# INFLUENCE NETWORK ANALYSIS ON SOCIAL NETWORK AND CRITICAL INFRASTRUCTURE INTERDEPENDENCIES

A Dissertation
Presented to
The Academic Faculty

by

Zixing Wang

In Partial Fulfillment
of the Requirements for the Degree
Ph.D. in Operations Research in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
DECEMBER 2020

# INFLUENCE NETWORK ANALYSIS ON SOCIAL NETWORK AND CRITICAL INFRASTRUCTURE INTERDEPENDENCIES

Approved by:

Dr. Chelsea White, co-Advisor
School of Industrial and Systems Engineering
*Georgia Institute of Technology*

Dr. Rachel Cummings, co-Advisor
School of Industrial and Systems Engineering
*Georgia Institute of Technology*

Dr. David Goldsman
School of Industrial and Systems Engineering
*Georgia Institute of Technology*

Dr. Andy Sun
School of Industrial and Systems Engineering
*Georgia Institute of Technology*

Dr. Marilyn Brown
School of Public Policy
*Georgia Institute of Technology*

Date Approved: 8/19/2020

# ACKNOWLEDGEMENTS

First of all, my deepest gratitude goes to my current advisors, Dr. Chip White and Dr. Rachel Cummings, who assisted and guided me during my last period in my Ph.D. study.

I also would like to thank Dr. Eva Lee, who advised me in most of my research.

I also acknowledge all my committee members for the invaluable opinions and suggestions.

Finally, to my wife and my parents who backed and assisted me always. It is my greatest happiness to have you all.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF SYMBOLS AND ABBREVIATIONS

| | |
|---|---|
| CI | Critical Infrastructures |
| CIC | Competitive Independent Cascade |
| CLT | Competitive Linear Threshold |
| GC | General Cascade |
| GT | General Threshold |
| IC | Independent Cascade |
| LT | Linear Threshold |
| MIN | Multi-layer Interdependency Network |
| MLTIN | Multi-layer Linear Threshold Network |
| WGTM | Weighted General Threshold Model |

# SUMMARY

Inspired by the social network, the influence network has been proved to be a powerful tool to analyze the influence propagation within a group of entities. An introduction of the topic is given in Chapter 1. In Chapter 2 of the thesis, a brief survey on some major results on the single-layer influence network analysis is presented, and we propose a new multi-layer influence network framework. In Chapters 3 and 4, we give two applications of the single-layer influence network, on social networks and cellular base station interdependency networks. In Chapter 5, we propose a new multi-layer linear threshold influence network to analyze the interdependency of critical infrastructure sectors in metro Atlanta and Florida. Summary and conclusions are presented in Chapter 6.

# CHAPTER 1. INTRODUCTION

The social network has been of interest to researchers for years. It is made of the social activity participants and the dyadic ties between select pairs these participants. An interesting related problem is that given a social network and we want to maximize the dissemination scope of some certain message, what is the best strategy? This problem, defined by Domingos and Richardson (2001) [22], Richardson and Domingos (2002) [51] and Kempe et al. (2003) [34], is known as the influence maximization problem.

Our thesis starts by looking into this problem. First, to address this problem mathematically, it is necessary for us to define some basic tools to describe the social network. Thus, we introduce the concept of graphs, which is used for almost all network analysis, before we start further discussion. A graph is a collection of objects and their relationships. Usually we call the objects in a graph vertices(or nodes) and their relationship as edges (or arcs). Definition 1.1 defines the graphs and related concepts mathematically.

**Definition 1.1 (Graphs).** Given a collection of objects $V$, a graph is a pair $G = (V, E)$ that includes the vertices set $V$ and the arcs set $E \subseteq V \times V$. If $E$ is unordered, then the graph is undirected; otherwise it is directed.

When applying the graph concepts to the social network, it is natural to assume that every vertex represents a social participant and every arc is the relationship between participants. The relationship has multiple interpretations. In the simplest case, it can represent the observed connections between participants, like the friends on Facebook or fans on Instagram. If we want to raise the relationship bar a little bit, we could let the arcs represent the capability of one participant to influence another, i.e. when the influencer takes some actions or reveals some opinions, it might cause others to do the same. In this dissertation, we would use the latter one as our standard to build arcs.

Before we step into maximizing influence, it is necessary to define formal influence spreading models that clearly describe how influences disseminate step by step. With the graph defined above, there are two issues left to address. The first issue is that obviously different pair of participants in the network would have different influencing capabilities. Thus it is necessary to assign a weight or probability to each arc to measure this capability, and such a weight or probability needs to have an appropriate interpretation. The second issue is that how to define how influence spreads on the weighted graph. After addressing these two issues, we can measure influence and maximize it. In Chapter 2, we will review the related literature and try to summarize a framework to address these two issues. In Chapter 3, we will study a church network scenario where we apply the influence network models.

Although the influence network model is designed for social network analysis, the concept of influence is not limited to only information and social influence. In Chapter 4, we apply the influence network models to analyze the interdependencies of cellular base stations, which are part of the critical infrastructure (CI) facilities. In this scenario, every node in the graph is a cellular base station and the influence is interpreted as the cascading effects generated from the failure of the originating stations. In Chapter 5, we further extend the framework to analyze multi-layer CI interdependencies. Given a set of CI facilities, we first classify them into sectors which represent energy, communication, transportation, etc.

Next, from the sectors we will define corresponding layers that include all the nodes affected by each sector. Then we will state how to construct influence network models on each layer and how they interact with each other. Finally, two examples for metro Atlanta, Georgia and the state of Florida are analyzed.

# CHAPTER 2. FUNDAMENTALS OF INFLUENCE NETWORK

The influence maximization problem, as described in the previous chapter, is important since it can support decision making in interfering social networks and other networks representing influence dissemination. In this chapter, we will review some major approaches to address and solve this problem. In the first section, we will give a formal definition of the problem and introduce two major influence network models. In the second section, we will review some major approaches to solve this problem based on the models in Section 1. Finally, in Section 3 we will review other notable methods for solution determination.

We begin by defining some useful graph notation. For directed graph $G = (V, E)$, we let $N^{in}(v) = \{u|(u, v) \in E\}$ and call it the in-nodes of $v$. The arcs ending in $v$ are called the in-arcs of $v$. The out-nodes and out-arcs are defined in the same manner.

## 2.1 Influence Network Models

We first explain an influence diffusion in Definition 1.1. Intuitively, given a directed graph $G = (V, E)$, there is an initial set which is influenced before any other nodes. This set is called the seed set and usually denoted as $S_0 \subseteq V$. In this thesis, we assume that all sets $V$ are finite. We assume that influence spreads from the seed set in discrete time steps; i.e. for step $t > 1$, there exists a node set series $\{S_t\}$ and $S_t \subseteq V, \forall t,$ where $S_t$ represents the influenced node set at time $t$. Influence diffusion models are graph models that explain how to determine $\{S_t\}$, see definition 2.1.

**Definition 2.1 (Influence diffusion models).** Given a finite directed graph $G = (V, E)$, influence diffusion models explain the scheme to decide node set series $\{S_t\}$, $S_t \subseteq V, \forall t$.

If $S_0 \subseteq S_1 \subseteq \ldots \subseteq S_t \subseteq \ldots \subseteq V$, then the influence diffusion model is called progressive. We mainly focus on progressive influence diffusion models in this thesis, but some non-progressive models will be reviewed in Section 2.3. If node $v \in S_t$, we call $v$ active at step $t$, otherwise we say it's inactive. Notice that for progressive models, since the graph is finite and $\{S_t\}$ is non-decreasing, series $\{S_t\}$ must converge. We let $S_\infty \equiv \lim_{t \to \infty} S_t$ and call $S_\infty$ the final active set.

The influence network models are special cases of the influence diffusion models in terms of specifying the process of deciding $\{S_t\}$. However, for social networks there is no decisive opinion of how people choose to accept new ideas and opinions because everyone might have his own criterion. Here we focus on two major influence network models proposed by Kempe et al. (2003) [34], which represent probably the most natural assumptions of the influence network. In the first model, the independent cascade model, if $u \in S_{t-1} \backslash S_{t-2}$, $(u, v) \in E$, and $v$ is inactive, then with a fixed probability $p(u, v)$, $v$ is active in step $t$. When multiple active source nodes try to influence an inactive node, the influencing between sources are independent. The reason we let $u \in S_{t-1} \backslash S_{t-2}$ rather than $u \in S_{t-1}$ is that every node can only activate its out-neighbors once, which happens in the next step after it becomes active. The influencing pattern assumption for this model is that when hearing some new information from others, some people would rely on the reasoning process to decide whether to accept it or not, different information sources would be independent since the reasoning is likely to vary.

In the second model, the linear threshold model, every node would choose a random variable $\theta_v$ which is uniformly distributed on [0,1] as its activating threshold. The node would be influenced if the sum of the weights of the incoming arcs from active nodes exceeds the threshold. This assumption reflects the situation that some people would accept the information if a certain number of close people have accepted it. The formal definitions are stated below.

**Definition 2.2 (Independent Cascade (IC) model).** Given a finite directed graph $G = (V, E)$, where $V$ denotes the vertices set and $E$ denotes the directed arcs(edges) set. Given the influence probability $p(u, v) \in [0,1]$ assigned to every arc, and the initial seed set $S_0 \in V$ as the input, generates the active sets $S_t$ for some $t \geq 1$ by the following randomized operation rule: At each $t$ such that $t \geq 1$, for every inactive nodes $v \in V \backslash S_{t-1}$, each node $u \in N^{in}(v) \cap (S_{t-1} \backslash S_{t-2})$ , $u$ would initate an activation attempt to v in the form of a Bernoulli trial with probability $p(u, v)$. If successful, $v$ would be added to $S_t$.

**Definition 2.3 (Linear Threshold (LT) model).** Given a finite directed graph $G = (V, E)$, where $V$ denotes the vertices set and $E$ denotes the directed arcs set, given the weight $w(e)$ assigned to every arc, and the initial seed set $S_0$ as the input, generates the active sets $S_t$ for some $t \geq 1$ by the following randomized operation rule: Initially, each node $v \in V$, independently selects a threshold $\theta_v \sim U[0,1]$, where $U(0,1)$ is the uniform distribution on interval $(0,1)$. At every time step $t \geq 1$, first set $S_t$ to be $S_{t-1}$; then for any inactive node $v \in V \backslash S_{t-1}$, if the total weight of the arcs from its active in-neighbors is at least $\theta_v$, i.e. $\sum_{u\in N^{in}(v)\cap S_{t-1}} w(u,v) \geq \theta_v$, then add $v$ to $S_t$ . It is required that $\forall v \in V$, $\sum_{u\in N^{in}(v)} w(u,v) \leq 1$.

For any progressive influence diffusion models, we can define an influence function $I(S_0): S_0 \to \mathbb{R}^+$ to summarize a positive number from the final active set $S_\infty$. Without special notation we will just use the expectation of the cardinality of the final active set as the influence function, i.e. $I(S_0) = \boldsymbol{E}(card(S_\infty)|S_0)$. If there are weights assigned for each node, then $I(S_0) = \boldsymbol{E}(\sum_{v\in S_\infty} w(v)\,|S_0)$, where $w: V \to \mathbb{R}_+$ is a weight function.

In the work of Kempe et al. (2003) [34], he also summarized two generalized models from two influence network models above. The general cascade model is a generalization of the independent cascade model, it replaced the influence probability $p(u,v)$ by a function $p_v(u,S): N^{in}(v) \times 2^{N^{in}(v)} \to [0,1]$, which means the influence probability of each in-node depends on the activated in-node set. The generalization of the linear threshold model is the general threshold model, which replace the activation condition $\sum_{u\in N^{in}(v)\cap S_{t-1}} w(u,v) \geq \theta_v$ by a general function $f_v: 2^{N_{in}(v)} \to [0,1]$. The definitions are given below.

**Definition 2.4 (General Cascade (GC) model).** Given a finite directed graph $G = (V,E)$, where $V$ denotes the vertices set and $E$ denotes the directed arcs set, and the initial seed set $S_0 \in V$ as the input, Every node $v$ has an activation function $p_v(u,S): N^{in}(v) \times 2^{N^{in}(v)} \backslash \{u\} \to [0,1]$, where generates the active sets $S_t$ for some $t \geq 1$ by the following randomized operation rule: At each $t$ such that $t \geq 1$, for every inactive nodes $v \in V \backslash S_{t-1}$, let $\{u_1,\ u_2,\ \dots\ u_n\} = N^{in}(v) \cap (S_{t-1} \backslash S_{t-2})$ , $u_1$ would first try to activate $v$ with probability $p_v(u_1, S)$, where $S = N^{in}(v) \cap S_{t-2}$. If $\{u_1,\ u_2,\ \ldots\ u_l\}$ fails to activate $v$ for some $l < n$, then $u_{l+1}$ would try to activate $v$ with probability $p_v(u_{l+1}, S \cup \{u_1,\ u_2,\ \ldots\ u_l\})$. The activation attempts are independent and if any nodes in $\{u_1,\ u_2,\ \ldots\ u_n\}$ activates $v$ then the $v$ is activated. The function $p_v(u, S)$ must be order-independent, which means for any $N^{in}(v),\ S_{t-1},\ S_{t-2}$ and $v$, the order of $N^{in}(v) \cap (S_{t-1} \backslash S_{t-2})$ should not change the final activation probability of $v$.

**Definition 2.5 (General Threshold (GT) model).** For a directed graph $G = (V, E)$, every node is assigned an activation function $f_v : 2^{N_{in}(v)} \rightarrow [0,1]$, $f_v$ is monotone and $f_v(\emptyset) = 0$. Each node $v$ select a threshold $\theta_v$ uniformly from [0,1], the thresholds are independent for different nodes. For a given initial set $S_0$, at every step $t \geq 1$, first set $S_t$ to $S_{t-1}$, for every node $v \in V \backslash S_{t-1}$, if $f_v\left(S_{t-1} \cap N_{in}(v)\right) \geq \theta_v$, then add $v$ to $S_t$.

Here we introduce another important concept, model equivalence, which is defined in definition 2.6 (Kempe et al. (2003) [34]). Two influence network models are equivalent if for any $t \geq 1$, if the active sets at every previous time step are the same, the distributions of the active set at time $t$ for two models are identical. This concept is important because in later sections we will see that it is easy to simulate the model or prove some properties of the model if we can find a simpler equivalent model for it.

**Definition 2.6.** For two influence network model, they are equivalent if both the conditions A and B are met:

**A**. For any $A_0, A_1, \ldots, A_t \subseteq V$, $Pr(S_1 = A_1, \ldots, S_t = A_t | S_0 = A_0)$ are zero in both models or nonzero in both models.

**B**. For any $A_0, A_1, \ldots, A_t \subseteq V$, $Pr(S_t = A_t | S_0 = A_0, S_1 = A_1, \ldots, S_{t-1} = A_{t-1})$ are the same in both models.

At the end of this chapter, we point out that the GC and GT model are equivalent Kempe et al. (2003) [34]. In fact, for GC model with activation function $p_v(u, S)$ and $S = \{u_1,\ u_2,\ \ldots\ u_n\}$, it is equivalent with GT model with activation function $f_v(S) = 1 - \prod_{i=1}^{n}(1 - p(u_i, A_{i-1}))$, where $A_i = \{u_1,\ u_2,\ \ldots\ u_i\}$ and $A_0 = \emptyset$. Conversely, for GT model with activation function $f_v(S)$, it is equivalent to a GC model with activation function $p_v(u, S) = \frac{f_v(S \cup \{u) - f_v(S)}{1 - f_v(S)}$ if $f_v(S) < 1$, if $f_v(S) = 1$, then $p_v(u, S)$ can be any number in [0,1] since $S$ is enough to activate $v$.

## 2.2 Properties of Influence Network Models and Influence Maximization

In this section we will discuss how to maximize the influence function for the models defined in the previous section. Firstly, we define the influence maximization problem:

**Definition 2.7 (Influence maximization).** Given an influence network model $G$, and a positive integer $k$, the influence maximization problem is max $I(S_0), s.t.\ |S_0| = k$.

Based on this definition, clearly there are influence maximization can be decomposed into two problems: how to calculate $I(S_0)$ and how to maximize it. We first visit the first question, since the influence function is determined by the final active set $S_\infty$ given $S_0$, it is sufficient to find methods to decide $S_\infty$. However, it is NP-hard to calculate $\boldsymbol{E}(\mathbf{1}_{\{v \in S_\infty\}} | S_0)$ for any $v \notin S_0$ for any influence network models introduced (Chen et al. (2010) [17], Wang et al. (2012) [58]) and thus calculating any influence function in the form of $\boldsymbol{E}(w(S_\infty) | S_0)$ is NP-hard. Researchers have come up with some heuristics to solve this problem. For IC and LT model, Kempe et al. (2003) [34] proposed **live-arc graphs** on which we can use Monte-Carlo method to simulate the influence function. The definitions are given below.

**Definition 2.8 (Live-arc graph of IC model).** given an IC influence network $G$=($V$,$E$) and arc influence probability $p(e)$. For every $e \in E$, keep the arc with probability $p(e)$, otherwise remove it. The resulting graph is called a live-arc graph of $G$.

**Definition 2.9 (Live-arc graph of LT model).** given an LT influence network $G = (V, E)$: For each $v \in V$, select one arc from all the arcs ending in $v$ with probability distribution $w(u, v)$ (no arc is chosen with probability $1 - \sum_{u \in N^{in}(v)} w(u, v)$ if $\sum_{u \in N^{in}(v)} w(u, v)$<1) . The resulting graph is called a live-arc graph of $G$.

Kempe et al. (2003) [34] proved that the live-arc graphs are equivalent to their corresponding influence network models. (see definition 2.6 for model equivalence). To estimate the influence function value of a given initial active set $S_0$, we use Monte-Carlo simulation, In each Monte-Carlo simulation round, an instance of live-arc graphs is generated and used to find the final active set. And using the set we can find the influence function value for this Monte Carlo iteration. After desired rounds of Monte Carlo simulation are done, the average influence function value is used as the final estimate.

However, the Monte Carlo simulation is very time-consuming for large graphs. Thus non-simulation-based heuristics are necessary for large-scale problems. For IC model, Wang et al. (2012) [58] proposed an Maximum Influence Arborescence (MIA) algorithm in which they trimmed all the paths with influence probability lower than a threshold.

Specifically, they define the maximum influence path from $u$ to $v$ as the path with highest influence probability among all the simple paths from $u$ to $v$, denoted by $MIP(u, v)$. Then they define the maximum influence in- arborescence $MIIA(v, \lambda) = \cup_{u \in V, pr(MIP(u,v)) > \lambda)} MIP(u, v)$, which is the collection of the maximum influence paths which has influence probability higher than $\lambda$. For any maximum influence path from $u$ to $v$, it must contain the maximum influence path from any intermediate nodes on this path to $v$. Thus MIIA is actually a forest where all trees ends in $v$. In this way to estimate the influence probability from any given set $S_0$ to $v$, we just need to start from $S_0$ and do a recursive calculation of the influence probability on each downstream node of $S_0$ on the MIIA of $v$ until the calculation ends in $v$. This method has no guarantee how far the results diverge from the real value, but it is very useful since it does not use Monte Carlo simulation.

