

## Event-Based Structural Change Detection in Urban-Scale Contact Network

Yuan Bai,<sup>1,2</sup> Bo Yang,<sup>1,2</sup> Rosalind M Eggo,<sup>3</sup> Zhanwei Du<sup>4</sup>

<sup>1</sup> College of Computer Science and Technology, Jilin University, Changchun, 130012, CHN

<sup>2</sup> Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, CHN

<sup>3</sup> London School of Hygiene & Tropical Medicine, London, WC1E 7HT, UK

<sup>4</sup> Department of Integrative Biology, University of Texas at Austin, Austin, 78705, USA

### Abstract

The detection of structural changes is an important task in analyzing network evolution, especially for interactions between people, that may be driven by external events. Existing work relies on snapshot data and misses out some key functions of networks. Here, we study contact network evolution where no snapshot data are available. In spite of the challenge, this study demonstrates how contact networks can be used to predict and control infectious disease epidemics. We first model structural changes in contact networks during the 2009 influenza pandemic in Hong Kong, and then present a probabilistic framework to address it, aiming to answer when and how the underlying structure changes, utilizing multiple data sources including demographic data, and epidemic surveillance data. The efficacy and public health utility of the method are demonstrated using both synthetic and real data.

### Introduction

Current methods for dynamical structure detection in network evolution mainly rely on snapshot data and miss out key functions of networks (Peel and Clauset 2015; Berlingerio et al. 2013). However, in many real-world applications, events occur on dynamical networks rather than structural snapshots (Peel and Clauset 2015; Berlingerio et al. 2013). For instance, during an epidemic of influenza, we usually do not know who interacted with whom, or how infection was transmitted by contacts, but can only observe are new cases reported in surveillance data. Therefore, snapshot-based methods cannot work in these scenarios. Whereas, it is still essential to understand how contact networks change during epidemics, because it will provide better understanding of epidemic patterns, and aid planning and evaluation of different intervention strategies.

In this paper, we focus on the task of event-based dynamical structure detection. By taking contact network as a case study, the specific problem to be addressed here is: **can we accurately detect when and how contact networks, which describe individuals' routine contact patterns, change over time based on reported cases?** The task is challenging for the following reasons: (1) contact networks of population are unobserved. Moreover, it is hard to obtain sufficient and high-quality data that directly and explicitly describe

human contact behaviors, especially on a large scale; (2) contact networks are heterogeneous with respect to different social and demographic scenarios; (3) contact networks often change at different stages of disease spread.

In this work, we will address these challenges by the following methods:

**Modeling:** use a mathematical model to characterize dynamic contact networks while considering their spatial-temporal features. We present a dynamic contact network metapopulation model, where nodes represent groups of homogeneous sub-populations (e.g., people of the same age or within the same locations), and weighted links represent the likelihood of contact between or within sub-population. Spatial heterogeneity is also incorporated into this model by considering different types of social settings mainly including household, school, workplace, and general community.

**Mining:** use novel data streams to infer when and how unobserved contact networks change during the 2009 pandemic in Hong Kong. We collate multiple data sources relevant to disease transmission, including demographic data, epidemiological data, and disease surveillance. We then propose a mining framework, by integrating a probabilistic inference model and an epidemic model, to detect when and how contact networks evolve.

## Models and Inference Methods

### Metapopulation-based Epidemic Model

We focus on epidemic spread by close person-to-person contact, i.e. as pandemic influenza is transmitted. An age-structured Susceptible-Infectious-Recovered (SIR) transmission model is used to describe the epidemic spreading on a contact network at a metapopulation level, in which all individuals are categorized to three states: *Susceptible* ( $S$ ), *Infected* ( $I$ ), and *Recovered* ( $R$ ).

