively long storylines, which in turn tends to become less intelligible. In our experiments, we find that short time windows of one or two time units are not only more computationally efficient, but also allow more coherent storylines than longer time windows.

4.3. Concept Ranking

In Subsection 4.1, r_i is calculated as the radius originating at the entrypoint from where the storyline should propagate. Within that range, many entities can be present, which requires a ranking strategy to determine an order in which entities should be investigated. For this purpose, two common approaches are to perform textual similarity based on methods such as *cosine similarity* [Radinsky and Horvitz 2013] or to compare the values of attributes from each entity [Lin 2008]. These approaches, however, are efficient on textually-rich sources, but not adequate for *Twitter* data, which are more often than not poorly described. Since our data representation is a graph of connected entities, we propose ranking as a variation of *PageRank* [Brin and Page 1998], which we extend as *ConceptRank* and explain it below.

Given a network of web pages, *PageRank* assigns the highest(lowest) importance to the most(least) referenced page(s), offset by the relevance of the referring page. It is given by:

$$PR_{(p_k)} = \left(\frac{1-\Gamma}{N}\right) + \Gamma \sum_{p \in Links(p_k)} \left(\frac{PR_{(p_i)}}{OL_{(p_i)}}\right) \tag{2}$$

where $PR(p_k)$ is the *PageRank* of page p_k, N is the total number of web pages, Γ is a user-defined damping factor in [0..1], $Links(p_k)$ is the set of links to page p_k, and $OL(p_i)$ is the number of outbound links from page p_i. Consider the concept graph of Fig. 8, where each node, instead of a web page, is assumed to be a spatially-tagged entity. It can be seen that T.TSARNAEV has the most **inbound links** (5), MIT has four, and S.COLLIER has only one. The other entities have none. Under *PageRank*, the most important entities (i.e., entities with the highest *PageRanks*) would be T.TSARNAEV, MIT, and S.COLLIER since they are the most connected entities.

One notable aspect of *PageRank* is that it does not differentiate relationships. Thus, in Fig. 8, "stop" and "drive" have the same influence in the *PageRank* calculation as does "following" or any other relationships. In terms of storytelling, this represents a deficiency because the types of interaction among entities relay strong information and should be accounted for. For example, persons seen around the blast site may hold clues to the bombing. However, students commuting to the MIT Campus from other directions most likely play no role in the bombing. Therefore the types of links influence the story and should be discriminated appropriately.

Given the above discussion, we propose *ConceptRank* not on web pages, but rather on entities, as follows. In a concept graph, the relevance of an entity is determined by a combination of both

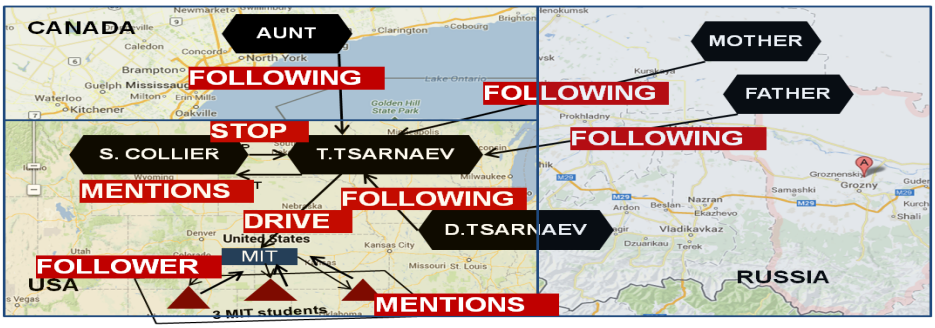

Fig. 8. Concept Graph with Mixed Relationships. Twitter features such as *following*, *follower*, and *mentions* are considered *implicit relationships*. Others, such as *stop* and *drive* are deemed *explicit*.

implicit and *explicit* relationships, as stated in *Definition* 3.3, but differentiated by their respective frequencies. Mathematically, we define *ConceptRank* as follows:

$$CR_{(e_k)} = \left(\frac{1-\Gamma}{N}\right) + \Gamma \sum_{p \in Links(p_i)} \left(\frac{CR_{(e_i)}}{\psi_{e_i}} + \frac{CR_{(e_i)}}{\Phi_{e_i}}\right) \tag{3}$$

where $CR_{(e_k)}$ is the *ConceptRank* of entity e_k, N is the total number of entities in the concept graph, Γ is the same damping factor as before, $Links(p_i)$ is the set of links to page p_i, ψ_{e_i} is the number of explicit outbound relationships of entity e_i, and Φ_{e_i} is the number of implicit outbound relationships of e_i. For all purposes, Φ_{e_i} can be viewed as a *Twitter*-specific parameter obtained from metadata relationships as outlined by the *Twitter features* of Subsection 3.2. In real datasets, explicit relationships are less prevalent while implicit relationships tend to abound, making them less useful in a ranking strategy. We follow with an illustration.

Consider the case in which law enforcement is investigating persons who were **stopped** by a cop, or anybody **driving** to the MIT Campus. The underlined words are the semantic constraints sought on text. The concept graph of Fig. 8 depicts a few interactions related to $N = 10$ entities. We set $\Gamma = 0.75$, which can be viewed as the initial *ConceptRank* value that every entity receives regardless of its connections. This parameter can be manipulated. For each entity i, we must first determine its number of implicit (Φ) and explicit (ψ) outbound relationships. S.COLLIER has one outbound relationship ($\overset{stop}{\rightarrow}$), which is explicit since it comes from *Twitter* text (not *Twitter* metadata), and no implicit ones. Thus its $\psi = 1$ and $\Phi = 0$. FATHER has only one outbound relationship ($\overset{mentions}{\rightarrow}$), which comes from *Twitter* metadata, and so is considered implicit. Thus its $\psi = 0$ and $\Phi = 1$. Table II summarizes the data for all entities, along with their *ConceptRank* (calculations not shown). What the *ConceptRank* values contribute is a ranked list such that the most relevant entities and their relationships can be weaved into a storyline. The ordering goes from highest to lowest values of *ConceptRank*, yielding the following ranking: T.TSARNAEV MIT S.COLLIER MIT students, since these entities have the highest values. The next four entities, (FATHER, MOTHER, AUNT, and D.TSARNAEV) have the same *ConceptRank*, in which case they can be inserted in any order. Given a different mix of implicit and explicit relationships, the ordering may change. In practical terms, *ConceptRank* favors the most well-connected entities, punishing the ones that are thinly-referenced in its spatial region. In the next section, we explain that only the top ranked entities (according to a threshold) are considered. All others are disregarded, preventing them from taking part in the story generation process.

ALGORITHM 2: Storyline Generation

inputs : Entity *entrypoint*, pruned and consolidated Time Matrix TM
output: List *storyline*

1: using *entrypoint* → get radius r from *spatial propagation* ;
2: *entities* ← identify entity set from radius r ;
3: compute *ConceptRank* for each e_k in *entities* ;
4: segregate each e_k into the appropriate tu_i of TM ;
5: **foreach** tu_i *and if* $|entities| < k$ **do**
6: | storyline ← add *top-k* entities in time order
7: **end**
8: storyline ← for each pair of entities, establish their relationship as their most frequent one ;
9: **if** *(storyline should proceed)* **then**
10: | **set** new *entrypoint* = e_{k+1};
11: | *iterate* → step 1
12: **end**
13: **output** *storyline*;

4.4. Spatio-Temporal Storyline Generation

We now put together the ideas in subsections 4.1, 4.2, and 4.3 to generate storylines. Algorithm 2 takes as input the user's desired entrypoint, and an appropriately pre-defined Time Matrix. The essential steps are as follows: obtain the radius of study and identify the entities in that radius (Lines 1 and 2); compute the *ConceptRank* of the found entities and allocate the most important ones to an appropriate time window according to their timestamps (Lines 3 and 4); using each time window, build the storylines with temporal ordering (Line 6); for each pair of entities, select a relationship to insert in between them (the most frequent relatioship is often appropriate) (Line 8); if the storyline is too short or incomplete, a new entrypoint is established as the next highest ranking entity above the top-k ones (Line 10). The process iterates (Lines 9 to 11), otherwise, the storyline is output (Line 13).