For LT model, Goyal et al. (2011) [30] proposed the SIMPATH algorithm. In fact, for a LT model on graph $G = (V, E)$ and arc weights $w(u, v)$, Let $r_{S,v}^{V}$ denote the probability that given whole node set $V$ and initial set $S$, $v$ is influenced. They proved that if $u$,$v$ are two nodes in $G$, $SP(u, v)$ are the set of all simple path from u to v in G (i.e. no loop), then $r_{\{u\},v}^{V} = \sum_{P \in SP(u,v)} w\ (P)$, where $w\ (P)$ is the product of the arc weights on $P$. Next, they prove that $I(S) = \sum_{u \in S} I^{V-S+\{u\}}(u)$, where $I(S)$ is the cardinality influence function on the original graph and $I^{V-S+\{u\}}(u)$ is the same influence function but on graph with $V - S + \{u\}$ as the node set and the original arcs between those nodes as the arc set. Finally, clearly we have $I^{V-S+\{u\}}(u) = \sum_{v \in V-S+\{u\}} r_{\{u\},v}^{V-S+\{u\}}$. This means that to calculate $I(S)$, we only need to find $I^{V-S+\{u\}}(u)$ for all $u \in S$, and $I^{V-S+\{u\}}(u)$ can be calculated by summing up $r_{\{u\},v}^{V-S+\{u\}}$ using $r_{\{u\},v}^{V} = \sum_{P \in SP(u,v)} w\ (P)$. However, since it is NP-hard to find all the simple paths between any two nodes in a general graph, it is necessary to set a threshold as a lower bound so that any simple paths with $w\ (P)$ smaller than the threshold are not considered.

So far we reviewed the algorithms and heuristics to calculate the influence function. Before stepping into the influence maximization of the previous influence network models, we introduce some important concepts for general set function optimization, which are useful for understanding the influence maximization algorithms introduced later. The first concept is the submodularity of set functions.

**Definition 2.10 (Submodularity of set functions).** Given a set function $f: S \rightarrow \mathbb{R}$, where $S \subseteq U$ is a set and $U$ is the universal set, the function is submodular if $\forall A \subseteq B \subseteq U$, and $v \in U \backslash B$, the following inequality holds: $f(A \cup \{v\}) - f(A) \geq f(B \cup \{v\}) - f(B)$.

Submodularity is a generalization of the convex function on $\mathbb{R}^n$, it means that adding an element to a small set would have more increments on the function value than adding the element to a larger set. Kempe et al. (2003) [34] proved that for both IC and LT influence network models, the influence function is nonnegative, non-decreasing and submodular. He also proposed a plain Greedy method for influence maximization. The method starts with $t = 0\ and\ S_t = \emptyset$.For $t \geq 1,$ it adds the node $\{argmax_x\ (f(\{x\} \cup S_{t-1}) - f(S_{t-1}))|x \in U - S_{t-1}\}$ to $S_{t-1}$ and let $S_t = S_{t-1}\ \cup \{x\}$, and repeats until the desired cardinality is reached. Nemhauser et al. (1978) [44] proved that for any nonnegative monotone submodular set function $f()$, a positive integer $k$ and the corresponding maximization problem max $f(S_0),\ s.t.\ |S_0| = k$. Let $S$ be the final set chosen by the

Greedy method and let $S^*$ be the optimal set. The following inequality must hold: The following inequality must hold: $f(S) \geq (1 - 1/e)f(S^*)$, where $e$ is the base on natural logarithm. This theorem provides a lower bound when applying the plain Greedy method to maximize any non-negative monotone submodular function.

In the plain greedy method we must evaluate the increments of adding each unchosen nodes to the chosen set, but Leskovec et al. (2007) [39] and Goyal et al. (2011) [29] argued that most of the evaluations are not necessary and they improved the plain greedy algorithm to the Lazy Greedy algorithm. In fact, by the property of submodular function, if $S' \subseteq S$, and $x, w \in V - S$, we must have $f(\{x\} \cup S) - f(S) \leq f(\{x'\} \cup S) - f(S')$. Thus, if we know $f(x, S') - f(S') \leq f(w, S) - f(S)$ , then we don't have to evaluate $x$ with previously chosen set $S$. In Greedy algorithm, $S'$ could be the set chosen at step $k$ and $S$ could be the set chosen at step $k + 1$. Therefore after evaluating $f(w, S)$, we will check if there is any $x,\ s.t.\ f(x, S') \leq f(w, S)$, where $S' \subseteq S$ is a intermediate set produced by Greedy method. Any such $x$ would be removed from candidate list to improve performance. The details of this Lazy Greedy algorithm is summarized in Algorithm 1.

**Algorithm 1.** The Lazy Greedy algorithm for the influence maximization

```
Input: $k$: size of returned set; $f$ :influence function
Output: selected subset with cardinality $k$
initialize $S = \emptyset$; priority queue $Q = \emptyset$; iteration= 1
for $i$ = 1 to n do
        $u.mg$=$f(u)$ (use Monte-Carlo or other heuristics here) , $u.i = 1$;
        insert element $u$ into $Q$ with $u.mg$ as the key
end for
while iteration $\leq$ $k$ do
       extract top (max) element $u$ of $Q$
       if $u.i$= iteration then
          $S = S \cup \{u\}$ ; $iteration = iteration + 1$;
      else
          $u.mg = f(u \cup S)$; $u.i = iteration$;
          re-insert $u$ into $Q$
      end if
end while
return $S$
```

## 2.3 Extensions of the Influence Network Models and Other Approaches for the Influence Maximization

So far we have reviewed major influence network models developed by previous researchers which will be used in all the following chapters. In the first part of this section, we introduce some extensions of the influence network models that we will use in the following chapters. In the second part, we review other models which try to solve similar problems. We will focus the advantages and defects of these models instead of the implementation details.

### *2.3.1 Extensions of the Influence Network Models*

In the influence maximization problem, we have limited resources to put influences on the initial active set $S_0$ and that is the reason why the size of $S_0$ is given and we try to decide which nodes to be selected for the active set. On some circumstances, we have a target on the final active set $S_\infty$ and the question is to decide the minimum cardinality and the elements of the initial active set $S_0$. This problem is called the minimum target set selection problem (MINTSS) by Goyal et al. (2013) [28]. Definition 2.11 defines the MINTSS problem.

**Definition 2.11 (Minimum target set selection problem).** Given an influence network model $G$ , and a positive number $r$, the minimum target set selection problem is min $card(S_0), s.t.\ I(S_0) \geq r$, where $card(S_0)$ is the cardinality function.

The MINTSS problem on IC and LT models with weighted influence function $I(S_0) = \boldsymbol{E}(\sum_{v \in S_\infty} w(v)\,|S_0)$ can be solved using a greedy method which is similar with the one we used in the influence maximization problem (Goyal et al. (2013) [28]). In iteration $t+1$ of the greedy method, the method adds $argmax_{v \in V-S_t} \frac{min(I(S_t \cup \{v\},\ r) - I(S_t)}{w(v)}$ to the selected set $S_t$ at iteration $t$ and let the new set be $S_{t+1}$, repeat until the desired influences $r$ is reached.

Another important model which we will use in chapter 3 is the competitive influence network models. The competitive version of IC and LT models are called CIC models and CLT models respectively (Borodin et al. (2010) [9], Budak et al. (2011) [11] and Chen et al. (2011) [16]). In these models every node has three states: positive, negative and inactive. The positive and negative influences follow the IC and LT mechanism to spread and every inactive node would only be influenced once. The definitions of CIC and CLT models are given below.

**Definition 2.12 (Competitive independent cascade (CIC) model).** Given a finite directed graph $G = (V, E)$ , the positive and negative influence probability $p^+(u, v),\ p^-(u, v) \in [0,1]$ for every arc, and the initial seed set $S_0^+,\ S_0^- \in V$ as the input, generates the active sets $S_t^+$ and $S_t^-$ for some $t \geq 1$ by the following randomized operation rule: At each $t$ such that $t \geq 1$, for every inactive nodes $v \in V \backslash S_{t-1}^+$, each node $u \in N^{in}(v) \cap (S_{t-1}^+ \backslash S_{t-2}^+)$ , $u$ would initate an activation attempt to v in the form of a Bernoulli trial with probability $p^+(u, v)$. Also, for every inactive nodes $v \in V \backslash S_{t-1}^-$, each node $u \in N^{in}(v) \cap (S_{t-1}^- \backslash S_{t-2}^-)$ , $u$ would initate an activation attempt to v in the form of a Bernoulli trial with probability $p^-(u, v)$. If only the positive attempt is successful, $v$ would be added to $S_t^+$, if only the negative attempt is successful, $v$ would be added to $S_t^-$. If both attempts are successful, a tie-breaking rule, which will be described later, will be used to decide which set $v$ is added to.

**Definition 2.13 (Competitive linear threshold (CLT) model).** Given a finite directed graph $G = (V, E)$, the positive and negative weights $w^+(u, v),\ w^-(u, v) \in [0,1]$ for every arc, and the initial seed set $S_0^+,\ S_0^- \in V$ as the input, generates the active sets $S_t^+$ and $S_t^-$ for some $t \geq 1$ by the following randomized operation rule: Initially, each node $v \in V$, independently selects two threshold $\theta_v^+,\ \theta_v^- \sim U[0,1]$, where $U(0,1)$ is the uniform distribution on interval $(0,1)$. At every time step $t \geq 1$, first set $S_t^+$ to be $S_{t-1}^+$ and $S_t^-$ to be $S_{t-1}^-$; then for any inactive node $v \in V \backslash (S_{t-1}^+ \cup S_{t-1}^-)$, if the total weight of the arcs from its positive in-neighbors is at least $\theta_v^+$ and the total weight of the arcs from its negative in-neighbors is less than $\theta_v^-$ , i.e. $\sum_{u \in N^{in}(v) \cap S_{t-1}^+} w^+(u, v) \geq \theta_v^+$ and $\sum_{u \in N^{in}(v) \cap S_{t-1}^-} w^-(u, v) < \theta_v^-$, then add $v$ to $S_t^+$. If $\sum_{u \in N^{in}(v) \cap S_{t-1}^+} w^+(u, v) < \theta_v^+$ and

$\sum_{u\in N^{in}(v)\cap S_{t-1}^-} w^-(u,v) \geq \theta_v^-$, then add $v$ to $S_t^-$. If both sum of weights are greater than the threshold then a tie-breaking rule is needed. It is required that $\forall v \in V$, $\sum_{u\in N^{in}(v)} w^+(u,v) \leq 1$ and $\sum_{u\in N^{in}(v)} w^-(u,v) \leq 1$.

As we explained in the definition, it is necessary to assume a tie-breaking rule when a node is activated simultaneously by both positive and negative influences. A type of tie-breaking rule is the fixed probability tie-breaking rules, which assign a fixed probability to the make the inactive nodes positive when both attempts succeed Borodin et al. (2010) [9]. When the probability is 1/0, the rule is called positive/negative dominance. Another natural tie-breaking rules is by the relative power of the positive and negative attempts (Chen et al. (2011) [16]), in CIC model, the positive activating probability is $\frac{\sum_{u\in N^{in}(v)\cap(S_{t-1}^+\setminus S_{t-2}^+)} p^+(u,v)}{\sum_{u\in N^{in}(v)\cap(S_{t-1}^+\setminus S_{t-2}^+)} p^+(u,v)+\sum_{u\in N^{in}(v)\cap(S_{t-1}^-\setminus S_{t-2}^-)} p^-(u,v)}$ while in CLT model positive activating probability is $\frac{\sum_{u\in N^{in}(v)\cap S_{t-1}^+} w^+(u,v)}{\sum_{u\in N^{in}(v)\cap S_{t-1}^+} w^+(u,v)+\sum_{u\in N^{in}(v)\cap S_{t-1}^-} w^-(u,v)}$. This rule is called proportional tie-breaking rule.

The similar live-arc graphs could be defined for both CIC and CLT models, in fact, in both models we can create two separate live-arc graphs using the definition 2.8 and 2.9 for positive influences and negative influences. The only problem is that two live-arc graphs might share some nodes, in this case we just check each shared node in which time step it is activated and assign it to the group that activate it first. If it is activated by both influences simultaneously, the tie-breaking rule is used to decide the assignment.

There are two influence functions defined for the CIC and CLT models, the positive and negative influence functions. As in the non-competitive case they are defined by the weight function of the final active set: $I^+(S_0^+,\, S_0^-) = \boldsymbol{E}(\sum_{v\in S_\infty^+} w(v)\,|S_0^+,\, S_0^-)$ and $I^-(S_0^+,\, S_0^-) = \boldsymbol{E}(\sum_{v\in S_\infty^-} w(v)\,|S_0^+,\, S_0^-)$. It is clear that $I^+(S_0^+,\, S_0^-)$ is non-decreasing with respect to $S_0^+$ given $S_0^-$ and $I^-(S_0^+,\, S_0^-)$ is non-decreasing with respect to $S_0^-$ given $S_0^+$. Budak et al. (2011) [11] gave the counter-example that such functions are generally not submodular. Despite this, they showed that for CIC model where $p^+(u,v) = p^-(u,v)$ for all arcs, and the tie-breaking rule is positive-negative dominance or proportional, the influence functions are submodular with respect to its corresponding initial active set. That is to say, the Algorithm 1 is applicable with the lower bound in this case. We will see the applications of such model in chapter 3.

### *2.3.2 Other Approaches for Influence Maximization*

The first model to introduce is the epidemic models. The epidemic models are used for study the epidemic disease in a certain area. Every individual in the system has at least two nodes, susceptible and infected, the simplest model that only involves these two states is called the SI model, if considering the recovery from the disease, the model is called the SIR model where a new recovered state is added (Kermack and McKendrick (1927) [35]). In these models, the number of individuals in these three states, denoted by $s, i$ and $r$, are interconnected by a system of ordinary differential equations related to the infection rate, recovery rate and time (Capasso and Serio (1978) [14], Ruan and Wang (2003) [55], Xiao and Ruan (2007) [59]). It is assumed in these models everyone in the system has interactions with all others, which means no specific connections between any pair of individuals. In this case, the model cannot identify different connection patterns of individuals. For example, if we know some individuals have more capabilities to influence others and we know who he might influence, this feature cannot be reflected by the analytical epidemic model. Another type of epidemic model is the agent-based simulation models (Parker (2007) [48], Bobashev et al. (2007) [8]), in these models the position and behavior of the individuals, which are treated as particle, can be modified but large-scale agent-based simulation consumes plenty of computational resources and usually lack analytical backgrounds.

Another tool to solve such problems is the Markov random fields (Domingos and Richardson (2001) [22], Richardson and Domingos (2002) [51]). In this model every node in an undirected graph is assigned a binary variable $X_v$ where 1 represent activated and 0 otherwise. The model resembles the influence network models since it requires that all $X_v$ are conditionally independent with non-neighbors given their neighbors, The core question is how to find the value of $Pr(X_v = 1|S_0)$ to get the expectation of the number of nodes influenced. Note that this is not a time-dependent model so that $S_t$ and $S_\infty$ do not exist. The probabilities $Pr(X_v = 1|S_0)$ are inferred recursively using the probabilities of neighbors of $v$. Since the model is built on an undirected graph, the connection between any pair of adjacent nodes is undirected, which restricts the model flexibility. To solve the model a system of linear or nonlinear equations needs to be analyzed, in some cases, the system is too complex to get a analytical solution.

Finally, the influence maximization problem can be modelled as a Markov decision process (MDP) (Yadav et al. (2015) [60]). MDP is a decision process which includes a state space $S$, an action space $A$, the transition probabilities $Pr(s_1|s_0, a_0)$ representing the probability that the state transited from $s_0$ to $s_1$ by taking action $a_0$ and the reward

$R(s_1|s_0, a_0)$ which means the reward from achieving state $s_1$ from $s_0$ using action $a_0$. For a influence maximization problem to select $k$ most influential nodes out of a graph with $n$ nodes, they let every state represents a possible state of all nodes, which includes $2^{\mathrm{n}}$ states, the actions are choosing $k$ nodes from $n$, and the reward $R(s_1|s_0, a_0) = I(s_1)$ which is the weighted influence function. The transition probabilities are not explicitly given but using the simulation or other heuristics to estimate, like we introduced before. To find the best $k$ nodes, they use a multi-armed bandit based approach to get the best action using the results of previous actions. The major challenges of this framework is that the state space and the action space are too big for large-scale problems.

# CHAPTER 3. A COMPUTATIONAL FRAMEWORK FOR INFLUENCE NETWORKS: APPLICATION TO CLERGY INFLUENCE IN HIV/AIDS OUTREACH

In this chapter, we introduce a sociology-based computational framework for independent cascade (IC) influence networks. The model construct is generic and is applicable to diverse social network analysis. We demonstrate its usage in calibrating the positive influence of church clergy in spreading HIV/AIDs information in a large metropolitan city. Five experiments are designed to contrast influence with respect to the interaction style between clergy and churchgoers. Competitive (CIC model) and non-competitive (IC model) knowledge dissemination are also analyzed. The generalized framework requires minimal regional data to establish the influence network. It provides useful policy insights for decision makers to determine effective avenues for information dissemination through community influencers.

## 3.1 Introduction

Infectious diseases such as HIV/AIDS, tuberculous and malaria pose challenges to global health. Although these contagious diseases have risen to the top of the international agenda in recent years, there remain major hurdles in combating and eradicating them effectively. In the United States alone, more than 1.2 million people are living with HIV, with 1 out of 8 unware of it (CDC (2018) [15]). Furthermore, statistics show that African Americans continue to bear the greatest burden of HIV: they represent 12% of the total population but account for 45% of the HIV diagnoses (CDC (2018) [15]).

The spread of many of these contagious diseases can be mitigated through changes in human behavior. It is well-known that strong social networks can encourage healthy behaviors.

Realizing the importance of early intervention, the Centers for Disease Control and Prevention (CDC) and public health leaders design HIV outreach programs for early preventive measures and treatment, especially for the vulnerable population. Unfortunately, the HIV/AIDS stigma still asserts a significant barrier for people to voluntarily seek disease prevention and treatment information. Thus, public health practitioners seek various strategies to reach out to high-risk individuals to overcome this prejudice.

Exploiting strong social networks can help disseminate knowledge and shape positive health behavior. In particular, religious community centers have long played significant roles in information dissemination. Hadaway and Marler (1998) [31] reported that about 20% of Americans go to church on a weekly basis. In a survey conducted in 2007, Khosrovani et al. (2008) [36] concluded that for the highly vulnerable population, the involvement of churches in providing information and education is very crucial.

Although the importance of churches in spreading disease prevention knowledge and reducing the stigma has long been recognized, few investigations have been carried out to calibrate the value of the aid from church leaders. With only limited resources for congregational HIV/AIDS education programs, it is beneficial to identify influential churches and clergymen and effective outreach efforts.