In the SIR model, the population is partitioned into  $n$  age groups, and  $N_i$  denotes the total number of individuals in age group  $i$ . Each infected individual in age group  $i$  changes their state to recovered with probability  $\gamma_i$ . For simplicity, we consider age independent recovery rate, so,  $\gamma_i = \gamma$ .  $1/\gamma$  denotes the duration of an infectious period. Let  $S_i(t)$ ,  $I_i(t)$  and  $R_i(t)$  denote the numbers of susceptible, infected, and recovered individuals in age group  $i$  at time  $t$ .  $S_i(t) \lambda_i(t)$  denotes the newly infected cases in age group  $i$  in the inter-

val  $[t, t + \Delta t]$ .  $\lambda_i(t)$  is formulated as:

$$\lambda_i(t) = \Delta t C \rho_i \sum_{j=1}^n \beta_j M_{ij} \frac{I_j(t)}{N_j} \quad (1)$$

where  $\rho_i$  is the susceptibility to infection of each individual in age group  $i$  and  $\beta_j$  is the probability of transmitting infection per effective contact for an infectious individual of group  $j$ .  $M_{ij}$  is the average frequency of effective contacts that an individual in age group  $i$  has with the individuals in age group  $j$ .  $C$  is a scaling factor and can be estimated by  $R_0$ , which is the average number of secondary cases that result from a single infected individual in an entirely susceptible population.

$\Delta t = 1/\gamma$  and  $Y_i(t)$  denotes the expected number of newly infected cases in age group  $i$  in the  $t$ -th generation (i.e. the infection period  $[t, t + 1/\gamma]$ ). The metapopulation based SIR epidemic model can be inferred as:

$$\begin{cases} Y_i(t) = \frac{1}{\gamma} C \rho_i \sum_{j=1}^n \beta_j M_{ij} \frac{Y_j(t-1)}{N_j} S_i(t) \\ S_i(t) = N_i - \sum_{\kappa=1}^{t-1} Y_i(\kappa) \end{cases} \quad (2)$$

Epidemiological parameters such as  $\rho_i$ ,  $\beta_i$ ,  $\gamma$  can be obtained from medical reports or serological tests. Next, we will discuss how to construct the contact network  $M$  to represent the frequency of effective contacts of different age groups.

### Dynamic Contact Network Model

Since infection transmission is not uniform across social settings due to the spatial and demographic heterogeneity of contacts (Cauchemez et al. 2004), we create a contact network  $M$  for different social settings based on demographic data. Here, we mainly focus on the social settings of households, schools, workplaces and general community contacts and simulate a virtual society of synthetic individuals (Fumanelli et al. 2012). Due to the lack of detailed information on how individuals interact in each unit of social setting, we assume that within a location, individuals mix homogeneously. The frequency of contacts within each of three social settings can be calculated as follows:

$$f_{ij}^Z = \begin{cases} \frac{1}{n_i^Z} \sum_{1 \leq l \leq N_i} \frac{z_j^l - \delta_{ij}}{\nu_Z^l - 1} & n_i^Z > 0, \nu_Z^l > 1 \\ 0 & \text{else} \end{cases} \quad (3)$$

$f_{ij}^Z$  is the frequency of contacts between age groups  $i$  and  $j$  within a social setting  $Z$ .  $N_i$  is the size of age group  $i$ .  $n_i^Z$  is the number of individuals in age group  $i$  with at least one contact in the setting  $Z$ . For each individual  $l$  in age group  $i$ ,  $\nu_Z^l$  denotes the number of individuals in setting  $Z$  that  $l$  belong to (e.g.,  $l$ 's household or school).  $z_j^l$  is the number of individuals in age group  $j$  in  $l$ 's setting  $Z$ .  $\delta_{ij}$  is the Kronecker delta function that allows an individual to be excluded from the set of his own contacts.

As individuals of different age groups mix randomly in general communities, such as shopping malls and leisure places, the frequency of contacts in the setting is simply proportional to the ratio of individuals by age group.

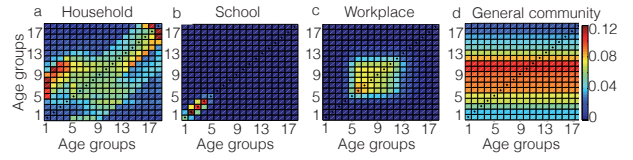


Figure 1: Four baseline contact matrices for household, school, workplace and general community.

