

Modeling Infection with Multi-agent Dynamics

Wen Dong¹, Katherine Heller², and Alex “Sandy” Pentland¹

¹MIT Media Laboratory
{wdong, sandy}@media.mit.edu

²Department of Brain and Cognitive Sciences, MIT
kheller@gmail.com

Abstract. Developing the ability to comprehensively study infections in small populations enables us to improve epidemic models and better advise individuals about potential risks to their health. We currently have a limited understanding of how infections spread within a small population because it has been difficult to closely track an infection within a complete community. The paper presents data closely tracking the spread of an infection centered on a student dormitory, collected by leveraging the residents’ use of cellular phones. The data are based on daily symptom surveys taken over a period of four months and proximity tracking through cellular phones. We demonstrate that using a Bayesian, discrete-time multi-agent model of infection to model real-world symptom reports and proximity tracking records gives us important insights about infections in small populations.

Keywords: human dynamics; living lab; stochastic process; multi-agent modeling.

1 Introduction

Modeling contagions in social networks can help us facilitate the spread of valuable ideas and prevent disease. However, because closely tracking proximity and contagion in an entire community over a substantial period of time was previously impossible, modeling efforts have focused on large populations. As a result, little could be said about how an individual can gain exposure to good contagion and avoid bad contagion through his or her immediate social network. This paper describes how a “common” cold spread through a student residence hall community, with information based on daily surveys of symptoms for four months and tracking the locations and proximities of the students every six minutes through their cell phones. This paper also reports how infection occurred – and how infection could have been avoided – based on fitting the susceptible-infectious-susceptible (SIS) epidemic model to symptoms and proximity observations. It combines epidemic models and pervasive sensor data to give individually-tailored suggestions about local contagion, and also demonstrates the necessity of extending the epidemic model to individual-level interactions.

Epidemiologists agree on a framework for describing epidemic dynamics – people in a population can express different epidemic states, and change their states according to certain events. Computing event rates requires only knowledge about the overall population at the present time. The susceptible-infectious-recovered (SIR) model, for example, divides the population into susceptible, infectious, and recovered sub-populations (or "compartments"). A susceptible person will be infected at a rate proportional to how likely the susceptible person is to make contact with an infected disease carrier, and an infected person will recover and gain lifetime immunity at a constant rate. Other compartmental models include the susceptible-infectious-susceptible (SIS) model for the common cold, in which infectious people become susceptible again once recovered, and the susceptible-exposed-infectious-recovered (SEIR) model, in which infected carriers experience an "exposed" period before they become infectious.

However, the availability of new data and computational power has driven model improvements, refining compartmental models that assume homogeneous compartments and temporal dynamics, leading to the development of the Epidemiological Simulation System (EpiSimS) that takes land use into account [1], and more recently simulations based on the tracking of face-to-face interactions in different communities [2345].

These simulations all show evidence in favor of an epidemic dynamics framework, and against the assumption of homogeneous relationships and homogeneous temporal dynamics. Using these kinds of algorithms with real-world symptom reports and proximity data could offer a much better understanding of how infection actually transfers from individual to individual, allowing for personalized contagion recommendations.

To understand the infection dynamics in a community at the individual level, we use the data collected in the Social Evolution experiment [6], part of which tracked "common cold" symptoms in a student residence hall from January 2009 to April 2009. The study monitored more than 80% of the residents of the undergraduate residence hall used in the Social Evolution experiment, through their cell phones from October 2008 to May 2009, taking daily surveys and tracking their locations, proximities and phone calls. This residence hall housed approximately 30 freshmen, 20 sophomores, 10 juniors, 10 seniors and 10 graduate student tutors. Researchers conducted monthly surveys on various social relationships, health-related issues, and status and political issues. They captured the locations and proximities of the students by instructing the cell phones to scan nearby Wi-Fi access points and Bluetooth devices every 6 minutes. They then collected the latitudes and longitudes of the Wi-Fi access points and the demographic data of the students. The data are protected by MIT COUIS and related laws.

