Applied Network Science

RESEARCH Open Access

Inferring network properties based on the epidemic prevalence

Long Ma, Qiang Liu* and Piet Van Mieghem

*Correspondence: Q.L.Liu@tudelft.nl
Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, PO Box 5031, 2600 GA, Delft, The Netherlands

Abstract
Dynamical processes running on different networks behave differently, which makes the reconstruction of the underlying network from dynamical observations possible. However, to what level of detail the network properties can be determined from incomplete measurements of the dynamical process is still an open question. In this paper, we focus on the problem of inferring the properties of the underlying network from the dynamics of a susceptible-infected-susceptible epidemic and we assume that only a time series of the epidemic prevalence, i.e., the average fraction of infected nodes, is given. We find that some of the network metrics, namely those that are sensitive to the epidemic prevalence, can be roughly inferred if the network type is known. A simulated annealing link-rewiring algorithm, called SARA, is proposed to obtain an optimized network whose prevalence is close to the benchmark. The output of the algorithm is applied to classify the network types.

Introduction

Graphs are the underlying structures of many systems and many dynamic processes on those systems can be modeled by a spreading process on their underlying graphs (Pastor-Satorras et al. 2015; Anderson et al. 1992; Harris 1974). The difference in the underlying graphs may lead to contrasting dissimilar behavior of the process. One well-known result is that the mean-field epidemic threshold of the spreading process vanishes with the size of the scale-free network (Pastor-Satorras and Vespignani 2001; Chatterjee and Durrett 2009), while the threshold of a sparse homogeneous network is non-zero. Another key difference is that a near-threshold spreading process is localized just above the threshold in a heterogeneous network, but delocalized in a homogeneous network (Goltsev et al. 2012; Liu and Mieghem 2019). Moreover, the autocorrelation of the infection state of each node in a regular graph is irrelevant to the curing rate in the steady state (Liu and Van Mieghem 2018). In a real scenario, reviewing of the spreading data of cholera in London in 1854 under the susceptible-infected-susceptible (SIS) model indicates that the trajectory of the prevalence reflecting network properties supporting the hypotheses that the Broad Street pump was the source of the cholera outbreak and that cholera does not spread via the air (Paré et al. 2018). Since the dynamics of different networks behave differently, the inverse question raises: "How much can we deduce about the underlying contact network by measuring the dynamics on the network?" The inverse question is meaningful when the direct measurement of the underlying graph is unavailable. For example, a disease control agency usually has the statistics of disease infection, but the underlying graph bearing the spreading of the disease is generally unknown.

Much work on the inverse problem exists (Mateos et al. 2019; Dong et al. 2019). Most of the papers focus on reconstructing the underlying graphs by measuring the time-dependent dynamical state of each node (Shandilya and Timme 2011; Berry et al. 2012; Timme and Casadiego 2014; Nitzan et al. 2017; Prasse and Van Mieghem 2018; Netrapalli and Sanghavi 2012; Myers and Leskovec 2010; Sefer and Kingsford 2015; Gomez Rodriguez et al. 2010). With the complete dynamics of each node, the network may be approximately reconstructed by different heuristic algorithms, e.g., the Bayesian methods (Friston 2002; Pajevic and Plenz 2009), the conflict-based method (Ma et al. 2015), statistical inference based method (Ma et al. 2018) and the compressed sensing or lasso methods (Shen et al. 2014). Different networked dynamical processes have been studied, such as the evolutionary game model (Han et al. 2015; Li et al. 2017), the SIS model (Shen et al. 2014) and the Ising model (Ma et al. 2018). Apart from reconstructing simple networks, there are many works on the reconstruction of the stochastic temporal networks (Li and Li 2017), multilayer networks (Mei et al. 2018), weighted networks (Ching et al. 2015) and directed networks (Hempel et al. 2011).

All of the above methods are based on the data from all or at least most nodes, but in real scenarios, individual-level observations of spreading are hard to obtain while most of the epidemic data are population-level (Shaman and Kohn 2009; Shaman et al. 2010). Motivated by the incompleteness of realistic situations, we study how much about the underlying network can be deduced *with incomplete measurements*. We assume that only the prevalence, which is the average fraction of infected nodes in the network is measured, but not the infection state of each node. Under this setting, network reconstruction does not seem possible, but inferring some network properties may be possible, in particular, when additional information apart from the prevalence is available. In this work, we confine ourselves to four types of classical network models: the scale-free (SF) graphs (Goh et al. 2001), the Barabási-Albert (BA) graphs (Barabási and Albert 1999), the Erdős-Rényi (ER) random graphs (Erdős and Rényi 1959) and the Watts-Strogatz (WS) small-world graphs (Watts and Strogatz 1998). The network size N of these networks considered in this work is not larger than 2000. Additionally, we focus on the SIS epidemic process on networks, which is one of the basic models resembling the dynamics of many networked systems and assume that the infection and curing rate of the SIS process are known. Under our assumptions, part of the network properties can be inferred, provided that the network type is additionally given. Furthermore, the network type among the four above-mentioned graphs can be identified, given the network size N and the average degree $E[D]$, which is also emphasized by recent work from a different approach (Di Lauro et al. 2019): the ER, regular and BA graphs are distinguished by the epidemic prevalence.

The paper is organized as following: In "The SIS process" section, we briefly review the SIS process on networks. In "Correlations between the SIS prevalence and network metrics" section, we evaluate the correlation between the network metric difference and the corresponding SIS prevalence difference given the network type. A high correlation implies that, if an estimated network, whose prevalence is close to the benchmark prevalence, can be found, then the metric of this estimated network may be also close to the metric of the benchmark network. We further verify the possibility of estimating the network metrics, whose differences are highly correlated with the prevalence difference. In "Distinction between network types" section, we propose a simulated annealing link-rewiring algorithm (SARA) to find a possible network whose prevalence is close to the benchmark. The output of the algorithm is applied to classify the network types. In "Estimating the topology of small networks and prevalence" section, we test the performance of SARA by inferring the structure of small networks and by forecasting the future trend of the prevalence. Finally, we conclude in "Summary" section.

