Reality Mining with Mobile Data: Understanding the Impact of Network Structure on Propagation Dynamics

Yuanfang Chen†, Noel Crespi†, Lei Shu‡, and Gyu Myoung Lee§

†Institut Mines-Télécom, Télécom SudParis, Evry, 91000, France
†Department of Computer Science, Université Pierre et Marie CURIE, Paris, 75005, France
‡Guangdong University of Petrochemical Technology, Maoming, 525000, China
§Liverpool John Moores University, Liverpool, L33AF, UK
yuanfang_chen@ieee.org, noel.crespi@mines-telecom.fr, lei.shu@ieee.org, g.m.lee@ljmu.ac.uk

Abstract. Recent studies have increasingly turned to graph theory to model Realistic Contact Networks (RCNs) for characterizing propagation dynamics. Several of these studies have demonstrated that RCNs are best described as having exponential degree distributions. In this article, based on the mobile data gathered from in-vehicle wireless devices, we show that RCNs do not always have exponential degree distributions, especially in dynamic environments. On this basis, a model is designed to recognize the structure of networks. Based on the model, we investigate the impacts of network structure on disease dynamics that is an important empirical study to the propagation dynamics. The time-varying infected number R is the important parameter that is used to quantify the disease dynamics. In this study, the prediction accuracy for R is improved by utilizing realistic structural knowledge mined by our recognition model.

Keywords: Reality Mining; Mobile Data; Structural Knowledge; Propagation Dynamics

1 Introduction

In recent years, there have been increasing efforts to uncover, model, and understand propagation processes arising over a wide variety of networks, e.g., propagation of infectious diseases [4][3], propagation of information [10][23][9][15], and even propagation of computer viruses [8][5]. Observing a propagation process, and quantifying and predicting the dynamics of the propagation, are important for: (i) reducing the transmission rate of an infectious disease, (ii) decreasing the number of infected individuals during an epidemic, (iii) allocating public health resources and responding to public health events, (iv) acquiring timely and accurate information, (v) capturing a new behavior or a new development

tendency from the propagation of information/knowledge, and (vi) controlling the number of infected nodes with the propagation of computer viruses. And these propagation processes arise over a wide variety of networks. It is necessary to figure out the impacts of network structure on the propagation dynamics, and automatically recognize the network structure based on a recognition model.

As an important aspect of propagation dynamics [25], the quantification and prediction of disease dynamics during epidemics [30][31] are very important in allocating public health resources and in responding to public health events. Underestimating the impact of a disease can lead to an inadequate public health response, while overestimating can lead to the misallocation of limited public health resources. The time-varying infected number R[1] can be used to quantify the disease dynamics during an epidemic, and a wide range of methods have been proposed to estimate or predict the parameter R [21][27][28][1][11] with time-based or network-based models. However, the existing methods are based on Exponential Networks (ENs)[2]. Compared with the ENs, Realistic Contact Networks (RCNs) [2] contain realistic structural knowledge that is helpful to improve the prediction accuracy for disease dynamics during an epidemic.

In this article, based on the mobile data gathered from in-vehicle and handheld wireless devices, we show that RCNs do not always have exponential degree distributions. On this basis, a model is designed to recognize the structure of networks, for mining the knowledge of network structure. With the model, we investigate the impacts of network structure on propagation dynamics. As the important empirical study for the propagation dynamics, we investigate the impacts of network structure on disease dynamics, and the key parameter R is used to quantify the disease dynamics.

The scientific contributions of this article are shown as follows:

- We compare RCNs with ENs, and measure the differences between them in their network structures with precise measurements.
- A model is designed to recognize the structure of networks.
- Real surveillance data is used to evaluate the prediction performance for R. And realistic structural knowledge is used into the prediction, which is mined and acquired by our recognition model.

The achieved main results of this article are: (i) RCNs do not always have exponential degree distributions, (ii) the structural knowledge from RCNs is helpful to improve the prediction accuracy for propagation dynamics, and (iii) as the basic and important structural knowledge for networks, degree distribution, is effective to reflect the structure of a network, and to improve the accuracy of predicting for the infected number R.

The remainder of this article is organized as follows. Section 2 introduces the preparatory work and methods of carrying out our study. In Section 3, fitting results are shown and discussed in detail, and these results are about fitting

[1] R is defined as the number of infected cases during an epidemic over time.

[2] In this study, the network with exponential degree distribution is named as "Exponential Network".

the network structure of RCNs into exponential, normal, poisson and power-law distributions. Based on these results, in Section 4, a model is designed to recognize the structure of networks. In Section 5, we investigate the impacts of network structure on propagation dynamics. With the structural knowledge of respective networks, the prediction accuracy for R on the RCNs and ENs is measured respectively, and the prediction results for R are compared with real surveillance data. As the background of this study, Section 6 provides related work. This article is concluded in Section 7.

2 Methods

Two types of networks and the real surveillance data of a disease outbreak are used in our study. For evaluating the impacts of the structural knowledge about networks on propagation dynamics, extensive experiments for a knowledge-based Susceptible-Infected-Recovered (SIR) model [29] are run on these networks.

2.1 Networks

Two types of networks are used: (i) Exponential Networks, and (ii) Realistic Contact Networks from the real physical world.

Exponential Networks. It has recently been demonstrated that empirical contact networks are best described as having exponential degree distributions [2].

Through analyzing empirical contact networks [2] and based on the analysis and proof of literature [1], a Bansal Network (BN) is implemented and used as the EN. In the BN, each pair is generated using an algorithm of Bansal *et al.* [2] (Greedy Rewiring Algorithm (Alg. 1)).

The probability mass function (pmf) of BN's degree distribution meets Eq.(1).

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0, \end{cases} \tag{1}$$

where $x \in [0, \infty)$ is the degree of a node, and $\lambda > 0$ is the key parameter of an exponential distribution, which is called "rate parameter". This can be described as: $X \sim Exp(\lambda)$, which means the random variable X has an exponential distribution.

