Information Sciences 379 (2017) 82–93

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Reality mining: A prediction algorithm for disease dynamics based on mobile big data

Yuanfang Chen[a,c,*], Noel Crespi[a], Antonio M. Ortiz[b], Lei Shu[c]

[a] Institut Mines-Télécom, Télécom SudParis, France
[b] R&D Department, Montimage, France
[c] Guangdong University of Petrochemical Technology, Guangdong, China

ARTICLE INFO

Article history:
Received 2 November 2015
Revised 22 July 2016
Accepted 29 July 2016
Available online 30 July 2016

Keywords:
Reality mining
Disease dynamics
Prediction algorithm
Mobile big data

ABSTRACT

Predicting disease dynamics during an epidemic is an important aspect of e-Health applications. In such prediction, Realistic Contact Networks (RCNs) have been widely used to characterize disease dynamics. The structure of such networks is dynamically changed during an epidemic. Capturing such kind of dynamic structure is the basis of prediction. With the popularity of mobile devices, it is possible to capture the dynamic change of the network structure. On this basis, in this study, we evaluate the impact of the network structure on disease dynamics, by analyzing massive spatiotemporal data collected by mobile devices. These devices are carried by the volunteers of Ebola outbreak areas. Based on the results of this evaluation, a model is designed to recognize the dynamic structure of RCNs. On the basis of this model, we propose a prediction algorithm for disease dynamics. By extensive experiments, we show that our algorithm improves the accuracy of the disease prediction.

1. Introduction

As an important aspect of e-Health [19–21], quantifying and even predicting disease dynamics during an epidemic [17,35,36,38] is very important to effectively allocate resources and to quickly make a response in a public health event. For the public health, underestimating the impact of a disease may lead to an inadequate response, while overestimating it, can lead to the misallocation on the limited resources.

The reproductive 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 R by mining surveillance data [4,16,24,31,34]. However, these existing methods, almost all of them, are based on the networks which are assumed to have special network structure, for example, the networks with exponential degree distributions. During an epidemic, the network structure of the relevant Realistic Contact Network (RCN) is dynamically changed along with the spread of a disease. The definition of the RCN is described in Definition 1.

* Corresponding author.
E-mail addresses: yuanfang.chen.2009@ieee.org, yuanfang.chen.tina@outlook.com (Y. Chen), noel.crespi@mines-telecom.fr (N. Crespi), antonio.ortiz@montimage.com (A.M. Ortiz), lei.shu@ieee.org (L. Shu).

[1] The number of cases generates in an infectious period, in an uninfected population.

http://dx.doi.org/10.1016/j.ins.2016.07.075
0020-0255/

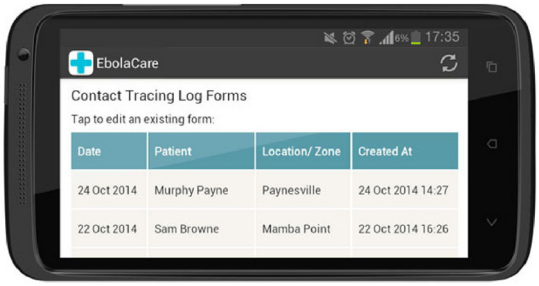

Fig. 1. A mobile device with the contact tracing application installed. This application is used to track the Ebola outbreak of West Africa [1]. It can track everyone who directly contacts with a sick Ebola patient. Such devices are carried by the volunteers of Ebola outbreak areas. The data collected by this application is shared with the WHO (World Health Organization), who is using information from hundreds of aid organizations to make big strategic decisions.

Definition 1. In the real physical world, a Realistic Contact Network consists of a group of people who can get in touch with each other. In this network, nodes represent people, and edges represent direct contact between two nodes. If there is an edge between two nodes, it means that there is physical contact between two individuals corresponding to the two nodes.

During an epidemic, capturing the dynamic change of the relevant RCN is helpful to improve the prediction accuracy of disease dynamics. With the popularity of mobile devices in public health [15], it is possible to acquire the dynamic change of network structure. An example is illustrated in Fig. 1.

In this study, we design a recognition model to dynamically acquire the structure knowledge of the relevant RCN during an epidemic. On the basis of this model, we propose a prediction algorithm to predict the parameter R.

The scientific contributions of this article are shown as follows:

- We evaluate the impact of network structure on disease dynamics, by analyzing the massive data collected by the mobile devices carried by volunteers.
- We design a recognition model to acquire the structure knowledge of an RCN. Three structure properties are used to design the model. The model design is based on the observations from the above evaluation.
- We propose a prediction algorithm. This algorithm uses the acquired structure knowledge by the recognition model.

There is a strong correlation among these three contributions: the model design is based on the evaluation, and the prediction algorithm uses the recognized structure knowledge by the recognition model as an important aspect to improve the prediction accuracy. These three contributions are linked together, and achieved one by one: evaluation → model design → prediction algorithm. Such a network-structure-based prediction algorithm improves the prediction accuracy, even when the network structure is dynamically changed.

The remainder of this article is organized as follows. Section 2 provides related work to elicit the research issue of this study. Section 3 introduces the RCN that is used to carry out our study. On the basis of the RCN, we evaluate the impact of the network structure on disease dynamics. Considering the results of this evaluation, we design a recognition model and show it in Section 4. Based on the recognition model, we propose a prediction algorithm in Section 5. Section 6 compares by extensive experiments our algorithm with the prediction algorithm not using a recognition model. Moreover, this section discusses the experimental results in detail. This article is concluded in Section 7.

2. Related work

This section provides a brief overview about disease dynamics, from the perspective of widely used models and prediction methods.

Epidemic models describe the spread of a communicable disease in a population. In these models, the individuals of a population are taken and placed into one of these three states: Susceptible, Infectious or Recovered (SIR). Modelling the transitions among these states generates the SIR model. This simple SIR model has been extended in a multitude of ways, e.g., by adding/deleting states, or by considering a special pattern of a transition. For example: (i) The SIS model [18]. For a disease with no immunity, infected susceptible individuals return to the susceptible state after recovering. (ii) Non-equilibrium transitions [22]. Replacing the homogeneous mixing hypothesis that any individual can contact with any other, non-equilibrium transitions assume that each individual has a certain number of contacts, which is reflected as the node's degree k in a contact network. The degrees of nodes in a network can be denoted as the degree distribution of the network. Degree distribution is the important structural property of a network. Different degree distributions bring different impacts on the transitions among different states of disease dynamics. The relevant studies on such impacts enable disease dynamics to be associated with a network, and to explore the impact of spatial contact patterns [9,11,32,33,37].