The SIS process

We consider the SIS process on an unweighed, undirected network without self-loops. In the network, all the nodes are divided into two compartments: infected nodes and susceptible (healthy) nodes. An infected node can infect each healthy neighbor with rate β and the infected node can be cured spontaneously with rate δ, both as Poisson processes. If we denote the infection state of node i at time t by a Bernoulli random variable $X_i(t)$, with $X_i(t) = 1$ being infected and $X_i(t) = 0$ being healthy, the exact SIS process of node i in an N-node network is governed by the following equation,

$$\frac{dE[X_i(t)]}{dt} = E\left[-\delta X_i(t) + [1 - X_i(t)]\,\beta \sum_{k=1}^{N} a_{ki} X_k(t) \right], \tag{1}$$

where $a_{ki} \in \{0, 1\}$ is the element of the adjacency matrix A of the network. In the brackets of the right-hand side of (1), the first term represents the curing process and the second term represents the infection process. If the effective infection rate $\tau \triangleq \beta/\delta$ is above an epidemic threshold, then the infection can persist in the network; below the threshold, the epidemic dies out exponentially fast for sufficiently long time. The endemic phase and all-healthy phase are identified by the time-dependent prevalence $y(t) = \frac{1}{N}\sum_{i=1}^{N} E[X_i(t)]$. In this paper, the SIS prevalence is generated by an event-driven simulation based on the Gillespie algorithm (Gillespie 1977; Liu and Van Mieghem 2017; St-Onge et al. 2019).

Correlations between the SIS prevalence and network metrics

Preliminaries

Two different networks may produce a similar prevalence, and thus we need to understand which network properties are important factors in the SIS process. If the SIS prevalence is sensitive to a specific network metric, then the prevalence generated by two networks with different values of this metric may be distinct. Assume that we have a benchmark network with a metric M_b and an estimated network with the metric M_e. If the time series of the prevalence on the benchmark and estimated networks are $\{y_b(i\Delta t)\}_{i=0,\ldots,T-1}$ and $\{y_e(i\Delta t)\}_{i=0,\ldots,T-1}$, respectively, then their correlation can be evaluated by computing the prevalence difference

$$\mathcal{D}_p \triangleq \frac{1}{T}\sum_{i=0}^{T-1} \left| y_e(i\Delta t) - y_b(i\Delta t) \right| \tag{2}$$

and the metric difference

$$\mathcal{D}_G \triangleq |M_e - M_b|. \tag{3}$$

If we have n corresponding realizations of the differences $(\mathcal{D}_{pi}, \mathcal{D}_{Gi})$ for $i = 1, \ldots, n$, then we can compute their correlation by the Pearson correlation coefficient (Van Mieghem 2014, p. 26),

$$\rho(\mathcal{D}_p, \mathcal{D}_G) \triangleq \frac{\sum\limits_{i=1}^{n}(\mathcal{D}_{pi} - \overline{\mathcal{D}_p})(\mathcal{D}_{Gi} - \overline{\mathcal{D}_G})}{\sqrt{\sum\limits_{i=1}^{n}(\mathcal{D}_{pi} - \overline{\mathcal{D}_p})^2}\sqrt{\sum\limits_{i=1}^{n}(\mathcal{D}_{Gi} - \overline{\mathcal{D}_G})^2}}. \tag{4}$$

Only if $\rho(\mathcal{D}_p, \mathcal{D}_G)$ approaches one, then the metric M and the prevalence $y(t)$ are highly correlated, which indicates that inferring the metric from the prevalence may be possible.

Evaluated network metrics

The graph metrics considered in this section are shown in Table 1.

The assortativity ρ_D, which is the degree correlation between connected nodes (Van Mieghem et al. 2010), can be calculated as

$$\rho_D = 1 - \frac{\sum_{i \sim j}(d_i - d_j)^2}{\sum\limits_{i=1}^{N} {d_i}^3 - \frac{(\sum\limits_{i=1}^{N} {d_i}^2)^2}{2L}}, \tag{5}$$

where d_i and d_j are the degrees of nodes at the end of a link $i \sim j$, and L is the number of links.

The average clustering coefficient C_G, which is the probability that the node pairs with same neighbors are also connected, can be computed as

$$C_G = \frac{1}{N}\sum_{i=1}^{N} C_i = \frac{1}{N}\sum_{i=1}^{N}\frac{2\blacktriangle_i}{d_i(d_i - 1)},$$

where $\blacktriangle_i$ is the number of triangles containing node i.

Some of the above metrics can be strongly correlated with the prevalence $y(t)$. For example, the epidemic threshold τ_c^{HMF} derived from the heterogeneous mean-field (HMF) approach (Pastor-Satorras et al. 2015) is

$$\tau_c^{\text{HMF}} = \frac{E[D]}{E[D^2]},$$

Table 1 Graph metrics

N	Network size (the number of nodes)
$E[D]$	Average degree
$E[D^2]$	Second moment of degree
$d_{\max}$	Largest degree
$E[H]$	Average shortest path length (the average hop-count)
$E[1/H]$	Global efficiency
λ_1	Spectral radius (the largest eigenvalue of the adjacency matrix)
μ_{N-1}	Algebraic connectivity (the second smallest eigenvalue of the Laplacian matrix)
ρ_D	Assortativity
C_G	Average clustering coefficient

where D is the degree of a randomly selected node and the epidemic threshold $\tau_c^{(1)}$ derived from NIMFA (Mieghem et al. 2009) is

$$\tau_c^{(1)} = \frac{1}{\lambda_1}.$$

Many graph metrics can also be bounded. For example, the average degree follows $E[D] \leqslant \lambda_1$ in connected graphs (Van Mieghem 2010) and the largest eigenvalue of the Laplacian matrix $\mu_1 \geqslant \frac{N}{N-1} d_{\max}$, while the algebraic connectivity is $\mu_{N-1} \leqslant d_{\min}$.