The nodes of BN are labeled $(1, ..., N)$, and an edge between two nodes indicates the presence of a transmission probability for a disease from one node to another. For example, there is a pair of nodes, i and j, $i \neq j$, the edge between them is $e_{\{i,j\}}$, and the transmission probability on the edge between i and j is given by $p_{\{i,j\}}$.

The input of Greedy Rewiring Algorithm is a connected and undirected network G, and the algorithm rewires edges until the degree distribution of the network becomes approximately exponential. In particular, the algorithm runs until the coefficient of variation ($\frac{standard\ deviation(sd)}{mathematical\ expectation(E[fx])}$) of the degree

distribution is less than 1 (this ensures an exponential distribution of network). The algorithm is described below and illustrated in Fig. 1.

Algorithm 1 Greedy rewiring algorithm

Input: A fully connected, undirected network G

1: select a random node i from the network G.
2: select a random edge $e_{\{i,j\}}$ from the network G such that the degree of node j is greater than one.
3: select a random edge $e_{\{j,m\}}$ from the network G, where the selected node m has the maximum probability of $k_m/\sum k_m$, and k_m means the degree of node m. Meanwhile $m \neq i$ and the node m is not the neighbor of node i.
4: If we find the appropriate node j and m, we remove the edge $e_{\{i,j\}}$ and add the edge $e_{\{i,m\}}$ to the network G.
5: The termination condition for re-building the network G is: $sd/E[fx] < 1$
   - sd is the standard deviation of the degree distribution
   - The degree distribution of network G is fitted into an exponential distribution fx
   - $E[fx]$ is the mathematical expectation of fx

Output: A network with exponential degree distribution

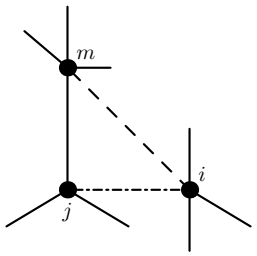


Fig. 1: Greedy rewiring process: node i is chosen at random, the edge $e_{\{i,j\}}$ is selected at random from the edges of node i, and the edge $e_{\{j,m\}}$ is selected at random from the edges of node j with probability proportion to the degree of node m. The edge $e_{\{i,j\}}$ (shown with a dotted line) is removed and the edge $e_{\{i,m\}}$ (shown with a dashed line) is added.

Realistic Contact Networks. Two RCNs from the real physical world are studied in this article.

Vehicle-based contact network (Fig. 2). There are 2483 nodes in this network with spatio-temporal GPS traces of vehicles, and the traces come from in-vehicle and GPS-enabled wireless devices. The network can be modelled as a dynamic graph G_t with the time-varying velocities of different traffic segments, and the velocities can be estimated using a combination of sources, including Automatic

Number Plate Recognition (ANPR) cameras, in-vehicle and GPS-enabled wireless devices and inductive loops built into road surfaces (a scenario is illustrated in Fig. 3).

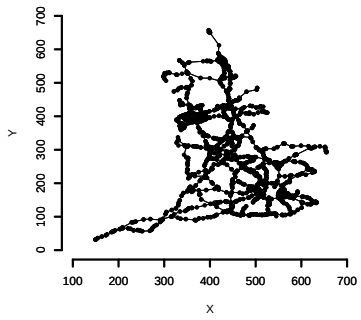


Fig. 2: Vehicle-based contact network. Using the coordinates of junctions of each traffic segment, the network can be built, where the black nodes are junctions, and the lines between these junctions are traffic segments that are with different traffic velocities.

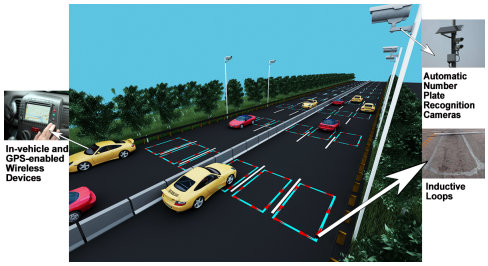


Fig. 3: A scenario of vehicle-based contact network. This network includes Automatic Number Plate Recognition (ANPR) cameras, in-vehicle and GPS-enabled wireless devices and inductive loops built into road surfaces. With this network, massive mobile data of different traffic segments can be gathered based on various sensor nodes and wireless devices. The mobile sensing data is gathered by the Highways Agency, in England.

The dynamic graph G_t can be described as follows. An undirected weighted graph $G_t = (V_t, E_t, W_t)$, where V_t is a set of n_t vertices with an online sequence of updates: (i) Delete($e_{\{i,j\}}$): delete the edge $e_{\{i,j\}}$ from E_t and corresponding vertices i and j from V_t; (ii) Insert($e_{\{i,j\}}$): insert the edge $e_{\{i,j\}}$ into E_t and

corresponding vertices i and j into V_t; (iii) Update($w_{\{i,j\}}$): update the weight $w_{\{i,j\}}$ related to the edge $e_{\{i,j\}}$ to W_t, and the weight $w_{\{i,j\}}$ is the velocity on the corresponding edge $e_{\{i,j\}}$. On the above (i), (ii) and (iii) basis, the graph G_t is updated, from $G_t = (V_t, E_t, W_t)$ to $G_{t+1} = (V_{t+1}, E_{t+1}, W_{t+1})$. It means that at different time points, with different velocities on different traffic segments, the transmission rates on these traffic segments are different. This vehicle-based contact network is time-varying.