For example, in a dynamic contact network with an arbitrary degree distribution, the real transition threshold of three states during an epidemic is: $\lambda_{th} = \frac{k}{k^2}$, where k is the average degree of the network, and k^2 is the average degree at the next moment of the dynamic network.

Table 1
A qualitative overview of the prediction methods on disease dynamics.

Prediction method	Addressed problem	Main idea	Open issues
Prediction based on online social networks [8,10]	Predicting the health of real-world populations in real time, to understand actual threats and ongoing disease outbreaks	Such a method used the geo-tagged status updates of traveling Twitter users to infer the properties of individual flow between cities.	(i) Use of Twitter data. First, Tweets are typed by users who are not experts: the correctness of their judgment for their suffered diseases cannot be guaranteed. Moreover, for example, influenza can be described by different words (even such description does not contain the keywords "flu" or "influenza"). Second, the observed data comes from a small fraction of travellers, and it is only partial data of these travellers. (ii) Based on our experience in Twitter data acquisition by the Twitter Search API, Tweets sent from mobile devices do not always have accurate GPS coordinates. And even if two Tweets can be located in the same location (with a certain locating error), we cannot make assure that the relevant physical individuals contact with each other and are within the communicable distance of a disease.
Prediction based on RCNs [4,14,26,29]	Such studies have proposed that the structure of contact networks impacts disease dynamics. And such impacts are different for different network structures. For the RCNs, there are three typical network structures that are widely used in studies. They are exponential, power-law [23] and random [7,30].	Constructing RCNs is by analyzing the actual surveillance data from disease outbreaks.	The exponential and power-law degree distributions have been found in many real-world complex networks [12,29]. However, the structure of a contact network is dynamically changed during an epidemic, and structure knowledge is very important to characterize the dynamics of the spread process on this contact network, so further work is necessary to improve the prediction accuracy of disease dynamics: mining the realistic structure knowledge acquired from the dynamic contact network in real time.

As another important aspect of the overview for disease dynamics, the previous studies on prediction methods are briefly classified and summarized in Table 1.

Through the brief overview, and the acquired open issues, our study therefore focuses on acquiring the realistic structure knowledge of the dynamic contact network during an epidemic, and then, considering the acquired structure knowledge, a prediction algorithm is proposed to predict the disease dynamics during the epidemic. About the structure knowledge, three structure properties are used in this study: clustering coefficient, degree distribution and degree correlation.

3. Impact evaluation

3.1. Realistic contact network

By processing surveillance data collected by the mobile devices which are carried by the volunteers of Ebola outbreak areas, we construct the RCN that is used in our study.

As a recent outbreak of disease, from 27th March, 2016, Ebola Virus Disease (it is commonly known as "Ebola") has killed 11,323 individuals, and the total number of cases has reached 28,646 [2]. Researchers generally believe that from a 2-year-old boy of Guinea to his mother, sister and grandmother (this is a contact network) Ebola rapidly spreads in West Africa since March 2014. A series of time-aware Ebola cases is collected by the WHO as well as the Ministries of Health of epidemic countries.

In this study, we select three groups of data from three typical outbreak countries, Guinea, Nigeria and Liberia, and seven regions of these three countries. Guinea is the source of this outbreak and has relatively high quantity of confirmed cases, Nigeria is far away from the source of the outbreak and has relatively low quantity of confirmed cases, and Liberia is close to the source of the outbreak and has high quantity of confirmed cases. And the seven regions are: Gueckedou of Guinea, Macenta of Guinea, Kissidougou of Guinea, Conakry of Guinea, Lagos of Nigeria, Port Harcourt of Nigeria and Monrovia of Liberia.

By using the above outbreak data, we construct a contact network. There are 941 nodes corresponding to different cases in this network (Fig. 2 illustrates a slice of the network). During an epidemic, with the spread of a disease, the contact network is gradually constructed, and by the order of time stamps of cases, we can clearly know the spread process of the disease. Such a contact network can be modelled as a dynamic graph G_t. In this network, there are four parameters: (i) Case ID. A unique number to indicate a case; (ii) Source ID. It indicates the ID of the infection source for a case; (iii) Date. The reported date of a case; (iv) Location. The coordinates (longitude and latitude) of a case.

The dynamic graph G_t can be described in detail as follows. An undirected weighted graph $G_t = (V_t, E_t, W_t)$, where V_t is a set of n_t vertices to indicate cases, E_t is the set of edges, and W_t is the set of weights. If there is an edge between vertex

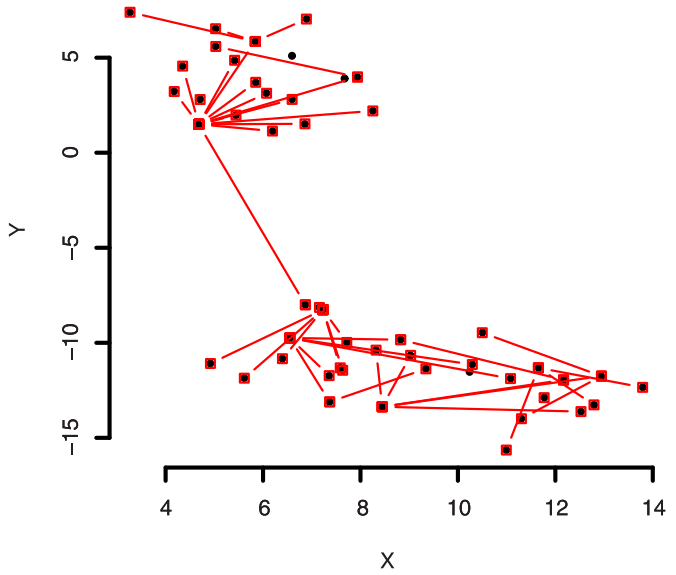

Fig. 2. A slice of a dynamic contact network. This slice displays 50 cases and their relationships (contact) from three typical countries and seven regions of the Ebola outbreak since 2014. The three countries considered are: Guinea, Nigeria and Liberia. The seven regions considered are: Gueckedou, Macenta, Kissidougou, Conakry, Monrovia, Lagos and Port Harcourt. The black dots are cases (suspected and confirmed), and if there is an edge between two dots, it means that there is contact between two corresponding individuals of two cases. We only display a slice of a contact network in this example, so there are isolated nodes. It means that there are black dots without connections.

i and vertex j, e_{ij}, it indicates that there is contact between the corresponding individuals of i and j. The weight w_{ij} is the transmission probability (p_{ij}) of a disease from vertex i to vertex j (on the corresponding edge e_{ij}).