Correlation analysis

For any pair of networks, the prevalence difference $\mathcal{D}_p$ and the metric difference $\mathcal{D}_G$ can be calculated based on (2) and (3). For each network metric, we calculate the correlations via (4) between a set of metric differences $\mathcal{D}_G$ and their corresponding prevalence differences $\mathcal{D}_p$ on four network models: the SF graphs (Goh et al. 2001), the BA graphs (Barabási and Albert 1999), the ER random graphs (Erdős and Rényi 1959) and the WS small-world graphs (Watts and Strogatz 1998). Specifically, the SF graphs are generated by the configuration model (Goh et al. 2001; Catanzaro and Pastor-Satorras 2005) and the degree exponent parameter γ is uniformly at random chosen in the interval $[2.5, 3.0]$ in this paper.

Specifically, we first randomly generate the four kinds of networks each with 100 realizations. The network sizes N and the average degrees $E[D]$ are chosen uniformly at random in the interval $[1000, 2000]$ and $[4, 12]$, respectively. The effective infection rate is set as $\tau = 3.0$, which is above the epidemic threshold of every network realization. Two kinds of initial state are chosen: $y_0 = 0.2$ or $y_0 = 1.0$, which means that 20% of the nodes are randomly chosen to be infected or all nodes are infected initially. For each network and initial state, a corresponding time series of the prevalence is obtained by averaging over 100 realizations of the SIS simulation. We mark the prevalence difference $\mathcal{D}_p$ under initial condition y_0 as $\mathcal{D}_p(y_0)$. We further denote the metric difference $\mathcal{D}_G$ for one specific metric as $\mathcal{D}_G(\text{metric})$. All metrics shown in "Evaluated network metrics" section are considered and the Pearson correlation coefficients $\rho\left(\mathcal{D}_p(y_0), \mathcal{D}_G(\text{metric})\right)$ are calculated by Eq. (4). The sample size of each correlation coefficient is $\binom{100}{2} = 4950$. Table 2 and Table 3 indicate that there are generally strong correlations between the difference of the prevalence $\mathcal{D}_p$ and the differences of the average degree $E[D]$, the second moment of degree $E[D^2]$, the average shortest path length $E[H]$, the global efficiency $E[1/H]$ and the spectral radius λ_1. A strong positive correlation indicates that the metric between two networks with the same network type can be similar if they have similar prevalence curves.

Table 2 Matrics with strong positive correlations

$\rho\left(\boldsymbol{\mathcal{D}}_p(y_0), \boldsymbol{\mathcal{D}}_G(\text{metric})\right)$	$\boldsymbol{\mathcal{D}}_G(E[D])$	$\boldsymbol{\mathcal{D}}_G(E[D^2])$	$\boldsymbol{\mathcal{D}}_G(\lambda_1)$	$\boldsymbol{\mathcal{D}}_G(E[H])$	$\boldsymbol{\mathcal{D}}_G\left(E\left[\frac{1}{H}\right]\right)$
ER graphs, $\boldsymbol{\mathcal{D}}_p(y_0 = 0.2)$	0.941	0.856	0.940	0.953	0.939
WS graphs, $\boldsymbol{\mathcal{D}}_p(y_0 = 0.2)$	0.877	0.826	0.921	0.952	0.958
BA graphs, $\boldsymbol{\mathcal{D}}_p(y_0 = 0.2)$	0.940	0.838	0.871	0.952	0.945
SF graphs, $\boldsymbol{\mathcal{D}}_p(y_0 = 0.2)$	0.944	0.612	0.561	0.861	0.823
ER graphs, $\boldsymbol{\mathcal{D}}_p(y_0 = 1.0)$	0.947	0.866	0.944	0.948	0.932
WS graphs, $\boldsymbol{\mathcal{D}}_p(y_0 = 1.0)$	0.905	0.818	0.927	0.952	0.954
BA graphs, $\boldsymbol{\mathcal{D}}_p(y_0 = 1.0)$	0.945	0.856	0.908	0.954	0.948
SF graphs, $\boldsymbol{\mathcal{D}}_p(y_0 = 1.0)$	0.948	0.631	0.459	0.792	0.783

Table 3 Matrics with weak positive correlations

$\rho\left(\boldsymbol{\mathcal{D}}_p(y_0), \boldsymbol{\mathcal{D}}_G(\text{metric})\right)$	$\boldsymbol{\mathcal{D}}_G(d_{\max})$	$\boldsymbol{\mathcal{D}}_G(C_G)$	$\boldsymbol{\mathcal{D}}_G(\mu_{N-1})$	$\boldsymbol{\mathcal{D}}_G(\rho_D)$	$\boldsymbol{\mathcal{D}}_G(N)$
ER graphs, $\boldsymbol{\mathcal{D}}_p(y_0 = 0.2)$	0.821	0.477	0.490	−0.014	−0.059
WS graphs, $\boldsymbol{\mathcal{D}}_p(y_0 = 0.2)$	0.805	−0.036	−0.002	0.624	−0.012
BA graphs, $\boldsymbol{\mathcal{D}}_p(y_0 = 0.2)$	0.386	0.358	0.854	0.595	−0.031
SF graphs, $\boldsymbol{\mathcal{D}}_p(y_0 = 0.2)$	0.398	0.182	0.657	0.013	−0.038
ER graphs, $\boldsymbol{\mathcal{D}}_p(y_0 = 1.0)$	0.856	0.525	0.524	0.082	−0.018
WS graphs, $\boldsymbol{\mathcal{D}}_p(y_0 = 1.0)$	0.807	−0.031	0.081	0.666	−0.039
BA graphs, $\boldsymbol{\mathcal{D}}_p(y_0 = 1.0)$	0.284	0.410	0.813	0.535	−0.003
SF graphs, $\boldsymbol{\mathcal{D}}_p(y_0 = 1.0)$	0.247	0.100	0.659	0.006	0.034

However, there are relatively weak correlations between the difference of the prevalence $\mathcal{D}_p$ and the differences of the network size N, the largest degree $d_{\max}$, the algebraic connectivity μ_{N-1}, the assortativity ρ_D and the average clustering coefficient C_G. Moreover, the initial state has very slightly influence on the correlations.