The above process may generate long storylines, which may become less intelligible. However, the point at which this iterative process should stop depends on one's own understanding of fact completeness. Insertion of new entities into the graph requires a check to see if the entity already exists, which is done in constant time. Range searches may perform from O(log N) to O(N) depending on the number of location overlaps. Computation of the *ConceptRank* affects only the inserted entities and the ones they link to either directly or indirectly. Fig. 9 shows the propagation of a storyline across four different regions in four iterations [t_i, t_{i+3}] of Algorithm 2. The entrypoints are represented by squares and the other entities by circles. At each iteration, the top 2 entities are linked followed by a new entrypoint, from where a new iteration begins. In this simple example, the four iterations generate one storyline composed of 12 entities (4 entrypoints + 8 other entities) and their relationships.

Table II. ConceptRank Illustration for the network of N=10 entities in Fig. 8 with a starting damping factor of Γ=0.75. Entites are ranked from highest to lowest values of *ConceptRank* $CR(e_i)$.

q=**stop,drive**		N = 10	Γ = 0.75	
i	Entity (e_i)	ψ	Φ	$CR(e_i)$
1	T.TSARNAEV	1 ($\overset{drive}{\rightarrow}$)	1 ($\overset{mentions}{\rightarrow}$)	0.0282
2	MIT	0	3 ($\overset{follower}{\rightarrow}$)	0.0276
3	S.COLLIER	1 ($\overset{stop}{\rightarrow}$)	0	0.0264
4	MIT Students (each)	0	1 ($\overset{mentions}{\rightarrow}$)	0.0250
5	MOTHER, AUNT, FATHER, D.TSARNAEV (each)	0	1 ($\overset{following}{\rightarrow}$)	0.0241

5. DISCUSSION

The generation of storylines as described previously is dependent upon a few parameters for which no general settings will achieve optimal results for every application. In the absence of a general solution, we address below each parameter and some of their implications. Further, we expose other challenges which can impact the efficiency of spatio-temporal techniques, and which the reader should be aware of.

5.1. Parameter Tuning

Below, we briefly discuss the parameters that affect spatio-temporal storytelling in the context of this research:

(1) **radius of study**: this parameter corresponds to the radius r in *Ripley's K* function. It determines the maximal area to be investigated, and, at least in theory, can be infinite. A long radius may find an unreasonably-high number of entities, raising concerns about computational complexity. A short radius may not find enough entities, all depending on the density of the dataset, and whether it is uniform or skewed in some way. In the context of *Twitter*, data tends to be very dense of entities and the radius is better kept short. It can then be increased in small increments, and experimented with. In our experiments, we start with a radius that encloses 1/4 of all entities for high-density datasets, and a longer radius that comprises 1/2 to 3/4 of all entities for increasingly sparse datasets. In general, this has been a good rule of thumb.

(2) **relationship types**: the *ConceptRank* calculation must be able to differentiate between *implicit* and *explicit* relationships, which is determined by the user as a pre-processing step. *Explicit* are the important relationships that contribute strong knowledge to the storylines. Often, they are far and few in between. *Implicit* relationships are informational, but not as important, and tend to occur frequently. Ideally, a good *ConceptRank* should get most of its value from the strong relationships (i.e., explicit) and less from the weak ones (i.e., implicit). For that to happen in the *ConceptRank* formula of Eq. 3, the number of *explicit* relationships should be much lower than the number of *implicit* ones. When the two are close, *ConceptRank* devolves into simple *PageRank*.

(3) **relationship between entities**: in the process of linking two entities, one challenging aspect is determining which relationship to link them with. There can be many options because the same entities can have different interactions at different times. For example, the dataset could have several connections between D.TSARNAEV and BOSTON, such as $\overset{drove}{\rightarrow}$, $\overset{went}{\rightarrow}$, or $\overset{walked}{\rightarrow}$ which begs the question of which one to select for the storyline. The suggested approach is

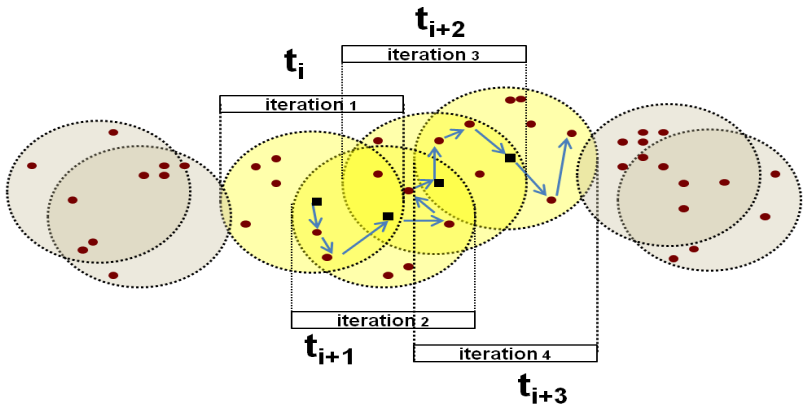

Fig. 9. Hypothetical generation of a storyline through four iterations of the algorithm (t_i through t_{i+3}). Each circle corresponds to one iteration. Squares represent entrypoints and dots represent entities. Each iteration begins at an entrypoint and connects two other entities, before a new entrypoint is considered.

to pick the most frequent relationship between the two entities (stemmed). This often returns reasonable results and is a simple task. For better accuracy, one may want to use an external tool such as WordNet [2013] to consolidate similar relationships into one, get the most frequent, and use it to link the entities.

(4) **damping factor**: this is the Γ parameter in the *ConceptRank* calculation of Eq. 3. For all effects, Γ assigns an initial value to every entity in the concept graph, such that every entity has an initial amount of relevance to begin with. This value should be set in the range [0,1]. The suggested direction is to set Γ closer to 1 such that most of the *ConceptRank* comes from the relationships, and not from the the initial value itself, which, for all purposes, is arbitrary.

(5) **time window**: a well-defined time unit should follow similar principles as the radius of study: it should not be so granular as to cover too few entities, but neither should it be so coarse as to encompass too many. Having too few entities per time unit may force many times units to be looked up, which is counter-productive. Having too many is unnecessary because a storyline is unlikely to include many entities (storylines are usually short). For *Twitter* data, a good time unit represents one day for high-visibility events. For slow-moving processes (e.g., an election), time units of 15 days to 1 month can be adequate.

5.2. Spatio-temporal Challenges

Storytelling is not one single analytical tool. Rather, it can be better described as a framework founded on small principles: entities that have a minimum amount of spatial and temporal proximity, whose connectivity can be measured mathematically, but whose social interactions must be justified semantically. This is what determines coherence, which this study tries to achieve at every level, both mathematically and semantically. Coherence, then, is highly sensitive to the steps that comprise the storytelling process. For this reason, we briefly describe below some of the challenges, pitfalls, and points of contention that the reader may face while generating storylines:

(1) Location and time extraction: Spatio-temporal storytelling requires entities to be explicity placed in space and time. Locations come in various flavors, such as *well-known places* (e.g., New York), *points of interest* (e.g., Statue of Liberty), common addresses (e.g., 123 Main Street), and geo-political boundaries (e.g., Cumberland County). They must be geocoded as latitude and longitude to allow for spatial operations. Ideally, granular locations (i.e., a specific address) should be used over coarser ones (i.e., a city) as they relay stronger accuracy. When many locations for the same entity can be identified, a simple, but often reasonable approach is to take the centroid of all locations as the approximate location for the entity in question. As for the time dimension, more than one timestamp can be present, in which case the latest one is often sufficient. Timestamps are less of a problem in near real-time data, such as recent tweets, because entities can be tagged with the tweet's issue time. For long-standing datasets, however, this assumption may not hold.

(2) Entity identification: Because entities represent the building blocks of storylines, correctly identifying and extracting them from raw data is imperative. This is often accomplished with third-party *NLP* tools (e.g., LingPipe [2013], Stanford NER [2013], Alchemy API [2013]), some of which perform better than others under different situations. As a consequence, different tools should be evaluated for maximum performance. In addition, there is the ongoing problem of *entity disambiguation*, in which case the same entity is described differently in various datasets, preventing them from being identified as a single element. As much as possible, the user should strive to pre-process these entities in an attempt to minimize ambiguity. Scientific literature in this field is plentiful, but we do not endorse any specific works in this study.

(3) Relationship binding: This is one of the most challenging aspects of storytelling. Entities must be connected to other entities through relationships, and deciding on what these relationships should look like directly impacts the computation of the *ConceptRank*. The reader should keep the following in mind. In any application, there should be a differentiation between more important and less important relationships. The more important ones should be limited in number

(we denote them as *explicit*), so they can contribute more to the calculation of the *ConceptRank* of Eq. 3. The vast majority should be left as less important (e.g., implicit) since they contribute less information. One pitfall of this view is that there is no clear way to designate what should and should not be important other than relying on the reader's own understanding. To compound this problem, a pair of entities can have multiple relationships. When this happens, we have relied on the most frequent relationship to bind the entities in the storyline. There are other equally-valid approaches, however, such as using the most recent one or the one whose entities are mostly spatially-proximal.

