INFECT: Infection Estimation in Social Networks

Biswarup Bhattacharya
University of Southern California
Los Angeles, CA
bbhattac@usc.edu

Iordanis Fostiropoulos
University of Southern California
Los Angeles, CA
fostirop@usc.edu

Negarsadat Abolhassani
University of Southern California
Los Angeles, CA
abolhass@usc.edu

Qing Dong
University of Southern California
Los Angeles, CA
qingdong@usc.edu

ABSTRACT

Assessing infection spread accurately to prepare effective mitigation strategies and develop treatment schedules is an important problem. Currently, health workers are interested in receiving accountable data in the shortest time possible to minimize the infection spread in the community. However, collecting data about the health status of people by considering relevant health data such as doctor visits takes time and it results in a gap between the infection burst and protection actions. In this paper, we visualized how contact networks may react to the introduction of infections and identified correlations between predicted health states and actual data using social media data. We first developed an SEIRS disease model which is reasonably accurate in performing realistic disease simulation. Using this realistic disease model, we optimized the disease parameters using historical training data from two social media sites Twitter and Flicker. Then we predicted health states of individuals and compared the results with the online activity or actual regional health information. We then evaluated the model on Los Angeles network and attempted to identify the key reasons behind an area being more prone to a disease. We considered influenza and used our model to perform evaluations on the 2018 flu season.

KEYWORDS

Health data, Infection prediction, SEIRS model

ACM Reference Format:
Biswarup Bhattacharya, Iordanis Fostiropoulos, Negarsadat Abolhassani, and Qing Dong. 2024. INFECT: Infection Estimation in Social Networks. In *Proceedings of Los Angeles, CA USA (DSSS).* ACM, New York, NY, USA, Article 4, 4 pages.

1 INTRODUCTION

Infectious diseases are a leading cause of death worldwide, particularly in low income countries. According to a study [], three infectious diseases were ranked in the top ten causes of death worldwide in 2016. Numerous studies have been conducted on assessing infection (contagion) states to analyze spread, prepare mitigation strategies and develop treatment schedules. Such studies are focused primarily on currently active (in the contact network) infectious diseases, e.g. influenza, tuberculosis. In this paper, we visualize how contact networks may react to introduction of infections and identify correlations between predicted health states and actual data. Our goal is to model the spread of diseases using mathematics to provide information for health workers about the levels of vaccination needed to protect a population. This graphic shows that when enough members of the population are immunized, they act as buffers against the spread of the infection to non-immunized people. Transmission of an infectious disease may occur through several pathways. However, for the purpose of this study, the direct contact of susceptible individuals with an infected one will be considered as the main transmission medium.

It is essential to model the spreading of infectious diseases on networks since it helps us to understand the spreading pattern of infectious diseases. It can help to decide whether a targeted vaccination program or quarantine is going to work. It also contributes to distinguish forces behind epidemic genesis due to the transmission of the infection which can enable us to design more effective prevention strategies. Moreover, there are not many infectious disease spreading models which build upon social networks with the rich context. By summarizing the disease-related online activity, we can further evaluate the performance of the proposed model.

The remainder of the paper is organized as follows. Section 2 summarizes the related work and background knowledge about health data. Section 3 introduces our approach and discusses each step in detail. In Section 4 we present our experimental results and discuss our datasets and networks. We conclude the paper by identifying future research directions in Section 5.

2 RELATED WORK & BACKGROUND

Several studies have been done on disease surveillance. Charles-Smith et al. [3] recommends identifying opportunities that enable public health professionals to integrate social media analytics into disease surveillance and outbreak management practice, which is precisely what this paper is seeking to work towards. Christakis and Fowler [4] found that social network analysis can predict flu outbreaks earlier than traditional tracking methods. Freifeld et al. [5] developed a system which is similar to what we plan to build. However, they considered only textual input and did not overlay disease simulations with actual activity to gather additional insights. Finally, Carroll et al. [2] called for the need of using data from different sources to gather public health insights – which this paper seeks to achieve.

3 PROBLEM SETUP

In this section, we first introduce notation for our problem. An individual can be in one of the following health states: S means that

DSSS, 2018
.