This paper makes the following contributions to the field of human behavior modeling: It is among the first to discuss the spread of flu symptoms, tracked daily with cellphone-conducted surveys over an entire community. It is also among the first to model the spread of flu symptoms by looking at proximity tracked by cell phones, paired with a repository of other cellphone-conducted surveys about activity, status, and demographics. Lastly, this paper introduces a multi-agent model that is compatible with compartmental epidemic models and can infer who infected whom and how to avoid catching the flu. The large quantity of behavioral data generated from pervasive computing technology provides the details necessary to shift social sciences research

Table 1. Probability of catching symptom = $1 - \alpha \times (-\beta \times \text{number of contacts with symptom})$, R^2 and p .

Symptom	α	β	R^2	p
Runny nose	1.013	0.024	0.52	0.04
Sadness	0.991	0.016	0.63	0.13
Stress	1.001	0.035	0.85	0.005
Nausea	0.993	0.006	0.94	0.11

from the level of large populations to individuals, and to enable social scientists to give more personalized advice.

The rest of the paper is organized as follows: In section 2 we describe the structure of face-to-face contact in the residence hall community, and the sensor data that captures this structure. In section 3 we introduce a Bayesian, multi-agent model, related to the Markov jump process, that not only simulates contagion but also makes inferences from observations. In section 4 we demonstrate that we can effectively predict new cases of symptoms, identify cases of symptoms even if students do not report them, and determine the students and contacts that are most critical for symptom-spreading. Hence, we show that the multi-agent model captures how symptoms of the common cold and the flu spread in a student dormitory community.

2 Contagion in Social Evolution Experiment

In the Social Evolution experiment, we offered students \$1 per day from 01/08/2009 through 04/25/2009 to answer surveys about contracting the flu, regarding the following specific symptoms: (1) runny nose, nasal congestion, and sneezing; (2) nausea, vomiting, and diarrhea; (3) frequent stress; (4) sadness and depression; and (5) fever. Altogether, 65 residents out of 84 answered the flu surveys, each of whom answered for half of the surveyed period. The correlation between stress and sadness is 0.39, while the correlations between other pairs are about 0.10.

The symptom self-reporting in the Social Evolution data seems to be compatible with what the epidemic model would indicate: symptoms other than runny nose are probabilistically dependent on that student’s friendship network. The durations of symptoms were about two days, and fit the exponential distribution well ($p \approx 0.6$ in Kolmogorov hypothesis testing). The chance of reporting a symptom is about 0.01, and each individual had a 0.006~0.035 increased chance of reporting a symptom for each additional friend with the same symptom (**Table 1**). These parameters are useful for epidemic simulation in the residence hall network, and for setting the initial values of fitting an epidemic model to real-world symptom observations and sensor data. The symptom surveys show some repeated infections, several clustered infections, the persistence of infections in larger clusters, and the persistence of infections caused by individuals who took longer to recover.

In this data, a student with a symptom had 3-10 times higher odds of seeing his friends with the same symptom (again, except for runny nose). As such, it makes sense

to fit the time-tested infection model with real-world data of symptom reports and proximity observations, and infer how friends infect one another through their contacts. In order to determine whether the higher odds could somehow be due to chance, we conducted the following permutation test to reject the null hypothesis that "the friendship network is unrelated to symptoms," and we can reject that null hypothesis with $p < 0.05$. The permutation test shuffles the mapping between the students and the nodes in the friendship network and estimates the probability distribution of the number of friends with the same symptom among all possible shuffling. If friendship networks were not related to the timing of when a student exhibits a symptom, then all mappings between the students and the nodes would be equally likely, and the number of friends with the same symptom would take the more likely values.

3 Modeling Infection Dynamics

In this section, we propose a discrete-time stochastic multi-agent SIS model, along with a corresponding inference algorithm to fit this multi-agent model to real-world data on proximity and symptom reporting. The inference algorithm does three things. First, it learns the parameters of the multi-agent model, such as rate of infection and rate of recovery. Second, it estimates the likelihood that an individual was infectious from the contact he had with other students, and from whether those others reported symptoms when the individual's symptom report is not available. Finally, it enables us to make useful predictions about contracting infections within the community in general.