The graph G_t is dynamic, so it is with a sequence of online updates: (i) Delete(e_{ij}): it deletes the edge e_{ij} from E_t, and the corresponding vertices i and j from V_t; (ii) Insert(e_{ij}): it inserts the edge e_{ij} into E_t, and the corresponding vertices i and j into V_t; (iii) Update(w_{ij}): it updates the weight w_{ij} that is related to the edge e_{ij}. On the basis of above (i), (ii) and (iii), 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 the spread of a disease, the active subnetworks are different. Such a contact network is time-varying.

3.2. Evaluation results

The Spatial Risk Model (SRM) [13] is used in this study to evaluate the impact of network structure on disease dynamics. It is a statistical model used in communicable diseases to estimate or predict the presence or incidence of infected cases within a particular geographical area. In this evaluation, we calculate the number of infected cases (reproductive number R). The parameter R is measured to quantify the disease dynamics during an epidemic. In this evaluation, we randomly sample 100 time periods from the real surveillance data to construct 100 different subnetworks (the details of the surveillance data have been introduced in Section 3.1, and a time period covers one day).

For evaluating the impact of network structure on disease dynamics, we use these three structure properties: clustering coefficient, degree distribution and degree correlation. These three structure properties strongly correlate with the degrees of nodes. In a network, the degrees of nodes are the most important and basic structure knowledge. The degree of a node indicates the number of adjacent edges on this node. For a real-world network, the relationships between nodes are complex. The degrees of nodes are helpful to characterize and model the real-world network.

Clustering coefficient. In graph theory, a clustering coefficient is a measurement of the degree to which vertices in a graph tend to cluster together. In this study, we use the definition proposed by Barrat [6] for the clustering coefficient of a network, which is a local node-level definition, and it is formulated as:

$$C_i^w = \frac{1}{s_i(k_i - 1)} \sum_{j,h} \frac{w_{ij} + w_{ih}}{2} a_{ij} a_{ih} a_{jh}, \tag{1}$$

where k_i is the degree of node i, s_i is the strength (summing up the edge weights of the adjacent edges of node i), a_{ij}, a_{ih} and a_{jh} are elements of the adjacent matrix of the network, and w_{ij} and w_{ih} are the weights of corresponding edges.

Degree distribution. Degree distribution is the probability distribution of nodes' degrees over a network. It provides the overall structure knowledge of the network. For displaying the relationship between degree distribution and disease dynamics (parameter R), we calculate the expectation of degree distribution, and this expectation can be denoted as:

$$E[X] = x_1 p_1 + x_2 p_2 + \ldots + x_k p_k, \tag{2}$$

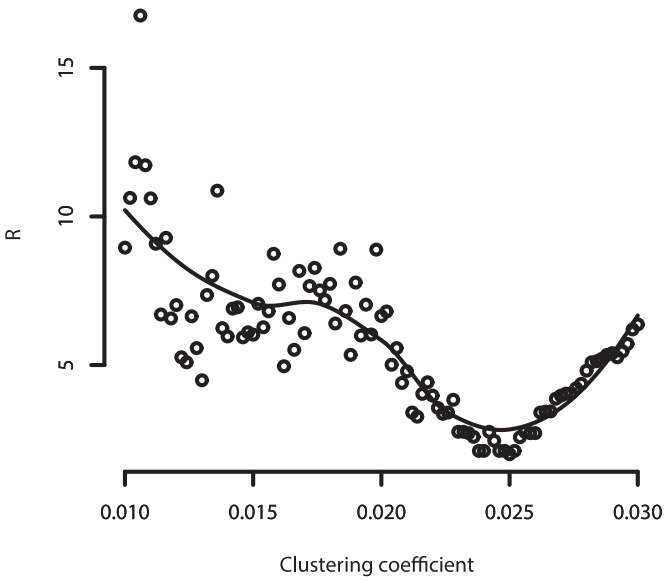

Fig. 3. Impact of the clustering coefficient on disease dynamics (parameter R).

where $X = \{x_1, x_2, \ldots, x_i, \ldots, x_k\}$ is the set of degrees of nodes, and $P = \{p_1, p_2, \ldots, p_i, \ldots, p_{k'}\}(p_1 + p_2 + \ldots + p_{k'} = 1$ *and* $k' \leq k)$ represents the probabilities of the degree values $\{x_1, x_2, \ldots, x_i, \ldots, x_k\}$, for example, if the set of degrees of nodes is $X = \{1, 1, 1, 2\}$, then the set of probabilities for this X is $P = \{\frac{3}{4}, \frac{1}{4}\}$, and $E[X] = 1 * \frac{1}{4} + 1 * \frac{1}{4} + 1 * \frac{1}{4} + 2 * \frac{1}{4} = \frac{5}{4}$.

Degree correlation. The degree correlation measures the level of the homophyly of a network. If a network has a high correlation, it means that the connected nodes of the network tend to have the same degree. In this study, we adopt the definition of M.E.J. Newman [27,28] to define degree correlation, and it can be denoted as:

$$r = \frac{1}{\sigma_q^2} \sum_{j,k} jk(\mathcal{E}_{jk} - q_j q_k), \tag{3}$$

where (i) $\sigma_q^2 = \sum_k k^2 q_k - [\sum_k k q_k]^2$, (ii) $\mathcal{E}_{jk}$ is the joint probability distribution of the degrees of the two nodes (j and k) at either end of a randomly chosen edge e_{jk}. This quantity is symmetric in an undirected network, $\mathcal{E}_{jk} = \mathcal{E}_{kj}$, and obeys these rules: $\sum_{j,k} \mathcal{E}_{jk} = 1$ and $\sum_j \mathcal{E}_{jk} = q_k$, (iii) $q_k = \frac{p_{k+1}}{\sum_{j \geq 1} p_j}$, where p_{k+1} is the degree of node $k+1$, and (iv) the value of "k" in Eq. (3) is set as the degree of the node k.