(4) Data pre-processing: Earlier in Subsection 4.1, we propose *Ripley's K function* to find dense regions where storylines can be investigated. To optimize the process, sparse datasets should be condensed by removing empty regions where few or no entities reside. Entities for which no location or time is available should be removed altogether. These steps can go a long way towards minimizing running time.

6. EMPIRICAL EVALUATION

Establishing a fair experimental measure for storylines is challenging for several reasons. First, since a story may not have an absolute end to it, the traditional *IR* notion of *recall* is not applicable, and thus we propose our own. Second, storylines often have different versions as told by different persons. Thus, *precision* cannot be established with certainty. Third, there are currently no associated benchmarks for this type of information retrieval task. With these in mind, we can still demonstrate the effectiveness of our work using a variation of methods which we explain below.

The experiments consider several datasets, run the storyline generation process, and corroborate them against published news articles. We also elaborate on the variation of parameters and explain their influence on the storylines themselves. The following tools were used in our implementation: Twitter API [2013] as the data source, Stanford NER [2013] for entity extraction, LingPipe [2013] for concept extraction, Neo4j [2012] to build the concept graph, Matlab for the *ConceptRank*, and Gephi [2013] for graph visualization. Using R-Tree indexing from PostGIS [2013], locations are indexed for spatial range queries according to Algorithm 1.

The following **metrics** are used: (1) in the first experiment (*Mexico Civil Unrest*), we use **recall** $= \frac{matches}{total\ \#\ GSR\ events}$ as the ratio of events that we can identify over the total number of events from our ground-truth dataset (GSR, which we explain later); (2) in the second experiment (*Brazil-WCS Olympics*), we use **recall** $= \frac{conf}{all}$, the fraction of generated storylines that can be confirmed by one or more news articles in the web ($conf$) and the total number of generated storylines (*all*); (3) **extension**: the number of evolving storylines that occur beyond the point of confirmation, but cannot be asserted by news articles. For example, the storyline about a person confirmed to be involved in event A, but not in event B. Event B (and its entities), in this case, represent an evolving story. Extension is important in exploratory analysis because it suggests other plot lines, whether legitimate or not.

In order to quantify how storylines are bound to one another, we also utilize the concept of ***coherence*** as defined in [Shahaf and Guestrin 2010]. It is a simple and intuitive method based on word occurrences that applies to textual datasets. While our approach is spatio-temporal, we are still interested in verifying that our algorithms do not suffer when mostly textual data is available, as opposed to times and locations. *Coherence* is given by $Coh(s) = \sum_{i=1}^{n-1} \sum_c 1(c \in d_1 \cap d_{i+1})$. Given a set of documents $D = \{d_1, ..., d_n\}$ (in our case, tweets), every time a concept c appears in temporally-ordered documents d_i and d_{i+1}, the coherence for storyline s increases by one unit. These values can then be normalized.

To provide a reasonable mix of content, the next subsections describe four sets of experiments with different purposes: the **Civil Unrest in Mexico** section applies spatio-temporal storytelling as a means to forecast events in simple, intuitive ways. The goal is to verify how far in advance certain events can be forecast before they are published in the news to indicate that storylines can be useful in different applications. The **Brazil WCS-Olympics** section seeks storylines based on a

Table III. 10 instances of civil unrest reported in the Gold Standard Report (GSR-IARPA). Each event is related to protests against educational reform in Mexico City in 2013 and other locations throughout the country. Only events located within 450 km of Mexico City are shown. For each reported event, the table shows its reporting source, originating location, and date it was published in the news. Storylines are then generated using the entire set of approximately 100,000 tweets, from where only the *Related Tweets* contribute entities to the storylines. The *Forecast Lead Time* column shows that the storyline from tweets are generated even before the news article is published, often days in advance.

	GSR Event	Reported By	Event Location	Published	# Related Tweets	Forecasting Storyline	Generated Date	Forecast Lead Time
1	SNTE Protesters block Eje Central; demand pension pay.	milenio.com	Mexico City	Jan-03-2013	5,422	**EDUCATION** fighting **SNTE** paying **SALARY** lower **FUNDS**.	Jan-02-2013	1 day
2	Teachers protest in Michoacan; demand Christmas pay.	milenio.com	Michoacan	Jan-04-2013	1,410	**FIGHT** break **STUDENT** distribute **FUNDS** sending **MORELIA**.	Jan-01-2013	3 days
3	Stop at Oaxaca University affect more than 20 thousand students.	milenio.com	Oaxaca	Jan-12-2013	2,051	**EDUCATION** halt **UNIVERSITY** remove **STUDENT**.	Jan-08-2013	4 days
4	SNTE professors at Aguascalientes will march against education reform.	lajornada.com	Aguascalientes	Jan-14-2013	1,960	**TEACHER** protest **EDUCATION** lower **FUNDS**.	Jan-10-2013	4 days
5	SNTE teachers walk in Veracruz against education reform.	milenio.com	Veracruz	Jan-17-2013	2,737	**TEACHERS** lose **FUNDS** remove **BUDGET** impact **EDUCATION**.	Jan-10-2013	7 days
6	Teachers block Morelia-Toluca in Zitcuaro.	lajornada.com	Zitacuaro	Jan-19-2013	734	**ROAD** blocked **PROTEST** include **TEACHERS** ask **FUNDS**.	Jan-11-2013	8 days
7	Several incidentes reported during SNTE's march.	milenio.com	Pachuca	Feb-01-2013	1,155	**FIGHT** breaks **CITY** drain **FUNDS**.	Jan-28-2013	4 days
8	Teachers march against labor reform in Tlaxcala.	milenio.com	Tlaxcala	Mar-14-2013	3,938	**EDUCATION** march **TEACHER** lower **BUDGET**.	Mar-02-2013	12 days
9	In Acapulco, SNTE teachers from San Marcos will march.	lajornada.com	Acapulco	Mar-14-2013	1,021	**TEACHERS** march **CITY** protest **EDUCATION**.	Mar-13-2013	1 day
10	SNTE teachers march in Atlixco.	milenio.com	San Pedro Atlixco	Mar-28-2013	2,760	**SNTE** march **TEACHERS** participate **PROTEST**.	Mar-20-2013	8 days

variation of semantic constraints, radius, and number of tweets that affect the number of generated storylines, recall, and coherence. It also investigates how far in advance the generation of a storyline can precede the publication of a related news article and the number of tweets compressed into a single storyline. The **Boston Bombings** section observes variations is number of tweets, time window size, and relationship types, and how they influence recall. Lastly, the **Syrian Civil War** part provides visualizations that are *Twitter*-centric, such as the concept graphs for locations and hashtags, and how they affect the generation of storylines.

6.1. Civil Unrest in Mexico (2013)

Storylines can be useful in many different applications. In this section, we take a lightweight view of *forecasting*, and discuss how storylines are able to foresee real-world developments before they are published in the news. Intuitively, *Twitter* data is available in near real time, while news reports lag behind for hours and sometimes days before their dissemination. Our goal, then, is to generate storylines, identify which real-world stories they relate to, and determine how far in advance the storyline "forecasts" the real event.

The dataset is composed of approximately 100,000 tweets related to *civil unrest*[1] in Mexico for the first three months of 2013. We target events related to *education reform*, which has provoked social strife in Mexico, and documented as part of the Gold Standard Report (GSR) from the *Intelligence Advanced Research Projects Activity* [IARPA 2013], which serves as our ground truth.

[1]civil unrest denotes an event of social impact, such as a strike or a protest. Violence does not have to be included.

Fig. 10. Spatial propagation of *education reform* protests. Starting from Mexico City, similar events are observed around the country. The map shows 10 of approximately 5,000 affected locations.

Table III shows a sample of 10 such events that we use for discussion. Each *GSR event* has an associated *reported by* source, an *event location*, and *published* date. For each *GSR event*, we show a *forecasting storyline* of no more than five entities and generated by our algorithm from its respective # of *related tweets*. A tweet is related to the storyline if it contains at least one entity also present in the storyline. The *generated date* of the *forecasting storyline* is the timestamp of the most recent related tweet. In the *forecasting storyline*, entities are bolded in uppercase, relationships are not. The *forecast lead time* is the time difference to the *published* date. The starting location is *Mexico City* from where we consider a radius of 450 km that includes other major cities shown in Fig. 10.