To summarize, if the type of the underlying graph is given, then inferring the network properties, whose differences $\mathcal{D}_G$ are highly correlated to the difference of the prevalence $\mathcal{D}_p$, is possible. A straightforward method is randomly generating the network realizations by the corresponding network model and selecting the one realization produces minimum prevalence difference $\mathcal{D}_p$.

Inferring network metrics given the network type

We further try to infer the network metrics based on the prevalence from a single realization of the SIS process given the network type. Specifically, for each network type, we first generate 1000 benchmark networks whose network sizes N and average degrees $E[D]$ are chosen uniformly at random in the interval $[200, 500]$ and $[4, 8]$, respectively. For each benchmark network, one corresponding benchmark prevalence is generated from only one realization of the SIS process.

We then try to estimate the network metrics of each benchmark network as follows. For each benchmark, 1000 networks with the same network type as the benchmark network are generated. The network sizes N and average degrees $E[D]$ of the generated networks are also chosen uniformly at random in the interval $[200, 500]$ and $[4, 8]$, respectively. The network with the smallest prevalence difference $\mathcal{D}_p$ to the benchmark is selected as the estimated network. The metrics of this estimated network are regarded as the estimated metrics of the benchmark network.

We measure the performance of the metric inference under the mean absolute error (MAE) and the mean squared error (MSE). The MAE and MSE for n underlying graphs is given by

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |M_{ei} - M_{bi}| \tag{6}$$

and

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (M_{ei} - M_{bi})^2, \tag{7}$$

where M_{ei} and M_{bi} denote the estimated and real metrics of the benchmark network G_i, $i = 1, 2, \cdots, n$.

Tables in the Additional file 1 show MAE and MSE of each network metric for different network types (the ER random graphs, the WS small-world graphs, the BA graphs and the SF graphs). For the treatment group, we calculate MAE and MSE of each metrics which are estimated by selecting the network whose prevalence is closest to the benchmark. For the control group, we calculate MAE and MSE of each metrics which are estimated by randomly generating a network whose network sizes N and average degrees $E[D]$ are chosen uniformly at random in the interval $[200, 500]$ and $[4, 8]$. For the network metrics whose differences are closely correlated with the prevalence difference $\mathcal{D}_p$, i.e., the average degree $E[D]$, the second moment of degree $E[D^2]$, the average shortest path length $E[H]$, the global efficiency $E[1/H]$ and the spectral radius λ_1, their MAE and MSE of the treatment group are much less than those of the control group, which indicates that these metrics can be roughly deduced based on the prevalence given the network type. However, for the network metrics whose differences are weakly correlated with the prevalence difference $\mathcal{D}_p$, i.e., the network size N, the largest degree $d_{\max}$, the algebraic connectivity μ_{N-1}, the assortativity ρ_D and the average clustering coefficient C_G, their MAE and MSE of the treatment group are close to those of the control group.

Distinction between network types

In this section, we try to distinguish the type of the underlying network given the time series of the prevalence $\{y_b(i\Delta t)\}_{i=0,\ldots,T-1}$, the network size N, the number of links L and the effective infection rate τ. We propose a simulated annealing link-rewiring algorithm (SARA) to optimize a network whose prevalence can be close to the input prevalence benchmark and the performance difference between different rewiring mechanisms in SARA can be applied to identify the graph type.

Simulated annealing link-rewiring algorithm

The basic principle of SARA is that the links of an estimated network are continually rewired based on different rewiring methods to minimize the prevalence difference D_p between the optimized network and the benchmark network.

The algorithm operates iteratively and a random network is initialized. In each iteration, the network will be renewed by rewiring the links of partial nodes. A new corresponding time series of the prevalence $\{y_e(i\Delta t)\}_{i=0,\ldots,T-1}$ can be generated by simulating the SIS process on the network and its difference D_p to the benchmark time series of the prevalence $\{y_b(i\Delta t)\}_{i=0,\ldots,T-1}$ is calculated. If the difference D_p decreases, then the rewired network will be accepted. If D_p increases, then the rewired network is accepted with an acceptance probability p and rejected with rejection probability $1 - p$ to prevent local optima. Moreover, a stable final converging result is obtained provided that the acceptance probability p decreases with the iterations. The final result of this algorithm is an estimated graph, whose corresponding prevalence is almost the same as the benchmark prevalence. Inspired by the generation processes of ER graphs and BA scale-free graphs, we consider two different rewiring methods: randomly connecting (RC) and preferential attachment (PA). In RC, the selected nodes are rewired uniformly at random to the rest of the nodes in the network, and in PA, the selected nodes are rewired to a node with probability proportional to the node's degree. The pseudo-code of SARA is shown in Algorithm 1.

