

Modeling and Predicting Incidence:  
Critical Systems Failures and Flu Infection Cases

Xinfeng Xu

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

Master of Science

in

Computer Science and Application

B. Aditya Prakash, Chair

Bert Huang

Gang Wang

Mar 26, 2019

Blacksburg, Virginia

Keywords: Data Mining, Infrastructure Networks, Contact Network, Diffusion,  
Epidemiology, Deep Learning

Copyright 2019, Xinfeng Xu

# Modeling and Predicting Incidence: Critical Systems Failures and Flu Infection Cases

Xinfeng Xu

(ABSTRACT)

Given several related critical infrastructure (CI) networks, such as power grid, transportation, and water systems, one crucial question emerges: how to model the propagation of failed facilities and predict their spread over time to the whole system? Given digital surveillance data, can we predict the impact of Influenza-Like Illness (ILI), including the percentage of outpatient doctors visits, the season duration, and peak? These two questions are related to modeling and predicting the incidence of different types of contagions. In the case of CI, the contagions are the failures of facilities. In the case of flu spread, the contagions are the infective ILI.

In this thesis, in the case of CI, we give a novel model of failure cascades and use it to identify key facilities in an optimization-based approach, called HotSpots. In the case of flu spread, we develop a deep neural network, EpiDeep, to predict multiple key epidemiology metrics. In both of these applications, we use the dynamics of propagation to develop better approaches.

By collaborating with Oak Ridge National Laboratory (ORNL) and working on the real CI networks provided by them, we find that HotSpots helps solve what-if scenarios. By using the digital surveillance data reported by the Centers for Disease Control and Prevention (CDC), we carry on experiments and find that EpiDeep is better than non-trivial baselines and outperforms them by up to 40%. We believe the generality of our approaches, and it can be applied to other propagation-based scenarios in infrastructure and epidemiology.

# Modeling and Predicting Incidence: Critical Systems Failures and Flu Infection Cases

Xinfeng Xu

(GENERAL AUDIENCE ABSTRACT)

Critical Infrastructure Systems (CIS), including the power grid, transportation, and gas systems, are essential to national security, economy, and political stability. Moreover, they are interconnected and are vulnerable to potential failures. The previous event, like the 2012 Hurricane Sandy, showed how these interdependencies can lead to catastrophic disasters among the whole systems. Therefore, one crucial question emerges: Given several related CIS networks: how to model the propagation of failed facilities and predict their spread over time to the whole system? Similarly, in the case of seasonal influenza, it always remains a significant health issue for many people in every country. The time-series of the weighted Influenza-like Illness (wILI) data are provided to researchers by the US Center for Disease Control and Prevention (CDC), and researchers use them to predict several key epidemiological metrics. The question, in this case, is: Given the wILI time-series, can we predict the impact of Influenza-Like Illness (ILI) accurately and efficiently?

Both of these questions are related to modeling and predicting the incidence of different types of contagions. Contagions are any infective trend which can spread inside a network, including failures of facilities, and popular news. In the case of CIS, the contagions are the failures of facilities. In the case of flu spread, the contagions are the infective ILI.

In this thesis, in the case of CI, we present a novel model of failure cascades and use it to identify critical facilities in an optimization-based approach. In the case of flu spread, we develop a deep neural network to predict multiple key epidemiology metrics. In both of these applications, we use the dynamics of propagation to create better approaches.

By collaborating with ORNL and working on the real CI networks provided by them, we find

that F-CAS captures the dynamics of the interconnected CI networks. In the experiments using the wILI data from CDC, we find that EpiDeep is better than non-trivial baselines and outperforms them by up to 40%. We believe the generality of our approaches, and it can be applied to other propagation-based scenarios in infrastructure and epidemiology.

# Acknowledgments

First, I would say thanks to my advisor B. Aditya Prakash. My CS research and this thesis would not have been done without his endless support, advice, and encouragement over these years. I appreciate that he gave me the chance to join his team as an “outsider” from the Physics Department. The research I have been done here not only increased my understandings and abilities in the Computer Science field but help and improve the research in my Astrophysics Ph.D. career.

Second, I would like to thank my committees: Bert Huang and Gang Wang. I appreciate their time and efforts for giving me advice and feedback to improve my thesis.

The work presented in chapter 3 was collaborated with Liangzhe Chen, Sangkeun Lee, Sisi Duan, Alfonso G. Tarditi, Supriya Chinthavali, while the work shown in chapter 4 was collaborated with Bijaya Adhikari, and Naren Ramakrishnan. I am grateful to have the opportunity to work with them, and I learned a lot from their supreme intelligence and skills. Then I would like to thank my girlfriend Ran Tu, for all her love and support during my time in Virginia Tech.

Finally, I would like to thank my parents for their help and encouragement over the years, without which my journey through life would have been difficult.

# Contents

List of Figures	ix
List of Tables	xii
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	2
1.1.1 Thesis Statement and Structure . . . . .	3
1.2 Overview . . . . .	4
1.2.1 Modeling and Predicting Failure Cascades in Infrastructure Networks	4
1.2.2 Modeling and Predicting Influenza Spread in Human-contact Networks	6
1.3 Contributions . . . . .	7
<b>2 Literature Review</b>	<b>9</b>
2.1 Failures Cascades in CI Networks . . . . .	9
2.2 Epidemiology and Influenza Spread . . . . .	10
<b>3 Failures Cascades in CI Networks</b>	<b>13</b>
3.1 Introduction . . . . .	14
3.2 Problem Formulations . . . . .	15

3.2.1	Network Structure	15
3.2.2	Failure Cascade Model F-CAS	17
3.2.3	Problem Definitions	19
3.3	Proposed Methods	21
3.3.1	Scenarios without loop	21
3.3.2	Scenarios with loop	23
3.4	Experiments	26
3.4.1	Baselines	27
3.4.2	Effectiveness (Q1)	27
3.4.3	Scalability (Q2)	28
3.4.4	Successful Case Studies (Q3)	29
3.5	Discussions and Summary	30
3.5.1	Discussion	30
3.5.2	Summary	32
<b>4</b>	<b>Modeling and Predicting the Influenza Spread</b>	<b>33</b>
4.1	Introduction	33
4.2	Problem Formulation	35
4.3	EpiDeep Model	36
4.4	Experiments	38

4.4.1	Initial Settings	39
4.4.2	National Predictions	40
4.4.3	Regional Predictions	42
4.4.4	Analysis of Delayed Data Arrival	43
4.5	Discussions and Summary	44
<b>5</b>	<b>Conclusions and Future Work</b>	<b>46</b>
	<b>Bibliography</b>	<b>48</b>

# List of Figures

1.1	<b>Left:</b> The figure which shows the connections of the infrastructure units in combined CI networks. <b>Right:</b> The corresponding description for each infrastructure unit. . . . .	5
1.2	Predictions of the Hurricane Sandy impact paths on NY state from our modeling. <b>Left:</b> The directly affected nodes by Hurricane Sandy are labeled as red. <b>Middle:</b> F-CAS simulation shows that there are more regions affected even before the failure cascade has a loop. The red areas are the affected regions. <b>Right:</b> F-CAS final results after a loop have been formed, which indicate a much larger failure. . . . .	6
1.3	<b>Left:</b> A visual representation of the EpiDeep model. Based on the historical surveillance data, we learn meaningful representations via deep-learning and forecast multiple key epidemiology metrics. <b>Right:</b> The comparison of the performances between the EpiDeep and other baselines. X-axis are the different CDC regions in US, while the Y-axis is the Root Mean Square Error between the model predictions and the actual data from CDC. . . . .	8
3.1	The examples of the Dominator tree. All power plant nodes are merged as the supernode $\mathbf{g}$ , then we construct the corresponding Dominator tree $D$ . . .	22
3.2	The pseudo-code of our HotSpot algorithm. . . . .	23
3.3	Left: the dominator trees for problems without loop. Right: the dominator trees for problems with loop (see section 3.3.2 for details). . . . .	24

3.4	Left: The Update function which can be used in scenarios with and without loops. Right: the $\text{Recur}^+$ scenarios which is used in the scenarios with loops. (see section 3.3.2 for details).	25
3.5	The number of nodes in each component of the CIS network for the four states used in the experiments.	26
3.6	The comparison of the performance of HotSpots to all other baselines in the OH state area. Panels (a) and (b) are for the <b>Tran-naive</b> scenario, while the panels (c) and (d) for the the <b>Trans-real</b> scenario. HotSpots outperform all baselines.	28
3.7	The time scale of HotSpots and how it scales with $k$ and $ V $ (see section 3.4.3 for details).	28
3.8	The case study of Hurricane Sandy’s impact paths. (a) the total area of directly affected by the Hurricane Sandy predicted by HotSpots when we overaly the constructed CIS network $G$ with the hurricane’s path. (b) Results from F-CAs when the failure cascades has not form a loop yet. (c) F-CAS simulation results after a loop has formed, which affects much larger areas. We mark the affected regions by the red boxes (the actual network is not shown due to security reasons, see section 3.4.4 for details).	30
3.9	The case study of the US NE 2003 blackout. We use the HotSpot algorithm to find the top five transmissions in the OH area, which is highlighted by the red circles.	31

4.1	(a) The architecture of the proposed EpiDeep model, including the clustering, encoder, and decoder module. (b) The structure of the encoder module, where the attention layer is added to the end. . . . .	37
4.2	The RMSE and MAPE performances for regional predictions of all three tasks for the 2016/17 season. EpiDeep consistently perform well comparing to other baseline. . . . .	42
4.3	EpiDeep’s performance on RMSE and MAPE for Future Incidence Predictions with simulated delayed data. The performance remains stable and has a trend to perform better when the time delay is shorter. . . . .	43

# List of Tables

4.1	The comparisons of the performance for all methods in forecasting the three tasks for epidemical seasons from 2010/11 to 2016/17. R, M, and LS are for RMSE, MAPE, and the average Log Score, respectively. A “–” sign represents that the method is not applicable in this prediction. <b>EpiDeep</b> consistently performs well in all forecasting tasks and outperforms all the baselines in the majority of the scenarios. . . . .	41
-----	--	----

# List of Abbreviations

CDC: Centers for Disease Control and Prevention

CIS: Critical Infrastructure Systems

ILI: Influenza-like Illness

ORNL: Oak Ridge National Laboratory

# Chapter 1

## Introduction

This thesis describes how we model and predict the propagation of contagions' incidences, including failure cascades in Critical Infrastructure (CI) networks and influenza cascades in Epidemiology related networks. In the case of CI, by building models to simulate the time-dependent cascades propagation, we are able to forecast the effect of the failure cascades (i.e., one or more key facilities break down as unexpected conditions) to the whole combined CI networks. In the case of flu spread, we build a deep-learning based model to handle the time-dependent Influenza-Like Illness (ILI) data, then predict the key epidemiological measurements, such as the percentage of outpatient doctors visits, the season onset, duration, infection peak time and peak values.

Both of our two topics are important as: (1). CI systems, such as transportation, water, and power grids are the fundamental basis for nation and individual security, public life, and economy. The failure cascades on these networks can cause a vast influence on people and the whole society. CI systems are usually a combination of multiple networks and are dynamically linked to each other. However, there lacks analysis on how to model and predict the failure cascades in such networks. (2). Recurrent outbreaks of ILI, such as flu, raise the needs to predict the spread of these diseases based on the historical ILI data. Introducing a deep-learning based method for modeling and predicting the influenza spread is helpful for people and the country to react appropriately to these ILI.

In this chapter, we describe the motivations of the thesis in Section 1.1. Then we show an

overview of the two topics in Section 1.2 and 1.3. Finally, in Section 1.4, we discuss our works' contributions and applications.

## 1.1 Motivation

Propagation of contagions is a common concept in daily life and happens in different areas such as infrastructure, epidemiology, biology, and social media. In critical infrastructure (CI) networks, one of the critical contagions types is failure cascade. The failure of broken pipes of a water network, failed instruments in the communication network and power outage of power-plants in energy networks will propagate along the infrastructure networks and affect a city, an area and even a vast part of the country. In epidemiology, one of the common cascade types is Influenza-Like Illness (ILI) spreading in epidemiology related networks, including human-contact networks and human-mobility networks. The recurrent outbreaks of ILI periodically affect the whole human population and remain as a serious worldwide threat to public health. There exist various analyses of diffusion propagation in various networks. However, the problems raised here are new and never been analyzed by our methods before. CI networks are usually mutually dependent on each other such that the analysis complexity increases dramatically. For example, the communication network relies on the inputs from the energy network and supplies of the water network. Therefore, the failures in one of the CI network will affect not only itself but also other related CI networks. The recent events, such as the 2012 hurricane Sandy, show how failure cascades can cause propagation in related CI networks and cause catastrophe to the whole combined system. However, traditional methods either only analyze the failure cascades in separate CI networks or did not model the dynamics of the combined system. Thus, a thorough model of the failure cascades in such networks is needed.

The Influenza-like-illness propagates in the human-contact network, but recent studies show that there is a close relationship between the influenza spread and the human mobility networks, such as long-distance airlines networks and short-distance commuting networks. There are various analyses to model the influenza spread. For the ILI problem, current methods work well with the solid geometrical and human-mobility network data, but less work has been done if lacking these data. Our proposed models handle this influenza spread problem by building models based on deep-learning neural networks, which is a different and "new" view.