## 3.2 Related Works and Our Contribution

### *3.2.1 Relate Works*

Religion plays an important role in American life. Most Americans believe in a deity, three-fourths pray at least weekly, and more than half attend religious services at least monthly (Barkan (2010) [6]). Numerous studies focus on the influences churches and religious workers have on regular church attendants. Khosrovani et al. (2008) [36] conducted a survey on African American churches in the metro Houston area in 2007 and concluded that although the attitude of churches has evolved over the last 25 years, the real disposition towards HIV remains passive and negative. On the other hand, 90% of church participants thought that churches should be involved in educating their congregation about HIV/AIDS prevention, helping ease the anxiety of HIV/AIDS carriers, and engaging high-risk individuals in counselling and seeking appropriate medical tests to learn of their health status. Bluthenthal et al. (2012) [7] found that while HIV and public health workers sought assistance from clergy (in Los Angeles County), educational outreach about HIV awareness and reducing HIV stigma were not high priorities for most religious congregations. There are some positive findings as well. For example, Moore et al. (2012) [41] reported that church leaders (in North Carolina) employed various approaches to communicate with congregants about HIV issues. Religious leaders play an important role in society and can potentially have a broad impact on HIV education. With the aid of modern social media, clergy can communicate with many congregants simultaneously, and hence can potentially generate broad coverage and positive impact.

In chapter 2, we have already reviewed the theoretical results given the social network graph. However, when facing a real world problem, we must formulate a graph representing the social network from the data first before applying the influence network models. Within social science, early work of Newman (2001) [45] analysed scientific collaboration networks where every researcher corresponds to a node, and two nodes are connected if the researchers co-authored a paper together. Since the emergence of social networks and user data, there has been tremendous interest in the phenomenon of influence propagation (Romero et al. (2011) [54], Adar and Adamic (2005) [1], Domingos and Richardson (2001) [22]). Most of these studies require a social graph with edges labelled with probabilities of influence between users. Goyal et al. (2010) [27] investigated where and how probabilities of influence between users were established from real social network data. The authors built models of influence from a social graph and a log of actions by its users using the Flickr data set consisting of a social graph with 1.3 million nodes, 40 million edges, and an action log consisting of 35 million tuples associated with 300 thousand distinct actions.

Besides these data-based approaches, notable results have been achieved in the field of sociology. Using a subset of the British population, Dunbar and Spoors (1995) [24] and Hill and Dunbar (2003) [32] first established the three social circle layers of individuals: the support clique, the sympathy group and the general acquaintance. He and his co-authors discussed the number of people in each group based on the number of kin and personality of the individual. Stiller and Dunbar (2007) [56] also investigated how memory capacity and theory of mind are involved in determining the size of social circles.

### *3.2.2 Our Influence Network Design and Contributions*

In this paper, we utilize sociology theory to design a general-purpose framework of a computational influence network. The model is flexible and can accommodate any type of data model analysis and model objectives. We present a scalable computational algorithm for determining a set of key influencers who can assert the maximum influence/effect within the network.

To demonstrate its applicability, we apply our influence network model to the clergy HIV/AIDs education outreach in a large metropolitan city. The input includes a set of churches, the number of regular participants, and the regional HIV/AIDS infection rate, estimated by zip codes. Our model will rank the churches (hence their influence power) based on their capability of spreading HIV/AIDS information positively and successfully. The results show a tradeoff between choosing larger churches versus choosing the churches located at sites with higher HIV/AIDs infection rates.

The model can be applied to other contagious diseases and public health outreach, or in the analysis of news and/or information spreading (rumors or facts). Our approach offers some novel features: (i) While most research related to social network construction focuses on collecting the data from online sources and building the network based on the data, our approach does not rely on specific data. This enables modeling of the social network impact among all groups, including those that are not as active online. (ii) Although the importance of church for spreading information in certain communities has been recognized, there is little research focusing on calibrating its importance mathematically. Our study facilitates policy-and decision making.

## 3.3 Methodology

In this section, we first present the construction of a generic influence network model. The social network is a generalized framework that incorporates some fundamental sociology features. The influence model is then applied to calibrate the importance of church leaders in the HIV/AIDS outreach. In the model, we estimate the HIV/AIDS infection rate in each of the sub-regions covered by a church using zip code information. Finally, we simulate the effects of spreading educational outreach information within each church. We measure their influence based on the number of HIV/AIDS infected churchgoers who are influenced positively by the clergy-led HIV/AIDS educational outreach.

### *3.3.1 The Influence Network Construction*

We choose independent cascade influence network model for our simulation here. The model consists of three elements: the node set $V$, the arc set $E$ and the weights on the arcs $p(e)$. Here, the network is considered as directed, even though friendship is usually a mutual relationship. We note that the abilities for a friend pair to influence each other are sometimes different. This is especially true when one side is in a leadership role while the other is in a follower role.

Dunbar and Spoors (1995) [24] and Hill and Dunbar (2003) [32] investigated the general social circle. The authors suggested that for most individuals, the social circle includes three layers: 1) the support clique, which only includes one’s closest friends and certain kin. The individual would seek advice and help only within the support clique when facing difficulties; 2) the sympathy group, defined as all those whose sudden death would be upsetting, this is the principal group of one’s social circle; and 3) the general acquaintance, these are the people that one would send a Christmas card. Table 1 shows the mean size, standard deviation for each of these three layers.

**Table 1 – Social network layers and their sizes.**

| Layers | Mean ($\mu$) | Std dev ($\sigma$) |
|---|---|---|
| Support clique | 4.72 | 2.95 |
| Sympathy group | 11.6 | 5.64 |
| General acquaintance | 124.8 | 34.69 |

Using the definition of these social circle layers, it can be deduced that one can assert only trivial influence on general acquaintance. In our construction, we first combine the support clique and the sympathy group together and assume that the combined group size is represented by a positive truncated normal distribution with mean $\mu$ and standard deviation $\sigma$. Specifically, let $(\mu_1, \sigma_1)$ and $((\mu_2, \sigma_2)$ be the mean and standard deviation of the support clique and the sympathy group respectively, then $\mu$ and $\sigma$ of the positive truncated distribution satisfy $\mu = \mu_1 + \mu_2$ and $\sigma = \sigma_1 + \sigma_2$. Alhough the sum of two independent normal distributions is still normal, for positive truncated distribution this property does not hold. However, if we let $X_1 \sim N(\mu_1, \sigma_1)$, and $X_2 \sim N(\mu_2, \sigma_2)$, using the data in Table 1, we have $Pr(X_1 < 0) = 0.055$ and $Pr(X_2 < 0) = 0.020$. Hence the error of using the property of normal distribution herein would be quite small and acceptable.

The arcs in the support clique and those in the sympathy group will be identified when we generate the influence probabilities. Even though the arcs in our model are directed, friendship is usually mutual. Thus, we assign a value q to represent the percentage of arcs for which the relationship is mutual. When assigning arcs to a certain node $v$, we first consider connecting the nodes that are the source of an arc directed to $v$.

Research has also shown that in reality, people are more willing to build relationships with trustworthy partners and extroverted people tend to have a larger social circle (Amichai-Hamburger and Vinitzky (2010) [3], Bravo et al. (2012) [10]).

Thus, instead of generating a random number from the positive truncated normal distribution, we first generate a $U(0,1)$ random variable $S(v)$ to represent the trustworthiness and networking ability of an individual $v$; we then use $F^{-1}$ $(S(v))$ to find the number of outgoing arcs from node $v$. Here, $F$ is the cumulative distribution function of positive truncated normal ($\mu$,$\sigma$) distribution.

An arc $(v_1, v_2)$ means that $v_1$ is able to influence $v_2$ with the probability $p(v_1, v_2)$. Given a set of nodes, Algorithm 2 generates the arcs and the resulting influence graph.

**Algorithm 2.** Influence graph construction

```
Input: Number of nodes $n$, mean $\mu$ and variance $\sigma$ for the size of the support clique and the sympathy group.
Output: The arc set $A$ (or the adjacency matrix $A$)
For each $v$ in $V$:
    count= 0;
    $C = N^{in}$(v)
      while (count < $F^{-1}$ $(S(v))$  {
        if  $(C = \emptyset)$ {
          Uniformly randomly choose an unconnected node $w$;
          Generate $u$ ~ Bernoulli($S(w)$), if $u = 1$ then add $(v, w)$ to $A$  and count++;
          }
        else {
          Choose a node $w \in C$.
          Generate $u$ ~ Bernoulli($q$), if $u = 1$, then add $(v, w)$ to $A$ and count++;
          Remove $w$ from $C$;
         }
      }
}
```

After generating the nodes, the influence probability on each arc is assigned. In our network, we only need to identify the arcs that represent the support cliques. To do this, a positive truncated normal distribution random number is used to fix the size of support cliques and then fit them into the existing arcs. We note that this algorithm generates a directed graph in which some of the weak relations in the social network are not included.

### *3.3.2 The Church Network*

We apply our generalized social network to churches. First we divide the entire metropolitan area into divisions with each division centered on a megachurch. The geographical boundaries of the divisions can be represented spatially with known population information from actual census data. We divide the population in one parish into three groups:

- **Group 1 (The clergy):** This group represents the clergy in the church; they have a good reputation among church participants.

- **Group 2 (The regular church participants):** This group represents regular church participants. They usually go to church on a weekly or monthly basis.

- **Group 3 (Non-church participants):** This group represents the people who do not attend the center church on a regular basis, it includes people who never go to church and/or people attending other smaller churches.

To construct the specialized church network, we first construct the network between individuals in Group 2 and 3. Then we connect the clergy in Group 1 to church participants in Group 2. To account for uncertainties in the degree of influences clergy have on participants, multiple connecting methods can be explored. For example, we could connect each clergy to each participant, or we could first generate the number of clergies from which a participant is acquainted with; then choose among these the most influential clergy to connect to the participant. The transition probability on the arcs connecting Group 1 and Group 2 equals the probability on the arcs representing the support clique for weekly church participants and the arcs representing the sympathy group for monthly churchgoers. An example of the church network is displayed in Figure 1.

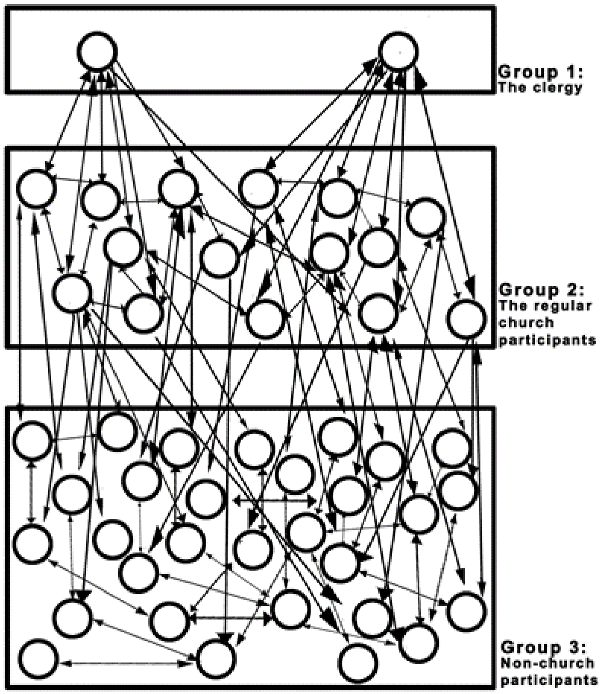

**Figure 1 – An example of the church network.**

### *3.3.3 The HIV/AIDS Infection Rate Estimate*

We design a zip code-based approach using the CDC HIV/AIDS infection rate data that are organized by zip codes to estimate the HIV/AIDS infection rate and assign the infections on the nodes of our network. The approach consists of three steps:

Step 1: Take the megachurch as the center, and construct a circle of radius $R$. This circle represents the geographical scope that the impact of the church could reach.

Step 2: For each zip code within this circle, calculate its distance to the church. Let $d_i$ denote the $i^{th}$ distance. In practice, we choose the zip code position on Google map to represent the zip code and use it to calculate its distance from the church.

Step 3: Let $I_k$ denote the zip code set of the kth circle, excluding the zip code for the church (k). The infection rate $r_k$ for this church region is calculated as follows.

$$r_k = \alpha f_k + (1 - \alpha) \sum_{i \in I_k} w_i f_i \tag{1}$$

Here $f_k$ denotes the CDC infection rate for the zip code that contains the church, $\alpha \in [0,1]$ is the weight placed on this zip code; and $w_i$ is calculated as

$$w_i = \frac{e^{-d_i}}{\sum_{i \in I_k} e^{-d_i}} \tag{2}$$

$-d_i$ is used as the relative weight of the $i^{th}$ zip code to reflect that the influence of the church would diminish exponentially with respect to distance.

### 3.4 Experimental Results

The city of Atlanta is used to demonstrate the model usage, and HIV/AIDS information spread analysis. Atlanta, with a population of 456,002 [2014 census], is experiencing a huge HIV/AIDS outbreak. Based on CDC data some areas in downtown Atlanta have an HIV/AIDS infection rate as high as some undeveloped African countries (AIDSVu (2019) [2]). Further, Georgia is ranked the 8th highest in church attendance among the 50 states and District of Columbia (Newport (2015) [46]), with 39% of Georgia residents attending church on a weekly basis; making Atlanta a perfect site to analyze the effects of disseminating HIV/AIDS information via churches.

### *3.4.1 The Independent Cascade Model*

There are 12 megachurches in the Atlanta area, all have over 1,000 regular participants (*Churches in metro Atlanta area* (2006) [18], *Fast Facts about American Religion* (2006) [25]). There are also many smaller churches. In this paper we infer the spreading effects by analyzing the megachurches only. Figure 2 shows the locations of the churches. The only data available to us is the number of weekly participants, each centered on a megachurch. For each division of the church, we apply our model (from Section III) to establish the influence network and estimate the HIV/AIDS infection rates. This method is iteratively applied to all the divisions to estimate the spreading effects to the entire city of Atlanta. By simulating the activities centered on each megachurch, we could estimate the spreading scope in the population radiated by the megachurches and use this to get an inference on the effects on the whole city. In this way, the problem of solving a graph with 400 thousand nodes involves solving independently each church partition. Inter-dependency effects (across partitions) can also be modeled.

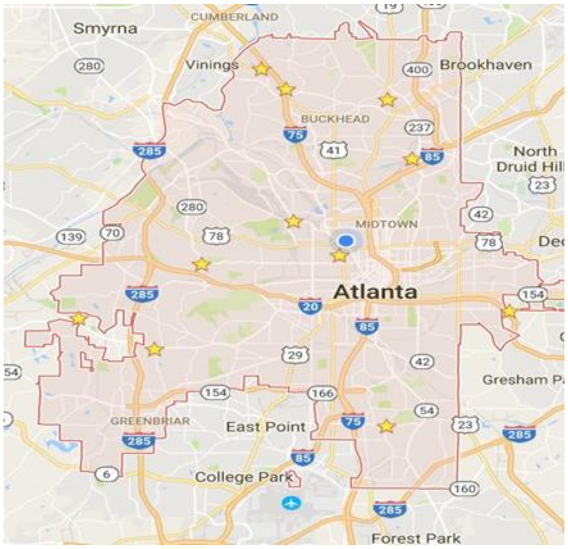

**Figure 2 – Megachurches in city of Atlanta.**

Our experiments intend to i) prove (and confirm) that in the church clergy-participant network, the most influential nodes for maximizing the overall influence are the clergy nodes; and ii) evaluate the overall effects under various interactions. This will assist church clergy and decision makers in strategizing the rules of engagement with the participants to optimize their overall influence.

We test three homogeneous models contrasting the different connection and interaction between clergy and participants. In model 1, each clergy member is connected to each participant. The subgraph connecting clergy and church participants is bipartite. In model 2, we randomly generate the number of clergies a participant knows and randomly assign them to an available clergy. The number of clergies a participant knows is a discrete uniform distribution from 1 to number of clergy. In model 3, each participant is only matched to one clergy member. This seems realistic since each churchgoer tends to have his/her trusted clergy. Each clergy member knows a binomial distributed number of participants. We also include two mixed / heterogeneous models (model 4 and 5), In model 4, the clergy and the participants are connected with equal proportion as in model 1, 2 and 3 respectively (i.e., 1/3 for each). In model 5, the mixing percentage of the three models are 25%, 25% and 50% respectively. In all models, only one kind of information is spreading on the network. In our test, the cardinality of our target set is equal to the number of clergies for each church. We run the test 1000 times for each church and take the average over the runs. The model parameters used across all models are shown in Table 2.

**Table 2 – Model parameters.**

| Parameter name | Value | Source |
|---|---|---|
| # of monthly participants/# of weekly participants | 24/39 | Newport (2015) [46] |
| The rate of clergy/total population | 0.002 | *Fast Facts about American Religion* (2006) [25] |
| Percentage of mutual friendships | 0.9 | assumption |
| Transition probability for support clique | 0.2 | assumption |
| Transition probability for sympathy group | 0.05 | assumption |
| Radius of the church influence circle | 5 km | assumption |
| $\alpha$ in Equation (1) | 0.5 | assumption |

The test results show that the most influential set exactly contains all clergy members. The results are expected since clergy affect / influence many more people than others. Thus, we focus on the influence propagation scope when all the clergies are influenced and served as the initial active set.

Figure 3 shows the results for the 5 models. The two horizontal axes are the number of weekly participants and the HIV/AIDS density in the area where the church is located. The vertical axis shows the simulation results on the number of HIV/AIDS participants who are influenced / affected positively by clergy efforts in disseminating information. We observe that results from model 1 and model 2 are quite similar and the most effective, and the results from model 3 are the least effective. The mixed cases return results that are bounded above and below by model 1 and model 3, respectively.

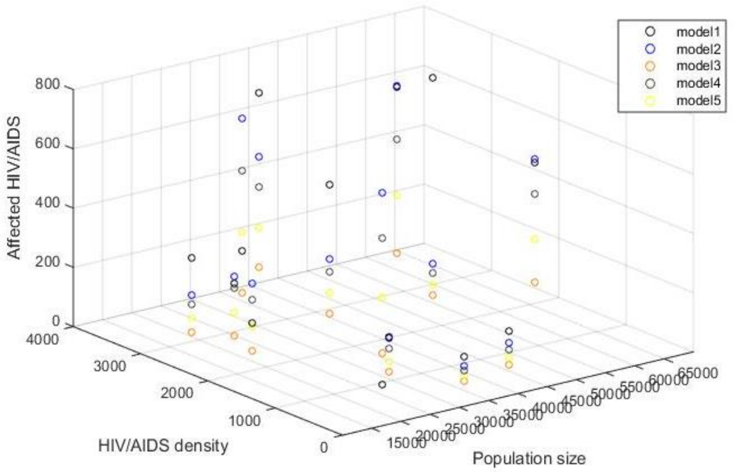

**Figure 3 – Affected HIV/AIDS population in each of the megachurches.**

Figure 4 shows the percentage of HIV/AIDS population influenced. The results show that regardless of the church population size, model 1 asserts the highest influence (55% versus 13% for model 3 respectively), due to the close relationship that participants have with each clergy. Model 2 (dark blue) shows that participants only need to know sufficient number of clergy (not all of them) to benefit from the outreach as well as in Model 1. Model 3 shows that if each participant only knows one clergy, the outreach will not be very successful. Mixed models probably present a more realistic connection pattern of the congregation. It is encouraging that they offer 27% to 43% positive outreach gain. These findings demonstrate that effective communication and interaction style must be explored to optimize clergy outreach efforts.