Discussion: Item 1 has a *forecasting storyline* of four entities (*education*, *SNTE*, *salary*, *funds*), generated from 5,422 tweets. These five entities are the ones of highest *ConceptRank*, and thus selected for the storyline. The relationships (*fighting*, *paying*, *lower*) are the most frequent ones between the adjoining entities. Note that storylines do not reflect stylized English language. Because they are linked based on spatial connectivity and time order, grammar rules cannot be easily enforced, though they often come out as properly-formed phrases. The *forecasting storyline* closely resembles the associated *GSR event*: even though they only share one entity, i.e. *SNTE*[2], both relay similar messages, that is, teachers protesting for higher wages. The most recent tweet in this set is dated Jan-02-2013, which when compared to the *GSR event*'s *published* date of Jan-03-2013 indicates that the *Twitter* storyline forecasts the real event by 1 day.

While the storyline of item 1 relates to Mexico City, the one of item 2 takes place farther away in the state of *Michoacan*. At first glance, the *forecasting storyline* is not related to the *GSR event* since they appear to have little in common. They are strongly connected, however, in two manners: the semantic closeness between "student fights" and "teacher protests", as well as by location. Most of the 1,410 related tweets have a location inside *Michoacan*, which is a state with many cities, one of them being *Morelia*, an entity that appears in the *forecasting storyline*. The *forecast lead time* is three days, showing that the algorithm was able to reconciliate tweets about fights due to educational factors, which were later reported by the media in the form of teacher protests. This example underscores the importance of location, which would otherwise make this linking difficult to justify.

We must emphasize the importance of the spatial aspect of this study, showing that all remaining items from 3 to 10 are highly-dependent on location. Note that none of those *forecasting storylines* have a location entity explicity stated. However, their related tweets do contain at least one metadata location that matches the location of the *GSR event*, and a timestamp that closely pre-dates the event's published date (within 30 days). This is particularly interesting in the case of item 9, whose

[2]Sindicato Nacional de Trabajadores de la Educacion.

Table IV. Recall results based on 9,304 *GSR events* in four different categories. Recall R1(R2,R3) denotes that the storyline matches the GSR event with one(two,three) common entinty in a designated radius, and within 30 days of the *GSR event*. Locations in Mexico are shown separately from other countries to illustrate the *education reform* scenario.

	Mexico			**Other Countries***		
GSR event type	**R1**	**R2**	**R3**	**R1**	**R2**	**R3**
01-civil unrest (employment, housing, resources, other policies)	0.617	0.552	0.230	0.553	0.507	0.420
02-vote (local, national elections)	0.586	0.430	0.418	0.490	0.430	0.367
03-infectious human illness (rare and common diseases, pandemic)	0.502	0.428	0.400	0.537	0.472	0.311
04-economy (currency exchange, stock market)	0.772	0.512	0.497	0.402	0.324	0.405
*Argentina, Brazil, Chile, Colombia, Costa Rica, Mexico, Panama, Peru, and Venezuela.						

GSR event is shown at *Acapulco*, but whose *forecasting storyline* does not reflect that location. But in fact, all of the 1,021 *related tweets* have a latitude/longitude that closely matches *Acapulco*. Also worth mentioning is the fact that there are a few entities very popular across the tweets, and as a consequence, appear commonly in the storylines. Four of them are *teachers*, *SNTE*, *protest*, and *funds*, which are commonly observed in *Aguascalientes*, *Veracruz*, *Zitacuaro*, *Pachuca*, *Acapulco*, and *San Pedro Atlixco*. The prominence of these entities as part of the storylines is very significant for a simple reason: it indicates that our spatio-temporal methodolody is able to find storylines about civil unrest related to the education reform in Mexico using location as a decisive factor. Even though Mexico City is the most prominent area, the algorithm also identifies other important locations where events occurred with similar entities as the ones in Mexico City, as shown in the circle of Fig. 10. Even more encouraging is being able to identify them several days ahead of time, such as the "march against labor reform" in Tlaxcala (item 8), which is shown in the corresponding *forecasting storyline* 12 days earlier as "education march lower budget".

Apart from civil unrest, the GSR dataset catalogs events of other natures. Table IV shows recall levels for our algorithm for 9,304 *GSR events* in four categories. Here, we define recall according to the equation: $recall = \frac{matches}{total\ \#\ GSR\ events}$, that is, the number of storylines that match at least one *GSR event* divided by the total number of *GSR events*. To deal with different parameters, we make a further refinement: in recall1 (R1), a storyline matches a *GSR event* under the following conditions: they must share at least one entity; the shared entity must be located within the investigated radius (e.g., 450 km in this case study); and the observation of the entity must pre-date the real event by no more than 30 days. In recall2 (R2) and recall3 (R3), at least 2 and 3 entities must be shared respectively. The radius and timestamp conditions remain the same as before. In short, we relate similar events that are close both spatially and temporally, as dictated by our methodology.

Table IV shows two sets of results: one for *Mexico*, since our previous forecast illustration was based on it; and one for other Latin American countries in order to add data variation. It shows a wide range in the recall values, with differing root causes:

— Storylines in R1 (i.e., the ones that share one entity with the *GSR event*) have good recall when only Mexico locations are considered, especially in the civil unrest and economy categories (0.617 and 0.772). The corresponding R1 values are significantly lower when other countries are included. The reason is the targeted events related to *education reform*, which is very prominent in Mexico, but not very common in the other countries. This gives us the first lesson: storytelling benefits when the scope is targeted. In other words, the examined topic should be specific enough to a region in order to maximize recall.
— When the match is increased from one to a minimum of two or three entities (i.e., going from R1 to R2 to R3), there is a significant drop in recall. Making the parameters increasingly more strict prevents storylines from being identified as a match to the *GSR event*. This can be seen in the *vote* event type, for which R3 is very low for Mexico (0.418), and even lower for other countries (0.367). The second lesson is that the requirement of having more matches tends to find very coherent storylines, but will almost always find very few of them. This effect can be relaxed by increasing the number of investigated locations.
— These experiments consider a radius of 450 km from the country's capital. It should be apparent that such a large radius would have different effects in large countries, such as Argentina and

Brazil, as opposed to smaller ones, such as Costa Rica and Panama. However, we find that large radii can be very appropriate in both situations, with one caveat: the application must be highly targeted for specific event types in order to avoid data explosion. For example, elections across the country would make sense when viewed in large areas. Therefore, the third lesson is that short radii may not find a significant number of entities that are global enough for forecasting. Local forecasting is certainly applicable, but may require more granular spatio-temporal reasoning.

— Data volumes and variation are important factors in recall levels. For example, in the case of event type 3 (*infectious human illness*), there are 1,916 *GSR events* for all countries, out of which 128 refer to Mexico. However, our dataset does not have a significant number of tweets related to diseases or health topics. As a consequence, the algorithm is unable to generate very good storylines that can be matched to *GSR*, causing the recall levels to be low. At its worst, recall is only 0.400 for Mexico, and 0.311 for other countries. Even though this is a data problem, and not a weakness of the approach, it brings up the fourth lesson: low data volumes and poor variation creates storylines that lack meaning and appear disconnected from real events.

In general, the recall levels shown in Table IV are very promising in the scope of spatio-temporal storytelling. Even when the values are low, they can be justified and remediated in different manners, such as by changing parameters (e.g., radius, countries), focusing on specific domains (e.g., vote, economy), relaxing or restricting match requirements, and verifying data volumes and variation. A well-tuned algorithm shows extremely high potential in a forecasting strategy.

6.2. Brazil WCS-Olympics (2014-2016)

In this experiment, we are interested in discovering storylines related to the planning of major sports events. They often report problems such as corruption, budget overruns, and construction delays, which provide a good testbed for our evaluation (the recent protests in Brazil make this case study particularly relevant). We consider our experiments successful if we can find a storyline with our algorithm, and confirm it with an indepedent news source. Just as important, however, is to explore the applicability of storylines to common analytical tasks. For this purpose, we take two directions: (1) observe the forecasting potential of the storylines, i.e., if a storyline points to an event, verify how far in advance the storyline was generated ahead of that event (similar to the Mexico experiments); (2) observe the compressibility potential of the storylines, i.e, verify how many tweets can be represented by a single storyline. We say that a storyline compresses a tweet if that tweet contains at least one entity that appears in the storyline (i.e., the tweet has at least one top-k entity, according to Algorithm 2).