### 1.1.1 Thesis Statement and Structure

**Taking dynamics of propagation into account, namely path-based failures for critical infrastructures, and seasonal inconsistency for flu spread, we can develop better incidence models and make more accurate predictions.**

One of the challenges in the case of CI is that failure does not always propagate locally from a single node to its neighborhood. For example, a substation would fail if there is no path from a valid power generator to this substation. We call this kind of failure as "path-based" failures, i.e., we need to check all the paths in the transmission network then decide if there are valid paths, which is a costly operation. Our proposed approach is able to take "path-based" failures into account and solve the failure propagation problem in the CI networks.

The primary challenge in the case of flu spread is the highly dynamic nature of influenza. The spread of flu is dependent on various conditions, e.g., weather patterns, dominant virus type, antibody levels, etc. Therefore, the similarities between different seasons are evolving, and simple algorithms can't capture this dynamic nature. Consequently, we built an approach that can explicitly leverage the evolving similarities between previous seasons and yet able

to make accurate predictions.

Following our thesis statement, we organize the thesis in two parts. In the first part of the thesis, we model the failure cascades in combined CI networks and analyze the failure propagations among the whole system. We predict the spreading of cascades and study the key failure maximization problems. In the second part of the thesis, we analyze the time-dependent ILI data series and use the deep-learning neural networks to model them. Then we extract the key epidemiology measurements. In all, we gain more understanding of the dynamic nature of contagions.

## 1.2 Overview

The structure of the thesis is as follows: First, we discuss our failure cascades model in chapter 3 to solve combined heterogenous CI networks. Then in chapter 4, we discuss how we use the deep-learning neural networks to model and predict the influenza spread. We show a brief overview of these two chapters in the following sections.

### 1.2.1 Modeling and Predicting Failure Cascades in Infrastructure Networks

In this section, we briefly describe our modeling of failure cascades in CI networks and several related applications. The work in this chapter has already been published in the ACM International Conference on Information and Knowledge Management (CIKM, 2017) [21].

The failure cascades happen in the combined heterogenous CI networks, and the structure is shown in the left panel of Figure 1.1, which has five connected infrastructure units, including

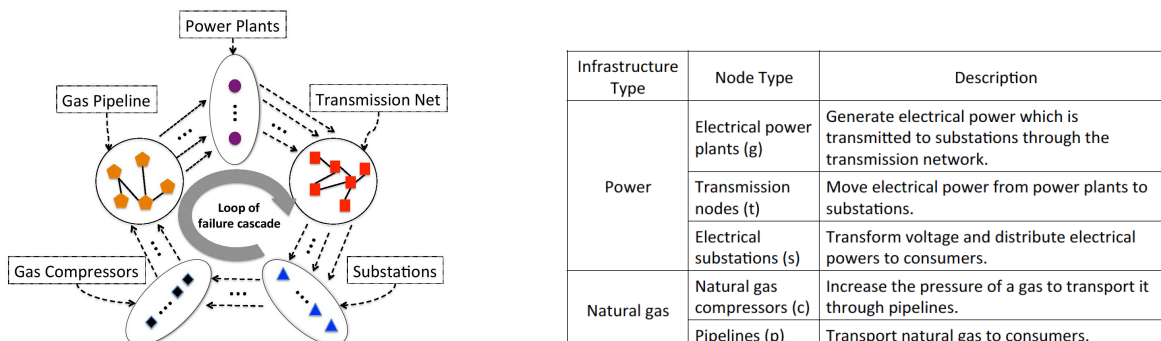


Figure 1.1: **Left:** The figure which shows the connections of the infrastructure units in combined CI networks. **Right:** The corresponding description for each infrastructure unit.

Gas Compressor, Gas Pipelines, Power Plants, Transmission Network, and Substations. The detailed descriptions are in the right panel of Figure 1.1.

The CI network systems we introduced above are vulnerable to potential failure cascades, since local failures may spread over the whole system and get amplified then lead to more extensive catastrophes. For example, nodes failing in the transmission network will cause the re-routing of the electrical power and may cause overload to other transmission lines, which increases the probability of failures in additional nodes. With those failure cascades spread and loop CI networks, further failures may happen and eventually affect the whole system.

In this work, we propose the F-CAS model which involves the modeling of each CI and the connections between them. First, we define the failure conditions for each CI components. For example, the substations fail when they do not have an existing path from the transmission network to an active power generator, due to the lack of electricity. Second, we propose two Independent Cascade (IC) style models, Trans-naive and Trans-real to handle the failures' influence on other parts of the system.

We applied our model to several applications and evaluated the effectiveness of it. (1)

We model the CI systems of NY state and estimate the impact of hurricane sandy. We show in Figure 1.2 about our successful simulation results. (2) We proposed a scalable and effective algorithm HotSpots to study the failure of maximization problems. We run multiple experiments based on the CIS data for different US states. The results show that our proposed method, HotSpots, are good than other non-trivial baselines in most of the cases.



Figure 1.2: Predictions of the Hurricane Sandy impact paths on NY state from our modeling. **Left:** The directly affected nodes by Hurricane Sandy are labeled as red. **Middle:** F-CAS simulation shows that there are more regions affected even before the failure cascade has a loop. The red areas are the affected regions. **Right:** F-CAS final results after a loop have been formed, which indicate a much larger failure.

## 1.2.2 Modeling and Predicting Influenza Spread in Human-contact Networks

In this section, we briefly describe the modeling and predictions of the influenza spread in national-wide human-contact networks. This work has been submitted and is under review.

Influenza remains a serious worldwide threat to public health. Centers for Disease Control and Prevention release the digital surveillance data each year and encourage data scientists to analyze and build models to predict the influenza spread. The overall architecture of our EpiDeep model is showing in the left panel of Figure 1.3.

We build our model, EpiDeep, based on a deep-learning neural network. We are given input

incidence as time-series (training data set), and the initial state of a new season ('query'), then our tasks are to train the deep-learning neural network to predict various metrics for the rest of the season. Our neural networks combine the Long Short Term Memory (LSTM) and Feed-Forward architectures to first learn lower-dimensional representations of all the temporal epidemic curves in the training dataset (including historical ILI curves, digital indicator curve, etc.), as well as the query. Based on the property of LSTMs and our adding of "Attention Mechanism," the model can automatically learn to forget/remember historical information selectively, even with sparse data.

Next, we design the neural networks to use the learn embeddings together with the query to predict various key epidemiology metrics. We then compare the EpiDeep model with other baselines about the performances. We show an example in the right panel of Figure 1.3), where we beat all other baselines in the future incidence predictions in three CDC regions. Overall, EpiDeep outperformer the non-trivial baseline: Empirical Bayes (EB), in three out of four metrics, including peak intensity, onset, and future incidence prediction by 16%, 14% and 40% on average.

## 1.3 Contributions

- In Chapter 3, Modeling and Predicting the Failure Cascades in CI Networks, the main contributions are:
  - We are the first to construct heterogeneous networks from real CI data provided by ORNL, and we describe the failures conditions in interconnected CI networks by taking advice from their domain experts.
  - We take the dynamic nature of path-based failures into account in our model and develop a new cascade model, F-CAS, which is tractable.

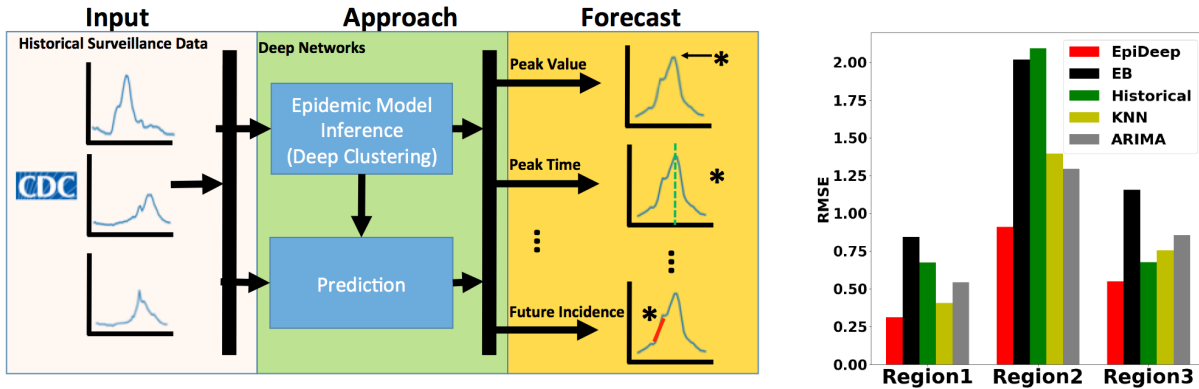


Figure 1.3: **Left:** A visual representation of the EpiDeep model. Based on the historical surveillance data, we learn meaningful representations via deep-learning and forecast multiple key epidemiology metrics. **Right:** The comparison of the performances between the EpiDeep and other baselines. X-axis are the different CDC regions in US, while the Y-axis is the Root Mean Square Error between the model predictions and the actual data from CDC.

- We develop an efficient algorithm, HotSpots, to discover the most important nodes in CI, whose failure may lead to the largest damage to the whole system. Thus, these are the key facilities which need to be “protected”.
- In Chapter 4, Modeling and Predicting the influenza spread in Human-contact Networks, the main contributions are:
  - We are the first to predict the influenza spread using ILI data by a deep-learning based neural network.
  - We are able to capture the evolving similarities between different flu seasons and embed them to make accurate forecasts.
  - We demonstrate the effectiveness of our model, EpiDeep, which outperforms other non-trivial baselines in most of the key epidemiology metrics.
  - These approaches have been used by the VT team to participate in the CDC flu challenge (2018 – 2019 flu season).

# Chapter 2

## Literature Review

Here we discuss the related works and give reviews of previous studies. First, we introduce the earlier studies in the case of CI networks in section 2.1. Then we discuss the previous researches in the case of flu spread in section 2.2.

### 2.1 Failures Cascades in CI Networks

Modern critical infrastructures (CIs) such as Power, Transportation, Communication etc are mutually dependent in non-trivial ways. These dependencies exist among multiple CIS network, therefore, failures from one CI network could potentially cascade to other related networks and cause wider influence than expected. Therefore, it is critical to study the failure cascades in CI networks. We discuss previous researches here in detail.

**Infrastructure Vulnerability Analysis:** Previous studies on vulnerability analysis and interdependency simulations include different approaches: e.g., empirical approach, agent-based approach, and system dynamics based approach (see [36] for a review). Moreover, there are a few methods based on mathematical frameworks [16]. However, most of these works focus on only two CI networks, where their interconnections are much easier than our frame. For example, the connections between a power grid network and a communication network has been studied in [37]; the interconnections between the power network and the water network are explored in [24]. Our framework is more general, where we focus on the

system including five connected CI components.

**Influence Maximization and Cascade Analysis:** Finding the best seed nodes which maximize influence is the target of the influence maximization problem. There are various studies on the Linear Threshold models and the standard Independent Cascade (IC) models[30]. Both of the models can reach a 63% approximation due to the submodularity of the objective functions. Other works have focused on developing more effective algorithms for the original problem [12], or trying to create continuous time models [23], and others with uncertainties [22]. However, almost all of these proposed methods assume that the failure cascades only locally among the network, which is not always the case in our problems. We have the path-based failures which are non-local cascade types, where the failure can travel long path to affect not just neighbors. Recently, Opera [18] finds that important nodes in CIS networks could maximally decrease the connectivity of the networks. However, their models are still “local”, which can not capture the dynamic nature of interconnected CIS networks.

## 2.2 Epidemiology and Influenza Spread

There are various former studies on Epidemiology. The fundamental epidemiological models divided the population into different types, including Susceptible, Infected and Recovered. Based on these, SIS [8, 43], and SIR [7] models are proposed, where the transition rates include infection rate, and recovery rate exist. However, none of these former studies considered the underlying network structure. [38] is the first to study the incidence problem of computer viruses in networks. There are also models that use epidemiological models for social networks include [27, 28].

Influenza which happens seasonally is a significant health issue that affects many people in

the world. The US National Center for Disease Control and Prevention (CDC) reports that there were more than 30,000 influenza-related hospitalizations in the last influenza season in the US alone [10]. Furthermore, the number of deaths results from the ILI are high in the past five seasons. These statistics reveal that it is essential to keep track of the ILI and make earlier decisions so that we can decrease the ILI incidence. Therefore there are various previous studies on this topic, and we discuss them in details here.

**Influenza Forecasting:** There have been various researches in developing methods for influenza forecasting recently. At least two different directions exist: (1) Statistical methods, e.g., [17, 20], fit a predefined statistical model on historical ILI data and use them as the basis for forecasting. These methods typically are too simple and hard to incorporate with domain knowledge like the epidemiological dynamics or require laborious feature engineering. Moreover, these statistical models often are "Black-box" and hard to provide any interpretation of the forecasts. This is critical since the interpretation can guide the researchers and provide the abilities to fine tune the models and make better predictions. (2) Mechanistic models, e.g., [40, 42, 48] are motivated by domain knowledge, and can they usually include various models based on human mobility. These models are good at providing interpretable forecasts. However, they require frequent calibration and are often too rigid to generalize well to make new predictions.