Moreover, the data for this network is gathered from all motorways and 'A' roads managed by the Highways Agency, in England. The data provides average velocities and traffic flow information for 15-minute periods since April 2009 on these motorways and roads. The data includes these variables: (i) Segment ID. A unique alphanumeric segment id represents a segment from one junction (intersection) to another junction; (ii) Date. There is a date for each record; (iii) Time Period. There are 96 time periods, 0-95, with 15-minute intervals, in a day (1440 minutes); (iv) Average Velocity. The average velocity (km/h) of vehicles on a traffic segment within a given 15-minute time period; (v) Segment Length. The length of a traffic segment (km).

Human-based contact network. There are 942 nodes in this network. With the wireless communication devices held by volunteers of epidemic areas, the volunteers report new cases (confirmed and suspected cases), corresponding locations, and relationships between these cases, and then, these reported cases with corresponding locations can be used to build the human-based contact network (an example is shown in Fig. 4). During an epidemic, the network is time-varying along with the propagation of an infectious disease, with the order of time stamps of reports. As the vehicle-based contact network, the human-based contact network can be modelled as a dynamic graph G_t. However, the weight $w_{\{i,j\}}$ is the transmission probability ($p_{\{i,j\}}$) of a disease from vertex i to vertex j (on the corresponding edge $e_{\{i,j\}}$). For this network, there are four variables: (i) Case ID. A unique number indicates a case; (ii) Source ID. A source id indicates the source of infection for a case; (iii) Date. It is the date that a case is reported; (iv) Location. It indicates the coordinates (longitude and latitude) of a reported case.

2.2 Outbreak Data

The outbreak data of Ebola in West Africa from March 2014, is used as real surveillance data to evaluate the prediction performance for R on RCNs and ENs.

As a latest outbreak of disease, until February 15, 2015, Ebola Virus Disease (EVD. It is commonly known as "Ebola") has killed 9380 people, and the total cases have reached 23253. Researchers generally believe that from a 2-year-old boy of Guinea to his mother, sister and grandmother (a human-based contact network), Ebola rapidly spreads in West Africa, from March 2014.

The reported Ebola cases with time series and location information are gathered by the World Health Organization (WHO), as well as the ministries of health

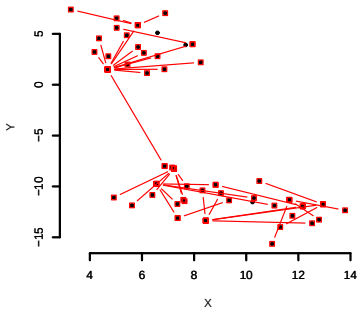


Fig. 4: An example of our human-based contact network. This example displays 50 cases and their relationships (contact), from three typical countries and seven regions of the Ebola outbreak in 2014. Three countries are: Guinea, Nigeria and Liberia. Seven regions are: Gueckedou, Macenta, Kissidougou, Conakry, Monrovia, Lagos and Port Harcourt. The black nodes of this network are cases (suspected and confirmed) and if there is an edge between two nodes, it means that there is contact between the individuals of the two cases.

of epidemic countries. And in this study, we select part of data from three typical outbreak countries, Guinea, Nigeria and Liberia. Guinea is the source of this outbreak and is with relatively high quantity of confirmed cases (2727, as of February 15, 2015), and Nigeria is far away from the source of the outbreak, and is with relatively low quantity of confirmed cases (19, as of February 15, 2015), and Liberia is close to the source of the outbreak, and is with high quantity of confirmed cases (3149, as of February 15, 2015). And seven regions of these three countries are: Gueckedou of Guinea, Macenta of Guinea, Kissidougou of Guinea, Conakry of Guinea, Monrovia of Liberia, Lagos of Nigeria, and Port Harcourt of Nigeria. And these variables are included in the outbreak data: (i) Case ID. A unique number indicates a case; (ii) Source ID. A source id indicates the source of infection for a case; (iii) Date. It is the date that a case is reported; (iv) Location. It indicates the coordinates (longitude and latitude) of a reported case.

2.3 Methods

A knowledge-based SIR model is used to evaluate the impacts of network structure on disease dynamics. As the results of this evaluation, the number of infected cases (infected number R) is calculated for each time period (different time periods have different network structures along with the propagation of a disease during an epidemic).

The SIR model is a model from epidemiology [13]. This model is developed to describe the propagation of an epidemic that occurs during a period of time.

The individuals of a contact network might be in three states: Susceptible (S), Infected (I) and Recovered (R). Susceptible individuals become infected at a given rate through contact with infected individuals. Infected individuals recover with a given rate and become recovered. The model is capable of showing the important parameter R which is measured to quantify the disease dynamics during an epidemic. The parameter R is the number of infected cases over time. In this study, we consider a knowledge-based SIR model with the knowledge of network structure.

Moreover, we consider different time periods (t), for our RCNs. For the vehicle-based contact network, the unit of time period is "15 minutes", and for the human-based contact network, the unit of time period is "day". And for comparing the impacts of different networks, the ratio $R_{A/B}$ is used to measure the different impacts of the network A and the network B. And for a network, the degree distribution is used to characterize and reflect the structure of the network.

3 Results and Analysis

To evaluate the impacts of network structure on disease dynamics, the basic and important structural knowledge of networks, degree distribution, is measured and compared for each network that is studied in this article.

In a network, the degree of a node is its most basic structural knowledge, and it indicates the number of adjacent edges of the node. The degree distribution is the probability distribution of these degrees over the network. It gives the overall structural information of the network. For a real-world network, the relationships between nodes are complex. The degree distribution is helpful to characterize and model a real-world network. On this basis, the structural knowledge of a complex network can be acquired and formulated. The formulated knowledge is effective for analyzing and solving network-related problems.

In this study, we analyze the degree distributions of RCNs in detail, by conducting maximum-likelihood fitting to fit the degree distributions of these networks into exponential, normal, poisson and power-law distributions [2], and calculating and comparing the estimated standard deviations and the estimated variance-covariance matrices of these fittings.

Vehicle-based contact network. Figure 5 illustrates the degree distribution of vehicle-based contact network.