The evaluation results are described and discussed as follows.

Evaluation results for the clustering coefficient. Fig. 3 illustrates the relationship between the structure property "clustering coefficient" and the parameter R.

From the results shown in Fig. 3, we observe that different values of clustering coefficients are corresponding to different R values. It means that there is a relationship between the network's clustering coefficient and disease dynamics, and this relationship cannot be ignored. Thus, when we design a prediction algorithm for disease dynamics, the clustering coefficient needs to be considered in order to improve the prediction accuracy.

Evaluation results for the degree distribution. Fig. 4 illustrates the relationship between the structure property "degree distribution" and the parameter R. Eq. (2) is used to calculate the expectation of network's degree distribution.

Comparing Figs. 3 and 4, there is a like-mirrored relationship about the changing trends of the two lines. It means that the two structure properties, the clustering coefficient and degree distribution of a network, bring different impacts to the parameter R. And the difference about the impacts is with a like-mirrored relationship.

Evaluation results for the degree correlation. Fig. 5 illustrates the relationship between the structure property "degree correlation" and the parameter R.

If the value of the degree correlation is positive in a network, it means that similar nodes (having the same degree) tend to connect together, and negative otherwise.

From the results illustrated in Fig. 5, we observe that the overall changing trend of R is non-monotonic decline, with the increase of the degree correlation. It means that with the spread of a disease, the homophyly of a contact network is getting higher, so there is a reductive trend about the new cases that are generated in an infectious period. The infectiousness of the disease is getting weaker as time goes by.

However, as a special and important point that is worth explaining in detail: with the increase of the degree correlation (the homophyly of a contact network is getting higher), the reproductive number R is decreasing over time, and this decreasing process is non-monotonic. At some points, the R values increase again to some extent. Based on SRM, there is a

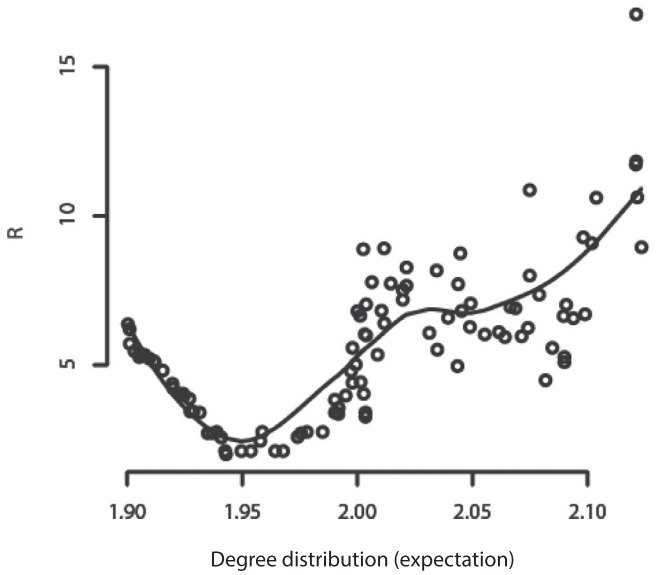

Fig. 4. Impact of the degree distribution on disease dynamics (parameter R).

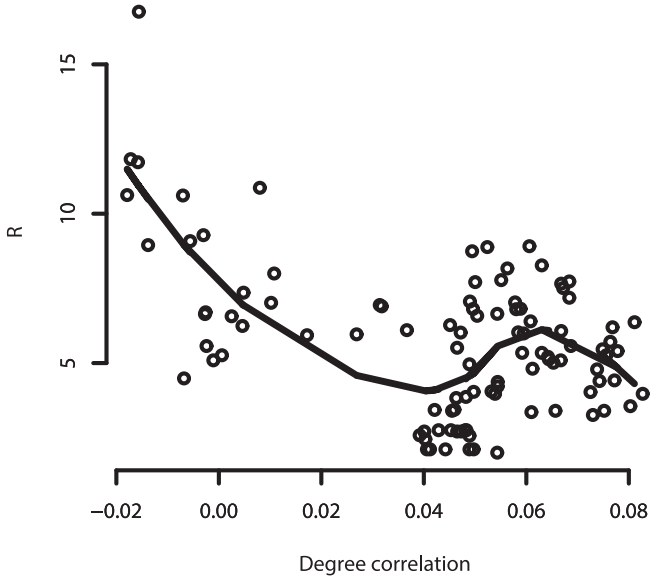

Fig. 5. Impact of the degree correlation on disease dynamics (parameter R).

recovery probability for each infected individual, and for the recovered individuals, even if they have recovered, they have contacted with other individuals of this network, so they still connect with these other individuals. When we calculate the new R value, the recovered individuals (but they are still connected to the network) are counted as new infectious cases. These recovered individuals make the values of R increase again at some value points of the degree correlation.

By these three formulae, Eqs. (4), (2) and (5), we can characterize and quantify the different impacts of three structure properties on the parameter R.

$$E[k_i] = E\left[\frac{\sum_{j,h} a_{ij}a_{ih}a_{jh}}{C_i^w s_i} + 1\right]. \tag{4}$$

We use the undirected weighted network in this study, without loss of generality, we assume $w_{ij} = 1$ and $w_{ih} = 1$, when there is an edge between i and j, and between i and h. Considering Eq. (1), we can deduce: $k_i = \frac{\sum_{j,h} a_{ij}a_{ih}a_{jh}}{C_i^w s_i} + 1$, and as the limit cases, when $C_i^w = 0$, $k_i = \infty$ and when $C_i^w = \infty$, $k_i = 1$. It means that there is a like-mirrored relationship between

"clustering coefficient" and "degree distribution", and this result can be observed in Figs. 3 and 4 as well.

$$E[\mathcal{E}_{jk}] = \sum_{j,k} jk\mathcal{E}_{jk} = r\sigma_q^2 + \sum_{j,k} jkq_jq_k, \tag{5}$$

where $E[\mathcal{E}_{jk}]$ is deduced from Eq. (3).