This dataset is composed of 14,460 tweets from where a concept graph of approximately 30,000 entities and 7,000 relationships was generated. We performed several runs, limiting the geographical radius to 450 km, which is the approximate distance between São Paulo and Rio de Janeiro, the two largest cities in Brazil. The initial entrypoint is an entity found in a tweet that has at least one of the semantic constraints, and geocoded within the 450 km radius. For example, we used "Fifa" as the entrypoint of the first storyline of Table V because it occurs in many tweets where the semantic constraint "finance" also occurred and those tweets were geocoded to Rio de Janeiro or vicinity. Note that the entrypoint is also the first entity of each storyline. For these storylines, we look for no more than 10 entities at a time ($top\text{-}k$=10). We set the semantic constraints, which are initially stemmed, as listed on the bottom of Table V, which shows 2 verifiable storylines identified in the process.

6.2.1. Discussion. We first explain how to interpret Table V, which contains two sample generated storylines numbered in the first column. The second column shows a few of the tweets that contributed to the generation of the storyline. The red entities in the tweets correspond to the large-font entities in the storyline. The blue relationships from the tweets are also part of the storyline. Column 3 shows the storyline itself along with the compression ratio. In other words, storyline 1 was generated by 854 tweets, 10 of which are shown in the "Sample List of Tweets" (column 2).

Table V. Generated storylines with corresponding published news articles. The *sample list of tweets* is a small subset of all tweets used in the generation of the corresponding *generated storyline*. In the *generated storyline*, entities are uppercased in bold type, whereas relationships appear in lowercase. Entrypoints are indicated by the small black squares. The compression ratio indicates the number of tweets that contributed at least one entity to the generation of the storyline.

	Sample List of Tweets (February 02 - April 15, 2013)	**Generated Storyline**	**Published News Article**
1	**@pkedit** **FIFA** pays big price for having **Havelange** propped up on the board. big examples of corrupt pols to Brazilian public. **@nolutha0209** **FIFA** is conveniently waiting for **replacement**, put **Havelange** house in order. **@jasondevos** bright minds with full backing of **FIFA** leader and the **Havelange** case. **@johnbranchnyt** **FIFA** board and soccer's quagmire of corruption and life after **Havelange** legacy. **@pitacodogringo** time for **fifa** to get rid of **havelange**. **@aldridge_12** ask Blatter where all that **money** go? expensive stadiums won't be ready, brazil tells **fifa**. **@migueldelaney** With **Havelange** oust looking more and more likely, a piece on why he wasn't best choice for the **money**. **@tumblingeek** Watch Mashable's Live Hangout with **Fifa** Soccer **president** defending **allegations** of corruption in Rio. **@devinwells15** lost cause when Switzerland ignores money trail just cause **fifa president** is involved in the **bribery**. **@giavanni_ruffin** big stadium big **money** has no accountability and **fifa** isn't different, just like Salt Lake City.	FIFA■ defending PRESIDENT rid HAVELANGE oust ALLEGATIONS ignore MONEY involved BRIBERY waiting REPLACEMENT. compression ratio: 1 storyline that compresses 854 tweets.	João Havelange resigns as honorary FIFA president amid World Cup bribery scandal ★ cbsnews.com 04-30-2013
2	**@ajelive** **police** cover against **protesters** all over, no place for football, one death already. **@tvh11** Rio de Janeiro maracana **stadium** will have **security** station ahead of 2014. **@robharris** **protesters** watched in **surveillance**, but few believe in **security** efforts. **@erguncaner** **police** being defeated, complicated mission, hospital under control after **stadium** violence. **@blakeharrison23** of course people will **protest** **brazil** needs hospitals. **@f365** **police** trying to boost **security**, some guy smashed a bank window ran away, was a masked guy. **@eltonnarabia77** haha, want to see what's gonna happen during the **Olympics**, **security** last priority @oliveirakk. **@canalglobonews** International experts arriving in **Brazil** for **Olympics** symposium on **security** practices. Officials to acquire **antiterrorism equipment**. **@samuelj29060** good piece on how **security** is so tight on football matches, like Israel and Turkey preparing for war not **Brazil** wc. **@mogdentelegraph** gareth bale forcing Wales to tighten **security** ahead of World Cup in **Brazil**, qualifiers coming up. fw.to/ONdCnIi	PROTESTERS■ watched SURVEILLANCE believe POLICE cover STADIUM boost SECURITY EQUIPMENT ANTITERRORISM OLYMPICS BRAZIL. compression ratio: 1 storyline that compresses 575 tweets.	Israel-made drones to protect FIFA soccer cup from terrorists. ★ jewishpress.com 02-25-2013
	semantic constraints = {*corruption*, *fraud*, *finance*, *olympics*, *football*, *purchase*, *construction*}		
	■ entrypoint entity ★ confirmed story		

The last column shows a published news article that best confirms the contents of the generated storyline. As explained earlier, the storyline is formed by linking entities from the tweets using relationships (verbs) also from the tweets. The selected relationships represent the most frequent ones, as explained earlier.

Our first storyline (item 1) refers to Mr. Havelange, a former president of FIFA, soccer's governing body. Many tweets mention some of the organization's dubious transactions with terms such as *bribery* and *corruption*. As a well-known entity, Mr. Havelange is highly referenced, which boosts his *ConceptRank*. Looking at his Twitter Storyline in Table V, it can be seen that he is linked from FIFA→PRESIDENT to →ALLEGATIONS→MONEY→BRIBERY→REPLACEMENT, which among the found entities, are the ones with the *top-10* highest *ConceptRank* values (we omit the relationships due to space constraints). A *cbsnews.com* article from April 30, 2013 corroborates the fact that he indeed resigned because of corruption allegations, which we denote with the symbol ★. The tweets that generated this storyline span the interval from April 02 to 15, 2013, which precedes the article publication date (April 30, 2013) by 15 days. This case study gives us two takeaways: first, storylines are able to efficiently compress large streams of information that appear seemingly disconnected (i.e., many tweets→few storylines). The summary is a contextual view of sports corrup-

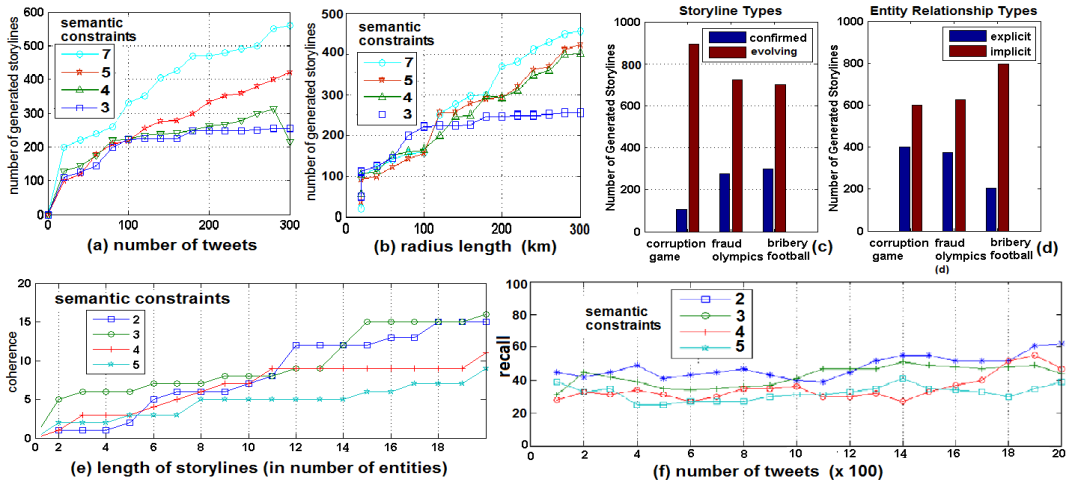

Fig. 11. Influence of different parameters on storytelling. (a) number of tweets on generated storylines (b) length of radius on generated storylines (c) confirmed/evolving stories (d) implicit/explicit relationships (e) length of storylines on coherence (f) number of tweets on recall.

tion in terms of their entities and social interactions; second, storylines have a strong forecasting potential, as it points many events to a potential outcome before that outcome takes place (e.g., bribery→resignation). This has far stronger usage than applications such as *Twitter Trends*, which is limited to displaying frequent keywords of the moment, but does not provide inference or reasoning.

