Modeling and Twitter-based Surveillance of Smoking Contagion

Gaurav Tuli

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

Doctor of Philosophy
in
Computer Science and Applications

Madhav V Marathe, Co-Chair
Samarth Swarup Co-Chair
Anil Vullikanti
Naren Ramakrishnan
Kiran Lakkaraju

May 01, 2015
Blacksburg, Virginia

Keywords: Tobacco Epidemic, Twitter-based Surveillance, Smoking-related Messaging, Electronic-cigarette, Networks, Control of Contagion Processes, Modeling and Simulation

Modeling and Twitter-based Surveillance of Smoking Contagion

Gaurav Tuli

(ABSTRACT)

Nicotine, in the form of cigarette smoking, chewing tobacco, and most recently as vapor smoking, is one of the most heavily used addictive drugs in the world. Since smoking imposes a significant health-care and economic burden on the population, there have been sustained and significant efforts for the past several decades to control it. However, smoking epidemic is a complex and "policy-resistant" problem that has proven difficult to control. Despite the known importance of social networks in the smoking epidemic, there has been no network-centric intervention available for controlling the smoking epidemic yet.

The long-term goal of this work is the development and implementation of an environment needed for developing network-centric interventions for controlling the smoking contagion. In order to develop such an environment we essentially need: an operationalized model of smoking that can be simulated; to determine the role of online social networks on smoking behavior; and actual methods to perform network-centric interventions. The objective of this thesis is to take first steps in all these categories. We perform Twitter-based surveillance of smoking-related tweets, and use mathematical modeling and simulation techniques to achieve our objective.

Specifically, we use Twitter data to: infer sentiments on smoking and electronic cigarettes; estimate the proportion of user population that gets exposed to smoking-related messaging that is underage; and identify statistically anomalous clusters of counties where people discuss about electronic cigarette a lot more than expected. In other work, we employ mathematical modeling and simulation approach to study how different factors such as addictiveness and peer-influence together contribute to smoking behavior diffusion, and also develop two methods to stymie social contagion. This lead to a total of four smoking contagion-related studies. These studies are just a first step towards the development of a network-centric intervention environment for controlling smoking contagion, and also to show that such an environment is realizable.

This work was partially supported by Defense Threat Reduction Agency's CNIMS grant HDTRA1-11-D-0016-0001, by Department of Energy grant DE-SC0003957, by National Institutes of Health's MIDAS grant 5U01GM070694-11, and by National Science Foundation's NetSE grant CNS-1011769.

Dedication

I dedicate my dissertation work to my loving Mother, Prabha Khanna, caring father Yash Pal Tuli, and beloved wife Asmita Gharat.

I have a special feeling of gratitude for my very humble parents. Their encouraging words and love kept me going during this journey. I especially salute my mother, a gold medalist academician of late 1960s, who herself couldn't complete her PhD 42 years ago due to familial and social pressure, but supported me through thick and thin to complete mine hiding all her pains and worries from me sitting 10,000 miles away.

I also dedicate this dissertation and give my sincere thanks to my wife, who was physically with me throughout this journey and has showered unconditional love and support over me. She used to hold my hands and comfort me whenever I used to feel low. I was extremely lucky to found her and have her by my side during this journey. I wish to be a partner as supportive and as understanding as hers.

Acknowledgments

I am very grateful to my committee. Each of them gave me valuable help and guided me along the way. A special thanks to my co-advisor, Dr. Samarth Swarup who introduced me to social contagion processes, and gave me ideas to pursue smoking contagion. In the last five years, not only he enhanced my technical strengths, but also instilled various professional skills in me. It was because of his countless hours of teaching and encouraging me, and patience throughout the process that made this possible.

My co-advisor, Dr. Madhav V Marathe, also guided and inspired me throughout this journey. Every interaction with him shaped my thesis and improved my quality of work. This thesis would not have been possible without him and without freedom, encouragement, and direction he has given me. I am also very grateful to him for providing such a stimulating and interactive environment at NDSSL.

Each of my colleagues at NDSSL deserves my deep gratitude. I especially like to thanks Dr. Christopher J Kuhlman for introducing me to complex contagions and computational simulations. He was virtually my third co-advisor who mentored and helped me in many of my studies. It was not possible without you, Chris. I also thank my academic sister Nidhi Parikh for always inspiring me and patiently listing to my boring practice talks. My sincere thanks also go to my Coffee Club friends Jae-Seung Yeom and Ashwin Aji, who kept me motivated and awake when it was needed the most.

My friends also played an extremely important role in this whole journey. My deepest gratitude goes to Abhijit and Sheetal for always being with me as my family. They stood by me in all the difficult times and encouraged me. I also have had an amazing group of friends at Virginia Tech that made this entire journey very enjoyable. I especially thank Jaideep Pandit, Balaji Subramaniam, and Hemant Bishnoi for countless hours of fun and late-night logical discussions. I will cherish all the memories with my friends for life.

I like to thank my parents and wife for their unconditional support and love. I thank my sisters Sonika, Priyanka, and Ruchika; and my brother-in-laws Pankaj, Vishrut, and Amit for their encouragement and unconditional support. Priyanka and Vishrut played a very important role from the beginning to this end, without which this was not possible. I also thank my nieces Simran, Suniti, and Anoushka; and nephews Daksh and Aarush for their good wishes and innocent talks that kept me pursuing my dream with a smiling face.

Last but not the least, I thank Virginia Tech for giving me a stimulating, encouraging and fun-filled environment to fulfill my dream. During my stay at Blacksburg, I not only hone

in my technical skills but also found my beautiful life partner and made many good friends for life. Thanks for bringing us all together.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Nicotine, in the form of tobacco smoking, chewing tobacco, and most recently in the form of vapors, is one of the most heavily used addictive drugs [139], and the leading preventable cause of disease, disability, and death across the globe [198]. Nearly six million deaths are associated with tobacco use each year globally — more than five million are due to the direct tobacco use and around 0.6 million are the result of breathing second-hand smoke [199]. Some regions in the world are more heavily affected by smoking epidemic than the others [64]. Although the bigger proportion of smoker population resides in the developing countries, WHO recent report shows that the 16% of the total mortality in both Americas and European region (highest in the world) are associated with tobacco use [198]. Smoking also imposes a significant health-care and economic burden on the population. In the United States, these burdens are estimated at $97 billion in productivity losses from premature death, and $96 billion in health-care expenditures annually [52].

To control this epidemic, there have been sustained and significant efforts for past several decades world wide [91, 141]. The countries are coming together [156], and as well working independently [55] to systematically curb the use of tobacco and overall practice of smoking. Most of these efforts are focused to protect adolescents from smoking initiation due to their vulnerability and known bad effects of nicotine on the developing brain [117, 163]. Despite these efforts, smoking prevalence among youth and adult smokers has declined very slowly. In the United States, it has only declined from 45% to 21% in the past 45 years [56, 62].

Controlling the smoking epidemic has become even more challenging due to social media

and introduction of new smoking products, such as electronic cigarette and hookah. In the age of Web 2.0, the online and offline social worlds of adolescents are merging at the highest rate. About 75% of the teens on Internet use social media (such as social networking sites, micro-blogging sites, forums etc) to connect and communicate with their friends [114]. Social media platform such as Twitter is becoming a venue were people talk and share about their smoking-related activities, and also get exposed to pro-smoking advertisements [85, 130]. Therefore due to such heavy smoking-related activity sharing and the current laxity in the restrictions on internet-based marketing, many of the users of these services may get heavily exposed to smoking-related messaging. Similarly, introduction of the novel smoking products in the market has been alluring the population to smoking initiation. For example, nicotine consumption via electronic cigarette and other vaping devices (henceforth collectively denoted as e-cig) has become three times more popular among adolescents within the last few years [23]. The strong evidence of their popularity has also been found in many survey studies [45].

These consideration provide us ample motivation to study the complex phenomenon of smoking contagion in more details. The initiation, continuation, and cessation of smoking has been reported to depend on multiple factors. Various studies have shown a strong correlations of smoking behavior, at least among adolescents and young adults, with three key factors: peer-influence, addictiveness, and exposure to pro- and anti-smoking marketing. Considering these factors, the tobacco control efforts by the governmental agencies generally aim for a broad impact either through policies changes (such as lowering nicotine levels or increasing taxes) or through public health campaigns (such as via television, print media, etc). There is also a big body of research in understanding the role of social networks in the spread of smoking behavior and their importance is well established. However, there has been no work reported so far for developing an environment for implementing network-centric interventions for controlling the smoking epidemic.

1.2 Thesis Overview and Contributions

The long-term goal of this work is the development and implementation of an environment needed for developing network-centric interventions for controlling the smoking epidemic. In order to develop such an environment we essentially need three main components: (1) surveillance tools and techniques to determine the role of online social networks on smoking behavior, (2) an operationalized model of smoking that can be simulated, and (3) actual methods to perform network-centric interventions. The objective of this thesis is to take first steps in all these categories. An illustration of long-term and achieved goals of this thesis are presented in Figure 1.1. We perform Twitter-based surveillance of smoking-related tweets, and use mathematical modeling and simulation techniques to achieve our objective.

Specifically, we take a data-driven machine learning approach and used Twitter data: to estimate the proportion of user population that gets exposed to smoking-related messaging that is underage, to infer sentiments on smoking and electronic cigarettes, and to identify

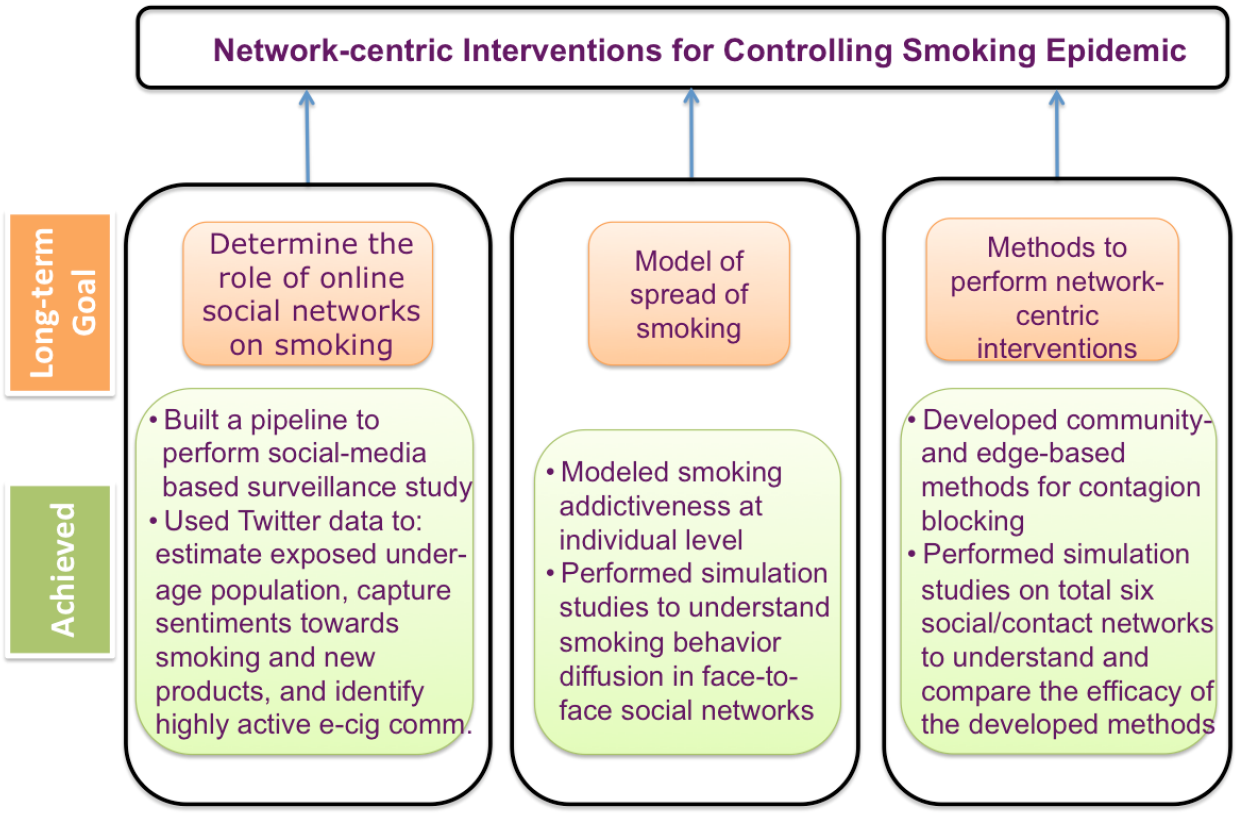


Figure 1.1: Long-term goals and overall contribution of the thesis

statistically anomalous clusters of counties where people discuss about electronic cigarette a lot more than expected. We present the details of a pipeline of software components that we built to perform Twitter-based surveillance studies for the thesis in the next section.

In other work, we employ mathematical modeling and simulation approach to study how different factors, such as addictiveness and peer-influence together contribute towards smoking behavior diffusion, and also develop two methods to stymie social contagion. This lead to a total of four smoking contagion-related studies — two Twitter-based surveillance studies, and two contagion modeling and simulation-based studies. These studies are just a first step towards the development of a network-centric intervention environment for controlling smoking contagion, and also to show that such an environment is realizable. I will discuss each of these studies and main contributions in brief next. A detail list of contributions of each study is presented in their respective chapters.

In our first Twitter-based surveillance studies, we focus on the adolescent Twitter users and investigate their exposure to smoking-related messaging. Since smoking is a social contagion that spreads through social fabric, understanding the effect of online social world on its diffusion is an important problem. Twitter is a extremely suited platform for such a study because teens' Twitter use has grown three folds in the past three years and tweets of 76% teen users are publicly available [114, 123]. In this study, we employ machine learning tools

and techniques to first identify pro-smoking and pro-marijuana tweets sent on Twitter. We then identify the users who get exposed to these tweets and infer their age group using supervised learning with Twitter data. Finally, we estimate the exposure rate for under age 18 users to smoking-related tweets employing a technique from the literature. Our analysis shows that smoking-related messages on Twitter are overwhelmingly pro-smoking, and also that a significant number of adolescent users receive and read multiple of these messages. However, a further study is needed to quantify the effect of pro-smoking messages on smoking initiation and continuation.

In the next study, we focus on electronic cigarette-related tweets to identify and analyze their hot-spots in the United states. A hot-spot of e-cig tweets can be defined as a statistically anomalous cluster of e-cig tweets in space and time i.e., a cluster that observe way more number of tweets than expected. Given the sudden multi-fold rise in the popularity of e-cig, it is important to identify the most active clusters of e-cig tweets and understand the sentiments and age-group they are composed of. We used spatiotemporal scanning [105] of non-commercial geo-tagged e-cig tweets to identify these clusters. This required us to first identify the non-commercial e-cig tweets in the US, and then to perform spatiotemporal scan statistics using the spatial location and time stamp of the tweets to obtain the anomalous clusters across space and time. At last, we analyze the sentiments of the tweets and age-group of the users inside these anomalous clusters. We again used machine learning tools and techniques to perform most of the tasks in the study. The results from this study suggest that three quarters of the e-cig tweet spatiotemporal hot-spots contain more pro-ecig tweets and more under-18 users as compared to the national averages of the same. Majority of these hot-spots are located on the west coast in the US.

The third study was designed to investigate the combined effect of peer-influence and addictive nature of tobacco smoking on the prevalence of smoking. Both peer-influence and addictiveness of smoking have been reported to have a big impact on smoking contagion. We study the effect of both the factors together by first developing a model that represents the addictivness of smoking behavior at individual level, and then performing an agent-based simulation of the model on a social network to replicate the effect of peer-influence. We represent population as network for the simulation study, where nodes represent individuals and edges corresponds to pairwise interactions. The simulation study of the proposed model was performed on a time-varying social network of a heavily cited longitudinal study in public health literature — *t*he Framingham Heart Study [49]. The results from the study show the presence of the two different *societal* thresholds to start and stop the smoking epidemic. This finding may explain the slow decline of smoking prevalence in the population. We also able to produce a qualitatively matching smoking prevalence curve using our model.

In the fourth and final study, we investigate a well-motivated problem of social contagion blocking. The literature on smoking contagion has shown that communities and connections play a very important role. For example, Christakis et al. [30] found that individuals start and stop smoking in groups, and also it is well established that some ties (or edges) are more important with respect to smoking initiation and cessation. Therefore, we investigate two primary themes for blocking social contagion in our study: (a) use community structure to contain a social contagion at community boundaries, and (b) consider weighted edges

in developing the technique to stymie the contagion. However, in the view of the non-approximability results we develop a set of heuristics for blocking the social contagions. The devised community-based blocking heuristic is hybrid in nature i.e., it uses both network structure (a proactive measure) and contagion dynamics (a reactive measure) to identify the critical nodes in the network. The edge-based blocking heuristic consider weights on edges in the network and a given budget to identify a critical edge set that leads to a small spread size. The heuristics are evaluated utilizing a rigorous set of computational (simulation) experiments of blocking the contagion propagation on the social networks from the literature . The selected social networks are at least five times greater in terms of numbers of nodes and an order of magnitude greater in numbers of edges than those used in previous contagion blocking studies.

1.3 Pipeline for Twitter-based Surveillance Studies

Another technical contribution of this work is a pipeline of the software components that were developed and used to perform the smoking-related surveillance studies on Twitter data. We built various components to handle all necessary steps, from data gathering through result visualization. Putting these components together gives us a pipeline that can be easily modified to perform surveillance studies on the data from any other social media platforms, or to accommodate semi-supervised or unsupervised learning tasks.

An illustration of the pipeline is shown in Figure 1.2. We briefly discuss the components of this pipeline in this section. The presented pipeline shows the steps involved in the supervised learning tasks for Twitter-based studies performed for this thesis.

1. **Data Ingestion and Archiving**

   The first component of the pipeline takes care of gathering and storing the data from the social media platforms for later use. It is one of the simplest yet most challenging task given the velocity, dynamicity, and volume of the data that can be generated by the social media platforms. Therefore, we need a smarter way to query and collect the data. For example, if we are gathering Twitter or Facebook data, and our data collecting queries are not properly formed, Terabytes of irrelevant data can be pushed into the system. Similarly, the whole system may fail if there is a sudden eruption of posts in certain area due to an unexpected event. Hence, the data should be gathered in small chunks using project-specific queries, and the system should be scalable and resilient to handle an unexpected volume of incoming data.

   We have gathered different types of data from Twitter using their APIs. It includes tweet data, user data or URL data. The tweet data is the tweet text along with various types of metadata; the user data consist of user details, such as user id, screen name, friends count, follower count etc.; and the URL data is mainly the information about the shared URL in the tweets. Based on a project requirement, we specifically collect:

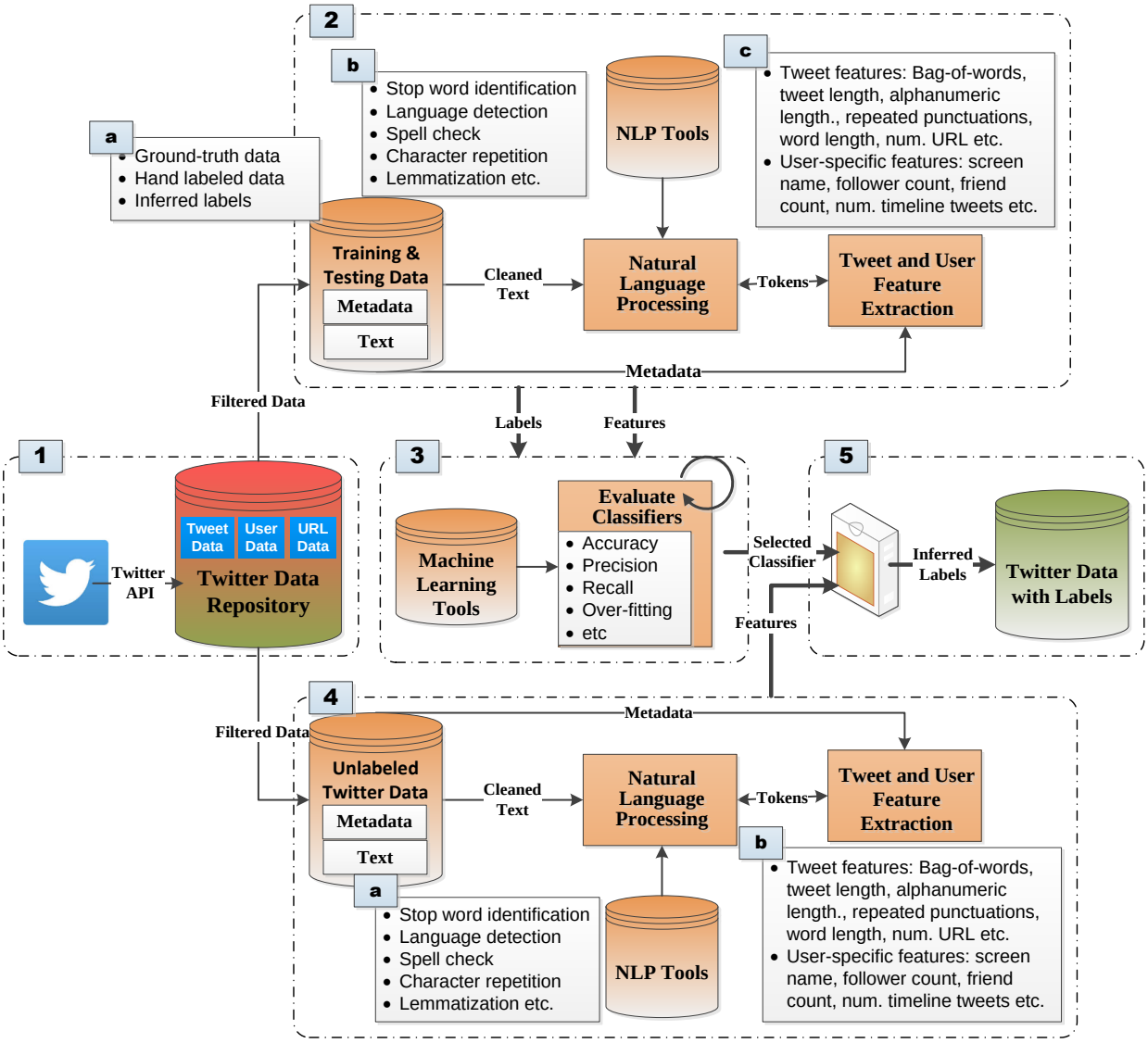


Figure 1.2: An illustration of the pipeline used in the Twitter-based surveillance studies in the thesis

tweets containing a set of keywords, or the timeline tweets of a set of users, or we can only collect personal or network information about a set of users.

The gathered data then gets archived in a way that is easy to access, that minimizes the memory footprint, and that expedites searching the data. The non-relational or NoSQL databases are well suited for storing social media data. They are often a good choice for storing tweets because they offer very quick write speeds, fast querying, and can easily distribute large data sets across a cluster of servers for parallel processing. NoSQL databases are also known for their ability to scale easily, which is very important for a system that need to store and handle an unexpected incoming data. MongoDB is most widely used and most recommended NoSQL databases. It is

a document-based database that uses documents instead of records in tables to store data. These documents look just like JSON objects with key-value pairs.

2. **Classifier Training**

   The next component is the most important part for the supervised learning pipeline. The primary goal of this pipeline is to learn the weights of the various features of the data and apply this learning to infer/predict the labels for the new data. This component has the following three sub-components.

   a. Label training data: First, we select the specific data about tweet or user profile from the Twitter data repository based on our application requirement. In our two Twitter-based studies, we used a combination of: tweet text, tweet posted date, user timeline tweet counts, user location, user language, profile creation date, and user friend and follower lists and counts.

   Next, we obtain the ground truth data (also called gold standard data) for the supervised-learning task. It is a set of data points for which we already know labels. These labels can either be already known or provided, or obtain using manual curation, or can be inferred using any other provided data (e.g,. inferred flu case count per county using over-the-counter medicine purchase records). We obtained the labeled data via in-house and Amazon Mechanical Turk-based manual curation.

   The labeled dataset is divided into two parts, and separately used for classifier training and testing. The majority (usually, 80-90%) of the labeled data is used for training the classifier, and the remaining portion is used for evaluating the performance of the classifier.

   b. Preprocessing data: If we are dealing with tweet text, then cleaning (e.g., removing non-ASCII characters) and preprocessing become essential given the expected variability and informality in tweet text. The preprocessing involves natural language processing (NLP), such as: *removing stop words*, *correcting colloquial words*, *removing repeating characters*, *spelling correction*, *performing lemmatization* etc. We also infer the language of the tweet since we focus primarily on English language tweets in our studies.

   Figure 1.3 presents an example of various NLP tasks that are usually performed on a tweet text. The underlined text at each step shows the replaced words from the previous step. As shown in the figure, first the colloquial words are replaced with English dictionary equivalent words. Next, the repeating characters were removed followed by spell check and punctuation removal. Finally the lemmatization is performed on each word followed by language detection and stop-word removal. We utilized a combination of in-house tools and standard NLP libraries for performing such tasks.

   c. Feature extraction: The next important step is extracting the features from data. Features are the basic properties of a data point that may help the machine-learning algorithms differentiate them. Examples of tweet features include: words in the tweet (also know as bag-of-words), tweet length, word length, number of repeated punctuation, number of alphanumeric lengthening, number of hashtags and URLs used,

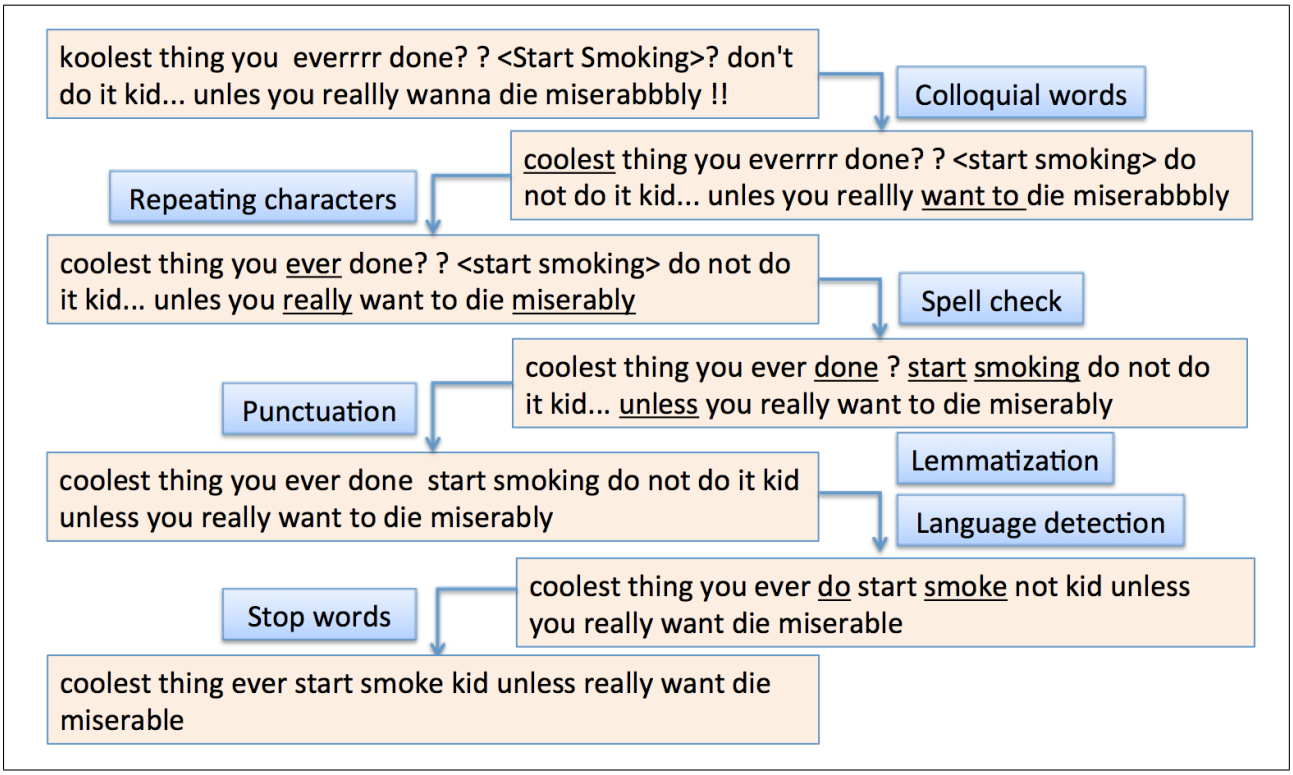


Figure 1.3: An example of tweet Natural Language Processing tasks

language of the tweet. Similarly, we can also gather user-specific features if the application requires it. These features may include: screen name, profile description, friends count, follower count, or total tweet counts.

Please note that the sequence of the NLP tasks is also very important to obtain the correct values for the features. For example, *counting and removing repeating characters* should be performed before the *spell check*, and *language detection* should be performed before the *stop words removal* because we use *stop words* present in a tweet for language detection.

3. **Classifier Testing and Evaluation**

   Once an initial (mostly big) set of features is extracted, test set tweets are used for assessing the performance of a classifier. Most of the classifiers are already implemented and available via standard machine learning libraries. This pipeline component evaluates a set of available classifiers. A classifier is evaluated based on how correctly it identifies the labels for the test dataset. The evaluation of a classifier is usually based on a performance metric that may include: accuracy, precision, recall, and F1 scores. Each classifier can also be checked for overfitting and underfitting at this point. The internal parameters of the classifiers can also be tuned for further performance gain.

   Please also note that only important features should be kept for learning, and the features not helping a lot should be discarded. Importance of the features can simply

be gauged by removing them one-by-one and noting the relative difference in the test results. If removing a feature reduces the performance of a classifier very much then it should be considered, otherwise it should be discarded.

The output of this pipeline component is a selected and configured classifier that will be used to infer the labels of unlabeled Twitter data.

4. **Unlabeled Data Feature Extraction**

   This component takes filtered, unlabeled Twitter data and produces features that can be used by the selected classifier to infer the labels for these tweets. By filtered data we mean a specific type of data (e.g., tweet text or user name) from the Twitter data repository. Note that most of the tasks performed with unlabeled data at this step are in parallel with the tasks performed at Component 2 with the labeled dataset.

   a. Preprocessing data: Same as before, the data is first cleaned to get remove unreadable, non-ASCII characters. Next, the cleaned Twitter data is provided for natural language processing. Again, a set of available NLP tools and in-house tools are used for this purpose. However, all the cleaning and preprocessing of the data should be performed exactly in same manner as it was performed at Component 2 with the labeled dataset.

   b. Feature extraction: Next, the features from the the Twitter data will be extracted. However unlike Component 2, only the selected feature set that was identified as most useful in Component 3 will be extracted. Again, the process of feature extraction should match exactly that performed in Component 2.

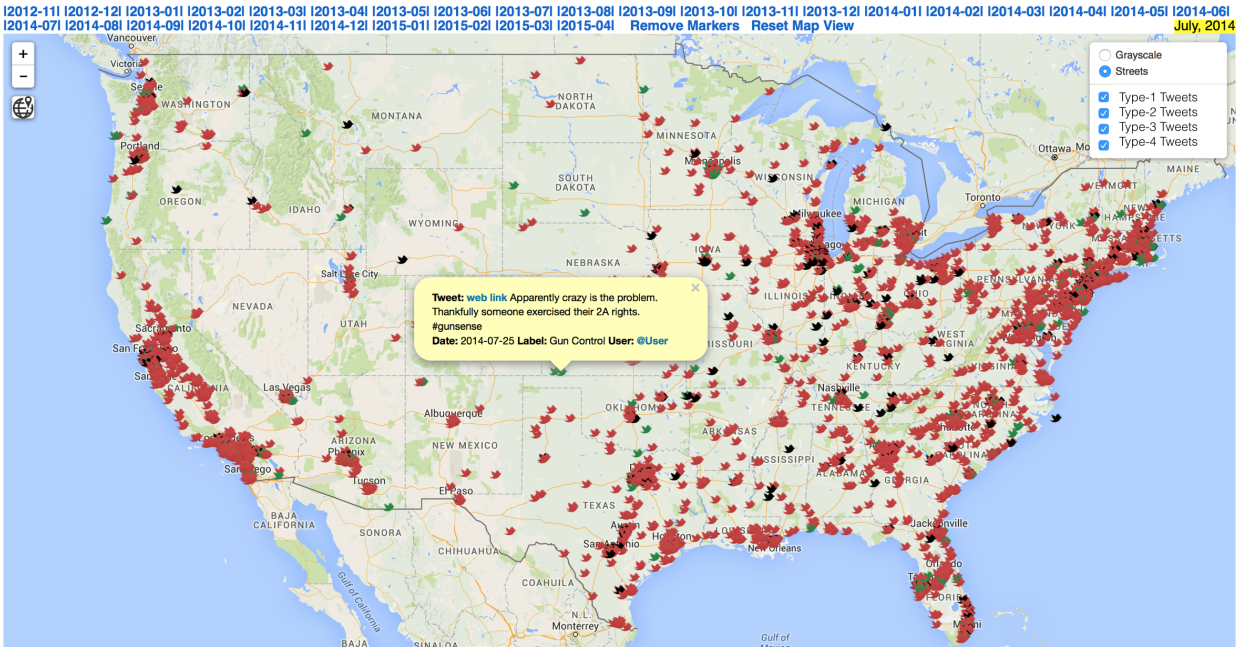


Figure 1.4: A web-based interactive tweet visualization tool used in our Twitter-based surveillance studies

5. **Data Labeling and Visualize Results**

   This is the final component of the pipeline. It combines all the efforts together to generate the labels for the unlabeled data, and help analyze the results. This component takes the selected classifiers, the labeled dataset, and features of the unlabeled dataset as inputs, and produces labels for the unlabeled Twitter Data.