Based on the contact frequency, the contact matrix (or the baseline) of each social setting  $Z$  is then constructed as  $M_{ij}^Z = q_i^Z f_{ij}^Z$ , where  $q_i^Z = n_i^Z / N_i$  is the probability of the individuals in group  $i$  have contacts with others. Fig.1 illustrates the baseline contact matrices of four social settings constructed using Hong Kong demographic data (www.census2011.gov.hk).

In order to obtain the contact network  $M$  in equation (2), we use  $\eta^Z$  to denote the fraction of effective contact within social setting  $Z$ . This value is responsible for disease spread, which will be affected by different factors. Then,  $M$  is defined as the linear combination of the baselines in different social settings:

$$M_{ij} = \sum_Z \eta^Z M_{ij}^Z \quad (4)$$

$\eta^Z$  is not constant and will be affected by external events over time. For example, during the period of an epidemic, the contact behaviors of human beings will be changed by their awareness to the outbreak and reactions to different intervention strategies such as closing schools or travel restriction. Then, the temporal feature of the contact network  $M$  can be characterized by the variation of  $\eta^Z$ . Correspondingly, the task of inferring the dynamic structure of contact from observed events can be transferred to estimate the coefficients  $\eta^Z$  from the time series of infected cases.

### Detecting Structural Changes in Contact Networks

We introduced a model to describe the structural changes in the contact network during an epidemic: the *contact-SIR* model. Next, we develop a probabilistic model to estimate the coefficients  $\eta^Z$  from surveillance data based on the proposed contact-SIR model.

We assume the available surveillance data consist of the reported cases of a specific disease, denoted by a matrix  $\mathbb{D} \in \mathbb{R}^{T \times n}$ ,  $t = 1, \dots, T$  and  $i = 1, \dots, n$ , with  $t$  indexing  $T$  infected periods and  $i$  indexing  $n$  age groups.  $D_i(t)$  denotes the number of newly infected cases in age group  $i$  at time  $t$ . The time  $T$  can be divided into some sliding windows with a fixed length  $\omega$ , which controls the granularity of the change points to be detected. The change-point detection can be used to estimate the times at which the contact pattern changes and the degree of changes. Small  $\omega$  give many (but possibly trivial) changes, while bigger  $\omega$  detects fewer changes. During the time window from the beginning to the end of a time interval, the structural changes in contacts are detected by calculating and identifying the fitting error of real surveillance data and simulated results provided

by the contact-SIR model. Let matrix  $\mathbb{Y} \in \mathbb{R}^{T \times n}$  denote the synthetic outbreak data generated by the contact-SIR model. Let  $\tau$  denote the last time of a time window, then the interval of time window can be denoted by  $[\tau - \omega + 1, \tau]$ . Note that one time unit in this window denotes one infection period (i.e.  $1/\gamma$ ). Let  $t_c$  denote a potential change point within the time window, the fitting error between the real data  $\mathbb{D}$  and the synthetic data  $\mathbb{Y}$  during the time window then can be measured by:

$$\varepsilon_{t_c} = \sum_{t=\tau-\omega+1}^{t_c-1} \sum_{i=1}^n \|D_i(t) - Y_i(t)\|_2^2 + \sum_{t=t_c}^{\tau} \sum_{i=1}^n \|D_i(t) - Y_i(t)\|_2^2 \quad (5)$$