From Fig. 5, we can observe that the degrees of nodes are not exponential distribution, in this real-world contact network. For figuring out the differences between them, the degree distribution of vehicle-based contact network and exponential, normal, poisson and power-law distributions, maximum-likelihood fitting is conducted to fit the degree distribution of vehicle-based contact network into exponential, normal, poisson and power-law distributions (Fig. 6), and then the estimated standard deviations and the estimated variance-covariance matrices of these fittings are measured to quantify "how many differences between two different degree distributions".

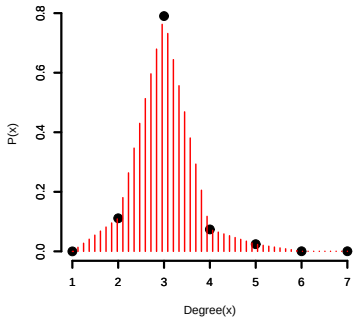


Fig. 5: Degree distribution of our vehicle-based contact network. There are 2483 nodes and 2500 edges in this network. The black spots are the probability distribution of nodes' degrees.

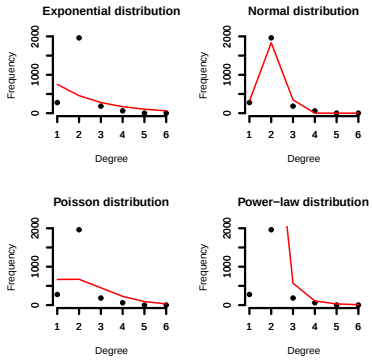


Fig. 6: Maximum-likelihood fitting of degree distributions. The degree distribution of vehicle-based contact network is fitted into exponential, normal, poisson and power-law distributions with maximum-likelihood fitting. The black spots display the probability distribution of nodes' degrees to the vehicle-based contact network, and the red lines are the corresponding fittings for exponential, normal, poisson and power-law distributions.

With these fittings that are shown in Fig. 6, corresponding parameter estimates can be calculated, for example, using maximum-likelihood fitting, the most likely value of parameter λ (rate parameter) is 0.4966, for the fitting with the exponential distribution. Corresponding estimated standard deviations and estimated variance-covariance matrices are measured, and these deviations and matrices are calculated by comparing with standard distributions that are with corresponding parameter estimates, for example, the exponential distribution with $\lambda = 0.4966$ is used as the standard distribution for the fitting with the

exponential distribution. These deviations and matrices show how many differences between two distributions. Moreover, the parameter estimates for different distributions from maximum-likelihood fitting, are listed as follows: (i) the rate parameter λ of exponential distribution is 0.4966, (ii) $\mu = 2.013693113$ and $\sigma = 0.539810394$ for the normal distribution, (iii) $\lambda = 2.013693$ for the poisson distribution, (iv) $xmin = 2$ and $\alpha = 5.785002$ for the power-law distribution.

Table 1 shows the estimated standard deviations and the estimated variance-covariance matrices of these fittings.

Table 1: Estimated standard deviations and estimated variance-covariance matrices for different fittings

Distribution	Standard deviation	Variance-covariance matrix		
Exponential	λ (rate parameter): 0.009965942		*rate parameter*	
		rate parameter	$9.932e-05$	
Normal	μ (mean): 0.010833103, σ (standard deviation (sd)): 0.007660161		*mean*	*sd*
		mean	0.0001173561	0.0000000000
		sd	0.0000000000	$5.867806e-05$
Poisson	λ (lambda): 0.02847792		*lambda*	
		lambda	0.000810992	
Power-law	$xmin + \alpha$: 0.006375843	NULL		

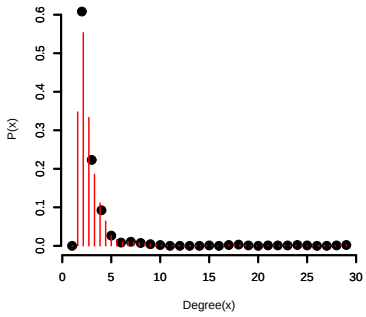
From the fitting results displayed in Fig. 6 and Tab. 1, the degree distribution of nodes for the vehicle-based contact network, is approximate to the power-law distribution with $xmin = 2$ and $\alpha = 5.785002$ and with the standard deviation 0.006375843.

Human-based contact network. Figure 7a illustrates the degree distribution of human-based contact network. On Fig. 7a basis, for figuring out the degree distribution of human-based contact network, maximum-likelihood fitting is conducted to fit the degree distribution of human-based contact network into exponential, normal, poisson and power-law distributions, and then the estimated standard deviations and the estimated variance-covariance matrices of these fittings are measured to quantify "how many differences between two different distributions". The results of fittings are illustrated in Fig. 7b.

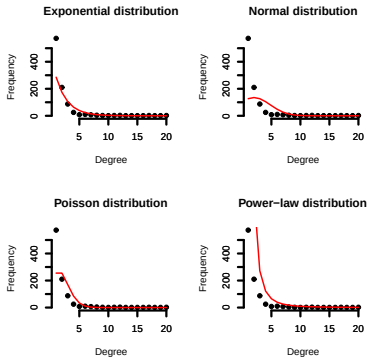
In Fig. 7, the results show that the degree distribution of human-based contact network is approximate to the exponential distribution with $\lambda = 0.50159915$.

The parameter estimates for different distributions from maximum-likelihood fitting are: (i) the rate parameter $\lambda = 0.50159915$ for the exponential distribution, (ii) $\mu = 1.99362380$ and $\sigma = 2.77914691$ for the normal distribution, (iii) $\lambda = 1.9936238$ for the poisson distribution, and (iv) $xmin = 2$ and $\alpha = 2.803973$ for the power-law distribution.

Table 2 shows the estimated standard deviations and the estimated variance-covariance matrices of these fittings.