   The generated labels can be further analyzed using simple charts and graphs, or via specifically designed visualization tools. An example of a visualization tool is shown in Figure 1.4. It is a web-based, interactive visualization tool that we developed and used in various Tweet-based studies. This tool renders tweets on maps using the latitude and longitude of the tweets. It uses different markers for different types of tweets to provide better visual cues, and allows to interactively visualize the tweet-level details. This tool also provides a feature to filter and study the tweets of a particular type, and allows to focus on the tweets from a selected geographical region.

1.4 Document Structure

The remaining manuscript is structured in the following way. In next chapter (Chapter 2), we present the motivation for this thesis. This chapter first provides a commentary on various studies from the literature that has explored the role of social networks in smoking initiation and smoking cessation, and then present our idea of building an environment for network-centric intervention for smoking contagion. Chapter 3 covers a wide variety of other related topics ranging from smoking epidemic and its control, to the interplay between social networks and health, to Twitter-based public health surveillance, and to social contagion modeling and behavior spread research.

Next, we will discuss each of the four studies in a separate chapter. Each of these chapters contain their own sections on the study-related background and motivation, methodology, results, contributions, and discussion.

In Chapter 4 and Chapter 5, we will discuss the two Twitter-based smoking surveillance studies. Both the studies are data-heavy, therefore we will first explain the data gathering and basic statistics of the data followed by feature extraction process, evaluation process of the classifiers, and the analysis of the results. Although the focus of these two Twitter-based studies is smoking-related messaging, the primary research questions that we addressed, and hence the type of data gathered and process of analysis were very different.

Next, we discuss the two contagion modeling and simulation studies. In Chapter 6, we introduce the individual-level structured resistance model of smoking addiction. This chapter first talks about the idea of modeling the addictive nature of the smoking behavior utilizing disease diffusion models in epidemiology, followed by the proposed model, simulation study setup and results. The Chapter 7, is organized little differently. We will first discuss the community-based blocking problem and our heuristics followed by the edge-based blocking

study and proposed heuristics. In both of these studies, we evaluate our developed heuristics by performing simulations of contagion propagation on social networks from the literature.

In the last chapter (Chapter 8), we provide a high-level discussion on the achieved goals and the future directions of the presented work. We discuss the potential extensions of our work in each of the three components needed for developing network-centric interventions for controlling the smoking epidemic.

Chapter 2

Motivation and Thesis Direction

2.1 The Role of Social Networks in Smoking Epidemic

The role of social networks in smoking initiation, continuation, cessation, and relapse has been examined by numerous studies. In this section, we briefly discuss some of these studies and their findings related to the effects of networks on smoking epidemic. The motive of this section is to review the efforts made to understand the importance of the social networks in the smoking epidemic.

Most of these studies were purely data-driven, where the correlation between the smoking status of an individual and various social network factors is gauged and analyzed. A primary theme for capturing the effect of networks on smoking contagion has been to record, analyze, and report the smoking behavior of the individuals with respect to the smoking behavior of their immediate social network neighbors. Another approach used in these studies was to perform social network analysis to determine whether, for example, the social status (or popularity) of an individual is associated with his/her smoking behavior. Next, we will discuss some of the studies in more detail.

2.1.1 Effect of Networks on Smoking Initiation and Continuation

The main focus of almost all the efforts in this area is to understand the effect of social factors, including social networks, on adolescent smoking initiation and experimentation. A set of studies from late 1990s found that adolescents whose parents, siblings, or friends

smoke are at increased risk of picking up smoking compared to their peers whose social network members do not smoke [89, 90, 142, 182]. Some evidence also suggests that new adolescents smokers are more likely to obtain their cigarettes from social network members who smoke [155]. Hence, adolescents whose social network members are smokers may have greater access to cigarettes, which could lead to regular smoking. It was also found that the smoking behavior of the members of social network is also one of the important determinants of the age of smoking initiation [180].

A study also examined the effect of both social bonding and ethnicity on adolescent smoking initiation [43]. The authors of this study found that social bonding was highly correlated with early age smoking initiation in African American adolescents as compared to their white counterparts. They also found that less exposure to pro-smoking social influences accounted for the lower rates of smoking at age 18 years for African Americans relative to Whites. These findings suggest that effect of social factors on smoking initiation varies based on ethnicity.

The relative effect of peer influence and peer selection on adolescent smoking and smoking susceptibility was analyzed in a study by Hall and Valente [74]. They reported that peers both impact smoking behavior and influence the development of friendship networks. They also found a direct effect of friendship selection made a year ago on the smoking and smoking susceptibility of the students. Hence, social network formed in the past can slowly but steadily impact risk-taking behavior, such as smoking, in adolescents.

Several studies have examined the effect of position of the individual in the social network on smoking behavior. In one of such study, Alexander et al. analyzed the associations of the peer social status (or popularity), best friend who smokes, and prevalence of smoking in the school on adolescents' current smoking behavior [5]. The study reported that the effect of popularity on smoking behavior is propositional to the smoking prevalence in the school. That means that the popular adolescents in schools with high (low) smoking rates are more (less) likely to be cigarette smokers than their less popular classmates. Also, the risk of current smoking was found significantly associated with the increasing rates of school smoking prevalence. This risk was also associated with having a peer network in which half the members smoked and one or two best friends smoked.

Another similar study found a very interesting result that both popular and isolated youth were likely to smoke cigarettes [187]. Authors of this study assert that since popular adolescents are well connected in school, they may be disproportionately exposed to the pro-smoking peer influences, whereas the isolated youth likely to make smoker friends outside of school. In this study, the popularity association with smoking susceptibility was found strongest for non-white boys, but they did not find enough evidence of interactions between popularity and gender.

Furthermore, a study analyzed the contribution of popularity in addition to other peer variables such as embeddedness in friendships and friendship quality to adolescent smoking involvement [46]. The authors found that the relationship between popularity and smoking involvement was negatively moderated by the smoking behavior of adolescents friends. This finding suggests that being less well known in adolescent networks and having a higher

proportion of friends who smoke relates to more smoking among adolescents.

In a more recent and rigorous analysis, Lakon et al. analyzed the relationship of the past month cigarette smoking with individual and classroom network indicators and a set of peer influence processes [108]. They also attempted to capture more dimensions of the construct of smoking related peer influences arising from friends both inside and outside of adolescents' schools. They utilized social network analysis and modeling to analyze the peer influence related to adolescents smoking uptake through three influence processes: 1) best friends' smoking behavior whose nominations were not restricted to school; 2) classroom best friend network smoking behavior; and 3) own perceived normative beliefs of their friends about drug use. The study reported that both in-degree centrality and being socially integrated in networks relates to more past month cigarette smoking. The study also finds some modest evidence that the peer influence from the best friend networks moderated the relationship between the reciprocity of ties and past month cigarette smoking behavior [108].

2.1.2 Effect of Networks on Smoking Cessation and Relapse

Similar to smoking initiation, the effect of social networks on smoking cessation has been well studied and understood. Many observational studies corroborated a strong importance of social support in the positive outcomes of smoking cessation and other health behaviors. In a study both higher levels of positive support and connectedness were found to be associated with smoking cessation efforts and relapse prevention [34]. Partner facilitation also emerged as the primary predictor of smoking cessation maintenance in a study that focused on newly abstinent females [35].

Another study assessed the role of three kinds of social support factors in smoking cessation and maintenance: support from a partner pertinent to quitting, perceptions of the availability of general support resources, and the presence of smokers in social networks [126]. The authors found that high levels of partner support and of the perceived availability of general support were associated with both cessation and with short-term maintenance of abstinence. The presence of smokers in the social network, on the other hand, was a hindrance to cessation maintenance, and was also the main differentiating factor between people who relapse and who stay long-term abstainers. The study also revealed that having a spouse who smokes or who is critical of attempts at cessation works as a barrier to cessation and abstinence maintenance.

In parallel to the previous results, a study by Chen et al. also found that being married to a nonsmoker and having less proportion of friends who are smokers are associated with cigarette smoking cessation [27]. A similar study in the United Kingdom around the same time by West et al. also found positive association of pressure from partner with attempts to quit smoking [196]. However, the same was not found to be significantly associated with successful attempts to quit smoking in this study.

Christakis and Fowler, later in 2008, again corroborated that the smoking status of friends,

spouse, and neighbors as well as education level affects both smoking initiation and cessation of individuals [30]. This study also revealed that although the overall prevalence of smoking decreased over time, those who remained smokers formed a group with each other and had few social ties with the non-smokers. The authors concluded that social niches emerge in the network with social norms that lead to the decision to quit or to continue smoking in groups. Using social network analysis they have also shown that individuals start and stop smoking in groups. Hence in order to maximize the cessation and abstinence maintenance, the social network effects should be leveraged rightly.

The effect of the social network on relapse has been noted in a recent study. Nguyen and Kohorn performed a survey study of women in postpartum hospital stay who quit smoking while pregnant to understand the reason for the high relapse rate in women smokers after pregnancy [135]. The authors reported that being enmeshed in a social network with prominent smoking norms and the risk of changing their relationships with smoker friends were among the top reasons for the relapse of smoking behavior after the baby was delivered. All the participants of the study emphasized the importance of their relationship with other smokers and the impending risk of losing the friendships if they quit smoking. The authors also suggested the development of a more targeted network-centric interventions to reduce postpartum smoking relapse.

In order to significantly increase the reach and efficacy of smoking cessation programs in the population, various web-based cessation programs have also been designed and evaluated [17, 112, 113, 168]. Given the known association of smoking cessation with social support, majority of such programs focus on social influence and information transfer via an online support group (or social network). QuitNet is one of the most popular and successful such online social networks focused on smoking cessation [32,67]. It has attracted a large amount of users and provided various services to support them in quitting. QuitNet's community feature allows multiple forms of social support. Users can also communicate through private e-mails or via one-to-many messaging in the forums. The success of QuitNet has lead to a more serious discussions to improve and develop such web-based smoking intervention programs [33].

2.2 Network-centric Interventions for Smoking Epidemic

The sustained efforts and initial findings presented in the previous section show that the social networks have been helping in fighting the smoking epidemic. They work through multiple mechanisms, such as social support, social influence, information sharing, and the transmission of social norms. This suggest us that applying these mechanisms systematically might further bring down the smoking prevalence and significantly reduce the number of new adolescent smokers in the population.

Motivated by this, the long-term goal of this thesis is the development and implementation of an environment needed for developing network-centric interventions for controlling the

smoking epidemic. In order to develop such an environment we mainly need three components that are discussed next. The objective of this work is to take first few steps in all these categories, and also to show that such an environment is realizable.

First, we need an operationalized model of smoking that can be simulated. There is no model of smoking epidemic yet that takes into account various smoking-related social and demographic factors. Such a model can be used in simulation studies to measure, tune, and demonstrate the efficacy of network-based interventions for smoking contagion. Epidemiological studies rely on such a formulations of the contagions diseases to understand and communicate the findings. Hence, such a model for smoking contagion will not only make the efforts more visible and studies more understandable, but will also provide the foundation for conducting more regress smoking intervention simulation studies.

Next, we need to be able to quantify the effect of online social networks on the smoking epidemic. All the previous smoking-related studies (discussed in the previous section) considered social exposure and bonding via only face-to-face friendship and social ties. However, the sudden outburst and adoption of online social media platforms, especially by adolescents, have given rise to the new fronts for smoking-related exposure, expressions, and reciprocity. Hence, smoking-related studies performed on social media platforms are pivotal for devising and implementing the network-centric smoking intervention techniques. Example of such studies may include: quantifying the smoking-related exposure on the new social media platforms, capturing the sentiments towards smoking and new products using the users' activities, and performing surveillance studies related to smoking epidemic on these platforms.

Lastly and most importantly, we also need the actual methods that will implement the devised network-centric interventions in the population. An interventions can be applied at different levels (or units) in a social network. For example, an intervention can be implemented at node-, edge- or community-level. Designing methods to operate at different intervention levels, and measuring and comparing their efficacy is extremely important to identify the most suitable network-centric interventions for smoking epidemic for a given situation.

In this thesis, we have taken a first step in all these three categories. We have laid few foundation stones for developing an environment that can help in: modeling social contagion, performing surveillance studies on social media, and devising and evaluating intervention techniques for dealing with smoking contagion.

Chapter 3

Literature Review

3.1 The Smoking Epidemic

The smoking epidemic is a complex phenomenon. Many factors contribute toward its initiation, continuation, and cessation. Some of the major factors that have been studied in depth include: age and gender: age and gender [71, 133, 143]; education and socio-economic status [80, 133, 200]; peer- and familial-influence [30, 63, 65, 82]; chemical dependence [13, 101]; exposure and accessibility [79, 92, 124, 152, 166]; and price and policies [12, 21, 87, 122, 129]. Next, we will discuss in more details about the effect of smoking on adolescents, its addictivity in general, and efforts by government agencies to curb its prevalence.

Among all the age groups, adolescents and young adults have been reported as most vulnerable to the smoking epidemic [199]. Of adults who smoke, 88% report that they started smoking before the age of 18 [184]. Being at a critical transitional phase in their lives with frequently changing social relationships with family and peers, teens become most vulnerable to tobacco use and other risky health behaviors [36, 95]. In order to understand and curb the various factors contributing towards the smoking epidemic, a large body of work is concentrates solely on tobacco usage by adolescents and young adults. Adolescent smoking has shown strong associations with peer- and familial-influence [188, 191], social contexts and position in network structure [107, 108, 163], tobacco outlet density [118], and smoking-related behavior exposure from advertisements or movies [37, 38].

Like any other addictive drug, cigarette smoking behavior becomes compulsive and difficult to cease even after people discover the substantial health benefits of quitting [136]. A study shows that 35 million smokers in the US express a desire for quitting smoking each year, but more than 85 percent of those who try to quit on their own relapse within a week [140].

There has been substantial and consistent efforts by governmental organizations to reduce the smoking prevalence in the population — mainly by policy changes to bring about public health interventions. Effective public health interventions include: raising the price of tobacco products [115,173]; smoke-free policies [50,156,197]; counter-marketing campaigns [15]; advertising restrictions [159]; access to treatment for tobacco use through insurance coverage [51]; and comprehensive approaches to prevent children and adolescents from accessing tobacco products.

3.2 Social Contagion Modeling and Network-based Studies on Behavior Spread

A contagion is any entity that can spread through a population. The basis of social contagion dictates that ideas, rumors, protests, information, and even behaviors can spread through a population in a way that is similar to the spread of infectious diseases spread [66, 72, 185]. In network-based contagion modeling, a population is treated as a network, where nodes represent people or other types of agents, and edges represent pairwise interactions among agents. Hence, nodes influence their distance-1 neighbors through their common edges [42]. Each node can be in one of two states, 0 (respectively, 1) meaning that a node does not (does) possess a contagion. If a node possesses a contagion, we implicitly assume that it is willing to pass it on.

In sociology, the models of contagion propagation for types of contagions discussed above are predominantly *progressive models* [69, 97, 164, 194]. Such models allow a node to transition only from state 0 to state 1; the transition from 1 to 0 is not permitted. The influence of contagion upon coming in contact with neighbors in the opposite state, is mainly largely captured using two basic models. First, the independent cascade model that specifies that each neighbor v of a node u will get one chance to infect u, after which v no longer remains influential on u [97]. Second, the linear threshold model that captures neighborhood influence in such a way that node transitions will occur only if a minimum threshold number or proportion of its neighbors already possess the contagion [70].

There are two main variants of this generic threshold-based model. When all nodes in a network require interaction with only one infected neighbor (or $threshold = 1$) to contract a contagion, the contagion is called *simple* contagion. Whereas, the contagion in called *complex* if at least one node in the network requires interaction with more than one infected node (or $threshold \geq 1$) [25]. This delineation has large impacts on population dynamics and on algorithms for controlling contagion processes [25, 103].

The spread of human behavior has also been studied as a contagion spread problem. Traditionally, observational studies are employed to understand the spread of human behavior through face-to-face social networks. This methodology mainly uses data from longitudinal

studies such as the Framingham Heart Study [49], and the Adolescent Health study [76] etc, and employ statistical techniques to analyze various factors that can affect the behavior spread. A wide variety of behavior spreads have been studied using such a methodology, including smoking [30, 195], alcohol consumption [158], health screening [96], drug use [125], and food consumption [144]. At the same time, there have been some attempts to perform experimental studies to measure the causal effect of social influence online [24, 58, 161], and to use online social networks to study and influence real-world behaviors [20, 44, 111, 174].

Since contagions are mostly undesirable, controlling them is a well-motivated and important problem [119, 186]. The majority of early work (e.g., [4]) used node removal techniques to block contagion transmission in network representations of populations. Deleting nodes from a network removes pathways through which a contagion can travel, thus inhibiting its diffusion. More recently (e.g., [172]), edge removal methods have been studied. There are many situations in which edge removal is a more pragmatic alternative than node removal. For example, in Twitter, person A_1 may stop following person A_2, thereby removing that tie of interaction; it is most often unrealistic to remove A_1 from Twitter. Similarly, political regimes may have the resources to remove or isolate individuals [171], but this approach has costs [165] and may not be politically viable.

3.3 Social Network, Health, and Social Media

Health has been studied in the context of social networks in numerous studies spanning multiple decades. As discussed by Berkman et al., social networks can affect human health through a variety of mechanisms, including: social influence (via norms and social control); social support; person to-person contacts; and access to information [14]. Network-based studies in the recent past (discussed in section 3.2) have also improved our understanding of how social networks influence the collective dynamics of health behavior. We should note that the effect of social network ties can be both protective and deleterious on health. For example, social influences are a primary factor in the adoption of healthy behaviors, such as compliance with diet and nutrition programs, maintenance of exercise routines, and adherence to preventive screening recommendations. On the other hand, social network effects can also lead to risky health behaviors, such as contracting a deadly disease, adoption of tobacco or other addictive substances, and indulgence in eating *junk* food.

Social media has become an integral part of day-to-day life among teens and young adults who are the most active groups of users on the Internet. Trends over the last decade in the United States show that teens have been consistently surpassing other age groups in Internet usage by a large margin [114]. Moreover, around three-quarters of the teens who go online also use social networking sites, Twitter, forums, or other blogging websites. Chou et al. performed an empirical study to identify the sociodemographic and health-related factors associated with social media users in the United States [28]. They found that among the population with Internet access, social media is popular independent of background, ethnicity, education, or acess to health care. Because it has such a high prevalence among

the population, social media is an attractive venue for research related to public health.

Social media is also transforming the public health landscape by providing valuable resources for health communication, online support groups, coaching for weight loss and smoking cessation, and health-related surveillance. Health information dissemination has now become more interactive and electronic, allowing both organizations and individuals to create, share and evaluate health-related information quickly [41, 162]. Social media also helps in creating well-connected health-consumer-centric communications networks, where users can share information, provide suggestions, and consult and post online rankings and reviews of health providers, hospitals, and drugs [8, 59, 170]. However, user-generated health-related information, reviews, and comments can sometimes be biased or influence-driven sometimes. Therefore, the consumers of such information should be wary and authorities should take preventive measures to limit the spread of "popular" but erroneous items online [151].

Support groups and peer-counseling have also extended into the virtual world. Coaching, abstention and emphathetic interaction between members are now mainly delivered using web-based tools. For example, web-based smoking cessation programs by *QuitNet*, *Free* and others provide status updates and peer-to-peer messaging to offer newly abstaining smokers support from members with years of abstention experience [31, 154]. Similarly, the health care industry is also slowly embracing the social media approaches [29]. For example, a primary care practice unit called *Hello Health* uses video chat, Twitter, and other Web 2.0 tools to check their patients and to communicate with them [77].

3.4 Twitter-based Public Health Surveillance

Social media platforms, especially Twitter, are rapidly becoming a key source for public health surveillance. The accessibility of vast amounts of freely available user-generated data that can be automatically collected and analyzed has made Twitter a significant object of study for health-related behaviors. Twitter has been used for various health-related surveillance studies, including: real-time monitoring of infectious disease [110]; understanding health behavior sentiments [160]; and analyzing sentiments toward emerging tobacco products [130]. A significant proportion of this work also looks at predicting health and emotional issues from Tweets and other behavior indicators such as depression [40] and post-partum depression [39].

Signals from Twitter has also been used for improving influenza forecasting. Paul et al. have demonstrated that influenza surveillance signals from Twitter can significantly improve forecasting [145]. They also found that Twitter data provides better forecasts as compared to Google Flu Trends data, hence validating the use of social media data sources for influence surveillance and forecasting. In a separate study, the same group presented an approach for distinguishing actual flu reporting tweets from concerned awareness-related tweets in order to improve the quality of influenza surveillance using Twitter [109]. Twitter data has also been used for examining flu trends at city level. Nagar et al., for example, used geotagged

tweets from New York City and validated the temporal predictability of daily tweets for emergency department visits for influenza-like illness [131].

A number of Twitter-based smoking-related surveillance studies have been reported in the literature. For example, Myslin et al. analyzed the content and sentiments of around 7,000 tweets using machine learning techniques [130]. They found a high prevalence of positive sentiments toward e-cig and other emerging tobacco products in the tweets, and also that in general sentiments about smoking were largely positive. Some of the studies have mainly focused on e-cig marketing on Twitter. Huang et al., for example, examined all e-cig-related tweets over two months and found that 90% of the tweets they gathered were commercial [85]. In addition, most of these tweets claimed health and smoking-cessation benefits.

Another set of studies analyzed the content of e-cig retail websites and performed surveys to measure online exposure of smoking-related messaging and advertisements. Grana et al., for example, gathered and examined the content of e-cig retail websites [68]. Similar to other studies, they found health-related benefits along with claims of enhancing social status by e-cig usage. Emery et al., on a different front, used an online survey of adults to analyze e-cig awareness, use, and information sharing and searching [45]. They found 86% of the subjects were aware of e-cigs, and that the tobacco users were twice as likely as non-users to have seen or heard information about e-cigs. In their study, Twitter was found to be a medium that is used by regular e-cig users 17% of the time for sharing and 9% of the time for searching e-cig related information.

A few studies have also assessed Twitter social networks revolving around smoking. For instance, Prochaska et al focused on smoking cessation messages [149]. They found the existence of several social networks revolving around smoking cessation; however, several of the accounts were non-active and the content of the messages was not consistent with clinical guidelines.

Chapter 4

Exposure of a vulnerable population to smoking-related tweets

4.1 Introduction

Tobacco use is the leading preventable cause of death in the United States, as per the CDC. The economic costs of tobacco use are estimated to run in the hundreds of billions of dollars every year, due to lost productivity and increased health care expenditures. Adolescence is the highest risk period for smoking initiation. Of adults who smoke, 88% report that they started smoking before age 18 [184]. According to the FDA's Center for Tobacco Products[1], each day in the US, 3300 kids under the age of 18 years smoke their first cigarette, and 700 kids become daily smokers.

With the recent legalization of medical and recreational marijuana use in various US states, there is also a need for understanding usage patterns and factors that influence initiation of marijuana use.

There are many factors associated with youth smoking initiation. Studies have shown that the smoking status of social network members and pro-tobacco marketing are both important determinants of the age of smoking initiation [180]. In the age of Web 2.0, the online and offline social worlds of adolescents are merging at the highest rate - about 75% of teens on the Internet use social media (such as social networking sites, micro-blogging sites, forums etc) to connect and communicate with their friends [114]. Since smoking is a social contagion that

[1] http://www.fda.gov/TobaccoProducts/ProtectingKidsfromTobacco

spreads through social fabric, understanding the effect of online social world on its diffusion is an important problem.

Tobacco marketing is heavily restricted in the U.S., to which the tobacco companies have responded by turning to novel forms of marketing to circumvent these restrictions [6, 78]. Restrictions are comparatively lax on Internet-based marketing so far, partly due to the relative anonymity afforded by the medium. This offers tobacco marketers a powerful means of circumventing current marketing restrictions intended to protect public health from tobacco-related disease. Several studies have shown that the Internet is being utilized as a "below the line" medium for exposing adolescents and young adults to tobacco promotions [57, 60, 147, 192]. However, it may also be possible to use anti-tobacco messaging to reduce receptivity to tobacco use [181]. In addition, there are strict restrictions on marijuana marketing in Colorado, including social media-based marketing. However, exposure to marijuana-related messages, even if they are not direct marketing messages, may have the effect of normalizing marijuana use.

Similarly, the presence of large-volumes of smoking-related messaging has the effect of countering the denormalizing strategy of tobacco control, even if the messages are not promoting specific brands of tobacco. The promotion of marijuana use has a normalizing effect for tobacco use as well. Nearly 60% of current and former tobacco users report marijuana use as well, and 90% of people who have ever used cannabis report having used tobacco as well [2].

Twitter is an interesting platform for such a study because teen Twitter use has grown three-fold in the past three years and the tweets of 76% of teen users are publicly available [114, 123]. Small-scale studies of smoking-related messages on Twitter have shown there is a high prevalence of positive sentiment [130] and, more generally, that Twitter can be used for tobacco surveillance in ways similar to how it is being used for infectious disease surveillance [110, 160]. However, there hasn't been a large-scale study of the volumes of pro- and anti-tobacco and pro- and anti-marijuana messaging on Twitter.

In the present work, we investigate the exposure of teens to smoking-related messaging on Twitter. We collected tweets corresponding to smoking-related keywords over a period of ten months to quantify the extent of smoking-related messaging. We train classifiers to discard the irrelevant tweets, and then categorize the remaining tweets into four classes: pro- and anti-tobacco, and pro- and anti-marijuana. We treat neutral tweets as pro-smoking because, as mentioned above, they have a normalizing effect that increases the likelihood of smoking initiation among non-smokers. We also train classifiers to identify the age group of Twitter users who are being exposed to these Tweets (and are also generating them). We focus on whether these users are over or under 18 years of age. Finally, we use a model of Twitter user behavior to estimate how many of these tweets are likely to be actually read [84].

Summary of results

We estimate that, in our data set, 69.2% of the tobacco-related tweets are pro-tobacco, and 94.3% of the marijuana-related tweets are pro-marijuana. We focus on 10 heavily tweeting "key" users for a deeper analysis. We extract a random subset of their followers ($N = 736$) and a random subset of the followers of those followers, termed second-degree followers ($N = 922$). We find that 36% of the first-degree followers and 33.5% of the second-degree followers are predicted to be under 18 years of age. We estimate that the first-degree followers are reading a median of 2.22 pro-tobacco and 3.36 pro-marijuana tweets/day from the key users, compared with only 0.39 anti-tobacco and 0.0 anti-marijuana tweets/day.

The rest of this chapter is organized as follows. We begin by describing our data set of smoking-related tweets. After that we present results from training classifiers to categorize these tweets into the four categories mentioned above. Then we describe how we generated a data set to infer the age of Twitter users from their tweet content, and present results of training classifiers. After that we describe the use of a stochastic model of Twitter user behavior to infer the rate at which a sample of users from our data set are reading smoking-related tweets.

4.2 Materials and Methods

4.2.1 Tweet classification

Description of data set

We gathered 1% of publicly available tweets matching the following keywords: *smoke, smoking, cig, tobacco*, and *marlboro* through the Twitter API. Note that "cig" would match *cigarette, cigar, ciggy* etc. as well. Tweets have been gathered continuously from March 1, 2013 to Jan 30, 2014. This gives us a collection of 106,127,613 tweets. Fig. 4.1 shows the overall counts of tweets, retweets, and unique IDs on each day.

On average, we received 315,856 tweets/day from 266,966.83 unique user IDs, of which 121,676.49 tweets were retweets. The maximum number of tweets on a given day were received on April 20, 2013 (seen as the highest spike in Fig. 4.1), which is a counterculture holiday in North America to celebrate cannabis consumption. The other main spike in Fig. 4.1 is seen on March 13, 2013, which was "no smoking day" in the UK.

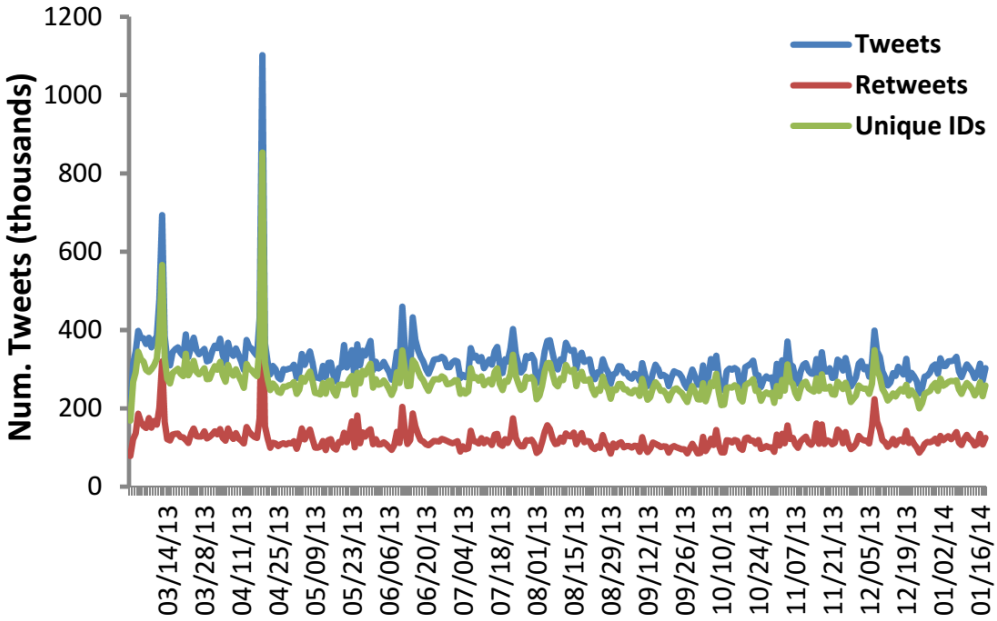


Figure 4.1: Overall counts of tweets, retweets, and unique IDs over time.

Categorizing tweets

We divide the tweet classification problem into a hierarchy of classification tasks as illustrated in Fig. 4.2. First we separate out irrelevant tweets from the relevant tweets (Task 1), then we identify tobacco-related vs marijuana-related tweets found in the relevant class (Task 2), and finally label pro- and anti- tweets in both tobacco (Task 3) and marijuana (Task 4) tweets sub-classes.

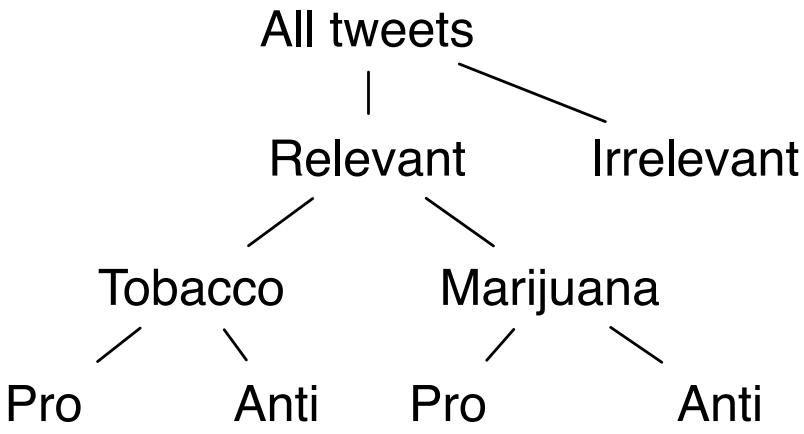


Figure 4.2: The hierarchy of classification tasks.

We evaluated four different learning algorithms for each of the four tasks using 10-fold cross-

Table 4.1: Sample tweets per label. PTT: *pro-tobacco tweets*, ATT: *anti-tobacco tweets*, PMT: *pro-marijuana tweets*, AMT: *anti-marijuana tweets*, and IRT: *irrelevant tweets*

Label (Total)	**Example Tweets**
PTT (1227)	- *I smoke to clear my mind..* - *More Doctors Smoke Camels Than Any Other Cigarette - http://t.co/PCiUSX5U3l*
ATT (957)	- RT @user_1: Don't smoke cigarettes; there are cooler ways to die. - I do wish I had never started smoking kinda
PMT (1173)	- *RT @user_2: We smoke dope all day all night* - *Look like Barbie, smoke like Marley...*
AMT (154)	- *7 years in jail for smoking weed here ???? Shit ain't fun* - *I'm so glad I don't smoke weed lol they tripping*
IRT (489)	- *I smell fire smoke again ..*. - *These firefighters are smoking hot lol*

validation on 4000 hand-annotated random tweets. The tweets were first hand-labeled into five classes based on only text i.e., we did not consider hyperlinks, emoticons or sarcasm while labeling the tweets. A sample of tweets and number of tweets per label class are shown in Table 4.1. Standard implementations of the learning algorithms, from the NLTK and Scikit-learn Python libraries, were employed [146], and randomly selected training(90%) and testing (10%) data sets were used per fold for the evaluation.

Each classification task uses *bag-of-words* as features that were gathered after pre-processing the tweets using standard NLP tools. The pre-processing involved (in sequence): removing hyperlinks, hashtags, and mentions, correcting colloquial words, removing repeating characters, spelling correction, removal of punctuation, performing lemmatization, and finally removing stop-words. We also accounted for overfitting and imbalanced data classes by tuning the classifier parameters.

A comparison between the accuracy of the four classifiers for the tweet classification tasks is shown in Table 4.2.a. We see that SVMs outperformed both Naïve Bayes (NB) and Maximum Entropy classifiers (MaxEnt), and that SVMs with linear kernel performing slightly better than the rbf kernel on all the four classification tasks. The aggregated accuracy for the tobacco-related tweet classification was 63.4% and that for marijuana was 72.3%. The average precision and recall values for each task using the selected *SVM-lin* classifier is presented in Table 4.2.b. All values but recall for *irrelevant* and *anti-marijuana* tweets are decent. Once reason for the low recall in these classes could be the limited number of hand-labeled tweets in these categories.