Applying Eqs. (2) and (4) to (5), we have:

$$\begin{aligned} \varepsilon_{t_c} &= \sum_{t=\tau-\omega+1}^{t_c-1} \sum_{i=1}^n \left\| D_i(t) - \frac{1}{\gamma} C \rho_i \sum_{j=1}^n \beta_j M_{ij}^t \frac{Y_j(t-1)}{N_j} S_i(t) \right\|_2^2 \\ &\quad + \sum_{t=t_c}^{\tau} \sum_{i=1}^n \left\| D_i(t) - \frac{1}{\gamma} C \rho_i \sum_{j=1}^n \beta_j M_{ij}^t \frac{Y_j(t-1)}{N_j} S_i(t) \right\|_2^2 \\ &= \sum_{t=\tau-\omega+1}^{t_c-1} \sum_{i=1}^n \left\| D_i(t) - \frac{1}{\gamma} C \rho_i \sum_{j=1}^n \beta_j \left( \sum_Z \eta^{(t,Z)} M_{ij}^Z \right) \frac{Y_j(t-1)}{N_j} S_i(t) \right\|_2^2 \\ &\quad + \sum_{t=t_c}^{\tau} \sum_{i=1}^n \left\| D_i(t) - \frac{1}{\gamma} C \rho_i \sum_{j=1}^n \beta_j \left( \sum_Z \eta^{(t,Z)} M_{ij}^Z \right) \frac{Y_j(t-1)}{N_j} S_i(t) \right\|_2^2 \end{aligned}$$

The true position of change point within the time window is estimated. Our probabilistic model determines the likelihood of each time  $t_c$  being a change point. Let the probability density of fitting error  $\varepsilon_{t_c}$  be:

$$p(\varepsilon_{t_c} | \theta) = \sum_{t=\tau-\omega+1}^{\tau} \alpha_t \phi(\varepsilon_t | \theta_t) \quad (6)$$

where  $\alpha_t$  denotes the probability of time  $t$  being a change point. Each component  $\phi(\varepsilon_t | \theta_t)$  denotes the probability density of fitting error  $\varepsilon_t$ , with parameter  $\theta_t = (\mu_t, \sigma_t^2, \eta_{\neq}^{(t,Z)})$ . In the mixture model of (6),  $p(\varepsilon_{t_c} | \theta)$  is actually the expectation of  $\phi(\varepsilon_t | \theta_t)$  over all times within the time window.  $\mu_t$  is the expectation of  $\varepsilon_t$ , and  $\sigma_t^2$  is the variance of  $\varepsilon_t$ . As defined above,  $\eta^Z$  denotes the fractions of effective contact contributed by social settings  $Z$ . Here,  $\eta^{(t,Z)}$  means  $\eta^Z$  at time  $t$  and  $\eta_{\neq}^{(t,Z)}$  represents a vector of  $(\eta^{(t-1,Z)}, \eta^{(t,Z)})$ . Note that, if time  $t$  is a change point, it will divide the time window  $[\tau - \omega + 1, \tau]$  into two phases,  $\eta^{(t-1,Z)}$  works for the first phase  $[\tau - \omega + 1, t - 1]$ , and  $\eta^{(t,Z)}$  works for the second one  $[t, \tau]$ .

We assume  $\varepsilon_t$  follows a zero-mean Gaussian distribution, i.e.,  $\varepsilon_t \sim N(0, \sigma_t^2)$ .  $\sigma_t$  is deemed as a predefined parameter to control the sparsity of  $\alpha^t$  (i.e., the frequency of change points). Now to estimate  $\theta_t$  is to estimate  $\eta_{\neq}^{(t,Z)}$ . As mentioned, our objective is to minimize the fitting error between observed data and synthetic data reproduced by the SIR-contact model with temporal parameters  $\eta$ . That is,

$$\eta_{\neq}^{(t,Z)} = \arg \min_{\eta_{\neq}^{(t,Z)}} \varepsilon_t \quad (7)$$

Given  $\alpha_t$  in Eq.6, the optimization can be solved by the interior-point algorithm (Byrd, Gilbert, and Nocedal 2000).

