Estimating Social Network Structure and Propagation Dynamics for an Infectious Disease

Louis Kim[1,2], Mark Abramson[1], Kimon Drakopoulos[2], Stephan Kolitz[1], and Asu Ozdaglar[2]

[1] Draper Laboratory, Cambridge, MA 02139, USA
{Lkim,kolitz,mabramson}@draper.com
[2] Massachusetts Institute of Technology, Cambridge, MA 02139, USA
{Kimondr,asuman}@mit.edu

Abstract. The ability to learn network structure characteristics and disease dynamic parameters improves the predictive power of epidemic models, the understanding of disease propagation processes and the development of efficient curing and vaccination policies. This paper presents a parameter estimation method that learns network characteristics and disease dynamics from our estimated infection curve. We apply the method to data collected during the 2009 H1N1 epidemic and show that the best-fit model, among a family of graphs, admits a scale-free network. This finding implies that random vaccination alone will not efficiently halt the spread of influenza, and instead vaccination and contact-reduction programs should exploit the special network structure.

Keywords: network topology, disease dynamics, parameter estimation.

1 Introduction

Many diseases spread through human populations via contact between infective individuals (those carrying the disease) and susceptible individuals (those who do not have the disease yet, but can catch it) [1]. These contacts form a social network, along which disease is transmitted. Therefore, it has long been recognized that the structure of social networks plays an important role in understanding and analyzing the dynamics of disease propagation [2]. In this paper, we present an algorithm to estimate the structure of the underlying social network and the dynamics of an infectious disease. Better understanding the social network and transmission parameters will help public officials devise better strategies to prevent the spread of disease.

Many previous studies of disease propagation assume that populations are "fully mixed," meaning that an infective individual is equally likely to spread the disease to any other susceptible member of the population to which he belongs [3]. In the same line of work, Larson et al. enriched the aforementioned models by incorporating different types of agents [4]. In these works, the assumption of "full mixing" allows one to generate nonlinear differential equations to approximate the number of infective individuals as a function of time, from which the behavior of the epidemic can be studied. However, this assumption is clearly unrealistic, since most individuals have contact with only a small fraction of the total population in the real world.

W.G. Kennedy, N. Agarwal, and S.J. Yang (Eds.): SBP 2014, LNCS 8393, pp. 85–93, 2014.

Building on this insight, a number of authors have pursued theoretical work considering network implications. These models replace the "fully mixed" assumption of differential equation-based models with a population in which contacts are defined by a social network [2, 5-9]. Nonetheless, to the best of our knowledge, there hasn't been work on inferring network structure characteristics from epidemics data.

Another strand of work employed large-scale experiments to map real networks by using various sources of data such as email address books, censuses, surveys, and commercial databases. However, this often requires extensive amount of time and resources collecting, manipulating, and combining multiple data sources to capture large size networks and estimate connections within those networks [10-12, 24, 25]. In this work, we use much lower dimensional data, temporal infection data, to infer the network characteristics assuming the network follows scale-free or small-world model.

The contribution of our paper is twofold. Firstly, we develop a method to extract the network structure from the observed infection data. Specifically, our approach assumes a parameterized network model and disease process parameters to simulate expected infection curve. Then, the algorithm greedily searches for the parameter values that will generate an expected infection curve that best fits the estimated real infection curve. We demonstrate that our suggested algorithm, assuming a scale-free network, closely estimates the network characteristics and disease dynamic parameters for the 2009 H1N1 influenza pandemic. Our results confirm that the network-based model performs better in estimating the propagation dynamics for an infectious disease compared to the differential equation-based models with the "fully mixed" assumption.

Secondly, given this finding we shed light on designing efficient control policies: For example, due to high asymmetry in degree distribution for scale-free graphs, degree vaccination will be superior to random vaccination in stopping the spread of disease.

The outline of the paper is as follows. In Section 2 we introduce the disease spread model and the proposed estimation algorithm. In Section 3 we evaluate the performance of the algorithm and suggest efficient control policies to mitigate the disease spread. In Section 4 we provide our conclusions.

2 Methods

In this section, we describe a discrete-time stochastic multi-agent SIR model, and propose a corresponding inference algorithm to fit the disease dynamics generated by the model to real H1N1 infection data. The inference algorithm learns the social network structure and key disease spread parameters, such as the rate of infection and the rate of recovery, for a given infectious disease. This enables us to make useful predictions about the contact network structure and disease propagation for similar types of diseases and allows us to devise appropriate control strategies.