Algorithm 1: Pseudo-code of the simulated annealing link-rewiring algorithm (SARA)

```
   Input  : $\{y(i\Delta t)\}_{i=0,...,T-1}$, $N$, $L$, $\tau$, initial temperature $V_{\mathrm{tmp}}$, cooling rate $0 < r < 1$
            and step length $S_N$
   Output: Estimated network $G_e$, final prevalence difference $\mathcal{D}_p$
1  An initial network is chosen uniformly at random from the set of all networks with
   $N$ nodes and $L$ links.
2  for iteration bound do
3      Uniformly randomly choose $N_c = \mathrm{round}(S_N \times \mathcal{D}_p)$ nodes.
4      Delete all links of each chosen node and then rewire these links to new neighbors.
5      If we randomly choose new neighbors without preference (RC), the probability
       $p_i$ that the rewired link is connected to a neighbor $i$ is $p_i = 1/(N - N_c)$, where
       node $i$ belongs to the $N - N_c$ uncollected nodes.
6      If we rewire links based on preferential attachment mechanism (PA) , the
       probability $p_i$ that the rewired link is connected to a neighbor $i$ is $p_i = d_i/\sum_j d_j$,
       where $d_i$ is the degree of node $i$ in residual network and the sum is made over all
       unselected nodes.
7      If $n$ isolated nodes appear after the rewiring process in step 5 or step 6, we
       remove $n$ links uniformly at random and rewire them to the isolated nodes
       based on the RC or PA mechanism, respectively. This step continues until there
       is no isolated node in the network.
8      Simulate the SIS process on the new network and calculate the prevalence
       difference $\mathcal{D}_2$ to the benchmark.
9      if $\mathcal{D}_2 < \mathcal{D}$ then
10         $\mathcal{D} \longleftarrow \mathcal{D}_2$; $G \longleftarrow G_2$;
11     else if $Exp(-(\mathcal{D}_2 - \mathcal{D})/V_{tmp}) > random(0, 1)$ then
12         $\mathcal{D} \longleftarrow \mathcal{D}_2$; $G \longleftarrow G_2$;
13     end
14     $V_{\mathrm{tmp}} = r \times V_{\mathrm{tmp}}$;
15 end
```

Distinction between the network types

We try to distinguish four kinds of graphs (the SF graphs, the BA graphs, the ER random graphs and the WS small-world graphs) based on the optimized prevalence curves generated by SARA. The experiment and the results are as follows. For each network model, we generate 100 network realizations with $N = 1000$ nodes and $L = 4000$ links as the benchmark networks. For the SF graphs, the degree exponent γ ranges in the interval $\gamma \in [2.5, 3.0]$. For the SW graphs, the rewiring probability $p_r \in [0.5, 1.0]$. The corresponding time series of the prevalence are obtained by averaging 10 realizations with effective infection rate $\tau = 1$, which is above the epidemic threshold for benchmark networks. For each benchmark graph realization and corresponding prevalence, we apply SARA with RC and PA mechanisms separately and obtain two corresponding prevalence differences $\mathcal{D}_{\mathrm{RC}}$ and $\mathcal{D}_{\mathrm{PA}}$ from the final output of the optimization, respectively. The performance difference between these two rewiring mechanisms provides a possibility of identifying the types of underlying graphs by applying different rewiring methods in SARA. We then try to classify the networks by the difference value $\mathcal{D}_{\mathrm{RC}} - \mathcal{D}_{\mathrm{PA}}$. Figure 1 shows that these four kinds of networks can be almost exactly classified by the difference value $\mathcal{D}_{\mathrm{RC}} - \mathcal{D}_{\mathrm{PA}}$. Indeed, $\mathcal{D}_{\mathrm{RC}} > \mathcal{D}_{\mathrm{PA}}$ for almost all SF and BA graphs while $\mathcal{D}_{\mathrm{RC}} < \mathcal{D}_{\mathrm{PA}}$ for almost all ER and SW graphs as shown in Fig. 1a. We exam the classification performance by the receivers operating characteristic (ROC) curve, which is a curve of the True Positive Rate (TPR)

$$R_{\mathrm{TPR}}(d) = \frac{N_{\mathrm{TP}}(d)}{N_{\mathrm{TP}}(d) + N_{\mathrm{FN}}(d)}$$

against the False Positive Rate (FPR)

$$R_{\mathrm{FPR}}(d) = \frac{N_{\mathrm{FP}}(d)}{N_{\mathrm{FP}}(d) + N_{\mathrm{TN}}(d)},$$

where d is the threshold of the difference value $\mathcal{D}_{\mathrm{RC}} - \mathcal{D}_{\mathrm{PA}}$, $N_{\mathrm{TP}}(d)$ is the number of true positives of $\mathcal{D}_{\mathrm{RC}} - \mathcal{D}_{\mathrm{PA}} > d$, $N_{\mathrm{FP}}(d)$ is the number of false positives of $\mathcal{D}_{\mathrm{RC}} - \mathcal{D}_{\mathrm{PA}} > d$.