Given  $\varepsilon_t$ , using the Expectation-Maximization (EM) algorithm,  $\alpha_t$  can be updated by the following rule:

$$\alpha_t \leftarrow \frac{\alpha_t \left[ \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left(-\frac{\varepsilon_t^2}{2\sigma_t^2}\right) \right]}{\sum_{t=\tau-\omega+1}^{\tau} \alpha_t \left[ \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left(-\frac{\varepsilon_t^2}{2\sigma_t^2}\right) \right]} \quad (8)$$

Given the probabilities  $\alpha$ , we can obtain the expectation of change point  $t_c$  and corresponding expectation of change significance  $\eta_{\neq}^Z$ , as follows:

$$\begin{cases} t_c = \sum_{t=\tau-\omega+1}^{\tau} \alpha_t t \\ \eta_{\neq}^{(t_c,Z)} = \sum_{t=\tau-\omega+1}^{\tau} \alpha_t \eta_{\neq}^{(t,Z)} \end{cases} \quad (9)$$

## Validations and Applications

We use both synthetic data and real data to test the proposed method. The tests with real-world data are also a demonstration of the potential applications of our method. Recently, (Yang et al. 2016) addressed the similar problem through a tensor deconvolution method. In the following validations, we also compare our method with the tensor-based method.

### Validations on Synthetic Data

We first generate synthetic surveillance data by simulating the outbreak of pandemic influenza H1N1 in Hong Kong in 2009. The main steps are summarized as follows. (1) According to the demographic data of Hong Kong, construct the baseline contact network in four social settings: households, schools, workplaces, and general community, as shown in Fig. 1. (2) Set the dynamic coefficient  $\eta^Z(t)$  for each setting. In this test, we set three change points, at time 10, time 20 and time 30, in time windows  $[6, 15]$ ,  $[16, 25]$  and  $[26, 35]$ , separately. Based on the settings and the coefficients, we obtain a dynamic contact network  $M$ . (3) Run the contact-SIR model on  $M$  to generate synthetic outbreak data ( $\gamma = 1/3$ ,  $R_0 = 1.5$ ,  $\beta_i = 1$ ,  $\rho_i = 2.6(0-19y)$ ,  $\rho_i = 1(\text{others})$ ) (Xia, Liu, and Cheung 2013; Wu et al. 2010). (4) Run the detection method, to detect the change points and change significance in terms of  $\eta^Z(t)$ . In the test, we set a sliding window with  $\omega = 10$ . Fig.2 shows the estimated values of  $\eta^Z$  for each setting by two methods. Compared with the tensor-based method, the results given by our method are closer to the simulated values.

### Fitting to Real-world Data

We apply our method to real-world data from the 2009 H1N1 pandemic spread in Hong Kong, to test whether our method can discover useful patterns that are consistent with the actual events. Epidemic surveillance data are provided by the Centre for Health Protection (CHP) of the Hong Kong Public Health Department (Xia, Liu, and Cheung 2013), which consist of confirmed infections during 66 infectious periods since the first case was identified on June 1st, 2009. Note that, there are no vaccine measures during the epidemic. We use our method to detect when and how the underlying structure changes. This time, we set a sliding window with  $\omega = 20$ .

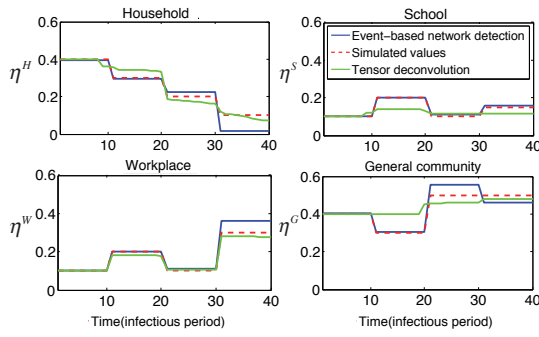


Figure 2: Test two methods with synthetic data.