**Time Series Forecasting:** Time-series predictions is a mature area where several methods from different perspectives [13, 41]. Recurrent neural networks [26, 33] are usually used in these works. However, these forecasting methods are not well-suited for the flu spread case since they are too specialized to capture the seasonal fluctuation or inconsistency in the wILI data. In contrast, we develop EpiDeep which can automatically embed, cluster and predict while keeping the flexibility to capture the seasonal inconsistency in the wILI data.

**Deep Learning in Epidemic Forecasting:** The cases of combining deep learning and

epidemic forecasting are rare but still exist. In [46], the authors used LSTM with geographical and climate constraints to make predictions. In [47], authors leverage LSTM to predict the activities of influenza by combining with Twitter data. However, data sparsity is also a significant issue in flu spread forecasts, and LSTM typically requires a large amount of data. Therefore, it does not fit straightforwardly here. Moreover, these models don't capture the dynamic nature of the flu spread, which we discuss it in detail in section 4.

# Chapter 3

## Failures Cascades in CI Networks

Modern critical infrastructures (CIs) such as Power, Transportation, Communication, etc are mutually dependent in non-trivial ways. These dependencies exist among multiple CIs network. Therefore, failures from one CI network could potentially cascade to other related networks and cause more extensive influence than expected. Moreover, these CIs are essential to the economy, public safety, and political stability. There are recent events, e.g., 2012 hurricane Sandy, showing how vital the interdependencies between different CI networks are. Moreover, it shows how moderate nature disasters could be signified in CI networks and cause more massive catastrophic failures to wider areas. Therefore, it is essential to analyze these CI networks and model the failure incidence among them.

However, traditional methods either too simple to analyze this task or ignored the dynamics of the whole dependent CI networks. Therefore, we study this path-based problem thoroughly and build the models to identify key facilities in CI networks. We first construct the CI networks based on the real data provided by ORNL. Then we study novel failure maximization problems to predict the spread of failure incidence. After that, we propose **Hotspots**, a scalable and efficient algorithm for identifying key facilities in CI networks. Finally, we build multiple experiments based on the actual CIs data for multiple US states and show that our method, **Hotspots**, has better performances than other non-trivial baselines. Moreover, it can give meaningful representations of the problems and can be used to provide situational-awareness during actual failures or disasters.

## 3.1 Introduction

Critical Infrastructure Systems (CIS) including Transportation, Gas, Communication and Energy systems are mutually dependent on each other in complex ways. For example, the Communication system depends on the electricity input from the Energy systems, while the Energy systems depend on the Water network for dissemination and disposition [45]. Furthermore, such dependencies can exist through multiple CIS. Therefore, the failures in one of the CIS networks can possibly propagate to the other networks and cause a more significant influence on multiple CIS.

One of the perfect examples of the failure's propagation among multiple CIS is the 2003 Northeastern US blackout [5]. There was a failure in 3 transmission lines, which caused a massive blackout impacting a wide area in multiple CIS. Initially, the power outage only caused a blackout in several U.S. states. However, this power outage propagated to other CIS, such as water/drinking, communication and transportation, and then led to a national-wide failure. In this event, nearly 50 million people were affected, and more than \$5 billion are lost. It has been reported that a significant blackout occurs every 4 months in the US, which can affect one million people or more [36]. This example shows that the vulnerability of a single CIS can propagate and be amplified among multiple CIS systems due to their complex interdependencies. Similarly, Hurricane Sandy is another example, which shows the propagation of failures to various CIS and slowed down the recovery of the area [14]. In all, how to model and simulate the interdependencies of CIS is a critical question, and this help realizes the idea of 'smart cities and nations'.

Since 2003, consistently efforts have been put into building the situational awareness related to vast areas and developing sophisticated decision support methods to help predict the propagation of impacts caused by extreme events like hurricane, earthquake, etc. [2]. For

example, the Oak Ridge National Lab (ORNL) developed detailed simulations to help predict how hurricane affect the power grid network and forecast the failures even before a hurricane hits the land [9]. This kind of situational awareness can assist the distributions panning of the available resources, and help protect critical infrastructures by unique constructions.

Most importantly, the critical and vulnerable nodes and links in the connected CIS networks should draw the most attention. Their failures will cause maximum damage to the whole system and can cause catastrophic disasters to national security, economy, and public safety [32, 39]. Some former researchers made efforts to this end. E.g., identifying high degree nodes that are more important than others [6], analyzing the vulnerability points of power systems [25], and modeling failures in infrastructure networks from data-mining view [18, 19]. However, none of these methods both work in connected CIS and take the dynamics of the failure propagation into considerations.

Here we introduce our approach to model failure propagation in connected CIS networks. By collaborating with ORNL, we first construct a heterogeneous system with 5 CI components. We then model the failure propagation among this heterogeneous network and study several problems, including finding the  $k$  most important nodes (the ‘hot spots’). Our model is novel and efficient comparing to non-trivial baselines and can predict the non-linear interaction among the different CIS.

## 3.2 Problem Formulations

### 3.2.1 Network Structure

By collaborating with ORNL and taking advises from the domain expert, we build our interested CIS networks by five essential components from the HSIP Gold data [1] and EIA

data [3] from the natural gas and power systems. The details of these five components are shown in figure 1.1. For the transmission network and the pipeline network, the nodes are interconnected through actual transmission lines/pipelines. However, the power plants, substations, and natural gas compressors are not interconnected among themselves. Overall, this network as the power plants consumes the natural gas and produces electric power, after that, the transmission network distributes the electricity to substations, then the substations deliver electricity to natural gas compressors, and at last gas compressors provide the natural gas to power plants through the pipelines. Clearly, here is a loop of resources (natural gas/electricity) within the network. Note that there exist different power plants that consume different types of fuels. But we only consider the ones which consume the natural gas for clarity sake.

The interlinks between different components are also important and we describe them in details here:

- **Substations** are connected to the closest transmission node to get the electricity. Moreover, each substation is also providing power to local facilities including the closest natural gas compressors. We assume that the service areas of substations do not overlap. Therefore, only one substation is connected to each natural gas compressor.
- **Power plants** get fuel from the nearest gas pipeline node. Moreover, it is connected to the closest transmission node to output the produced electricity.
- **Natural gas compressors** are connected to the closest pipeline in order to provide pressure to move the natural gas to power plants. It is also connected to the closest substations to get the electricity.

The whole network structure is summarized as the left panel of figure 1.1. The final constructed heterogeneous graph can be described as  $G(V, E)$ , where  $V = \{V_g, V_t, V_s, V_c, V_p\}$  contains the five CI components mentioned above, and  $E = \{E_t, E_p, E_{inter}\}$  represents the

edges appeared in transmission network, pipeline network, and the interlinks between different components, respectively. Note that these edges have directions, which are either provided from the data we have or from the feed-supply connections as we describe above.

### 3.2.2 Failure Cascade Model F-CAS

The failures that happened in one component of the CIS network can get amplified through failure cascade to other parts. Therefore, the CIS network is vulnerable to potential attacks or failures. For example, the failure of natural gas compressors may introduce the inabilities of some power plants to work and then cause wide-spread power outage. The failures of some transmission nodes can also increase the power flow and may introduce overload to nearby nodes or transmission lines.

This kind of failure cascade is complex to model due to the interconnections between various components. Previous studies are using high-performance supercomputers utilizing sophisticated paralleled algorithms. However, they still need a long time ( $\sim 10$  days) to get accurate simulations even for only the power system [1]. Therefore, simplifications are needed here. We propose, F-CAS, which is a failure cascade model and aims at capturing the most important dynamics and estimates the failure cascades then extracts the critical nodes. We start with the failures in transmission networks. Then the failures would propagate to other components according to their connections. For each component, we define their failure conditions in our CIS network here:

- **Substations** would fail when they have no usable connections through the transmission network to any active power plants, due to the lack of electricity.
- **Natural gas compressors** would fail when the connected substation fails, due to the shortcoming of electricity.

- **Power plants** would fail when none of the active natural gas compressors are connected to it, due to the inadequate of gas or fuel.
- **Pipelines** would fail only when they are damaged by natural disasters. We remove the bad pipelines from the network before the analysis.
- **Transmission node** would fail if it has an overload. It is due to the re-routing of electrical current inside the transmission networks. We proposed two models to describe the scenario of overloadings here:
  - *Tran-naive*: When a certain node in the transmission network fails, its children in the network would need to get the electricity from other nodes, which we call them as co-parents. The failed node increases the output of its co-parents and therefore, increases the possibility of overloading of its co-parents. Based on this assumption, we first identify the parents-children relations as well as the co-parents-children relations for the whole network. Once a transmission node fails, we assign a certain increase of possibility of failure for its co-parents. This possibility can be described as:

$$e_{ij} = \begin{cases} c & \text{if } t_i \text{ and } t_j \text{ share a child} \\ 0 & \text{otherwise} \end{cases} \quad (3.1)$$

where  $c$  is a constant possibility value/weight. Therefore, the failure cascade propagating in the transmission network is similar to the Independent Cascade (IC) model on the co-parent transmission network.

- *Tran-real*: When a node fails, we assign a constant possibility/weight to the co-parents above. However, this is not realistic. To judge if the co-parent actual fails, we will need to compare its loading to its capacity. Therefore, the possibility of failing of node  $t_j$  if

its parent node  $t_i$  failed is:

$$e_{ij} = \frac{\sum_{x \in Cs[Par(t_j)-t_i]} Load(x)}{\sum_{x \in [Par(t_j)-t_i]} Capacity(x)} \quad (3.2)$$

where  $Par(t_j)$  are the parents of node  $t_j$ ,  $Cs[Par(\cdot)]$  are the combination of the child nodes from all the parent nodes found in the set  $Par(\cdot)$ . We look at all parents of  $t_j$  (ignore itself:  $t_i$ ) and calculate the ratio between the total load consumed by their children and the capacities of these parents. When this ratio is close to 1, the possibility of failures of the parents is high, and they would not be able to provide enough power to  $t_j$ . Therefore,  $t_j$  fails.

**Novelty:** Even though simplified, we still catch important dynamics in each of the components in CIS networks, which cannot be captured by the former cascade-style based model in the literature [22, 30]. For example, in the situation of path-based failures in the substations. In standard failure propagation models including IC, the failures would propagate from one failed node to another through the connecting edges. However, in F-CAS, a substation could fail due to a path-based/ non-close failure of other transmission nodes. Our model does catch the most important dynamics, and interconnections happened in the whole CIS network.

### 3.2.3 Problem Definitions

As mentioned above, in order to improve the network's reliability and efficiency, it is essential to find the critical nodes whose failure can cause maximum damage. We are interested in this problem here and define the following failure maximization problems based on this idea.

#### PROBLEM 1 (Max-Sub)

**Given** the constructed heterogeneous network, say  $G$ , the developed failure cascade model, F-CAS, and a number  $k$

**Find** the best initial set to fail, say  $S^*$ , which has  $k$  transmission nodes, such that the expected number of the failed substation are maximized after the network is stable, i.e.,

$$S^* = \underset{S}{\operatorname{arg\,max}} \mathbb{E}[\#s|S] \quad (3.3)$$

Where  $\#s$  is the final number of failed substations after the propagation of failures started from the  $k$  transmission failures. We focus on the failure of substations since this is the point which directly related to consumers in the region. By adding the transmission nodes into consideration, we extend the **Max-Sub** to **Max-SubBus** as follows.

#### **PROBLEM 2 (Max-SubBus)**

**Given** the constructed heterogeneous network, say  $G$ , the developed failure cascade model, F-CAS, and a number  $k$

**Find** the best initial set to fail, say  $S^*$ , which has  $k$  transmission nodes, such that the expected number of failed substation + **failed nodes in the transmission bus network** are maximized after the network is stable, i.e.,

$$S^* = \underset{S}{\operatorname{arg\,max}} \mathbb{E}[\#s + \#t|S] \quad (3.4)$$

Similarly,  $\#t$  is the final number of failed transmission nodes after the propagation of failures started from the  $k$  transmission failures.

### 3.3 Proposed Methods

It is challenging to solve **Max-Sub** and **Max-SubBus** since several reasons. 1) The failures not only cascade locally from one node to its neighbor but also affect in longer distances as described above. 2) The failures would propagate in the CIS network until convergence, which makes it harder to solve. We have proved that both **Max-Sub** and **Max-SubBus** are NP-hard problems, and they are more general than the NP-hard problem mentioned in [30], which is the influence maximization for IC.

Therefore, we first try to simplify the original problem, where the failures do not propagate into a loop. In this case, we update the network  $G$  into  $G'$ , where only three components are needed: power plants, transmission networks, and substations. The failures in substations would not introduce more failures in the power plants to cause a loop. After this simplified scenario, we introduce the complex problem where the failure cascade does form a loop.

#### 3.3.1 Scenarios without loop

Here we consider  $G'$  which has 3 instead of 5 CI networks, and there are directional cascades from power plants to the substations. At the first step, we merge all the power plants together as a supernode  $\mathbf{g}$  [see sub-panel (a), (b) in figure 3.1]. This super node receives all the connections from the merged power plant nodes. The failure conditions for substations change accordingly to be: a substation only fails when itself does not have a valid path to  $\mathbf{g}$ . After that, we construct a dominator tree rooted at this supernode  $\mathbf{g}$  [see sub-panel (c) in figure 3.1 [15]].