Further investigation points to a second storyline corresponding to the sequence PROTESTERS → SURVEILLANCE → POLICE → STADIUM → SECURITY → EQUIPMENT → ANTITERRORISM → OLYMPICS → BRAZIL. The term *protesters* is now the new entrypoint, which intuitively makes allusion to concerns about current protests taking place in Brazil, and the equipment that will be used for antiterrorism. This storyline comprises tweets from February 02 to February 20, 2013, and confirmed by a news article published on February 25, 2013. The storyline has a 5-day lead on the news article, which announced that drones would be acquired as antiterrorism equipment. This storyline is a summary of 575 tweets. These results are encouraging because they guide the storytelling process into uncharted, non-obvious, but relevant threads that are valuable for exploratory analysis.

6.2.2. Effect of the number of semantic constraints on the number of generated storylines. Starting a story requires one to designate concepts of interest such as “attack” or “theft”. These semantic constraints are crucial not only because they influence the storytelling process from a semantic perspective, but also because they limit the search space. Therefore, we are interested in observing how the number of semantic constraints affect the number of generated storylines for our dataset. Fig. 11(a) shows that with 7 semantic constraints, it takes 300 tweets to genereate ~570 storylines. With 3 semantic constraints, the same can be achieved with about 240 tweets. Fewer *semantic constraints* tend to include a higher number of entities, while generating a higher number of storylines, which can make reasoning difficult. The converse tends to generate less storylines, but may miss relevant ones. Therefore a trade-off arises. Note that the number of generated storylines is highly dependent on how well connected the tweets are, such as users mentioning one another or heavy usage of the same hashtags. In our case, we obtain good results using no more than 7 *semantic constraints* at a time, after which the storylines become too restrictive and less understandable.

6.2.3. Effect of the radius length on the number of generated storylines. The radius of study is important since we hypothesize that a story evolves sequentially in space and time, and thus growing in small steps may be justifiable. In practice, a short radius will tend to find a lower number of entities than a large radius. However, there is no immediate evidence that short radii are advantageous over

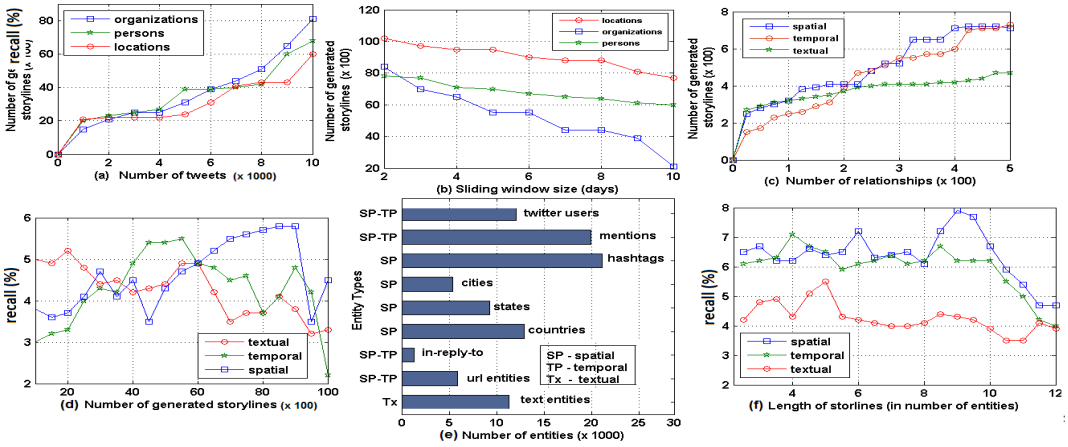

Fig. 12. Influence of entity types on storytelling. (a) by number of tweets (b) by time window size (c) by relationship types (d) recall by entity types (e) entity distribution (f) recall by length of storyline

longer ones in terms of finding better or more complete stories. In Fig. 11(b), we vary the radius of study up to 450 km, and observe the number of generated storylines for different numbers of semantic constraints. At 100 km, for example, approximately 170 storylines are generated for both 4 and 5 semantic constraints. Going to 3 semantic constraints, ~220 storylines are generated, even though we expected this number to increase more sharply. The interesting fact is that the number of generated storylines does not increase significantly simply due to radius variation. The reason goes back to the semantic constraints which tend to put a bound on the number of storylines in spite of the increase in the radius. Our tests also points to other reasons. Since we deal with *Twitter* data, we often encounter imprecise locations, which cannot always be identified correctly. For example, state and city names are relatively simple to disambiguate, but streets are particularly more challenging. Thus in some of our experiments, we also allow city-wide distances.

6.2.4. Effect of the length of storylines on coherence. Coherence has to do with the sequence of concepts observed across related tweets. Intuitively, longer storylines are more likely to share concepts, and thus have higher coherence. However, growing the storylines too long has adverse effects, as mentioned previously. Using this metric, we try to observe if higher coherence indeed correlates to good stories. We generate Fig. 11(e) which corresponds to the dataset of Table V, which was verified manually, and serves as ground truth. In the graph, coherence of storylines of length 10 is between 5 and 7 for any number of semantic constraints. Lower than length 10, the coherence remains fairly similar across lines, while greater than 10, it varies more widely, but still does not increase significantly. This fact lets us establish one observation: for our dataset, keeping the storylines short is preferable because it finds enough shared concepts. Longer storylines do not add much coherence. We find that storylines of 5 entities and 4 links are very efficient under the characteristics of *Twitter* data, and that is what we used to generate Table V, which represents 3 scenarios confirmed independently.

6.2.5. Effect of the number of tweets on recall. One of the most contentious points of storytelling is that of *recall*, which does not follow tradition *Information Retrieval*, and thus is difficult to quantify. In this study, we establish it from a subset of our data, a small sample of which is shown in Table V. Every time a storyline is generated, recall is increased if the storyline can be confirmed in the news media. Fig. 11(f) shows some of our results for up to 2000 generated storylines. In the best case, recall reached 61% with a high number of semantic constraints (7). The worst case is 27% from 100 generated storylines of 4 semantic constraints. Recall remains relatively constant throughout,

which possibly has a reason: the semantic constraints selected, such as "corruption", "fraud", and "footbal" seem to find many instances in the data sources, while the radius length seems to have been appropriate. While our recall numbers appear low in some cases, we note that the results are very positive for two reasons: first, many of our storylines cannot be found in the news, and thus not counted in favor of recall, even though many of them appear legitimate. Second, *Twitter* data is extremely noisy, highly abbreviated, and grammatically incorrect, which promotes poor identification of entities and their proper insertion in the concept graph.

6.2.6. Storyline and relationship types. The vast majority of our data generates storylines that cannot be easily confirmed in an automated manner. This is illustrated in Fig. 11(c), where the red bars (evolving storylines) shows a much higher number of instances than the blue bars (confirmed storylines), for several variations of semantic constraints. This remains true in our experiments, and while it is not exactly unexpected, it has a reason: first, our dataset is not simply about corruption and sports, but rather ranges over any types of topics, which makes it challenging to find storylines that make sense to the human mind. Fig. 11(f) shows that most relationships are identified as *implicit* as opposed to *explicit* ones. They have a significant influence on the calculation of the *ConceptRank*, with explicit being better. One of our goals has been to maximize the number of confimed stories as well as the number of explicit relationships. Even though our algorithms are general in purpose and our dataset is very sparse, our results are extremely encouraging. Given more specific applications along with related data, our algorithm can achieve even better results.

6.3. Boston Bombings (2013)

In this experiment, we are interested in the influence of different entity types on the spatial and temporal evolution of a storyline. A spatial entity is one for which location can be obtained, while temporal entities are associated with a timestamp. When neither location nor time is available, the entity is simply denoted as textual. Fig. 12(e) shows the distribution of 9 entity types from the dataset of 10,000 tweets used in this part of the experiments. Most of those entities are spatial, some are temporal, and only tweet content is considered textual. We use semantic constraints "Boston" and "bombing" to generate storylines of 5 entities in length. We claim that space and time are important to storytelling, so we investigate the influence of spatio-temporal entities on the storylines. Our results are given below.

6.3.1. Effect of entity types on the number of generated storylines. *Twitter* data are often rich on people's names (users, mentions, reply-to), but in many instances poor on organizations, locations, and textual content. However, we also observe that a wide mix of distributions is not unusual. Fig. 12(a) shows the number of generated storylines when a single type of spatial entity is considered (e.g., person-person connections). Organizations alone, for instance, generate over 8,000 storylines with all 10,000 tweets, whereas locations generate 6,000 storylines, indicating that entities such as "FBI" were found consistently. Apart from the graph, persons tend to generate more storylines because *Twitter* user names are unique and easy to identify. Organizations, on the other hand, are much less frequent, and are observed in bursts. In this dataset, for instance, "police" occurs in thousands of tweets where "Boston" also occurs, possibly because of the extensive number of retweets after the bombing events of April 15. Locations are not always present, but tend to be very reliable in finding *evolving* stories, whereas organizations are easier to find in the news, and thus appropriate for confirmed stories.