Discrete-time stochastic multi-agent SIS model to fit real-world infection dynamics:

- Input:
 - A dynamic network, $\{G_t = (N, E_t): t\}$, where nodes representing people, bi-directional edges $E_t = \{(n_1, n_2): n_1 \text{ is near } n_2 \text{ at time } t\}$ representing "near-by" relation, and
 - Hyperparameters which provide prior information about: α — the probability that infectious persons outside of the network makes a susceptible person within the network infectious, β — the probability that an infectious person within the network makes a susceptible nearby person infectious, and γ — the probability that an infectious person becomes susceptible. The above variables are all assumed to be distributed according to beta distributions defined by these given hyperparameters.
 - Hyperparameters which define the prior probability of observing various symptoms depending on whether or not a person currently has a cold.
- Output: a matrix structure $\{X_{n,t}, Y_{n,t}: n, t\}$ indexed by time t and node n . The state $X_{n,t}$ of node n at time t is either 0 (susceptible) or 1 (infected). The symptom $Y_{n,t}$ of node n at time t is probabilistically dependent on the state of node n at time t .
- Procedure:
 - Initialize all parameters using their prior distributions, and assume that all people are susceptible at time $t = 1$.

- For each subsequent time $t + 1 = 2, \dots, T$ we assume the following generative model:
 - An infectious person becomes susceptible with probability γ , according to a Bernoulli distribution. If the Bernoulli trial is a success (the infectious person is now susceptible), $X_{n,t+1}$ is set, deterministically, accordingly, and the resulting symptoms $Y_{n,t+1}$ are set stochastically, from their probability distribution, conditioned on $X_{n,t+1}$.
 - Infectious persons within and outside of the network contribute to turning a susceptible person infectious, and the contributions happen independently:
 - Person n becomes infectious via contact with another infectious person in their network at time t . Each infectious contact, as specified by G_t , infects n with probability β , according to a Bernoulli distribution.
 - Person n is infected by someone outside the network, with probability α , according to a Bernoulli distribution.
- Set $X_{n,t+1}$ accordingly if any of the above Bernoulli trials is a success (a susceptible person is now infectious). Also set $Y_{n,t+1}$ stochastically, from its probability distribution, conditioned on $X_{n,t+1}$.

The probability of seeing a state sequence/matrix $\{X_{n,t}: n, t\}$ is therefore

$$\begin{aligned}
 & P(\{X_{n,t}: n, t\}, \alpha, \beta, \gamma) \\
 &= P(\alpha)P(\beta)P(\gamma) \prod_n P(X_{n,1}) \prod_{t,n} P(X_{n,t+1}|(X_{n',t}), \alpha, \beta, \gamma) \\
 &= P(\alpha)P(\beta)P(\gamma) \prod_{t,n} \gamma^{1X_{n,t}=1 \cdot 1X_{n,t+1}=0} \cdot (1 - \gamma)^{1X_{n,t}=1 \cdot 1X_{n,t+1}=0} \\
 &\quad \cdot \left(\alpha + \beta \cdot \sum_{(n',n,t) \in E} X_{n',t} \right)^{1X_{n,t}=0 \cdot 1X_{n,t+1}=1} \\
 &\quad \cdot \left(1 - \alpha - \beta \cdot \sum_{(n',n,t) \in E} X_{n',t} \right)^{1X_{n,t}=0 \cdot 1X_{n,t+1}=0}
 \end{aligned}$$

We employ a Gibbs sampler to iteratively sample infectious/susceptible state sequences, sample events conditioned on state sequences, and sample parameters. This provides an algorithm for performing inference in the above generative model. We can infer values of states $\{X_{n,t}: n, t\}$, and even missing values in symptoms $\{Y_{n,t}: n, t\}$, conditioned on the values of $\{Y_{n,t}: (n, t) \in obs.\}$ which we observe, and the interaction network $\{G_t: t\}$. An in depth description of our model and inference algorithm, and further discussion can be found in [7].

The SIS model describes infection dynamics in which the infection doesn't confer long-lasting immunity, and so an individual becomes susceptible again once recovered. The common cold has this infection characteristic.