After we get the dominator tree structure of the network, each substations are then connected to the  $\mathbf{g}$  by a path  $P$ . Therefore, the only scenario for a substation to work is that all nodes

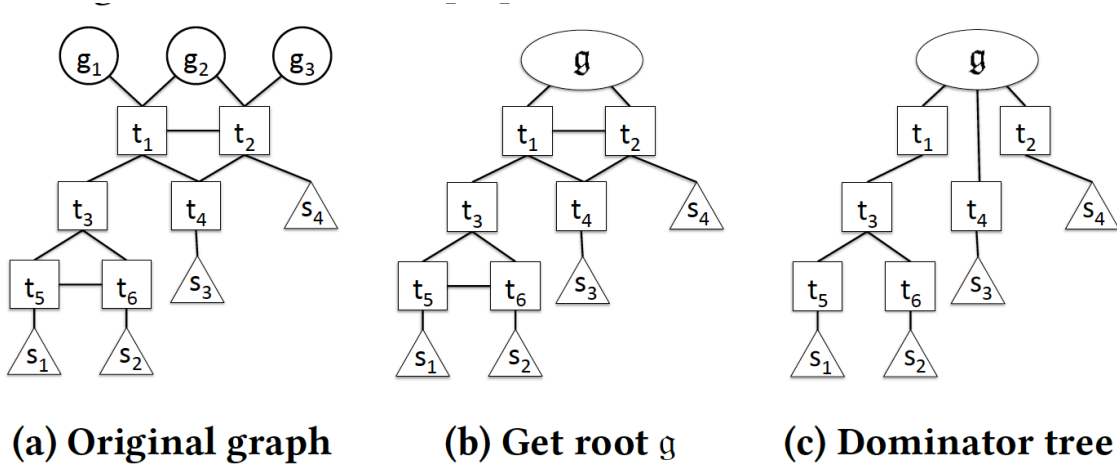


Figure 3.1: The examples of the Dominator tree. All power plant nodes are merged as the supernode  $g$ , then we construct the corresponding Dominator tree  $D$ .

in this paths are still working, e.g., the possibility for a substation node  $s_i$  to work inside the dominator tree is:

$$Pr(s_i|S) = 1 - \prod_{t_j \in P_i} (1 - Pr(t_j|S)) \quad (3.5)$$

where  $t_j$  is the transmission node between  $g$  and  $s_i$ . We assume the independence of each  $Pr(t_i|S)$  in this case. Therefore, the result is only a lower limit estimation to the actual  $Pr(t_i|S)$ . Under this assumption, the total number of failures in substations in the case of **Max-Sub**, **Max-SubBus** without loop are:

$$\mathbb{E}[\#s|S]' = \sum_{s_i} Pr(s_i|S) = T - \sum_{s_i} \prod_{t_j \in P_i} (1 - Pr(t_j|S)) \quad (3.6)$$

$$\mathbb{E}[\#s + \#t|S]' = \sum_{s_i} Pr(s_i|S) + \sum_{t_i} Pr(t_i|S) = T - \sum_{s_i} \prod_{t_j \in P_i} (1 - Pr(t_j|S)) + \sum_{t_i} Pr(t_i|S) \quad (3.7)$$

We find that both  $\mathbb{E}[\#s|S]'$  and  $\mathbb{E}[\#s + \#t|S]'$  have nice properties which are sub-modular as well as monotonically non-decreasing in terms of  $S$ . These are important properties which

we can use to solve them here:

---

**Algorithm 2** HOTSPOTS framework

---

**Input:**  $G'$ , F-CAS,  $k$ ,  $m$

**Output:** A set  $S$  of  $k$  nodes

- 1:  $S = \{\}$
  - 2: Merge all power plants into  $g$ , and construct a dominator tree  $D$  rooted at  $g$  for  $G'$
  - 3: **while**  $|S| < k$  **do**
  - 4:   **for** each  $t_i \notin S$  **do**
  - 5:     Estimate  $\Pr(t|S)$ ,  $\Pr(t|S \cup t_i)$  using F-CAS simulation.
  - 6:      $\delta(t_i) = \text{Update}(\Pr(t|S \cup t_i), \Pr(t|S), D, g)$
  - 7:      $t^* = \arg \max \delta(t_i)$ , add  $t^*$  to  $S$
  - 8: **Return**  $S$
- 

Figure 3.2: The pseudo-code of our HotSpot algorithm.

**HotSpot Framework:** We present the pseudo-code our of HotSpot framework in figure 3.2. Since both  $\mathbb{E}[\#s|S]'$  and  $\mathbb{E}[\#s + \#t|S]'$ , this let us to build an algorithm in time scale of  $(1 - 1/e)$  for optimizing them [34]. HotSpot adopts a greedy algorithm to iteratively select nodes which at each step the node maximize the marginal gain of the objective function, above.

### 3.3.2 Scenarios with loop

Now we move to the full CIS network  $G$  with all the 5 components and the failure now can propagate from substation network back to pthe ower network and finally to substation network again to form a loop. We first calculate the failure possibility of each power plant  $g_i$  when some of the substations ( $S$ ) fail:

$$Pr(g_i|S) = 1 - \prod_{c_j \in A_i} (1 - Pr(Sub(c_j)|S)) \quad (3.8)$$

where  $A_i$  is the natural gas compressor network nodes, which provide the gas to power plant  $g_i$ , and  $Sub(c_j)$  represents the the substations which are connected to  $c_j$ . Basically, similarly to equation 3.5. We find all the natural gas compressors which are linked to this power plant, and all substations which are connected to these natural gas compressors. The failing possibility of  $g_i$  is therefore accordingly determined.

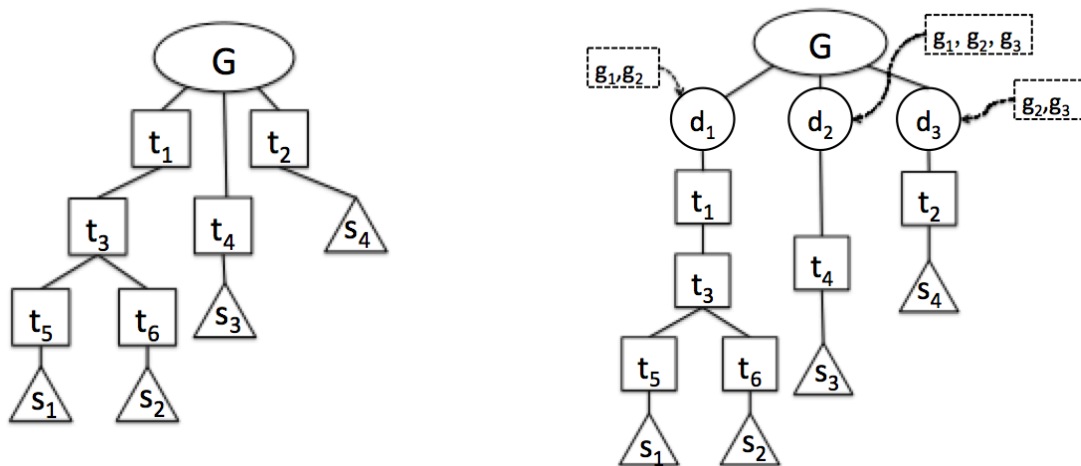


Figure 3.3: Left: the dominator trees for problems without loop. Right: the dominator trees for problems with loop (see section 3.3.2 for details).

Then we update the constructed dominator tree mentioned above according to the failure propagated from the power plants. Before we take all power plants into one super node  $g$  (see the left panel of figure 3.3). However, this is not able to describe the case when each power plant has different links to the substations. Therefore, we create a dummy node  $D$  for each set of power plants which generate electricity to this tree branch (see the right panel of figure 3.3). The failing possibility of one dummy node  $d_i$  is then as

$$Pr(d_i|S) = 1 - \prod_{g_j \in B_i} (1 - Pr(g_j|S)) \quad (3.9)$$

where  $B_j$  represents all the power plants which provide power to the the transmission nodes,  $t_i$ , in this path.

Finally we can rewrite the equations 3.6 and 3.7 as:

$$\mathbb{E}[\#s|S] = T - \sum_{s_i} \prod_{t_j \in P_i} (1 - Pr(t_j|S)) \cdot \prod_{g_j \in B_i} \prod_{c_y \in A_i} (1 - \hat{P}r(Sub(c_y)|S)) \quad (3.10)$$

$$\mathbb{E}[\#s + \#t|S] = T - \sum_{s_i} \prod_{t_j \in P_i} (1 - Pr(t_j|S)) \cdot \prod_{g_j \in B_i} \prod_{c_y \in A_i} (1 - \hat{P}r(Sub(c_y)|S)) + \sum_{t_i} Pr(t_i|S) \quad (3.11)$$

Moreover, we prove that the above objective functions 3.10 and 3.11 are also sub-modular as well as monotonically non-decreasing in  $S$  [2]. Therefore, the full problems of **Max-Sub** and (Max-SubMax) can be solved by adopting the same  $(1-1/e)$  approximations mentioned in the algorithm in figure 3.2. We add the Update and Recur algorithm to handle the new dominator tree (see algorithms in figure 3.4).

---

**Algorithm 3** *Update*( $\cdot$ )

---

**Input:**  $Pr(t|S \cup t^*), Pr(t|S), D/D^+, g$

**Output:**  $\delta(t^*)$

//For the without-loop version  
1: Return  $Recur(Pr(t|S \cup t^*), D, g) - Recur(Pr(t|S), D, g)$   
//For the with-loop version  
2: Initialize all  $Pr(s|S \cup t^*), Pr(s|S), Pr(d|S \cup t^*), Pr(d|S)$  as 0  
3: **while**  $Pr(s|S \cup t^*), Pr(s|S)$  is changing **do**  
//Traverse  $D^+$  to update  $Pr(s|S \cup t^*), Pr(s|S)$   
4:  $Recur^+(Pr(s|S \cup t^*), Pr(t|S \cup t^*), D^+, g, 1)$   
5:  $Recur^+(Pr(s|S), Pr(t|S), D^+, g, 1)$   
6: Update  $Pr(d|S \cup t^*), Pr(d|S)$  using Eq. 10  
7: Return  $\mathbb{E}[\#s|S \cup t^*] - \mathbb{E}[\#s|S]$  using Eq. 11 (similarly for **Max-SubBus**)

---

**Algorithm 5**  $Recur^+(Pr(s|S), Pr(t|S), D^+, x, v)$ 


---

**Input:**  $Pr(s|S), Pr(t|S), D^+, x, v$  (the probability that none of the previous nodes fail)

**Output:** Update the values of  $Pr(s|S)$

1: **if**  $x$  is a substation **then**  
2:  $Pr(x|S) = 1 - v$   
3: **else**  
4:  $kids = \{\text{child nodes of } x \text{ in } D^+\}$   
5: **for each kid in kids do**  
6:  $Recur^+(Pr(s|S), Pr(t|S), D^+, kid, v * (1 - Pr(x|S)))$

---

Figure 3.4: Left: The Update function which can be used in scenarios with and without loops. Right: the  $Recur^+$  scenarios which is used in the scenarios with loops. (see section 3.3.2 for details).

Note that the overall time complexity of our HotSpot framework can be calculated as :

$$T = O[kV_t(mE_t + lV_t + lV_s)] \quad (3.12)$$

where  $V_t$ ,  $V_s$  are the number of transmission and substation nodes, respectively,  $E_t$  is the number of edges contained in the transmission network,  $l$  is the total number of loops happened in the whole failure propagation until the convergence, and  $m$  is the number of simulations to get the failing probabilities for transmission nodes.

## 3.4 Experiments

We build multiple experiments to compare the performance of our algorithms to various baselines. For these experiments, we treat the load and the capacity of all transmission nodes as a fixed value. The actual load and capacity information needs the infrastructure data and can easily be adopted by updating the related values in F-CAS.

**Datasets:** We build the heterogeneous CIS network using the data from ORNL. Data from four different US states are used: Pennsylvania (PA), Tennessee (TN), Ohio (OH), and Florida (FL). We show the details of these states CIS network in figure 3.5.

Node Type	TN	PA	FL	OH
Power Plants	11	45	79	35
Transmission Nodes	206	224	253	105
Electrical Substations	489	831	1590	806
Gas Compressors	105	291	45	189
Pipelines	387	5667	624	7641

Figure 3.5: The number of nodes in each component of the CIS network for the four states used in the experiments.

### 3.4.1 Baselines

We compare the performance of several baselines to our algorithm. Since there are no existing algorithms which were built for solving the same problems defined as **Max-Sub** and **Max-SubBus**, we show two related algorithms and generate several variations based on our algorithms using different selection strategies.

- **OPERA**: This is a former study that is aimed at choosing  $k$  critical nodes to maximize the connectivity of a network [18]. They use the number of triangles in the network to measure the connectivity of the network. We adopt this method in the transmission network.
- **NETSHIELD**: This is an immunization method that is targeted at minimizing the epidemic threshold of the network [44]. We use it to select the best  $k$  nodes from the transmission network by this method's ranking results.
- **DEGREE**: We pick  $k$  nodes from the transmission network by the highest degrees.
- **PAGERANK**: We pick  $k$  nodes from the transmission network by the largest pageranks.
- **RANDOM**: We pick the transmission nodes in totally random.