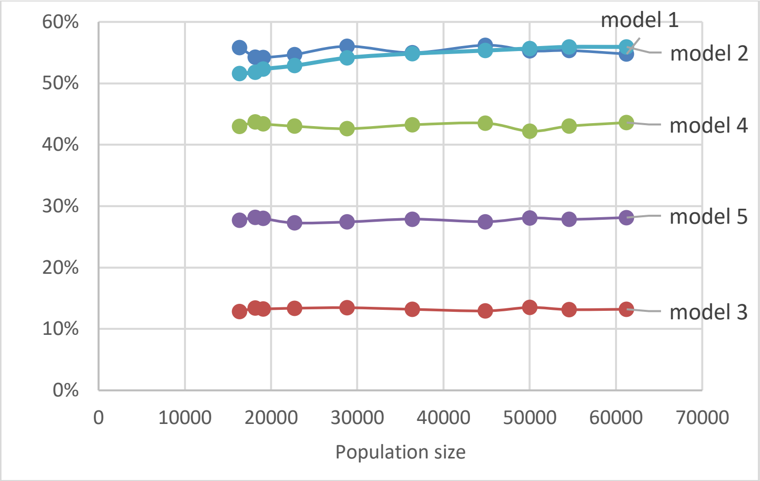

**Figure 4 – Percentage of positively affected HIV/AIDS population.**

Our findings also suggest that regardless of the church size, the percentage of the HIV/AIDS population that is positively influenced remains rather steady. In fact, the percentage of all people (regardless of HIV/AIDS infection status) influenced by the church also remains relatively constant. If n denotes the population size in a certain church area with h the percentage of HIV/AIDS infected population; and p the size of HIV/AIDS population who are influenced positively at the end, our findings suggest that $p = \lambda n h$ for some positive scalar $\lambda < 1$. This leads to an interesting conclusion: we can estimate a rough ranking of the churches in any area by simply ranking the product of the number of participants and the HIV/AIDS infection rate in this area. Strategically, public health leaders can determine in this order the allocation of resources in reaching out to the churches.

### *3.4.2 Competitive Model*

In reality, as with social media, where there are positive messages and misleading and/or negative messages, there may be opposite information countering the information that we want to spread. We would like to investigate the net effect of this competitive information messaging. We use a competitive independent cascade (CIC) model defined in Definition 2.12 with proportional tie-breaking rule to model this competitive outreach environment. In a simplistic case, two types of information are labeled: one positive and one negative. The positive information is what we would like to spread and influence the community positively; while the negative information is countering our effort.

There are usually stable groups holding the positive and negative opinions respectively. A node is said to be stable when it holds the positive or negative opinion before the information dissemination starts. In our simulation for the competitive model, we assume initially 1%, 5%, 10% and 15% population hold positive opinion while the similar number hold opposite opinion. Our aim is to compare the spreading effects with and without HIV/AIDs outreach by the clergy.

Figure 5 contrasts the clergy effect against the rate of HIV/AIDS population that are positively influenced under the assumption of model 3 (that every participant only knows one clergy). When the endogenous percentages of people holding opposite opinions is relatively high (at 15%), the effect of the clergy outreach is minimal and dominated by the network effect of opposing stable groups. The clergy's influence becomes more significant when most participants have neutral opinion. When only one percent of population hold opposite opinion, the positive influence of the clergy has a two-fold increase. This affirms that religious leaders play an important role in society and can potentially have a broad impact on HIV education. We observe that the rates do not fluctuate much, showing that when other variables are fixed, the size of the church congregation does not play a leading role in the scope of information propagation. This again supports the linear relationship that we observe in the non-competitive case.

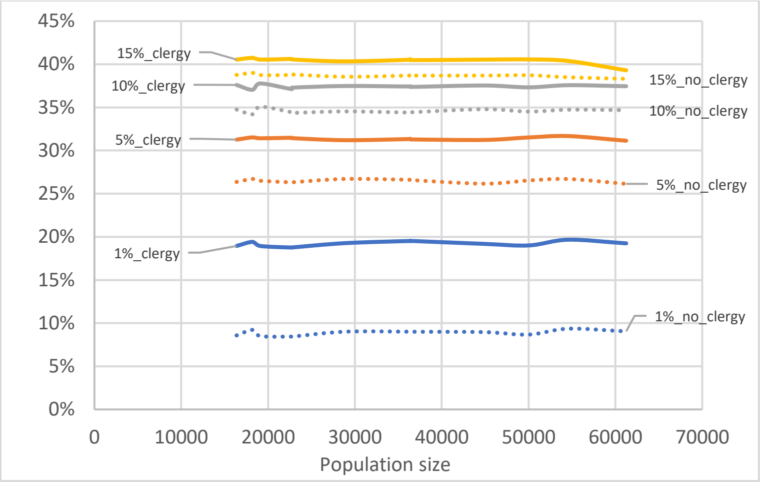

**Figure 5 – Competitive model: Percentage of positively affected HIV/AIDS population with (solid curves) or without (dotted curves) clergy outreach (when each participant is connected to one clergy only.**

Figure 6 analyzes the roles of the congregational workers as the size of the stable group varies. The horizontal-axis is the percentage of positive stable group in the population (same as the percentage of negative stable group) and the vertical-axis is the percentage of HIV/AIDS population who are affected by positive information.

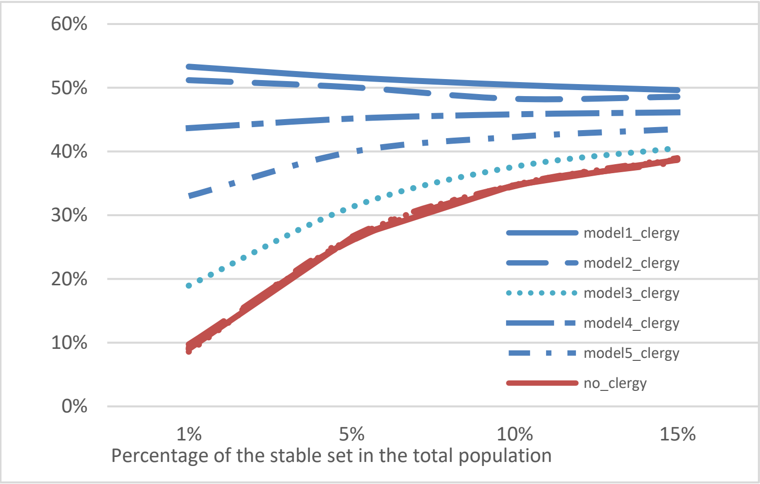

**Figure 6 – Competitive model: Percentage of affected HIV/AIDS population for the largest church in Alanta. The x-axis corresponds to the endogenous percentage of people holding positive (and negative opinions).**

Recall that in model 4, 1/3 of the participants are connected to all clergy, another 1/3 of the participants are connected to a random number of clergy while the remaining 1/3 are connected to only one clergy. In model 5, the weights of the 3 become 0.25, 0.25 and 0.5.

The results for the 5 models without involvement of clergy (orange curves) are virtually identical. When clergy are not involved, the connection style is not important.

For the 5 models with clergy HIV outreach, model 1 and model 2 show a slight decreasing trend with respect to the stable group size. This means that in models 1 and 2 clergy are somewhat more influential when the stable group size is small. The gap between having clergy versus no clergy are very significant, although diminish as the size of the stable set increases. This clearly confirms the important role of clergy, especially when the church community is close-knit and participants and clergy know each other well.

Models 3, 4 and 5 show increasing trends, showing that clergy become important as the stable group becomes bigger. These models have looser interaction networks. This shows that when the participants have strong opinions, it starts to spill over with the clergy's outreach.

Contrasting the with (blue) and without clergy (orange) trend across all models, we can see from Figure 6 that the gap between the two curves diminishes when the size of the stable set expands, which corresponds to our intuition that the importance of churches in spreading information decreases when the original propagation sources are ample. However, the positive role of clergy in disseminating knowledge remains significant (despite strong opinion) when participants know and interact with multiple clergy. The effect is most limiting when each participant only knows one clergy.

### 3.5 Conclusions

In this paper, we propose an approach to rank churches based on their capability to spread HIV/AIDS information when limited data is available. Since the resources that public health leaders have for church educational programs are limited, choosing the sites to affect the maximum (positive) influence is practical and essential. Although we focus on HIV/AIDS experiments in our validation, the computational framework is generic and is applicable to diverse social network analyses, including public health disease trending and/or in spreading social or other information.

The method presented has three novelties. First, it builds a generalized influence network based on sociological human behavior theory. The model does not require specific group data and/or a local social network construct. Second it estimates the disease infection rate through spatial and census information. Third, it models the network effect among all people (clergy and participants) by coupling sociology theory of social circles. The computational algorithm simulates combination network effect and measures the net positive outcome.

We implement and analyze the model for effective HIV/AIDS knowledge dissemination for 12 megachurches in the city of Atlanta. When no competing information is present, we discover that while different connection/interaction structures between clergy and regular church participants would affect the degree of knowledge spread, the percentage of HIV/AIDS infected people who would eventually learn the disease information with the help of the church stays constant, regardless of the church size.

The same results apply to the competitive model (when both positive and negative information exist among the participants and the population). Thus, one can design a rough ranking by simply ranking the product of the church size and the local HIV/AIDS infection rate. This offers a practical policy for public health HIV intervention and education.

When there is competing information, it becomes clear that clergy assert more significant (and positive) influence among church participants when the church community is close-knit and participants know multiple clergy. Even when each partcipant knows only one trusted clergy, clergy does play an important role in the information spread, and their significance is more evident when the strong opinion group is small. Churches can organize social activities to facilitate active participants' interaction to a broad group of church members to optimize their effect. When participants interact with multiple churches, the overall impact of clergy in promoting HIV/AIDS knowledge becomes more significant. Compared to previous results for influence networks, our algorithm is scalable and can be used to analyze any population size.

# CHAPTER 4. INTERDEPENDENCY ANALYSIS IN CELLULAR STATION NETWORK

The communication sector is one of the 16 critical infrastructure sectors identified by The Department of Homeland Security (DHS). Cellular station networks are a vital part of the sector on which almost all private businesses, organizations and governments rely. In this chapter, we try to answer the question that if the resources held by policy makers are restricted, how can we find the most critical cellular stations whose collapse would affect most cellular stations in the network. To solve the problem, we present a two-stage framework to analyze the cascading effects in the cellular station network using the linear threshold influence network, where the stations are modeled as the nodes and the station loads variation are modeled as the influences. A case study based on the cellular station network in the United States is explored as experiments.

## 4.1 Introduction

Critical infrastructures (CI) are critical components for economy operations. The disruption of CI could have debilitating effects on private businesses and governments. The importance of the CI security and resilience has been identified by U.S. government (Presidential Policy Directive, White house, 2013[21]). Additionally, the CI are highly dependent on each other so that a malfunction of any component could lead to other component failures, which we call cascading effects. For example, in August 2016, a power outage struck the Delta Airlines data center in Atlanta, causing data loss, this failure spread to the air transportation sector since most of the Delta Airlines could not depart due to loss of data. Such an interruption can quickly propagate to many other airports (Dastin (2016) [19]).

The communication sector is a CI on which many other sectors, such as transportation, commercial and governmental facilities and finance, rely (*National Infrastructure Protection Plan (NIPP) Communications Sector-Specific Plan for 2015* (2015) [43]). Among all the components of the communication sector, the cellular base station network is one of the most important part. During a crisis, some towers in the network may stop functioning, the users' devices will have to use other nearby towers, causing congestions. In such a situation, nearby users may also suffer service outages. Thus, cascading effects arise.

In this chapter, we will present methods to summarize a LT influence network from cellular station network data as well as a two-stage approach to investigate the cascading effects. The first stage is only necessary when the problem scale is large. We perform a geographical clustering on all the nodes to form sub-networks. In the second stage, we construct a linear-threshold influence network to simulate the congestion propagation.

## 4.2 Related Works

The critical infrastructure interdependency was first investigated in 2001 (Rinaldi et al. (2001) [53]), the paper classified for CI interdependency: physical, cyber, geographical and logical. The classification gave a good reference for most of the paper afterwards. Their subsequent paper (Rinaldi (2004) [52]) summarizes the likely methods for the interdependency analysis.

There are three directions for modelling interdependency: simulation based, analytics based and data based (Ouyang (2014) [47]). Among simulation-based approaches, Dudenhoeffer et al. (2006) [23] designed an agent-based approach to simulate the and used genetic algorithm to decide the CI components to protect/restore. Johansson and Hassel (2010) [33] model the CI interdependency as a network and simulate the flows when removing edges to find strains added to the network. Zio and Sansavini (2011) [61] made a notable approach that model the interdependency as load transfers that failed nodes would transfer its node to adjacent nodes, however they do not have realistic experiments to test how well the model works. For analytics based approaches, Lee II et al. (2003) [38]and Lee II et al. (2007) [37] modeled the provision interdependency as a multi-commodity network flow problem and gave a mixed-integer programming (MIP) formulation to solve it. Svendsen and Wolthusen (2007) [57] designs another multi-commodity flow formulation but assigns a response function for each arc and each resource, where some of the resources can be buffered. A drawback of using network problem for this is that it can only model the provision interdependency while other types of interdependency, like geographical, do exist. Data-based models are some methods designed based on special data forms. For instance, Ramachandran et al. (2015) [49] summarizes the geospatial data to find the CI components that would affect most other CI components geographically. Reilly et al. (2015) [50] assumed that each CI sector is managed by a certain governmental or private department and explored the externality of the policy taken by some departments as interdependency.

There are three directions for modelling interdependency: simulation based, analytics based and data based (Ouyang (2014) [47]). Among simulation-based approaches,

Dudenhoeffer et al. (2006) [23] designed an agent-based approach to simulate the and used genetic algorithm to decide the CI components to protect/restore. Johansson and Hassel (2010) [33] model the CI interdependency as a network and simulate the flows when removing edges to find strains added to the network. Zio and Sansavini (2011) [61] made a notable approach that model the interdependency as load transfers that failed nodes would transfer its node to adjacent nodes, however they do not have realistic experiments to test how well the model works. For analytics based approaches, Lee II et al. (2003) [38]and Lee II et al. (2007) [37] modeled the provision interdependency as a multi-commodity network flow problem and gave a mixed-integer programming (MIP) formulation to solve it. Svendsen and Wolthusen (2007) [57] designs another multi-commodity flow formulation but assigns a response function for each arc and each resource, where some of the resources can be buffered. A drawback of using network problem for this is that it can only model the provision interdependency while other types of interdependency, like geographical, do exist. Data-based models are some methods designed based on special data forms. For instance, Ramachandran et al. (2015) [49] summarizes the geospatial data to find the CI components that would affect most other CI components geographically. Reilly et al. (2015) [50] assumed that each CI sector is managed by a certain governmental or private department and explored the externality of the policy taken by some departments as interdependency.

## 4.3 Methods

### *4.3.1 Model Selection*

The modern cellular network consists of many small calling areas where each area is served by a cellular base station. The base stations are transceivers connecting several other devices to one another and/or to a wider area. The base stations would exchange information when the communication is made. When physical or cyber attack paralyzes a tower, the users who are sending or receiving signal of it need to seek working towers nearby, and thus aggravate the burdens of them.

Our model tries to answer such a question: Given a number $K$, which represents the number of stations to which our resources can be allocated, which $K$ stations in a given area, if attacked, could affect the largest number of stations in the network. By answering this question, we should be able to know that the failure of these stations would lead to the most serious loss and thus the protection would have the highest efficiency.

In the IC influence network model each node attempt to activate the adjacent inactivated node independently, which means for any inactivated node, even one activated node in its in-node set could succeed in activating it. However, that contradicts the facts that the load transfer is accumulative in cellular base station network. The incoming signal must exceed a certain threshold (max power) for the base stations to stop taking new users. The LT influence network model fits much better since the threshold $\theta_v$ can be interpreted as $\frac{\text{Max load} - \text{current load}}{\text{Total load incoming when nearby towers are down}}$ of node $v$.

### *4.3.2 Learning the Parameters of the LT Influence Network*

After choosing the influence network model, the next step would be defining the network elements. Naturally we let each node denote a base station in the cellular network. To find arc set $E$, we define a max reach distance $R$ for all base stations, it is the distance from the most distant user to the base station. For any base stations $w$ and $v$, if their distance is less than $2R$, we assume that there might exist user who originally uses $w$ but has to use $v$ when $w$ is paralyzed. Thus an arc should go from $w$ to $v$, and vice versa.

The final step is to learn the weights on arcs. Theoretically, for each node $v$, let $L_v^M$ be the maximal load, $L_v^C$ be the current load and $L_{(u,v)}$ be the load going from $u$ to $v$ when node $u$ is down. We should have

$$\theta_v = \frac{L_v^M - L_v^C}{\sum_{(u,v)\in E} L_{(u,v)}}$$

Notice that $\theta_v$, as we have defined, is U$[0,1]$ random variable. This is because the $L_v^C$ and $L_{(u,v)}$ are constantly changing. Meanwhile, We assume that $\sum_{(u,v)\in E} L_{(u,v)} \geq L_v^M - L_v^C$, which means that if all the adjacent nodes around $v$ are down, $v$ would also be down due to high loads. With these assumptions, we let weight be

$$w(u,v) = \frac{L_{(u,v)}}{\sum_{u\in N^{in}(v)} L_{(u,v)}}$$

so that for any $v$, $\sum_{u\in N^{in}(v)} w(u,v) = 1$. $\theta_v < \sum_{u\in N^{in}(v)} w(u,v) = 1$ corresponds to the assumption that $v$ is down if all the adjacent nodes around $v$ are down.

Thus, to calculate $w(u,v)$, we only need $L_{(u,v)}$, While we assume that we know the geographical positions of base stations, we first randomly generate the users' locations uniformly on the whole target area. For each user, associate it to the nearest station, which should be the base station it uses. Except the nearest station, we also find the second nearest station for each user, which represents the station that the user would connect when the primary station is down. In this way, $L_{(u,v)}$ is the number of users who uses $u$ as the primary station and use $v$ as the secondary station. By definition of $w(u,v)$, we don't care the absolute value of $L_{(u,v)}$ but only how much percentage it takes in $\sum_{u\in N^{in}(v)} L_{(u,v)}$. Thus the number of virtual users used does not matter as long as it is sufficiently large.

### *4.3.3 Two-Stage Framework to Analyze the Influence Network*

After getting the LT influence network, the problem next is to find the nodes that can influence the greatest number of nodes in the network. For large networks, the Greedy-MC method would take very long time to solve. As for the Simpath method, although the heuristics is scalable, it has no theoretical lower bound and the would perform badly when the network size is big and complex. In view of this, we designed a two-stage framework to analyze this problem.

Stage 1 is forming sub-networks by clustering. To compare two methods on different sizes of network, we need to break the complex large network into small networks. Since the influences cannot spread over long range, it is reasonable to cluster geographically, the clustering method we used is K-means++ method by Arthur and Vassilvitskii (2007) [5]. In stage 1 we only cluster nodes, the arcs between nodes which belong to different clusters would be removed.

In stage 2, we select the nodes to protect from each sub-network. Suppose that we get $n$ sub-networks in stage 1 and need to select $K$ nodes. For each sub-network got from stage 1, we select $\left\lfloor\frac{K}{n}\right\rfloor$ or $\left\lceil\frac{K}{n}\right\rceil$ nodes by Greedy-MC method and Simpath method respectively so that the total number of selected nodes equals to $K$. This method could ensure the final node set distribution is generally uniform on the target area. It is worth pointing out that the choice $n$ implies a tradeoff between less artificial restrictions on the original network and higher precision on each sub-network. Thus, the user should try various $n$ values to find the one that gives the best results.