Observing the above evaluation results and three formulae, Eqs. (4), (2) and (5), we are clear on this: these three formulae indeed can characterize and quantify the different impacts of three structure properties on the parameter R. We can learn the following points:

- From Eq. (4), C_i^w is inversely proportional to the degree of node i (k_i), and comparing Figs. 3 and 4, their changing trends have a like-mirrored relationship, so we can make this conclusion: Eqs. (1) and (2) can model and reflect the different impacts of the clustering coefficient and degree distribution on the parameter R.
- From Eq. (5), we know: for a network, r is proportional to the joint degree distribution of the network ($\sum_{j,k} jk\mathcal{E}_{jk}$), rather than the degree of a single node. So comparing Figs. 4 and 5, their changing trends are different, and they do not have an obvious relationship.

4. Recognition model

The structure of a dynamic network is time-varying. It means that in a dynamic network the network structure is different at different time points. The RCN is a kind of typical dynamic network. In such a network, if we want to study how the network structure impacts disease dynamics, and even predict the disease dynamics, the dynamic recognition of network structure is necessary. In this section, we design a recognition model to recognize the dynamic structure of an RCN.

There are two components in this recognition model: (i) measure. This component calculates the values of three structure properties, and (ii) knowledge acquisition. The measured values from the first component need to be combined together to reflect and quantify the structure knowledge of a network, and then the structure knowledge can be used into the prediction of disease dynamics (measuring and predicting the value of R).

This model can be formulated and described as follows:

- Measure. As the important structure properties to reflect the structure of a network, in this study, three structure properties are used, and they are: clustering coefficient (C_i^w), degree distribution ($E[X]$) and degree correlation (r). The formulae to measure the values of these structure properties are shown in Eqs. (1), (2) and (3), respectively. Using these three formulae, three values can be calculated, and these values reflect and quantify the structure knowledge of the network, and by this quantification, the structure knowledge can be used into the design of the prediction algorithm for disease dynamics.
- Knowledge acquisition. Based on the measured values of structure properties, the combination formula can be denoted as:

$$M(C_i^w, E[X], r) = E[k_i] + E[X] + E[\mathcal{E}_{jk}], \tag{6}$$

where these three parts, $E[k_i]$, $E[X]$ and $E[\mathcal{E}_{jk}]$ have been introduced in Eqs. (4), (2) and (5), and Eqs. (4) and (5) are deduced from Eqs. (1) and (3) by converting and unifying the different measurement parameters (C_i^w and r) to degree-relevant parameters (k_i and $\mathcal{E}_{jk}$).

5. Prediction algorithm

On the basis of the above recognition model, we propose a prediction algorithm for disease dynamics. The algorithm consists of two parts: (i) acquiring structure knowledge by using our recognition model, and (ii) estimating the parameter R by using SRM.

Begin Prediction Algorithm:
First part: Acquiring structure knowledge
Input: a contact network of a disease outbreak, G_t
$C_i^w = transitivity(G_t, type = c(local))$ ▷ The function $transitivity(.)$ is used to calculate the clustering coefficient of each node $i \in V_t$.
$S_i = graph.strength(G_t)$ ▷ The function $graph.strength(.)$ is used to calculate the strength of each node $i \in V_t$.
==========
▷ This for-loop structure is used to calculate the degree of each node $i \in V_t$, k_i.
for (i in $1 : |V_t|$) **do**
  $tmp = \frac{\sum_{j,h} a_{ij}a_{ih}a_{jh}}{C_i^w s_i} + 1$ ▷ tmp is a temporary variable.
  $K = combine(K, tmp)$ ▷ K is the set of k_i ($i \in V_t$).
end for
==========

▷ This for-loop structure is used to calculate $E[k_i]$ that is the expectation of K.
for (i in 1 : $|K|$) **do**
  $E[k_i] = E[k_i] + K[i] * K[i]/sum(K)$
end for
==========
▷ This for-loop structure is used to calculate $E[X]$ that is the expectation of degree distribution.
for (i in 1 : $|dg|$) **do** ▷ $dg = degree(G_t)$ is the degree distribution of network G_t.
  $E[X] = E[X] + dg[i] * dg[i]/sum(dg)$ ▷ $dg[.]$ is used to indicate the degree of a node.
end for
==========
$r = assortativity.degree(G_t, directed = FALSE)$ ▷ The function $assortativity.degree(.)$ is used to calculate the degree correlation of each node $i \in V_t$.
==========
$E_t = get.edges(G_t)$ ▷ The function $get.edges(.)$ is used to get all edges of network G_t.
==========
▷ This for-loop structure is used to calculate the two parts of $E[\mathcal{E}_{jk}]$: (i) $\mathcal{E}_1$: σ_q^2, and (ii) $\mathcal{E}_2$: $\sum_{j,k} jkq_jq_k$.
for (i in 1 : $|E_t|$) **do**
▷ The function $which(.)$ is used to get the numbers of the two nodes jointed by an edge.
  $j = which(V_t == E_t[i, 1])$
  $k = which(V_t == E_t[i, 2])$
  $q_j = \frac{p_{j+1}}{\sum_{k\geq 1} p_k}$ ▷ Calculate q_j.
  $q_k = \frac{p_{k+1}}{\sum_{j\geq 1} p_j}$ ▷ Calculate q_k.
▷ Calculate the two parts of σ_q^2: (i) σ_1: $\sum_k k^2q_k$, and (ii) σ_2: $\sum_k kq_k$.
  $\sigma_1 = \sigma_1 + dg[k] * dg[k] * q_k$
  $\sigma_2 = \sigma_2 + dg[k] * q_k$
  $\mathcal{E}_2 = \mathcal{E}_2 + dg[j] * dg[k] * q_j * q_k$ ▷ Calculate $\sum_{j,k} jkq_jq_k$.
end for
$\sigma_q^2 = \sigma_1 - [\sigma_2]^2$
$E[\mathcal{E}_{jk}] = r * \sigma_q^2 + \mathcal{E}_2$
==========
$M = \frac{E[k_i]+E[X]+E[\mathcal{E}_{jk}]}{3}$ ▷ Combine these three expectations: $E[k_i]$, $E[X]$ and $E[\mathcal{E}_{jk}]$.
==========
Second part: Estimating the parameter R
SRM($\beta = M$, $\gamma = M$, $no.iteration$) ▷ "$no.iteration$" is used to indicate the number of iterations. SRM: Spatial Risk Model
==========
Output: a predicted value of the parameter R, related to input network G_t
End

As the important parameters of SRM, β and γ, (i) β is a non-negative scalar, the infection probability of an individual. This individual is susceptible and has a single infected neighbor. The infection probability of a susceptible individual with n infected neighbors is $n * \beta$, and (ii) γ is a positive scalar, the recovery probability of an infected individual, and if this individual recovers from a disease, this individual will be disconnected with other individuals, in a contact network.