Using the trained SVM-linear classifiers, we found that around 68% of the total 106 Million collected tweets were tobacco related, and that more than 69% of this fraction were pro-

Table 4.2: a) Classifiers' accuracy comparison for hierarchical tweet classification tasks 1 through 4 using bag-of-words features. b) Average precision, recall, and F1-score using the selected *SVM-lin* classifiers for the two classes Cl-1 and Cl-2, in each of 4 tasks. Task 1 is *relevant vs. irrelevant*, Task 2 is *tobacco vs. marijuana*, Task 3 is *pro- vs. anti-tobacco*, and Task 4 is *pro- vs. anti-marijuana*.

		Classifier			
		NB	ME	SVM-lin	SVM-rbf
1	Min	0.870	0.870	0.880	0.873
	Max	0.913	0.915	0.930	0.925
	Avg	0.888	0.891	**0.900**	0.898
2	Min	0.817	0.831	0.869	0.871
	Max	0.863	0.874	0.920	0.923
	Avg	0.843	0.858	**0.899**	0.893
3	Min	0.734	0.757	0.752	0.729
	Max	0.803	0.803	0.807	0.812
	Avg	0.767	0.778	**0.783**	0.771
4	Min	0.856	0.795	0.841	0.841
	Max	0.902	0.894	0.924	0.932
	Avg	0.876	0.839	**0.894**	0.881

(a)

		SVM-lin		
		Precision	Recall	F1
1	Cl-1	0.901	0.995	0.945
	Cl-2	0.859	0.221	0.348
2	Cl-1	0.883	0.967	0.923
	Cl-2	0.935	0.790	0.855
3	Cl-1	0.805	0.812	0.808
	Cl-2	0.756	0.745	0.750
4	Cl-1	0.912	0.974	0.942
	Cl-2	0.590	0.286	0.377

(b)

tobacco tweets. Marijuana-related tweets were around 94% pro-marijuana. The summary of the results is shown in Table 4.3, and the per-month distribution of the relevant tweets in the four classes is shown in Figure 4.3.

4.2.2 Identifying the exposed population

To identify the Twitter users that get exposed to such smoking related tweets, we explored the follower network of users who heavily and regularly sent the smoking related tweets. We used Twitter API and extracted a large subset of the followers (upto two degree-of-separation) and their timeline tweets. Due to Twitter's stringent data limits, we employed

Table 4.3: Summary of classification results for smoking-related tweets. Counts and fractions of tweet in each class. RvI: Relevant vs. Irrelevant, TvM: Tobacco vs. Marijuana, PTvAT: Pro- vs. Anti-Tobacco, PMvAM: Pro- vs. Anti-Marijuana.

Categories	**Class 1 count(%)**	**Class 2 count(%)**
RvI	99,212,819 (93.5)	6,914,507 (6.5)
TvM	67,306,393 (67.8)	31,906,426 (32.2)
PTvAT	46,591,523 (69.2)	20,714,870 (30.8)
PMvAM	30,087,140 (94.3)	1,819,286 (5.7)

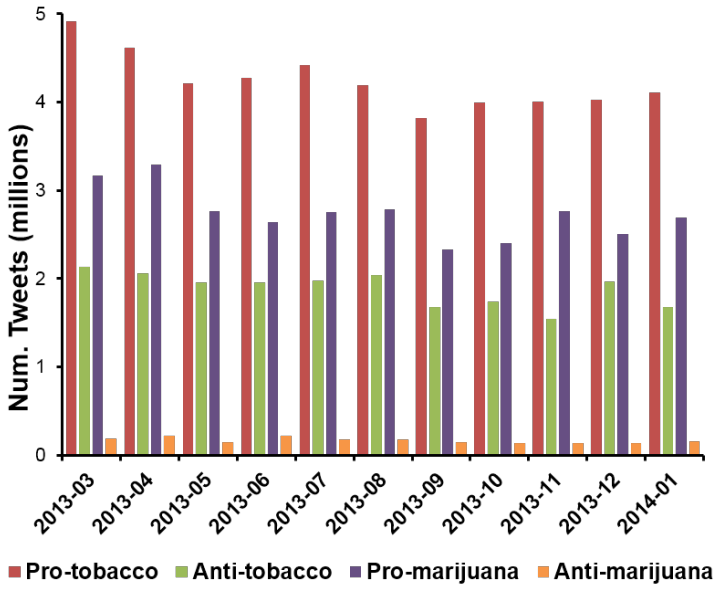


Figure 4.3: Number of smoking-related tweets per month in the four classes.

the following technique to extract the follower network and timeline tweets:-

(i) First, we identified 100 *key users* who consistently tweet heavily about smoking using the gathered tweets from March 01 to May 31, 2013 (*window 1*).

(ii) Then we started extracting the follower network of these users until two degree-of-separation i.e, their followers and followers of their followers. We could extract follower network of only 84 *key users* from *window 1* because some of the accounts were suspended, protected, or had no followers. Also, we extracted at most 25,000 followers per user for both the hops due to rate limits.

(iii) Towards the end of the tweet collection process, we again identified 100 top tweeters using tweets from Sept 01 to Nov 30, 2013 (*window 2*), and found 15 *key users* that are common between *window 1* and *window 2*. We assume that these users were consistently tweeting throughout the tweet collection period and therefore their followers are the one who got exposed to smoking-related tweets the most.

(iv) We then randomly selected 1000 heavily tweeting (i.e.,users with at least 3200 timeline tweets) direct followers or also known as one-degree-of-separation (1-DoS) followers of the *key users* out of total 43,788 extracted followers. However in manual verification of the *key users* later, we found that five profiles were either bots or heavy non-english tweeters. Removing the followers of these users left us with 828 1-DoS followers out of total 36,396. Similarly, we randomly selected 1000 heavily tweeting two-degree-of-separation (or 2-DoS) followers out of total 1,849,181 unique followers of 828 1-DoS followers.

(v) Finally, we downloaded upto 3200 most recent tweets from the Twitter timeline of the selected 1-DoS and 2-DoS followers of the *key users* for inferring their age as discussed in

Table 4.4: Tweet categories for the ten selected key users. Note that rows may not sum to 100 because percentage of irrelevant tweets are not shown. The maximum in each row is bolded.

	Tobacco		**Marijuana**	
	Pro %	Anti %	Pro %	Anti %
User 1	36.84	6.32	**55.79**	0.0
User 2	**54.51**	7.7	5.38	0.13
User 3	**61.14**	15.39	3.83	0.0
User 4	2.39	0.34	**93.86**	1.71
User 5	27.1	4.8	**67.63**	0.0
User 6	**59.2**	18.68	2.98	0.08
User 7	12.57	**83.86**	2.75	0.03
User 8	**57.08**	31.32	1.11	0.0
User 9	29.05	**60.4**	3.71	0.39
User 10	**45.62**	41.63	3.42	0.18

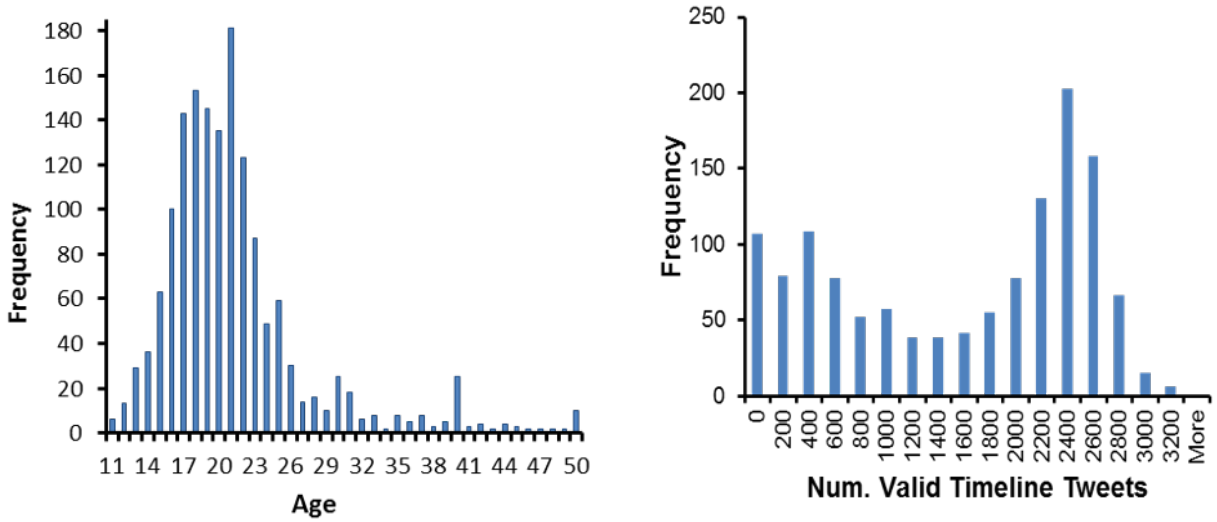
the next section.

We also used the trained classifiers from the previous section to classify the tweeting behavior of the ten selected key users. For each of them, we give the fraction of their tweets that are pro- and anti-tobacco and marijuana in Table 4.4. Note that, in keeping with recommendations for ethical social media research [157], we have suppressed the user names. We find that, of the 10 key users, five are pre-dominantly pro-tobacco, two are pre-dominantly anti-tobacco, and three are pre-dominantly anti-marijuana.

4.2.3 Identifying the vulnerable population

The aim of the present work is to investigate the exposure of adolescents under the age 18 to smoking-related messaging over Twitter. The Twitter user profile is quite limited, only containing fields for name, location, website, and description. Some users may choose to reveal their age in the description, but in general we have to resort to inference methods to identify age. Various approaches have been tried for this problem [3, 134, 150].

We employed an approach that make use of only tweets to infer the age of the users. Our technique focuses only on English tweets by a user, and infer user age into *under18* and *over18* age-group classes. The *under18* class includes users from ages 11 to 17 inclusive, and the *over18* class comprises of users from ages 18 to 50 inclusive. The various steps involves in this process are presented next.



(a) Happy Birthday to Me Tweets

(b) Valid Tweets per Users

Figure 4.4: Frequency distributions of: (a) *HBTM* tweets for ages 11 to 50, and (b) valid tweets across all users.

Collecting tweet corpus for age classification

Similar to [3], we searched for tweets matching two templates using Twitter's search API: "happy Nth birthday to me" and "happy birthday to me #N", where N was replaced successively with numbers 11 through 50. This gave us 368 unique English speaking users in *under18* class, and 1037 users in *over18* class. We collectively call these users "Happy Birthday to Me" (or *HBTM*) users. A distribution of *HBTM* users is shown in Figure 4.4.a. The histogram shows that users from ages 16 to 23 wished themselves birthday on Twitter more often compared to other age groups, and that the maximum number of users were returned for age 21. We also observe local peaks at the ages 30, 40 and 50.

For each of these *HBTM* users, we downloaded up to 3200 most recent tweets from their Twitter timeline. Each user is now represented by a collection of its timeline tweets, and can further be used as a training instance for supervised learning to infer the age of the other Twitter users as discussed next.

Age related feature extraction

For the classification, we first filter the tweets (and consequently users) that are suitable for the learning process and then gather the features from these tweets to infer the age-group labels. A tweet is *valid* for learning only if it is in English, and contains one *stop* word and two *regular* words. *Stop* words are needed to identify language of a tweet, and *regular* words helps in feature extraction. Since we focus on English tweets, we remove all non-english language tweets.

Table 4.5: Numeric features captured from the timeline tweets. For the first seven features, we also took the mean across non-zero instances.

Numeric Feat	**Description**
urls	avg num of times urls used
hashtags	avg num of times hashtags used
mentions	avg num of times mentions used
contractions	avg num of times word contractions used
repPuncts	avg num of times repeated punctuation used
anLengths	avg num of times alpha-numeric lengthing used
spellMistakes	avg num of time spelling mistakes were made
avgWordLeng	avg word length
words	avg num of words used (i.e., tweet length)
tweetsPerDay	avg num of tweets per day

A frequencty distribution of valid timeline tweets for all the extracted *HBTM* users is shown in Figure 4.4.b. Although the frequency of valid timeline tweets is maximum for 2200-2399 bucket, we choose to consider only 150 valid tweets out of 400 most recent timeline tweets per users for feature extraction. It is because age is a dynamic characteristic of an individual that changes over time, and very old tweets might not be a good representatives of the current age of a user. We also trained and compared the classifiers (discussed in next subsection) using 100 and 200 valid tweets but the performance was best with 150 valid tweets. We did not consider the users with less than 150 valid tweets. Discarding such users left us with 167 *under18* and 621 *over18* users for the training purposes.

We processed the valid tweets to count and remove: urls, hashtags, mentions, alphanumeric lengthnings, word contractions, repeated punctuation, and stop words. The spelling mistakes were also counted but were not corrected in order to preserve the word selection behavior of the users. By doing this, we gathered 17 numeric features per user and 32,421 bag-of-words in total to construct the training and testing data for the supervised learning. A list of numeric features and their brief description is shown in Table 4.5. These features were obtained by first adding them across all the 150 valid tweets of a user and then taking mean. For the first seven features, we also computed the mean only over the tweets in which that particular feature was used at least once (later shown with prefix *nz_*). The mean over all tweets captures the tendency of a user to use a feature, where as mean over only non-zero instances measures the frequency of a feature when used.

Age classifier evaluation

We evaluated SVM and Random Forest classifiers using 10-fold cross-validation on the bag-of-words (BoW) and numeric features separately. We also employed a classifier stacking technique similar to [150] in order to combine the two types of features. A standard implementation of the algorithms from the Scikit-learn Python libraries were used [146]. Since

the data in the two age classes is unbalanced, we again performed the evaluation with a wide range of classifier parameters. We selected the parameter set that gave us a balanced classifier i.e., a classifier with reasonably good accuracy, less overfitting and high precision and recall for the smaller data class.

The evaluation results for age-classification are shown in Table 4.6. We have only compared the average results for a balanced classifier for each classifier-feature set combination. Results show that SVM classifiers with linear kernel again outperformed random forest classifiers using both BoW and numeric feature set. However, stacking the two selected balanced classifiers did not further improve the results. The exact same values for the the classifier with BoW and stacked features show us that numeric features did not contribute enough for the age inference.

Lists of top five most informative features for age classification identified by the selected classifiers using bag-of-words(BoW) and numeric features are shown in Table 4.7. The most informative words show the clear difference in the word selection by the two age-groups. *Under18* users, for example, talk more about school and use words such as cute, hate etc more often than then their counter part. Similarly, the *over18* age-group users talk more about work, class, drinking etc more often then the *under18*.

Also, the most important numeric features of the two age-groups in the Table 4.7 suggest that users in *under18* class use similar number of urls, word contraction, and mentions on average per tweet as compared to *over18* class users. On the other hand for the *over18* class, number of hastags used, word length, and number of words used on average per tweet

Table 4.6: A summary of accuracy of the balanced results for two classifiers using bag-of-words (BoW), numeric and stacked features. *Acc. Train* and *Acc. Test* are the average accuracy on training and testing data respectively. *Prec. Under18* and *Rec. Under18* denote the average precision and recall values for the class *Under18*. We see that SVM with a linear kernel perform the best for both type of feature. However, stacking the continuous features did not help in further improving the results.

		Classifier	
		SVM-lin	**randForest**
BoW	Acc. Train	0.951	0.965
	Acc. Test	**0.848**	0.780
	Prec. Under18	0.627	0.404
	Rec. Under18	0.631	0.117
Numeric	Acc. Train	0.757	0.965
	Acc. Test	**0.728**	0.711
	Prec. Under18	0.303	0.286
	Rec. Under18	0.230	0.248
Stacked	Acc. Train	0.951	
	Acc. Test	**0.848**	
	Prec. Under18	0.627	
	Rec. Under18	0.631	

Table 4.7: First five most informative bag-of-word (BoW) and numeric features used by trained *SVM-lin* classifier to infer age of twitter users in *under 18* and *over 18* age-groups. The prefix *nz_* in numeric features represents the feature where the mean was taken only over the tweets in which that feature was used at least once.

	under 18	***over 18***
BoW		
	school	work
	cute	class
	hate	drink
	actually	keep
	justin	senior
Numeric		
	nz_urls	hashtags
	contractions	avgWordLength
	nz_mentions	words
	nz_repPuncts	nz_anLengths
	repPuncts	anLengths

Table 4.8: Summary of age classification results for exposed population. 1-DoS and 2-DoS are the selected one- and two- degree of separation followers of the *key users*.

Twitter Users	***under 18*** **count(%)**	***over 18*** **count(%)**	**Total**
1-DoS Follower	265 (36.0)	471 (64.0)	736
2-DoS Follower	309 (33.5)	613 (66.5)	922

proved to be better identifiers. Also more interestingly, *under18* users on average tend to use repeated punctuation in their tweets in similar fashion, whereas pattern of using alphanumeric lengthening in tweets on average is similar for *over18* users.

Infer exposed followers' age

The selected classifier was used to infer the age of the one- and two-degree of separation followers of the key users. We ran the classifier on the timeline tweets of the exposed population of 1-DoS and 2-DoS that we discussed in the Section 4.2.2. The results of the age-classification are presented in Table 4.8. We found that a substantial fraction of both 1-DoS (36%) and 2-DoS (33.5%)followers of the *key users* are under age 18. We could infer the age of only 736 out of 828 1-DoS and 922 out of 1000 2-DoS followers because the remaining users did not have 150 valid tweets in their 400 most recent timeline tweets.

4.2.4 Modeling tweet reading behavior

To model user behavior, we follow the work of Hogg et al. [84]. Their stochastic model assumes that a Twitter user is in one of four states: away from Twitter ($Away$), visiting Twitter ($Visit$), reading tweets ($Read$), or responding to tweets by retweeting, replying, etc. ($Respond$).

We wish to estimate the rate at which a user, u, sees tweets from a chosen "key user". We refer to this as the *exposure rate*.

We will proceed as follows. We will estimate the probability that the user u has received L more tweets after the key user's tweet when u actually checks his Twitter feed. We will use an estimate of the probability that u reads past the L^{th} tweet in his feed to obtain our required exposure rate. This model is illustrated in Figure 4.5.

The rate at which a user, u, receives tweets from his friends is, on average,

$$R(u) = N_f(u)R'_f(u),$$

where $N_f(u)$ is the number of friends of u, and $R'_f(u)$ is the average tweet rate of u's friends. Calculating $R'_f(u)$ would require obtaining the tweets of all the friends of each user u, which is very time-consuming due to Twitter's rate limitations. Therefore, following [84], we estimate $R(u)$ as,

$$R(u) = N_f(u)\overline{R'},$$

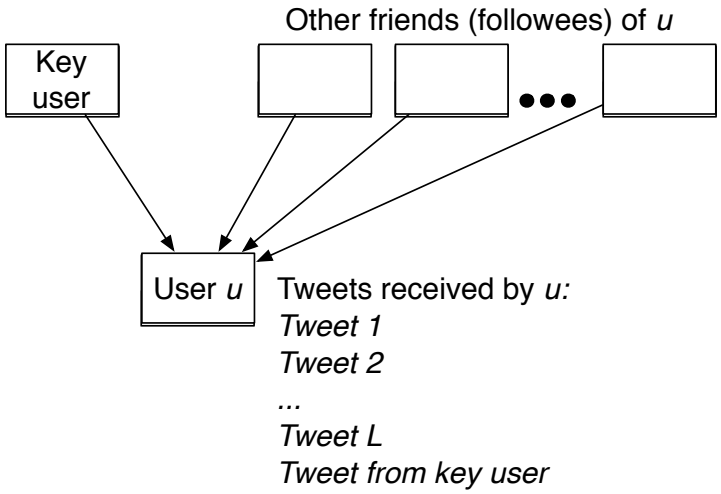


Figure 4.5: Tweet reading model. To estimate the rate at which user u reads tweets from the key user, we have to account for the tweets arriving from other friends (followees) of u. If there are L tweets that have arrived after the key user's tweet, before u checks his Twitter feed, we need to estimate the probability that u will read past L tweets to see the key user's tweet.

where $\overline{R'}$ is a measure of the average tweet rate of all users. The tweet rate distribution is heavy-tailed, as shown in Figure 4.6.a, so we use the median and not the mean. In the dataset considered, we obtain $\overline{R'} = 12.0877$ tweets/day.

Next we need to estimate the number of times user u checks his Twitter feed each day, i.e., his visit rate. Conservatively, we assume that this is just $R'(u)$, the rate at which u tweets. In practice, it is possible that a user checks his Twitter feed sometimes without sending out a tweet, and also that he sometimes sends out a tweet without reading any tweets in his timeline. Since we don't have any information available about the rates of these behaviors, we use $R'(u)$ for u's visit rate.

Suppose that when user u checks his Twitter feed, he has received a tweet from a key user, followed by L more tweets. The probability that the key user's tweet is read by u can be written as [84],

$$P_{read}(u) = \sum_{L} P_{rec}(L|u) P_{view}(L), \tag{4.1}$$

where $P_{rec}(L|u)$ is the probability that user u receives L more tweets after the tweet from the key user, and $P_{view}(L)$ is the probability that u goes past L items in his feed.

Hogg et al. [84] approximate $P_{rec}(L|u)$ as a competition between two Poisson processes with constant rates: one process corresponding to the random arrivals of tweets into user u's timeline, and the other corresponding to u's random visits. The first Poisson process has rate $R(u)$. Since we are assuming that the rate at which u visits Twitter is the same as his rate of tweeting, the second Poisson process has rate $R'(u)$. User u will see L tweets in his Twitter feed on his next visit if the first Poisson process has L "arrivals" before the second Poisson process has one arrival.

This is a straightforward question. It is solved by merging the two processes into one with rate $R(u) + R'(u)$. The probability that the next arrival is from the first Poisson process is given by

$$p_1 = \frac{R(u)}{R(u) + R'(u)}. \tag{4.2}$$

Now the probability of seeing L tweets before the next visit is the same as seeing L "successes" before the first failure, i.e., a geometric distribution. Therefore,

$$P_{rec}(L|u) = p_1^L (1 - p_1).$$

The mean of this distribution is $p_1/(1 - p_1)$, which is equivalent to $R(u)/R'(u)$. This makes intuitive sense, because we expect that, on average, the rate at which tweets "accumulate" in the Twitter feed of user u is the ratio of the rate at which the tweets arrive and the rate at which u checks his Twitter feed.

To estimate the second term in equation 4.1, $P_{view}(L)$, Hogg et al. reference the "law of surfing" [18], which says that the probability a user views m items in a list before stopping

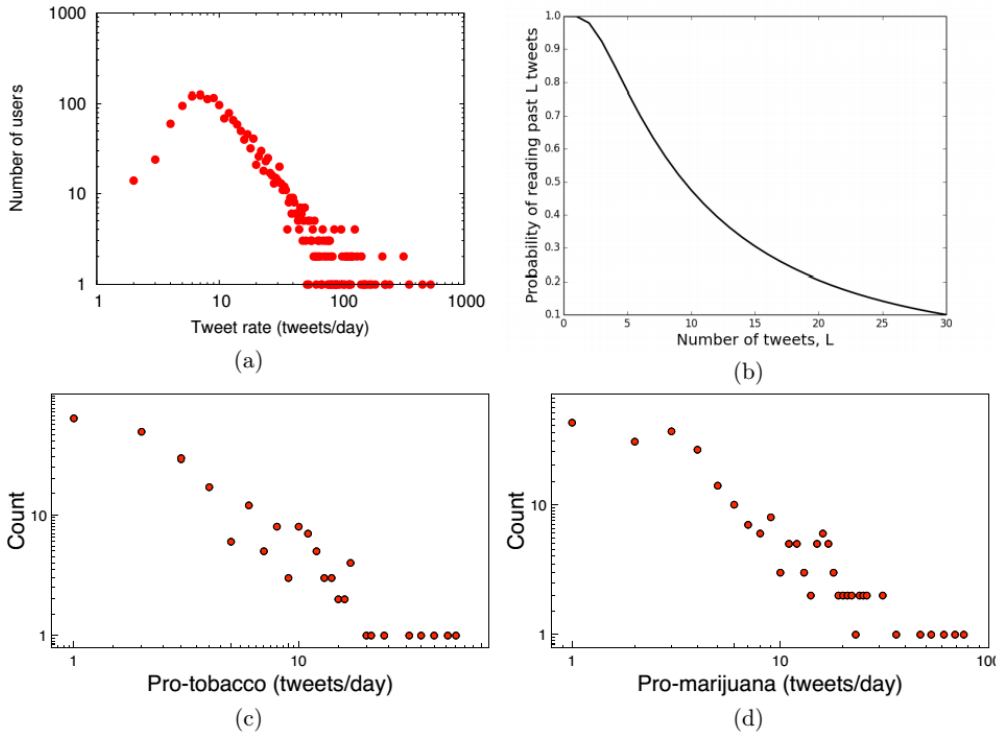


Figure 4.6: (a) The distribution of tweet rates per user. (b) The probability distribution that a user will read beyond L tweets in his Twitter feed [18, 83, 84]. (c,d) Estimated tweet exposure distributions in tweets/day for *under 18* first-degree followers of the key users to pro-tobacco and pro-marijuana tweets. Exposure rates are binned by rounding to the nearest integer.

has an inverse Gaussian distribution,

$$P(m) = e^{-\frac{\lambda(m-\mu)^2}{2m\mu^2}} \sqrt{\frac{\lambda}{2\pi m^3}}.$$

We need the fraction of users who view at least $L + 1$ tweets, for which we have to take the complementary cumulative distribution of $P(m)$, which is also known as the survival function. The parameters are chosen to be $\mu = \lambda = 14$, based on [83, 84]. The resulting probability distribution is shown in Figure 4.6.b.

We can now combine the results using equation 4.1 to compute an estimate of the probability that user u will read a key user's tweet. As an example, suppose that u has 99 friends who tweet at average rate $\overline{R'} = 12.0877$ tweets/day, and that u himself tweets at the same rate, which means $R'(u) = 12.0877$ visits/day. In this case we get, from equation 4.2, $p_1 = 0.99$. Plugging everything into equation 4.1 gives us $P_{read}(u) = 0.128$. This means that even with 98 competing tweeters, a key user's tweets will be read by an average follower with probability 0.128.

Table 4.9: Estimated number of tweets of the key users read by their first degree followers in each category per day. "All" refers to the entire set of selected first-degree followers, ignoring those whose accounts are protected or have been suspended. "Under 18" refers to those first-degree followers whom we have classified as being under 18 years of age.

		Tobacco		**Marijuana**	
		Pro	Anti	Pro	Anti
All	Mean	4.27	0.74	6.41	0.0002
	Stdev	5.97	1.02	9.06	0.002
	Median	2.38	0.41	3.54	0.0
Under 18	Mean	4.67	0.81	7.06	0.0003
	Stdev	6.71	1.15	10.17	0.004
	Median	2.22	0.39	3.36	0.0

Next we combine the results from the preceding sections to estimate the rate at which users who are under 18 years of age are seeing pro-tobacco and pro-marijuana tweets.

4.3 Combining the Results

For each user, u, who is a follower of one of our 10 selected "key users", we estimate the probability that they read a tweet from the key user in one day. In equation 4.2, we replace $R(u)$ with $N_f(u)\overline{R'} - R(u_{key})$, where $R(u_{key})$ is the tweet rate of the key user that u follows. Then we calculate $P_{read}(u)$ as in the example in the previous section. Multiplying $P_{read}(u)$ by the corresponding key user's tweet rate gives us the estimated average number of tweets by the key user that u reads in a day.

However, not all of the key users' tweets are relevant. Some are not tobacco or marijuana-related. Therefore in each case, we multiply the key user's tweet rate by the fraction of their tweets that are pro- and anti-tobacco and marijuana. Taking the mean across all first-degree followers gives the mean exposure rate to each category of tweet for first-degree followers of the key users.

We do the same calculations for just the first degree followers of the key users. The results for both sets of users are summarized in Table 4.9.

The standard deviations in Table 4.9 are so high because the distributions are skewed to the right (i.e., heavy tailed), as shown in Figures 4.6.c and 4.6.d. Also note that these are the estimates of the exposure to just the tweets from the key users. Other followees of the first-degree followers may also be tweeting about tobacco and marijuana. Thus the estimates in Table 4.9 are under-estimates of the actual exposure rates of the first-degree followers to tobacco and marijuana-related tweets.

4.4 Contributions and Discussion

Our contributions are listed below. This study is under review for publication in a journal [177].

In this work we analyzed a large data set of tweets obtained through the Twitter API over the period March 1, 2013 to Jan 19, 2014. We trained machine learning classifiers to categorize the tweets into pro- and anti-tobacco and marijuana categories. We found that tobacco-related tweets tend to be somewhat pro-tobacco: 69.2%. We also found that marijuana-related tweets are overwhelmingly pro-marijuana: 94.3%. One caveat to note here is that the set of keywords we chose could skew these numbers. In particular, there are many slang words associated with marijuana, which were not included in our search terms. Doing a more comprehensive search and analysis of the resulting data is an important direction for future work.

We chose 10 heavily tweeting "key users" to focus on for detailed analysis. We crawled their follower network to two degrees of separation and selected a random subset of their first-degree and second-degree followers.

To identify underage Twitter users, we created a different data set by search for "happy birthday" tweets. Using this data set, we trained machine learning classifiers to predict whether a user is over or under 18 years of age based on their tweet content. This trained classifier was then used to classify the previously selected followers. We found that 36% of the first-degree and 33.5% of the second-degree followers are predicted to be under 18 years of age. This means that a significant number of adolescents are being exposed to pro-tobacco and pro-marijuana messaging on Twitter, which could be a cause for concern.

Quantifying the extent to which online exposure to pro-tobacco and pro-marijuana messaging effects smoking initiation and maintenance is an open question. In particular, a study like ours needs to be supplemented with a survey designed to elicit information about the effects of social media messaging on smoking behavior. This is an important direction for future research.

Finally, we used the model of Hogg et al. [84] to estimate the number of tweets in each category that the first-degree followers of the key users might be reading. Since the distributions are heavy-tailed, we suggest that the median is the most meaningful estimate. We find that underage first-degree followers are exposed to a median of 2.22 pro-tobacco tweets/day and 3.36 pro-marijuana tweets/day. Once again, this number is skewed by our choice of key users, though only to an extent. We chose the most heavily tweeting and persistent key users for our analysis, so the numbers we have calculated are representative of the most common exposure. Another caveat here is that these number are an *under-estimate*. This is because we don't know how many other followees of these users are also tweeting about tobacco and marijuana use.

Overall, we have shown how to address an important public health question through machine learning-based analytics of Twitter data. Tobacco and marijuana use are complex phenomena, affected by many factors. Efforts to limit the initiation of underage populations

into these behaviors need to take a multi-pronged approach. Our study suggests that understanding and regulating social media may be an important part of this approach, given the extent of pro-tobacco and pro-marijuana messaging to found, and the significant presence of underage populations on social media, in positions where they can be heavily exposed to these messages.

Chapter 5

Find and Analyze the Hotspots of Electronic Cigarette-related Tweets

5.1 Introduction and Motivation

Electronic cigarettes and other vaping devices (referred to as e-cig hereafter) are battery powered devices that deliver nicotine in the form of heated vapor. These devices are collectively called electronic nicotine delivery systems (or ENDS). The popularity and use of these products have grown enormously in the past few years. For example, King et al. found that e-cig awareness and use have doubled among adults between 2010 and 2013 [100]. Morover, the Center for Disease Control and Prevention (CDC) has reported that e-cig use tripled among middle and high school students between 2013 and 2014 [23].

Given the sudden multi-fold rise in the popularity of e-cig, it is important to identify spatially the regions where they are most activvely used. In this study, we used e-cig-related tweets to identify and analyze e-cig hotspots in the United States. A hot-spot of e-cig tweets can be defined as a statistically anomalous cluster in space and time where considerably higher numbers of e-cig tweets are observed than expected. High e-cig related activities on Twitter, such as e-cig information searching and sharing and heavy e-cig marketing and promotion, makes Twitter a very good platform for an e-cig surveillance study.

We used spatiotemporal scanning [105] of non-commercial geotagged e-cig tweets to identify these clusters. This required us to first identify the non-commercial e-cig tweets in the US and then to perform a spatiotemporal analysis using the spatial location and time stamp of

the tweets to identify the anomalous clusters across space and time. Using machine learning tools and techniques, we also analyzed the sentiments of the tweets and the age group of the users within these anomalous clusters.

Spatial scan statistics is widely used across many disciplines to identify anomalous clusters. It is used heavily in the surviallance of diseases such as: respiratory infectious [128]; food- and water-borne diseases [121]; sexually transmitted diseases [81]; vector-borne diseases [61]; and cancer [73]. It is also used for spatiotemporal analysis in, for example, studies related to: suicide [93]; natural disaster [167]; criminology [202]; forestry [48], and history and archeology [193]. Most recently, Twitter data has been used for examining flu trends at the city level using spatiotemporal scanning. Nagar et al used geotagged tweets from New York City, and validated the temporal predictability of daily tweets for visit to emergency departments for influenza-like illness [131].

A few Twitter-based e-cig surveillance studies have been reported in the literature. For example, Myslin et al. analyzed the content and sentiments of around 7,000 tweets using machine learning techniques [130]. They found a high prevalence of positive sentiments toward e-cig and hookah and also that sentiments were largely positive about smoking itself. Some of the studies have focused mainly on e-cig marketing on Twitter. Huang et al., for example, examined all e-cig-related tweets over a two-month period and found that 90% of these tweets were commercial [85]. They discovered that the tweets were not only overwhelmingly commercial, but also that most of them claimed health and smoking-cessation benefits. Grana et al. gathered and examined the content of e-cig retail websites [68] and found similar health-related benefits listed on the websites. Emery et al., on a different front, used an online survey of adults to analyze e-cig awareness, use, and information sharing and searching [45]. They found that 86% of the subjects were aware of e-cig, and that tobacco users were twice as likely as non-users to have seen or heard information about e-cig. In their study, Twitter was also found to be a medium that is used by regular e-cig users 17% of the time for sharing and 9% of the time for searching e-cig related information.