## 4.4 Experiments and Sensitivity Analysis

### *4.4.1 Data and Experiments*

We test our model on the cellular station data set on Homeland Infrastructure Foundation-Level Databased provided by U.S. Department of Homeland Security (DHS (2018) [20]). The dataset includes the geographical locations of 23498 cellular towers in the United States.

We set max reach distance $R$= 5km to formulate the LT influence network as we discussed in part 4.3.2. When applying the two-stage framework to analyze the LT influence network, we choose $K$=100, i.e. we want to pick 100 most influential nodes. The $n$ value we try on both Simpath and MC-Greedy method are [10, 20, 25, 33, 40, 50, 60, 70, 80, 90, 100]. Generally, we choose the grid size to be 10, for $n$ values 25 and 33, we try it because it allows exactly 4 and 3 nodes in each cluster and Greedy method performance descends when number of nodes increases due to dependency in network. For Simpath method, we also try skipping stage 1, i.e. no clustering, to compare the results. We did not do this for MC-Greedy since the method is not scalable. In MC-Greedy method, every MC simulation is set at 1000 rounds.

The results are presented in Figure 7 and Figure 8. In Figure 7 we plot the final affected nodes for two methods and various $n$ values. Each curve represents a method. From the figure we can see that for each method, when the $n$ values change, the number of final affected nodes are not monotone but with several turning points. For Simpath method the maximal value appears at $n = 33$ while for MC-Greedy method it appears at $n = 60$. For most of $n$ values less than 40, the Simpath is better than MC-Greedy while for $n$ values over 40, the MC-Greedy outperforms. This is possibly because when choosing more nodes, the MC simulation requires more rounds to be accurate and thus the MC-Greedy gives bad results. The results by Simpath method without clustering is the worst among all results which justifies the necessity of stage 1. Last but not least, the overall optimal $n$ is 60 when using MC-Greedy method.

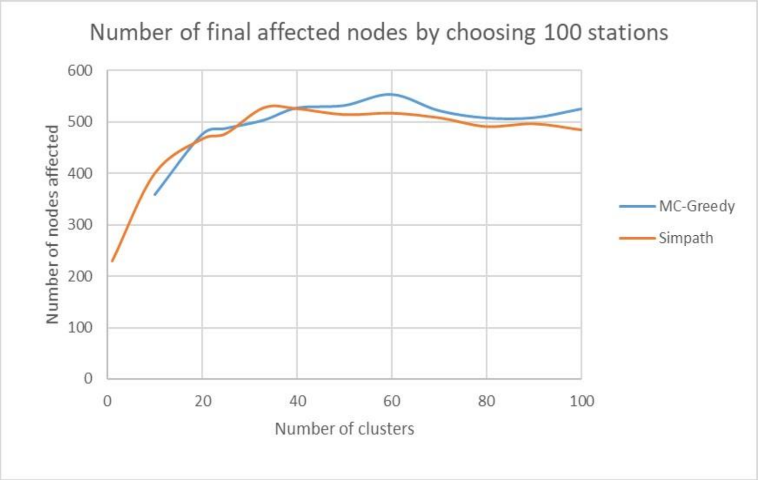

**Figure 7 – Number of final affected nodes when choosing 100 stations.**

Figure 8 shows the running time for two methods. As a scalable approach, Simpath is much faster for all $n$ values.

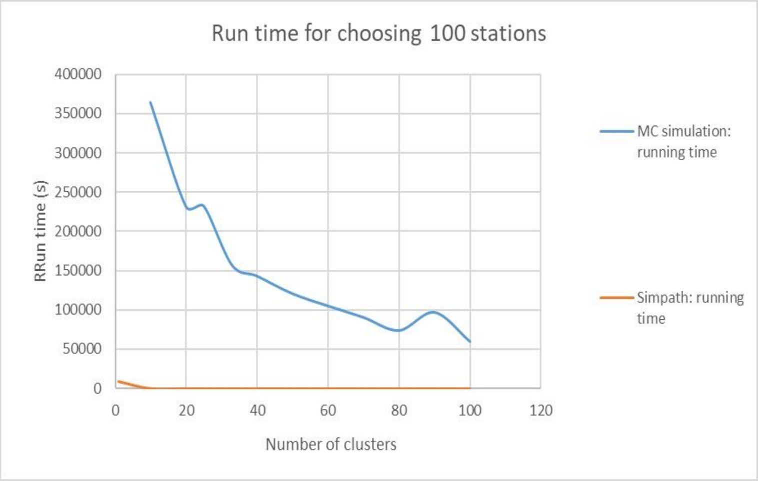

**Figure 8 – Running time for choosing 100 stations.**

### *4.4.2 Sensitivity Analysis*

After getting the results of both methods, a natural question is that whether two methods give similar results, i.e. most of the chosen stations overlapped. Table 3 shows the number of stations overlapped for various n values. The overlapped nodes are fluctuating around 20, which means only about 20% of the nodes overlapped.

**Table 3 – Number of overlapped nodes across two methods.**

| Number of clusters | Overlapped nodes |
|---|---|
| 10 | 23 |
| 20 | 15 |
| 25 | 14 |
| 33 | 13 |
| 40 | 15 |
| 50 | 24 |
| 60 | 20 |
| 70 | 16 |
| 80 | 18 |
| 90 | 23 |
| 100 | 22 |

However, it is also possible that two nodes chosen by two methods are close rather than exactly overlap. To check this, we plotted two set of figures. The first set includes Figure 9, Figure 10, and Figure 11, which compare the nodes chosen by two methods in 100, 50 and 33 clusters, respectively. From the figures we see that for 100 clusters, the chosen nodes by two methods are not geographically close. For 50 and 33 clusters, even though the number of exactly overlapped nodes are not greater, the nodes chosen by two methods are close with each other. We assume the reason is that the sub-networks divided by 100 clusters are not favorable at all, like they represent some area with very low population density. Since it is forced to choose 1 node from each sub-network, some chosen nodes from such sub-network would not appear in 50 or 33 clusters.

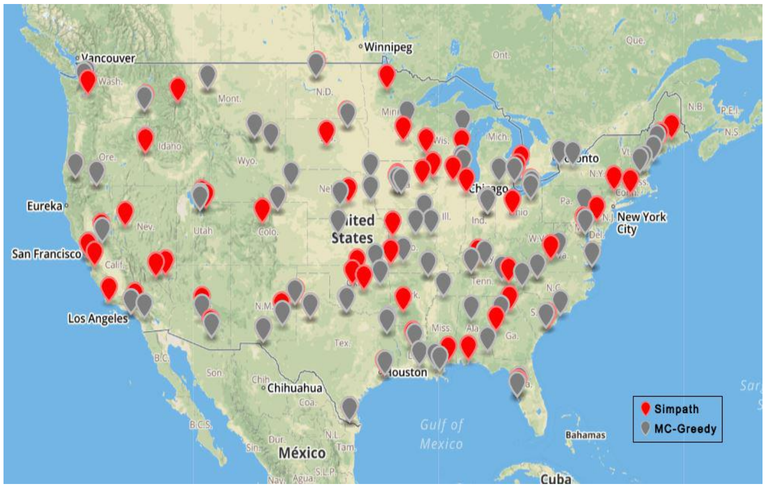

**Figure 9 – Nodes chosen by two methods with 50 clusters.**

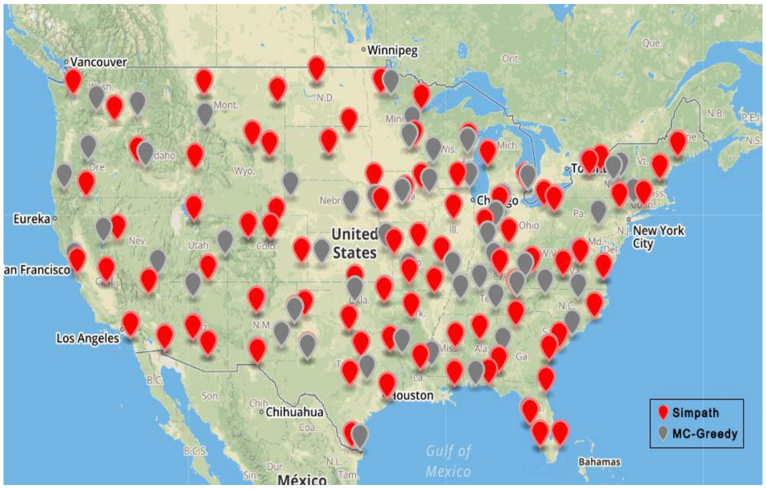

**Figure 10 – Nodes chosen by two methods with 100 clusters.**

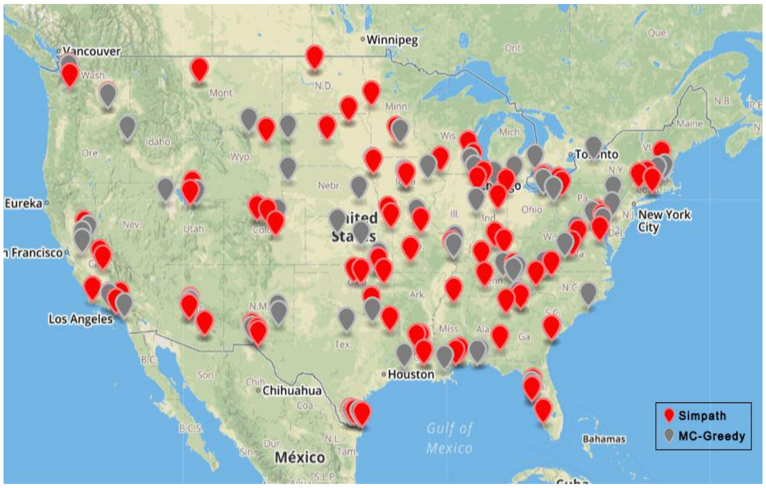

**Figure 11 – Nodes chosen by two methods with 33 clusters.**

Another sensitivity analysis we have done is to check that for the same method and different number of clusters, whether the same set of nodes are chosen or not. The results are shown in the second set of figures, which includes Figure 12 and Figure 13. As we have analyzed above, for both methods, the nodes from 33 and 50 clusters are close to each other while the nodes from 100 clusters outstand.

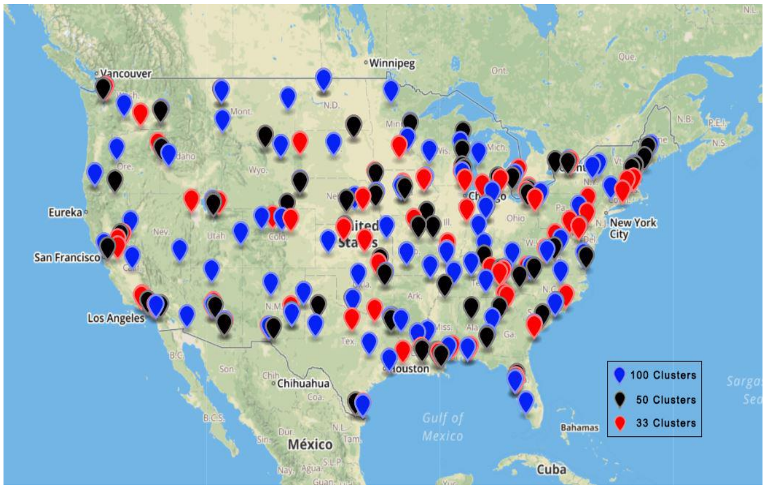

**Figure 12 – Nodes chosen by 33, 50 and 100 clusters by MC-Greedy method.**

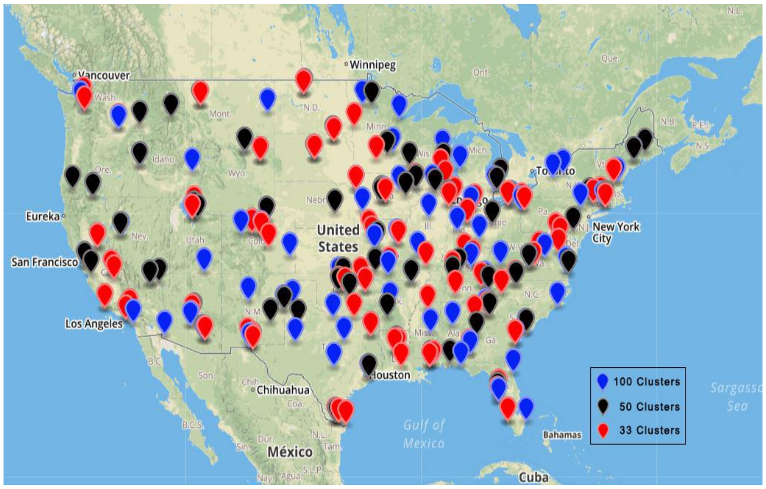

**Figure 13 – Nodes chosen by 33, 50 and 100 clusters by Simpath method.**

## 4.5 Conclusions

In this chapter, we presented a method to formulate the cellular station network as an linear-threshold influence network and a two-stage framework to analyze the LT influence network. The method was tested on the cellular station network within the U.S. We analyzed the results to identify the optimal partition in the network and made the sensitivity analysis to compare different influence network analyzing method and partitions. The results would give a good reference for the policy makers on how to allocate the limited resources to protect the communication infrastructures more efficiently. The work for next step would include extending the network formulation and analysis framework to other critical infrastructures (CI) and developing a multi-layer influence network model to analyze the interdependencies across CI sectors. The related work is presented in chapter 5.

# CHAPTER 5. MULTI-LAYER INFLUENCE NETWORK MODELLING ON CRITICAL INFRASTRUCTURE INTERDEPENDENCIES

In this chapter, we extend the linear threshold influence network model to the multi-layer case and use this new tool in a two-stage framework to analyze the cascading effects in the CI interdependency network. This is an extension of chapter 4 where we only explore the single-layer LT influence network. We explore two experiments on the metro Atlanta area and the state of Florida as the applications of the model.

## 5.1 Introduction

In this paper, we extend the single-layer linear threshold influence model that we have introduced in chapter 2 and 4 to the multi-layer case in order to study the scope of cascading failure effects from each sector. As we have done in chapter 4, we model the CI interdependency network as a directed graph $G = (V, E)$. The vertices (or nodes) in $V$ represent the CI facilities, such as roads, electricity transmission substations, and cellular base stations. A directed edge (or arc) exists in $E \subseteq V \times V$ if the failure of the originating node can cause a failure of the ending node, thereby capturing the possibility of cascading failures in the network. The nodes in $V$ are partitioned according to their CI sector—such as transportation, communication, or energy—according to the Presidential Policy Directive 21 (PPD-21)[21], , and we use these sectors to define corresponding layers in the network. We build a linear threshold model on each layer and assume the cascading effects are independent across each layer. We show that our new multi-layer model has the same properties of the single-layer model, e.g., submodularity and live-edge graph equivalence. In the experiments section, two scenarios, for metro Atlanta and the state of Florida, are explored.

We omit the literature review for this chapter as the contents are closely related to chapter 4 and relative literature are already reviewed.

## 5.2 Multi-Layer Interdependency Network

In practical CI domains, nearly all systems rely on external resources to function. For example, the transportation network that includes railways, airlines, and roads requires electricity to operate. Similarly, the electricity network that includes generating stations, transmission lines, and substations requires transportations systems for maintenance workers to access and service the facilities. As the method we used in chapter 4, We use a weighted influence network $G = (V, E)$ to model these interdependencies, where nodes represent CI facilities, and directed edges represent dependence on the child node on resources from the parent node. Therefore, the failure of one CI facility has the potential to cascade through the network to other facilities. Each edge is assigned a weight, which represents the fraction of required resources that come from the parent node.

To define the structure of a multi-layer interdependency network (MIN), we need to introduce the concepts of *sectors* and *layers*. We partition the nodes of $V$ into sectors—where $C_i$ denotes the *i*-th sector—based on the CI sector to which that facility belongs (e.g., transportation, communication, energy). For each sector, we define a corresponding layer that consists of both nodes and edges. The layer includes all edges originating from its sector, as well as all nodes associated with these edges. Each edge will belong to exactly one layer, but nodes may belong to multiple layers due to cross-sector dependencies. As a concrete example, a cellular base station belongs to the communication sector but requires resources such as electricity and roads (for access), and thus it would also belong to the electricity and transportation layers as a sink node.

**Definition 5.1 (Multi-layer interdependency network (MIN)).** Given a finite directed graph $G = (V, E)$ and a partition $\{C_i$ of $V$, i.e. $C_i \cap C_j = \emptyset, \forall i \neq j$ , $\cup_i C_i = V$, the graph $G = \{V, E, \{C_i\}\}$ is a multi-layer interdependency network (MIN) with sectors $\{C_i\}$. For each sector $C_i$, let $E_i = \{(v,u)|v \in C_i\}$ and $V_i = \{v|\exists u, s.t.\ (u,v) \in E_i\ or\ (v,u) \in E_i\}$. The subgraph $G_i = (V_i, E_i)$ is a layer of $G$ corresponding to sector $C_i$.

Next we define a linear threshold influence model on a MIN. Each layer can be treated as a single-layer linear threshold influence network, and influence—in our setting, CI failures—can propagate across layers. For example, if a hospital (public facilities sector) does not receive enough electricity (energy sector) to operate, then the hospital may have to temporarily close.

**Definition 5.2 (Multi-layer linear threshold influence network (MLTIN)).** Given a MIN $G = (V, E, \{C_i\})$, for each node $v \in V$and each layer $i$such that $v \in V_i$, a threshold $\theta_v^i$ is selected uniformly in [0,1], and all $\theta_v^i$ are independent of each other. Every edge $e$ is assigned a weight $w(e)$ satisfying $\sum_{u \in N^{in}(v) \cap C_i} w(u,v) \leq 1$ for all $v \in V$ In every layer the influence spreads independently using the single-layer linear threshold model defined in Definition 2.3. If a node becomes active in one layer, it will become active in all layers. That is, a node $v$ will become active if $\sum_{u \in N^{in}(v) \cap C_i \cap S_{t-1}} w(u,v) \geq \theta_v^i$ for any layer $i$. Every node is assigned a positive weight $h(v)$, and the influence function $\sigma(S)$ is the expected total weight of nodes in the final active set. That is, $\sigma(S_0)$=$E[\sum_{v\in S_\infty} h(v)]$. The MLTIN is denoted by $G = (V, E, \{C_i\}, w, h)$, or $G = (V, E, \{C_i\})$ when $w$ and $h$ are clear from context.

Figure 14 shows an example of a MLTIN for comparison to a single-layer linear threshold network. The top graph is a single-layer linear threshold network with edge weights. The bottom graph is a MLTIN, where sector 1 contains nodes {A, B, C} and sector 2 contains nodes {D, E, F}. Layer 1 of this graph includes nodes {A, B, C, D, E} and all red edges, while the layer 2 includes nodes {B, C, D, E, F} and all blue edges.