Two significant change points for all four settings are detected (Fig.3(a)), which are near July 1st (change point 1) and August 25 (change point 2). Two significant change points for all four settings are detected (Fig.3(a)), which are near July 1st (change point 1) and August 25 (change point 2). This finding corresponds well to two types of events: school closing due to the start of the summer vacation, and school reopening, when schools return for the autumn term after the summer holiday. As reported, the government announced the closure of all primary schools, kindergartens on June 23 and reopened them after summer holiday on September 1. We find that during the period of school closing, the fraction of effective contacts in school began to reduce and the relative fraction in household began to increase. During the period of school reopening, the result is the opposite. Based on the estimated  $\eta^Z$ , the contact networks are constructed for the three stages separated by the two main change points (Fig.3(a)). With the detected dynamic network method, we reproduced the H1N1 outbreak in Hong Kong and the result is shown in Fig.3(b). As comparison, we also plot the simulation results obtained by the tensor deconvolution-based method and a static method. In the static method, we adopt the coefficients of  $\eta^H$  (0.31),  $\eta^S$  (0.24),  $\eta^W$  (0.16),  $\eta^G$  (0.29), as suggested by (Xia, Liu, and Cheung 2013). The pattern obtained by our method fits best to the observed data.

## Conclusion

When analyzing dynamic networks, a key task is to determine when and how their underlying structure changes with time, or, how networks transit from one relatively steady state to another relatively steady state. We introduced a new method to detect dynamic contact network changes based on observed disease transmission events. Our main contributions are two-fold: (1) raised the problem of event-based dynamical structure detection when network snapshots are not available; (2) proposed a framework to address this issue by combining multiple data sources in a flexible way. The efficacy and the applications of the method were shown using both synthetic and real data.

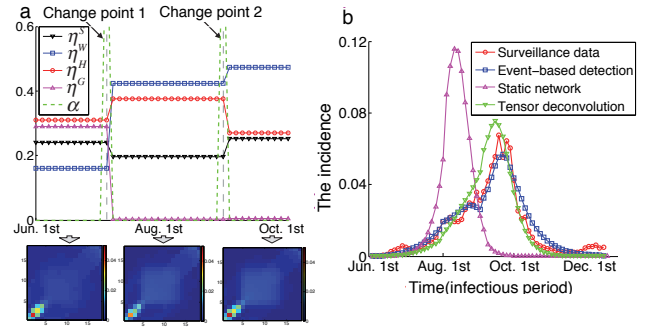


Figure 3: (a) The dynamics of social contact in four settings during 2009 Hong Kong H1N1 outbreak. Two primary time change points are detected, corresponding to the two peaks of  $\alpha$ -values, which reveal two events of schools closing and autumn term starting. (b) The real pattern of H1N1 outbreak and the simulations obtained by three methods.

## Acknowledgment

This work was supported in part by National Natural Science Foundation of China under grants 61572226,61373053.

## References

- Berlingerio, M.; Coscia, M.; Giannotti, F.; Monreale, A.; and Pedreschi, D. 2013. Evolving networks: Eras and turning points. *Intelligent Data Analysis* 17(1):27–48.
- Byrd, R. H.; Gilbert, J. C.; and Nocedal, J. 2000. A trust region method based on interior point techniques for nonlinear programming. *Mathematical Programming* 89(1):149–185.
- Cauchemez, S.; Carrat, F.; Viboud, C.; Valleron, A.; and Boelle, P. 2004. A bayesian mcmc approach to study transmission of influenza: application to household longitudinal data. *Statistics in medicine* 23(22):3469–3487.
- Fumanelli, L.; Ajelli, M.; Manfredi, P.; Vespignani, A.; and Merler, S. 2012. Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread. *Plos Computational Biology* 8(9):e1002673–e1002673.
- Peel, L., and Clauset, A. 2015. Detecting change points in the large-scale structure of evolving networks. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- Wu, J. T.; Cowling, B. J.; Lau, E. H. Y.; Ip, D. K. M.; Ho, L. M.; Tsang, T.; Chuang, S. K.; Leung, P. Y.; Lo, S. V.; and Liu, S. H. 2010. School closure and mitigation of pandemic (h1n1) 2009, hong kong. *Emerging Infectious Diseases* 16(3):538–41.
- Xia, S.; Liu, J.; and Cheung, W. 2013. Identifying the relative priorities of subpopulations for containing infectious disease spread. *Plos One* 8(6):e65271.
- Yang, B.; Pei, H.; Chen, H.; Liu, J.; and Xia, S. 2016. Characterizing and discovering spatiotemporal social contact patterns for healthcare. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.