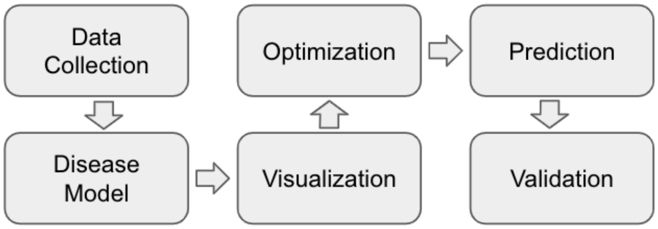

Figure 1: Overview of our approach

the individual is susceptible to disease (healthy), E means that the individual has been exposed and has latent disease, and I means that the individual is infected, and R means that the individual has permanently recovered (or immunized) from the disease.

3.1 Disease Model and Network

Disease Model: We adopt a SEIRS model [9] for modeling the disease dynamics. Tuberculosis and many other infectious diseases follow a SEIRS pattern, where treated individuals can relapse or become reinfected. The disease dynamics are therefore given by:

$$\begin{aligned}\text{Susceptible } (S) &\xrightarrow{\alpha} \text{Exposed } (E)\\ \text{Exposed } (E) &\xrightarrow{\beta} \text{Infected } (I)\\ \text{Infected } (I) &\xrightarrow{c} \text{Recovered } (R)\\ \text{Recovered } (R) &\xrightarrow{\gamma} \text{Susceptible } (S)\end{aligned}$$

In the context of a graph of individuals, α is the edge-wise fixed probability of a susceptible (S) individual (node) being exposed (E) to the disease from an infected (I) neighbor, β is the fixed probability of an exposed (E) individual (node) becoming infected (I), c is the probability of an infected (I) individual (node) voluntarily seeking and successfully completing treatment (R), and γ is the probability of a cured individual returning to the susceptible S stage. We assume that the treatment takes place in one time period, where a period represents the duration needed for a complete treatment regimen (~ half a year for TB). Here, α can be considered to be correlated to the infectiousness of a disease since that is the triggering probability of individuals converting from S to E state. Similarly, β can be considered as a proxy to the potency of the contagion, i.e. the ability of the contagion to develop into a full-blown infection.

Network Model: Individuals are part of a contact network, and infection spreads via the edges in the network. There are n individuals and $N(i)$ denotes neighbors of individual i in the network. The network structure (graph) is known from the beginning ($t = 0$). Each individual (node) in the network is in one of the health states $\{S, E, I, R\}$. Let s_i^t denote the state of individual i at time t. At $t = 0$, all nodes are deemed to be in the S (healthy) state. At $t = 0^+$, k% of the nodes are infected with the contagion.

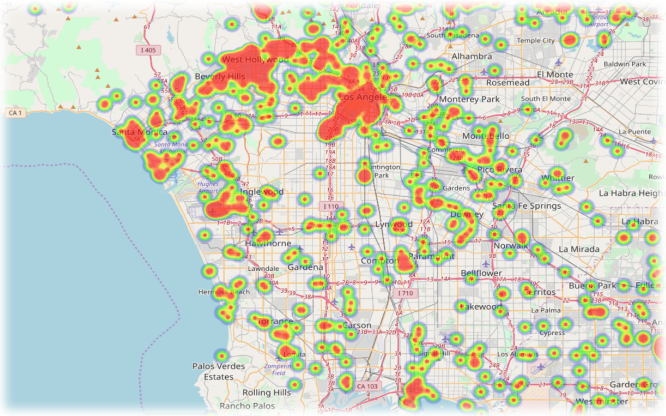

Figure 2: Flu outbreak in LA area

The actual probability of an individual undergoing a change in the health state is given by:

$$T = \begin{array}{c} \\ S \\ E \\ I \\ R \end{array} \begin{array}{c} \begin{array}{cccc} S & E & I & R \end{array} \\ \begin{bmatrix} q_j & 1-q_j & 0 & 0 \\ 0 & 1-\beta & \beta & 0 \\ 0 & 0 & 1-c & c \\ \gamma & 0 & 0 & 1-\gamma \end{bmatrix} \end{array},$$

$$\text{and } q_j = (1-\alpha)^{|\{k \in N(j) \mid s_k^t = I\}|}$$

The rows denote the from state and the columns denote the to state. The transition probabilities follow the disease dynamics described earlier. In particular, q_j captures the probability that node j does not become exposed from his infected neighbors $\{k \in N(j) \mid s_k^t = I\}$. We assume E individuals do not seek treatment voluntarily since their disease is latent. For model simplicity, we assume S individuals cannot directly transition directly to I or R state. This is not an extreme assumption for many diseases, where the overall transition durations can be much longer than the round length. This model is inspired from the model described by Bhattacharya et al. [1].

3.2 Visualization

In this section, we visualize what the actual disease progression looks in the data that we have. Therefore; given a set of disease parameters, we observe how the actual disease progression looks in a particular location.