2.1 Data

We obtained data from state health departments, including the weekly percentage of all hospitalizations and outpatient visits resulting from influenza-like illness (%ILI) over the 2009-2010 flu seasons [14, 15]. Each point on the %ILI temporal curve represents

the percentage of the total number of hospitalizations and outpatient visits that are specific to H1N1. We also obtained total estimated cases of H1N1 and total estimated H1N1 related hospitalizations from the Center for Disease Control [15]. Using these data, we estimated the number of H1N1 infections for each week as follows:

Assuming that the flu wave first grows then declines after the peak of the infection while the number of non-H1N1 hospitalizations remains relatively stable, we estimate the number of non-H1N1 hospitalizations. Finally, the above allow us to estimate the number of H1N1 related hospitalizations during each period.

Given the number of H1N1 related hospitalizations during each period and the total estimated cases of H1N1, we can estimate the number of H1N1 infections at each period, assuming that the number of H1N1 infections are proportional to the number of H1N1 related hospitalizations [4]. (Refer to the Epidemic curve estimation section of [4] for more details.)

We used the estimation method described above to estimate the infection curve for the state of Massachusetts, in which the estimated true infection curve includes the effects of vaccines as administered. Figure 1 shows the estimated temporal infection curve and the temporal curve of vaccines as administered [17].

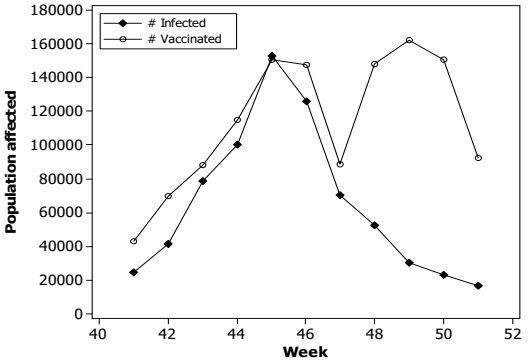

Fig. 1. The infection curve estimated from %ILI data and the number of vaccines administered during the observation period (October, 2009 – December, 2009) for Massachusetts

2.2 Disease Process

We employ a variation of susceptible-infective-removed (SIR) model first proposed by Kermack and McKendrick [13]. Individuals in the network, represented by nodes, are assigned one of the three states: the susceptible state (S) in which individuals are not infected but could become infected, the infective state (I) in which individuals are currently carrying the disease and can spread the disease to susceptible individuals upon contact, and the recovered state (R) in which individuals have either recovered from the disease and have immunity or have died. Edges connecting nodes in the network indicate contacts between individuals – contacts may occur through conversation between friends, co-workers, family members, etc. Alternatively, contacts can also occur between two strangers passing by chance. When a contact occurs between

an infective individual and a susceptible individual, the susceptible will become infective with probability β. Each infective individual recovers from the disease and becomes immune with probability δ after a period of time, a week in our case.

There exists a population of individuals,V connected by a graph $G = (V, E)$. We define $X_i(t) \in \{S, I, R\}$ to be the state of individual $i \in V$ at time t. And let $\eta_i(t) = \sum_{(j \setminus i) \in E} \mathbb{I}\, X_j(t) = I$, where $\mathbb{I}$ is the indicator function, denote the number of infected neighbors the individual i has at time t. Then given that the individual i is susceptible at time t, he will become infected at time t+1 with the following probability:

$$\mathbb{P}(X_i(t+1) = I | X_i(t) = S) = 1 - (1 - \beta)^{\eta_i(t)} \quad (1)$$

since with probability $(1 - \beta)^{\eta_i(t)}$ all infection attempts fail. Given that the individual i is infected at time t, he will recover at time t+1 with the probability:

$$\mathbb{P}(X_i(t+1) = R | X_i(t) = I) = \delta. \quad (2)$$

We assumed independence in infection attempts between neighbors. We also assume that if a susceptible individual is vaccinated, then he or she will immediately become immune to the disease and the individual's state will change to recovered state. Given a network and set of initial infections, the disease propagation process can be simulated according to the described probabilities.

2.3 Estimation Algorithm

The estimation algorithm uses the disease process described above to simulate infections. The simulated results are compared to the real H1N1 infection data, and we optimize over the network and disease spread parameters to obtain a best-fit simulated curve. The purpose of the algorithm is to find network characteristics, such as degree distribution, and disease spread parameters, β and δ that will help us make useful predictions about the network and how the disease spreads within the community.

Many real-world social networks such as citation networks, internet and router topologies, sexual contact networks are expected to exhibit small-world or scale-free properties [9-12, 18-20]. We tested both small-world and scale-free networks in our algorithm for the contact network.