(a) Degree distribution of our human-based contact network. There are 942 nodes and 938 edges in this network. The black spots are the probability distribution of nodes' degrees.



(b) Maximum-likelihood fitting of degree distributions. The degree distribution of human-based contact network is fitted into exponential, normal, poisson and power-law distributions with maximum-likelihood fitting. The black spots display the probability distribution of nodes' degrees to the human-based contact network, and the red lines are the corresponding fittings for exponential, normal, poisson and power-law distributions.

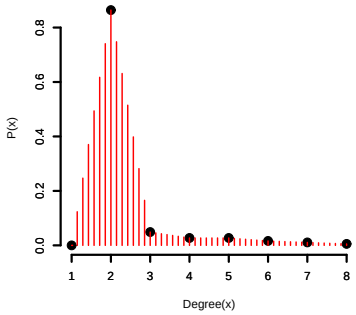
Fig. 7: Degree distribution and maximum-likelihood fitting for our human-based contact network.

Table 2: Estimated standard deviations and estimated variance-covariance matrices for different fittings

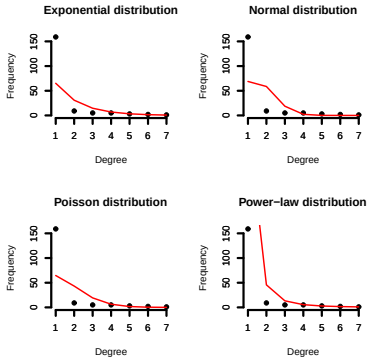
Distribution	Standard deviation	Variance-covariance matrix
Exponential	λ (rate parameter): 0.01635166	$rate\ parameter$: 2.673769e − 04
Normal	μ (mean): 0.09059760, σ (standard deviation (sd)): 0.06406218	$mean$: 0.008207925, 0.000000000; sd: 0.000000000, 0.004103963
Poisson	λ (lambda): 0.0460285	$lambda$: 0.002118623
Power-law	$xmin + \alpha$: 0.03831463	NULL

Comparing the estimated standard deviations and estimated variance-covariance matrices listed in Tab. 2, the minimum standard deviation for these fittings is 0.01635166. This minimum standard deviation is corresponding to the exponential distribution with the rate parameter $\lambda = 0.50159915$.

However, based on the descriptions for the networks that are studied in this article, the human-based contact network is time-varying along with the propagation of an infectious disease. As an example, the analysis results of the subnetwork that is with 96 time periods of August 26th, 2014[3], are shown in Fig. 8 and Tab. 3.



(a) Degree distribution for the subnetwork of human-based contact network. There are 96 time periods of August 26th, 2014 in this network. The black spots are the probability distribution of nodes' degrees.



(b) Maximum-likelihood fitting of degree distributions. The degree distribution for the subnetwork of human-based contact network is fitted into exponential, normal, poisson and power-law distributions with maximum-likelihood fitting. The black spots display the probability distribution of nodes' degrees to the subnetwork, and the red lines are the corresponding fittings for exponential, normal, poisson and power-law distributions.

Fig. 8: Degree distribution and maximum-likelihood fitting for the subnetwork of human-based contact network.

With the fittings for the subnetwork of human-based contact network, the parameter estimates for different distributions are: (i) the rate parameter $\lambda = 0.74796748$ for the exponential distribution, (ii) $\mu = 1.33695652$ and $\sigma = 1.00841216$ for the normal distribution, (iii) $\lambda = 1.33695652$ for the poisson distribution, and (iv) $xmin = 1$ and $\alpha = 3.041947$ for the power-law distribution.

Table 3 shows the estimated standard deviations and the estimated variance-covariance matrices of the fittings for the subnetwork.

[3] This subnetwork is obtained by a time-based sample. It is the contact network of this day, August 26th, 2014.

Table 3: Estimated standard deviations and estimated variance-covariance matrices for different fittings

Distribution	Standard deviation	Variance-covariance matrix
Exponential	λ (rate parameter): 0.05514089	*rate parameter*: 0.003040518
Normal	μ (mean): 0.07434113, σ (standard deviation (sd)): 0.05256712	*mean*: 0.005526604, 0.000000000; *sd*: 0.000000000, 0.002763302
Poisson	λ (lambda): 0.08524123	*lambda*: 0.007266068
Power-law	$xmin + \alpha$: 0.02865438	NULL

From the fitting results for the subnetwork of August 26th, 2014, which are listed in Tab. 3, the degree distribution of the subnetwork is approximate to the power-law distribution with $xmin = 1$ and $\alpha = 3.041947$.

With the above detailed analyses on the structure of networks, this fact can be observed: network structure is different to different networks, and is time-varying to dynamic networks.

4 Recognition Model of Network Structure

Because network structure is different to different networks, and the network structure is time-varying to dynamic networks, it is necessary to recognize the structure of a network, for analyzing the propagation dynamics on the network.

Our recognition model consists of: fitting, selection and parameter adjustment, and it can be formulated and described as follows:

- As the first step of model, the fitting is to fit the structure of a network into exponential, normal, poisson and power-law distributions with maximum-likelihood fitting, and the fitting calculates the parameter estimates and standard deviations to corresponding distributions. The parameter estimates and standard deviations to corresponding distributions, can be denoted as: (i) pe_{exp} and sd_{exp} for the exponential distribution, (ii) pe_{norm}, $sd_{\mu,norm}$ and $sd_{\sigma,norm}$ for the normal distribution, (iii) pe_{pois} and sd_{pois} for the poisson distribution, and (iv) pe_{pl} and sd_{pl} for the power-law distribution.
- And then, the selection is to select an approximate distribution by comparing the calculated standard deviations of four distributions. This step is denoted as:
  $\min\{sd_{exp}, sd_{norm} = \frac{sd_{\mu,norm}+sd_{\sigma,norm}}{2}, sd_{pois}, sd_{pl}\}$.
- Finally, our model uses the standard deviation of the selected approximate distribution to adjust the degree distribution function of the selected approximate distribution, and the selected approximate distribution is with the corresponding parameter estimate calculated by the fitting of first step.