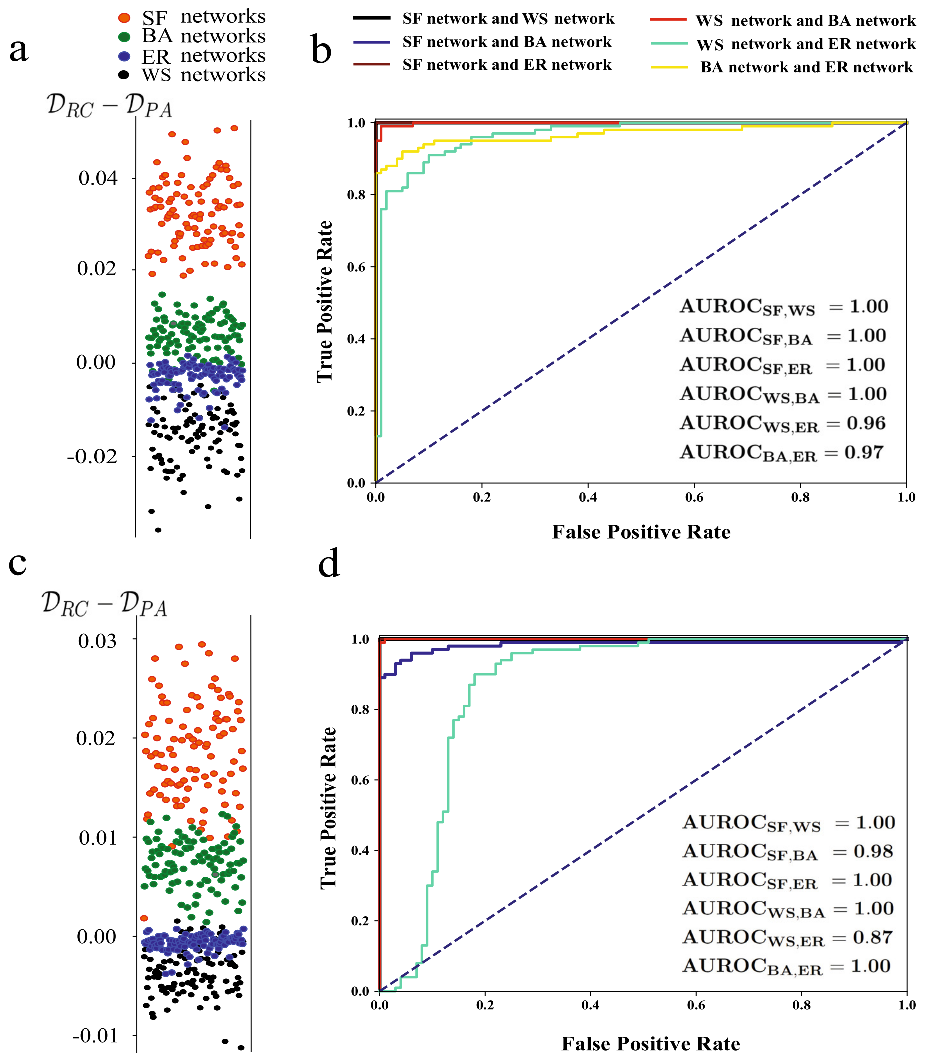

Fig. 1 The classification of network types. **a** and **b**: The results for the initial state $y_0 = 0.2$. **c** and **d**: The results for the initial state $y_0 = 1.0$. The time series of the prevalence are obtained by averaging over 10 realizations

The denominators $N_{\mathrm{TP}}(d)+N_{\mathrm{FN}}(d)$ and $N_{\mathrm{FP}}(d)+N_{\mathrm{TN}}(d)$ are the number of real positives and real negatives of $\mathcal{D}_{\mathrm{RC}} - \mathcal{D}_{\mathrm{PA}} > d$, respectively.

The area under the ROC curve ($AUROC$) depicts the accuracy of classification. If $AUROC = 1$, then the classification is perfect. In Fig. 1b and Fig. 1d, the ROC curves of the difference value $\mathcal{D}_{\mathrm{RC}} - \mathcal{D}_{\mathrm{PA}}$ between any two kinds of networks show that these networks can be distinguished almost exactly.

Estimating the topology of small networks and prevalence

The network output of SARA: an example

In this section, we test the feasibility of approximately reconstructing small graphs from the prevalence. We show example output of SARA under the benchmark of a small tree network and a small wheel network. In SARA, the initialized networks are chosen uniformly at random from all networks with the same number of nodes and links as the benchmark networks. The rewiring methods are selected to be the one with a smaller difference of the prevalence in the output. As shown in Fig. 2, the main features of the benchmark networks are captured fairly well by the final output of SARA.

Forecast the future trend of epidemic prevalence

Any benchmark prevalence from either homogeneous or heterogeneous networks can be fitted well by SARA. Therefore, we can further analyze the feasibility of predicting the future prevalence evolution by fitting the few initial prevalence observations.

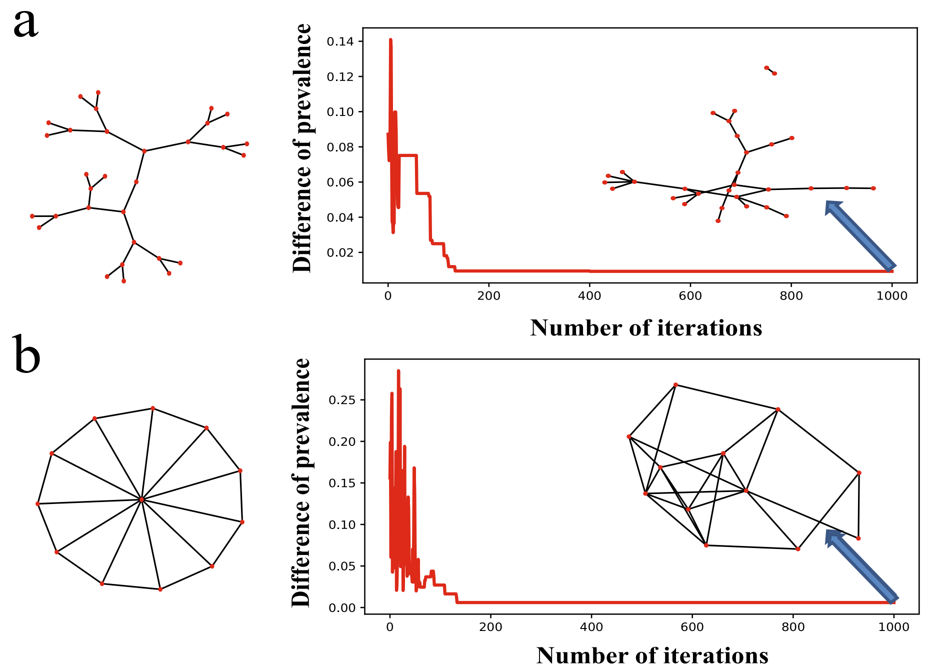

Fig. 2 The reconstruction of a tree network and a wheel network. The left parts are the benchmark underlying network and the right parts are the estimated networks. The curves are the difference of prevalence against the number of iterations. The difference of prevalence is already small when the number of iterations is around 150. The prevalence curves are obtained by averaging 500 realizations and only the central node is infected initially

We fit only the initial part (10%) of the time series of the prevalence $\{y(i\Delta t)\}_{i=0,\ldots,\lfloor T/10\rfloor-1}$ generated by four different benchmark networks and compare the whole prevalence output of the algorithm with the benchmark prevalence. RC rewiring is applied for ER and WS graphs, and PA rewiring is applied for BA and SF graphs. As shown in Fig. 3a about the ER and WS graphs, the estimated prevalence (dashed curves) are close to the benchmark (solid curves). However, as shown in Fig. 3b, the prediction is inaccurate for BA and SF graphs.