### 3.4.2 Effectiveness (Q1)

To test the effectiveness, we compare the performance of our algorithm to the baselines as mentioned above in figure 3.6. We show only the OH state results for clarity sake (other states' results are similar). Our algorithm HotSpots outperforms all baselines in both scenarios, e.g., **Max-Sub** and **Max-SubBus**. This is because HoSpots not only captures the connections between different components in the CIS network but take considerations of the dynamics of failure propagation.

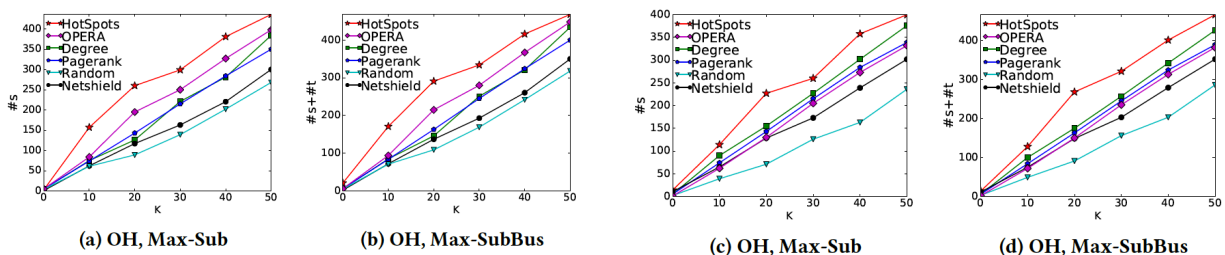


Figure 3.6: The comparison of the performance of HotSpots to all other baselines in the OH state area. Panels (a) and (b) are for the **Tran-naive** scenario, while the panels (c) and (d) for the the **Trans-real** scenario. HotSpots outperform all baselines.

### 3.4.3 Scalability (Q2)

The time scale of algorithms is also important in our case. Fast and in time predictions can provide more time for the government to handle the infrastructure failures. We evaluate here how HotSpots scales if size of initial set,  $k$ , or the size of the whole system,  $|V|$ , varies. As shown in figure 3.7, our algorithm scales close to linearly to  $k$  and close to quadratically to  $|V|$ . Note that in all of our scenarios, HotSpots finished in less than half an hour in picking the top-50 nodes.

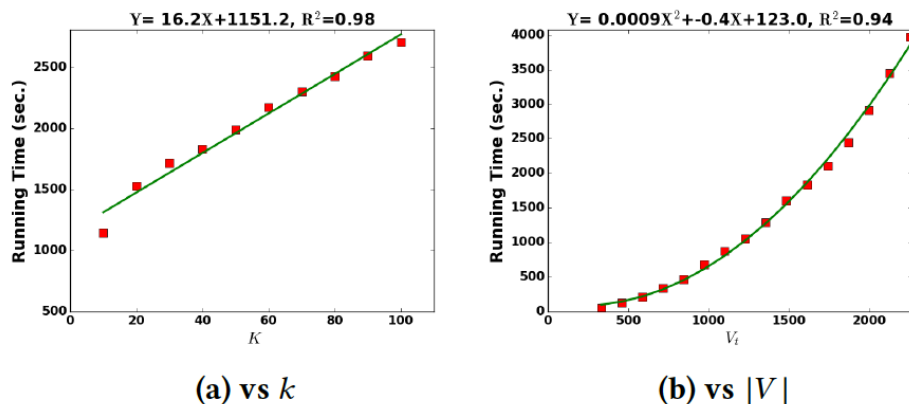


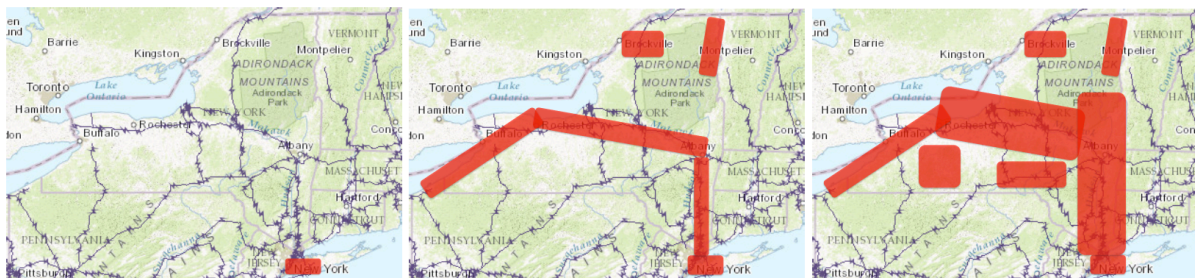
Figure 3.7: The time scale of HotSpots and how it scales with  $k$  and  $|V|$  (see section 3.4.3 for details).

### 3.4.4 Successful Case Studies (Q3)

We apply the HotSpots algorithm to the real case studies with the data provided by ORNL and others from the website. Since the HSIP Gold data from ORNL is not open to the public, we can't plot the network directly due to national security reasons. However, in our visualizations, we instead use the general maps and transmission wires found from the publicly available data of the US Energy Information Administration (EIA) [3].

**Heterogeneous networks in the case of Hurricane Sandy:** To capture the physical, geographical and cyber interdependencies among different CIS components, we need to construct the heterogeneous networks. We overlay track of hurricane sandy from Hurricane Mapping [4] to our constructed full CIS heterogeneous network in figure 3.8. We show in three panels how the hurricane first touch the New York (NY) city in panel (a) and in our F-CAS model it affects 227 nodes in G. By using these 227 nodes as the initial conditions, F-CAS can predict the further failure cascades as in panel (b) and (c). Before the cascade forms a loop, 487 nodes are failing, while after the cascade forms a loop, there are 712 nodes failing. We draw the area of failures in the map by red boxes, and these results match existing hurricane assessment tools, such as OCIA[3], EARRS [9], and HEADOUT [2].

**Comparison with the 2003 National Blackout:** Another in-depth case study is the US NE 2003 blackout which was found to originate from a single high voltage transmission line in northern OH failed due to overheating [5]. This again reveals the vulnerability of CIS networks and finding and protecting the critical nodes in the network is critical. Therefore, we run our algorithm HotSpots to pick the "best" 5 vulnerable nodes in the OH region and the results are shown in figure 3.9. By comparing our results to the actual blackout reports, we find that one critical node (the right-top one) we identified is only one hop away to the actually failed transmission lines. This is the point which failed in 2003 and triggered the



(a) 227 nodes initially affected in NY city (b) 487 nodes fail before the cascade loop (c) 712 nodes fail after a loop of cascade

Figure 3.8: The case study of Hurricane Sandy’s impact paths. (a) the total area of directly affected by the Hurricane Sandy predicted by HotSpots when we overlay the constructed CIS network  $G$  with the hurricane’s path. (b) Results from F-CAs when the failure cascades has not form a loop yet. (c) F-CAS simulation results after a loop has formed, which affects much larger areas. We mark the affected regions by the red boxes (the actual network is not shown due to security reasons, see section 3.4.4 for details).

whole blackout. This finding strengthens the accuracy of HotSpot and can suggest the DHS NIPP [4] to maintain or protect these critical nodes to prevent such catastrophes. On the contrary, other baselines do not find this critical node.

## 3.5 Discussions and Summary

### 3.5.1 Discussion

- Dominator tree is a popular data structure which is usually used in software engineering. Moreover, since it is only an approximation to the original graph, in the case of **Max-Sub** and **Max-SubBus** problem, dominator tree provides a lower bound of the objective functions, i.e., Dominator tree method will underestimate the final failed substation nodes. However, the advantage is 1) Dominator tree can catch important dynamics in the graph by the immediate-dominator pairs; 2) Dominator tree simplifies the original NP-hard problem into a solvable and optimizable problem. 3) By comparing with the case studies,

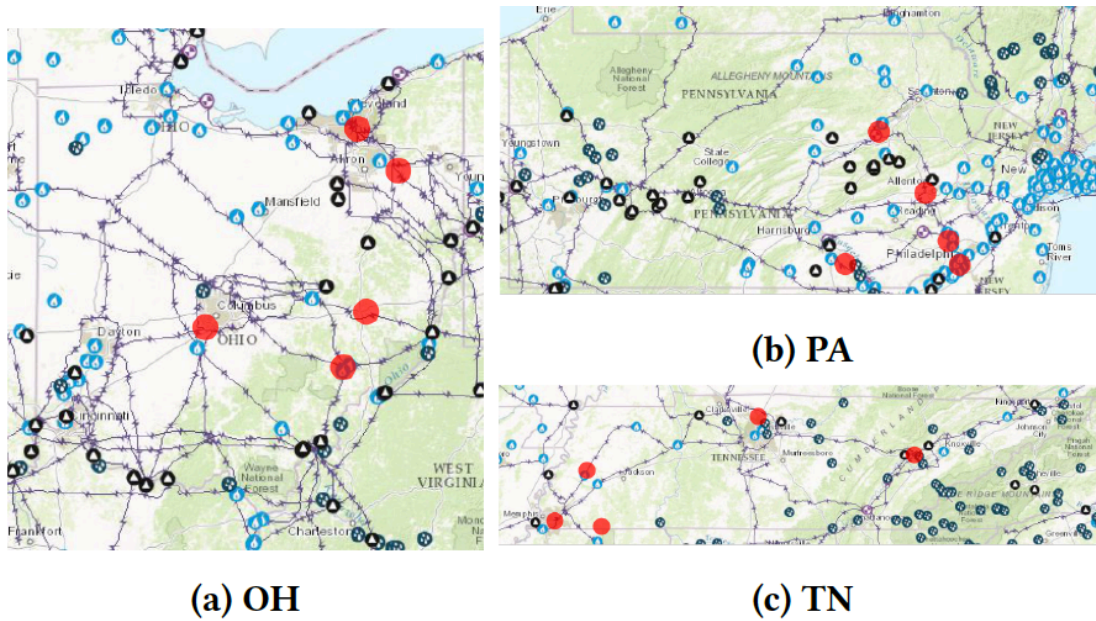


Figure 3.9: The case study of the US NE 2003 blackout. We use the HotSpot algorithm to find the top five transmissions in the OH area, which is highlighted by the red circles.

we find that our HotSpots method by taking the dominator tree approximation can make accurate and meaningful predictions. Therefore, we can conclude that the dominator tree is an effective and efficient data structure for our problems, which already enables high-quality results.

- Another question is how robust our method is to the original data. Since we got the infrastructure data from ORNL and construct the networks by taking advice from their domain experts, the first assumption is that the data and networks are close to accurate. Moreover, the important nodes usually have more links and are better described in the infrastructure networks. Therefore, it is less possible for the data to have “errors” near the important nodes. However, if the data has “errors” in some of the nodes or edges, what will happen to the predictions made by our HotSpots method? In our dominator tree approximations, the closer the “errors” to the root node, the stronger that our method will be affected. Since the “close-to-root-errors” will affect the whole branch in the dominator

tree graph. Moreover, our case studies show that we get meaningful and close predictions based on current data. Therefore, we conclude that our data can represent most of the real CI structures. However, the robustness of the data is an important work and needs in-depth studies in the future.

### 3.5.2 Summary

As far as we know, the proposed F-CAS model and the HotSpots algorithm are the first attempt in modeling and analyzing heterogenous CIS networks with five CI components. Moreover, additional components can be easily added to this framework. The CIS network is generated to help the network-based analysis and easily visualize the propagation of failures. As shown above, our failure cascade model F-CAS runs fast and captures both the path-based failures and neighbor-based failures, which are not modeled by any other cascade models yet. Note that this kind of path-based failure is not restricted within the transmission network. Similarly, in the electrical distribution networks, overload patterns and path-based failures also exist. Moreover, in the SCADA system (supervisory control and data acquisition, see [5]), the communication servers give signals to control other facilities through the networks including the WAN and LAN. If these networks failed due to some issues, message overflow could happen (similar to the reported overloading effect in F-CAS), and it can cause further issues to the entire system. In all, F-CAS and HotSpots can be used to a vast range of CIS networks to model and find the essential nodes among the system.

# Chapter 4

## Modeling and Predicting the Influenza Spread

In this chapter, we propose EpiDeep, a deep neural network, to learn meaningful representations of the trend of influenza incidence and get accurate predictions for several key epidemiology metrics, including future incidences, peak intensity and peak time. We present extensive experiments on forecasting influenza-like illnesses (ILI) in the United States based on the data from the Centers for Disease Control and Prevention (CDC). Our results show that EpiDeep is successful at learning meaningful embeddings and can track the evolution of ILI among seasons. In all, our approach outperforms non-trivial baselines by up to 50%.

The structure of this chapter is as follows: First, we give some introductions to the ILI and the CDC flu challenge in section 4.1. In section 4.2, we formalize the problem and we present our model in section 4.3. In section 4.4, we show the experiments and compare them to other non-trivial baselines. We summarize our findings and results in section 4.5.

### 4.1 Introduction

Seasonal influenza is a major global health issue that affects many people across the world. Centers for Disease Control and Prevention (CDC) reports that there were 30,453 laboratory-confirmed influenza related hospitalization in 2017/18 influenza season in the United States

alone. According to the same CDC estimate, the 2017-18 season saw a record number of deaths due to influenza in the past five seasons. These statistics reveal that despite years of efforts, accurately predicting key indicators of influenza season and employing counter-measures remains a major challenge.