Input

The inputs to the algorithm include:

- A parameterized disease spread network structure, where nodes represent people and undirected edges represent contacts between people. In our simulations, the network structure is assumed to be either scale-free or small-world, though the algorithm could be applied to other network structures.
- Initial values of the network parameters, the average degree, k_0 and, for the small-world network structure, p_0, the probability of a long-range contact.
- Initial values of the disease process parameters β_0 and δ_0.
- Real temporal infection data to fit the model generated infection dynamics.
- Data on vaccines administered (if administered).

Output

- The algorithm outputs a simulated expected infection curve, which fits the real data as closely as possible, and the network and infection parameters used to generate the expected infection curve.

Procedure

Begin with the given initial values of the social network and disease spread parameters: k_0, p_0, β_0 and δ_0. Let Δk, Δp, $\Delta\beta$ and $\Delta\delta$ be the amounts by which k, p, β and δ are changed at each step in the optimization. Let $\hat{e}_i$ and e_i each denote the number of infections for the real infection curve and the estimated expected infection curve at time i, respectively. We define the error, E, between the simulated expected infection curve and the true infection curve as:

$$E = \sum_{i=1}^{\max.period} |\hat{e}_i - e_i| \tag{3}$$

Repeat the following steps until the error can no longer be reduced by changes to the parameters (we define the optimal output parameters as k^*, p^*, β^* and δ^*):

1. Given k_0, p_0, β_0 and δ_0, search in all possible directions to find a direction that improves E. That is, evaluate E at all possible combinations of k, p, β and δ, where $k \in \{k_0, k_0 + \Delta k, k_0 - \Delta k\}$, $p \in \{p_0, p_0 + \Delta p, p_0 - \Delta p\}$, $\beta \in \{\beta_0, \beta_0 + \Delta\beta, \beta_0 - \Delta\beta\}$ and $\delta \in \{\delta_0, \delta_0 + \Delta\delta, \delta_0 - \Delta\delta\}$. Evaluate E by doing the following for each set of parameters:
   (a) Generate the network according to the given network type (small-world or scale-free) and network characteristics (k, the average degree for a scale-free network; k, the average degree and, p, the short-cut probability for a small-world network).
   (b) Simulate R realizations of the disease process. For each realization, initialize the disease simulation infection by assigning N_I nodes to the infected states, where N_I is the number of people infected at the beginning of the observation period in the data. We assume that the initial infected nodes are selected uniformly at random from among all the nodes.[1] Update the disease states for each time period, according to the disease process parameters and the vaccine administration data (We assume that those who receive vaccines at each time period are chosen uniformly at random).
   (c) Generate an expected infection curve by averaging the number of infected individuals at each time period over the R realizations of the disease process.
   (d) Calculate E.
2. Determine which search direction resulted in the minimum error. Update k_0, p_0, β_0 and δ_0 to the values of k, p, β and δ that achieved the lowest sum of residuals. The algorithm is summarized as a flow chart in Figure 2.

[1] Commonly used assumption in epidemic simulation. [25]

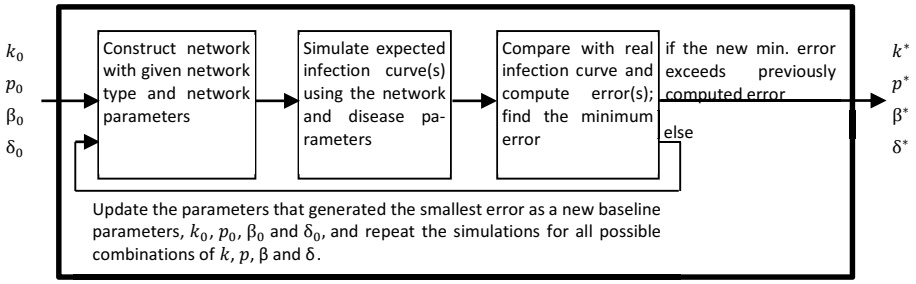

Fig. 2. Flow diagram showing the estimation algorithm details

3 Results

We have applied our algorithm to data from the 2009 H1N1 outbreak to demonstrate how our algorithm finds realistic network and infection parameters that can approximate the dynamics of an infectious disease. We scaled down the population size by a factor of 10,000 uniformly at random in order to reduce the computation time for infection simulation. In our simulations, we set `R`, the number of realizations per set of parameters, equal to 1,000. This effects how well the simulated curve approximate the expected curve. We began our search with relatively large values of Δk, Δp, $\Delta\beta$ and $\Delta\delta$ (changes in average degree, long-range connection probability, infection probability and recovery probability, respectively) and then manually decreased them as the sums of residuals began to converge. Specifically, initially $\Delta k = 10$, Δp, $\Delta\beta$, $\Delta\delta = 0.1$. We narrowed the search by reducing Δk to 1 and Δp, $\Delta\beta$ and $\Delta\delta$ to 0.01.