Unlike previous Twitter-surveillance studies on e-cig that either looked at small datasets or were focused on commercial tweets, we analyzed a reasonably large number of non-commercial geotagged e-cig tweets from the United States spanning two years. As these tweets were geotagged by users, our dataset was especially suited to spatial analysis. None of the previous e-cig studies have conducted a spatial analysis using Twitter data. The results from this study suggest that three-quarters of the spatiotemporal hotspots for e-cig tweet contain more pro-ecig tweets and more under-18 users compared to the national averages. Also, the majority of these hot-spots are located on the west coast of the US.

5.2 Materials and Methods

5.2.1 Description of the Dataset

Using the Twitter Streaming API, we collected tweets containing geographic metadata. This search was capped by Twitter's API limit of 1% of total tweet bandwidth. A tweet contains geographic metadata when either a users explicitly specify the tweet location using a mobile device or mentions about a place in the tweet. The date range used for this analysis was October 15, 2012 to October 15, 2014.

We filtered this dataset further using the following e-cig related keywords: *e-cig*, *electronic cig*, *vape*, *vaping*, and *ehookah*. This gave us a total 83,708 e-cig-related tweets, among which 62,894 tweets fell within continental US bounding box. The distribution of the filtered tweets over time is shown in Figure 5.1. A rise in the number of tweets can be observed over the 24 months with a global maximum in May, 2014. We believe that this increase is due to both an increase in popularity of e-cig and a changed data collection technique.

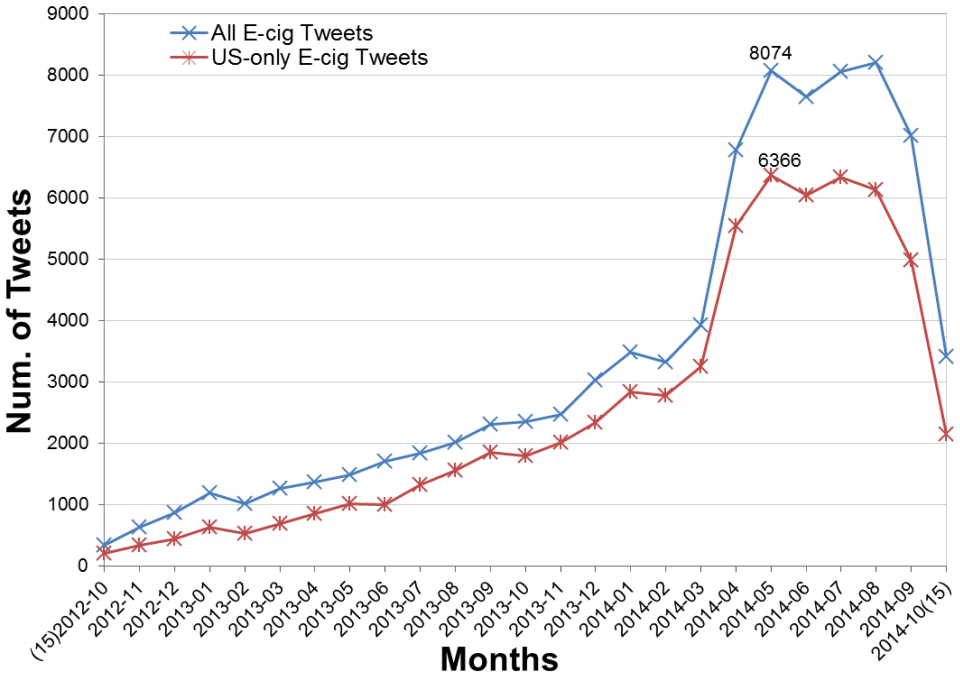


Figure 5.1: Overall counts of geotagged e-cig tweets per month.

5.2.2 Identifying Non-commercial E-cig Tweets

Our focus is on non-commercial tweets alone because they allow us to study the sentiment and population of the general public tweeting about e-cig. As previous studies have found that the majority of e-cig tweets are commercial [85], an important step was to identify non-commercial tweets from the collected e-cig tweets. We employed supervised learning techniques to identify the non-commercial tweets in the collected dataset.

Training data

First, we replaced the hyperlinks in the tweets with the hyperlink information using a combination of domain name, user name, and photo tags. The following format was used:

a) A photo sharing using a hyperlink is represented using the "<photo>" tag, such as [<instagram.com><photo>] if an Instagram picture was shared.

b) Hyperlinks leading to domain "twitter.com" may contain users' screen name or both screen name and photo id. Such hyperlinks are represented as [<twitter.com> <userScreenName>] or [<twitter.com><userScreenName><photo>].

c) Only the domain name is mentioned for the hyperlinks for which no other information could be inferred.

d) The tag "<urlError >" was used when a domain name was unable to be retrieved from the hyperlink. We used LongURL API to obtain the hyperlink information [120].

Next, we chose 5,000 random e-cig tweets and had these labeled on Amazon mechanical turk (AMT). The scorers were instructed to label the tweets based on the text and the hyperlink information where this was present. A tweet could be labeled in one of three categories: *commercial*, *non-commercial*, and *irrelevant*. A *commercial* tweet is a tweet that contains promotional text related to e-cigarette or other vaping products or a tweet that appears to be from a commercial website; a *non-commercial* tweet shares personal liking, disliking, experiences, habits, and current activities related to e-cig; and the *irrelevant* class contains tweets that are not about e-cig, tweets with not enough information for labeling, or tweets that are not in English.

The average agreement between two scorers for the given labeling task was 83%. This gave us 292 *commercial*, 3,744 *non-commercial*, and 114 *irrelevant* tweets. A sample of tweets per label class are shown in Table 5.1. Since we are interested in the *non-commercial* tweets, we merged the *commercial* and *irrelevant* classes into a single class (*merged-irrelevant*) for training the classifiers. This merging gave us more balanced training data for the binary classification task.

E-cig related feature extraction

We gathered features for classifier training by pre-processing the tweets using standard NLP tools and counting various entities in the tweets. First, the URL information tags were converted into strings based on tag types to make them more machine readable. For example, tags [<instagram.com><photo>] and [<twitter.com><userScreenName><photo>] were converted into strings *domainame_photo* and *domainame_username_photo* respectively. We then removed the '#' character from the hashtags but preserved the hashtag string and fully removed the mentions from the text. We also counted the number of occurrences of urls, hashtags, and mentions in a tweet, and appended strings in the text to note one or more occurrences. For example, presence of a single hashtag in the tweet was noted using *hasone_hashtag*, while more than one mention was noted using *hasmore_mentions*.

We also identified the presence of *self* and *other* words in the tweets using the word list presented in [109], and noted this by appending *has_self* or *has_other* to the tweet. Finally, we modified the tweet text by correcting: alphanumeric lengthnings, word contractions, and repeated punctuation, as well as removing the stop words. We then created uni-, bi- and tri-grams from all these gathered features, and selected 15,000 of the most useful n-grams for analyzing and evaluating the classifiers. We identified the non-English tweets using stop words and discarded these because our focus in this study is only on English tweets.

Table 5.1: Sample e-cig tweets per label. comm-ecig: *commercial* e-cig tweets, ncom-ecig: *non-commercial* e-cig tweets, and irr-ecig: *irrelevant* e-cig tweets.

Label (Total)	**Example Tweets**
comm-ecig (292)	*- Don't cut yourself short in this vape lyfe! COMING SOON! FIRST OFFICIAL E LIQUID/E JUICE [<instagram.com><photo>]* *- Do you think E-cigarettes are effective? Find out the popular votes [<www.tellwut.com>] @user_1*
ncom-ecig (3744)	*- Smoking an e-cig while walking in a mall #areuretarded #ulookit* *- I'm at The Daily Vape [<www.swarmapp.com><photo>]* *- @user_2 vape me*
irr-ecig (114)	*- E-cigarette* *- Kopi biskuit vaping,tiap pagi begini sarapan nya,seminggu paling tipes -_- (at CIMB NIAGA) [<urlError><photo>]*

Table 5.2: Some of the most informative features used by the trained *SVM-lin* classifier to label *non-commercial* and *merged-irrelevant* tweets. The *merged-irrelevant* class consists of tweets from *commercial* and *irrelevant* classes.

non-commercial	***merged-irrelevant***
hasone_url hasmore_mentions	kedai (means shop in Malay)
ban	has_other hasone_url hasone_mention
domainame_username_photo	vape hahaha
domainame has_self	vape hose
vaping	free

Classifier evaluation

A set of SVM classifiers was evaluated for the binary classification task — non-commercial vs merged-irrelevant— using 10-fold cross-validation and randomly selected training(90%) and testing (10%) data for each fold. A standard implementation of the algorithms from the Scikit-learn Python library was used [146]. As the data in the two classes was unbalanced, we performed the evaluation with a wide range of classifier parameters and selected the parameter set that gave us a balanced classifier. A balanced classifier is a classifier that give us reasonably good accuracy, less overfitting, and high precision and recall for both data classes.

Using this process, we selected an SVM classifier with a linear kernel for the given classification task. This classifier gave us average accuracy of 0.919. Average precision, recall, and F1 score for the *non-commercial* class (*merged-irrelevant* class) were 0.956 (0.543), 0.955 (0.539), and 0.955 (0.535), respectively. A set of the most informative features used by this classifier to distinguish the two classes is shown in Table 5.2.

The features in the two classes are very different from each other. The most informative features for the *non-commercial* class, such as "*hasone_url hasmore_mentions*" and "*domainame_username_photo*", suggest that a single URL along with more than one mention in a tweet, as well as a shared Twitter picture in a tweet, are good identifiers of a non-commercial e-cig tweets. A user talking about herself (captured using *has_self*), and words such as "ban" and "vaping" are also very informative for labeling this class. We verified our data and found that there are many non-commercial tweets where a user share a picture with other users using mobile phone applications such as Swarm [169] or Instagram [88]. Swarm also allows users to check in at places and share their location on social media.

The informative features for the *merged-irrelevant* class suggest that features such as the trigram "*has_other hasone_url hasone_mention*" and the words "kedai" and "free" are the good identifiers of this class. If we observe closely, the tri-grams and the words exhibit the essence of advertising tweets: the tweets address other users (captured by feature *has_other*); company's products are promoted by sharing its URL (*hasone_url*)and Twitter handle (*hasone_mention*); and the deals and promotions (*free*) are shared. The feature "kedai", meaning shop in Malay, is a good identifier of commercial (non-English) tweets.

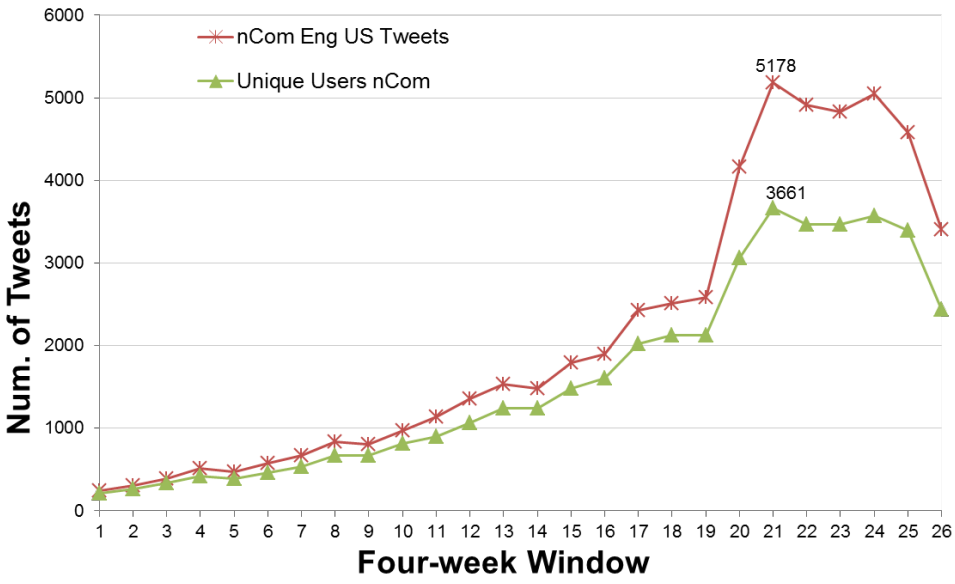


Figure 5.2: Non-commercial e-cig tweets and unique users over four-week time windows. Total 54,597 tweets and 41,593 unique users.

Non-commercial e-cig tweets in the US

For unlabeled e-cig tweets, as with the training data we replaced the hyperlinks with the appropriate tags and preprocessed them to obtain various features. Using the selected classifier on this dataset, we found that 54,597 (87%) of the e-cig tweets from the continental US were non-commercial. We aggregated the tweets in four-week windows and plotted them along with unique users in Figure 5.2. A total of 41,593 unique users posted these tweets. In order to adjust the day-of-week effect, we aggregated daily data into weeks and then clubed these into groups of four to obtain non-overlapping, month-equivalent time windows for the analysis.

5.2.3 Spatiotemporal Scan of Non-commercial E-cig Tweets

The next important step in our study was to identify the statistically significant anomalous clusters (or hotspots) of non-commercial e-cig tweets in the US. We used spatial scan statistics for this purpose. This technique is widely used in epidemiology for disease surviallance studies to monitor or detect specific diseases as discussed earlier.

Detection of anomalous spatial clusters involves a series of steps, beginning with obtaining case counts and population-at-risk data for a set of spatial locations and then performing randomization testing to identify statistically significant clusters of these locations. A brief overview of these steps can be found in [132]. These steps are based on the spatial scan

statistics approach of Kulldorff [105].

Obtaining case counts and population-at-risk

In a disease surveillance study of, for example, influenxa-like illness the case counts are obtained by collecting the number of doctor visits for influenza-like illness per county or zip code after every given time window. Similarly, the population-at-risk are the individuals in the county or zip code who have the potential to contract the flu in the given time period. For our study, the number of unique users tweeting about e-cig from a particular county is equivalent to the case count, and the individuals who have the potential to tweet about e-cig from that county are equivalent of the population-at-risk.

We calculated the case counts and population-at-risk for each county per four-week time window (hereafter referred to as a time window). Only tweets that fell within the continental US bounding box, and where country code equaled "US" and language code equaled "en" were used. The population centroids of each county [183] were treated as a *spatial points* at which spatial scanning was later performed.

The e-cig tweet case counts were obtained by aggregating e-cig tweets at county level for each time window in the following way. First, we identified which county each tweet belonged to using the tweet's geo-coordinates and county polygons. Next, we assigned the unique user count of these tweets per county to the population centroid of each county. This is equivalent to obtaining the number of unique disease cases per county for four-week period. Since we needed a set of *spatial points* to perform the scan statistics, the counts were assigned to the population centroids.

The population-at-risk count were obtained by sampling all the HealthMap geotagged tweets time window by time window. We first collected a random 2% sample of the HealthMap geotagged tweet data and then calculated the population-at-risk per county using the the same approach as when computing e-cig tweet case counts. We gathered a total of 20,323,775 tweets and 8,384,647 unique users by sampling the data for the study period. The distribution of the tweets and users is shown in Figure 5.3.

Model of data and score function

The discrete Poisson model [105] was used for the spatiotemporal analysis of e-cig tweet clusters. With this model, the number of cases at each *spatial point* is Poisson-distributed. This means that under the null hypothesis the disease rate is uniform everywhere and that the number of cases at each *spatial point* is proportional to the population-at-risk at this point. Therefore, the goal of scan statistics is to find the *spatial points* (or clusters of *spatial points*) S where the disease rate is higher inside than outside.

We adopted the most common hypothesis testings approach that was also used in [132]. As

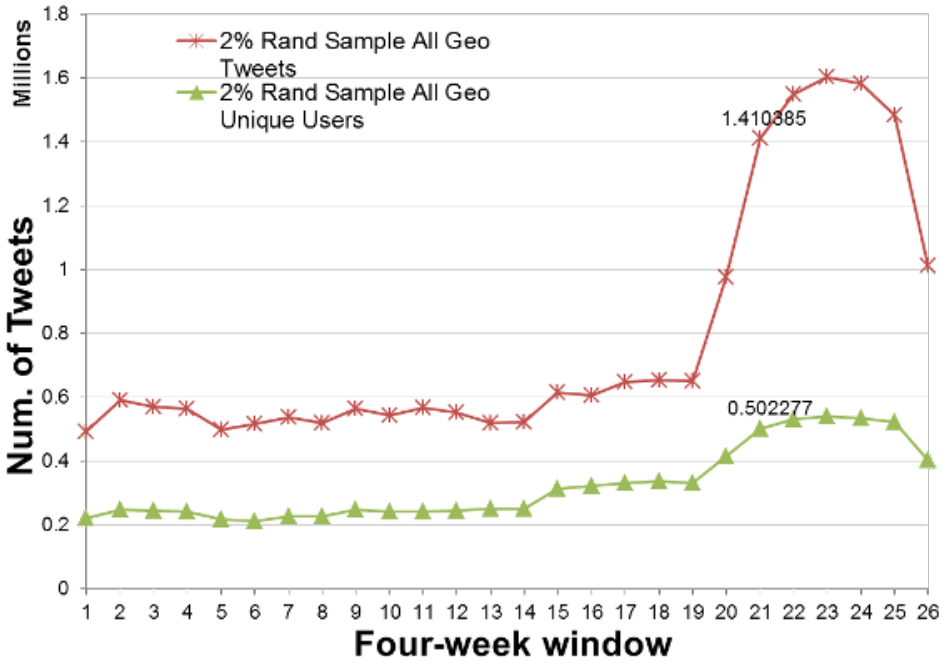


Figure 5.3: Two percent random sample of all HealthMap geotagged tweet from continental US per four-week time windows. Total 20,323,775 tweets and 8,384,647 unique users.

per this approach, we first compute a score *F(S)* for each *spatial point S* using the likelihood ratio with the maximum likelihood estimates of any free parameters, and then calculating the statistical significance of clusters by randomization testing using a large number of replicas of the data.

5.2.4 Analysis of anomalous spatiotemporal clusters

After identifying the statistically anomalous cluster, the next step was to analyze these with respect to pro-e-cig sentiments and users in the under-18 age group. We used supervised machine learning techniques to perform the sentiment analysis of the tweets and to infer the users' age. The process to identify the pro-e-cig sentiment is discussed next.

Identifing pro-e-cig sentiments in non-commercial tweets

Training data

We selected all 3,744 hand-labeled *non-commercial* e-cig tweets from the previous AMT task and had these labeled as *pro-ecig*, *anti-ecig*, and *notSure-ecig* on AMT. The *pro-ecig* tweets

Table 5.3: Sample *non-commercial* e-cig tweets per sentiment labeling classes: *pro-ecig*, *anti-ecig*, and *notSure-ecig*.

Label (Total)	**Example Tweets**
pro-ecig (1929)	*- E-cig till it dies!* *- I'm at GASPANIC bar And Restaurant / THE VAPE SHACK[<www.swarmapp.com>]* *- LAME @user_1: I think I really lost my vape pen. This sucks.*
anti-ecig (307)	*- I will never understand why people smoke pot/ vape.* *- Kid in front of me at graduation is wearing a baseball hat with a tassel and smoking an e-cigarette. WORST.*
notSure-ecig (343)	*- I just experienced the weirdest talk about electronic cigarettes* *- my mom is sitting here smoking an e-cigarette in the middle of the airport and everyone's staring.* *- vape pens are stupid af but they're so addicting*

are the tweets that show a positive opinion, sentiment, or experience about e-cig, and also the tweets that share and/or support the activities associated with e-cig usage. The *anti-ecig* tweets are the tweets that opposes, discourages, or does not support e-ciga usage. The category *notSure-ecig* includes all the tweets where identifying the polarity of opinion about the e-cig usage is difficult. This can be the case where the content of a tweet is insufficient to make a call on its polarity, or where a tweet contains both positive and negative sentiments.

The average agreement between the pairs of scorers for the given labeling task was 69%. The labeling gave us 1,929 *pro-ecig*, 307 *anti-ecig*, and 343 *notSure-ecig* tweets. There were 1,165 tweets for which no-agreement was found between the scorers. A sample of tweets per label class are shown in Table 5.3. As we are interested in analyzing the anomalous e-cig tweet clusters with respect to *pro-ecig* tweets, we merged the *anti-ecig* and *notSure-ecig* classes into a single class (*merged-other*) for training the classifiers similar to the last time. This merging also helped us in balancing the training data to an extent for the binary classification task.

Feature extraction and classifier evaluation

The tweet preprocessing and features extraction process for this supervised learning task was very similar to that discussed in Section 5.2.2. We appended the hashtag string and URL information strings; the number of occurrences of urls, hashtags, and mentions was noted; the presence of *self* and *other* words was captured; and also the text of the tweetswas modified in the same way as before. However, in addition to these features, we also extracted positive and negative emoticons in the tweets using the emoticon library similar to the one presented in [1]. The presence of positive and negative emoticons was noted using *has_emotpos* and

Table 5.4: Some of the most informative features used by the trained *SVM-lin* classifier to label *pro-ecig* and *merged-other* tweets. The *merged-other* class consists of tweets from *anti-ecig* and *notSure-ecig* classes.

pro-ecig	***merged-other***
has_self	has_other
vaping	cigarettes
love	hate
has_emotpos	gay
quit smoking	ban
domainame_username_photo	wonder

has_emotneg respectively.

Using these features, we evaluated a set of SVM classifiers for this binary classification task. The 10-fold cross-validation and parameter tweaking was performed in the same manner as discussed in section 5.2.2. Using this process, we selected an SVM classifier with average accuracy of 0.798. Average precision, recall, and F1 score for the *pro-ecig* (*merged-other*) class using this classifier were 0.847 (0.622), 0.892 (0.525), and 0.868 (0.565), respectively.

Some of the most informative features used by the classifier to distinguish between the two classes are shown in Table 5.4. There is a visible difference in the type of words used in the tweets of the two classes. For example, "vaping", "love", and "quit smoking" are good identifiers of the *pro-ecig* class, whereas words, such as "hate", "gay", "ban", and "wonder" identify well the *merged-other* class well. Similarly, more tweets where people talk about themselves (*has_self*) or use positive emoticons (*has_emotpos*) tend to belong to the *pro-ecig* class, whereas tweets that refer to other people (*has_other*) tend to be from the *merged-other* class.

Age-group inference of the users

In order to infer the users' age group, we used the technique discussed in [177]. The classifiers were trained and evaluted using the features collected from the 150 "valid" tweets from the most recent 400 timeline tweets of users who tweeted "happy Nth birthday to me" or "happy birthday to me #N", where N was their age. The average accuracy of the classifier was 0.848, and the average precision and recall for identifying the *under-18* class was 0.627 and 0.631 respectively. Please refer to [177] for more details.

Figure 5.4: Anomalous clusters on west coast of the United States. A circle in the map represents the presence of multiple counties in a cluster. Top five cluster are denoted using arrows.

5.3 Results

5.3.1 Anomalous clusters

We performed spatiotemporal retrospective analysis of the e-cig tweet clusters using SaTScan [106]. The scan was performed using cylindrical windows with a circular geographic base and with height corresponding to time. For each *spatial point* at center, the size of the circular geographical window varied between zero and either a 50 Km radius or upto 10% of the population-at-risk, whichever reached first. Similarly, the height of the cylinder i.e., temporal window size, was varied between one and five time windows (i.e., from 4 to 20 weeks). The statistical significance of the spatiotemporal clusters was computed using the Monte Carlo randomization testing with 999 data replicates.

A list of the top-twenty clusters arranged in non-increasing log-likelihood ratio is presented in Table 5.5. A cluster may consist of multiple adjacent counties not exceeding 50km radius and may spread across one to five consecutive time windows denoted by "Start-End" time, Also, a county name is shown for each cluster for spatial reference. The majority of these clusters is located on the west coast of the United States, as can be seen in Figure 5.4. We did not consider clusters with less than 50 case counts in our analysis.

The case counts and population-at-risk counts used by SaTScan for analysis are also shown in the table. Note that the case counts and population-at-risk counts were computed by identifying the unique users of tweets for each county. The counts shown in the table are the aggregated counts per cluster using the counts of each county that belong to that cluster. The last two columns show the total e-cig tweets and unique users per cluster. We analyzed the clusters by classifying these tweets and users into various classes. The analysis results are presented next.

Table 5.5: Top twenty statistically significant spatiotemporal anomalous clusters. A cluster may comprise of multiple adjacent counties and multiple time windows as per the parameter set used with SaTScan. One time window is four week.

Id	**County(or one of the counties)**	**Counties in the cluster**	**Start-End time**	**Pop-at-risk count**	**Case count**	**Total e-cig tweets**	**Total unique users**
1	Los Angeles, CA	1	18-22	13779	881	1219	764
2	Orange, CA	1	21-25	3756	439	586	373
3	King, WA	4	21-25	3867	387	546	314
4	Alameda, CA	5	21-25	6471	547	925	470
5	Riverside, CA	1	22-26	2363	224	317	181
6	Multnomah, OR	3	20-24	1735	202	303	164
7	Erie, NY	2	18-22	1688	163	219	138
8	Fresno, CA	2	20-24	4520	440	571	395
9	Clark, NV	1	10-14	3227	162	215	137
10	Hennepin, MN	8	17-21	3544	213	266	187
11	San Diego, CA	1	21-25	3473	269	352	230
12	San Joaquin, CA	1	18-22	619	76	111	64
13	Ventura, CA	1	21-25	868	104	114	95
14	San Bernardino, CA	1	22-26	2360	184	251	151
15	Hidalgo, TX	2	19-23	740	85	110	72
16	Collin, TX	3	17-21	1970	133	152	117
17	Davis, UT	4	20-24	1481	125	149	104
18	Kittitas, WA	1	20-24	978	116	161	102
19	Placer, CA	2	15-19	1558	92	110	76
20	Aransas, TX	4	19-23	665	68	80	61

5.3.2 Analysis with respect to pro-ecig sentiments

Next, we considered the tweets that belonged to these clusters and analyzed their sentiments. For each cluster, we used the classifier discussed in section 5.2.4 to label the e-cig tweets into *pro-ecig* and *merged-other* classes. There were a total of 6,757 e-cig tweets from the top twenty anomalous clusters. Figure 5.5 shows the fraction of e-cig tweets that was found to be *pro-ecig* for each cluster. We also inferred the sentiments of all the e-cig tweets together and found that the national average of the fraction was 0.777. Results show that 15 out of 20 clusters contains tweets that were above the national average rate in terms of of pro-ecig sentiments.

5.3.3 Analysis with respect to users under age 18

As with the tweets, the users of these cluster were analyzed with respect to their age. We used the age classifier briefly discussed in Section 5.2.4 to infer the age group of e-cig tweet users in these clusters. We inferred the age group of a total of 4,195 unique users, and plotted the fraction of users aged *under-18* for the clusters as shown in the Figure 5.6. We also identified the unique users from all the e-cig tweets and inferred their age group. The national average fraction of unique e-cig tweet users in the *under-18* class was 0.19. Again, 15 out of 20 clusters were found to be have more than the national average fraction of users under age 18.

5.4 Discussion and Contributions

The contributions and future directions of this study are listed below. This study is under preparation for submission to a journal [176].

In this work, we analyzed a large data set of geotagged e-cig-related tweets obtained over two years from October 15, 2012 to October 15, 2014. We trained machine learning classifiers to first filter non-commercial tweets from the dataset. We found that more than 87% of the collected e-cig tweets were non-commercial. The overwhelming presence of non-commercial tweets in the dataset can be attributed to the type of e-cig tweet used in the study. We think the data make sense because tweets are usually geotagged by non-commercial Twitter users but not by an advertiser or a business account.

We combined spatiotemporal scan statistics and machine learning tools and techniques to infer the age of users and the sentiments of tweets from the highly active clusters of e-cig tweets in space and time. To our knowledge, this is the first study of its kind to analyze the hotspots of e-cig-related messaging in social media. It is important to note that we chose a generic set of keywords representing e-cig-related used in the study may be a small subset

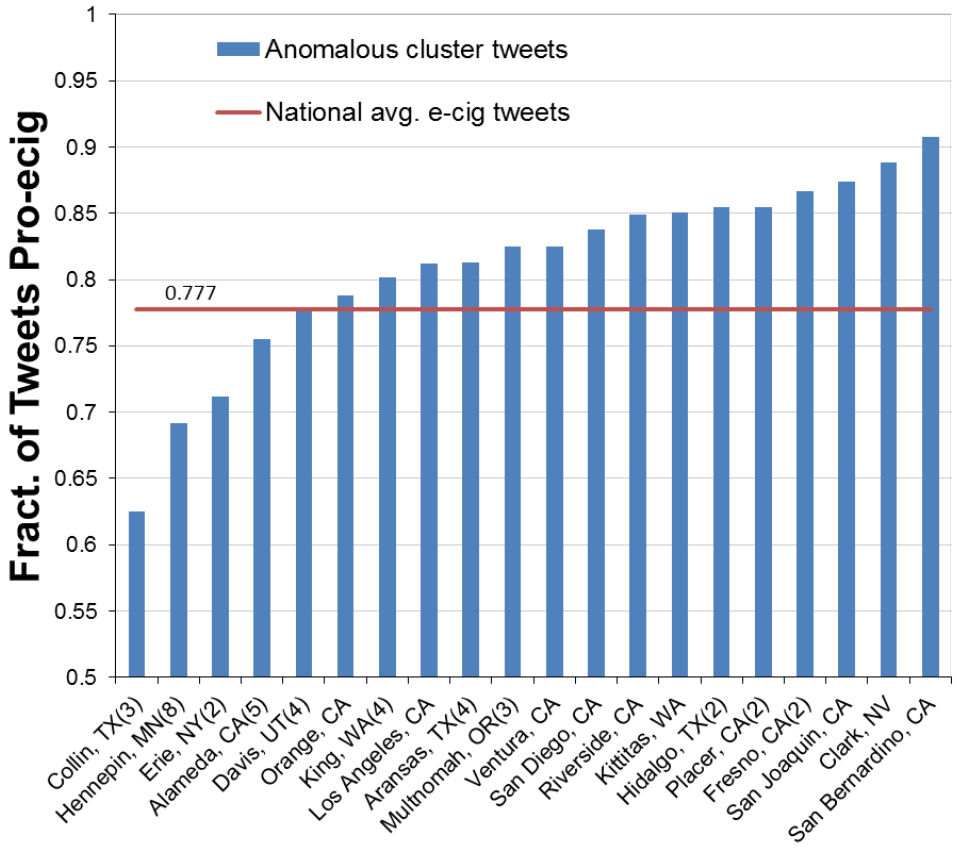


Figure 5.5: Fraction of e-cig tweets that are *pro-ecig* for each anomalous cluster. The clusters are arranged in increasing order of this fraction. Each cluster is denoted using a representative county name and number of counties it consists of, if it is more than one.

of all the e-cig-related communication over Twitter. Our results should be read carefully considering this caveat.

A potential extension would be to study the relationship between exposure to messages related to e-cig and related products and adoption and continuing use of these products. Both e-cig-related messaging and exposure to such messaging can impact smoking behavior and vice versa. To address this, a survey-based study should be performed primarily focusing on the younger population. Asking questions about new smoking products will be very helpful in understating and validating e-cig popularity in the identified hotspot communities.

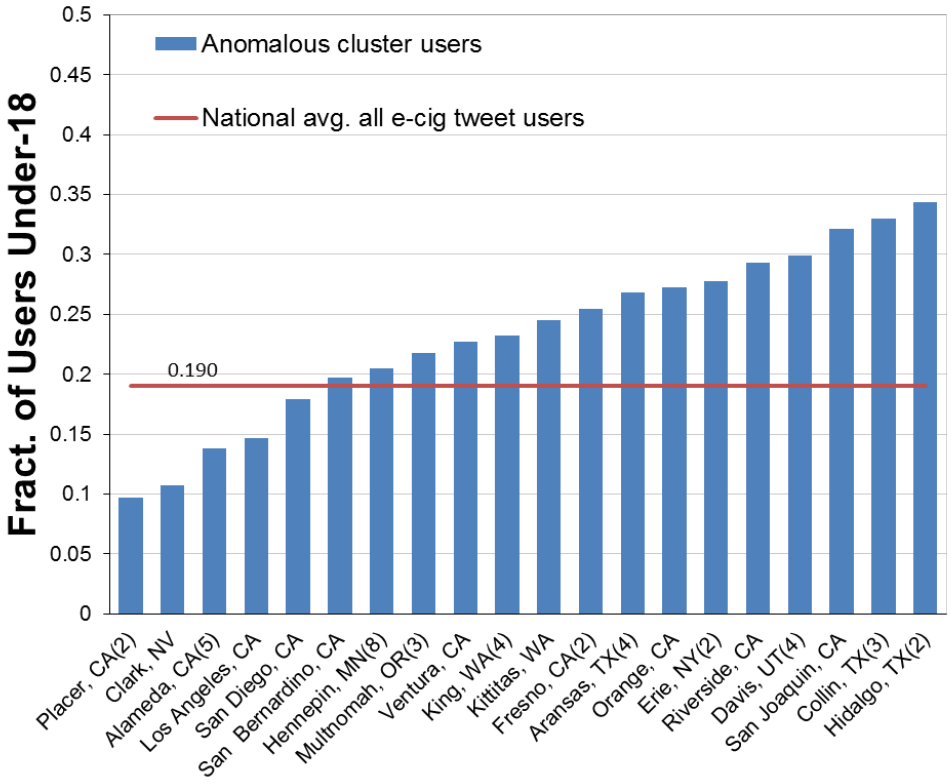


Figure 5.6: Fraction of e-cig tweet users that are labeled *under-18* for each anomalous cluster. The clusters are arranged in increasing order of this fraction. Each cluster is denoted using a representative county name and number of counties it consists of, if it is more than one.