To encourage research teams to make accurate real time forecasts, CDC has been hosting ‘FluSight’ challenge for seasonal influenza forecasting at the national and regional levels [10, 11]. i.e., challenge provides four forecasting targets, namely Future Incidences, Seasonal Peak Intensity, Seasonal Peak Time and Onset Week. We will discuss each forecasting targets in detail in the next section. CDC also provides historical data collected by the Outpatient Influenza-like Illness Surveillance Network (ILINet) to aid in forecasting. ILINet consists of more than 3,500 outpatient healthcare providers all over the United States. Each week the healthcare providers voluntarily report the total number of Influenza-like Illness (ILI) related visits. Then the CDC compiles the reports and releases percentage weighted ILI (wILI) for national and regional levels with a delay of two weeks. The forecasting targets for the FluSight challenge are in terms of the wILI incidence curve. The ILINet data, which provides weekly wILI incidence curves for each season since 1997/98 is publicly available.

A key challenge in accurate forecasting here is the dynamic nature of the influenza season [35]. For example, wILI incidence for season  $x$  can be similar to that of season  $y_1$  at the beginning of the season, but could end up being very different at the later stages. None of the existing methods can catch this property of the influenza, and we try to learn from this and include the dynamic nature of the influenza into our deep-learning method.

## 4.2 Problem Formulation

The CDC FluSight challenge provides a standard goal of predictions. Therefore, we follow the rules and build our models based on these standards. There are three epidemiological metrics to forecast: Future Incidence, Seasonal Peak Time, and Seasonal Peak Intensity. Here we describe these metrics [13]:

- **Future Incidence:** This is forecasting the short term future incidence, e.g., the next one to four weeks ahead of the latest wILI data. Due to the delay of ILINet data (two weeks late), researchers, at week  $t$  of one season, actually are required to make predictions of wILI values for the week  $t-1$ ,  $t$ ,  $t+1$ , and  $t+2$ . These predictions are critical since CDC and prepare and distribute resources among the country two weeks before an outbreak.
- **Seasonal Peak Intensity:** This is forecasting the maximum intensity of influenza, i.e., the maximum wILI values, in the current season. For the nation, it is important since the amount of medical-related resources are directly connected to the seasonal peak intensity.
- **Seasonal Peak Week:** This is forecasting the time when seasonal peak intensity is observed. From the definition of the CDC, the peak week is when the wILI values, rounded to the first decimal, reach the highest. Similar to peak intensity, correct forecasting of peak week will allow CDC to distribute resources over the country ahead of the possible outbreak.

At week  $t+2$ , we have the available data as time-series from CDC:  $\mathcal{Y}_c = \{y_c^1, y_c^2, \dots, y_c^t\}$ , which represents the former stage of the current season  $c$  till week  $t$ . Each  $y_c^i$  are the wILI data for the week  $i$ . Therefore, our goal is to predict all three targets for the season  $c$  given  $\mathcal{Y}_c$ :

### PROBLEM 1. EPIDEMICPREDICTION

**Given:** a time-series  $\mathcal{Y}_c = \{y_c^1, y_c^2, \dots, y_c^t\}$ , which represents the current season  $c$  until the week  $t$ .

**Predict:**

- **Task 1:** Future Incidence Prediction:  $\forall_{i=t+1}^{t+4} y_c^i$
- **Task 2:** Peak Intensity Prediction:  $\max y_c^i \forall_{i=1}^T$ , where  $T$  represents the last week of the current season  $c$ .
- **Task 3:** Peak Time Prediction:  $\arg \max_i y_c^i \forall_{i=1}^T$ , where  $T$  represents the last week of the current season  $c$ .

### 4.3 EpiDeep Model

We are given historical wILI data, and they are viewed as the ‘training set’, on which we propose a deep neural network based approach to solve the above Problem 1. At the high level, our model tries to encode the similarities between the observed part of the current season to the past seasons. Therefore, the model can learn from previous data and make accurate forecasts. Overall, the model is trained by the set of historical wILI data until last season  $Y = \{\mathcal{Y}_1, \mathcal{Y}_2, \mathcal{Y}_3, \dots, \mathcal{Y}_{c-1}\}$ . Once the model is well trained, we predict the mentioned key epidemiology metrics for the current season. Based on the literature [13], the first week of the season  $y_i$  is defined as the 20th week of the year.

The primary challenge when encoding the similarities between the past seasons and the current seasons  $y_c$  is that we only observe till week  $t$  for the current season. Therefore, simple ideas are not applicable here, e.g., calculating the distance between curves. We solve this issue in several steps. We show the overall architecture of EpiDeep in figure 4.1 and we discuss them in details here.

First, we embed the similarities between the current season and past seasons limited till week  $t$  of each season. It is accomplished by the ‘Query Length Data Clustering’ module. We

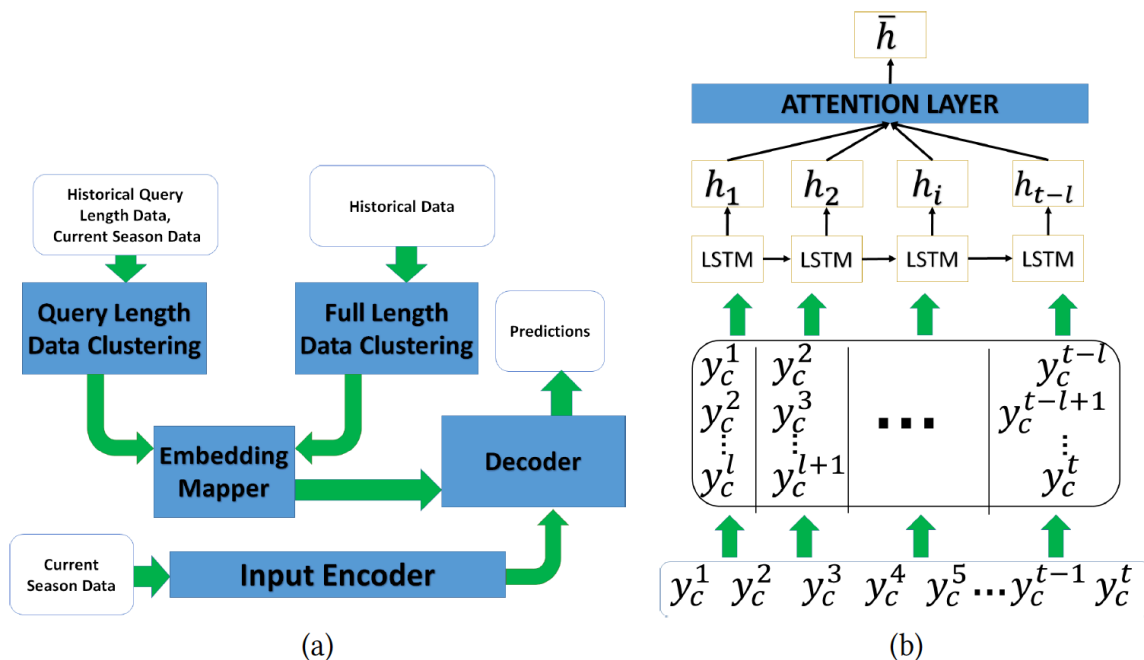


Figure 4.1: (a) The architecture of the proposed EpiDeep model, including the clustering, encoder, and decoder module. (b) The structure of the encoder module, where the attention layer is added to the end.

treat the observed period of the current season,  $y_c$ , as a ‘query’ to the past seasons, where they have the fully observed season-length historical wILI data. Then we get the snippets of the historical data till the observed week  $t$ . The module learns the feature representation of these historical data comparing with current season  $y_c$ . In order to capture the similarities between the current season and past seasons, this module learns embeddings for each season with  $y_c$  and then cluster them such that similar seasons are embedded together. However, since the query length data is only part of the available information, we still need to consider the full-length historical data.

Therefore, secondly, we learn the embeddings to capture the similarities between the full-length historical data without taking consideration the current season. The ‘Full Length Data Clustering’ module accomplishes this and embeds the full data into a continuous space

so that the clustering of the embeddings are meaningful. Then we learn a mapping function to map from the query length space to the full-length historical data space. In this way, we can learn embeddings of the current season with taking consideration of the fully observed historical wILI data.

The last part when embedding the inputs is the ‘Input Encoder’ module, which provides a simple representation of the current season  $y_c$ . This module is mainly focused on extracting important information from any snippets in  $y_c$ . Finally, the ‘Decoder’ module combines all the outputs from these modules and makes predictions for the required three epidemiological metrics.

For EpiDeep, we first pre-train the clustering of historical data and the mapping layers and then jointly train the entire model. The adaptive moment estimation (Adam) optimization algorithm [31] was used to determine the model parameters. The automated differentiation package in Pytorch [17] is used in our model.

For the CDC FluSight challenge, we follow the same architecture to EpiDeep. However, the challenge requires binned probabilistic predictions. Therefore, we assign probabilities to each result bin pre-defined by the CDC.

## 4.4 Experiments

In order to test the performance of our EpiDeep model, we build several experiments comparing to multiple non-trivial baselines. We discuss them in detail here.

### 4.4.1 Initial Settings

We describe the initial settings for our experiments. All experiments are conducted using a 4 Xeon E7-4850 CUY with 512GB of 1066 Mhz main memory. The paper and code are available for only academic purposes <sup>1</sup>.

**Data:** The wILI historical data from CDC are collected and used in all experiments. These data and the instructions are available online <sup>2</sup>. For each of the predictions made here, only the historical data observed former than the time of forecasting is adopted.

**Baselines:** There exist various baselines which can be able to forecast the influenza incidence. However, a vast majority of them require additional data inputs such as twitter feeds, human-mobility data, and weather data. Since the only inputs for EpiDeep are the historical wILI data, we choose to compare with those non-trivial baselines which can forecast given only wILI data. The details of these baselines are described here:

- **Hist:** It is one straight-forward way for influenza incidence forecasting. We compute the average of past seasons until the predicted week  $t$  for each season. The average values are used as predictions.
- **ARIMA:** It is an auto-regressive method which can be used for predicting time-dependent data. It is popular in multiple fields, including transportation and epidemics. Here we adopt ARIMA (7,0,1) since it has the best performance.
- **KNN:** It is the K-nearest-neighbor method, where we choose the closest (most similar)  $k$  historical seasons and average them as the predictions. Many reported approaches adopted similar ideas by calculating the closest past season [46].
- **LSTM:** It is one of the popular machine learning methods: Long Short Term Memory network. This is the simplest version [46] with no additional data inputs except the

---

<sup>1</sup><https://bit.ly/2GmEjqD>

<sup>2</sup><https://gis.cdc.gov/grasp/fluview-/fluportaldashboard.html>

historical wILI data. This model is similar to EpiDeep without the embeddings and the attention model.

- **EB:** It stands for an Empirical Bayes framework, which has been reported in multiple papers, e.g., [13]. It makes predictions based on the fitting of the past season to the current season, while the previous seasons are translated by model parameters. This is the model which performs best in the past few seasons and has already won several of the CDC FluSight challenges. It is interesting to see the comparison of EB to EpiDeep.

**Key Epidemiology Metrics:** As mentioned in section 4.2, three important metrics are required by the CDC and used in this paper. There are also other discussions about the metrics for evaluating models for epidemic forecasting [41].

- **RMSE:** The root mean squared error, which is defined as:  $\text{RMSE} = \sqrt{\frac{\sum_{i=1}^N e_i^2}{N}}$ . One important characteristic of RMSE is that it penalizes extreme violations from the average values.
- **MAPE:** The mean absolute percentage error, which is defined as  $\text{MAPE} = \frac{1}{N} \sum_{i=1}^N \left| \frac{e_i}{y_i} \right|$ . The difference of MAPE from RMSE is that it does not penalize the extreme violations.
- **Log Score:** The CDC FLuSight Challenge requires the probabilistic predictions. In given result intervals, the reported values are the log score which is defined as  $\log(p, i) = -\log(p_i)$  for the probability assigned to the bin  $i$ .

#### 4.4.2 National Predictions

One of the main tasks required by the CDC FluSight challenge is to predict the epidemiology metrics at the national level. We compare the performance of EpiDeep to all baselines at the US national level beginning from season 2010/11 until the season 2016/17 here. In order to have enough input data for each season, the predictions only start after reaching the

Table 4.1: The comparisons of the performance for all methods in forecasting the three tasks for epidemical seasons from 2010/11 to 2016/17. R, M, and LS are for RMSE, MAPE, and the average Log Score, respectively. A “–” sign represents that the method is not applicable in this prediction. EpiDeep consistently performs well in all forecasting tasks and outperforms all the baselines in the majority of the scenarios.