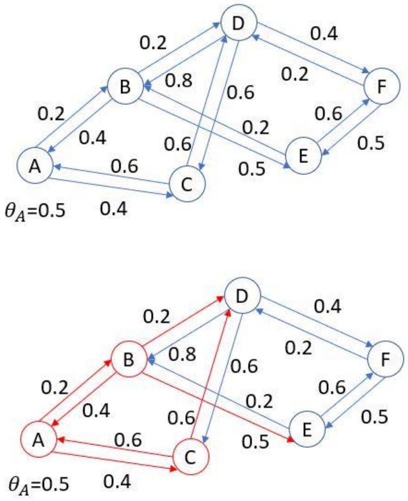

**Figure 14 – Linear threshold network and MLTIN.**

Before we dig into the properties of the MLTIN we just defined, we review some results we introduced in Chapter 2. Kempe et al. (2003) [34] proved that for both the IC and LT models, the influence function is nonnegative, non-decreasing, and submodular. And the Greedy algorithm achieves a multiplicative $(1 - 1/e)$-approximation to the optimal solution given these properties of the influence function Nemhauser et al. (1978)

[44]. Mossel and Roch (2007) [42] extended these results to the Weighted General Threshold Model (WGTM) with more general influence functions of the form $\sigma(S_0)=E[g(S_\infty)]$, for any nonnegative set function $g(S)\colon 2^V \rightarrow [0,\infty)$. The WGTM is nothing but the General Threshold Model we introduced in Definition 2.5 with every node assigned a positive weight $h(v)$. Note that for the linear threshold model, $g(S) = |S|$. Mossel and Roch (2007) [42] showed that if both $f_v(\cdot)$ and $g(\cdot)$ are monotone and submodular, then $\sigma(\cdot)$ is monotone and submodular and can be approximately optimized by the Greedy algorithm.

MLTINs are designed for the analysis of the network interdependencies that motivate our research. We next show that the MLTIN is an instance of a Weighted General Threshold Model where every node is assigned in Definition 2.5 (Lemma 5.1), and additionally that its influence function satisfies technical conditions ensuring that the Greedy method will achieve a $(1-1/e)$-approximation for the influence maximization problem for an MLTIN (Theorem 5.1).

**Lemma 5.1.** MLTIN is a weighted general threshold network with$f_v(S) = 1 - \prod_i (1 - \sum_{u \in C_i \cap S} w(u,v))$ and $g(S) = \sum_{u \in S} h(u)$, where $w(u,v)$ are the weights on edge $(u,v)$ and $h(u)$ is the weight on node $u$.

**Proof.** First, we check the conditions to activate an uninfluenced node in both MLTIN and WGTM. $\forall\, G = \{V, E, \{C_i\}\}$, an uninfluenced node $v \in V$ and an influenced set $S \subseteq N_{in}(v)$, the condition to influence $v$ for MLTIN is $\exists i, s.t. \sum_{u \in C_i \cap S} w(u,v) \geq \theta_v^i$, and the condition to influence $v$ for WGTM with activation function $f_v(A) = 1 - \prod_{i=1}^{n} (1 - \sum_{u \in C_i \cap A} w(u,v))$ is $1 - \prod_{i=1}^{n} (1 - \sum_{u \in C_i \cap S} w(u,v)) \geq \theta_v$, where $\theta_v \sim U[0,1]$.

It suffices to show that given the same influenced set $S \subseteq N_{in}(v)$, the probabilities to influence $v$ are the same for both models. It is straightforward to show that for both models, the influence probabilities are $1 - \prod_{i=1}^{n}(1 \ - \sum_{u \in C_i \cap S} w(u, v))$. Thus, the two models are equivalent.

Comparing $\sigma(S_0)$ between MLTIN and the weighted general threshold model implies that $g(S) = \sum_{u \in S} h(u)$in MLTIN. ■

Theorem 5.1 shows that both $f_v(\cdot)$ and $g(\cdot)$ are monotone and submodular are satisfied for MLTINs, and hence the Greedy algorithm provides a $(1 - 1/e)$ - approximation for the influence maximization problem on MLTINs (Nemhauser et al. (1978) [44]).

**Theorem 5.1.** For MLTIN, $\sigma(S_0) = E[\sum_{u \in S_\infty} h(u)]$ is nonnegative, monotone increasing, and submodular.

**Proof.** It suffices to show both $f_v(\cdot)$ and $g(\cdot)$ in MLTIN are monotone increasing and submodular. From Lemma 1, we know $f_v(S) = 1 - \prod_i (1 - \sum_{u \in C_i \cap S} w(u, v))$ and $g(S) = \sum_{u \in S} h(u)$. It is clear that $g(\cdot)$ is monotone increasing and submodular since $g(S) = \sum_{u \in S} h(u)$ is a linear function (thus submodular) and for all $u$, $h(u)$ is positive, so $g(S)$ is increasing.

For $f_v(S)$, note that $\forall A \subseteq B \subset N^{in}(v)$, $\sum_{u \in C_i \cap A} w(u, v) \leq \sum_{u \in C_i \cap B} w(u, v)$ and thus $f_v(A) \leq f_v(B)$. To prove submodularity, $\forall\ G = \{V, E, \{C_i\}$ and $\forall v \in V$, let $A \subseteq B \subset N^{in}(v)$, and $\forall u \in N^{in}(v) \backslash B$. Assume that there are a total of $n$ sectors in $G$. Let $A_i = A \cap C_i$ and $B_i = B \cap C_i$. Assume that $u \in C_k$. We have:

$$f_v(A) = 1 - \prod_{i=1}^{n}(1 \ - \sum_{s\in A_i} w(s,v))$$

$$= 1 - \prod_{i=1, i\neq k}^{n} (1 - \sum_{s\in A_i} w(s,v)) \cdot (1 - \sum_{s\in A_k} w(s,v))$$

$$= 1 - M\left(1 - \sum_{s\in A_k} w(s,v)\right)$$

Here we let $M = \prod_{i=1,i\neq k}^{n}(1 - \sum_{s\in A_i} w(s,v))$. Similarly, we have:

$$f_v(A \cup u) = 1 - M\left(1 - \sum_{s\in A_k \cup u} w(s,v)\right)$$

Thus:

$$f_v(A \cup u) - f_v(A) = Mw(u,v)$$

We define $N = \prod_{i=1,i\neq k}^{n}(1 - \sum_{s\in B_i} w(s,v))$. The same equation applies:

$$f_v(B \cup u) - f_v(B) = Nw(u,v)$$

Notice that we have $A_i \subseteq B_i$, so $M \geq N$, which implies $f_v(A \cup u) - f_v(A) \geq f_v(B \cup u) - f_v(B)$. Thus, $f_v()$ is submodular, and hence σ() is monotone increasing and submodular. ■

Theorem 5.1 thus implies that the Greedy method for influence maximization and its improved form such as CELF and CELF++ of Goyal et al. (2011) [29]. will achieve a $(1 - 1/e)$-approximation to the optimal set.

We next define the live-arc graph for MLTIN, which we will use in our experiments in Section 5.3 to improve computational efficiency.

**Definition 5.3 (Multi-layer live-arc graph**). Given a MILTIN $G = (V, E, \{C_i\})$ with layers $\{G_i = (V_i, E_i)\}$5. and arc weights $w(u, v)$, the multi-layer live-arc graph of $G$ is created as follows: Independently for each node $v$ and layer $G_i$, sample one $u \in N^{in}(v) \cap V_i$ with probability $w(u, v)$ (no node is chosen w.p. $1 - \sum_{u \in N^{in}(v) \cap V_i} w(u, v)$ if $\sum_{u \in N^{in}(v) \cap V_i} w(u, v)$<1). Only the edge $(u, v)$ is remains in the live-arc graph and all other arcs are removed. We denote the resulting multi-layer graph $G_L = (V, E_L)$. Given the seed set $S_0 \subseteq V$, for each $t \geq 1$, the nodes in $S_{t-1}$ will activate all inactive out-nodes in $G_L$. That is, the set of nodes that become active at time $t$ is: $\{v | v \notin S_{t-1}, u \in S_{t-1}, (u, v) \in E_L\}$.

Since we already use $S_t$ for MLTIN, we let $R_{G_L}^t(S_0)$ denote the active set for the multi-layer live-arc graph for $t \geq 1$.

**Theorem 5.2.** A MLTIN $G = \{V, E, \{C_i\}\}$ is equivalent to its live-arc graph using Definition 2.6. That is,

$$Pr(S_t = A_t | S_0 = A_0, \dots, S_{t-1} = A_{t-1})$$

$$= Pr\left(R_{G_L}^t(S_0) = A_t \middle| \begin{matrix} R_{G_L}^1(S_0) = A_1, \dots, R_{G_L}^{t-1}(S_0) \\ = A_{t-1}, \;\; S_0 = A_0 \end{matrix}\right)$$

$$\forall t > 0, A_0, \dots, A_{t-1}, A_t \subseteq V$$

**Proof.** $\forall t, A_0, \dots, A_{t-1}, A_t$. we only consider the case where $A_0 \subseteq A_1 \subseteq \dots \subseteq A_{t-1} \subseteq A_t$ and if $A_k = A_{k+1}$, all subsequent sets are the same. Otherwise both probabilities are 0. First, we consider the MLTIN, $\forall v \in A_t \backslash A_{t-1}$:

$$Pr(v \text{ activated at time } t\ |S_0 = A_0, \dots, S_{t-1} = A_{t-1})$$

$$= Pr(v \text{ activated by } A_{t-1} | v \text{ not activated by } A_{t-2})$$

$$= \frac{Pr(v \text{ activated by } A_{t-1}, v \text{ not activated by } A_{t-2})}{Pr(v \text{ not activated by } A_{t-2})}$$

$$= 1 - \frac{Pr(v \text{ not activated by } A_{t-1})}{Pr(v \text{ not activated by } A_{t-2})}$$

$$= 1 - \frac{\prod_{i=1}^{n}(1 - \sum_{u \in A_{t-1} \cap C_i} w(u,v))}{\prod_{i=1}^{n}(1 - \sum_{u \in A_{t-2} \cap C_i} w(u,v))}$$

Meanwhile, $\forall v \in V \backslash A_t$:

$$Pr(v \text{ not activated at } t | S_0 = A_0, \dots, S_{t-1} = A_{t-1})$$

$$= Pr(v \text{ not activated by } A_{t-1} | v \text{ not activated by } A_{t-2})$$

$$= \frac{Pr(v \text{ not activated by } A_{t-1})}{Pr(v \text{ not activated by } A_{t-2})}$$

$$= \frac{\prod_{i=1}^{n}(1 - \sum_{u \in A_{t-1} \cap C_i} w(u,v))}{\prod_{i=1}^{n}(1 - \sum_{u \in A_{t-2} \cap C_i} w(u,v))}$$

Thus, in MLTIN:

$$Pr(S_t = A_t | S_0 = A_0, \dots, S_{t-1} = A_{t-1})$$

$$= \prod_{v \in A_t \backslash A_{t-1}} (1 - \frac{\prod_{i=1}^{n}(1 - \sum_{u \in A_{t-1} \cap C_i} w(u,v))}{\prod_{i=1}^{n}(1 - \sum_{u \in A_{t-2} \cap C_i} w(u,v))})$$

$$\cdot \prod_{v \in V \backslash A_t} \frac{\prod_{i=1}^{n}(1 - \sum_{u \in A_{t-1} \cap C_i} w(u,v))}{\prod_{i=1}^{n}(1 - \sum_{u \in A_{t-2} \cap C_i} w(u,v))} \qquad (3)$$

Next, we check the same probability in the multi-layer live-edge graph, $\forall v \in A_t \backslash A_{t-1}$ :

$$Pr(v \text{ is reached by } S_0 \text{ in } t \text{ steps, but not in } t\ -1 \text{ steps}|$$
$$S_0 = A_0, .., R_{G_L}^{t-1}(S_0) = A_{t-1})$$
$$= Pr(v \text{ is reached by } S_0 \text{ in } t \text{ steps, but not in } t\ -1\ steps\ |$$
$$v \text{ is not reached by } S_0\ in\ t -\ 1 \text{ steps})$$
$$= 1 - \frac{Pr(\forall u \in A_{t-1}, (u,v)\ is\ not\ live)}{Pr(\forall u \in A_{t-2}, (u,v)\ is\ not\ live)}$$
$$= 1 - \frac{\prod_{i=1}^{n}(1 - \sum_{u\in A_{t-1\cap C_i}} w(u,v))}{\prod_{i=1}^{n}(1 - \sum_{u\in A_{t-2\cap C_i}} w(u,v))}$$

Meanwhile, $\forall v \in V \backslash A_t$:

$$Pr(v\ is\ not\ reached\ by\ S_0 \text{ in } t \text{ steps}|\ S_0\ = A_0, \dots, R_{G_L}^{t-1} = A_{t-1})$$
$$= \frac{Pr(\forall u \in A_{t-1}, (u,v)\ is\ not\ live)}{Pr(\forall u \in A_{t-2}, (u,v)\ is\ not\ live)}$$
$$= \frac{\prod_{i=1}^{n}(1 - \sum_{u\in A_{t-1\cap C_i}} w(u,v))}{\prod_{i=1}^{n}(1 - \sum_{u\in A_{t-2\cap C_i}} w(u,v))}$$

Thus, in the multi-layer live-edge graph:

$$Pr\left(R_{G_L}^{t}(S_0) = A_t \middle| S_0 = A_0, \dots, R_{G_L}^{t-1}(S_0) = A_{t-1}\right)$$
$$= \prod_{v\in A_t \backslash A_{t-1}} (1 - \frac{\prod_{i=1}^{n}(1 - \sum_{u\in A_{t-1\cap C_i}} w(u,v))}{\prod_{i=1}^{n}(1 - \sum_{u\in A_{t-2\cap C_i}} w(u,v))}$$

$$\cdot \prod_{v \in V \setminus A_t} \frac{\prod_{i=1}^{n}(1 - \sum_{u \in A_{t-1} \cap C_i} w(u,v))}{\prod_{i=1}^{n}(1 - \sum_{u \in A_{t-2} \cap C_i} w(u,v))} \tag{4}$$

Note Equations (3) and (4) are the same expressions. Thus the required equality holds. ■

Theorem 5.2 enables us to use the multi-layer live-arc graph defined in Definition 5.3 to efficiently estimate the influence function $\sigma(S_0)$. In each round of Monte Carlo simulation for a given MLTIN $G = (V, E, \{C_i\})$ and seed set $S_0$, the live-arc graph should be independently generated, and the total weight of nodes connected to $S_0$ provides a single-round estimate of $\sigma(S_0)$. Finally, these estimates should be averaged across all simulation rounds to get the final estimate of the influence $\sigma(S_0)$.

## 5.3 Experiments

We now apply our new framework on two real-world scenarios involving Metro Atlanta and Florida. In each scenario we choose some facilities from four CI sectors: Energy, Communication, Transportation, and Commercial/Public Facilities. We choose these sectors because their existence is physical and thus relatively easy to identify the interdependencies. For each scenario we built a MLTIN described in section IV to analyze the interdependencies. Our objective is to identify the most important nodes to protect. We begin by running the lazy greedy method by Goyal et al. (2011) [29] to expand the target set. To choose the local optimum at each step of the greedy method, we use the multi-layer live-edge graph proposed in section IV to do the Monte-Carlo simulation.

The data set we used for the experiments are accessible by public. The data only include the geographical locations of the facilities, where point-type facilities (substations, cellular base stations) are described as a geographical point and path-type facilities (roads, electricity transmission lines) are described by two coordinates. The end of path-type facilities may not exactly match the coordinates of point-type facilities, which means some slight modifications are made in pre-processing. See Table 3 and Table 4 for specific data sources. The commercial and public facilities data, including gas stations, are collected using Google Maps API.

### *5.3.1 Network Construction*

It is necessary to explain how we build the MLTIN given the CI facilities before presenting the experimental results. The CI facilities included are summarized in Table 4.

**Table 4 – CI facilities included in our experimental scenarios.**

| CI sector | Facilities included |
|---|---|
| Transportation | Major road intersections, Major roads |
| Communication | Cellular base stations |
| Energy | Electricity substations, Electricity transmission lines, Gas stations |
| Commercial/Public Facilities | City halls, Hospitals, Colleges/Universities, Emergency Medical services, Fire stations, Schools |

Notice that in Table 4 there are some facilities serving as the arcs/paths in its original network, like roads and electricity transmission lines. However, in our model it is also turned into vertices since they are also CI facilities.

We divide the CI sectors into three categories: the sectors with pre-built routes, the sectors without pre-built routes, and the sectors with only ending facilities. Meanwhile, in each layer there are within-sector edges that connect nodes in the same sector, and cross-sector edges that connects nodes in different sectors.

Our first task is to identify the within-sector edges and their weights. The first category includes the cellular station network, where the nodes do not rely on pre-built lines/routes to interact with each other. In the cellular station network we define a max reach distance $R$. If the distance between two stations is less than $2R$, we let there be an edge connecting both stations. This implies that users would turn to the other station if one of the stations is down. In both scenarios we let $R = 20km$ as it is the average coverage radius for a general macrocell cellular base station (*Mobile Base Stations* (2012) [40]). The weight on the edge from node $u$ to $v$ is set as $w(u, v) = \frac{L_{(u,v)}}{\sum_{u \in N^{in}(v)} L_{(u,v)}}$, where $L_{(u,v)}$ is the load transferred from node $u$ to $v$ when $u$ is down. This can be estimated using the area that $u$ is the closest node and $v$ is the second closest node. We made this assumption since the real load transfer from one station to another is not included in the data.

The road network and electricity network are considered in the second category. The electricity transmission lines and roads are treated as nodes in the network. If the flow data are given, like the AADT (average annual daily traffic) for roads (FDOT (2019) [26]), we let the weight of edge from $(u, v)$ to $u$ be $\frac{flow_{(u,v)}}{\sum_{v:(u,v) \in E} flow_{(u,v)}}$ ($(u, v)$ are edges in the original sector network, but vertices in MLTIN), considering that each edge should have flow-based influences on its vertices. Conversely, the weight from $u$ to $(u, v)$ is 0.5. If the flow data are not given, The weights on the edges are constant for each $u$ and sum to 1.

Finally, we need to determine the cross-sector arcs. The weights for these arcs are summarized in Table 5. For each component in the destination sector, there are 'count' number of edges going from the nearest components of the 'origin' sector, each edge is weighted by the number in the 'weight' column. For example, the first row in the Table 5 represents that for each substation, there are two edges from the nearest transportation nodes to it, each carries a weight of 0.05, implying that there is 0.9 probability that this substation would not be influenced by transportation directly at all. We acknowledge that the choices of weights might make significant difference in the results. Thus, we perform a sensitivity analysis on these weights in Appendix A. Note that there are no arcs originating from facilities sector since they are all end users. Before we show our results, Table 6 shows the weight assumptions of nodes for both scenarios. Recall that every node is assigned a weight to represent its relative importance in the network.