Based on the above explanations for β and γ, we observe that these two parameters β and γ are all degree-related. So β and γ can be used to reflect the structure knowledge of a contact network in the SRM model.

6. Evaluation

6.1. Experimental setting

A series of time-aware Ebola cases[2] (941 nodes) and their relationships (938 edges) are collected and used to evaluate the performance of our prediction algorithm. These cases and relationships come from the outbreak of Ebola in West Africa from March 2014. The network structure of the corresponding RCN is dynamically changed during this outbreak, with the spread of the Ebola virus disease. On this basis, we construct nine networks according to the time stamps of cases. These networks are $G_t = \{G_1, G_2, \ldots, G_9\}$, where $t = 1, 2, \ldots, 9$ is corresponding to nine weeks of August (four weeks), September

[2] Using the wireless devices carried by volunteers, the cases and relationships (contact) can be tracked and recorded. These devices are GPS-enabled, and the reported records have time stamps.

Table 2
Detailed information of nine dynamic networks, G_t ($t = 1, 2, \ldots, 9$).

Number of nodes	Number of edges	λ (standard error)	Time period
333	330	0.5045455 (0.02764892)	from 1st August to 7th August
340	337	0.504451 (0.0273577)	from 8th August to 14th August
340	337	0.504451 (0.0273577)	from 15th August to 21st August
531	528	0.5028409 (0.02182144)	from 22nd August to 31st August
647	644	0.5023292 (0.0197486)	from 1st September to 7th September
707	704	0.5021307 (0.01888457)	from 8th September to 14th September
779	776	0.501933 (0.01798362)	from 15th September to 21st September
916	913	0.5016429 (0.01657475)	from 22nd September to 30th September
941	938	0.5015991 (0.01635166)	from 1st October to 7th October

(four weeks) and October (one week) 2014. The detailed information of these nine networks is shown in Table 2. And we use Graphx of Spark [3] to process these networks.

In Table 2, we use this process to obtain the values of λ and corresponding standard errors, for each network: we conduct a maximum-likelihood fitting to fit the degree distribution of each network into an exponential distribution, and then we can obtain the corresponding values of λ and standard errors on the fitted exponential distribution, for each network. The maximum-likelihood fitting uses a maximum-likelihood estimation [25] to estimate the values of distribution parameters. Moreover, the sizes of the nine networks are increasing gradually: from 333 nodes (the first week of August) to 941 nodes (the first week of October), with the spread of Ebola virus disease.

For comparison, the SRM-based prediction algorithm that does not integrate a recognition model is used in our experiments. It has recently been demonstrated that empirical human contact networks are best described as having exponential degree distributions [4,5]. Thus, the prediction algorithm that does not integrate a recognition model is based on the networks which have exponential degree distributions.

Moreover, we compare the predicted results of two algorithms with the real values of reproductive number R. These real values are counted from the collected outbreak data. By this real comparison, we can clearly know the performance of the prediction algorithms.

6.2. Experimental results and discussion

Fig. 6 illustrates experimental results. We compare the prediction accuracy of our algorithm and the prediction algorithm that does not integrate a recognition model. And we compare the predicted results of two algorithms with the real values of R as well. Extensive experiments are conducted: for the dynamic network of each time period, the parameter of SRM, $no.iteration = 100$. It means that the two algorithms are iterated 100 times to get the average of predicted values of R, in each time period.

By analyzing the comparative results illustrated in Fig. 6, we get two observations, and specially we discuss how the proposed algorithm improves prediction accuracy?

- In terms of the predictive performance for the parameter R, our prediction algorithm performs better than the prediction algorithm that does not integrate a recognition model. The relevant standard deviations[3] of R's real values and R's predicted values are 100.133, 8.271831 and 3.146532, so the variation of the predicted values calculated by our algorithm is closer to the variation of real values. The values corresponding to Fig. 6 and relevant three standard deviations are listed in Table 3.
  The prediction performed by our algorithm is based on acquiring the structure knowledge of a contact network. By the acquired structure knowledge, the more reasonable values of β and γ can be set in SRM. On the basis of such a parameter setting, we can obtain better prediction results.
- In each time period, the deviation between $\mathfrak{R}$ and $\mathfrak{I}$ is different from the deviation between $\mathfrak{R}$ and $\mathfrak{U}$. $\mathfrak{R}$ denotes the real value, $\mathfrak{I}$ is the predicted value calculated by our algorithm, and $\mathfrak{U}$ indicates the predicted value calculated by the algorithm that is used to compare with our algorithm. The deviations between the real value and the predicted values (from two algorithms), for each time period, are listed in Table 4.

As shown in Table 4, our algorithm has a smaller predicted deviation to the real values. Our algorithm is based on the structure recognition of the RCN during an epidemic, and the infectious disease spreads through this RCN. Thus, based on recognizing the real disease transmission network, and using the calculated values by our recognition model, we can obtain more authentic values of β and γ to SRM. On this basis, our prediction algorithm achieves a smaller predicted deviation to the real value, for each time period. It also means that our algorithm improves prediction accuracy.

Moreover, by analyzing the deviations between real values and predicted values (Table 4), we observe that there are big deviations between real values and predicted values, for example, 298.40895 and 305.5625. Why some deviation values are

[3] In statistics, the Standard Deviation (SD) is a measure that is used to quantify the amount of variation or dispersion of a set of data values.