Chapter 6

Combined Effect of Addiction Dynamics and Peer Influence on Smoking Epidemic

6.1 Introduction

Nicotine, in the form of cigarette smoking or chewing tobacco, is one of the most heavily used addictive drugs, and the leading preventable cause of disease, disability, and death in the U.S. [139]. It imposes a significant health-care burden on the population. The economic costs of smoking in the United States are estimated at $193 billion annually ($97 billion in productivity losses from premature death and $96 billion in health-care expenditures).

Like any other addictive drug, cigarette smoking behavior becomes compulsive and difficult to cease even after knowing the substantial health benefits of quitting [137]. Recent studies show that 35 million smokers express a desire for quitting smoking each year, but more than 85 percent of those who try to quit on their own relapse within a week [140]. Nevertheless, despite sustained and significant efforts by governmental and non-governmental institutions, smoking prevalence among youth and adult smokers has only declined slowly from 45% to 21% in the past 45 years [53, 54, 138] (see figure 6.1).

It has been repeatedly shown that smoking behavior is contagious, i.e., that peer influence (including family members) is the strongest factor in both initiation and cessation of smoking

[30, 63, 65, 82]. From an epidemiological viewpoint, the SIS model seems appropriate for modeling the contagion of smoking behavior, since smokers who quit can relapse. Here, the S state (which stands for "Susceptible") corresponds to non-smokers and the I state ("Infected") stands for smokers. However, the slow decline of smoking prevalence is puzzling from this perspective, as we shall discuss in the next section.

Our main contribution in this work is to introduce an extension to the SIS model, which we call the *structured resistance model*, to account for the addictive nature of smoking behavior. In this model we have multiple S and I states corresponding to increasing levels of addiction. We present this model in section 7.1.4 and we present simulations with this model on the Framingham Heart Study social network [49] in section 7.1.6. We end with a discussion of the model and possible directions for future work.

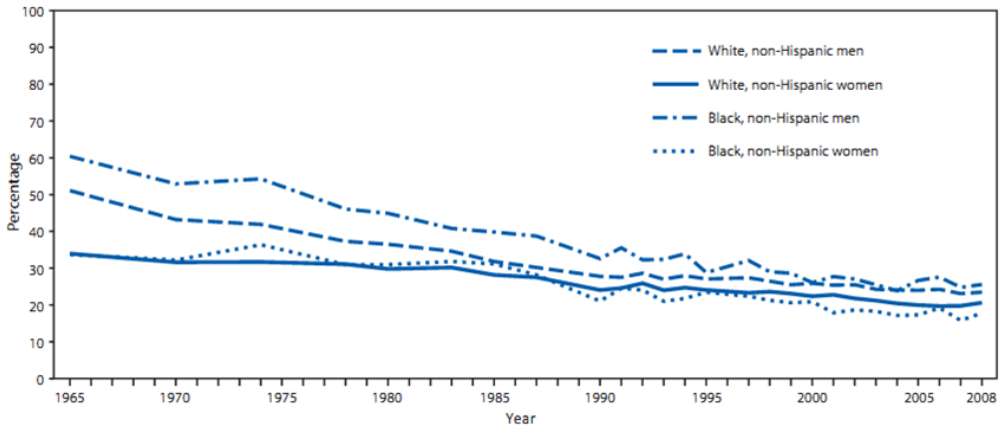


Figure 6.1: Smoking prevalence has declined slowly over the course of four decades. Source: CDC (*http://www.cdc.gov/mmwr/pdf/other/su6001.pdf*, p. 109).

6.2 Modeling the Smoking Epidemic

For the standard SIS model of epidemics, it is well known that there is an epidemic threshold, β_c, such that if the probability of infection is below β_c, the only steady state is the disease free state, while for values of the probability of infection above β_c, the only steady state corresponds to a finite fraction of the population being in the I state [19]. Figure 6.2 illustrates this scenario numerically. It can also be calculated analytically for some classes of networks.

If we assume that the effect of anti-smoking efforts is to reduce the infectiousness of smoking, then we would expect the prevalence of smoking to reduce fairly sharply, and to disappear once the infectiousness becomes low enough. This expectation stands in contrast to the reality depicted in figure 6.1, which shows that after much effort, the prevalence of smoking has declined steadily, but very slowly over more than four decades.

This puzzling contrast suggests that the SIS model is not quite right for understanding the contagion of smoking behavior. We argue that it misses the defining feature of smoking behavior, which is that smoking is addictive. Therefore we need a new model.

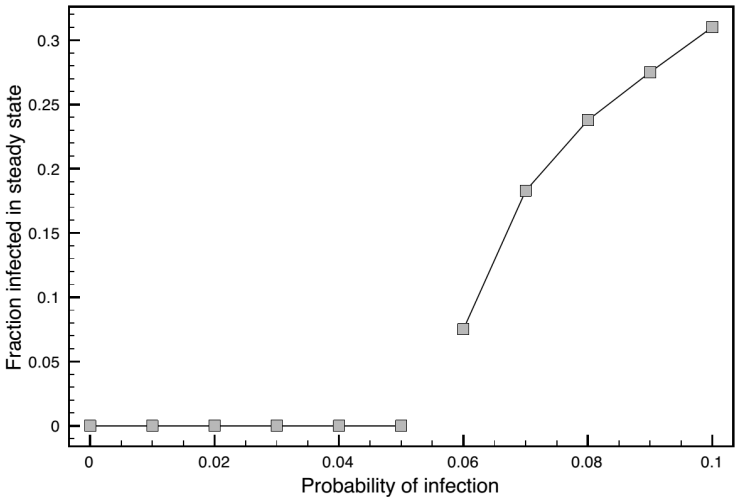


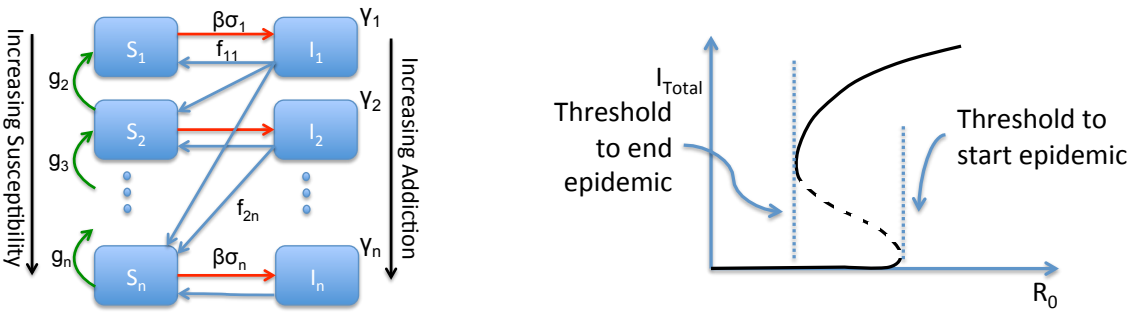
Figure 6.2: Bifurcation diagram for the SIS model.

6.3 The Structured Resistance Model

To model the dynamics of addictive behavior, we adapt a model developed by Reluga et al. [153]. Their study investigates the dynamics of disease immunity. We invert its semantics to model resistance to addictive behavior. This *structured resistance model* is shown in figure 6.3a. The multiple S states correspond to increasing levels of susceptibility to the behavior, and the multiple I states correspond to increasing levels of addiction. Initially, an individual starts out in state S_1, and moves to state I_1 upon adopting the behavior. The probability of this transition is given by $\beta\sigma_1$, where β is a multiplier on all $S \to I$ transitions and will be taken to correspond to R_0. The rate at which individuals quit, i.e., transition from an I state to an S state, is given by the corresponding γ_i parameter. Crucially, since the behavior is addictive, transitions from I_i only go to $S_{j>i}$. This means that when an individual quits, his susceptibility level is at least as high (and possibly higher) than it was before, but never lower. The probability of making the transition $I_i \to S_{j>i}$ is given by f_{ij}. The only way to recover to a lower level of susceptibility is via the $S_j \to S_{j-1}$ transitions, the probability of which is given by g_j. This is meant to model the fact that if an individual stays free of the addictive behavior for a long time, his level of susceptibility can decrease.

For a fully mixed population, the state update equations can be written as follows [153]. for $j = 1, ..., n$. We assume that $g_1 = g_{n+1} = 0$, since these transitions correspond to states that do not exist in the model.

Note that an individual makes an $S \to I$ transition only if one or more of his neighbors are in an I state. However, an individual makes the $S_j \to I_j$ transition with probability $\beta\sigma_j$, no matter *which* I state its neighbor is in. In other words, the contagion spreads from individuals in I states to individuals in S states, but does not depend on the details of which I states the spreaders are in. This is why we have an I_{Total} term in the equations instead of terms for each of the I states.



(a) The Structured Resistance Model (b) Schematic of a backward bifurcation

Figure 6.3: The structured resistance model is shown on the left. Some parameters are not marked for clarity, but these correspond to the ones that are shown. See text for details. A schematic of a backward bifurcation is shown on the right. The solid lines indicate stable steady states and the dashed line indicates an unstable steady state.

From these equations, it can be shown that when the quantity,

$$Q = -\sigma_1 - \sum_{j=2}^{n} \frac{\beta\sigma_1}{g_j}\left(\frac{\sigma_1}{\gamma_1} - \frac{\sigma_j}{\gamma_j}\right)\left(1 - \sum_{k=1}^{j-1} f_{1k}\right) \tag{6.1}$$

is positive and increasing in β, then the epidemic bifurcation is a backward bifurcation. This means that the bifurcation looks like in figure 6.3b (the "threshold to start epidemic"). There are actually multiple bifurcations and multiple steady states in this scenario. The solid lines in figure 6.3b indicate the stable steady states and the dashed portion indicates an unstable steady state.

The bifurcation diagram can be effectively divided into three regions. From $R_0 = 0$ to the "threshold to end epidemic", there is only one stable steady state, which corresponds to the entire population being in the susceptible state. In this range, the contagion does not take off, no matter what the initial state may be. Similarly, from the "threshold to start epidemic", for higher values of R_0, there is only one stable steady state, which corresponds to the contagion becoming endemic, i.e. if even only very few individuals are initially in an I state, a finite fraction will be in the I states in the long run. In between these two regions, we have a region where there are two stable steady states and one unstable steady state. In this region, if the initial state starts above the dashed curve, the population will move to the upper steady state, while if it starts below the dashed curve, the population will move to the lower steady state.

Practically, this means that for a new addictive behavior to become endemic in the population, its "infectiousness" must be higher than the threshold to start the epidemic (which is higher), but once it becomes endemic, for efforts to counter it to be successful, they must succeed in reducing the infectiousness of the behavior to below the threshold to end the

epidemic (which is lower). Intuitively, this means that significantly more effort might be required to end the epidemic than expected.

The equations above describe the behavior of the model in a fully-mixed population, or equivalently on a fully-connected network. In the next section, we do simulations to investigate the model on a more realistic network. Since we are interested in understanding the decline in smoking prevalence over a period of decades, we use a dynamic network that has been constructed from the Framingham Heart Study [49], which is a longitudinal study spanning precisely that period.

6.4 Simulations

Since the increasing levels in the structured resistance model represent increasing levels of susceptibility and addiction, we choose the probability of $S_i \to I_i$ to be increasing with i, and the recovery rates γ_i to be decreasing with i. The f_{ij} values, which control the $I_i \to S_j$ transitions, are chosen to be zero when $j < i$, as mentioned earlier. The entire set of parameters is shown in table 6.1. It turns out that any set of parameters chosen according

Table 6.1: Parameters for simulations with the structured resistance model.

Level, i	Infection probability, σ_i	Recovery rate, γ_i	f_{i1}, f_{i2}, f_{i3}	Resistance waxing rate, g_i
1	0.05	0.7	0.4, 0.4, 0.2	0.0
2	0.5	0.5	0.0, 0.7, 0.3	0.2
3	0.7	0.3	0.0, 0.0, 1.0	0.1

to these conditions will result in the model exhibiting a backward bifurcation. We can verify this is the case by substituting these parameter values into equation 6.1.

We conduct simulations using the Framingham Heart Study (FHS) social network. The FHS is a longitudinal study that gathered data on many health characteristics and health behaviors. The social network is a time-varying network spanning the years 1971-2008 (i.e., starting with the "offspring cohort" of the FHS). The offspring cohort were recruited into the study at an early age (the lowest age in the data is 5 years), thus giving us a network with children and adolescents as well as adults. Edges in the network correspond to various social and familial relationships. For the present work, we assume each edge to be an undirected edge along which the contagion can spread in either direction. For each edge present in the network, the data provides a start month and an end month. Thus edges are present at different times and for different durations. The degree distribution for the union graph (which assumes all the edges are present for all time) is shown in figure 6.4. The Framingham Heart Study data has been used in other instances of the study of the spread of smoking [30].

The bifurcation diagram for the structured resistance model on the Framingham Heart Study

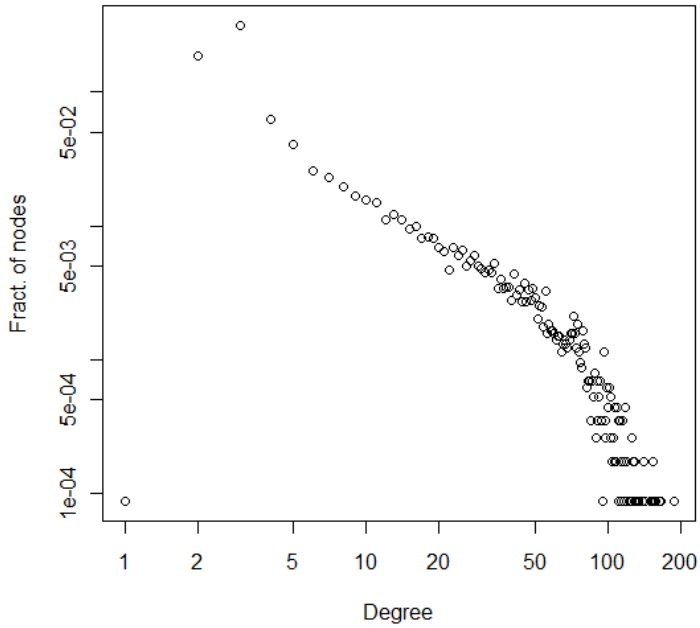
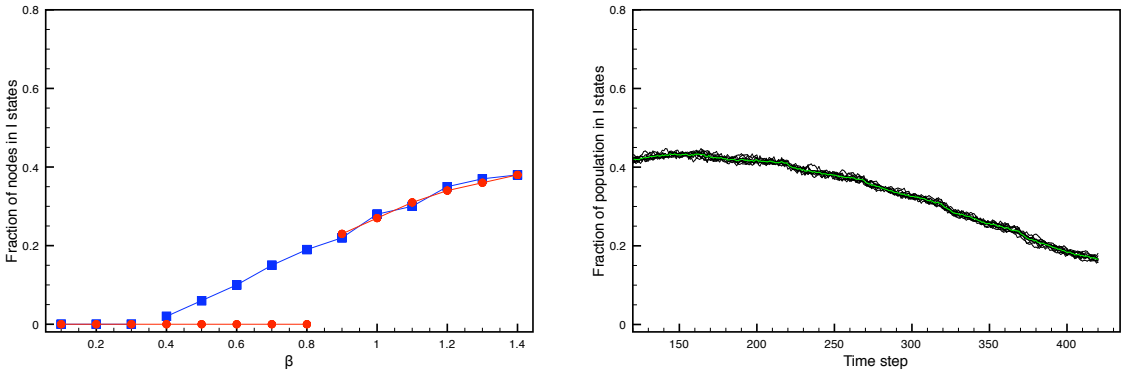


Figure 6.4: The degree distribution of the Framingham Heart Study union graph. It is not scale-free.

social network is shown in figure 6.5a, using the parameters in table 6.1. The diagram is obtained by doing simulations with two different initial conditions. In the first condition, the number of initial infections is very small, and in the other, the number of initial infections is large. In the region where there are two stable steady states, the first initial condition converges to the lower steady state and the second to the upper one. The lower threshold, shown in blue, corresponds to the upper steady state, while the upper threshold, shown in red, corresponds to the lower stationary state (the backward bifurcation). We see that there is a large gap between the two thresholds, which suggests that once the behavior is endemic, a large amount of effort is required (β has to be brought down a lot) to entirely eliminate the behavior from the population. Note that the unstable equilibrium is not shown because it is hard to determine numerically.

Figure 6.5b shows a sample epicurve obtained as follows. We initialize the population by randomly setting 5% of the nodes to be in state I_1 while the rest are in state S_1. The value of β is chosen to be 1.3, which is well above the upper threshold. We run the model until it reaches a stationary state, which corresponds to about 42% of nodes in I states. Then β is decreased slowly to simulate increasing awareness of the dangers of smoking and increased resistance to initiation. This causes the proportion of nodes in I states to decrease along the blue curve in the bifurcation diagram, which results in a slow decline in smoking prevalence. The epicurve is plotted from this point on in figure 6.5b. This qualitatively matches the empirical data reported by the CDC (figure 6.1).



(a) Bifurcation diagram

(b) Epicurve

Figure 6.5: The bifurcation diagram for the structured resistance model on the Framingham Heart Study social network is shown on the left. A sample epicurve is shown on the right, which exhibits the slow decline in smoking prevalence as β is decreased after time step 120. The green curve shows the average of the black curves (which show individual simulation results).

6.5 Contributions and Discussion

We have presented a multi-level SIS model, called the structured resistance model, that follows a set of rules to capture the dynamics of addictive behavior. Levels in the model correspond to increasing susceptibility and addiction to a behavior. We performed an agent-based simulation study using the proposed model on the social network of one of most heavily cited study in public health community. The edges in the network were used to replicate the peer influence for the diffusion of smoking behavior through the network. This model exhibits a backward bifurcation, which suggests a possible reason for the slow decline of smoking prevalence in the United States. Using simulations, we were able to numerically show the presence of two different steady states. This study was published in a conference [178]

This basic model can be extended in various ways. There are factors other than peer influence that affect smoking behavior, such as socioeconomic status and marital status [22], access to cigarettes and exposure to advertising [79], and prices and policies [116]. Data about all these factors can be included into an agent-based model driven by the basic structured resistance model. Detailed synthetic information environments can be constructed by fusing data about these behaviors with other data sets on demographics, locations, and activities to build a complete picture of the ecology of a smoker [9].

Smoking is a complex, "policy-resistant" problem. We believe that mathematical modeling and simulation-based approaches are essential to understanding such systems and to achieving lasting social benefits.

Chapter 7

Blocking smoking contagion

7.1 Community-based Blocking

7.1.1 Background and Motivation

A **contagion** is any entity that can spread through a population. Examples include online and face-to-face information, innovations, emotions, Twitter tweets, and trust (e.g., [66, 72, 185]). There are many circumstances where halting contagion propagation is desirable, including calming a mob [70], stopping the dissemination of leaked information [26], impeding the spread of an ideology or opinion [201], squelching a mass movement, and interrupting the communication of adversaries [7].

In this work, we study blocking of contagions such as joining a mass protest and rioting, as well as those cited above, that spread according to a popular propagation model from the sociology literature, the **progressive threshold model** [70, 97, 164, 194]. In this model, a population is treated as a network, where nodes represent people or other types of agents, and edges represent pairwise interactions among agents. Hence, nodes influence their distance-1 neighbors through their common edges. Each node can be in one of two states, 0 (respectively, 1) meaning that a node does not (does) possess a contagion. If a node possesses a contagion, we implicitly assume that it is willing to pass it on. The model allows a node to transition only from state 0 to state 1; the transition from 1 to 0 is not permitted. A node transitions from state 0 to state 1 if at least a threshold θ number of its neighbors already possess the contagion; hence, this model captures neighborhood influence. We call state 0 (1) the **unaffected** (**affected**) state.

Following [25], we investigate the contagion dynamics of two types of threshold systems. A **simple contagion** is one in which $\theta = 1$ for all nodes in the population, while a **complex contagion** is one in which $\theta > 1$ for at least one node. This delineation has large impacts on population dynamics and on algorithms for controlling contagion processes [25, 103].

7.1.2 Motivation for Our Approach

Node-based contagion blocking in the progressive threshold model consists of identifying nodes—which we call **critical** or **blocking** nodes—whose states remain frozen at 0; they do not transition to state 1. Thus, they do not assist their neighbors in transitioning to state 1. This is equivalent to removing these nodes and their incident edges from a network. The goal is to select as few nodes as possible in order to prevent as many nodes as possible from reaching state 1. The motivation for selecting *few* nodes is that convincing (or forcing) a node to remain in state 0 has a cost, and one seeks to minimize the cost of blocking.

We classify these node selection criteria as either proactive or reactive. **Proactive methods** identify critical nodes based solely on graph structure. **Reactive methods** take into account network structure and the dynamics of the contagion process. Proactive methods are attractive precisely because they do not depend on a particular dynamics model. Thus, from the practical perspective of policy development, intervention planning can be done without dynamics information. Reactive methods require more information to identify critical nodes, but it has been demonstrated that at least one reactive method is far better at blocking complex contagion spread than several state-of-the-art proactive methods on three well-known social networks [102]. Reactive methods may also require greater execution times than proactive methods to compute sets of blocking nodes.

Based on these findings, we seek a hybrid method of specifying critical nodes that has the advantage of being driven by network structure (i.e., is proactive), but incorporates contagion dynamics to increase its effectiveness (reactive). Our approach is to break a network into small clusters, with each node residing in exactly one cluster. Consistent with most working definitions of a community, we assume that the nodes within a cluster are relatively well connected and that the number of edges between clusters is relatively small. Because of the anticipated larger number of internal edges, we further assume that a contagion will propagate through a cluster relatively quickly and hence that we are unable to block it from doing so. Consequently, we seek to *contain* a contagion at cluster (community) boundaries. The edges that span clusters (i.e., **external edges**) and incident nodes are recorded. Seven such edges are shown schematically in Figure 7.1; e.g., two are $\{v_1, v_5\}$ and $\{v_3, v_8\}$. We then apply reactive blocking methods to these boundary regions of communities. We describe the reactive blocking method later (Section 7.1.5).

These ideas are summarized in Figure 7.2. The first two phases consider only graph topology; no dynamics are involved. The third phase utilizes dynamics information. Note that this approach also has the benefit that if the dynamics model changes, or multiple models must be considered, the first two phases are executed only once for all models.

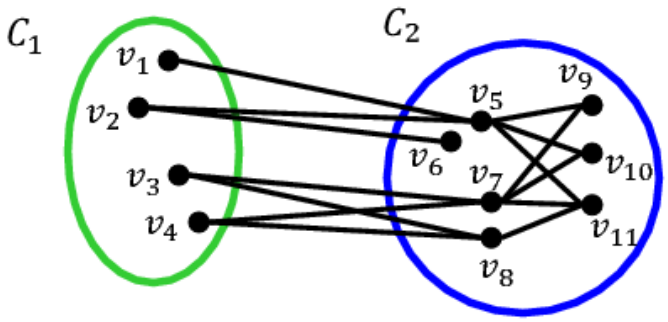


Figure 7.1: Schematic of two clusters or communities C_1 and C_2, each containing a subset of graph nodes, connected by the edges shown.

Cluster size is currently being investigated. Since each cluster that is seeded with a contagion or that a contagion reaches is effectively sacrificed, the size of a cluster should be small. The implications of "sacrificed nodes" are problem-specific and some applications may only be able to tolerate the sacrificing of relatively smaller community sizes. In epidemiological studies, for example, virus outbreaks that reach 10% to 30% of a population are routinely deemed significant, suggesting that these community sizes would not be acceptable for blocking viruses. Ideally, cluster size would be a tunable parameter.

The point of this study is to summarize initial steps taken in the implementation of these procedures. For example, computing clusters of scale-free networks is recognized as a difficult problem. In this work, we use one community detection method to segment scale-free networks, and we do not adhere to our desired guideline of limiting all cluster sizes to some specified S. That is part of future work. Using computed clusters, we complete phases 2 through 4 of Figure 7.2. We use simulation and a node-blocking scheme to perform these tasks.

- *Phase 1*: Decompose network G into clusters C_i of specified maximum size S.
- *Phase 2*: For each pair of clusters (C_i, C_j), determine external edges $e_{i,j} = \{v_i, v_j\}$ $(i \neq j)$ spanning these communities and the incident nodes $v_i \in C_i$ and $v_j \in C_j$.
- *Phase 3*: For specified (i) outbreak locations (i.e., communities C_i) and (ii) contagion dynamics model, use a scheme to block contagion propagation across these edges $e_{i,j}$ and thereby select the critical nodes.
- *Phase 4*: Evaluate the effectiveness of the critical nodes.

Figure 7.2: High-level approach to blocking contagion dynamics.

7.1.3 Contributions

Our contributions are listed below. This study was published at an AAAI workshop [175].

1. A hybrid approach to complex contagion blocking. Our approach utilizes a fast method [16] to identify communities or clusters in a graph. A critical node selection (CNS) algorithm uses these results to target node selection at cluster boundaries, obviating the need to analyze the entire graph.

2. Modular nature of approach. Our method is naturally customizable to evaluate different dynamics models. Here, we demonstrate proof-of-concept using a classic threshold model, but the the method can also be used for the independent cascade (IC) model [97]. As dynamics models change for a particular graph, clusters need not be recomputed; only a CNS algorithm need be executed.

3. Comparisons of blocking among three networks. We apply the method to three social networks from the literature, ranging up to 80000 nodes and 500000 edges. Seeding of most of the clusters generated by the community detection method [16] produces little contagion diffusion, or diffusion that can be readily blocked. However, there are some larger communities that generate widespread diffusion. We investigate some of these larger communities and show that large numbers of blocking nodes are required to halt diffusion, particularly for simple contagions. We find for these networks that the difficulty in blocking diffusion in large communities increases with increasing average degree d_{ave}, but that d_{ave} is less important when threshold is small (e.g., $\theta = 1, 2$). The ranking of networks in order of ease (or difficulty) of blocking diffusion is not straight-forward because we observe crossover behavior. For example, data for $\theta = 3$ shows much less contagion propagation in Slashdot than in Facebook for smaller values of blocking nodes. However, to halt all diffusion, Slashdot requires more blocking nodes.

Organization. The rest of the chapter is organized as follows. In Section 7.1.4, the modeling approach is described; this is used in the simulations and the CNS algorithm. Section 7.2.2 contains related work. In Section 7.1.5, the CNS algorithm is overviewed. Section 7.1.6 contains our experimental results, and conclusions comprise Section 7.2.8.

7.1.4 Model of Contagion Dynamics

First we formalize the model we use for contagion dynamics. The model is implemented in software to perform the simulations of this study. We give examples of contagion dynamics and of blocking. Finally, we formalize the problem we are trying to solve: minimizing the number of affected nodes.

We use discrete dynamical systems (e.g. [11]) to model contagion propagation on social networks; i.e., the dynamics are discrete in time and discrete in node states. Let **B** denote the Boolean domain {0,1}. A **graph dynamical system** (GDS) **S** over **B** is a triple

$\mathcal{S} = (G, \mathcal{F}, W)$, where

(a) $G(V, E)$, an undirected graph with node set V and edge set E where $n = |V|$ and $m = |E|$, represents the underlying social network on which a contagion propagates,

(b) $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$ is a collection of functions in the system, with f_i denoting the **local transition function** associated with node v_i, $1 \leq i \leq n$, and

(c) W is the **update scheme** that specifies the execution sequence of the f_i.

Each node of G has a state value from $\mathbf{B}$. Each function f_i specifies the local interaction between node v_i and its distance-1 (i.e., adjacent) neighbors in G. We use the convention that a node is not a neighbor of itself, but it is not a consideration for the model employed here. In this paper, function f_i at node v_i is a **progressive threshold function**, characterized by a non-negative integer denoted by θ_i. The function f_i is specified as follows:

(a) If the state of v_i is 0, then

(i) f_i is 1 if at least θ_i of v_i's neighbors are in state 1;

(ii) otherwise, the value of f_i is 0.

(b) If the state of v_i is 1, then f_i is 1 for all combinations of inputs.

Thus, θ_i is called the **threshold** of v_i and represents the minimum number of neighbors of v_i that must be in state 1 for v_i to change from 0 to 1.

The update scheme we use throughout this work is the **synchronous update scheme**, meaning that to compute the states of nodes at time t, all inputs to the f_i $(1 \leq i \leq n)$ are quantities at time $(t-1)$. We provide an example momentarily to make this concrete, and we note that there are other update schemes [11]. The synchronous update scheme means that the GDS is a **synchronous dynamical system** (SyDS), and we use SyDS henceforth to emphasize the synchronous update approach; i.e. $\mathbf{S}$ is an SyDS.

A **configuration** $\mathcal{C}(t)$ of an SyDS at any time is an n-vector $(s_1, s_2, \ldots, s_n)$, where $s_i \in \mathbf{B}$ is the state of v_i. A single SyDS transition from one configuration to another can be expressed by the following pseudocode, where each of the two steps is executed in parallel, but the steps themselves are executed serially.

for each node v_i **do in parallel**
(i) Compute the value of f_i. Let s_i' denote this value.
(ii) Update the state of v_i to s_i'.
end for

Critical Set Problem for a Threshold SyDS

We now state the problem that we are trying to solve, following [103].

Small Critical Set (SCS)

Instance: An SyDS **S** over **B**, where each f_i is a progressive threshold function; a set I of seed nodes; and an upper bound β (i.e., budget) on the number of critical nodes.

Requirement: A critical set B with $|B| \leq \beta$ and of all subsets of $V - I$ of size at most β, the removal of B from V leads to the smallest number of affected nodes.

Note that this problem is the same as that implicitly addressed by a host of blocking papers whether they use high degree approaches, centrality approaches, or others; i.e., the goal is to minimize the number of nodes that acquire a contagion. This problem for simple contagions is known to be NP-hard [47], and hence is also hard for complex contagions. It is shown in [103] that it is NP-hard to obtain a ρ-approximation for complex contagions. Here, we devise an approach for blocking complex contagions based on community structure.

7.1.5 Experimental Procedures

Our experimental procedures follow the approach given in Figure 7.2. In Phase 1, we use the Louvain method [16] for computing communities. The output consists of (*node ID, community ID*) pairs. In Phase 2, the graph G and the results from Phase 1 enable us to easily identify all edges that span communities, such as those depicted in Figure 7.1; e.g., edge $e_{2,6} = \{v_2, v_6\}$. In Phase 3, we use the progressive threshold model, and for this work, we take the thresholds of all nodes to be the same in a diffusion instance. This makes it easier to discern threshold effects in results. Each diffusion instance takes all nodes in one community as seed nodes. Dynamics are simulated for each seed set and for two time steps to determine the nodes in neighboring communities that contract the contagion. For example, in Figure 7.1, if all nodes of C_1 are seeded, the simulation will identify the nodes in C_2 that get affected. This gives the edges that span communities and that transmit contagion. A budget β on the number of critical nodes is specified. A node-based blocking algorithm is used to compute the critical nodes from among all nodes affected at time 1. The diffusion instance is then rerun, but now the critical nodes are incorporated and simulations are run until a fixed point is reached. The final spread fraction; i.e., the total fraction of affected nodes, is then computed. In Phase 4, we run all diffusion instances without critical nodes until a fixed point is reached, so that we can evaluate the effectiveness of blocking nodes by comparing spread fractions with and without critical nodes.

The covering-based blocking algorithm in [103] is used to compute critical nodes. Here, we provide an example of how the algorithm works using Figure 7.1. Assume community C_1 is seeded and all nodes v_i have $\theta_i = 2$. At $t = 1$ of a simulation, contagion spreads to a neighboring community, C_2, and v_5, v_7, and v_8 are affected because each has at least two neighbors in C_1; v_6 is not affected. At $t = 2$, v_9, v_{10}, and v_{11} are affected because each has at least two neighbors in state 1. The CNS algorithm uses these data as inputs. If $\beta \geq 3$, the three nodes affected at $t = 1$ are identified as critical nodes and the covering algorithm terminates.

Consequently, to explore the CNS algorithm further, let us assume instead that $\beta = 2$. The nodes affected at $t = 1$ are evaluated as to how they affect nodes at $t = 2$. The node that contributes to the greatest number of nodes affected at $t = 2$ is selected as a critical node; ties are broken arbitrarily. Both v_5 and v_7 contribute to three nodes (v_9, v_{10}, and v_{11}). Assume the algorithm selects v_5 (at random). Now, with v_5 as critical, nodes v_9 and v_{10} cannot get affected (they now only have v_7 contributing to their possible transitions, so their thresholds are not met and they will not transition to state 1). Now v_7 and v_8 contribute to the one remaining affected node (v_{11}). Again, since there is a tie, one node is selected at random, say v_8. The critical set $B = \{v_5, v_8\}$, with $|B| \leq \beta$ as required, will block signifcant contagion propagation for all $t > 1$. The final spread size is the number of nodes in C_1, which are seed nodes, plus one, for v_7; v_7 gets affected and is not a blocking node. Note in this case that v_7, although contributing to the maximum number of affected nodes at $t = 2$, is not chosen as critical. Obviously, there are cases, particularly when θ and β are small, where the critical set will be insufficient to thwart all diffusion.

We briefly address the issue of other models beyond the deterministic progressive threshold model studied here. The IC model [97] is a variant of a $\theta = 1$ model. Therefore, one could run deterministic $\theta = 1$ dynamics to provide the requisite input data to the CNS algorithm. With the computed blocking nodes, one could then simulate (stochastic) IC dynamics to assess the efficacy of the blocking nodes in thwarting IC contagion propagation. This is an example of how one strategy for determining blocking nodes can be used with multiple dynamics models.