4 Experimental Result

In this section we model the contagion which existed in the residence hall community. We estimate, at the community level, the parameters of susceptible-infectious-susceptible (SIS) infection dynamics. At the individual level, we describe the results of using the Gibbs sampling algorithm to fit the discrete-time multi-agent SIS infection dynamics to symptom observations.

We took several steps to calibrate the performances of the multi-agent model and support vector classifier on synthetic data. First, we synthesized 50 time series – each 128 days long – from the Bluetooth proximity pattern in the Social Evolution data and different parameterizations. Then, we randomly removed the infectious/susceptible data from 10% of the population, added noise to the remaining data in each time series, and averaged the performances on inferring the held-out data corresponding to each method and parameterization.

We ran Gibbs samplers for 10,000 iterations, got rid of the initial 1000 burn-in iterations, and treated the remaining 9000 iterations as samples from the posterior distribution. We trained the support vector classifier from another 1000-day time series synthesized using the right parameterization, and used the number of infectious contacts yesterday, today, and tomorrow as a feature. We assigned different weights to the “infected” class and the “susceptible” class to balance the true prediction rate and the false prediction rate.

All methods can easily identify 20% of infectious cases in the missing data with little error, but the model-based method using our dynamic multi-agent system consistently performs better than the support vector classifier. Less noise in symptom observation and in the individuals’ contact networks significantly improves the performance of inferring missing data, as shown through the ROC (receiver operating characteristic) curves in the left panel of **Fig. 1**. An ROC curve indicates better performance if it correctly predicts more positive cases and incorrectly predicts fewer negative cases, or equivalently if it is closer to the top-left corner, or it has the larger area below.

The support vector classifier performs worse – especially in identifying the isolated infectious cases in the missing data – because it assumes that its cases are i.i.d (identical and independently distributed) and because including the temporal structure of epidemic dynamics into the features is not an easy task. The support vector classifier also assumes that we either already have enough training data or can synthesize training data. This assumption generally cannot be satisfied for the kinds of problems we are interested in here.

In order to infer the latent common cold time series that best fits the multi-agent SIS model from dynamical Bluetooth proximity information and symptom self-report in the Social Evolution data using our Gibbs sampler, we extracted the hour-by-hour proximity snapshot over the 107 days we were monitoring symptoms and interpolated the hourly symptom report as the submitted daily symptom report. We assumed that the symptoms are probabilistically independent given the common cold state. We ran the Gibbs sampler for 10,000 iterations, removed the first 1000 burn-in iterations, and took the rest as samples of the posterior probability distribution of common cold states conditioned on symptom self-reports.

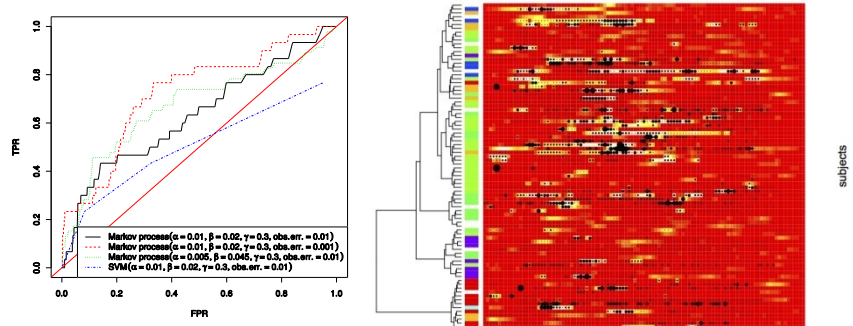


Fig. 1. (Left) Less observation error (obs.err.=0.001) and better knowledge about network ($\beta = 0.045$) lead to better trade-off between true positive rate (TPR) and false positive rate (FPR). The support vector classifier has worse trade-off between TPR and FPR than the multi-agent Markov model. (Right) An agent-based model can infer common cold states, and captures infection from symptom self-reports and proximity network. Sizes of black dots represent the number of symptoms reported, ranging from zero symptoms to all symptoms, and no black dot means no self-report.