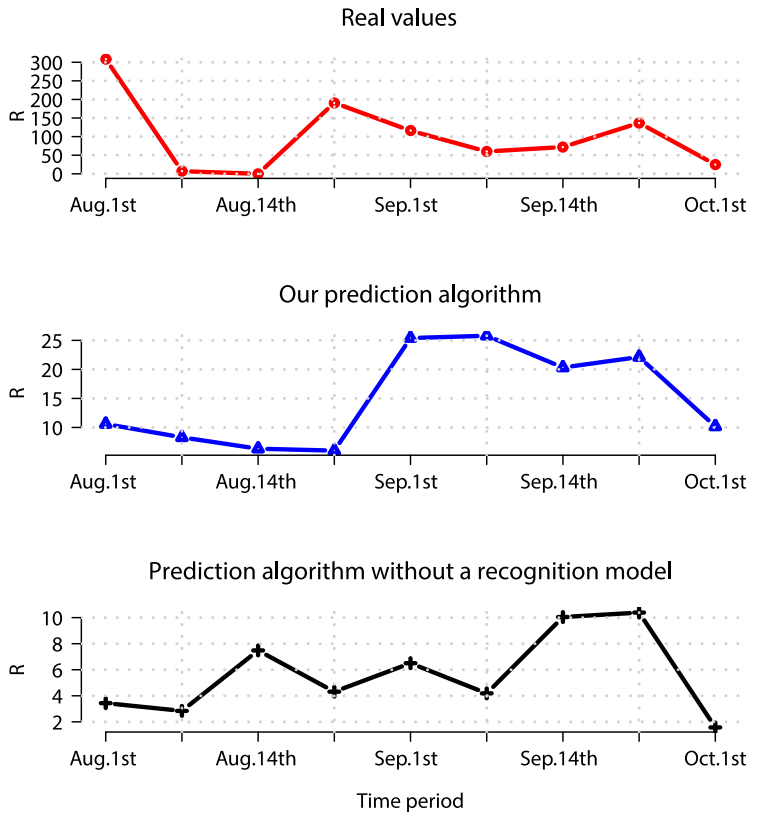

Fig. 6. Comparative experiment results. Nine time periods from August to October are selected, and then nine dynamic networks are constructed. In each time period, based on the corresponding dynamic network, the predicted results from two algorithms are illustrated in the second and third sub-figures, respectively. The first sub-figure provides the real values of R, which are counted from the collected outbreak data.

Table 3
Values corresponding to Fig. 6 and relevant three standard deviations.

	Real value	Our prediction algorithm	Prediction algorithm that does not integrate a recognition model
	309	10.59105	3.4375
	7	8.299401	2.842105
	0	6.337838	7.482759
	191	6.009524	4.315315
	116	25.39712	6.503185
	60	25.79491	4.186813
	72	20.32645	10.04628
	137	22.13499	10.3887
	25	10.17526	1.571429
Standard deviation	100.133	8.271831	3.146532

Table 4
Deviations between $\mathfrak{R}$ and $\mathfrak{I}/\mathfrak{U}$.

Deviation between $\mathfrak{R}$ and $\mathfrak{I}$	Deviation between $\mathfrak{R}$ and $\mathfrak{U}$	Time period
298.40895	305.5625	from 1st August to 7th August
1.299401	4.157895	from 8th August to 14th August
6.337838	7.482759	from 15th August to 21st August
184.990476	186.684685	from 22nd August to 31st August
90.60288	109.496815	from 1st September to 7th September
34.20509	55.813187	from 8th September to 14th September
51.67355	61.95372	from 15th September to 21st September
114.86501	126.6113	from 22nd September to 30th September
14.82474	23.428571	from 1st October to 7th October

so big? During the outbreak of an epidemic, we can only obtain a subnetwork of the complete RCN to carry out prediction, based on the collected outbreak data, because of the limited number of mobile devices carried by volunteers. It makes the number of individuals in the subnetwork to be less than the total number of individuals (there are contact among these individuals during an epidemic). On the basis of this reason, the predicted values calculated by two algorithms are much less than the real values. If we can acquire a complete RCN during an epidemic, the prediction accuracy of our algorithm can continue to be improved. However, in addition to the absolute value of R, the changing trend is very important to reflect the prediction accuracy of an algorithm as well. From Fig. 6, the prediction results of our algorithm can better follow the real changing trend.

7. Conclusion

In this article, we have evaluated the impact of network structure on disease dynamics, and based on the structure knowledge mined by our recognition model, a prediction algorithm has been proposed. The evaluation about "impact" is based on the mobile data collected from the real physical world. As a key result of this evaluation, we have observed that in a dynamic environment the network structure of an RCN varies in different time periods, and different structures have different impacts on disease dynamics. On this basis, we have designed a model to recognize the structure of a network. By using the model, we have proposed a prediction algorithm to predict the disease dynamics during an epidemic. By evaluating and comparing the accuracy of prediction for the time-varying reproductive number R, we have verified that our prediction algorithm can improve the prediction accuracy by using realistic structure knowledge that is mined by our recognition model. Moreover, in the comparison, the predicted results for R by two algorithms (our algorithm and the algorithm used to compare with our algorithm) have been compared with the real values of R. Such real values are counted from the collected surveillance data.

Acknowledgment

This work is supported by 2013 Special Fund of Guangdong Higher School Talent Recruitment, Educational Commission of Guangdong Province, China Project No. 2013KJCX0131, Guangdong High-Tech Development Fund No. 2013B010401035, National Natural Science Foundation of China Grant No. 61401107.