Degree distribution functions and the detailed process of adjustment are introduced as follows:

(i) The degree distribution functions of exponential, normal, poisson and power-law distributions:

- The degree distribution function of exponential distribution is: $f(x;\lambda) = \lambda e^{-\lambda x}(x \geq 0)$.
- The degree distribution function of normal distribution is: $f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
- The degree distribution function of poisson distribution is: $f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$.
- The degree distribution function of power-law distribution is: $f(x;xmin,\alpha) = \frac{\alpha-1}{xmin}(\frac{x}{xmin})^{-\alpha}$.

(ii) The detailed process of adjustment:

Based on (i) above degree distribution functions, and (ii) the parameter estimates and standard deviations to corresponding distributions, the adjusted degree distribution functions can be obtained and these adjusted degree distribution functions reflect the structure of real networks. The adjusted degree distribution functions to corresponding distributions are listed in Eq.(2).

$$
\begin{aligned}
&f(x;(\lambda \pm sd_{exp})) = (\lambda \pm sd_{exp})e^{-(\lambda \pm sd_{exp})x}(x \geq 0),\\
&f(x;(\mu \pm sd_{\mu,norm},(\sigma \pm sd_{\sigma,norm})\\
&= \frac{1}{(\sigma \pm sd_{\sigma,norm})\sqrt{2\pi}} e^{-\frac{(x-(\mu \pm sd_{\mu,norm}))^2}{2(\sigma \pm sd_{\sigma,norm})^2}},\\
&f(x;(\lambda \pm sd_{pois})) = \frac{(\lambda \pm sd_{pois})^x e^{-(\lambda \pm sd_{pois})}}{x!},\\
&f(x;(xmin \pm sd_{pl}),(\alpha \pm sd_{pl}))\\
&= \frac{(\alpha \pm sd_{pl})-1}{(xmin \pm sd_{pl})}(\frac{x}{(xmin \pm sd_{pl})})^{-(\alpha \pm sd_{pl})}.
\end{aligned}
\tag{2}
$$

An example is provided to explain the adjustment. The human-based contact network is approximate to the exponential distribution with $\lambda = 0.50159915$, and the standard deviation from the rate parameter λ of this exponential distribution is 0.01635166, so the degree distribution function of this human-based contact network can be denoted as: $f(x;0.50159915 \pm 0.01635166) = (0.50159915 \pm 0.01635166)e^{-(0.50159915 \pm 0.01635166)x}(x \geq 0)$. And the degree distribution function can be used to reflect the network structure of this human-based contact network.

5 Evaluation

We investigate the impacts of network structure on propagation dynamics. With the structural knowledge of respective networks, the prediction accuracy for R on the RCNs and ENs is measured respectively, and the prediction results for R are compared with real surveillance data.

Knowledge-based SIR model. For a SIR model, the following differential equations represent this model:

$$\begin{aligned} \frac{dS}{dt} &= \delta R - \beta SI, \\ \frac{dI}{dt} &= \beta SI - \gamma I, \\ \frac{dR}{dt} &= \gamma I - \delta R, \end{aligned} \tag{3}$$

where β is the rate at which susceptible individuals contract the disease when exposed to infection, γ is the rate at which infected individuals recover from the disease and δ is the rate at which recovered individuals lose immunity and become susceptible again.

The important parameter I in Eq.(3) indicates an individual is infected, and is used to calculate the infected number R. In this study, our SIR model is based on the knowledge of network structure. For our knowledge-based SIR model, the important parameter I is formulated in Eq.(4).

$$I = \beta_0 + \beta_1 f(x), \tag{4}$$

where $f(x)$ is the degree distribution function of a network, and it can be acquired by our recognition model.

Parameter configuration of experiments. Based on the description of BN, a BN is an exponential network. For the comparability with RNs, the values of rate parameter λ for BNs are set to: (i) 0.4966 corresponding to our vehicle-based contact network, and (ii) 0.50159915 corresponding to our human-based contact network. And the number of nodes: (i) the BN with $\lambda = 0.4966$, is 2483, and (ii) the BN with $\lambda = 0.50159915$, is 942.

We repeat the process 100000 times for each network in our experiments with different randomly selected individuals. We use the average number of infected cases across all 100000 realizations as the value of R for each network.

Prediction accuracy for R. Based on our knowledge-based SIR model, extensive experiments are run on different networks, and on these experiments basis, the infected number R can be predicted, and the parameter R is time-varying to reflect the propagation dynamics of a disease. And the prediction results for R from different networks are compared with real surveillance data, to show network structure impacts the propagation dynamics on the network. And utilizing realistic structural knowledge can help to improve the prediction accuracy for R that is used to reflect the propagation dynamics on a network. Figure 9 illustrates: (i) the prediction results, and (ii) the comparison of the prediction results and real surveillance data.