Season	10/11			11/12			12/13			13/14			14/15			15/16			16/17		
Method	R	M	LS	R	M	LS	R	M	LS	R	M	LS	R	M	LS	R	M	LS	R	M	LS
<b>Task 1: Future Incidence Prediction</b>																					
Hist	1.29	0.4	39.82	0.44	0.21	49.57	1.4	0.36	46.92	0.77	0.23	54.88	1.12	0.26	61.96	0.65	0.23	39.0	0.9	0.22	53.12
ARIMA	0.65	<b>0.15</b>	–	<b>0.28</b>	<b>0.12</b>	–	0.89	<b>0.17</b>	–	0.65	<b>0.12</b>	–	0.88	0.17	–	0.42	0.13	–	0.67	0.15	–
KNN	0.76	0.26	57.32	1.57	0.71	75.11	0.81	0.24	75.97	1.06	0.37	76.86	<b>0.61</b>	0.24	75.1	0.98	0.36	80.41	0.65	0.2	77.75
LSTM	0.92	0.36	40.12	0.72	0.32	58.41	0.93	0.32	54.46	1.25	0.40	79.76	0.82	0.19	64.45	0.78	0.33	56.97	0.98	0.31	72.06
EB	0.81	0.31	43.29	0.97	0.5	60.11	1.04	0.24	65.42	0.67	0.24	58.37	0.87	0.21	61.8	0.93	0.39	47.79	1.06	0.32	56.61
EpiDeep	<b>0.59</b>	0.17	<b>26.61</b>	0.36	0.16	<b>32.2</b>	<b>0.68</b>	<b>0.17</b>	<b>29.89</b>	<b>0.45</b>	<b>0.12</b>	<b>36.03</b>	0.73	<b>0.15</b>	<b>41.01</b>	<b>0.41</b>	<b>0.13</b>	<b>29.75</b>	<b>0.58</b>	<b>0.15</b>	<b>35.15</b>
<b>Task 2: Peak Intensity Prediction</b>																					
Hist	1.39	0.31	∞	1.0	0.42	∞	2.55	0.42	∞	0.78	0.17	1.18	1.94	0.32	1.2	0.78	0.22	0.93	<b>0.54</b>	<b>0.11</b>	∞
ARIMA	2.47	0.52	–	0.64	0.25	–	4.01	0.65	–	2.76	0.6	–	3.93	0.65	–	1.55	0.41	–	2.82	0.54	–
KNN	1.31	0.28	∞	3.28	1.35	∞	0.61	0.1	∞	0.9	0.19	32.47	<b>0.18</b>	<b>0.03</b>	∞	1.49	0.4	53.73	0.58	0.09	∞
LSTM	1.43	0.32	89.91	1.35	0.56	77.84	1.94	0.51	56.92	0.84	0.17	22.41	2.83	0.42	0.94	1.42	0.37	43.77	1.98	0.38	95.41
EB	<b>0.96</b>	0.21	<b>60.73</b>	<b>1.21</b>	<b>0.48</b>	82.16	2.48	0.41	<b>42.86</b>	1.1	0.24	0.46	2.3	0.38	0.45	<b>0.24</b>	<b>0.06</b>	<b>11.24</b>	1.41	0.28	64.29
EpiDeep	0.99	<b>0.2</b>	71.4	1.59	0.6	<b>71.03</b>	<b>1.36</b>	<b>0.21</b>	71.43	<b>0.56</b>	<b>0.1</b>	<b>0.37</b>	2.26	0.37	<b>0.34</b>	0.46	0.11	18.64	0.87	0.15	<b>43.9</b>
<b>Task 3: Peak Time Prediction</b>																					
Hist	17.0	0.3	∞	21.0	0.33	∞	8.0	0.15	0.67	6.0	0.12	0.57	5.0	0.1	0.51	13.0	0.21	∞	7.0	0.12	0.95
ARIMA	33.53	0.58	–	37.1	0.58	–	27.3	0.51	–	27.08	0.51	–	26.93	0.5	–	38.69	0.62	–	34.8	0.59	–
KNN	12.0	0.21	∞	5.9	0.09	∞	16.82	0.32	0.47	16.82	0.32	0.33	11.0	0.21	0.14	<b>6.6</b>	<b>0.1</b>	∞	10.17	0.17	∞
LSTM	7.09	0.14	64.13	5.6	0.08	81.32	9.35	0.24	1.48	9.73	0.22	0.29	19.24	0.41	21.25	11.45	0.23	50.44	8.85	0.32	88.4
EB	1.09	<b>0.02</b>	60.7	<b>4.5</b>	<b>0.07</b>	78.6	<b>5.4</b>	<b>0.1</b>	<b>0.3</b>	<b>1.0</b>	<b>0.02</b>	<b>0.2</b>	<b>1.4</b>	<b>0.02</b>	<b>0.2</b>	8.04	0.11	75.0	<b>6.3</b>	<b>0.1</b>	64.2
EpiDeep	<b>1.0</b>	<b>0.02</b>	<b>33.2</b>	5.1	0.08	<b>29.2</b>	6.0	0.12	0.33	6.0	0.12	0.26	3.71	0.05	0.28	10.3	0.16	<b>29.6</b>	6.65	0.11	<b>21.6</b>

epidemiological week 40 until the next year’s week 20 (the same epidemiological season). Moreover, for the Peak Intensity and Peak Time forecasting tasks, we only predict until we observe the actual peak from the data. We compare the results in table 4.1. The ARIMA method does not predict probabilistic values. Therefore, we do not report the log score of the ARIMA baseline.

As presented in the table, EpiDeep performs better than all the baselines in the majority of scenarios. Notably, EpiDeep outperforms the best baseline EB in two out of the three tasks, i.e., Future Incidence Prediction and Peak Intensity Prediction. This is a big success since EB performs the best in most of the FluSight challenge in the past several epidemical seasons. Overall, EpiDeep has the best performance in 17 out of 21 scenarios for Future Incidence Prediction, in 10 out of 21 scenarios for Peak Intensity Prediction, and in 7 out of 21 scenarios for Peak Time Prediction. When EpiDeep is not the best, it is close to the second/third best ones. Moreover, simpler baselines including Hist, KNN, and ARIMA have large fluctuations in their prediction performances. They sometimes perform very well but

most of the time is consistently wrong. On the other hand, EpiDeep, LSTM and EB all have stable performance in all tasks, and EpiDeep beats the other two in most of the scenarios.

### 4.4.3 Regional Predictions

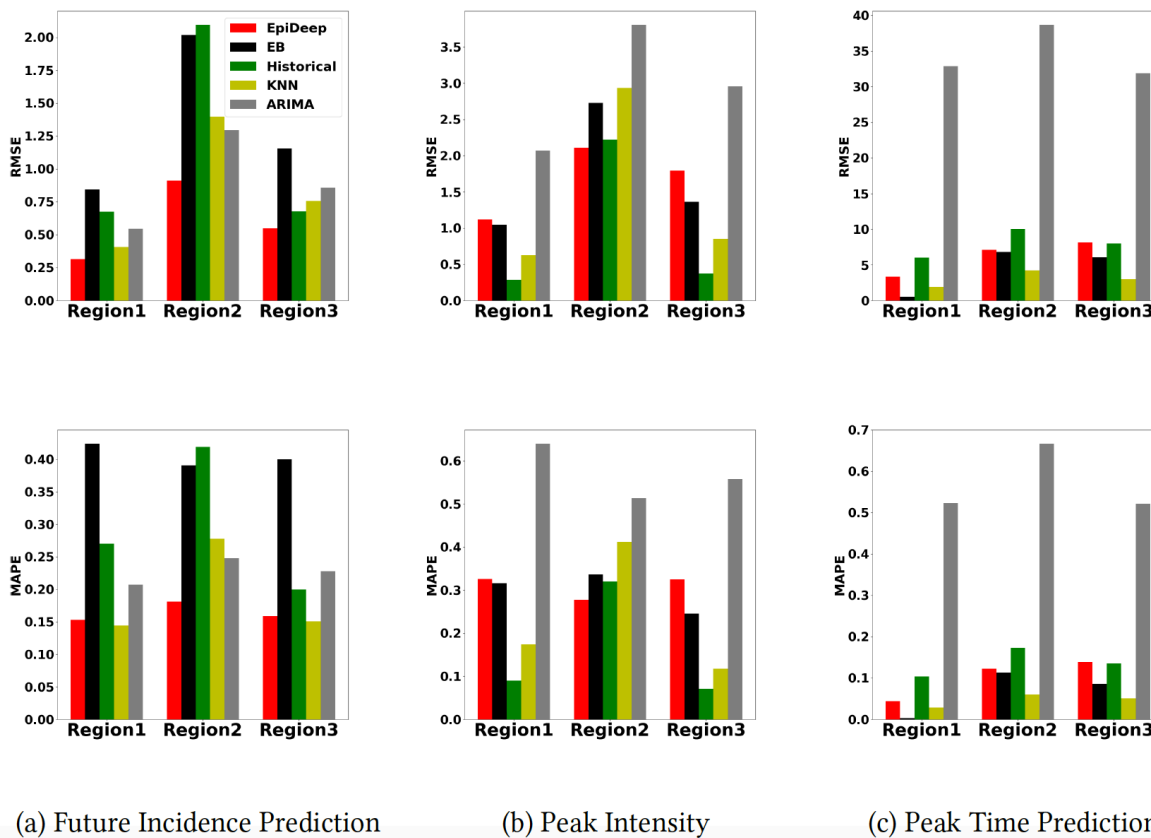


Figure 4.2: The RMSE and MAPE performances for regional predictions of all three tasks for the 2016/17 season. EpiDeep consistently perform well comparing to other baseline.

Besides the Nation Predictions, CDC also requires the researchers to predict the wILI trends in different regions. US has been divided into 10 HHS regions by the US Department of Health and Human Services. CDC provide the wILI data for each of these regions every week. Therefore, we apply our method and other baselines to these data and make forecasts for different HHS regions. Notably, the influenza patterns can be different for different regions,

since their position, population, and weather are different from each other and can affect the trends of influenza.

Overall, we find that EpiDeep consistently perform well in all regional predictions. We show an example of the predictions in HHS region 1, 2, and 3 for the 2016/17 season in figure 4.2. We found that EpiDeep outperforms other baselines in the Future Incidence Prediction, and are comparable to be the best in the other two tasks.

#### 4.4.4 Analysis of Delayed Data Arrival

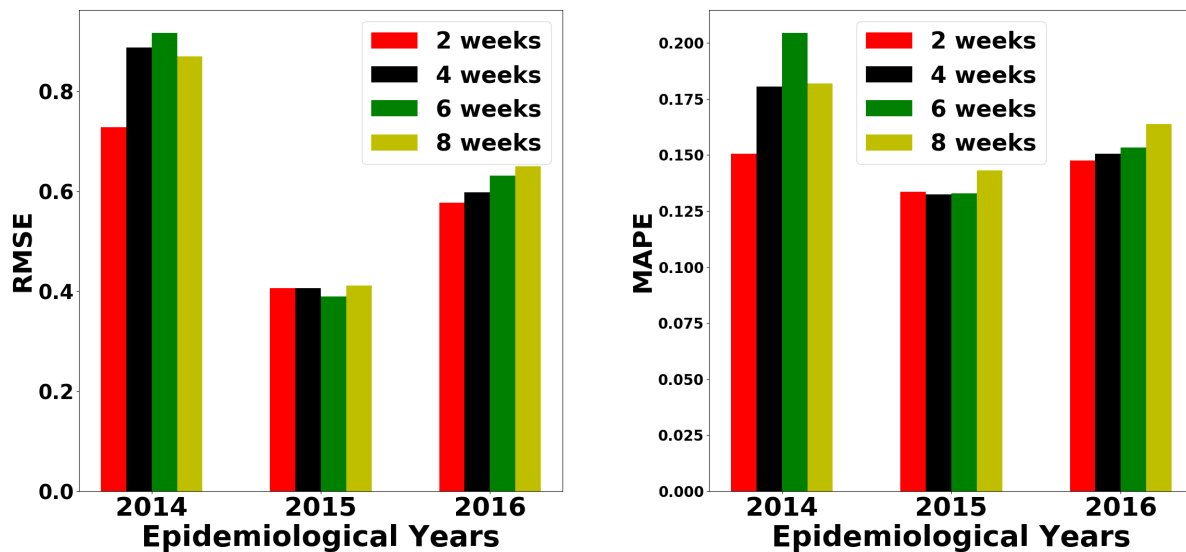


Figure 4.3: EpiDeep’s performance on RMSE and MAPE for Future Incidence Predictions with simulated delayed data. The performance remains stable and has a trend to perform better when the time delay is shorter.

As mentioned in Section 4.1, the wILI data reported by the CDC has a time delay of about 2 weeks. An interesting experiment is to check if EpiDeep performs well when longer time delays are added to the data input. Will the performance remain stable or drop drastically when longer delays exist?

We set the experiments to check this question, and the results are shown in figure 4.3. We leveraged **EpiDeep** to work on Future Incidence Prediction with data’s time delay of 2, 4, 6, and 8 weeks. The experiments were done in three epidemic seasons, i.e., 2014/2015, 2015/2016, and 2016/17 and at the national level. We observed that **EpiDeep** consistently perform well even the time delay reaching 8 weeks. This confirms the fact that **EpiDeep** can learn from the longtime trend and make stable predictions even when there is a more significant delay in data arrival.

## 4.5 Discussions and Summary

In this section, we discussed our method to predict the spread of influenza in ILI incidence. We proposed a novel deep learning model **EpiDeep** to learn from the historical epidemic data and then compare with the observed current season to make better predictions. Comparing with other non-trivial baselines, **EpiDeep** outperforms them in the majority of scenarios for all three tasks. Moreover, we found that **EpiDeep** can make stable predictions when there is a longer delay time in the data arrival.