References

[1] Contact tracing app for the ebola outbreak in west africa, http://www.appsagainstebola.org/#the-app.
[2] Ebola situation reports, http://apps.who.int/ebola/ebola-situation-reports.
[3] Graphx of spark, http://spark.apache.org/docs/latest/graphx-programming-guide.html#overview.
[4] G.M. Ames, D.B. George, C.P. Hampson, A.R. Kanarek, C.D. McBee, D.R. Lockwood, J.D. Achter, C.T. Webb, Using network properties to predict disease dynamics on human contact networks, Proc. R. Soc. B (2011) 1–7.
[5] S. Bansal, B.T. Grenfell, L.A. Meyers, When individual behaviour matters: homogeneous and network models in epidemiology, J. R. Soc. Interf. 4 (16) (2007) 879–891.
[6] A. Barrat, M. Barthelemy, R. Pastor-Satorras, A. Vespignani, The architecture of complex weighted networks, Proc. Nation. Acad. Sci. U S A 101 (11) (2004) 3747–3752.
[7] J. Bartlett, M.J. Plank, Epidemic dynamics on random and scale-free networks, ANZIAM J. 54 (1–2) (2012) 3–22.
[8] S. Brennan, A. Sadilek, H. Kautz, Towards understanding global spread of disease from everyday interpersonal interactions, in: Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, AAAI Press, 2013, pp. 2783–2789.
[9] A. Cardillo, C. Reyes-Suárez, F. Naranjo, J. Gómez-Gardeñes, Evolutionary vaccination dilemma in complex networks, Phys. Rev. E 88 (3) (2013) 032803.
[10] L.E. Charles-Smith, T.L. Reynolds, M.A. Cameron, M. Conway, E.H. Lau, J.M. Olsen, J.A. Pavlin, M. Shigematsu, L.C. Streichert, K.J. Suda, et al., Using social media for actionable disease surveillance and outbreak management: A systematic literature review, PloS one 10 (10) (2015) e0139701.
[11] M.E. Craft, Infectious disease transmission and contact networks in wildlife and livestock, Phil. Trans. R. Soc. B 370 (1669) (2015) 20140107.
[12] G.F. De Arruda, A.L. Barbieri, P.M. Rodríguez, F.A. Rodrigues, Y. Moreno, L. da Fontoura Costa, Role of centrality for the identification of influential spreaders in complex networks, Phys. Rev. E 90 (3) (2014) 032812.
[13] L. Eisen, R.J. Eisen, Using geographic information systems and decision support systems for the prediction, prevention, and control of vector-borne diseases, Ann.Rev.Entomol. 56 (2011) 41–61.
[14] S. Eubank, H. Guclu, V.A. Kumar, M.V. Marathe, A. Srinivasan, Z. Toroczkai, N. Wang, Modelling disease outbreaks in realistic urban social networks, Nature 429 (6988) (2004) 180–184.
[15] Y.J. Fan, Y.H. Yin, L. Da Xu, Y. Zeng, F. Wu, Iot-based smart rehabilitation system, IEEE Trans. Ind. Inf. 2 (10) (2014) 1568–1577.
[16] C. Groendyke, D. Welch, D.R. Hunter, A network-based analysis of the 1861 hagelloch measles data, Biometrics 68 (3) (2012) 755–765.
[17] H. Heesterbeek, R.M. Anderson, V. Andreasen, S. Bansal, D. De Angelis, C. Dye, K.T. Eames, W.J. Edmunds, S.D. Frost, S. Funk, et al., Modeling infectious disease dynamics in the complex landscape of global health, Science 347 (6227) (2015). aaa4339–1–aaa4339–10
[18] J. Juang, Y.-H. Liang, Analysis of a general sis model with infective vectors on the complex networks, Physica A 437 (2015) 382–395.
[19] A. Kumar, G. Hancke, A zigbee-based animal health monitoring system, Sensors J. IEEE 15 (1) (2015) 610–617.
[20] L. Li, R.-L. Ge, S.-M. Zhou, R. Valerdi, Guest editorial: Integrated healthcare information systems, Information Technology in Biomedicine, IEEE Transactions on 16 (4) (2012) 515–517.
[21] D.A. Luke, K.A. Stamatakis, Systems science methods in public health: dynamics, networks, and agents, Ann. Rev. Public Health 33 (2012) 357–376.
[22] J. Marro, R. Dickman, Nonequilibrium phase transitions in lattice models, second, Cambridge University Press, 2005.
[23] S. Meyer, L. Held, et al., Power-law models for infectious disease spread, Ann. Appl. Stat. 8 (3) (2014) 1612–1639.
[24] Z. Mukandavire, S. Liao, J. Wang, H. Gaff, D.L. Smith, J.G. Morris, Estimating the reproductive numbers for the 2008–2009 cholera outbreaks in zimbabwe, Proc. Nation. Acad. Sci. 108 (21) (2011) 8767–8772.
[25] I.J. Myung, Tutorial on maximum likelihood estimation, Journal of mathematical Psychology 47 (1) (2003) 90–100.
[26] M. Newman, C.R. Ferrario, Interacting epidemics and coinfection on contact networks, PloS one 8 (8) (2013) e71321.
[27] M.E. Newman, Assortative mixing in networks, Phys. Rev. Lett. 89 (20) (2002) 208701.
[28] M.E. Newman, Mixing patterns in networks, Phys. Rev. E 67 (2) (2003) 026126.
[29] A. Saumell-Mendiola, M.Á. Serrano, M. Boguñá, Epidemic spreading on interconnected networks, Phys. Rev. E 86 (2) (2012) 026106.

[30] Y. Shang, Mixed si(r) epidemic dynamics in random graphs with general degree distributions, Appl. Math. Comput. 219 (10) (2013) 5042–5048.
[31] T. Stadler, R. Kouyos, V. von Wyl, S. Yerly, J. Böni, P. Bürgisser, T. Klimkait, B. Joos, P. Rieder, D. Xie, et al., Estimating the basic reproductive number from viral sequence data, Mole.Biol. Evol. 29 (1) (2012) 347–357.
[32] G.Q. Sun, A. Chakraborty, Q.-X. Liu, Z. Jin, K.E. Anderson, B.-L. Li, Influence of time delay and nonlinear diffusion on herbivore outbreak, Commun. Nonlinear Sci. Numer. Simulation 19 (5) (2014) 1507–1518.
[33] G.-Q. Sun, Q.-X. Liu, Z. Jin, A. Chakraborty, B.-L. Li, Influence of infection rate and migration on extinction of disease in spatial epidemics, J. Theor. Biol. 264 (1) (2010) 95–103.
[34] W.E.R. Team, et al., Ebola virus disease in west africaí–the first 9 months of the epidemic and forward projections, N. Engl. J. Med. 371 (16) (2014) 1481–1495.
[35] G.M. Vazquez-Prokopec, D. Bisanzio, S.T. Stoddard, V. Paz-Soldan, A.C. Morrison, J.P. Elder, J. Ramirez-Paredes, E.S. Halsey, T.J. Kochel, T.W. Scott, 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) (2013) e58802.
[36] M. Woolhouse, How to make predictions about future infectious disease risks, Philosoph. Trans. R. Soc. B 366 (1573) (2011) 2045–2054.
[37] C.-y. Xia, Z. Wang, J. Sanz, S. Meloni, Y. Moreno, Effects of delayed recovery and nonuniform transmission on the spreading of diseases in complex networks, Physica A 392 (7) (2013) 1577–1585.
[38] W. Yang, M. Lipsitch, J. Shaman, Inference of seasonal and pandemic influenza transmission dynamics, Proc. Nation. Acad. Sci. 112 (9) (2015) 2723–2728.