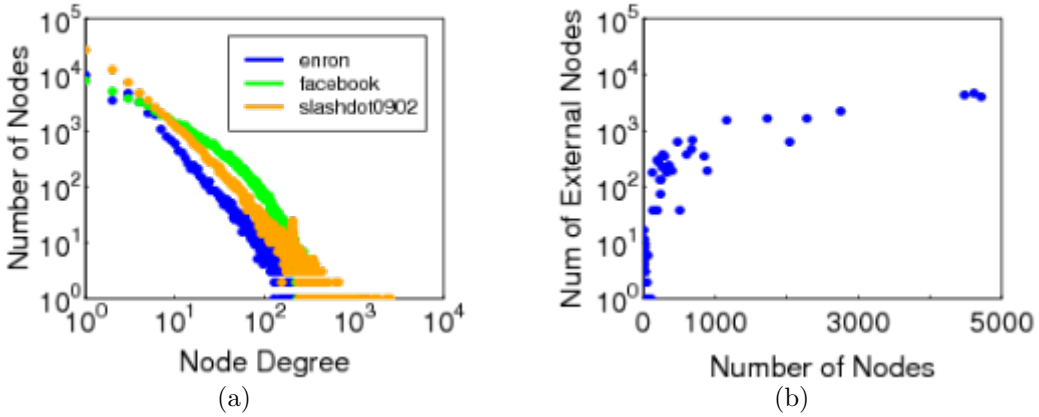
7.1.6 Experimental Results

Table 7.1 lists selected characteristics of the giant components of networks studied. Figure 7.3a provides degree distributions for the three networks. Table 7.2 lists the parameter values used with the procedures described in Section 7.1.5. Threshold values as large as 10 are easily justified, as thresholds of this magnitude were inferred from adolescent smoking data (where peer influence is known to play a large role in smoking initiation in teenagers) [75,82]. It seems reasonable that in order to join a strike or demonstration against a government, where one risks losing her job or imprisonment, thresholds may be even greater. Values of β were chosen to investigate a range of behaviors.

Table 7.1: Networks used in experiments; n and m are numbers of nodes and edges, d_{ave} is average degree, and n_c is the number of communities determined using [16]; all correspond to the giant component.

Networks	n	m	d_{ave}	n_c
Enron [94]	33696	180811	10.7	183
Facebook [189]	63392	816886	25.8	56
Slashdot0902 [94]	82168	504230	12.3	376

Based on the example in Section 7.1.5, it is useful to know, for a given community C_i, how



(a) (b)

Figure 7.3: (a) Degree distributions for the networks of this study. (b) For each of the 183 communities in the Enron network, the number of nodes in other communites adjacent to the extenal edges of a community (i.e., number of external nodes) as a function of number of nodes in a community.

Table 7.2: Parameters and values used in experiments.

Networks	**Threshold Values**	**Numbers of Critical Nodes**
Enron, Facebook, Slashdot0902	1, 2, 3, 4, 5, 10	0, 10, 10^2, 10^3, 10^4

many nodes in other communities are adjacent to all external edges of C_i. These are the **external nodes**. (In Figure 7.1, C_1 has 4 external nodes, v_5 through v_8.) These data for the Enron network are given in Figure 7.3b, where the abscissa is the number of nodes in a community and the ordinate is the number of external nodes.

We will investigate particular communities in each network. These are listed in Table 7.3, where n' is the number of nodes in the community (and % of nodes is also given), and n_{en} and n_{ee} are the numbers of external nodes and edges.

Figure 7.4 contains four plots of Enron data with the same axes and legend, each corresponding to a different number of critical nodes. In Figure 7.4a, $\beta = 0$. Consider the threshold-1 (abbreviated "thr=1") curve. For this θ and β, 183 diffusion instances are simulated, one diffusion instance for each community. Therefore, each curve in each plot is composed of 183 data points. The diffusion instance for community C_i means that all nodes for that community are seed nodes, and all other nodes are initially in state 0. The final spread fraction is computed for each diffusion instance, and then plotted in increasing numerical order. The green curve for threshold-1 is horizontal at an ordinate of 1. This is because we are using only the giant component of the Enron network and thus the graph is connected. So even one seed node will result in a spread fraction of 1.0 when there are no blocking nodes; i.e., the contagion will propagate through the network. For threshold-2, the curve has an ordinate of zero until an abscissa value of about 0.73. At that point, the spread fraction increases to about 0.6, and remains relatively constant thereafter. This means that for $\theta = 2$, 73% or

Table 7.3: Particular communities in the networks that are seeded and evaluated.

Networks	**Comm ID**	n' (%)	n_{en}	n_{ee}
Enron	42	4715 (14%)	4066	21145
Enron	0	2042 (6%)	639	902
Facebook	6	17326 (27%)	17138	63775
Facebook	16	3663 (6%)	7704	13755
Slashdot0902	3	22269 (27%)	21526	113207
Slashdot0902	40	4954 (6%)	10350	23684

about 133 of the 183 communities generate very small spread sizes when seeded; about 27% of communities produce significant contagion propagation. Overall, this plot indicates that a smaller fraction of communities (e.g., < 0.3) generates appreciable diffusion. The number of communities that can spread contagion decreases as θ increases. In successive plots, where β increases to 10, 100, and 1000, there remains communities for all thresholds investigated that produce appreciable spreading.

That 1000 critical nodes does not block all diffusion from a community is not surprising in light of other work. It has been shown experimentally [103] that seeding just 10 to 20 well-connected nodes in realistic scale-free networks can require 1000 or more critical nodes to completely halt all diffusion; i.e., prevent all unaffected nodes from becoming affected. Here, the larger communities contain thousands of nodes that are seeds.

Data for Facebook and Slashdot are qualitatively similar to those in Figure 7.4. The takeaway here is that contagions starting in the majority of communities can be readily blocked. However, diffusion starting from larger communities are much harder to block. Our procedures must be refined to deal with these larger communities, which we now further investigate.

We examine two Enron communities in more detail that have very large and intermediate connectivities to other communities. Community 42 (C_{42}) has the greatest number of nodes (4715), the second largest number of external edges (21145), and third largest number of external nodes (4066). Community C_0 is an intermediate sized community; see Table 7.3. These communities were selected based on the data in Figure 7.3b. From this point forward, the communities that we investigate are provided in Table 7.3.

Figure 7.5a depicts the spread fraction as a function of time for simulations where the nodes of C_{42} are the seed nodes; there are no critical nodes. Increasing the threshold drives down the spread size. This is a mechanism of retarding contagion propagation in itself. For example, if an external agent can take action, or threaten to take action, with significant adverse consequences for the nodes, then it may cause nodes to increase their thresholds

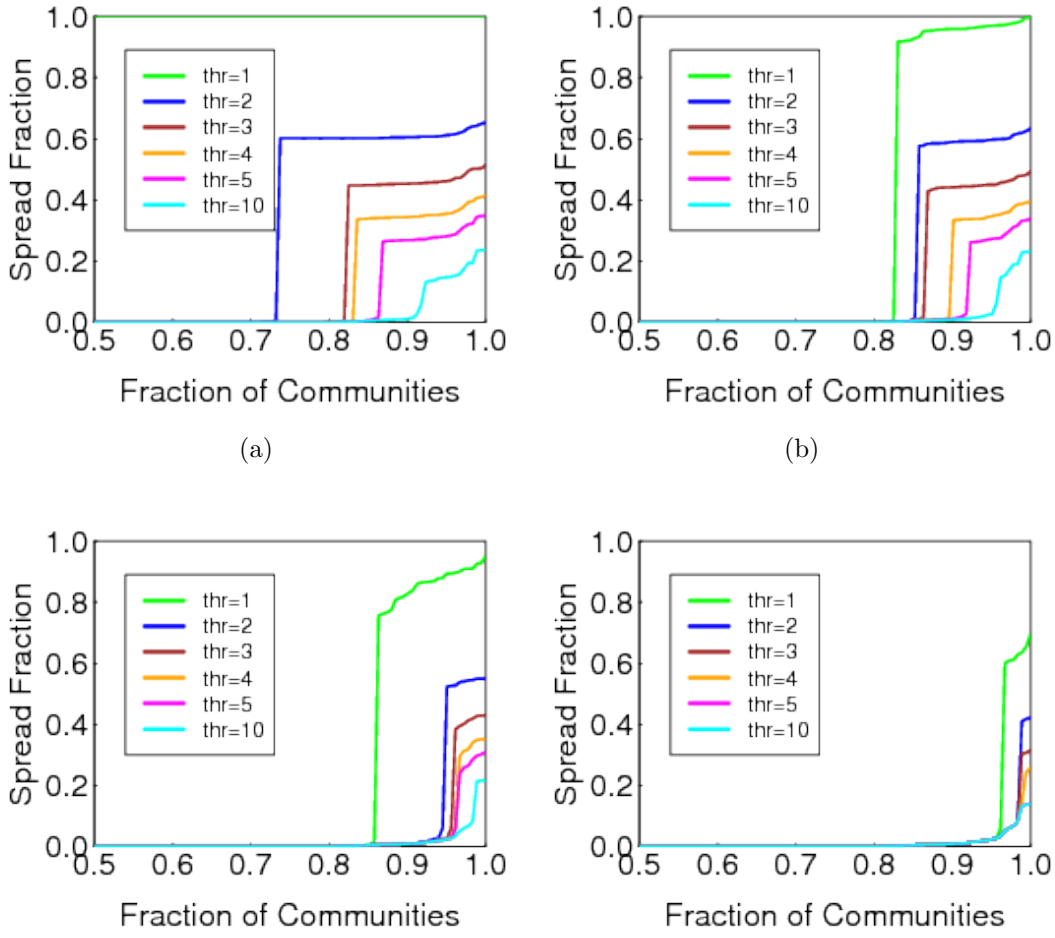


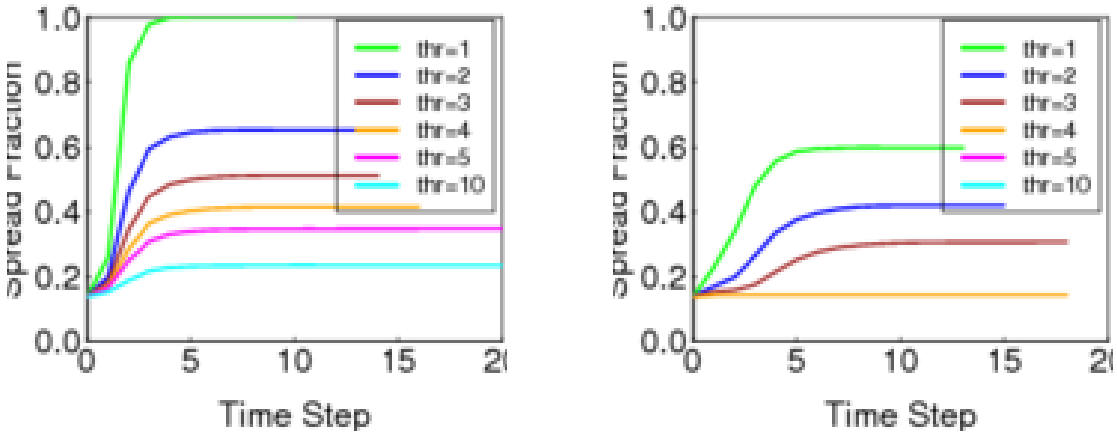
Figure 7.4: Final spread fraction at the end of a simulation when all nodes of a community are seeded. For the Enron network, there are 183 communities. Spread sizes are arranged in increasing numerical order for each curve in each plot. Each curve corresponds to a homogeneous threshold used for all nodes. The numbers of critical nodes in the plots are: (a) zero, (b) 10, (c) 100, and (d) 1000.

(i.e., it may require greater influence to convince a node to acquire a contagion). Such an example could be governments of countries with a history of imprisoning dissidents.

Figure 7.5b shows data for the same conditions, except that now $\beta = 1000$. All contagion diffusion is stymied for $\theta \geq 4$ (the curves for $\theta = 4, 5$ lay underneath the light blue curve for $\theta = 10$). The horizontal curve for $\theta = 10$ corresponds to the size of C_{42}.

Figure 7.6 depicts analogous data for the Facebook network, for Community 6 in Table 7.3. With an average degree 2.5 times that of Enron, it more readily propagates complex contagions.

Figure 7.7a shows the final spread fraction for C_{42} of the Enron network as a function of β

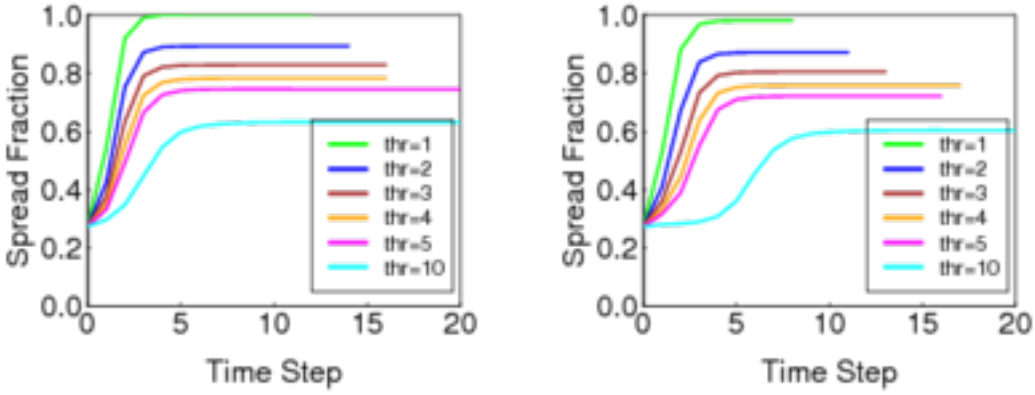


(a) (b)

Figure 7.5: Spread size as a function of time in the Enron network when all nodes of the largest community C_{42} are seeded: (a) no critical nodes and (b) $\beta = 1000$ critical nodes are used. Each curve corresponds to all nodes possessing a single threshold.

and θ. It is clear that $\beta = 1000$ will halt $\theta = 4, 5, 10$ complex contagions, but that more than 1000 critical nodes are required to halt contagions with lesser thresholds. Figure 7.7b shows analogous data for C_0, a smaller community than C_{42}. Now as few as 100 critical nodes will halt complex contagions originating within C_0, but simple contagions require greater numbers of critical nodes. These data illustrate that the numbers of blocking nodes required to halt diffusion decrease with increasing θ, and since assigning or converting nodes to blocking nodes has a cost, taking into account complex contagion diffusion can reduce the cost of blocking it, and can more effectively assign as blocking the β critical nodes.

In Figure 7.7 and in subsequent plots, n_{en} provides an upper bound on β. This bound is realized; i.e., $\beta = n_{en}$, when all external nodes are set critical. In this case, no node beyond



(a) (b)

Figure 7.6: Spread size as a function of time in the Facebook network when all nodes of the largest community C_6 are seeded: (a) no critical nodes and (b) $\beta = 1000$ critical nodes are used. Each curve corresponds to all nodes possessing a single threshold.

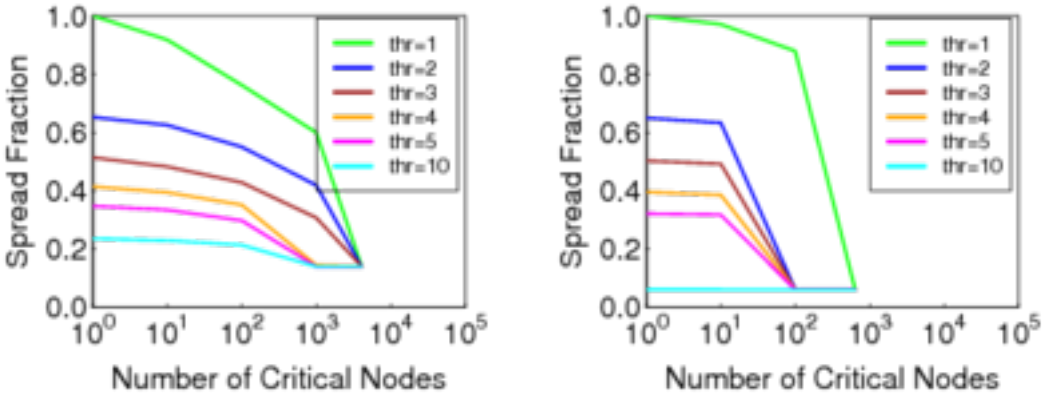


Figure 7.7: Final spread size as a function of the number of critical nodes and homogeneous node threshold in the Enron network: (a)seeding of the largest community C_{42}, and (b) seeding of an intermediate-sized community C_0.

the seed community can become affected.

Figure 7.8 provides analogous data for Facebook. Clearly, increasing average degree pushes the curves up and to the right; i.e., for a given θ and β, the spread fractions are greater. With Figure 7.8b, we can compare the numbers of critical nodes required to halt all diffusion for $\theta = 1$ and $\theta = 10$. The data show that a factor of 77 more critical nodes are required to halt simple contagions than $\theta = 10$ contagions (at most 100 critical nodes will halt $\theta = 10$ diffusion, but $\beta = n_{en} = 7704$ nodes are required to halt $\theta = 1$ diffusion).

Finally, Figure 7.9 provides data for thresholds 1, 3, and 10, for Enron (in green), Facebook (in blue), and Slashdot (in red), for both large and intermediate community sizes. Generally, Facebook is the most difficult to block and Enron the easiest. However, $\theta = 3$ data in Figure 7.9b shows a crossover of Facebook and Slashdot data. Slashdot produces smaller

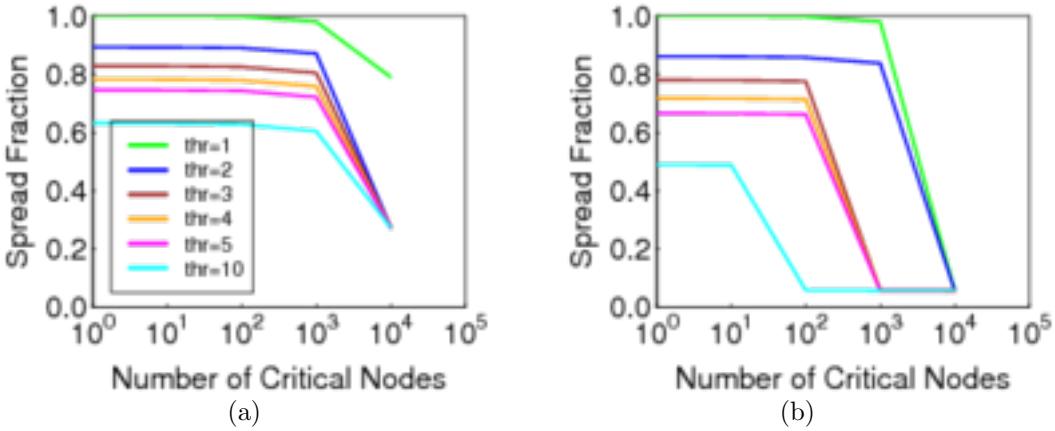
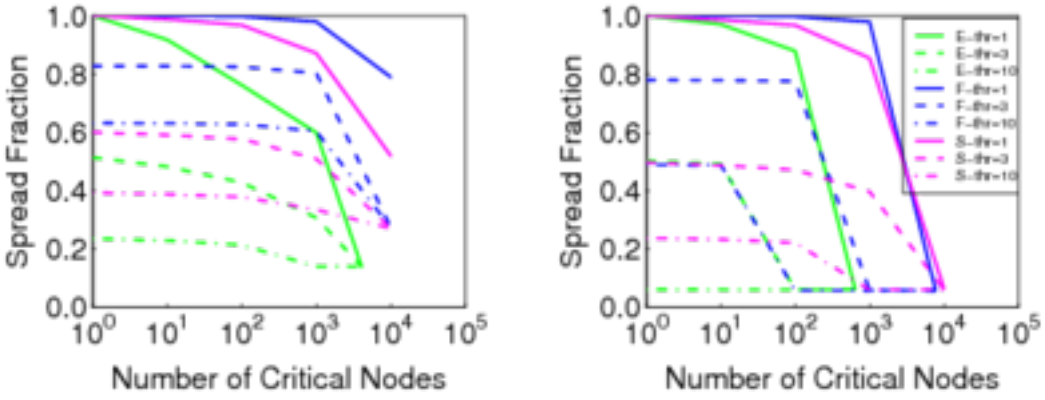


Figure 7.8: Final spread size as a function of the number of critical nodes and homogeneous node threshold in the Facebook network: (a) seeding of the largest community C_6, and (b) seeding of an intermediate-sized community C_{16}.



(a) (b)

Figure 7.9: Final spread size as a function of the number of critical nodes and homogeneous node threshold in three networks: (a) seeding of the largest communities, and (b) seeding of intermediate-sized communities whose sizes are 6% of nodes.

spread fractions for $\beta \leq 100$, but requires more critical nodes to block all diffusion.

7.1.7 Conclusion and Future Work

In this work we propose a method of blocking simple and complex contagions using a combination of proactive and reactive methods. This combination first exploits graph structure to segment a network and then targets cluster boundaries for computing critical nodes for a particular dynamics model. We show differences in blocking behavior among three social networks.

There are many avenues for future work. Of primary interest is to augment the current community detection algorithm to obtain smaller clusters, which should enable blocking of contagions with smaller numbers of critical nodes. We will also apply the method to larger graphs.

7.2 Edge-based Blocking

7.2.1 Background and Motivation

Controlling contagions, such as false rumors, leaked information, or social unrest, by limiting their spread is a well motivated and important problem [119, 186]. The majority of early work (e.g., [4]) used node removal techniques to block contagion transmission in network

representations of populations. Deleting nodes from a network removes pathways through which a contagion can travel, thus inhibiting its diffusion. More recently (e.g., [172]), edge removal methods have been studied, and edge-based blocking is the focus of this work.

There are many situations in which edge removal is a more pragmatic alternative than node removal. For example, in Twitter, person A_1 may stop following person A_2, thereby removing that tie of interaction; it is most often unrealistic to remove A_1 from Twitter. Similarly, political regimes may have the resources to remove or isolate individuals [171], but this approach has costs [165] and may not be politically viable. Recent work [86] describes how two opposing countries, C_1 and C_2, intervene to influence leaders of a third country C_3. This is, in effect, an attempt by each of C_1 and C_2 to sever the opposing country's ties to C_3. From network considerations alone, edge removal is a more surgical approach than node removal: removing one node v results in deleting $d(v)$ edges, where $d(v)$ is the degree of v.

7.2.2 Related Work

Table 7.4 provides a perspective on node- and edge-based classes of methods for stymieing contagion propagation. Each class provides schemes based on graph structure, and on dynamics models. However, to our knowledge, no work exists on edge blocking techniques that use both dynamics models and information on contagion outbreak locations to block diffusion. This work fills that void.

Table 7.4: Sample studies on blocking contagions, showing how our work fills a void in edge-based blocking methods.

Basis of Blocking Method	**Blocking Nodes**	**Blocking Edges**
Graph Structure	[4]	[203]
Dynamics Models	[148]	[172]
Dynamics and Initial Conditions	[103]	Our work.

Most edge-blocking studies focus on simple contagions, using models that incorporate, for example, independent cascade (IC) and variants of susceptible-infected-recovered (SIR) dynamics [148, 172, 203]. (A node [person] may contract a **simple contagion** through one interaction with a person who already possesses it; this is a 1-threshold model. A **complex contagion** requires interactions with at least two such people.) An exception is a study [99] of the linear threshold (LT) model; however, that work contains no evaluation of the blocking method in terms of how its deleted edges reduce contagion spread. Here, we focus on threshold-based models, which are well-motivated in the social science literature [25, 70, 194]. Evidence for the existence of progressive complex contagions—which is the type we model here—continues to mount; e.g., through data mining of people's behavior and evaluation of social movements [66, 179]. (A **progressive** model [97] means that once a node contracts a contagion, it remains with the node.) We study both simple and complex contagions, and

show interesting differences between the two. We cite other references in relation to our results, in subsequent sections.

7.2.3 Contributions

Below, we summarize our main contributions. For more theoretical results, please refer to [104].

Heuristics. In view of the non-approximability result for the basic edge-based blocking problem, we develop a practical edge-covering heuristic to block both simple and complex contagions for directed, weighted and unweighted graphs. We also introduce a straightforward heuristic for weighted graphs. To evaluate our main heuristic, we perform computational experiments of contagion propagation on social networks from the literature that are at least five times greater in terms of numbers of nodes and an order of magnitude greater in numbers of edges than those used in previous studies. These networks are a detailed human contact network of Montgomery County, Virginia and two Facebook networks. We compute the numbers of nodes that contract a contagion (we call this the **spread size**) with and without blocking edges. We provide what we believe are the first comparisons of state-of-the-art edge-based blocking methods by comparing our results with those of other methods. In total, we evaluate 12 combinations of networks and blocking heuristics, which to our knowledge is the biggest study of its kind. Results show that our edge-covering method is more effective in blocking simple and complex contagions, for both unweighted and weighted graphs.

Experiments. We provide a small set of experimental results to understand the behavior and limitations of our edge covering heuristic. For example, we demonstrate how increasing the number of seed nodes (i.e., nodes initially possessing a contagion) can increase the probability of cascade (i.e., widespread diffusion) in the presence of blocking edges. We also illustrate a somewhat surprising result of network structure effects: it can be much more difficult to block contagions in a network with a far less average degree and a far less average clustering coefficient.

Generalizability. Finally, the models we employ in this study are deterministic. However, all theoretical and experimental results are also *directly* applicable to stochastic progressive threshold models, where a node i contracts a contagion with some probability $p_i > 0$ once its threshold θ_i is met.

7.2.4 Weighted Edge Blocking Problem

In formulating the problems considered in this paper, we use terminology from the context of information propagation in social networks. The problem statements can be readily extended to other contexts. As mentioned earlier, we say that a node is **affected** if its final state is 1; otherwise, the node is **unaffected**. We provide two formulations of the problem of blocking

a contagion through edge removals.

In this formulation, we assume that each edge e has a nonnegative cost c_e. If a set B of edges is chosen as the blocking set, then the total cost c_B of the blocking set is given by $c_B = \sum_{e \in B} c_e$. This leads to the following problem where we seek to minimize the number of new affected nodes subject to a budget on the blocking cost.

Small Weighted Critical Edge Set (SWCES)

Instance: A social network represented by the SyDS $\mathcal{S} = (G(V, E), \mathcal{F})$ over , with each function $f \in \mathcal{F}$ being a threshold function; the set I of nodes which are initially in state 1 (i.e., the set of **seed nodes**); an upper bound β on the cost of the blocking set.

Requirement: A critical set B of edges with $c_B \leq \beta$ such that among all edge subsets with cost at most β, the removal of B leads to the smallest number of new affected nodes.

The SWCES problem was first formulated in [47] for the case of simple contagions, where each node computes a 1-threshold function. They showed that the problem is NP-hard and presented a bicriteria approximation that violates the budget by a constant factor and approximates the number of new affected nodes by another constant factor. We consider the problem for complex contagions where one or more nodes may have threshold values of 2 or more. We show that, if the budget constraint cannot be violated, the problem cannot be approximated to within any factor $\rho \geq 1$, unless $\mathbf{P} = \mathbf{NP}$.

7.2.5 Complexity of Weighted Edge Blocking

We will show the non-approximability result for the weighted edge blocking problem SWCES for complex contagions. Throughout this section, we use the terms "blocking set" and "critical set" synonymously.

As mentioned earlier, the problem was shown to be **NP**-hard for simple contagions in [47]. However, that proof relies on the fact that any blocking set of edges for simple contagions must disconnect the graph. It is not difficult to see that the condition does not hold for complex contagions. Therefore, the proof in [47] cannot be directly extended to the complex contagion case.

Assuming that the bound β on the cost of the critical set cannot be violated, for any $\rho \geq 1$, there is no polynomial time ρ-approximation algorithm for the SWCES problem for complex contagions, unless $\mathbf{P} = \mathbf{NP}$.

Proof Sketch: Suppose $\mathcal{A}$ is a ρ-approximation algorithm for the SWCES problem for complex contagions for some $\rho \geq 1$. Without loss of generality, we can assume that ρ is a positive integer. We will show that $\mathcal{A}$ can be used to efficiently solve **3SAT**, which is known to be **NP**-hard [127].

Given an instance I of 3SAT, we construct an instance of SWCES as follows. We first

describe how the the node and edge sets of the underlying graph $G(V, E)$ are constructed.

Description of the node set V:

1. V has two special nodes a and b, which are the seed nodes; the initial states of all other nodes are 0.
2. For each variable $x_i \in X$,
   (a) there are two nodes p_i and q_i, corresponding to the literals x_i and $\overline{x_i}$ respectively; and
   (b) there is a set R_i of ρn nodes.
3. For each clause $C_j \in C$,
   (a) there is a node w_j in V; and
   (b) there is a set T_j of ρn nodes.

Thus, V has a total of $2 + n(2 + \rho n) + m(1 + \rho n) = \rho n^2 + \rho mn + 2n + m + 2$ nodes.

Description of the edge set E:

1. For each i, $1 \leq i \leq n$, the four edges $\{a, p_i\}$, $\{a, q_i\}$, $\{b, p_i\}$ and $\{b, q_i\}$ are in E. The weight of each of these $4n$ edges is 1.
2. For each i, $1 \leq i \leq n$, nodes p_i and q_i are joined to all the ρn nodes in R_i. The weight of all these edges is $n + 1$.
3. Consider each clause C_j ($1 \leq j \leq m$). Suppose C_j contains the three literals l_1, l_2 and l_3. Then node w_j (corresponding to C_j) is joined to the nodes corresponding to the three literals of C_j; further, w_j is also joined to all the nodes of T_j. The weights of all the edges introduced in this step is also $n + 1$.

The threshold for each of the nodes w_1, w_2, ..., w_m is 3. The threshold for each node in $\bigcup_{j=1}^{m} T_j$ is 1. The thresholds for all other nodes are 2. (Thus, the system models a complex contagion.) The budget on the cost of the blocking set is chosen as n.

This completes the construction. It is easy to see that the construction can be carried out in polynomial time.

Using this construction it is possible to prove that Algorithm $\mathcal{A}$ produces a blocking set of cost at most n leading to at most ρn new affected nodes if and only if there is a solution to the given instance of 3SAT. The first step is to show that when there is a satisfying assignment for the 3SAT instance, there is a blocking set of cost n, which ensures that the number of new affected nodes is n. Second, we show that when the 3SAT instance is not satisfiable, then regardless of which blocking set B of cost at most n is chosen, the number

of new affected nodes exceeds ρn. The detailed proof of both steps is available in the longer version [104].

Therefore, by running $\mathcal{A}$ on the resulting instance of SWCES, and checking the number of new affected nodes, we can decide whether or not the given instance of 3SAT is satisfiable. Since $\mathcal{A}$ runs in polynomial time, the theorem follows. □

The algorithm for the problem shown in Figure 7.10 can be implemented to run in $O(|V| + |E|)$ time. □

The proof of the above proposition is available in the longer version [104].

7.2.6 Heuristics

In the experimental investigations of contagion blocking, the goal is to (approximately) solve the SWCES Problem of Section 7.2.4. As described above, this problem is formally hard, and no approximation algorithm possesses a non-trivial performance guarantee unless $\mathbf{P} = \mathbf{NP}$. Therefore, we formulate a heuristic—the edge-covering heuristic (ECH)— to solve the problem, and its implementation is used in Section 7.2.7.

The heuristic consists of two parts. In the first part, the dynamics are simulated on the network, up to and including time T, according to the model of Section 7.1.4. The times at which nodes become affected are recorded. Let S_i be the set of all nodes that are affected at time i, with $S_0 = I$. These data, along with edge costs and blocking budget β, are used to compute the blocking set B, subject to the cost constraint $c_B \leq \beta$, as described next.

The algorithm marches through the simulation time steps, and at each time i performs the following computations in seeking a solution. First, it determines whether the total cost of all edges used to transport contagion to affected nodes $v_j \in S_i$ is less than β. If so, then these edges constitute B. Otherwise, the necessary number of least cost edges required to save each affected node is computed. This total cost is a node property, and nodes are arranged in increasing order of these costs. These nodes are saved, in order, by removing the identified minimum cost edges. At each time, either all nodes in S_i are saved, or they are not. The former eventuality is a solution. If the latter holds, the algorithm moves to the next time $i + 1$ and repeats the computations. If no solution is found over all i, the solution at the time with the least number of remaining affected nodes is chosen for B. The algorithm is presented in Figure 7.10.

Input: A SyDS $\mathcal{S}$ including the underlying graph $G(V, E)$ and a threshold θ_j for each node $v_j \in V$; a set $I \subset V$ of seed nodes; a cost w_e for each edge $e \in E$; and a cost budget β for blocking edges.

Output: A set $B \subseteq E$ of edges such that the total cost c_B of B is $c_B \leq \beta$, and whose removal from E leads to a small number of affected nodes.

Steps of the Algorithm:

1. Simulate $\mathcal{S}$ for T time steps and store the computed sets S_i of newly affected nodes at time i; $1 \leq i \leq T$.
2. **For** $i = 1$ **to** $T - 1$ **do**
   (a) **For** each $v_j \in S_i$, compute R_{ji}, the set of all edges used to transmit contagion to v_j. Let α_j be the total cost of all edges in R_{ji}.
   (b) **If** $\sum_{v_j \in S_i} \alpha_j \leq \beta$, **then** set $B = \bigcup_{v_j \in S_i} R_{ji}$ and **return** the solution B.
   (c) **For** each $v_j \in S_i$ **do** the following. Compute $\eta_j = \max\{|R_{ji}| - \theta_j + 1, 0\}$, the number of edges incident on v_j to block. Order the edges $e \in R_{ji}$ in increasing cost w_e order. Compute the cost $c_j = \sum_{k=1}^{\eta_j} w_k$ of the η_j least-cost edges. Let C_j be the *ordered* set of the η_j low cost edges.
   (d) Order the $C_j \subset E$, for all $v_j \in S_i$, in increasing cost c_j order. Break ties by giving priority to greater degree nodes.
   (e) Set $B = \emptyset$; set $c_B = 0$; set the solution flag $s = 1$; set $l_s = 0$ (the number of saved nodes).
   (f) **For** each C_j in increasing cost order **do** the following. **If** $c_j + c_B \leq \beta$, set $c_B = c_B + c_j$; set $B = B \bigcup C_j$; and increment l_s. **Else** continue to add the next least cost edge $e \in C_j$ to B and add w_e to c_B until $c_B > \beta$; reset the solution flag $s = 0$; break out of the **for** loop of (2f).
   (g) **If** $s == 1$ **then return** B. **Else** set the remainder $h_i = |S_i| - l_s$; set $H_i = B$.
3. Determine the earliest time i at which h_i is minimum; set $B = H_i$; and **return** B.