From Fig. 9, we acquire: with realistic structural knowledge of networks, the prediction accuracy for R is improved, and network structure impacts propagation dynamics. For comparing the impacts of different networks, the ratio $R_{A/B} = \frac{R_A}{R_B}$ is calculated, at different time points, respectively, to measure the different impacts of the network A and the network B. First, we use A to denote real surveillance data, B to denote our vehicle-based contact network, C to

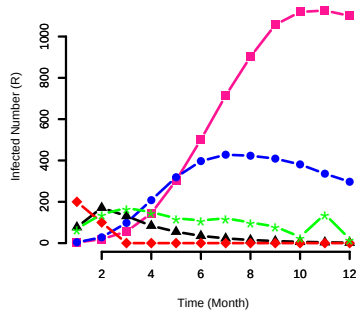


Fig. 9: Comparison of prediction results and real surveillance data. (i) The black line with triangular spots displays the acquired result by mining real surveillance data. (ii) The green line with star spots is the prediction result on our vehicle-based contact network. (iii) The red line with diamond-shaped spots is the prediction result on our human-based contact network. (iv) The pink line with square spots is the prediction result on the BN with $\lambda = 0.4966$ and 2483 nodes. (v) The blue line with circular spots is the prediction result on the BN with $\lambda = 0.50159915$ and 942 nodes.

denote our human-based contact network, D to denote the BN with $\lambda = 0.4966$ and 2483 nodes, and E to denote the BN with $\lambda = 0.50159915$ and 942 nodes. And then, $R^t_{A/B}$ denotes the ratio for network A and network B, at the t^{th} time point. Finally, $R^t_{B/A}$, $R^t_{C/A}$, $R^t_{D/A}$ and $R^t_{E/A}$ ($t = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$ (12 months)) are calculated and listed in Tab. 4.

Table 4: Ratios of R to measure the different impacts of two different networks on propagation dynamics

	$R^t_{B/A}$	$R^t_{C/A}$	$R^t_{D/A}$	$R^t_{E/A}$
t=1	0.91	2.6	0.039	0.052
t=2	0.82	0.58	0.11	0.16
t=3	1.3	0	0.43	0.75
t=4	1.8	0	1.7	2.5
t=5	2.2	0	5.5	5.8
t=6	3.1	0	14	11
t=7	5.2	0	31	19
t=8	6.7	0	60	28
t=9	8	0	105.7	40.9
t=10	5	0	186.5	63.5
t=11	35	0	281.5	84
t=12	10	0	550	148.5

The ratios listed in Tab. 4, are all different, so the impacts of these networks on propagation dynamics are different.

6 Related Work

6.1 Propagation Dynamics

Understanding the propagation processes arising over a wide variety of network structures is very important to mine useful knowledge about how a behavior on a network to impact the nodes of the network, and even is helpful to model the behavior. In recent years, there is an increasing effort to study propagation dynamics based on a variety of networks. Recent achievements can be divided into two categories based on different types of networks:

- The propagation dynamics of information on social networks [14][17]. In the information propagation of social networks, exponential and power-law models that reflect network structure have been widely used to model the dynamics of propagation [9][26]. Not only the network structure but also the prior probabilities of activation of edges [22] or the transmission rates of networks [7] are used to study the propagation dynamics of information on social networks.
- The propagation dynamics of real phenomena on contact networks. The contact networks describe the real relationships between individuals/systems of the physical world. Based on the real relationships from the physical world, the propagation dynamics on these networks is different from the propagation dynamics on social networks. With the development of the IoT (Internet of Things) and the help of various sensors and wireless devices, some researchers have paid their attention to this propagation dynamics, and have obtained some achievements: (i) for the propagation of infectious diseases [18][24][12][6], and (ii) for the propagation of contaminants [16]. Analyzing and studying the dynamics of propagation between individuals/systems can help us to understand and control the propagation dynamics on these real networks.

Some previous achievements assume networks to be static so that information propagates over these networks that their structures remain constant over time, and these achievements consider that different networks possess similar network structures and the structures of different networks can be modelled into unified models, e.g., exponential models and power-law models.

6.2 Disease Dynamics

As an important aspect of propagation dynamics, the disease dynamics on contact networks has been widely studied.

The quantification and prediction of disease dynamics during epidemics [30][31][20] are very important to public health [19] in allocating public health resources and in responding to public health events.

The infected number R can be used to quantify the disease dynamics during epidemics. For studying the quantized disease dynamics, a wide range of methods have been proposed to estimate or predict R [21][27][28][1][11] based on the assumptions of network structure, e.g., the contact networks for the spread of disease are best described as having exponential degree distributions [2].

However, realistic contact networks are not always and absolutely with the assumptions of network structure (e.g., exponential degree distributions). For improving the accuracy of estimating and predicting for R during an epidemic on a network, the realistic structure of the network needs to be mined.

7 Conclusion

In this article, we have mined the impacts of network structure on propagation dynamics through studying the disease dynamics that is an important aspect of propagation dynamics. Our study is based on the mobile data gathered from the real physical world, and with the mobile data, two RCNs are built, and as a comparison, we have implemented exponential networks using the greedy rewiring algorithm that is proposed by Bansal *et al.*. Exponential networks are widely used into RCN-based studies, and it has been demonstrated that the RCNs are best described as having exponential degree distributions. As a key result of this study, we have observed that RCNs do not always have exponential degree distributions, especially in dynamic environments. On this result basis, we have designed a model to recognize the structure of a network. Based on the model, we have investigated the impacts of network structure on propagation dynamics with evaluating and comparing the accuracy of prediction for the time-varying infected number R. In this comparing, the prediction results for R from different networks are compared with real surveillance data. From this investigation, we have obtained another key result of this study, the structure of a network impacts the propagation dynamics related on this network, and the prediction accuracy for R can be improved by utilizing realistic structural knowledge mined by our recognition model.

Acknowledgments

This work is supported by Guangdong University of Petrochemical Technology's Internal Project No.2012RC106, Educational Commission of Guangdong Province, China Project No. 2013KJCX0131, Guangdong High-Tech Development Fund No. 2013B010401035, 2013 Special Fund of Guangdong Higher School Talent Recruitment, National Natural Science Foundation of China under Grant 61401107, 2013 Top Level Talents Project in "Sailing Plan of Guangdong Province", and 2014 Guangdong Province Outstanding Young Professor Project.