3.1 Estimation Algorithm Results

Figure 3 shows the resulting infection curves generated by the algorithm, compared to the estimated infection curve from data and from using differential equations with "fully-mixed" assumption. In addition to the error measure described above, we used the difference in total expected number of infections,

$$\left|\sum_{i=1}^{max.\ period} \hat{e}_i - \sum_{i=1}^{max.\ period} e_i\right| \quad (4)$$

and the difference in peak number of infections to compare the curves:

$$|\hat{e}_w - e_w|, \text{ where } w \text{ represents the time period of infection peak} \quad (5)$$

For the small-world network, the estimated parameter values were 8 for average degree, 0 for short-cut probability, .2 for infection probability (β), and .35 for recovery probability (δ). The error measured was close to 25 infections, which is 35% of total number of infections. Compared with the data-generated infection curve, the simulated infection curve for the small-world network had an 8% lower expected total number of infections and a 43% lower expected peak infections. Overall, the small-world network model did not provide a good fit to the estimated infection curve from data.

On the other hand, the best-fit infection curve generated using a scale-free network fits the data well. The estimated parameter values were 2 for average degree (k), .43 for infection probability (β), and .62 for recovery probability (δ). Measured error was about 9 infections, constituting 12% of the total infections. Compared to the estimated infection curve from data, we measured a 1.3% difference in the expected total number of infections and a 2.1% difference in the expected peak infections. This result is a significant improvement over the estimation under the "fully-mixed" assumption, which had a measured error of 26 infections (36% of total infections), a 25.7% difference in the expected total number of infections, and a 10% difference in the expected peak infections.

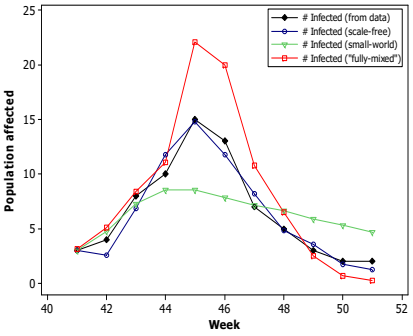

Fig. 3. Best fit curves generated by the algorithm using small-world network (green) and scale-free network (blue) compared to the estimated infection curve from data (black) and from using differential equations with "fully-mixed" assumption (red).

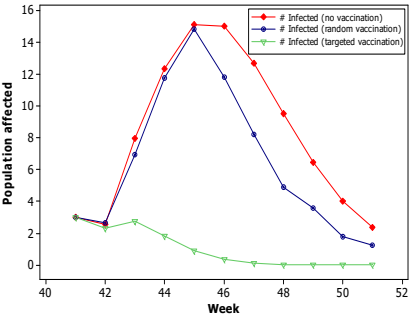

Fig. 4. A simulated curve that closely reflects the estimated infection curve from data with random vaccination scheme (blue) is compared to simulated curves with no vaccination (red) and degree-based targeted vaccination (green).

3.2 Control

Understanding the network structure and disease dynamics facilitates the adoption of efficient control measures to contain or stop the propagation of the disease on the network. Many studies have shown that, since scale-free networks have some nodes

with a very large number of connections compared to the average degree, targeting those high-degree nodes to vaccinate will effectively reduce the propagation of the disease [21-23].

Now that our algorithm has successfully learned contact network and disease dynamics parameters, we can use the model to study the effect of different control methods. Figure 4 compares the disease processes with no vaccination, random vaccination, and targeted vaccination (where we selectively vaccinate those individuals with high degree). These results validate the claim that vaccinating high-degree nodes with very large connections is effective in stopping the disease propagation. Randomly vaccinating individuals reduced the expected number of infections by about 22%, whereas targeting highly connected nodes for vaccination reduced the expected number of infections by around 88%.

4 Conclusions

Understanding the network structure and the disease dynamics on the network has important implications both for refining epidemic models and for devising necessary control measures in order to effectively utilize resources to prevent the spread of disease. The spread of infection is often complex to analyze due to the lack of information about the contact network on which it occurs. This paper showcases a methodology to learn network and disease propagation parameters of an infectious disease, H1N1 influenza. The findings for H1N1 give us useful insight into the infection dynamics of similar diseases and assist in analyzing effect of different vaccination policies. We hope that this study will benefit future efforts in infection prevention.