Figure 7.10: Details of the edge-covering heuristic (ECH) for the SWCES problem.

7.2.7 Experimental Results

Networks

Giant components of the networks of this study are summarized in Table 7.5. These were chosen for their ranges in traits; e.g., there is a 6× variation in average degree and an order of magnitude difference in numbers of edges. They are also bigger by at least 5× in numbers n of nodes, and more so in the numbers m of edges, compared to networks evaluated in other edge blocking studies [98,99,172,203]. All networks are taken as undirected to foster greater diffusion and hence to more stringently evaluate the blocking methods. Degree distributions are given in Figure 7.11.

Network MONT-VA is a social contact network for Montgomery County, Virginia, which is constructed from detailed data (including Census data, activity surveys, and geo-spatial data) and models [10]. Individual agent movements are computed, from which are generated pairwise interactions at particular times and locations. Edge weights are durations of pairwise interactions, in seconds.

FB-1 and FB-2 are two networks constructed using the Facebook data made available by [190]. FB-1 is a friendship network of a subset of Facebook users, i.e., there is an edge between two users if they are friends. This network is unweighted and undirected. FB-2 is an interaction network that shows which user pair from FB-1 interacts via wall posts (during the period Sep. 26, 2006 to Jan 22, 2009). It is also an undirected network but has the count of wall posts between user pairs as edge weights.

Table 7.5: Network characteristics.

Network	n	m	d_{ave}	C_{ave}	**k-core for seeds**
MONT-VA	77,528	1,967,714	50.8	0.395	20
FB-1	63,392	816,886	25.8	0.222	20
FB-2	43,953	182,384	8.30	0.111	10

Experimental Procedures

Table 7.6 contains the parameters of our test procedures. We focus on small numbers n_s of seed nodes, small thresholds θ, and small values β of blocking edges because, as will be demonstrated, unit changes in values can cause significant changes in results. Numbers of blocking edges vary with networks, and hence we give a range.

Following [103], we take seed nodes from a high k-core subgraph of each network (which is the subgraph in which all nodes have degree at least k) such that each seed set induces a connected subgraph on the original network. An anchor seed node is first selected at

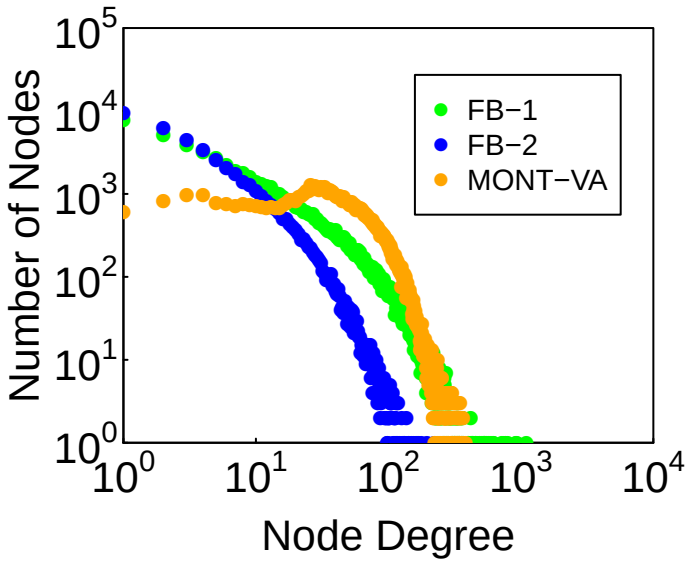


Figure 7.11: Degree distributions of the three networks.

Table 7.6: Experimental parameters.

Networks	n_s	θ	β
MONT-VA, FB-1, FB-2	2, 3, 5, 10, 20	1, 2, 3, 5	$\{0, \ldots, 1500\}$

random, and the seed set is randomly grown by adding nodes adjacent to current seeds. Where possible, we select seeds from the 20-core to foster greater diffusion and thereby tax the heuristics. However, FB-2 did not have a sufficiently sized 20-core so we used the largest core possible (e.g., of about 3000 nodes) so that seed sets would have minimal or no overlap. For each value of n_s, we produced 100 different seed sets. We specify uniform (i.e., the same) thresholds for all nodes in a simulation to make it easier to reason about results; heterogeneous thresholds can be readily accommodated.

For a given (n_s, θ) pair, 100 diffusion instances were simulated; one for each seed set. All diffusion is taken to be deterministic. However, stochasticity enters through the selection of seed node sets, as described above. Simulation results for $\beta = 0$ are fed into our ECH blocking algorithm to compute blocking edges. We developed an MPI worker-pool-based implementation such that 100 sets of blocking edges are typically computed in about one minute (many take about 20 seconds) using 11 worker processes on Dell 6100 12-core compute nodes with a Qlogic QDR Infiniband interconnect. Simulations are then repeated, but now with the inclusion of blocking edges. We compare final spread sizes with and without blocking edges to determine their effectiveness in thwarting contagion diffusion.

Note that for $\theta > 1$, with no blocking edges, it is not a certainty that the contagion will reach all nodes, or that any diffusion will take place at all. Consider $n_s = \theta = 2$. It is possible, and indeed happens, that there is no node in the set $V - I$ that is adjacent to both seed nodes. In

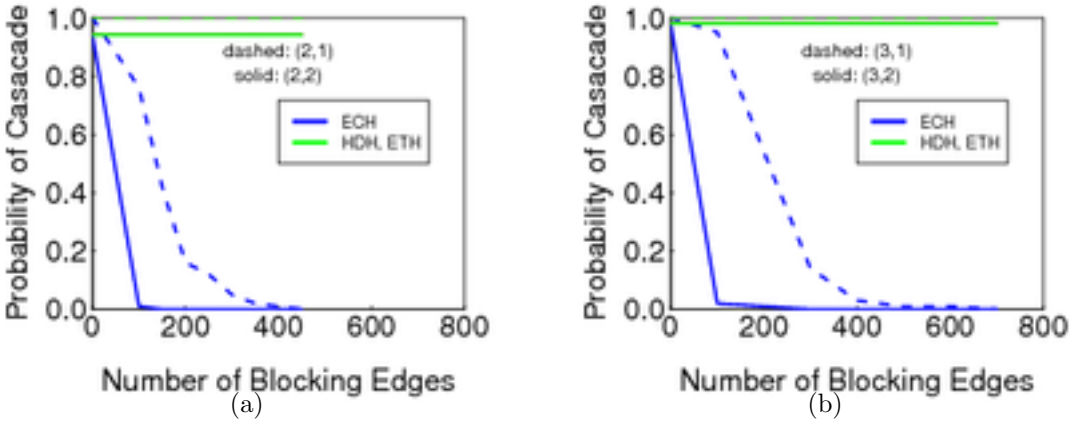


Figure 7.12: Data for the unweighted MONT-VA network, showing probability of cascade p_c versus β, for the (n_s, θ) pairs in the plots: (a) $n_s = 2$; (b) $n_s = 3$. The threshold-1 data for ETH and HDH are at $p_c = 1.0$.

this situation, the contagion will not propagate. Hence, there is a probability of widespread diffusion (i.e., of a cascade) p_c that is predicated on the dynamics model, network, and initial conditions. Effective blocking edges will significantly decrease the probability of a cascade compared to $\beta = 0$ results; i.e., the goal of blocking edges is to make $p_c \to 0$.

Finally, we note that our results presented in Sections 7.2.7 and 7.2.7 apply to the case where a node in state 0 transitions with probability < 1 when a threshold-number of neighbors are affected. To see this, simply set the probability of transition to 1.0, and run this set of procedures. A successful set of blocking nodes will also stop stochastic diffusion.

Comparison With Other Heuristics

We first briefly describe the other heuristics that we study along with our own. The epidemic threshold-based heuristic (ETH) [172] computes the left and right eigenvectors corresponding to the left and right maximum eigenvalues of the network adjacency matrix. Note that the adjacency matrix may be unweighted (i.e., contain appropriate entries of 1.0), or may be weighted. This yields a left eigenvector u and a right eigenvector w, and the weight of an edge $\{i, j\}$ is assigned the product $u(i) \cdot w(j)$. Edges with the greatest "eigenproduct" are selected as blocking edges. The high degree heuristic (HDH) [203] computes the edge weight as the product of the degrees of the two incident nodes. Greatest degree-product edges are selected for blocking. Since HDH is not suited for graphs with edge weights, we use instead for these networks a greedy algorithm that selects the blocking set B of maximum cardinality such that $c_B \leq \beta$.

Figure 7.12 provides data for the unweighted MONT-VA network. Data for ECH are provided in blue while the those for ETH and HDH are shown in green. The left plot is for

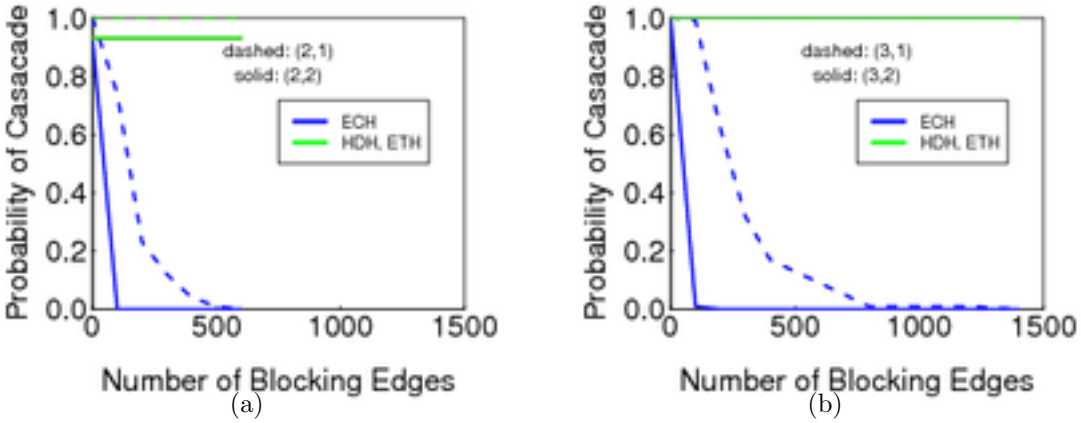


Figure 7.13: Data for the unweighted FB-1 network, showing probability of cascade versus β, for the (n_s, θ) pairs in the plots: (a) $n_s = 2$; (b) $n_s = 3$. The ETH and HDH produce $p_c = 1.0$ for all conditions for $n_s = 3$, and for threshold-1 for $n_s = 2$.

$n_s = 2$ while the right is for $n_s = 3$, and the dashed (solid) curves are for threshold 1 (2). The ordinate is computed by noting how many of the 100 diffusion instances result in wide-spread diffusion. (We will demonstrate later that it is clear whether or not *widespread diffusion* occurs.) We use small seed sets to represent isolated outbreaks. Clearly, the ECH is more effective at blocking diffusion than ETH and HDH, since it produces lesser p_c values (lower curves are better). For example, for $(n_s, \theta) = (2, 2)$ in the left plot, $\beta = 200$ with the ECH will completely block all diffusion (i.e., $p_c = 0$), while ETH and HDH generate $p_c = 0.94$. ETH and HDH do not reduce the probability of cascade from that for the $\beta = 0$ case, and hence their data curves coincide.

Analogous data for the unweighted FB-1 network are provided in Figure 7.13 and again it is apparent that our ECH is more effective in blocking contagion propagation. Data for the weighted MONT-VA network and the weighted FB-2 networks are provided in Figures 7.14 and 7.15, respectively. Note that the abscissa label accounts for the total cost of blocking, for weighted networks. Once again, we see the same outcome. In all of these cases, ETH and HDH do not reduce the probability of a cascade from that for $\beta = 0$, so that p_c remains close to 1.0, while ECH, with a moderate number of blocking edges, halts all diffusion. The results in these plots are representative of those for the space of conditions in Table 7.6.

Basic Behavior and Parametric Studies

We now turn to investigating the behavior of the ECH blocking method and network dynamics.

Figure 7.16 is produced by taking the final spread fraction for each diffusion instance of a group of 100 instances—so the instances vary only in the composition of seed sets for a

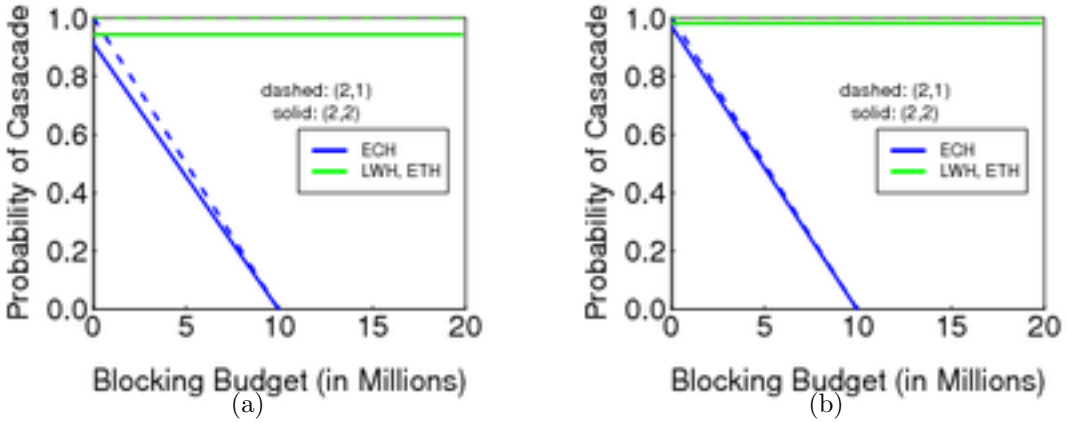


Figure 7.14: Data for the weighted MONT-VA network, showing probability of cascade versus β, for the (n_s, θ) pairs in the plots: (a) $n_s = 2$; (b) $n_s = 3$. The ETH and HDH give $p_c = 1.0$ for threshold-1.

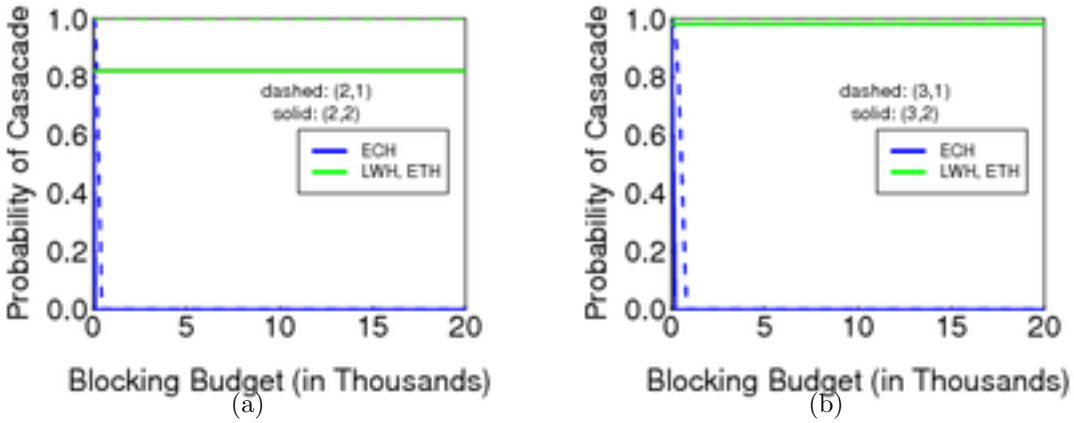


Figure 7.15: Data for the weighted FB-2 network, showing probability of cascade versus β, for the (n_s, θ) pairs in the plots: (a) $n_s = 2$; (b) $n_s = 3$. The ETH and HDH give $p_c = 1.0$ for threshold-1. ECH stops all diffusion with a $\beta < 1000$.

fixed n_s—and arranging these spread fractions in non-decreasing numerical order. Sharp transitions in the curves denote a sudden jump in the final spread size. From these data, probabilities of cascades are computed. For example, the probability of cascade for $(n_s, \theta) = (3, 3)$ is 0.38 (because 62 of the 100 diffusion instances produce no diffusion). These are the data used to compute p_c in the previous section.

In Figure 7.17, p_c is plotted against n_s for $\theta = 3, 5$ and $\beta = 100$. We converted FB-2 to an unweighted graph to compare with FB-1 and MONT-VA. As expected, for a fixed θ, p_c increases as n_s increases. However, there is a network structure effect in these data that is seemingly counterintuitive. From inspection of Table 7.5, the average degree and clustering coefficient are roughly twice as great for MONT-VA as for FB-1. Furthermore, for larger

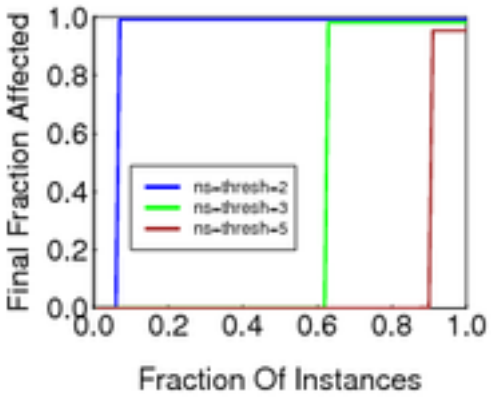


Figure 7.16: Simulation results for the unweighted MONT-VA network for (n_s, θ) pairs (1,1), (2,2), and (3,3). For each set of conditions, the final fractions of affected nodes are plotted in increasing numerical order. The abscissa value at which each curve rises sharply gives the probability of a cascade.

k-cores, in the range of $k = 20$ to 30, the sizes of the cores for MONT-VA are about $5\times$ those of FB-1, meaning that MONT-VA is more well-connected. With this information, it is natural to expect that MONT-VA more effectively spreads contagion. Yet, Figure 7.17 shows the opposite result. The explanation, we contend, is contained in Figure 7.11. FB-1 has a group of nodes with greater degrees than those in MONT-VA, and these nodes can drive contagion through the network. Thus, there are clearly network structure-dynamics interactions that complicate making generalizations across networks.

7.2.8 Conclusions and Future Directions

We formulated an edge blocking problems and contrasted them with node-based ones. We devised a heuristic for it and compared our heuristic to other methods from the literature and demonstrated that it provides significantly improved blocking performance. Future work includes additional complexity results for edge based schemes, and formulations of new heuristics.

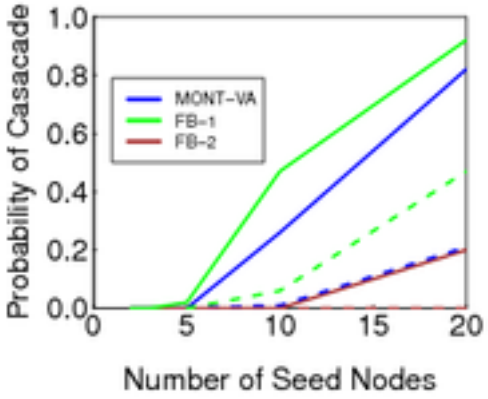


Figure 7.17: Simulation results for three unweighted networks, for fixed values of threshold 3 (solid) and 5 (dashed) and $\beta = 100$. As seed set size increases, the probability of cascade increases.

Chapter 8

Conclusions and Future Direction

In this chapter, we provide a high-level discussion on the achieved goals and the potential extensions of this thesis. The primary objective of this thesis is to take first step in developing an environment for network-centric interventions for controlling the smoking epidemic. The three main long-term goals to develop such an environment are: (1) determine the role of online social networks on smoking behavior, (2) create an operationalized model of smoking that can be simulated, and (3) devise actual methods to perform network-centric interventions. We contributed towards each of these three long-term goals by performing four studies — two Twitter-based surveillance studies, and two contagion modeling and simulation-based studies. An illustration of the long-term and achieved goals, and the future direction in each of the three main component of this thesis are shown in Figure 8.1. Next, we provide a high-level discussion and future directions of the studies divided as per long-term goals.

We performed Twitter-based surveillance studies of smoking-related tweets in order to measure the role of online social networks on smoking epidemic. We built a pipeline of software components to perform such surveillance studies. This pipeline handles all necessary steps, from data gathering through result visualization for performing social media-based studies. Then we utilized this pipeline to obtain Twitter data and estimated the exposed under-age Twitter user population to smoking-related messaging. We also used the pipeline to capture the Twitter users' sentiments towards tobacco smoking and electronic cigarettes as well for identifying highly active electronic cigarettes communities in the United States.

Through these studies, we have shown how to address important public health question through machine learning-based analytics of Twitter data. Efforts to limit the smoking initiation of underage populations into these behaviors need to take a multi-pronged approach. Our studies suggest that understanding and regulating social media may be an important part of this approach. We also combined spatiotemporal scan statistics and machine learning tools and techniques to infer the age of users and sentiments of tweets from the highly active

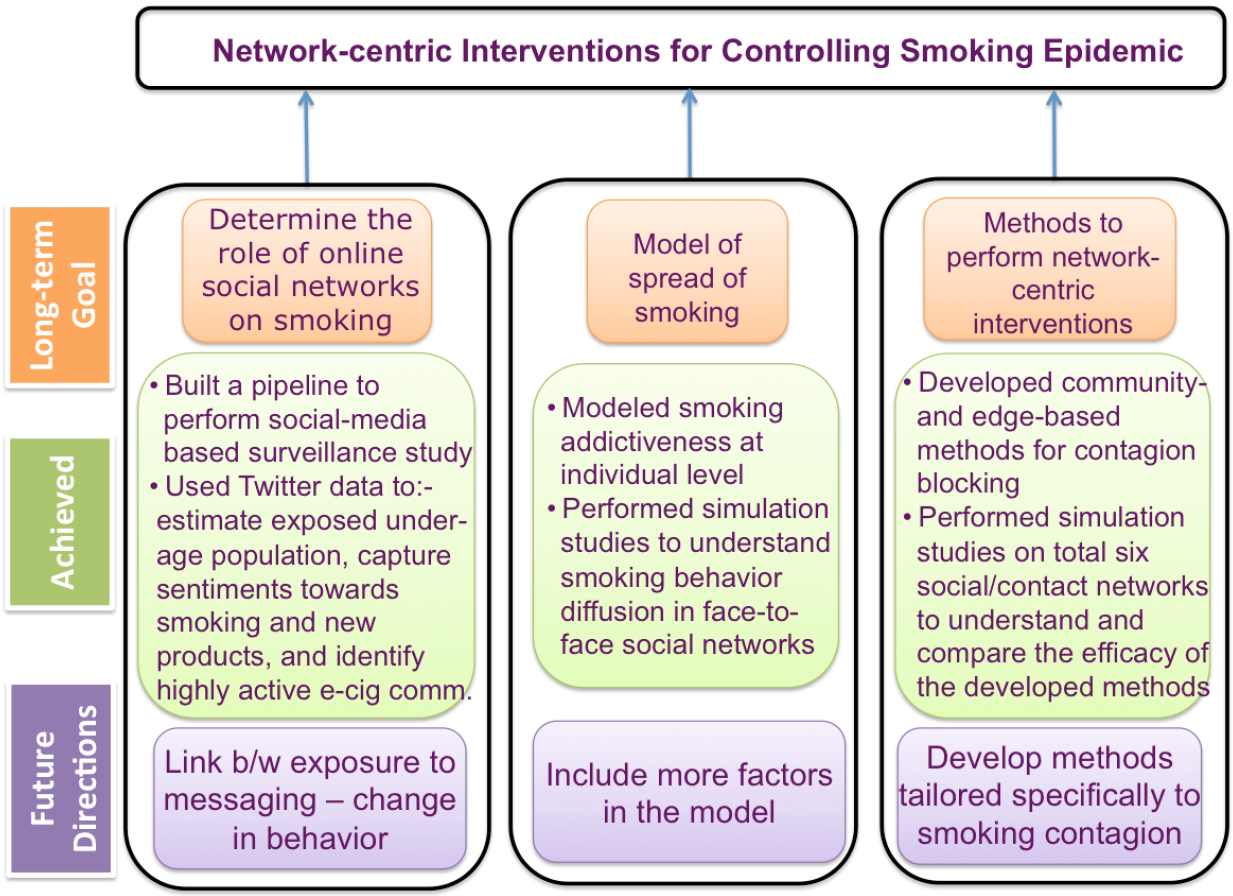


Figure 8.1: Overview of the contributions and potential future directions of the thesis

clusters of e-cig tweets in space and time. To our knowledge, this is the first study of its kind to analyze the hot-spot of e-cig related messaging over social media

A potential extension to this component can be a study the relationship between exposure to smoking-related messaging over social media and change in smoking behavior. Both smoking-related messaging and exposure to such messaging can impact smoking behavior or vice verse. A survey-based study should be performed primarily focusing on the young population that collects data on smoking behavior and social media usage, and then studies the correlation between the two. Asking specific questions on new smoking-related products such as e-cigs and e-hookahs will be very helpful in understating and validating the popularity of e-cig and other new products in certain communities.

The pipeline of the software components can also be improved in various manners. Some new features that can be added to the components of the pipeline are: allow on-the-fly search terms inclusion and exclusion to easily obtain new posts/comments from social media, provide full-text-search capability in the database that helps in performing a quick analytic on the extracted and processed data, or include new visualization tools that provide both aggregated and point data view of the social media posts on the map.

We employed mathematical modeling and simulation approach to built a model of smoking addictiveness at individual level. Our main contribution in this area is to introduce an extension to a infectious disease contagion model (i.e., SIS model), to account for the addictive nature of smoking behavior. We also study the effect of both addictiveness and peer influence together by performing an agent-based simulation of the model on a social network to replicate the effect of peer-influence. Overall, we showed that smoking epidemic can be modeled and simulated as an infections disease epidemic with minor changes.

The model of smoking addictiveness can be extended to represent the complex phenomenon of smoking contagion. Factors such as socioeconomic status, marital status, acess to cigarettes, and prices and policies should be included to the model. Data about all these factors can be included into an agent-based model driven by the presented model. Detailed synthetic information environments can be constructed by fusing data about these behaviors with other data sets on demographics, locations, and activities to build a complete picture of the ecology of a smoker.

Lastly, we used contagion modeling and simulation approach to investigate a well-motivated problem of social contagion blocking. In the light of findings from the literature that individuals start and stop smoking in groups, and that some edges are more important with respect to smoking initiation and cessation, we develop two methods to stymie both simple and complex (social) contagion. Our main contributions was that the community blocking heuristic is hybrid in nature. Therefore, it is driven by network structure and also incorporates contagion dynamics when a contagion is spreading to increase its effectiveness. For edge-based contagion blocking, we develop a practical edge-covering heuristic to block social contagion on a wide variety of networks, i.e., directed, weighted, and unweighted graphs. To evaluate our main heuristic, we perform computational experiments of contagion propagation on six social networks from the literature, some of which are five times greater in terms of numbers of nodes and an order of magnitude greater in numbers of edges than those used in previous studies.

The smoking contagion blocking needs to be studied more. The blocking techniques we devised were applicable to a general form of social contagion. We need to develop methods that are tailored specifically to smoking contagion. For this we need to gather community-by-community better estimates of: active smokers, peer-influence, mode of influence, sentiments towards existent and new smoking products, availability of such products etc. When such a data gets fused with a more mature model of smoking contagion, we can devise more robust network-centric interventions for controlling the smoking epidemic.

Bibliography

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38. Association for Computational Linguistics, 2011.

[2] Arpana Agrawal, Alan J Budney, and Michael T Lynskey. The co-occurring use and misuse of cannabis and tobacco: A review. *Addiction*, 107(7):1221–1233, 2012.

[3] Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors. In *Proceedings of the AAAI Conference on Weblogs and Social Media*, pages 387–390, 2012.

[4] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.

[5] Cheryl Alexander, Marina Piazza, Debra Mekos, and Thomas Valente. Peers, schools, and adolescent cigarette smoking. *Journal of adolescent health*, 29(1):22–30, 2001.

[6] Stacey J Anderson, Pamela M Ling, and Stanton A Glantz. Implications of the federal court order banning the terms "light" and "mild": What difference could it make? *Tobacco Control*, 16:275–279, 2007.

[7] Ashwin Arulselvan, Clayton W Commander, Lily Elefteriadou, and Panos M Pardalos. Detecting critical nodes in sparse graphs. *Computers & Operations Research*, 36(7):2193–2200, 2009.

[8] Nancy L Atkinson, Sandra L Saperstein, and John Pleis. Using the Internet for health-related activities: Findings from a national probability sample. *Journal of Medical Internet Research*, 11(1), 2009.

[9] Christopher Barrett, Stephen Eubank, Achla Marathe, Madhav V Marathe, Zhengzheng Pan, and Samarth Swarup. Information integration to support policy informatics. *The Innovation Journal*, 16(1):article 2, 2011.

[10] Christopher L Barrett, Richard J Beckman, Maleq Khan, VS Anil Kumar, Madhav V Marathe, Paula E Stretz, Tridib Dutta, and Bryan Lewis. Generation and analysis

of large synthetic social contact networks. In *Winter Simulation Conference*, pages 1003–1014. Winter Simulation Conference, 2009.

[11] Christopher L Barrett, Harry B Hunt, Madhav V Marathe, SS Ravi, Daniel J Rosenkrantz, and Richard E Stearns. Complexity of reachability problems for finite discrete dynamical systems. *Journal of Computer and System Sciences*, 72(8):1317–1345, 2006.

[12] Joseph E Bauer, Andrew Hyland, Qiang Li, Craig Steger, and K Michael Cummings. A longitudinal assessment of the impact of smoke-free worksite policies on tobacco use. *American Journal of Public Health*, 95(6):1024, 2005.

[13] NL Benowitz. Clinical pharmacology of nicotine: Implications for understanding, preventing, and treating tobacco addiction. *Clinical Pharmacology & Therapeutics*, 83(4):531–541, 2008.

[14] Lisa F Berkman, Thomas Glass, Ian Brissette, and Teresa E Seeman. From social integration to health: Durkheim in the new millennium. *Social science & medicine*, 51(6):843–857, 2000.

[15] Lois Biener. Anti-tobacco advertisements by Massachusetts and Philip Morris: What teenagers think. *Tobacco Control*, 11(suppl 2):ii43–ii46, 2002.

[16] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[17] Beth C Bock, Amanda L Graham, Christopher N Sciamanna, Jenelle Krishnamoorthy, Jessica Whiteley, Rosa Carmona-Barros, Raymond S Niaura, and David B Abrams. Smoking cessation treatment on the Internet: Content, quality, and usability. *Nicotine & Tobacco Research*, 6(2):207–219, 2004.

[18] Rafal Bogacz, Eric Brown, Jeff Moehlis, Philip Holmes, and Jonathan D Cohen. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review*, 113(4):700, 2006.

[19] Marián Boguná, Romualdo Pastor-Satorras, and Alessandro Vespignani. Epidemic spreading in complex networks with degree correlations. In *Proceedings of the XVIII Sitges Conference on Statistical Mechanics, Lecture Notes in Physics, Springer, Berlin*, pages 127–147. 2003.

[20] Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, 2012.

[21] John Britton and Ilze Bogdanovica. Tobacco control efforts in Europe. *The Lancet*, 381(9877):1588–1595, 2013.

[22] Ulla Broms, Karri Silventoinen, Eero Lahelma, Markku Koskenvou, and Jaakko Kaprio. Smoking cessation by socioeconomic status and marital status: The contributions of smoking behavior and family background. *Nicotine and Tobacco Research*, 6(3):447–455, 2004.

[23] Centers for Disease Control and Prevention. Tobacco use among middle and high school students — United States, 2011-2014. Technical Report MMWR 64(14):381-385, 2015.

[24] Damon Centola. The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197, 2010.

[25] Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. *American Journal of Sociology*, 113(3):702–734, 2007.

[26] Deepayan Chakrabarti, Yang Wang, Chenxi Wang, Jurij Leskovec, and Christos Faloutsos. Epidemic thresholds in real networks. *ACM Transactions on Information and System Security (TISSEC)*, 10(4):13–1 – 13–26, 2008.

[27] Ping-Hsin Chen, Helene Raskin White, and Robert J Pandina. Predictors of smoking cessation from adolescence into young adulthood. *Addictive Behaviors*, 26(4):517–529, 2001.

[28] Wen-Ying Sylvia Chou, Yvonne M Hunt, Ellen Burke Beckjord, Richard P Moser, and Bradford W Hesse. Social media use in the United States: Implications for health communication. *Journal of medical Internet research*, 11(4), 2009.

[29] Wen-ying Sylvia Chou, Abby Prestin, Claire Lyons, and Kuang-yi Wen. Web 2.0 for health promotion: Reviewing the current evidence. *American Journal of Public Health*, 103(1):e9–e18, 2013.

[30] Nicholas A Christakis and James H Fowler. The collective dynamics of smoking in a large social network. *New England journal of medicine*, 358(21):2249–2258, 2008.

[31] Nathan K Cobb, Amanda L Graham, and David B Abrams. Social network structure of a large online community for smoking cessation. *American Journal of Public Health*, 100(7):1282–1289, 2010.

[32] Nathan K Cobb, Amanda L Graham, Beth C Bock, George Papandonatos, and David B Abrams. Initial evaluation of a real-world Internet smoking cessation system. *Nicotine & tobacco research: official journal of the Society for Research on Nicotine and Tobacco*, 7(2):207, 2005.

[33] Nathan K Cobb, Amanda L Graham, M Justin Byron, and David B Abrams. Online social networks and smoking cessation: A scientific research agenda. *Journal of medical Internet research*, 13(4), 2011.