Our proposed method was designed to overcome particular challenges in influenza forecasting cases. The first challenge is the data sparsity issue. Since the ILINet data from CDC became available after 1997, there are only around 20 flu seasons which can be used for predictions. Therefore, we add the deep clustering modules to help embed similar seasons together and finally combine with the LSTM results to make better predictions. This challenge may happen in other deep learning studies, where we can adopt the similar method, i.e., find the inner structures of the data (in our case, the periodic seasons) and learn the similarities/differences to help the learning processes.

The second challenge is the lack of other related data, e.g., the transportation data, and

human-contact information. **EpiDeep** already makes good predictions without these data. Moreover, in the future, we may want to take these data into considerations and jointly embed them to make even better predictions. Also, the geographical relations of different HHS regions can also be viewed as constraints to the loss functions of **EpiDeep**. Though notably, **EpiDeep** could already discover many of these relationships automatically from the historical wILI data.

# Chapter 5

## Conclusions and Future Work

In conclusion, this thesis focuses on modeling and predicting incidence in two different scenarios, i.e., critical systems failures and flu infection cases.

In chapter 3, we construct the F-CAS model to capture the interconnectivity and dynamics of the CIS networks. After that, we proposed the HotSpots algorithm to identify key facilities in CIS networks. As it is shown in several case studies, our algorithm outperforms all baselines and present meaningful results comparing to the actual infrastructure failure cases.

In chapter 4, we build a deep-learning based algorithm, EpiDeep, to model the spread of influenza-like illness. We then predict several key epidemiological metrics following the rules from the CDC. Our experiment results show that EpiDeep is successful at learning meaningful embeddings and can track the evolution of ILI among seasons. In all, our approach outperforms non-trivial baselines by up to 50%.

**Future Work:** We take the dynamic nature of propagation into account and build models to forecast the propagation of contagions in different cases. Moreover, our studies and methods can be applied to similar and other different fields. E.g.,

- For the case of the CIS network, the results from our proposed approaches can serve as a useful input for other high-performance but time-consuming simulations. Moreover, applying HotSpots among other similar CIS networks would also get meaningful results, such as the water network, transportation network, and the SCADA system. As men-

tioned in section 3.5.1, we constructed the networks based on the ORNL data and adopted the approximation of dominator trees. Since our case studies show accurate and meaningful results, the “error” introduced from the data and the dominator tree approximation should be minor. However, a more systematic study and approach to investigating the robustness of these two points is needed and is an interesting future work.

- For the case of the ILI infection scenario, we would like to add other key epidemiology metrics into the predicts, e.g., the seasonal onset defined from the CDC. Moreover, the EpiDeep has the potential to add additional information into the framework. For example, there are social media data, weather data, etc. EpiDeep can jointly embed multiple heterogeneous data sources and leverage them for more accurate and meaningful predictions. On the other hand, the ILI data usually has the geographical structure, and we can also take this information as constraints to the EpiDeep.
- In other cases of contagions, the dynamics of propagation can also be helpful. For example, the propagation of rumors/ideas through social media network is also dynamic, which are dependent on communication styles, personal influence, etc. Taking the dynamics into account can help maximize/minimize the spread of rumors/ideas. In other cases of diffusion models, deep learning based methods still exist. E.g., recent studies in [29], where they build the deep learning based algorithms to study the problem of cascade prediction utilizing only two types of (coarse) information, i.e., which node is infected and its corresponding infection time.

# Bibliography

- [1] Homeland security infrastructure program (HSIP). <https://gii.dhs.gov/HIFLD/hsip-guest>.
- [2] Hurricane electrical assessment damage outage tool (HEADOUT). [http://www.gss.anl.gov/wp-content/uploads/2015/09/MORS\\_Presentation\\_Talaber\\_WG21\\_and\\_WG31\\_060415.pdf](http://www.gss.anl.gov/wp-content/uploads/2015/09/MORS_Presentation_Talaber_WG21_and_WG31_060415.pdf).
- [3] U.S. energy information administration (EIA). <https://www.eia.gov/>.
- [4] U.S. hurricane mapping. <https://hurricanemapping.com/>.
- [5] Spencer Abraham, Herb Dhaliwal, R John Efford, Linda J Keen, Anne McLellan, John Manley, Kenneth Vollman, Nils J Diaz, Tom Ridge, et al. *Final report on the august 14, 2003 blackout in the united states and canada: Causes and recommendations*. US-Canada Power System Outage Task Force, 2004.
- [6] Réka Albert, István Albert, and Gary L. Nakarado. Structural vulnerability of the north american power grid. *Phys. Rev. E*, 69:025103, feb 2004.
- [7] Roy M. Anderson and Robert M. May. *Infectious diseases of humans : dynamics and control / Roy M. Anderson and Robert M. May*. Oxford University Press Oxford ; New York, 1991.
- [8] N. T. J. BAILEY. *The mathematical theory of infectious diseases and its applications. 2nd edition*. Number 2nd edition. Charles Griffin Company Ltd 5a Crendon Street, High Wycombe, Bucks HP13 6LE., 1975.

- [9] Alan M Barker, Eva B Freer, Olufemi A Omitaomu, Steven J Fernandez, Supriya Chinthavali, and Jeffrey B Kodysh. Automating natural disaster impact analysis: An open resource to visually estimate a hurricane's impact on the electric grid. In *South-eastcon*, pages 1–3, 2013.
- [10] Matthew Biggerstaff, David Alper, Mark Dredze, Spencer Fox, Isaac Chun-Hai Fung, Kyle S Hickmann, Bryan Lewis, Roni Rosenfeld, Jeffrey Shaman, Ming-Hsiang Tsou, et al. Results from the centers for disease control and prevention's predict the 2013–2014 influenza season challenge. *BMC infectious diseases*, 16(1):357, 2016.
- [11] Matthew Biggerstaff, Michael Johansson, David Alper, Logan C Brooks, Prithwish Chakraborty, David C Farrow, Sangwon Hyun, Sasikiran Kandula, Craig McGowan, Naren Ramakrishnan, et al. Results from the second year of a collaborative effort to forecast influenza seasons in the united states. *Epidemics*, 2018.
- [12] Christian Borgs, Michael Brautbar, Jennifer Chayes, and Brendan Lucier. Maximizing social influence in nearly optimal time. In *SODA*, pages 946–957, Philadelphia, PA, USA, 2014.
- [13] Logan C Brooks, David C Farrow, Sangwon Hyun, Ryan J Tibshirani, and Roni Rosenfeld. Flexible modeling of epidemics with an empirical bayes framework. *PLoS computational biology*, 11(8):e1004382, 2015.
- [14] William N. Bryan. *Hurricane Sandy Situation Report*. U.S. Department of Energy Office of Electricity Delivery & Energy Reliability, 2012.
- [15] Adam L. Buchsbaum, Haim Kaplan, Anne Rogers, and Jeffery R. Westbrook. A new, simpler linear-time dominators algorithm. *ACM Trans. Program. Lang. Syst.*, 20(6):1265–1296, 1998.

- [16] Sergey V. Buldyrev, Nathaniel W. Shere, and Gabriel A. Cwilich. Interdependent networks with identical degree of mutually dependent nodes. *Physical Review*, 83(1), 2011.
- [17] CDC. Summary of the 2009-2010 influenza season. <https://www.cdc.gov/flu/pastseasons/0910season.html>, 2010. Accessed: 2018-11-05.
- [18] Chen Chen, Jingrui He, Nadya Bliss, and Hanghang Tong. On the connectivity of multi-layered networks: Models, measures and optimal control. In *ICDM*. IEEE, 2015.
- [19] Chen Chen, Hanghang Tong, Lei Xie, Lei Ying, and Qing He. Fascinate: Fast cross-layer dependency inference on multi-layered networks. In *KDD*. ACM, 2016.
- [20] Liangzhe Chen, KSM Tozammel Hossain, Patrick Butler, Naren Ramakrishnan, and B Aditya Prakash. Syndromic surveillance of flu on twitter using weakly supervised temporal topic models. *Data mining and knowledge discovery*, 30(3):681–710, 2016.
- [21] Liangzhe Chen, Xinfeng Xu, Sangkeun Lee, Sisi Duan, Alfonso G. Tarditi, S Chinthavali, and B Aditya Prakash. Hotspots: Failure cascades on heterogeneous critical infrastructure networks. pages 1599–1607, 11 2017.
- [22] Wei Chen, Tian Lin, Zihan Tan, Mingfei Zhao, and Xuren Zhou. Robust influence maximization. In *KDD*. ACM, 2016.
- [23] Hadi Daneshmand, Manuel Gomez-Rodriguez, Le Song, and Bernhard Schoelkopf. Estimating diffusion network structures: Recovery conditions, sample complexity & soft-thresholding algorithm. In *ICML*, pages 793–801, 2014.
- [24] Leonardo Dueñas-Osorio, James I. Craig, and Barry J. Goodno. Seismic response of critical interdependent networks. *Earthquake Engineering and Structural Dynamics*, 36(2):285–306, 2007.

- [25] Ajendra Dwivedi and Xinghuo Yu. A maximum-flow-based complex network approach for power system vulnerability analysis. *IEEE Transactions on Industrial Informatics*, 9(1):81–88, 2013.
- [26] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012, 2009.
- [27] Michaela Goetz, Jure Leskovec, Mary McGlohon, and Christos Faloutsos. Modeling blog dynamics. In *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, California, USA, May 17-20, 2009*, 2009.
- [28] Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th International Conference on World Wide Web, WWW '04*, pages 491–501, New York, NY, USA, 2004. ACM.
- [29] Mohammad Islam, Sathappan Muthiah, Bijaya Adhikari, B Aditya Prakash, and Naren Ramakrishnan. Deepdiffuse: Predicting the 'who' and 'when' in cascades. 10 2018.
- [30] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network.
- [31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [32] Yu-Shuai Li, Da-Zhong Ma, Hua-Guang Zhang, and Qiu-Ye Sun. Critical nodes identification of power systems based on controllability of complex networks. *Applied Sciences*, 5(3):622–636, 2015.
- [33] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

- [34] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions i. *Mathematical Programming*, 14(1):265–294, 1978.
- [35] Elaine O Nsoesie, Richard Beckman, Madhav Marathe, and Bryan Lewis. Prediction of an epidemic curve: A supervised classification approach. *Statistical communications in infectious diseases*, 3(1), 2011.
- [36] Min Ouyang. Review on modeling and simulation of interdependent critical infrastructure systems. *Reliability Engineering System Safety*, 121:43 – 60, 2014.
- [37] Marzieh Parandehgheibi and Eytan Modiano. Robustness of bidirectional interdependent networks: Analysis and design. *arXiv preprint arXiv:1605.01262*, 2016.
- [38] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200–3203, 2001.
- [39] Arunabha Sen, Anisha Mazumder, Joydeep Banerjee, Arun Das, and Randy Compton. Identification of k most vulnerable nodes in multi-layered network using a new model of interdependency. In *INFOCOM WKSHPs*, pages 831–836. IEEE, 2014.
- [40] Jeffrey Shaman, Edward Goldstein, and Marc Lipsitch. Absolute humidity and pandemic versus epidemic influenza. *American journal of epidemiology*, 173(2):127–135, 2010.
- [41] Farzaneh Sadat Tabataba, Prithwish Chakraborty, Naren Ramakrishnan, Srinivasan Venkatramanan, Jiangzhuo Chen, Bryan Lewis, and Madhav Marathe. A framework for evaluating epidemic forecasts. *BMC infectious diseases*, 17(1):345, 2017.
- [42] James D Tamerius, Jeffrey Shaman, Wladimir J Alonso, Kimberly Bloom-Feshbach, Christopher K Uejio, Andrew Comrie, and Cécile Viboud. Environmental predictors of

- seasonal influenza epidemics across temperate and tropical climates. *PLoS pathogens*, 9(3):e1003194, 2013.
- [43] Norman T.J Bailey. The mathematical theory of infectious diseases and its applications. *SERBIULA (sistema Librum 2.0)*, 34, 01 1975.
- [44] Hanghang Tong, B Aditya Prakash, Charalampos Tsourakakis, Tina Eliassi-Rad, Christos Faloutsos, and Duen Horng Chau. On the vulnerability of large graphs. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 1091–1096. IEEE, 2010.
- [45] U.S. Department of Energy. The Water-Energy Nexus: Challenges and Opportunities, June 2014.
- [46] Siva R Venna, Amirhossein Tavanaei, Raju N Gottumukkala, Vijay V Raghavan, Anthony Maida, and Stephen Nichols. A novel data-driven model for real-time influenza forecasting. *bioRxiv*, page 185512, 2017.
- [47] Svitlana Volkova, Ellyn Ayton, Katherine Porterfield, and Courtney D Corley. Forecasting influenza-like illness dynamics for military populations using neural networks and social media. *PloS one*, 12(12):e0188941, 2017.
- [48] Qian Zhang, Nicola Perra, Daniela Perrotta, Michele Tizzoni, Daniela Paolotti, and Alessandro Vespignani. Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model. In *Proceedings of the 26th International Conference on World Wide Web*, pages 311–319. International World Wide Web Conferences Steering Committee, 2017.