References

1. Ames, G.M., George, D.B., Hampson, C.P., Kanarek, A.R., McBee, C.D., Lockwood, D.R., Achter, J.D., Webb, C.T.: Using network properties to predict disease

dynamics on human contact networks. Proceedings of the Royal Society B: Biological Sciences pp. 1–7 (2011)
2. Bansal, S., Grenfell, B.T., Meyers, L.A.: When individual behaviour matters: homogeneous and network models in epidemiology. Journal of the Royal Society Interface 4(16), 879–891 (2007)
3. Belik, V., Geisel, T., Brockmann, D.: Natural human mobility patterns and spatial spread of infectious diseases. Physical Review X 1(1), 011001 (2011)
4. Charaudeau, S., Pakdaman, K., Boëlle, P.Y.: Commuter mobility and the spread of infectious diseases: application to influenza in france. PloS one 9(1), e83002 (2014)
5. Chen, Z., Gao, L., Kwiaty, K.: Modeling the spread of active worms. In: INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies. vol. 3, pp. 1890–1900. IEEE (2003)
6. Colizza, V., Barrat, A., Barthelemy, M., Valleron, A.J., Vespignani, A.: Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. PLoS medicine 4(1), e13 (2007)
7. Du, N., Song, L., Yuan, M., Smola, A.J.: Learning networks of heterogeneous influence. In: Advances in Neural Information Processing Systems. pp. 2780–2788 (2012)
8. Garetto, M., Gong, W., Towsley, D.: Modeling malware spreading dynamics. In: INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies. vol. 3, pp. 1869–1879. IEEE (2003)
9. Gomez Rodriguez, M., Leskovec, J., Krause, A.: Inferring networks of diffusion and influence. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1019–1028. ACM (2010)
10. Gomez-Rodriguez, M., Leskovec, J., Krause, A.: Inferring networks of diffusion and influence. ACM Transactions on Knowledge Discovery from Data (TKDD) 5(4), 21 (2012)
11. Groendyke, C., Welch, D., Hunter, D.R.: A network-based analysis of the 1861 hagelloch measles data. Biometrics 68(3), 755–765 (2012)
12. Hufnagel, L., Brockmann, D., Geisel, T.: Forecast and control of epidemics in a globalized world. Proceedings of the National Academy of Sciences of the United States of America 101(42), 15124–15129 (2004)
13. Keeling, M.J., Rohani, P.: Modeling infectious diseases in humans and animals. Princeton University Press (2008)
14. Lappas, T., Terzi, E., Gunopulos, D., Mannila, H.: Finding effectors in social networks. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1059–1068. ACM (2010)
15. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 497–506. ACM (2009)
16. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 420–429. ACM (2007)
17. Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, M.W.: Statistical properties of community structure in large social and information networks. In: Proceedings of the 17th international conference on World Wide Web. pp. 695–704. ACM (2008)
18. Lipsitch, M., Cohen, T., Cooper, B., Robins, J.M., Ma, S., James, L., Gopalakrishna, G., Chew, S.K., Tan, C.C., Samore, M.H., et al.: Transmission dynamics and control of severe acute respiratory syndrome. Science 300(5627), 1966–1970 (2003)

19. Luke, D.A., Stamatakis, K.A.: Systems science methods in public health: dynamics, networks, and agents. Annual Review of Public Health 33, 357–376 (2012)
20. Martín, G., Marinescu, M.C., Singh, D.E., Carretero, J.: Leveraging social networks for understanding the evolution of epidemics. BMC systems biology 5(Suppl 3), S14 (2011)
21. Mukandavire, Z., Liao, S., Wang, J., Gaff, H., Smith, D.L., Morris, J.G.: Estimating the reproductive numbers for the 2008–2009 cholera outbreaks in zimbabwe. Proceedings of the National Academy of Sciences 108(21), 8767–8772 (2011)
22. Myers, S., Leskovec, J.: On the convexity of latent social network inference. In: Advances in Neural Information Processing Systems. pp. 1741–1749 (2010)
23. Myers, S.A., Zhu, C., Leskovec, J.: Information diffusion and external influence in networks. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 33–41. ACM (2012)
24. Riley, S., Fraser, C., Donnelly, C.A., Ghani, A.C., Abu-Raddad, L.J., Hedley, A.J., Leung, G.M., Ho, L.M., Lam, T.H., Thach, T.Q., et al.: Transmission dynamics of the etiological agent of sars in hong kong: impact of public health interventions. Science 300(5627), 1961–1966 (2003)
25. Rodrigue, M.G., Leskovec, J., Balduzzi, D., Schölkopf, B.: Uncovering the structure and temporal dynamics of information propagation. Network Science 2(01), 26–65 (2014)
26. Rodriguez, M.G., Schölkopf, B.: Submodular inference of diffusion networks from multiple trees. arXiv preprint arXiv:1205.1671 (2012)
27. Stadler, T., Kouyos, R., von Wyl, V., Yerly, S., Böni, J., Bürgisser, P., Klimkait, T., Joos, B., Rieder, P., Xie, D., et al.: Estimating the basic reproductive number from viral sequence data. Molecular biology and evolution 29(1), 347–357 (2012)
28. Team, W.E.R.: Ebola virus disease in west africathe first 9 months of the epidemic and forward projections. N Engl J Med 371(16), 1481–95 (2014)
29. Tomé, T., Ziff, R.M.: Critical behavior of the susceptible-infected-recovered model on a square lattice. Physical Review E 82(5), 051921 (2010)
30. Vazquez-Prokopec, G.M., Bisanzio, D., Stoddard, S.T., Paz-Soldan, V., Morrison, A.C., Elder, J.P., Ramirez-Paredes, J., Halsey, E.S., Kochel, T.J., Scott, T.W., et al.: Using gps technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment. PloS one 8(4), e58802 (2013)
31. Woolhouse, M.: How to make predictions about future infectious disease risks. Philosophical Transactions of the Royal Society B: Biological Sciences 366(1573), 2045–2054 (2011)