**Table 5 – Weight assumptions for the cross-sector edges.**

| origin | dest | weight | count |
|---|---|---|---|
| trans | sub | 0.05 | 2 |
| trans | gas | 0.2 | 2 |
| sub | trans | 0.2 | 1 |
| sub | gas | 0.8 | 1 |
| trans | faci | 0.2 | 2 |
| sub | faci | 0.8 | 1 |
| cell | trans | 0.1 | 1 |
| cell | sub | 0.1 | 1 |
| trans | cell | 0.05 | 2 |
| sub | cell | 0.2 | 1 |
| cell | faci | 0.1 | 1 |

**Table 6 – Node weight assumptions for Florida scenario.**

| Facilities | Reason |
|---|---|
| Cellular base stations | Population/ number of stations*2, 2 represents the places people go per day. |
| Road intersections | sum of AADT of all adjacent roads/2*1.5, 1.5 is the people per vehicle and 2 represents each AADT is shared by 2 intersections. |
| Roads | AADT on road*1.5 |
| Electricity transmission lines | Population/<br>number of lines*2, 2 is the places people go per day |
| Electricity substations | Total transmission line weights connected/2 |
| All facilities, including gas stations | population*1%/number of certain type of facilities, assuming 1% people needs it on a certain day |

### *5.3.2 Metro Atlanta Scenario*

In this scenario, we consider the metro Atlanta region which is defined as the area inside Interstate 285. The number of facilities in each CI sector is listed in Table A.1 in Appendix A. The map of the facilities, with and without the base map, is displayed in Figure 15 and Figure 16. The map processing API is QGIS 2.18 and the base map is Openstreetmap. The total number of nodes is 1145, and the number of transportation nodes, electricity nodes, cellular base stations nodes and other nodes are 152, 210, 2 and 781 respectively. We do not use any clustering method on the nodes since the network size is relatively small.

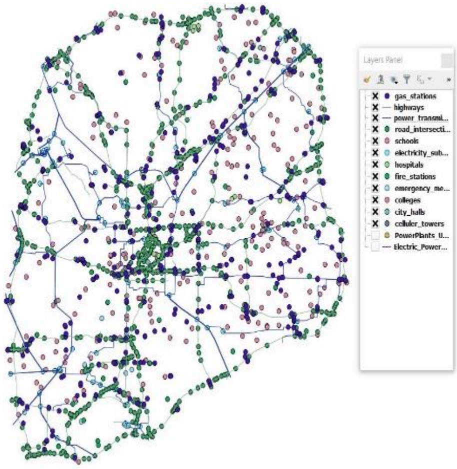

**Figure 15 – Facility map in metro Atlanta without base map.**

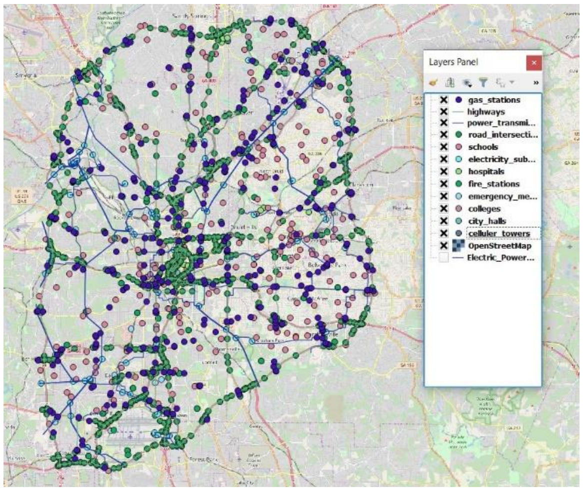

**Figure 16 – Facility map in metro Atlanta with base map.**

Our experiments choose 10 to 100 most influential nodes from the CI facility set with 10 as the step size. By ‘influencial’, we mean that if these nodes are removed, the removal would have the greatest negative impact on network performance. We let the number of chosen nodes in the target set change to see if the percentages of chosen nodes in each sector are relatively constant. The results are shown in Figure 17, Figure 18 and Figure 19. In Figure 17 we plotted the results choosing 10 nodes on the map. We observed that the selected nodes are generally uniformly scattered on the map, this is reasonable because if a node is already selected, selecting other nodes nearby would have diminished influence. In Figure 18 we show the percentage of nodes chosen in the transportation and electricity sectors. We do not show selected nodes in other two sectors because there too few chosen. When the number of selected nodes is over 30, the percentage of nodes chosen in each sector becomes relatively stable. And the percentage of nodes chosen in both sectors are both about 50%. In Figure 19 we show the percentage of total influenced nodes, we can see that even if the approach we used might not find an optimal solution, the influence curve is still concave, i.e. the sub-optimal solution preserves submodularity.

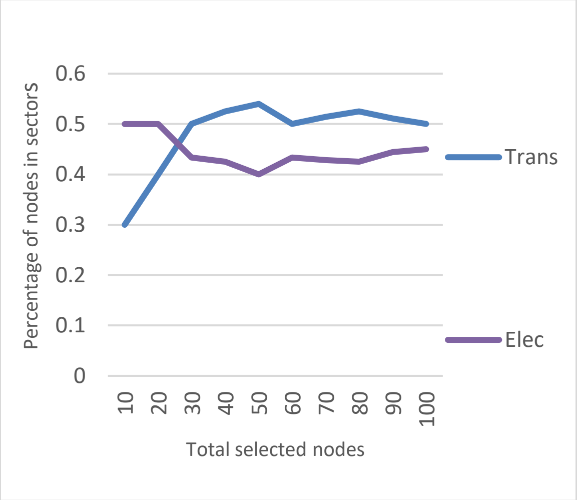

**Figure 17 – Percentage of nodes chosen in two sectors.**

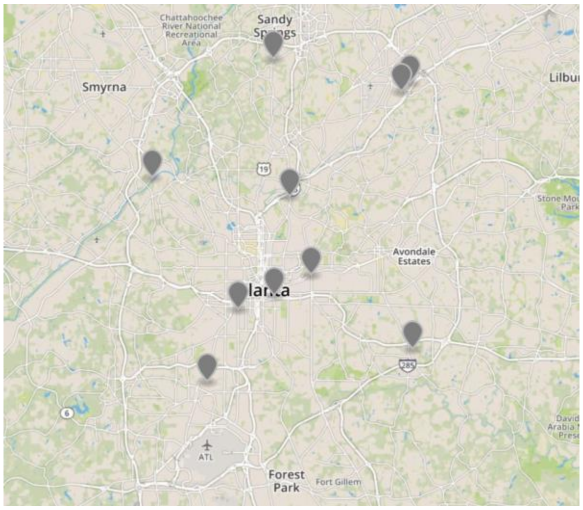

**Figure 18 – 10 selected nodes in the metro Atlanta scenario.**

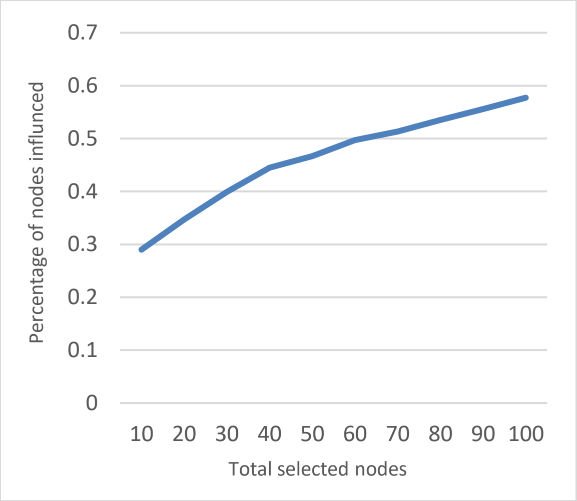

**Figure 19 – Percentage of influenced nodes.**

Also, as we mentioned above, the choices of the weights of the arcs were assumed to play an important role in the final results. Thus, we tried different weights options as a sensitivity analysis. The results are shown in Appendix B.

### *5.3.3 Florida Scenario*

In the second scenario, we apply our model to a larger geographical area than in the first scenario. The facilities included are displayed in Table A.2 in Appendix A. Note that we have included fewer categories of facilities relative to the first scenario but more facilities per category.

Since the Monte Carlo method is sensitive to the size of the problem and this scenario includes about 60K nodes. We use the K-means++ clustering (Arthur and Vassilvitskii (2007) [4]) to perform clustering before running the greedy method. This clustering method will cluster all nodes geographically and make sure each cluster contains a similar number of nodes. We do not consider it important if the sector distribution in each cluster matches the original data. In each cluster we choose the same number of nodes to form the final chosen set. We remark that in the Atlanta scenario, the chosen nodes scatter evenly on the map, and hence we anticipated that the clustering method would produce similar results. For this scenario, we now investigate when the number of the clusters vary, how the total weights of the influenced nodes change, and how many nodes in each sector will be selected. If there are no significant changes of percentage of nodes chosen in each sector, we can conclude that it is beneficial to have more clusters to reduce computational burden.

In this experiment we select the 1000 most influential nodes in the network and set the Monte Carlo simulation round as 1000. Figure 20 shows the percentage of influenced nodes with respect to the number of clusters used.

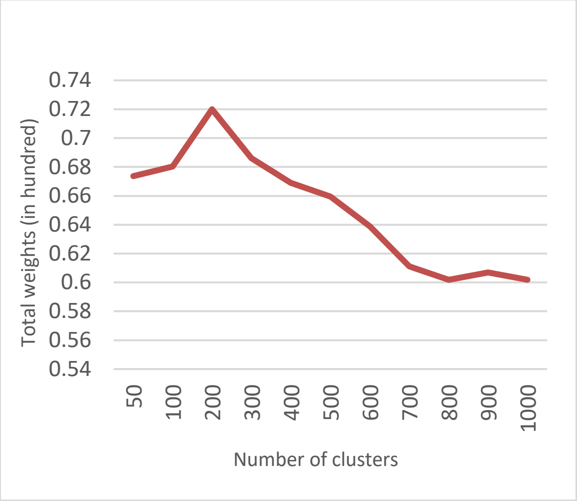

**Figure 20 – Percentage of influenced nodes with respect to the number of clusters used.**

In general, about 60% to 72% percent of the nodes (weighted) are influenced and the peak appears near 200 clusters. One of the probable causes for the low tail on the high number of clusters is that more clusters forces the selection of nodes due to stricter cluster constraints. It is also possible that the Greedy method can achieve better results for small size graphs, which makes the results for a small number of clusters unfavorable. In Figure 21 we plot how many nodes are selected in each sector with respect to the number of clusters. In general the number of nodes chosen in each sector is steady, this, together with Figure 20, justifies our guess that clustering would have little impacts to the final results. Over half of the nodes are from the electricity sector, which we expect since the electricity sector is necessary for all other facilities to operate (reflected by high edge weights from the electricity sector) and electricity is provided by only one electricity substation in our model. The sector in which the least nodes are selected is the communication sector since we only have the data of 662 cellular base stations. However, if we check the percentage of the facilities chosen out of all candidates in the sector, which is displayed in Figure 22, the smallest percentage is associated with the transportation sector, as access to any facility is usually not constrained to a single road. We also notice a big plunge for the percentage of cellular base stations chosen. This is because when the number of clusters are large, many clusters have no cellular base stations (662 stations in total while maximum 1000 clusters). Thus, in some clusters we only choose one cellular base station while we are supposed to choose more. And in some other clusters we can only choose nodes in other sector since there are no cellular base stations.

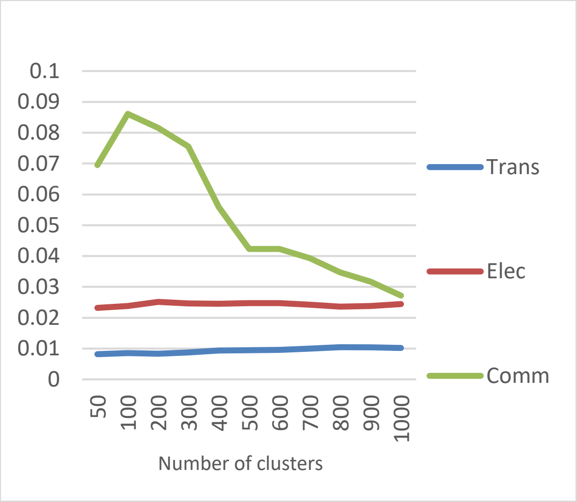

**Figure 21 – Percentage of nodes selected in each sector.**

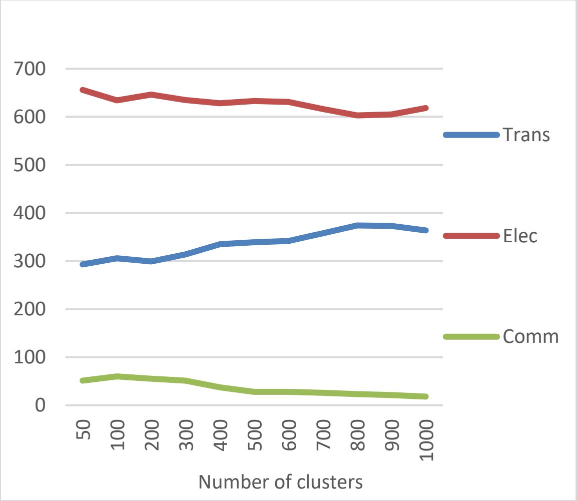

**Figure 22 – Nodes selected in each sector.**

## 5.4 Conclusions and Future Works

In this paper we have presented a new framework to analyze CI interdependency. This framework is a multi-layer extension of the classical linear threshold influence network. Our work has two novelties. First, it gives the definitions of the multi-layer interdependency network (MIN) and the multi-layer linear threshold influence network (MLTIN) based on the MIN. Second, we prove that the Greedy-based approach used for the linear threshold influence network can also be applied to the new framework in determining a lower bound.

The new framework is then applied to two scenarios – the metro Atlanta scenario and the Florida scenario. We have the following observations from the scenarios. First, the selected facilities tend to scatter uniformly on the map. Secondly, though the electricity nodes are chosen the most in terms of quantity, the communication sector has the highest percentage of chosen nodes out of all nodes in the sector.

Although we used the clustering methods to make the framework feasible for large-scale networks, other heuristics that do not use Monte Carlo simulation can still be investigated in the future (e.g., extending the Simpath algorithm presented by Goyal et al. (2011) [30]for the linear threshold network to MLTIN). Multiple approaches could be applied to some scenarios to compare the results and run times. On the other hand, some researchers proposed some interesting arguments towards the single-layer influence network. For example, Buldyrev et al. (2010) [12] found that the number of nodes needed to paralyze the interdependent network is higher than the single-layer network. Buldyrev et al. (2011) [13] studied a problem of failure of two interdependent networks in the case of identical degrees of mutually dependent nodes (correspondently coupled networks (CCN)). They found that the percentage of nodes needed to paralyze CCN is smaller than randomly coupled network. Our another future work is to test if these findings are true in our MLTIN.

# CHAPTER 6. SUMMARY

In this thesis, we focused on the widely studied influence network models and their applications and extensions. We began by introducing stochastic diffusion models and two common influence network models, Independent Cascade (IC) and Linear Threshold (LT). From these most basic models, we reviewed two more general models, General Cascade (GC) and General Threshold (GT), and introduced the concept of model equivalence. After introducing GC and GT, we introduced the influence maximization problem based on influence network models. To solve this problem, some important properties of IC and LT models were reviewed, i.e, their equivalence to respective live-arc graphs and submodularity. From these properties we reviewed the Lazy Greedy algorithm that utilizes these properties to find a solution to the influence maximization problem with a lower bound guarantee. At the end of Chapter 2, we introduced some extensions of the influence network models, such as the minimum target set selection problem and the competitive models.

In Chapter 3 and 4, we presented two applications of the single-layer IC and LT model. First, in Chapter 3, to mitigate the spread of HIV through human behavior, we build an IC model on a church-based social network to spread disease prevention knowledge and reduce the stigma of the disease. Our contribution is mainly on how to build the IC model with limited data to reflect the church features. In the experimental part, we found that while different connection/interaction structures between clergy and regular church participants would affect the degree of knowledge spread, the percentage of HIV/AIDS infected people who would eventually learn the disease information with the help of the church stays constant, regardless of the church size and whether there is information competition in the model. In Chapter 4, we used the LT model to study the cascading effects in the cellular base station network and attempted to find the stations which failure would cause the most damage in the network. In the experiments, we applied the framework to the U.S. national cellular base station network, since this network has over 20K stations. We used the K-means clustering method to first cluster the stations before finding the most critical ones to protect. We applied both the Monte Carlo method and the Simpath method (Goyal et al. (2011) [30]) and compared the results. In the sensitivity analysis, we compared the results across different number of clusters to check if the chosen stations are the same. It turns out that though only about 20% of the stations overlapped, most stations that do not overlap are close to each other when the number of clusters are small.

As an extension of Chapter 4, we designed a brand new multi-layer linear threshold network (MLTIN) in Chapter 5 to analyze the interdependency among multiple critical infrastructure sectors. We proved that such a network is a special case of the weighted general threshold network (WGTN) and preserves the submodularity and equivalence to the corresponding live-arc graphs. In the experimental part, we applied the new framework to two scenarios, metro Atlanta and Florida. In the metro Atlanta scenario, the network is unclustered and we observed that when the number of nodes chosen is high (over 30) in our model, the number of nodes chosen in each sector kept relatively constant. We also observed that the final influences from the chosen initial active set are submodular, even if the methods may not find the optimal solution. For the Florida scenario, since the network is very large, we used the K-means clustering method before we started selecting the nodes. As we found in Chapter 4, the best number of clusters is at the middle of the curve, using

30-50 clusters gave better results than using none or 100 clusters. This is because the Greedy method may not perform well for large networks and too many clusters may put too many restrictions on the nodes selection, which adversely affects the quality of nodes chosen in each cluster.

In summary, we applied the influence network models to solve the problem of finding the most influential entities in a network in social networks and critical infrastructure interdependencies. We hope that our results could provide useful references to the policy makers in order to better utilize resources to prevent disease outbreaks and cascading effects.

# APPENDIX A. CRITICAL INFRASTRUCTURES IN SECTION 5.3

In this appendix, we listed the critical infrastructure studied in metro Atlanta scenario and Florida scenario in Section 5.3.

**Table A.1 – Critical infrastructures in metro Atlanta scenario.**

| CI sectors | Facilities | Numbers |
|---|---|---|
| Transportation | Major roads | 59 |
| | Major road intersections | 93 |
| Energy | Electricity transmission lines | 114 |
| | Electricity substations | 96 |
| | Gas stations | 377 |
| Communication | Cellular base stations | 2 |
| Commercial and public facilities | City halls, Hospitals,<br>Colleges,<br>EMS,<br>Fire stations,<br>Schools | 406 |

**Table A.2 – Critical infrastructure in Florida scenario.**

| CI sectors | Facilities | Numbers |
|---|---|---|
| Transportation | Major roads | 19625 |
| | Major road intersections | 16183 |
| Energy | Electricity transmission lines | 22050 |
| | Electricity substations | 3253 |
| Commercial and public facilities | Hospitals | 152 |
| Communications | Cellular base stations | 662 |

# APPENDIX B. SENSITIVITY ANALYSIS FOR METRO ATLANTA SCENARIO

In this appendix we show the sensitivity analysis for the metro Atlanta scenario. Specifically, we let the weights of edges from transportation sector to other sectors increase up to 0.15 with 0.05 step size. And we do the same thing for the electricity sector. We do not do sensitivity analysis for the other two sectors. For the communication sector, there are too few cellular base stations in the area (only 2), and they are usually chosen in the first 10 nodes. So we do not expect any difference if we change the weights of edges from them. For commercial and public facilities, there are no edges from them at all.