[34] Sheldon Cohen, Edward Lichtenstein, Karen Kingsolver, Robin Mermelstein, John S Baer, and Thomas W Kamarck. Social support interventions for smoking cessation. 1988.

[35] H Catherina Coppotelli and C Tracy Orleans. Partner support and other determinants of smoking cessation maintenance among women. *Journal of consulting and clinical psychology*, 53(4):455, 1985.

[36] Candace Currie, Cara Zanotti, Antony Morgan, Dorothy Currie, Margaretha de Looze, Chris Roberts, Oddrun Samdal, Otto Smith, and Vivian Barnekow. *Social determinants of health and well-being among young people*. Number 6. 2012.

[37] Madeline A Dalton, James D Sargent, Michael L Beach, Linda Titus-Ernstoff, Jennifer J Gibson, M Bridget Ahrens, Jennifer J Tickle, and Todd F Heatherton. Effect of viewing smoking in movies on adolescent smoking initiation: A cohort study. *The Lancet*, 362(9380):281–285, 2003.

[38] Ronald M Davis, Elizabeth A Gilpin, Barbara Loken, K Viswanath, and Melanie A Wakefield. The role of the media in promoting and reducing tobacco use. *Health*, 98:4302, 1998.

[39] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 3267–3276, New York, NY, USA, 2013. ACM.

[40] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICSWM 2013)*, 2013.

[41] Dan Dumbrell and Robert Steele. The changing nature of health information dissemination through the role of social media. *Applied Mechanics and Materials*, 411:110–114, 2013.

[42] David Easley and Jon Kleinberg. Networks, crowds, and markets. *Cambridge Univ Press*, 6(1):6–1, 2010.

[43] Phyllis L Ellickson, Maria Orlando, Joan S Tucker, and David J Klein. From adolescence to young adulthood: Racial/ethnic disparities in smoking. *American Journal of Public Health*, 94(2):293–299, 2004.

[44] Nicole B Ellison, Charles Steinfield, and Cliff Lampe. The benefits of Facebook "friends": Social capital and college students use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168, 2007.

[45] Sherry L Emery, Lisa Vera, Jidong Huang, and Glen Szczypka. Wanna know about vaping?vPatterns of message exposure, seeking and sharing information about e-cigarettes across media platforms. *Tobacco Control*, 23(suppl 3):iii17–iii25, 2014.

[46] Susan T Ennett, Robert Faris, John Hipp, Vangie A Foshee, Karl E Bauman, Andrea Hussong, and Li Cai. Peer smoking, other peer attributes, and adolescent cigarette smoking: A social network analysis. *Prevention Science*, 9(2):88–98, 2008.

[47] Stephen Eubank, VS Anil Kumar, Madhav V Marathe, Aravind Srinivasan, and Nan Wang. Structure of social contact networks and their impact on epidemics. volume 70, page 181. American Mathematical Society, 2006.

[48] Songlin Fei. Applying hotspot detection methods in forestry: A case study of chestnut oak regeneration. *International Journal of Forestry Research*, 2010.

[49] Manning Feinleib, William B Kannel, Robert J Garrison, Patricia M McNamara, and William P Castelli. The Framingham Offspring Study. Design and preliminary data. *Preventive Medicine*, 4(4):518–525, 1975.

[50] Caroline M Fichtenberg and Stanton A Glantz. Effect of smoke-free workplaces on smoking behaviour: Systematic review. *BMJ: British Medical Journal*, 325(7357):188, 2002.

[51] Centers for Disease Control and Prevention. Coverage for tobacco use cessation treatments. Technical report, Centers for Disease Control and Prevention, 2003.

[52] Centers for Disease Control and Prevention. Smoking-attributable mortality, years of potential life lost, and productivity losses - United States, 2000-2004. Technical Report MMWR 2008 57(45):1126-28, Centers for Disease Control and Prevention, 2008.

[53] Centers for Disease Control and Prevention. Cigarette use among high school students – United States, 1991-2009. Technical Report MMWR 2010 59:797801, Centers for Disease Control and Prevention, 2010.

[54] Centers for Disease Control and Prevention. Early release of selected estimates based on data from the 2010 National Health Interview Survey: Current smoking, 2010.

[55] Centers for Disease Control and Prevention. Comprehensive smoke-free laws 50 largest U.S. cities, 2000 and 2012. Technical Report MMWR 61(45):914-917, Centers for Disease Control and Prevention, 2012.

[56] Centers for Disease Control and Prevention. Current cigarette smoking among adults — United States, 2005–2012. Technical Report MMWR 63(02):29-34, Centers for Disease Control and Prevention, 2012.

[57] Susan R Forsyth and Ruth E Malone. "I'll be your cigarette — Light me up and get on with it": Examining smoking imagery on youtube. *Nicotine & Tobacco Research*, pages 1–7, 2010.

[58] James H Fowler and Nicholas A Christakis. Cooperative behavior cascades in human social networks. *Proceedings of the National Academy of Sciences*, 107(12):5334–5338, 2010.

[59] Kay-Lambkin Frances. Social influence, addictions and the Internet: The potential of web 2.0 technologies in enhancing treatment for alcohol/other drug use problems. *Journal of Addiction Research & Therapy*, 2012.

[60] Becky Freeman and Simon Chapman. British American Tobacco on Facebook: Undermining Article 13 of the global World Health Organization Framework Convention on Tobacco Control. *Tobacco Control*, 19:e1–e9, 2010.

[61] Jean Gaudart, Belco Poudiougou, Alassane Dicko, Stéphane Ranque, Ousmane Toure, Issaka Sagara, Mouctar Diallo, Sory Diawara, Amed Ouattara, Mahamadou Diakite, et al. Space-time clustering of childhood malaria at the household level: A dynamic cohort in a mali village. *BMC Public Health*, 6(1):286, 2006.

[62] Surgeon General. *Reducing the health consequences of smoking: 25 years of progress*. US Department of Health and Human Services, 1989.

[63] Stephen E Gilman, Richard Rende, Julie Boergers, David B Abrams, Stephen L Buka, Melissa A Clark, Suzanne M Colby, Brian Hitsman, Alessandra N Kazura, Lewis P Lipsitt, et al. Parental smoking and adolescent smoking initiation: An intergenerational perspective on tobacco control. *Pediatrics*, 123:e274–e281, 2009.

[64] Gary A Giovino, Sara A Mirza, Jonathan M Samet, Prakash C Gupta, Martin J Jarvis, Neeraj Bhala, Richard Peto, Witold Zatonski, Jason Hsia, Jeremy Morton, et al. Tobacco use in 3 billion individuals from 16 countries: An analysis of nationally representative cross-sectional household surveys. *The Lancet*, 380(9842):668–679, 2012.

[65] Myong-Hyun Go, Harold D Green Jr., David P Kennedy, Michael Pollard, and Joan S Tucker. Peer influence and selection effects on adolescent smoking. *Drug and Alcohol Dependence*, 109:239–242, 2010.

[66] Sandra Gonzalez-Bailon, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. The Dynamics of Protest Recruitment Through an Online Network. *Nature Scientific Reports*, pages 1–7, 2011. DOI: 10.1038/srep00197.

[67] Amanda L Graham, Nathan K Cobb, Linda Raymond, Stewart Sill, and Joyce Young. Effectiveness of an Internet-based worksite smoking cessation intervention at 12 months. *Journal of Occupational and Environmental Medicine*, 49(8):821–828, 2007.

[68] Rachel A Grana and Pamela M Ling. smoking revolution: a content analysis of electronic cigarette retail websites. *American Journal of Preventive Medicine*, 46(4):395–403, 2014.

[69] Mark Granovetter. The Strength of Weak Ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[70] Mark Granovetter. Threshold Models of Collective Behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.

[71] Liv Grøtvedt and Knut Stavem. Association between age, gender and reasons for smoking cessation. *Scandinavian Journal of Public Health*, 33(1):72–76, 2005.

[72] Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, pages 491–501, New York, NY, USA, 2004. ACM.

[73] Britt Gustafsson and John Carstensen. Space–time clustering of childhood lymphatic leukaemias and non-hodgkin's lymphomas in sweden. *European Journal of Epidemiology*, 16(12):1111–1116, 2000.

[74] Jeffrey A Hall and Thomas W Valente. Adolescent smoking networks: The effects of influence and selection on future smoking. *Addictive Behaviors*, 32(12):3054–3059, 2007.

[75] Kathleen Mullan Harris. The national longitudinal study of adolescent health (Add Health), waves i and ii, 1994-1996; wave iii, 2001-2002, 2008. Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill 2008.

[76] Kathleen Mullan Harris, Francesca Florey, Joyce Tabor, Peter S Bearman, Jo Jones, and J Richard Udry. The national longitudinal study of adolescent health: Research design. *Carolina Population Center, University of North Carolina at Chapel Hill*, 2003.

[77] Carleen Hawn. Take two aspirin and tweet me in the morning: How Twitter, Facebook, and other social media are reshaping health care. *Health Affairs*, 28(2):361–368, 2009.

[78] Yogi H. Hendlin, Stacey J. Anderson, and Stanton A. Glantz. "Acceptable rebellion": Marketing hipster aesthetics to sell Camel cigarettes. *Tobacco Control*, 19:213–222, 2010.

[79] Lisa Henriksen, Ellen C Feighery, Nina C Schleicher, David W Cowling, Randolph S Kline, and Stephen P Fortmann. Is adolescent smoking related to the density and proximity of tobacco outlets and retail cigarette advertising near schools? *Preventive medicine*, 47(2):210–214, 2008.

[80] Rosemary Hiscock, Linda Bauld, Amanda Amos, Jennifer A Fidler, and Marcus Munafò. Socioeconomic status and smoking: A review. *Annals of the New York Academy of Sciences*, 1248(1):107–123, 2012.

[81] Brooke A Hixson, Saad B Omer, Carlos del Rio, and Paula M Frew. Spatial clustering of HIV prevalence in Atlanta, Georgia and population characteristics associated with case concentrations. *Journal of urban health*, 88(1):129–141, 2011.

[82] Beth R. Hoffman, Steve Sussman, Jennifer B. Unger, and Thomas W. Valente. Peer influences on adolescent cigarette smoking: A theoretical review of the literature. *Substance Use and Misuse*, 41:103–155, 2006.

[83] Tad Hogg and Kristina Lerman. Social dynamics of Digg. *EPJ Data Science*, 1(1):1–26, 2012.

[84] Tad Hogg, Kristina Lerman, and Laura M. Smith. Stochastic models predict user behavior in social media. *Human Journal*, 2(1):25–39, 2013.

[85] Jidong Huang, Rachel Kornfield, Glen Szczypka, and Sherry L Emery. A cross-sectional examination of marketing of electronic cigarettes on Twitter. *Tobacco Control*, 23(suppl 3):iii26–iii30, 2014.

[86] Benjamin WK Hung, Stephan E Kolitz, and Asuman Ozdaglar. Optimization-based influencing of village social networks in a counterinsurgency. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 10–17. Springer, 2011.

[87] Corinne G Husten and Lawrence R Deyton. Understanding the tobacco control act: Efforts by the us food and drug administration to make tobacco-related morbidity and mortality part of the usa's past, not its future. *The Lancet*, 381(9877):1570–1580, 2013.

[88] Instagram. Instagram website. Accessed: 2015-02-01.

[89] Christine Jackson. Initial and experimental stages of tobacco and alcohol use during late childhood: relation to peer, parent, and personal risk factors. *Addictive Behaviors*, 22(5):685–698, 1997.

[90] Christine Jackson and Lisa Henriksen. Do as i say: parent smoking, antismoking socialization, and smoking onset among children. *Addictive Behaviors*, 22(1):107–114, 1997.

[91] Prabhat Jha and Frank J Chaloupka. *Curbing the epidemic: Governments and the economics of tobacco control*. World Bank Publications, 1999.

[92] Lloyd D Johnston, Patrick M O'Malley, and Yvonne M Terry-McElrath. Methods, locations, and ease of cigarette access for american youth, 1997–2002. *American Journal of Preventive Medicine*, 27(4):267–276, 2004.

[93] Phillip Jones, David Gunnell, Stephen Platt, Jonathan Scourfield, Keith Lloyd, Peter Huxley, Ann John, Babar Kamran, Claudia Wells, and Michael Dennis. Identifying probable suicide clusters in wales using national mortality data. *PloS one*, 8(8):e71713, 2013.

[94] Jure Leskovec website, 2011. `http://cs.stanford.edu/people/jure/`.

[95] Jon D Kassel, Laura R Stroud, and Carol A Paronis. Smoking, stress, and negative affect: Correlation, causation, and context across stages of smoking. *Psychological bulletin*, 129(2):270, 2003.

[96] Nancy L Keating, A James O'Malley, Joanne M Murabito, Kirsten P Smith, and Nicholas A Christakis. Minimal social network effects evident in cancer screening behavior. *Cancer*, 117(13):3045–3052, 2011.

[97] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[98] Masahiro Kimura, Kazumi Saito, and Hiroshi Motoda. Minimizing the spread of contamination by blocking links in a network. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, 2008.

[99] Masahiro Kimura, Kazumi Saito, and Hiroshi Motoda. Solving the contamination minimization problem on networks for the Linear Threshold Model. In *PRICAI 2008: Trends in Artificial Intelligence*, pages 977–984. Springer, 2008.

[100] Brian A King, Roshni Patel, Kimberly Nguyen, and Shanta R Dube. Trends in awareness and use of electronic cigarettes among us adults, 2010-2013. *Nicotine & Tobacco Research*, 17(2):219–27, 2015.

[101] Judith M Knapp, Christine L Rosheim, Edward A Meister, and Thomas E Kottke. Managing tobacco dependence in chemical dependency treatment facilities: A survey of current attitudes and policies. *Journal of Addictive Diseases*, 12(4):89–104, 1993.

[102] Chris J. Kuhlman, V. S. Anil Kumar, Madhav V. Marathe, S. S. Ravi, and Daniel J. Rosenkrantz. Inhibiting diffusion of complex contagions in social networks: Theoretical and experimental results. *Data mining and knowledge discovery*, 29(2):423–465, 2015.

[103] Chris J Kuhlman, VS Kumar, Madhav V Marathe, SS Ravi, and Daniel J Rosenkrantz. Finding critical nodes for inhibiting diffusion of complex contagions in social networks. In *Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part II*, pages 111–127. Springer-Verlag, 2010.

[104] Chris J Kuhlman, Gaurav Tuli, Samarth Swarup, Madhav V Marathe, and SS Ravi. Blocking simple and complex contagion by edge removal. In *IEEE International Conference on Data Mining series (ICDM)*, pages 399–408. IEEE, 2013.

[105] Martin Kulldorff. A spatial scan statistic. *Communications in Statistics-Theory and methods*, 26(6):1481–1496, 1997.

[106] Martin Kulldorff. Satscan-software for the spatial, temporal, and space-time scan statistics. *Boston: Harvard Medical School and Harvard Pilgrim Health Care*, 2010.

[107] Cynthia M Lakon, John R Hipp, and David S Timberlake. The social context of adolescent smoking: A a systems perspective. *American Journal of Public Health*, 100(7):1218, 2010.

[108] Cynthia M Lakon and Thomas W Valente. Social integration in friendship networks: The synergy of network structure and peer influence in relation to cigarette smoking among high risk adolescents. *Social Science & Medicine*, 74(9):1407–1417, 2012.

[109] Alex Lamb, Michael J Paul, and Mark Dredze. Separating fact from fear: Tracking flu infections on twitter. In *HLT-NAACL*, pages 789–795, 2013.

[110] Vasileios Lampos, Tijl De Bie, and Nello Cristianini. Flu detector – tracking epidemics on Twitter. In J. L. Balcázar and et al., editors, *Proceedings of ECML PKDD*, volume 6323, Part III of *LNAI*, pages 599–602, Berlin Heidelberg, 2010. Springer Verlag.

[111] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: The coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[112] Leslie Lenert, Ricardo F Muñoz, John E Perez, and Aditya Bansod. Automated e-mail messaging as a tool for improving quit rates in an Internet smoking cessation intervention. *Journal of the American Medical Informatics Association*, 11(4):235–240, 2004.

[113] Leslie Lenert, Ricardo F Muñoz, Jackie Stoddard, Kevin Delucchi, Aditya Bansod, Steven Skoczen, and Eliseo J Pérez-Stable. Design and pilot evaluation of an Internet smoking cessation program. *Journal of the American Medical Informatics Association*, 10(1):16–20, 2003.

[114] Amanda Lenhart, Kristen Purcell, Aaron Smith, and Kathryn Zickuhr. Social media & mobile Internet use among teens and young adults. millennials. *Pew Internet & American Life Project*, 2010.

[115] David T Levy, Andrew Hyland, Cheryl Higbee, Lillian Remer, and Christine Compton. The role of public policies in reducing smoking prevalence in california: Results from the california tobacco policy simulation model¡/i¿. *Health Policy*, 82(2):167–185, 2007.

[116] Lan Liang, Frank Chaloupka, Mark Nichter, and Richard Clayton. Prices, policies and youth smoking, May 2001. *Addiction*, 98(Suppl 1):105–122, 2003.

[117] Yue Liao, Zhaoqing Huang, Jimi Huh, Mary Ann Pentz, and Chih-Ping Chou. Changes in friends and parental influences on cigarette smoking from early through late adolescence. *Journal of Adolescent Health*, 53(1):132–138, 2013.

[118] Sharon Lipperman-Kreda, Joel W Grube, and Karen B Friend. Local tobacco policy and tobacco outlet density: Associations with youth smoking. *Journal of Adolescent Health*, 50(6):547–552, 2012.

[119] Yang-Yu Liu, Jean-Jacques Slotine, and Albert-Lszl Barabsi. Controllability of complex networks. *Nature*, 473:167–173, 2011.

[120] LongURL. LongURL API version 2.0 documentation. Accessed: 2014-10-27.

[121] Francisco J Luquero, Cunhate Na Banga, Daniel Remartínez, Pedro Pablo Palma, Emanuel Baron, and Rebeca F Grais. Cholera epidemic in guinea-bissau (2008): the importance of place. *PloS one*, 6(5):e19005, 2011.

[122] Judith Mackay, Bungon Ritthiphakdee, and K Srinath Reddy. Tobacco control in asia. *The Lancet*, 381(9877):1581–1587, 2013.

[123] Mary Madden, Amanda Lenhart, Sandra Cortesi, Urs Gasser, Maeve Duggan, Aaron Smith, and Meredith Beaton. Teens, social media, and privacy. *Pew Research Center. http://www. pewinternet. org/˜/media//Files/Reports/2013/PIP_TeensSocialMediaandPrivacy. pdf*, 2013.

[124] William J McCarthy, Ritesh Mistry, Yao Lu, Minal Patel, Hong Zheng, and Barbara Dietsch. Density of tobacco retailers near schools: Effects on tobacco use among students. *American Journal of Public Health*, 99(11):2006, 2009.

[125] Sara C Mednick, Nicholas A Christakis, and James H Fowler. The spread of sleep loss influences drug use in adolescent social networks. *PloS one*, 5(3):e9775, 2010.

[126] Robin Mermelstein, Sheldon Cohen, Edward Lichtenstein, John S Baer, and Tom Kamarck. Social support and smoking cessation and maintenance. *Journal of consulting and clinical psychology*, 54(4):447, 1986.

[127] R Garey Michael and S Johnson David. *Computers and Intractability: A Guide to the Theory of NP-completeness*. 1979.

[128] George E Moore, Michael P Ward, Martin Kulldorff, Richard J Caldanaro, Lynn F Guptill, Hugh B Lewis, and Lawrence T Glickman. A space–time cluster of adverse events associated with canine rabies vaccine. *Vaccine*, 23(48):5557–5562, 2005.

[129] Laurence Moore, Chris Roberts, and Chris Tudor-Smith. School smoking policies and smoking prevalence among adolescents: Multilevel analysis of cross-sectional data from wales. *Tobacco Control*, 10(2):117–123, 2001.

[130] Mark Myslín, Shu-Hong Zhu, Wendy Chapman, and Mike Conway. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research*, 15(8), 2013.

[131] Ruchit Nagar, Qingyu Yuan, Clark C Freifeld, Mauricio Santillana, Aaron Nojima, Rumi Chunara, and John S Brownstein. A case study of the new york city 2012-2013 influenza season with daily geocoded twitter data from temporal and spatiotemporal perspectives. *Journal of medical Internet research*, 16(10), 2014.

[132] Daniel B Neill and Andrew W Moore. Anomalous spatial cluster detection. In *Proceedings of the KDD 2005 Workshop on Data Mining Methods for Anomaly Detection*, 2005.

[133] KJ Neufeld, DH Peters, M Rani, S Bonu, and RK Brooner. Regular use of alcohol and tobacco in India and its association with age, gender, and poverty. *Drug and alcohol dependence*, 77(3):283–291, 2005.

[134] Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. "How old do you think I am?": A study of language and age in Twitter. In *Proc. ICWSM*, 2013.

[135] Stephanie N Nguyen, Isabelle Von Kohorn, Dena Schulman-Green, and Eve R Colson. The importance of social networks on smoking: Perspectives of women who quit smoking during pregnancy. *Maternal and child health journal*, 16(6):1312–1318, 2012.

[136] U.S. Department of Health and Human Services. The health benefits of smoking cessation: A report of the Surgeon General. Atlanta: Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health.

[137] U.S. Department of Health and Human Services. The health benefits of smoking cessation: A report of the Surgeon General. Atlanta: Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health.

[138] U.S. Department of Health and Human Services. Reducing the health consequences of smoking: 25 years of progress. a report of the Surgeon General. U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. Technical Report DHHS Publication No. (CDC) 89-8411, U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 1989.

[139] National Institute on Drug Abuse. Infofacts: Cigarettes and other tobacco products, 2011.

[140] National Institute on Drug Abuse. Tobacco addiction. *research report series*, 2011.

[141] World Health Organization et al. *WHO report on the global tobacco epidemic, 2008: The MPOWER package*. Geneva: World Health Organization, 2008.

[142] Lisbet Øygard, KNUT-INGE KLEPP, Grethe S Tell, and Odd D Vellar. Parental and peer influences on smoking among young adults: ten-year follow-up of the oslo youth study participants. *Addiction*, 90(4):561–569, 1995.

[143] Meri Paavola, Erkki Vartiainen, and Pekka Puska. Smoking cessation between teenage years and adulthood. *Health Education Research*, 16(1):49–57, 2001.

[144] Mark A Pachucki, Paul F Jacques, and Nicholas A Christakis. Social network concordance in food choice among spouses, friends, and siblings. *American Journal of Public Health*, 101(11):2170, 2011.

[145] Michael J Paul, Mark Dredze, and David Broniatowski. Twitter improves influenza forecasting. *PLoS currents*, 6, 2013.

[146] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[147] Donna A Perez, Anne C Grunseit, Chris Rissel, James Kite, Trish Cotter, Sally Dunlop, and Adrian Bauman. Tobacco promotion "below the line": Exposure among adolescents and young adults in NSW, Australia. *BMC public health*, 12(1):429, 2012.

[148] B. Aditya Prakash, Deepayan Chakrabarti, Michalis Faloutsos, Nicholas Valler, and Christos Faloutsos. Threshold Conditions for Arbitrary Cascade Models on Arbitrary Graphs. In *Proceedings of the 11th IEEE Conference on Data Mining (ICDM 2011)*, pages 537–546, 2011.

[149] Judith J Prochaska, Cornelia Pechmann, Romina Kim, and James M Leonhardt. Twitter = quitter? An analysis of Twitter quit smoking social networks. *Tobacco Control*, 21(4):447–449, 2012.

[150] Delip Rao and David Yarowsky. Detecting latent user properties in social media. In *Proceedings of the NIPS Workshop on Machine Learning for Social Computing*, Whistler, BC, Canada, Dec 2010.

[151] Anand Reddi. Health information and the like. *Science*, 342(6164):1315, 2013.

[152] Lorraine R Reitzel, Ellen K Cromley, Yisheng Li, Yumei Cao, Richard Dela Mater, Carlos A Mazas, Ludmila Cofta-Woerpel, Paul M Cinciripini, and David W Wetter. The effect of tobacco outlet density and proximity on smoking cessation. *American Journal of Public Health*, 101(2):315, 2011.

[153] Timothy C Reluga, Jan Medlock, and Alan S Perelson. Backward bifurcations and multiple equilibria in epidemic models with structured immunity. *Journal of Theoretical Biology*, 252:155–165, 2008.

[154] Amanda Richardson, Amanda L Graham, Nathan Cobb, Haijun Xiao, Aaron Mushro, David Abrams, and Donna Vallone. Engagement promotes abstinence in a web-based cessation intervention: Cohort study. *Journal of medical Internet research*, 15(1), 2013.

[155] Nancy A Rigotti. Reducing the supply of tobacco to youths. *Regulating Tobacco*, page 143, 2001.

[156] Luis R Rivero, James L Persson, David C Romine, John T Taylor, Theron C Toole, Christopher J Trollman, and William W Au. Towards the world-wide ban of indoor cigarette smoking in public places. *International journal of hygiene and environmental health*, 209(1):1–14, 2006.

[157] Caitlin M Rivers and Bryan L Lewis. Ethical research standards in a world of big data. *F1000 Research*, 3(38), 2014.

[158] J Niels Rosenquist, Joanne Murabito, James H Fowler, and Nicholas A Christakis. The spread of alcohol consumption behavior in a large social network. *Annals of Internal Medicine*, 152(7):426–433, 2010.

[159] Henry Saffer and Frank Chaloupka. The effect of tobacco advertising bans on tobacco consumption. *Journal of health economics*, 19(6):1117–1137, 2000.

[160] Marcel Salathé, Duy Q Vu, Shashank Khandelwal, and David R Hunter. The dynamics of health behavior sentiments on a large online social network. *EPJ Data Science*, 2(1):1–12, 2013.

[161] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856, 2006.

[162] Daniel Scanfeld, Vanessa Scanfeld, and Elaine L Larson. Dissemination of health information through social networks: Twitter and antibiotics. *American journal of infection control*, 38(3):182–188, 2010.

[163] David R Schaefer, Steven A Haas, and Nicholas J Bishop. A dynamic model of us adolescents smoking and friendship networks. *American Journal of Public Health*, 102(6):e12–e18, 2012.

[164] Thomas C Schelling. *Micromotives and macrobehavior*. WW Norton & Company, 1978.

[165] David A. Siegel. Non-disruptive tactics of suppression are superior in countering terrorism, insurgency, and financial panics. *PLoS ONE*, 6:e18545–1–e18545–6, 2011.

[166] Katherine Clegg Smith, Frances Stillman, Lee Bone, Norman Yancey, Emmanuel Price, Precilla Belin, and Elizabeth Edsall Kromm. Buying and selling loosies in baltimore: The informal exchange of cigarettes in the community context. *Journal of Urban Health*, 84(4):494–507, 2007.

[167] Joanne R Stevenson, Christopher T Emrich, Jerry T Mitchell, and Susan L Cutter. Using building permits to monitor disaster recovery: A spatio-temporal case study of coastal mississippi following hurricane katrina. *Cartography and Geographic Information Science*, 37(1):57–68, 2010.

[168] Victor J Strecher, Jennifer B McClure, Gwen L Alexander, Bibhas Chakraborty, Vijay N Nair, Janine M Konkel, Sarah M Greene, Linda M Collins, Carola C Carlier, Cheryl J Wiese, et al. Web-based smoking-cessation programs: results of a randomized trial. *American journal of Preventive Medicine*, 34(5):373–381, 2008.

[169] SwarmApp. Swarm website. Accessed: 2015-02-01.

[170] Rosemary Thackeray, Benjamin T Crookston, and Joshua H West. Correlates of health-related social media use among adults. *Journal of medical Internet research*, 15(1), 2013.

[171] Leonard Thompson. *A History of South Africa*. Yale University Press, New Haven, CT, 3 edition, 2001.

[172] Hanghang Tong, B Aditya Prakash, Tina Eliassi-Rad, Michalis Faloutsos, and Christos Faloutsos. Gelling, and melting, large graphs by edge manipulation. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 245–254. ACM, 2012.

[173] Joy Townsend. Price and consumption of tobacco. *British Medical Bulletin*, 52(1):132–142, 1996.

[174] Amanda L Traud, Eric D Kelsic, Peter J Mucha, and Mason A Porter. Comparing community structure to characteristics in online collegiate social networks. *SIAM review*, 53(3):526–543, 2011.

[175] Gaurav Tuli, Chris J Kuhlman, Madhav V Marathe, SS Ravi, and Daniel J Rosenkrantz. Blocking complex contagions using community structure. In *Proc. of Multiagent Interaction Networks (MAIN2013) Workshop of AAMAS*, 2013.

[176] Gaurav Tuli, Madhav V Marathe, Jared Hawkins, Clark Freifeld, John Brownstein, and Samarth Swarup. Find and analyze the hotspots of non-commercial electronic cigarette tweets. Manuscript under preparation for publication, 2015.

[177] Gaurav Tuli, Madhav V Marathe, Kiran Lakkaraju, and Samarth Swarup. Analyzing the exposure of vulnerable population to smoking-related messages on twitter. Manuscript submitted for publication, 2015.

[178] Gaurav Tuli, Madhav V Marathe, SS Ravi, and Samarth Swarup. Addiction dynamics may explain the slow decline of smoking prevalence. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 114–122. Springer, 2012.

[179] Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. Structural Diversity in Social Contagion. *Proceedings of the Naitonal Academy of Sciences (PNAS 2012)*, 109(9):5962–5966, 2012.

[180] Jennifer B Unger and Xinguang Chen. The role of social networks and media receptivity in predicting age of smoking initiation: A proportional hazards model of risk and protective factors. *Addictive Behaviors*, 24(3):371–381, 1999.

[181] Jennifer B Unger, Tess Boley Cruz, Darleen Schuster, June A Flora, and C Anderson Johnson. Measuring exposure to pro- and anti-tobacco marketing among adolescents: Intercorrelations among measures and associations with smoking status. *Journal of Health Communication*, 6:11–29, 2001.

[182] Jennifer B Unger, C Anderson Johnson, Jacqueline L Stoddard, Elahe Nezami, and Chou Chih-Ping. Identification of adolescents at risk for smoking initiation: validation of a measure of susceptibility. *Addictive Behaviors*, 22(1):81–91, 1997.

[183] U.S. Census Bureau. Centers of population by county, 2010. Accessed: 2015-01-20.

[184] U.S. Department of Health and Human Services. *Preventing Tobacco Use Among Youth and Young Adults: A Report of the Surgeon General*. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, Office on Smoking and Health, Atlanta, 2012.

[185] Thomas W Valente. Network models of the diffusion of innovations. *Computational & Mathematical Organization Theory*, 2(2):163–164, 1996.

[186] Thomas W Valente. Network Interventions. *Science*, 337:49–53, 2012.

[187] Thomas W Valente, Jennifer B Unger, and C Anderson Johnson. Do popular students smoke? the association between popularity and smoking among middle school students. *Journal of Adolescent Health*, 37(4):323–329, 2005.

[188] Andrea Villanti, Marc Boulay, and Hee-Soon Juon. Peer, parent and media influences on adolescent smoking by developmental stage. *Addictive Behaviors*, 36:133–136, 2011.

[189] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P. Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN'09)*, August 2009.

[190] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P Gummadi. On the evolution of user interaction in Facebook. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 37–42. ACM, 2009.

[191] Frank Vitaro, Brigitte Wanner, Mara Brendgen, Catherine Gosselin, and Paul L Gendreau. Differential contribution of parents and friends to smoking trajectories during adolescence. *Addictive Behaviors*, 29(4):831–835, 2004.

[192] Olivia Ann Wackowski, M Jane Lewis, and Cristine D Delnevo. Qualitative analysis of Camel Snus' website message board–Users' product perceptions, insights and online interactions. *Tobacco Control*, 20:e1, 2011.

[193] Fahui Wang, John Hartmann, Wei Luo, and Pingwen Huang. Gis-based spatial analysis of tai place names in southern china: an exploratory study of methodology. *Geographic Information Sciences*, 12(1):1–9, 2006.

[194] Duncan J Watts. A Simple Model of Global Cascades on Random Networks. *(PNAS)*, 99(9):5766–5771, 2002.

[195] Ming Wen, Heather Van Duker, and Lenora M Olson. Social contexts of regular smoking in adolescence: Towards a multidimensional ecological model. *Journal of adolescence*, 32(3):671–692, 2009.

[196] Robert West, Andy McEwen, Keith Bolling, and Lesley Owen. Smoking cessation and smoking patterns in the general population: a 1-year follow-up. *Addiction*, 96(6):891–902, 2001.

[197] WHO. *Smoke-free movies: From evidence to action.* World Health Organization, Geneva, Switzerland, 2009.

[198] WHO. *WHO Global Report: Mortality attributable to tobacco.* World Health Organization, Geneva, Switzerland, 2013.

[199] WHO. *WHO Report on the global tobacco epidemic.* World Health Organization, Geneva, Switzerland, 2013.

[200] Marilyn A Winkleby, Darius E Jatulis, Erica Frank, and Stephen P Fortmann. Socioeconomic status and health: How education, income, and occupation contribute to risk factors for cardiovascular disease. *American Journal of Public Health*, 82(6):816–820, 1992.

[201] E Yildiz, D Acemoglou, A Ozdaglar, A Saberi, and A Scaglione. Discrete opinion dynamics with stubborn agents. *OPRE-2011-01-026*, 2011.

[202] April M Zeoli, Jesenia M Pizarro, Sue C Grady, and Christopher Melde. Homicide as infectious disease: Using public health methods to investigate the diffusion of homicide. *Justice quarterly*, 31(3):609–632, 2014.

[203] Hai-Feng Zhang, Ke-Zan Li, Xin-Chu Fu, and Bing-Hong Wang. An efficient control strategy of epidemic spreading on scale-free networks. *Chinese Physical Review Letters*, 26:068901–1–068901–4, 2009.