Summary

We study the feasibility of inferring properties of the underlying graphs based on the SIS prevalence. Pearson's correlations (4) between the differences of prevalence and the network metrics are evaluated. Given network type, the difference of the epidemic prevalence is highly related to the differences of some network metrics, such as the average degree $E[D]$, the second moment of degree $E[D^2]$, the average shortest path length $E[H]$, the global efficiency $E[1/H]$ and the spectral radius λ_1. If the network type is known, then these metrics can be roughly estimated by finding a network whose prevalence curve is close to the benchmark. To distinguish the network type, we further propose an algorithm SARA, which can find a network whose epidemic prevalence is close to the benchmark. Given network size and the number of links, four network types (the SF graphs, the BA graphs, the ER random graphs and the WS small-world graphs) can be classified by different rewiring methods combined with SARA. Visually, the output network of SARA captures the features of small benchmark networks well. Finally, we show that it is possible to predict the later prevalence from the initial stage prevalence for homogeneous networks.

The prevalence in the SIS model resembles the population-level observations. Population-level observations lose details of nodal infection but may still provide information about the underlying network. In real scenarios, the population-level observations are available for many different infectious diseases, such as influenza, Ebola virus disease, Zika virus disease, etc. Disease control agencies may take advantage of the population-level observations to understand the detailed spreading pattern, further forecast the outbreaks more accurately and control the diseases more efficiently. For example, a small

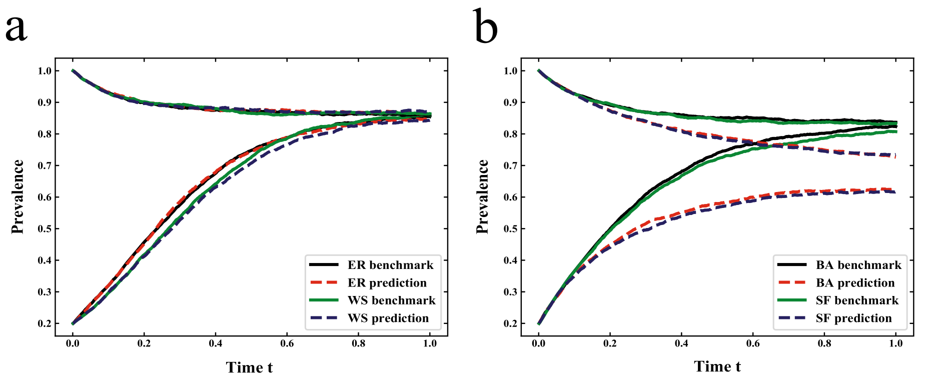

Fig. 3 Forecast the future trend of epidemic prevalence. **a** The results about ER graphs and WS graphs. **b** The results about BA graphs and SF graphs. Two kinds of initial states are chosen: $y_0 = 0.2$ and $y_0 = 1.0$. The time series of the prevalence are the mean of 10 propagation

diameter of the network inferred by the population-level observations implies that modern transportation plays a role; a large clustering coefficient means that spreading is effectively exploring a community or geographical area; using the initial stage prevalence, it is possible to approximately reconstruct the small-size local network containing the initial infections. Limitations like those in our experiments, such as the demanding of extra parameters apart from the prevalence, still exist, but on the other side, additional known knowledge, e.g. population distribution, may be available and helps the inference of the network properties.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1007/s41109-019-0218-0.

Additional file 1: Supplementary material for inferring network properties based on the epidemic prevalence.

Abbreviations

AUROC: The area under the ROC curve; BA: Barabási-Albert; ER: Erdős-Rényi; FPR: False positive rate; HMF: Heterogeneous mean-field; MAE: The mean absolute error; MSE: The mean squared error; NIMFA: The *N*-intertwined mean-field approximation; PA: Preferential attachment; RC: Randomly connecting; ROC: Receivers operating characteristic; SARA: Simulated annealing link-rewiring algorithm; SF: Scale-free; SIS: Susceptible-infected-susceptible; TPR: True positive rate; WS: Watts-strogatz

Acknowledgements

We thank Marcus Märtens for providing simulation tools.

Authors' contributions

PVM supervised the research. LM and QL designed the algorithm. LM implemented the experiment and drafted the manuscript. QL and PVM reviewed and revised the manuscript critically. All authors read and approved the final manuscript.

Funding

LM and QL are supported by China scholarship council.

Availability of data and material

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 8 February 2019 Accepted: 2 October 2019
Published online: 29 October 2019

References

Anderson RM, May RM, Anderson B (1992) Infectious Diseases of Humans: Dynamics and Control, Vol. 28. Wiley Online Library

Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512

Berry T, Hamilton F, Peixoto N, Sauer T (2012) Detecting connectivity changes in neuronal networks. J Neurosci Methods 209(2):388–397

Catanzaro M, Pastor-Satorras R (2005) Analytic solution of a static scale-free network model. Eur Phys J B-Condensed Matter Compl Syst 44(2):241–248

Chatterjee S, Durrett R (2009) Contact processes on random graphs with power law degree distributions have critical value 0. Ann Probab 37(6):2332–2356

Ching ES, Lai P-Y, Leung C (2015) Reconstructing weighted networks from dynamics. Phys Rev E 91(3):030801

Di Lauro F, Croix J, Dashti M, Berthouze L, Kiss I (2019) Network inference from population-level observation of epidemics. arXiv preprint arXiv:1906.10966

Dong X, Thanou D, Rabbat M, Frossard P (2019) Learning graphs from data: A signal representation perspective. IEEE Signal Process Mag 36(3):44–63

Erdős P, Rényi A (1959) On random graphs I. Publ Math Debrecen 6:290–297

Friston KJ (2002) Bayesian estimation of dynamical systems: an application to fMRI. NeuroImage 16(2):513–530

Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. J Phys Chem 81(25):2340–2361

Goh K-I, Kahng B, Kim D (2001) Universal behavior of load distribution in scale-free networks. Phys Rev Lett 87(27):278701