6.3.2. Effect of time window size on the number of generated storylines. As stated previously, a *time window* is a set of entities constrained in a time interval, such as *one day* or *one week*. Its size can be best estimated according to application. In the case of *Twitter*, hour-based time windows can be appropriate at times, but often, day windows suffice. In our experiments, windows are daily and static (i.e., uniform size). In general, if the time window is small, fewer entities are included in it, and less storylines occur. Larger time windows have the opposite trend. *Twitter*, however, breaks this assumption. When a high-visibility event occurs, a few days can be much more lively than the rest of

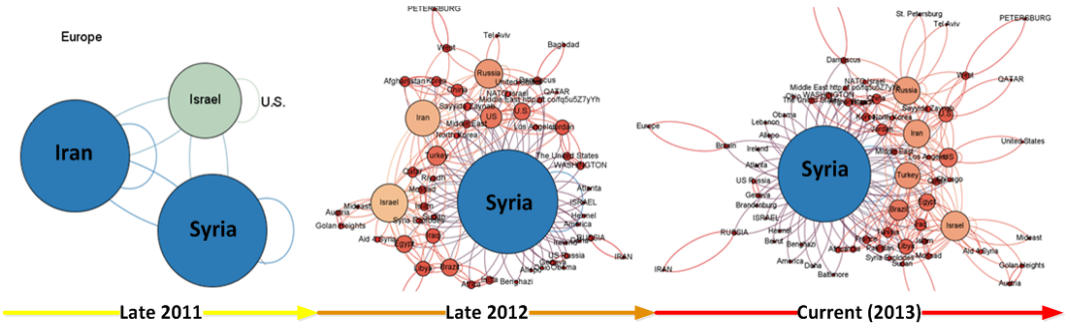

Fig. 13. Spatio-temporal propagation of the *Syrian Civil War*. In late 2011, only a few countries paid close attention to the conflict. The number of related locations to **Syria** surged considerably throughout 2012, continuing in an upward trend in 2013. Size indicates the importance of the node in the graph.

the month. Fig. 12(b) illustrates exactly that: for the *Boston bombings*, organizations had only 2,000 storylines in a 10-day time window, but hit over 8,000 storylines on a separate two-day time window, a significant change. Persons and locations saw a less dramatic trend, though also considerable. This graph indicates that, using this dataset, storylines would be better approached with an organization entrypoint due to its richness in connectivity, leaving other types as alternatives.

6.3.3. Frequency of Relationship Types. In many cases, it is desirable to determine the nature of the relationships that bind entities. A high number of connections among spatial (temporal) entities should not surprisingly make a case for spatial (temporal) propagation. One would assume that spatial relationships (e.g., "person-person" or "location-location") are more common than temporal relationships (e.g., "bombing on April 15" to "arrest on July 16"). For *Twitter* data, however, this does not always materialize. Fig. 12(c) illustrates the number of generated storylines if we consider different types of relationships. It shows that spatial and temporal relationships are not far apart, but both tend to be more prevalent than textual ones. To be fair, we used *LingPipe*, one of the well-established tools in *Natural Language Processing*, to make sure textual relationships were identified consistently. This results further strengthens our hypothesis to pursue *Twitter* content from a spatio-temporal perspective.

6.3.4. Recall Considerations. Given a dataset, it is important to determine if spatio-temporal propagation truly deliver higher-recall storylines than textual ones. For this experiment, we generate 10,000 storylines isolated by entity types (e.g., spatial, temporal, and textual), and use the same **recall** metric as explained earlier. We observe the following. As the number of generated storylines grow, Fig. 12(d) shows that recall maintains a higher trend under spatial and temporal entities than under textual ones. This is often the case for a reason: in our approach, locations are geocoded into their latitude and longitude, allowing many data points to coalesce into a single entity. When combined with a timestamp, they result in storylines that are strongly bound to specific locations and times, and thus can be more easily confirmed as legitimate. Upon further investigation, we realize that our dataset has not only an abundance of "Boston" instances, but also of many other nearby nameplaces: "Watertown", "MIT", "Massachusetts", "Boylston", "Newton", "Sommerville", "Waltham", among others. When these terms are taken textually (i.e., not geocoded) they become separate entities, less connected, and thus fall out of *ConceptRank*'s favor. In effect, they become "invisible" in a traditional storytelling sense, but are consistently identified in a spatio-temporal setting. In terms of storyline length, Fig. 12(f) also describes higher recall for spatio-temporal entities under varying storyline lengths. In this case, however, short stories do not necessarily equate to higher recall. We surmise that the algorithm has been able to find high numbers of well-connected entities such that linking them still maintains semantic cohesiveness.

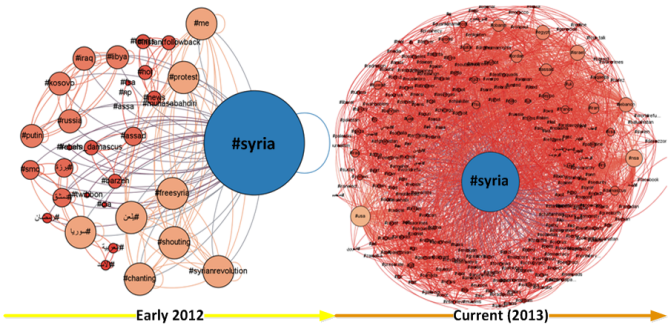

Fig. 14. Temporal propagation of *Twitter* hashtags during the *Syrian Civil War*. In early 2012, the **#Syria** hashtag started a slow uptrend that included a few other hashtags. As the conflict gained international attention, the number of co-mentions that included **#Syria** multiplied significantly. The high impact of the conflict is reflected by the high hashtag density in the period Jan-Apr 2013.

6.4. Syrian Civil War (2011-2013)

In this experiment we are interested in spatio-temporal evolution of entities and events in the ongoing *Syrian Civil War*. At the time of this writing, it is an armed conflict between forces loyal to the ruling Ba'ath Party and those seeking to oust it. The protesters demanded the resignation of President Bashal al-Assad. Although the conflict was originally confined within Syria's borders, it gained international attention with countries such as France, United States and Russia among others intervening to resolve the issue.

The concept graph consists of approximately 17,000 entities and 6,500 relationships. We perform several runs, starting with Damascus as the center, and allowing the radius to go long enough to include the Middle East, Europe and North America. We use the semantic constraints {*protests*, *chemical weapons*, *rebels*, *civilwar*, *Ba′athParty*} for storylines of length 5. Fig. 13 shows the spatio-temporal evolution of the *Syrian Civil War*. The dataset from late 2011 shows that only a few countries in close proximity to Syria, such as Iran and Israel, were paying attention to the Syrian protests. Over time the protests grew into a full-fledged civil war with increasing global impact extending into late 2012 and up until April 2013, the time of this writing. The latest stories regarding the use of chemical weapons have been corroborated by various leading news agencies around the world.

6.4.1. Temporal propagation of Hashtags. An advantage of the concept graph approach is that it allows us to perform detailed analysis by filtering the graph for specific types of entities and relationships. In addition, maintaining spatio-temporal information on all entities and relationships enables us to take a snapshot of the graph in different locations at any point in time. In this case, we are interested in the temporal propagation of the *Twitter* hashtags related to the *Syrian Civil War*. A hashtag is a form of metadata tag, or simply put, a word or phrase prefixed with the symbol #. Hashtags are neither registered nor controlled by any one group of users, and one of their most powerful effects is to group messages, since one can search for a hashtag and obtain the set of messages that contain it. When promoted by enough individual users, they can become "trend" and attract more users to the discussion. Over time, users discussing a specific hashtag typically start relating multiple hashtags, thus relating other discussions. In the case of the *Syrian Civil War*, we notice that **#Syria**, **#Iran**, **#Assad**, and others start getting mentioned together as the civil war gains international attention. Two such snapshots are shown in Fig. 14, where the Syria hashtag suffers an explosion of co-mentions with many others from early 2012 to mid 2013. The significant difference in the number of hashtags between the two snapshots indicates the importance that the *Syrian Civil War* gained over time. We use this feature of the hashtags to evolve the concept graph which allows us to improve our storytelling capabilities.