References

1. Newman, M.E.J.: The spread of epidemic disease on networks. Phys. Rev. E 66, 16128 (2002)
2. Moore, C., Newman, M.E.J.: Epidemics and percolation in small-world networks. Phys. Rev. E 61, 5678–5682 (2000)
3. Anderson, R.M., May, R.M.: Infectious diseases of humans. Oxford University Press, Oxford (1991)
4. Larson, R.C., Teytelman, A.: Modeling the effects of H1N1 influenza vaccine distribution in the United States. Value in Health 15, 158–166 (2012)
5. Pastor-Satorras, R., Vespignani, A.: Epidemic spreading in scale-free networks. Phys. Rev. Lett. 86, 3200–3203 (2001)
6. Kuperman, M., Abramson, G.: Small world effect in an epidemiological model. Phys. Rev. Lett. 86, 2909–2912 (2001)
7. Keeling, M.J., Eames, K.T.D.: Networks and epidemic models. J. R. Soc. Interface 2, 295–307 (2005)
8. Eubank, S.: Network based models of infectious disease spread. Jpn. J. Infect. Dis. 58, S9–S13 (2005)
9. Liljeros, F., Edling, C.R., Åmaral, L.A.N., Stanley, H.E., Aberg, Y.: The web of human sexual contacts. Nature 411, 907–908 (2001)

10. Salathé, M., Kazandjieva, M., Lee, J.W., Levis, P., Feldman, M.W., Jones, J.H.: A high-resolution human contact network for infectious disease transmission. Proc. Natl Acad. Sci. USA 107, 22020–22025 (2010)
11. Dong, W., Heller, K., Pentland, A.(S.): Modeling Infection with Multi-agent Dynamics. In: Yang, S.J., Greenberg, A.M., Endsley, M. (eds.) SBP 2012. LNCS, vol. 7227, pp. 172–179. Springer, Heidelberg (2012)
12. Newman, M.E.J., Forrest, S., Balthrop, J.: Email networks and the spread of computer viruses. Phys. Rev. E 66, 035101 (2002)
13. Kermack, W., McKendrick, A.: A contribution to the mathematical theory of epidemics. Proc. R. Soc. A 115, 700–721 (1927)
14. Seasonal influenza (flu). Center for Disease Control and Prevention (2010), http://www.cdc.gov/flu/weekly/
15. CDC estimates of 2009 H1N1 influenza cases, hospitalizations and deaths in the United States, April 2009 – March 13, 2010. Center for Disease Control and Prevention (2010), http://www.cdc.gov/h1n1flu/estimates/April_March_13.htm
16. Weekly influenza update, May 27, 2010. Massachusetts Department of Public Health (2010), http://ma-publichealth.typepad.com/files/weekly_report_05_27_10.pdf
17. Table and graph of 2009 H1N1 influenza vaccine doses allocated, ordered, and shipped by project area. Center for Disease Control and Prevention (2010), http://www.cdc.gov/h1n1flu/vaccination/vaccinesupply.htm
18. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships of the internet topology. Comput. Commun. Rev. 29, 251–262 (1999)
19. de Solla Price, D.J.: Networks of scientific papers. Science 149, 510–515 (1965)
20. Newman, M.E.J.: Networks: an introduction. Oxford University Press (2010)
21. Albert, R., Jeong, H., Barabasi, A.-L.: Error and attach tolerance of complex network. Nature 406, 378–382 (2000)
22. Pastor-Satorras, R., Vespignani, A.: Immunization of complex networks. Phys. Rev. E 65, 036104 (2002)
23. Dezsõ, Z., Barabási, A.-L.: Halting viruses in scale-free networks. Phys. Rev. E 65, 055103(R) (2002)
24. Barrett, C.L., Beckman, R.J., Khan, M., Kumar, V.A., Marathe, M.V., Stretz, P.E., Dutta, T., Lewis, B.: Generation and analysis of large synthetic social contact networks. In: Rossetti, M.D., Hill, R.R., Johansson, B., Dunkin, A., Ingalls, R.G. (eds.) Proceedings of the 2009 Winter Simulation Conference. IEEE Press, New York (2009)
25. Bisset, K., Chen, J., Feng, X., Kumar, V.A., Marathe, M.: EpiFast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In: Proceedings of the 23rd International Conference on Supercomputing (ICS), pp. 430–439 (2009)