The necessity for visualization is due to the different goals of studies that result in various interpretations of the data. For the scope of this work, we do not try to provide suggestions for decisions, but provide a method to visualize forecasting that will aid decision makers' in actionable results.

The visualization will provide information about where the outbreaks are happening in a way that will assist a decision maker to focus resources in high risk areas. Forecasting will determine where new outbreaks are more likely to happen.

The progression can vary drastically with respect to the parameters. Figure 2 visualize flu out breaks for the following parameters: $\alpha = 0.01$, $\beta = 0.1$, $c = 0.3$, $\gamma = 0.5$. This figure shows only the progression of only the infected nodes. The idea is to optimize the parameters of the disease model to simulate the progression to the actual disease spread, collected from activity data. So, our goal is to achieve a visualization which is as close to the actual disease progression as possible. As an example, we can predict the flu season this year based on what it looked like in past years and then optimize the α, β, c, γ values to the values which most closely mimic flu progression.

3.3 Optimization and Prediction

The goal in in this phase is to find the optimal values for our disease model parameters in order to predict the health states of the individuals in a population. To achieve this goal, we introduced three main approaches which are discussed in the following sections.

3.3.1 Grid Search. The traditional way of performing optimization is grid search which is an exhaustive searching through the parameters space of a learning algorithm with respect to an error function. Then the fitted model should be evaluated which is mostly done by cross-validation. We define the scope and the initial values for the parameters in this model and then fit the model using the available data from ground truth and social media.

We perform the grid search within the disease parameters space and check which values conform closest to the ground truth. We use a loss function to measure each candidate model distance to the ground truth.

3.3.2 Deep Learning. Deep neural networks have shown promising results for various clinical prediction tasks such as diagnosis, mortality prediction, predicting duration of stay in hospital [6][8]. Therefore, we leverage deep learning for training our model and predicting the infection spread.

(1) Normal LSTM model: By definition, clinical medical data consist of multi-variate time series of observations and more recent data has higher impact on the current status of the population's health. In particular, LSTM models are designed for analysis of time series data. In LSTM model, time is considered as an important feature, which controls the data in the model. We calculate α, β, c, and γ parameters distances with the actual data and compare at various acceptance thresholds depending on the disease type. The accepted parameters are reported as the optimal values. However, we can modify the normal LSTM model to improve the accuracy of our approach.
(2) Staged LSTM+encoder-decoder model: We model each season/year as a separate LSTM. The output of the (i-1)th LSTM is fed into an auto-encoder architecture and becomes an input to the ith LSTM. This architecture can be defined for each year which means auto-encoder steps can be defined at season level. Then we can make a matrix of auto-encoders and LSTMs to model the parameters. The goal of this model is to incorporate seasonal data and also predict at different level of granularity.

3.3.3 Transfer Learning. The problem with the previous two models is that disease parameters are dependent on the network at some level, whereas they are disease model and they should be network agnostic. For a single network model, this has no impact on the prediction and the result. However, the model is not generalizable to other networks which means complete re-training required for any new network. It is contradicting with the goal of introducing Disease models since α, β, c, and γ for a disease should be the same in every network. Thus, we need to define another method to learn the true value of these parameters, so that we can simply transfer it to a new network. The challenge is that it is hard to decouple α, β, c, γ from the network characteristics. To address this challenge we use the notion of a two-stage ensemble transfer learning model. First, we need to abstract the network information to a few distributions. Then, we can perform some sort of expectation–maximization(EM) method to extract the true α, β, c, and γ values. Finally, given the α, β, c, and γ values, we transfer this disease model to a new network. Then we incorporate the new network information and predict the outcome. For example, we transfer flu disease model from Los Angeles network to San Fransisco network. This process is not trivial and requires further investigation. The generalization offered by this model makes the model valuable and practical. We will focus on this general model in our future work.

4 EXPERIMENTS

To evaluate our proposed model, we considered flu infection spread in Los Angeles. Our goal is to predict the flu outbreaks in 2018 using previous years data. We used this case as a proof of concept to show that our model can provide valuable information to health workers.

4.1 Setup

We implemented the model shown in figure 1. We collected data from social media which is discussed in details in the next section. We trained our model for flu disease and optimized the parameters using LSTM. We compared the results with the with actual 2018 flu numbers collected from activity data, CDC data and Google trends.