6.4.2. Effects of concepts on generated storylines. The high volume of tweets in our datasets generate a high number of storylines that cannot be easily confirmed in an automated manner. This is

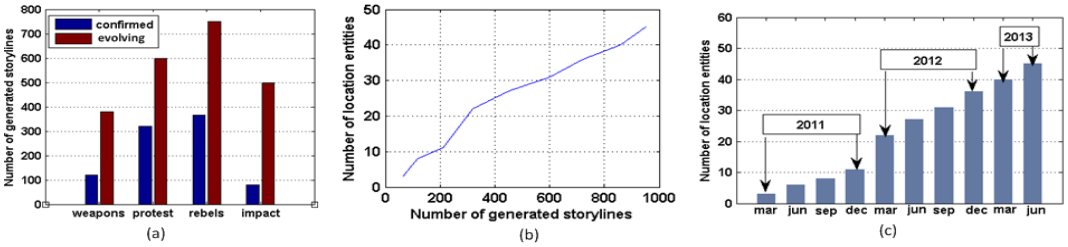

Fig. 15. Effect on storytelling (a) by concepts (b) by locations (c) by location propagation over time

illustrated in Fig. 15 (a), where the blue bars (confirmed stories) show a much lower number of instances than the red bars (evolving stories). Selection of semantic constraints also influence this result. The abstract concept *impact* seems to have a very high proportion of evolving stories. However. these stories could further evolve into real news events in the future. If we focus the concepts more on the conflicts, we get a higher proportion of confirmed stories. This is illustrated in the bars for *protest* and *rebels*. The stories around *weapons* have recently seen an upsurge due to the recent allegations of use of chemical weapons.

6.4.3. Discussion. In order to discover stories of interest, it is imperative that we limit the search space to specific locations. For example, the concepts discussed in the previous section (*impact, protest, rebels, weapons*), at first glance, are very general, until we combine them with locations specific to the *Syrian Civil War*. If not bounded by the location (or radius), the semantic constraint *protest* could generate stories for the Syrian protests, Mexico protests, or any others around the world. Hence, the importance of the spatial component of our proposed work. In the context of the *Syrian Civil War*, we fix the starting radius to the capital Damascus, increasing it gradually to find connecting stories. Fig. 15 (b) shows that the number of stories increases with an increase in the number of locations.

The experimental results have helped characterize the importance of the spatio-temporal aspects on story coherence and recall. We also find that proper understanding of the distribution of entities along with their relationships can help capture richer information content that could otherwise be missed. These extensive experiments demonstrate the potentially-high usability of our methods, which fill a gap not currently addressed in the existing storytelling literature.

7. CONCLUSION

In studying social interactions from spatio-temporal propagation, we have been able to generate dynamic real-world storylines from *Twitter* sources. Our approach establishes ranking based on different relationship types, and has proven effective on ill-formed datasets. Because we treat spatial distribution as an integral factor of our algorithms, we have been able to identify dense regions where storylines develop. Further, our approach establishes time-coherent entity connections that otherwise may have been more challenging from purely textual approaches that do not consider the myriad locations such as the ones affected by the *Syrian Civil War*. The experiments demonstrated our high potential for exploratory analysis using other current events such as the *Mexico Civil Unrests*, the *2016 Olympics*, and the *Boston Bombings*. For future work, we plan to investigate more systematic methods of grounding the true "goodness" of generated storylines, and metrics for storyline similarity. Eventually, our objective is to establish storytelling as a robust tool for entity reasoning in a wide range of application domains.

REFERENCES

Alchemy API 2013. (2013). Retrieved August 01, 2013 from http://www.alchemyapi.com/

Sergey Brin and Larry Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30 (1998), 107–117.

Jeffrey Chan, James Bailey, and Christopher Leckie. 2008. Discovering correlated spatio-temporal changes in evolving graphs. *Knowledge and Information Systems* 16 (2008), 53–96.

Jeffrey Chan, James Bailey, and Christopher Leckie. 2009. Using graph partitioning to discover regions of correlated spatio-temporal change in evolving graphs. *Intelligent Data Analysis* 13 (2009), 755–793.

Esther Galbrun and Pauli Miettinen. 2012. Siren: An interactive tool for mining and visualizing geospatial redescriptions. In *KDD'12*. 1544–1547.

Betsy George, James Kang, and Shashi Shekhar. 2009. Spatio-temporal sensor graphs (STSG): A data model for the discovery of spatio-temporal patterns. *IDA* 13 (2009), 457–475.

Gephi 2013. https://gephi.org/. (2013).

Georg Groh, Florian Straub, and Benjamin Koster. 2012. Spatio-temporal small worlds for decentralized information retrieval in social networking. In *ACM GIS'12*. 418–421.

M. Shahriar Hossain, Christopher Andrews, Naren Ramakrishnan, and Chris North. 2011. Helping Intelligence Analysts Make Connections. In *Workshop on Scalable Integration of Analytics and Visualization (AAAI '11)*. 22–31.

M. Shahriar Hossain, Patrick Butler, Naren Ramakrishnan, and Arnold Boedihardjo. 2012a. Stortytelling in Entity Networks to Support Intelligence Analysts. In *KDD'12*. 1375–1383.

M. Shahriar Hossain, Joseph Gresock, Yvette Edmonds, Richard Helm, Malcolm Potts, and Naren Ramakrishnan. 2012b. Connecting the Dots between PubMed Abstracts. *PLoS ONE* 7, 1 (2012).

IARPA 2013. Open Source Indicators Program (OSI). (2013). Retrieved September 20, 2013 from http://www.iarpa.gov/solicitations_osi.html

Jigsaw 2013. Jigsaw project. (2013). Retrieved August 15, 2013 from http://www.gvu.gatech.edu/ii/jigsaw/

Hyunmo Kang, Catherine Plaisant, Bongshin Lee, and Benjamin Bederson. 2007. NetLens: iterative exploration of content-actor network data. *Information Visualization* 6 (2007), 18–31.

Jon Kleinberg. 1998. Authoritative sources in a hyperlinked environment. In *SIAM '98*. 668–677.

Deept Kumar, Naren Ramakrishnan, Richard F. Helm, and Malcolm Potts. 2008. Algorithms for Storytelling. *IEEE TKDE* 20, 6 (2 2008), 736–751. DOI: http://dx.doi.org/10.1145/1188913.1188915

Dekang Lin. 2008. An information-theoretic definition of similarity. In *ICML '08*. 296–304.

LingPipe 2013. http://alias-i.com/lingpipe/index.html. (2013).

Massimo Marchiori. 1997. The quest for correct information on Web: Hyper search engines. In *In WWW '97*. 1225–1235.

Geraldine Del Mondo, M.A. Rodriguez, Christophe Claramunt, Loreto Bravo, and Remy Thibaud. 2013. Modeling consistency of spatio-temporal graphs. *Data and Knowledge Eng.* 84 (2013), 59–80.

Neo4j 2012. Neo4j - http://www.neo4j.org/. (2012).

Palantir 2013. (2013). Retrieved July 21, 2013 from http://www.palantir.com/

PostGIS 2013. http://postgis.net/. (2013).

Kira Radinsky and Eric Horvitz. 2013. Mining the web to predict future events. In *WSDM '13*. 255–264.

Naren Ramakrishnan, Deept Kumar, Bud Mishra, Malcolm Potts, and Richard F. Helm. 2004. Turning CARTwheels: an alternating algorithm for mining redescriptions. In *KDD '04*. 266–275.

Thomas Reed and Keith Gubbins. 1973. *Applied statistical mechanics: thermodynamic and transport properties of fluids*. Butterworth-Heinemann, Boston, Massachusetts.

Sentinel Visualizer 2013. (2013). Retrieved August 01, 2013 from http://www.fmsasg.com/

Dafna Shahaf and Carlos Guestrin. 2010. Connecting the dots between news articles. In *KDD'10*. 623–632.

Dafna Shahaf, Carlos Guestrin, and Eric Horvitz. 2012a. Metro Maps of Science. In *KDD'12*. 1122–1130.

Dafna Shahaf, Carlos Guestrin, and Eric Horvitz. 2012b. Trains of Thought: Generating Information Maps. In *WWW'12*. 899–908.

Stanford NER 2013. http://nlp.stanford.edu/software/CRF-NER.shtml. (2013).

Scott Turner. 1994. *The creative process: A computer model of storytelling and creativity*. Psychology Press, 122–123.

Twitter API 2013. https://dev.twitter.com/docs/api. (2013).

WordNet 2013. (2013). Retrieved January 28, 2013 from http://wordnet.princeton.edu/

Received Month YYYY; revised Month YYYY; accepted Month YYYY