The right panel of **Fig. 1** shows the (marginal) likelihood of the daily common-cold states of individuals. Rows in this heat map are indexed by subjects, arranged so that friends go together, and are placed side by side with a dendrogram that organizes friends hierarchically into groups according to the distance between the individuals and groups. Different colors on the leaves of the dendrogram represent different living sectors in the student dorm. Columns in this heat map are indexed by date in 2009. Brightness of a heat-map entry indicates the likelihood of being infectious. The brighter a cell is, the more likely it is that the corresponding subject is infectious on the corresponding day. Sizes of black dots represent the number of reported symptoms, ranging from zero symptoms to all symptoms. When a black dot doesn't exist on the corresponding table entry, the corresponding person didn't answer the survey on the corresponding day.

This heat map shows clusters of common cold happenings. When interpersonal proximities happened in larger social clusters, symptom clusters lasted longer and involved more people. A study of the heat map also tells us what the Gibbs sampler does in fitting the multi-agent SIS model to the symptom report: a subject often submitted flu-symptom surveys daily when he was in a "susceptible" state, but would forget to submit surveys when he was in the "infectious" state. The Gibbs sampler will nonetheless say that he was infectious for these days, because he was in the infectious state before and after, an infectious state normally lasts four days, and many of his contacts were in the infectious state as well. A subject sometimes reported symptoms when none of his friends did in the time frame. The Gibbs sampler will say the he was in the susceptible state, because the duration of the symptom reports didn't agree with the typical duration of a common cold, and because his symptom report was isolated in his contact network.

The inferred infectious state from symptom reports and hourly proximity networks normally lasts four days, but could be as long as two weeks. A student often caught a cold 2 ~ 3 times from the beginning of January to the end of April. The bi-weekly

searches of the keyword “flu” from January 2009 to April 2009 in Boston – as reported by Google Trends – explains 30% of variance in the number of (aggregated) bi-weekly common cold cases inferred by the Gibbs sampler, and network size explains another 10%.

The timing of different symptoms with regard to the inferred common cold cases follows interesting patterns. Stress and sadness normally began three days before the onset of a stretch of infectious state, and lasted two weeks. Runny nose and coughing began zero to two days before the onset of a symptom report and ended in about seven days, and they have similar density distributions. Fever normally occurred on the second day after the onset of a stretch of infectious state, and lasted for about two days. Nausea often happened four days before the onset of reaching an infectious state, then disappeared and reappeared again at the onset.

5 Conclusions

The study of infection in a small population has important implications both for refining epidemic models and for advising individuals about their health. The spread of infection in this context is poorly understood because of the difficulty in closely tracking infection in a complete community. This paper showcases the spread of an infection centered on a student dormitory, based on daily symptom surveys over a period of four months and on proximity tracking through resident cellular phones. It also demonstrates that fitting a discrete-time multi-agent model of infection with real-world symptom self-reports and proximity observations give us useful insight in infection paths and infection prevention.

6 References

1. S. Eubank, H. Guclu, V. Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429:180–4, 2004.
2. L. Isella, J. Stehle, A. Barrat, C. Cattuto, J. Pinton, and W. Van den Broeck. What’s in a crowd? analysis of face-to-face behavioral networks. *J Theor Biol*, 271:166–180, 2010.
3. M. Salathe, M. Kazandjieva, J. Lee, P. Levis, M. Feldman, and J. Jones. A high-resolution human contact network for infectious disease transmission. *Proc Natl Acad Sci (USA)*, 107:22020–22025, 2010.
4. L. Hufnagel, D. Brockmann, and T. Geisel. Forecast and control of epidemics in a globalized world. *Proc Natl Acad Sci USA*, 101:15124–9, 2004.[18]
5. J. Stehle, N. et. al., “Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees,” *BMC Medicine*, 9(1):87, 2011.
6. W. Dong, B. Lepri, and A. Pentland. Modeling the coevolution of behaviors and social relationships using mobile phone data. *Proc ACM MUM 2011*.
7. W. Dong, K. Heller and A. Pentland. Modeling Infection with Multi-agent Dynamics. arXiv 1201.xxxx [cs.MA, cs.SI].