4.2 Datasets

To create realistic contact networks, we collected data from Twitter, Flickr, and Google trends. The goal of the data collection is to generate a realistic human network, that is a network of nodes simulating how, when and where people interact. Therefore, we can model a virus spread on a realistic network. We also defined a ground truth of infected individuals to be able to perform parameter tuning for our model and validating the results.

For the realistic human network we collected geo-tagged tweets and photos. Total collected location nodes are 526,000 globally out of which 174,573 are unique from 2011 to 2017.

For our ground truth we use data from Google Trends API using flu related keywords and geo-location based flu related tweets. That could include queries like flu symptoms, flu remedies and related terms.

It has already been validated that Google Trends models after real flu outbreaks fairly well [7]. We also evaluated how well our tweet data performs and agrees with Google trends and official outbreak numbers as reported by data.gov.

For Google trends we collect data every 24 hours and we are able to get the search results for flu at 30 minute basis. For tweet data we collected stream data continuously of generic keywords as well as flu related keywords.

4.3 Results

The predicted values for flu disease model for flu outbreaks in 2018 are: $\alpha = 0.15$, $\beta = 0.8$, $c = 0.5$, and $\gamma = 0.5$. The infection prediction loss is equal to 7.64 which shows our approach did a fairly well job in predicting these values.

5 CONCLUSIONS

In this paper, we proposed a new approach to predict infection spread in a particular location. Our model provides health workers with valuable information about the health status of people as fast as possible. It also eliminates the need for collecting information from different health centers to calculate the infection trend and locate the high risk areas. Using live data as collected from Twitter and Google Trends helps in providing forecasting information to decision makers in the shortest time possible. This model is applicable to all infectious diseases and locations. We evaluated our approach using LSTM learning on flu spread in Los Angeles area. We created the realistic human network using Twitter and Flicker data and applied LSTM on this network. Results showed that our approach is fast and can predict infection spread with high accuracy. For the future work, we focus on generalization of our approach to make it transferable to wide range of networks.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Ferrara and Dr. Morstatter for providing valuable insights.

The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions. The work is supported by the National Natural Science Foundation of China under Grant No.: 61273304 and Young Scientists' Support Program (http://www.nnsf.cn/youngscientists).

REFERENCES

[1] Biswarup Bhattacharya, Han Ching Ou, Arunesh Sinha, Sze-Chuan Suen, Bistra Dilkina, and Milind Tambe. 2018. TRACE: Algorithmic ACTS for Preventing the Spread of Recurrent Infectious Diseases on Networks.

[2] Lauren N Carroll, Alan P Au, Landon Todd Detwiler, Tsung-chieh Fu, Ian S Painter, and Neil F Abernethy. 2014. Visualization and analytics tools for infectious disease epidemiology: a systematic review. *Journal of biomedical informatics* 51 (2014), 287–298.

[3] Lauren E Charles-Smith, Tera L Reynolds, Mark A Cameron, Mike Conway, Eric HY Lau, Jennifer M Olsen, Julie A Pavlin, Mika Shigematsu, Laura C Streichert, Katie J Suda, et al. 2015. Using social media for actionable disease surveillance and outbreak management: a systematic literature review. *PloS one* 10, 10 (2015), e0139701.

[4] Nicholas A Christakis and James H Fowler. 2010. Social network sensors for early detection of contagious outbreaks. *PloS one* 5, 9 (2010), e12948.

[5] Clark C Freifeld, Kenneth D Mandl, Ben Y Reis, and John S Brownstein. 2008. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *Journal of the American Medical Informatics Association* 15, 2 (2008), 150–157.

[6] Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, and Aram Galstyan. 2017. Multitask Learning and Benchmarking with Clinical Time Series Data. *ArXiv e-prints*, Article arXiv:1703.07771 (March 2017), arXiv:1703.07771 pages. arXiv:stat.ML/1703.07771

[7] Min Kang, Haojie Zhong, Jianfeng He, Shannon Rutherford, and Fen Yang. 2013. Using Google Trends for Influenza Surveillance in South China. *PLOS ONE* 8 (01 2013), 1–6. https://doi.org/10.1371/journal.pone.0055205

[8] Zachary Chase Lipton, David C. Kale, Charles Elkan, and Randall C. Wetzel. 2015. Learning to Diagnose with LSTM Recurrent Neural Networks. *CoRR* abs/1511.03677 (2015). arXiv:1511.03677 http://arxiv.org/abs/1511.03677

[9] P Van den Driessche, M Li, and J Muldowney. 1999. Global stability of SEIRS models in epidemiology. *Canadian Applied Mathematics Quarterly* 7 (1999), 409–425.