Goltsev AV, Dorogovtsev SN, Oliveira JG, Mendes JF (2012) Localization and spreading of diseases in complex networks. Phys Rev Lett 109(12):128702

Gomez Rodriguez M, Leskovec J, Krause A (2010) Inferring Networks of Diffusion and Influence. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York. pp 1019–1028. https://doi.org/10.1145/1835804.1835933

Han X, Shen Z, Wang W-X, Di Z (2015) Robust reconstruction of complex networks from sparse data. Phys Rev Lett 114(2):028701
Harris TE (1974) Contact interactions on a lattice. Ann Probab 2(6):969–988
Hempel S, Koseska A, Kurths J, Nikoloski Z (2011) Inner composition alignment for inferring directed networks from short time series. Phys Rev Lett 107(5):054101
Li X, Li X (2017) Reconstruction of stochastic temporal networks through diffusive arrival times. Nature Commun 8:15729
Li J, Shen Z, Wang W-X, Grebogi C, Lai Y-C (2017) Universal data-based method for reconstructing complex networks with binary-state dynamics. Phys Rev E 95(3):032303
Liu, Q, Van Mieghem P (2017) Evaluation of an analytic, approximate formula for the time-varying sis prevalence in different networks. Phys A: Stat Mech Appl 471:325–336
Liu Q, Van Mieghem P (2018) Autocorrelation of the susceptible-infected-susceptible process on networks. Phys Rev E 97(6):062309
Liu Q, Mieghem PV (2019) Network localization is unalterable by infections in bursts. IEEE Transactions on Network Science and Engineering:1–1. https://doi.org/10.1109/TNSE.2018.2889539
Ma L, Han X, Shen Z, Wang W-X, Di Z (2015) Efficient reconstruction of heterogeneous networks from time series via compressed sensing. PloS one 10(11):0142837
Ma C, Chen H-S, Lai Y-C, Zhang H-F (2018) Statistical inference approach to structural reconstruction of complex networks from binary time series. Phys Rev E 97(2):022301
Mateos G, Segarra S, Marques AG, Ribeiro A (2019) Connecting the dots: Identifying network structure via graph signal processing. IEEE Signal Process Mag 36(3):16–43
Mei G, Wu X, Wang Y, Hu M, Lu J-A, Chen G (2018) Compressive-sensing-based structure identification for multilayer networks. IEEE Trans Cybernet 48(2):754–764
Mieghem PV, Omic J, Kooij R (2009) Virus Spread in Networks. IEEE/ACM Trans Netw 17(1):1–14. https://doi.org/10.1109/TNET.2008.925623
Myers S, Leskovec J (2010) On the Convexity of Latent Social Network Inference. In: Adv Neural Inf Proc Syst. Curran Associates Inc., USA Vol. 2. pp 1741–1749. http://dl.acm.org/citation.cfm?id=2997046.2997090
Netrapalli P, Sanghavi S (2012) Learning the graph of epidemic cascades. SIGMETRICS Perform Eval Rev 40(1):211–222. https://doi.org/10.1145/2318857.2254783
Nitzan M, Casadiego J, Timme M (2017) Revealing physical interaction networks from statistics of collective dynamics. Sci Adv 3(2):1600396
Pajevic S, Plenz D (2009) Efficient network reconstruction from dynamical cascades identifies small-world topology of neuronal avalanches. PLoS Comput Biol 5(1):1000271
Paré PE, Liu J, Beck CL, Kirwan BE, Başar T (2018) Analysis, estimation, and validation of discrete-time epidemic processes. Transactions on Control Systems Technology, IEEE:1–15. https://doi.org/10.1109/TCST.2018.2869369
Pastor-Satorras R, Vespignani A (2001) Epidemic spreading in scale-free networks. Phys Rev Lett 86(14):3200
Pastor-Satorras R, Castellano C, Van Mieghem P, Vespignani A (2015) Epidemic processes in complex networks. Rev Modern Phys 87(3):925
Prasse B, Van Mieghem P (2018) Exact Network Reconstruction from Complete SIS Nodal State Infection Information Seems Infeasible. IEEE Trans Netw Sci Eng:1-1. https://doi.org/10.1109/TNSE.2018.2872511
Sefer E, Kingsford C (2015) Convex Risk Minimization to Infer Networks from probabilistic diffusion data at multiple scales. In: 2015 IEEE 31st International Conference on Data Engineering. IEEE. pp 663–674. https://doi.org/10.1109/ICDE.2015.7113323
Shaman J, Kohn M (2009) Absolute humidity modulates influenza survival, transmission, and seasonality. Proc Nat Acad Sci 106(9):3243–3248
Shaman J, Pitzer VE, Viboud C, Grenfell BT, Lipsitch M (2010) Absolute humidity and the seasonal onset of influenza in the continental united states. PLoS Biol 8(2):1000316
Shandilya SG, Timme M (2011) Inferring network topology from complex dynamics. New J Phys 13(1):013004
Shen Z, Wang W-X, Fan Y, Di Z, Lai Y-C (2014) Reconstructing propagation networks with natural diversity and identifying hidden sources. Nature Commun 5
St-Onge G, Young J-G, Hébert-Dufresne L, Dubé LJ (2019) Efficient sampling of spreading processes on complex networks using a composition and rejection algorithm. Comput Phys Commun 240:30–37. https://doi.org/10.1016/j.cpc.2019.02.008
Timme M, Casadiego J (2014) Revealing networks from dynamics: an introduction. J Phys A: Math Theoret 47(34):343001
Van Mieghem P (2014) Performance Analysis of Complex Networks and Systems. Cambridge University Press, Cambridge
Van Mieghem, P, Wang H, Ge X, Tang S, Kuipers FA (2010) Influence of assortativity and degree-preserving rewiring on the spectra of networks. Eur Phys J B 76(4):643–652
Van Mieghem P (2010) Graph Spectra for Complex Networks. Cambridge University Press, Cambridge
Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world'networks. Nature 393(6684):440

Publisher's Note