The first thing we are concerned about is whether changes in the weights of the edges would result in significant changes in the percentage of nodes chosen in each sector. Figure B.1 and Figure B.2 show the percentage change when the total selected nodes increase. In these figures, Trans/o and Elec/o represent the percentage of nodes chosen in both sectors with original weight assumption, while Trans/trans+0.1 represents the percentage of nodes chosen in transportation sector when we increase all the weights of the edges from transportation sector to other sectors by 0.1. From the figures we can see that increasing the weights by 0.1 would results in the same level increments in the percentage, even though there are some outliers when total number of nodes selected is too few. Thus, we do need to pay attention to the choices of weights of edges in practice. Figure B.3 and Figure B.4 displayed the percentage of nodes chosen in each sector if the number of selected nodes is fixed at 50. In Figure B.3 the bars at 0.05 represent the percentage of nodes in both sectors when we increase the weights of edges from transportation nodes to other sectors by 0.05. And the same rule applies to Figure B.4 where we change the weights of edges from electricity sector. As we see in Figure B.1 and Figure B.2, in general increasing weights of a certain sector would increase the percentage of that sector, but the effects are not monotone in our experiments.

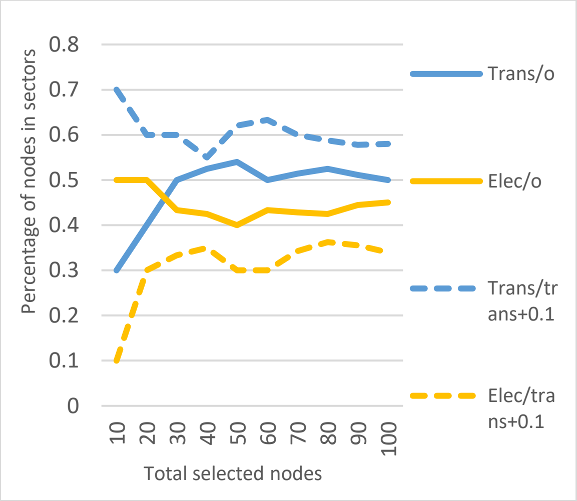

**Figure B.1 – Percentage of nodes chosen in two sectors with respect to the weights of electricity edges.**

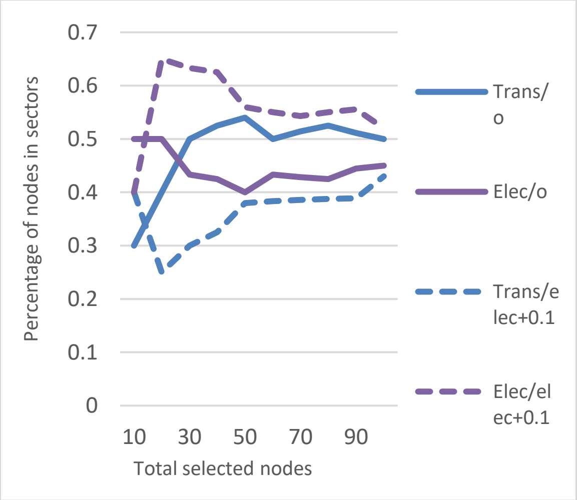

**Figure B.2 – Percentage of nodes chosen in two sectors with respect to the weights of transportation edges.**

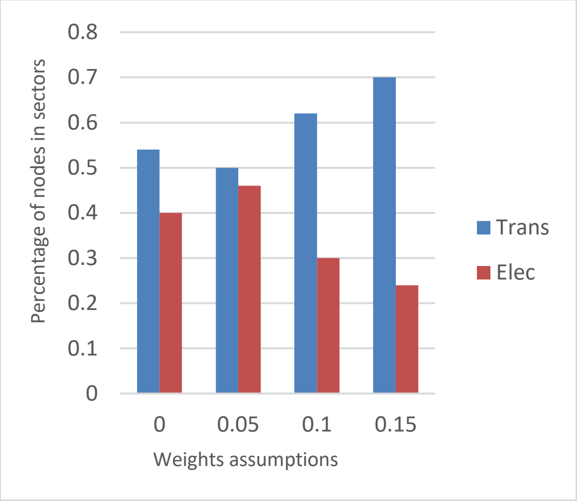

**Figure B.3 – Percentage of nodes chosen in two sectors with respect to the weights of electricity edges for 50 nodes selected in total.**

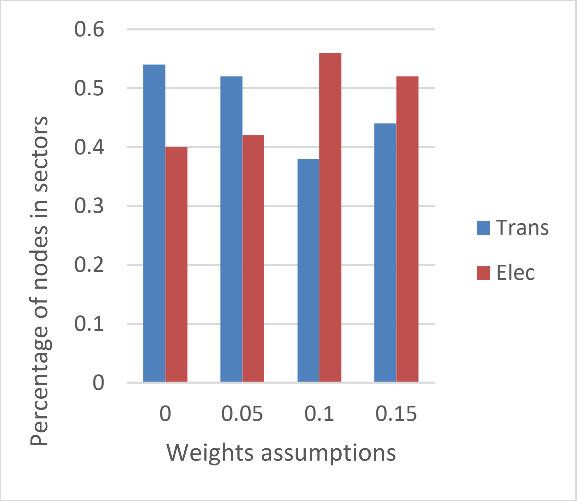

**Figure B.4 – Percentage of nodes chosen in two sectors with respect to the weights of transportation edges for 50 nodes selected in total.**

Another thing we want to investigate is the percentage of final influenced nodes. The results are shown in Figure B.5 and Figure B.6. The curve labelled as ‘0’ is the influences in original weights and 0.05 represents increasing the all the weights of the edges from a sector to other sectors by 0.1. From the figures we can get that increasing the weights do have significant effects on the final influenced nodes, and such effects are more notable when the number of selected nodes increases. On the other hand, all curves are concave in both figures, implying the submodularity property of the model.

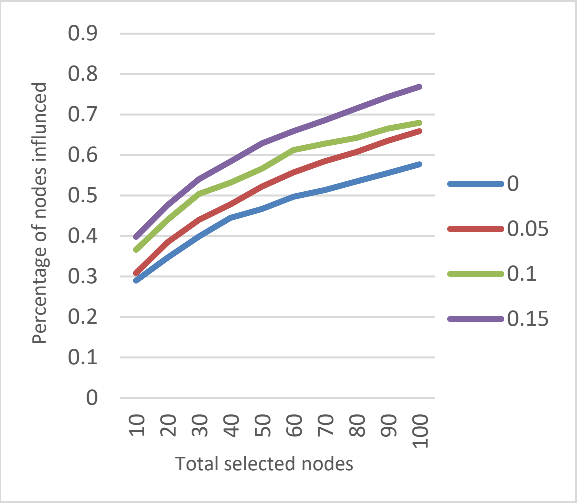

**Figure B.5 – Percentage of final influenced nodes with respect to the weights of transportation edges.**

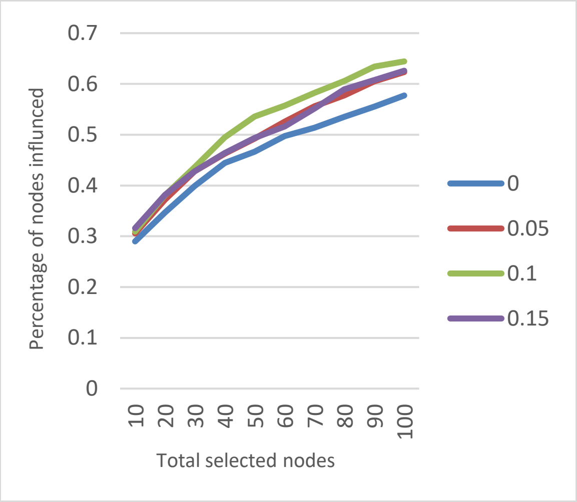

**Figure B.6 – Percentage of final influenced nodes with respect to the weights of electricity edges.**

# REFERENCES

1. Adar, E. and Adamic, L.A. *Tracking information epidemics in blogspace*. in *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*. 2005.

2. AIDSVu. *HIV Local Data: Atlanta*. 2019; Available from: https://aidsvu.org/local-data/united-states/south/georgia/atlanta/.

3. Amichai-Hamburger, Y. and Vinitzky, G., *Social network use and personality.* Computers in Human Behavior, 2010. **26**(6): p. 1289-1295.

4. Arthur, D. and Vassilvitskii, S., *k-means++: the advantages of careful seeding*, in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. 2007, Society for Industrial and Applied Mathematics: New Orleans, Louisiana. p. 1027–1035.

5. Arthur, D. and Vassilvitskii, S. *k-means++: The advantages of careful seeding*. in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. 2007. Society for Industrial and Applied Mathematics.

6. Barkan, S.E., *Sociology: Understanding and changing the social world, comprehensive edition*. 2010, New York: Flat World Knowledge, Inc.

7. Bluthenthal, R.N., Palar, K., Mendel, P., Kanouse, D.E., Corbin, D.E., and Derose, K.P., *Attitudes and beliefs related to HIV/AIDS in urban religious congregations: Barriers and opportunities for HIV-related interventions.* Social Science & Medicine, 2012. **74**(10): p. 1520-1527.

8. Bobashev, G.V., Goedecke, D.M., Feng, Y., and Epstein, J.M. *A Hybrid Epidemic Model: Combining The Advantages Of Agent-Based And Equation-Based Approaches*. in *2007 Winter Simulation Conference*. 2007.

9. Borodin, A., Filmus, Y., and Oren, J. *Threshold Models for Competitive Influence in Social Networks*. in *Internet and Network Economics*. 2010. Berlin, Heidelberg: Springer Berlin Heidelberg.

10. Bravo, G., Squazzoni, F., and Boero, R., *Trust and partner selection in social networks: An experimentally grounded model.* Social Networks, 2012. **34**(4): p. 481-492.

11. Budak, C., Agrawal, D., and El Abbadi, A. *Limiting the spread of misinformation in social networks*. in *Proceedings of the 20th international conference on World wide web*. 2011. ACM.

12. Buldyrev, S.V., Parshani, R., Paul, G., Stanley, H.E., and Havlin, S., *Catastrophic cascade of failures in interdependent networks.* Nature, 2010. **464**(7291): p. 1025-1028.

13. Buldyrev, S.V., Shere, N.W., and Cwilich, G.A., *Interdependent networks with identical degrees of mutually dependent nodes.* Physical Review E, 2011. **83**(1): p. 016112.

14. Capasso, V. and Serio, G., *A generalization of the Kermack-McKendrick deterministic epidemic model.* Mathematical Biosciences, 1978. **42**(1): p. 43-61.

15. CDC. *HIV in the United States and Dependent Areas*. 2018; Available from: https://www.cdc.gov/hiv/statistics/overview/ataglance.html.

16. Chen, W., Collins, A., Cummings, R., Ke, T., Liu, Z., Rincon, D., Sun, X., Wang, Y., Wei, W., and Yuan, Y. *Influence maximization in social networks when negative opinions may emerge and propagate*. in *Proceedings of the 2011 siam international conference on data mining*. 2011. SIAM.

17. Chen, W., Yuan, Y., and Zhang, L. *Scalable Influence Maximization in Social Networks under the Linear Threshold Model*. in *2010 IEEE International Conference on Data Mining*. 2010.

18. *Churches in metro Atlanta area*. 2006; Available from: http://hirr.hartsem.edu/cgi-bin/mega/db.pl?db=default&uid=default&view_records=1&ID=*&sb=4&State=GA.

19. Dastin, J. *Power outage at Delta causes flight cancellations, delays.* 2016; Available from: https://www.reuters.com/article/us-delta-air-outages/power-outage-at-delta-causes-flight-cancellations-delays-idUSKCN10J0VP.

20. DHS. *Cellular tower data*. 2018; Available from: https://hifld-geoplatform.opendata.arcgis.com/datasets/cellular-towers.

21. Directive, P.P., *21 (2013) Critical infrastructure security and resilience.* The White House Office of the Press Secretary.

22. Domingos, P. and Richardson, M., *Mining the network value of customers*, in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. 2001, Association for Computing Machinery: San Francisco, California. p. 57–66.

23. Dudenhoeffer, D.D., Permann, M.R., and Manic, M. *CIMS: A framework for infrastructure interdependency modeling and analysis*. in *Proceedings of the 38th conference on Winter simulation*. 2006. Winter Simulation Conference.

24. Dunbar, R.I.M. and Spoors, M., *Social networks, support cliques, and kinship.* Human Nature, 1995. **6**(3): p. 273-290.

25. *Fast Facts about American Religion*. 2006; Available from: http://hirr.hartsem.edu/research/fastfacts/fast_facts.html#megamap.

26. FDOT. *Florida Traffic Online*. 2019 [cited 2018 10/17]; Available from: https://tdaappsprod.dot.state.fl.us/fto/.

27. Goyal, A., Bonchi, F., and Lakshmanan, L.V.S., *Learning influence probabilities in social networks*, in *Proceedings of the third ACM international conference on Web search and data mining*. 2010, Association for Computing Machinery: New York, New York, USA. p. 241–250.

28. Goyal, A., Bonchi, F., Lakshmanan, L.V.S., and Venkatasubramanian, S., *On minimizing budget and time in influence propagation over social networks.* Social Network Analysis and Mining, 2013. **3**(2): p. 179-192.

29. Goyal, A., Lu, W., and Lakshmanan, L.V.S., *CELF++: optimizing the greedy algorithm for influence maximization in social networks*, in *Proceedings of the 20th international conference companion on World wide web*. 2011, Association for Computing Machinery: Hyderabad, India. p. 47–48.

30. Goyal, A., Lu, W., and Lakshmanan, L.V.S. *SIMPATH: An Efficient Algorithm for Influence Maximization under the Linear Threshold Model*. in *2011 IEEE 11th International Conference on Data Mining*. 2011.

31. Hadaway, C.K. and Marler, P.L., *Did you really go to church this week? Behind the poll data.* The Christian Century, 1998. **115**(14): p. 472.

32. Hill, R.A. and Dunbar, R.I.M., *Social network size in humans.* Human Nature, 2003. **14**(1): p. 53-72.

33. Johansson, J. and Hassel, H., *An approach for modelling interdependent infrastructures in the context of vulnerability analysis.* Reliability Engineering & System Safety, 2010. **95**(12): p. 1335-1344.

34. Kempe, D., Kleinberg, J., and Tardos, É., *Maximizing the spread of influence through a social network*, in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2003, Association for Computing Machinery: Washington, D.C. p. 137–146.

35. Kermack, W.O. and McKendrick, A.G., *A contribution to the mathematical theory of epidemics.* Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character, 1927. **115**(772): p. 700-721.

36. Khosrovani, M., Poudeh, R., and Parks-Yancy, R., *How African-American ministers communicate HIV/AIDS-related health information to their congregants: a survey of selected black churches in Houston, TX.* Mental Health, Religion & Culture, 2008. **11**(7): p. 661-670.

37. Lee II, E.E., Mitchell, J.E., and Wallace, W.A., *Restoration of services in interdependent infrastructure systems: A network flows approach.* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2007. **37**(6): p. 1303-1317.

38. Lee II, E.E., Wallace, A., Mitchell, J., Mendon, D., and Chow, J., *Managing disruptions to critical interde-pendent infrastructures in the context of the 2001 worldtrade center attack.* Beyond September, 2003. **11**: p. 165-198.

39. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., and Glance, N., *Cost-effective outbreak detection in networks*, in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2007, Association for Computing Machinery: San Jose, California, USA. p. 420–429.

40. *Mobile Base Stations*. 2012; Available from: https://mobilenetworkguide.com.au/mobile_base_stations.html.

41. Moore, D., Onsomu, E.O., Timmons, S.M., Abuya, B.A., and Moore, C., *Communicating HIV/AIDS Through African American Churches in North Carolina: Implications and Recommendations for HIV/AIDS Faith-Based Programs.* Journal of Religion and Health, 2012. **51**(3): p. 865-878.

42. Mossel, E. and Roch, S., *On the submodularity of influence in social networks*, in *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. 2007, Association for Computing Machinery: San Diego, California, USA. p. 128–134.

43. *National Infrastructure Protection Plan (NIPP) Communications Sector-Specific Plan for 2015*, D.o.H. Security, Editor. 2015.

44. Nemhauser, G.L., Wolsey, L.A., and Fisher, M.L., *An analysis of approximations for maximizing submodular set functions—I.* Mathematical Programming, 1978. **14**(1): p. 265-294.

45. Newman, M.E.J., *Scientific collaboration networks. I. Network construction and fundamental results.* Physical Review E, 2001. **64**(1): p. 016131.

46. Newport, F., *Frequent church attendance highest in Utah, lowest in Vermont.* Gallup. February, 2015. **17**: p. 2015.

47. Ouyang, M., *Review on modeling and simulation of interdependent critical infrastructure systems.* Reliability engineering & System safety, 2014. **121**: p. 43-60.

48. Parker, J. *A flexible, large-scale, distributed agent based epidemic model*. in *2007 Winter Simulation Conference*. 2007.

49. Ramachandran, V., Shoberg, T., Long, S., Corns, S., and Carlo, H., *Identifying Geographical Interdependency in Critical Infrastructure Systems Using Open Source Geospatial Data in Order to Model Restoration Strategies in the Aftermath of a Large-Scale Disaster.* International Journal of Geospatial and Environmental Research, 2015. **2**(1): p. 4.

50. Reilly, A.C., Samuel, A., and Guikema, S.D., *"Gaming the System": Decision Making by Interdependent Critical Infrastructure.* Decision Analysis, 2015. **12**(4): p. 155-172.

51. Richardson, M. and Domingos, P., *Mining knowledge-sharing sites for viral marketing*, in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2002, Association for Computing Machinery: Edmonton, Alberta, Canada. p. 61–70.

52. Rinaldi, S.M. *Modeling and simulating critical infrastructures and their interdependencies*. in *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*. 2004.

53. Rinaldi, S.M., Peerenboom, J.P., and Kelly, T.K., *Identifying, understanding, and analyzing critical infrastructure interdependencies.* IEEE Control Systems Magazine, 2001. **21**(6): p. 11-25.

54. Romero, D.M., Galuba, W., Asur, S., and Huberman, B.A. *Influence and Passivity in Social Media*. 2011. Berlin, Heidelberg: Springer Berlin Heidelberg.

55. Ruan, S. and Wang, W., *Dynamical behavior of an epidemic model with a nonlinear incidence rate.* Journal of Differential Equations, 2003. **188**(1): p. 135-163.

56. Stiller, J. and Dunbar, R.I.M., *Perspective-taking and memory capacity predict social network size.* Social Networks, 2007. **29**(1): p. 93-104.

57. Svendsen, N.K. and Wolthusen, S.D., *Connectivity models of interdependency in mixed-type critical infrastructure networks.* Information Security Technical Report, 2007. **12**(1): p. 44-55.

58. Wang, C., Chen, W., and Wang, Y., *Scalable influence maximization for independent cascade model in large-scale social networks.* Data Mining and Knowledge Discovery, 2012. **25**(3): p. 545-576.

59. Xiao, D. and Ruan, S., *Global analysis of an epidemic model with nonmonotone incidence rate.* Mathematical Biosciences, 2007. **208**(2): p. 419-429.

60. Yadav, A., Marcolino, L.S., Rice, E., Petering, R., Winetrobe, H., Rhoades, H., Tambe, M., and Carmichael, H. *Preventing HIV spread in homeless populations using PSINET*. in *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015.

61. Zio, E. and Sansavini, G., *Modeling interdependent network systems for identifying cascade-safe operating margins.* IEEE Transactions on Reliability, 2011. **60**(1): p. 94-101.