Inferring network properties based on the epidemic prevalence

Long Ma, Qiang Liu, and Piet Van Mieghem

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.

1 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 [1–3]. 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 [4, 5], 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 [6, 7]. Moreover, the autocorrelation of the infection state of each node in a regular graph is irrelevant to the curing rate in the steady state [8]. 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 [9]. 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

---

*Long Ma-2@tudelft.nl
†Q.L.Liu@tudelft.nl
‡P.F.A.VanMieghem@tudelft.nl
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 [10, 11]. Most of the papers focus on reconstructing the underlying graphs by measuring the time-dependent dynamical state of each node [12–20]. With the complete dynamics of each node, the network may be approximately reconstructed by different heuristic algorithms, e.g., the Bayesian methods [21, 22], the conflict-based method [23], statistical inference based method [24] and the compressed sensing or lasso methods [25]. Different networked dynamical processes have been studied, such as the evolutionary game model [26, 27], the SIS model [25] and the Ising model [24]. Apart from reconstructing simple networks, there are many works on the reconstruction of the stochastic temporal networks [28], multilayer networks [29], weighted networks [30] and directed networks [31].

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 [32, 33]. 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 [34], the Barabási-Albert (BA) graphs [35], the Erdős-Rényi (ER) random graphs [36] and the Watts-Strogatz (WS) small-world graphs [37]. 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 [38]: the ER, regular and BA graphs are distinguished by the epidemic prevalence.

The paper is organized as following: In Section 2, we briefly review the SIS process on networks. In Section 3, 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 Section 4, 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 Section 5, 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 Section 6.
2 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 $\beta$ and the infected node can be cured spontaneously with rate $\delta$, 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 [39],

$$\frac{dE[X_i(t)]}{dt} = E[-\delta X_i(t) + (1 - X_i(t))\beta \sum_{k=1}^{N} a_{ki} X_k(t)]$$  \hspace{1cm} (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 $\tau \equiv \beta/\delta$ 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 [40]. 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 [41–43].

3 Correlations between the SIS prevalence and network metrics

3.1 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

$$D_p \triangleq \frac{1}{T} \sum_{i=0}^{T-1} |y_e(i\Delta t) - y_b(i\Delta t)|$$  \hspace{1cm} (2)

and the metric difference

$$D_G \triangleq |M_e - M_b|.$$  \hspace{1cm} (3)

If we have $n$ corresponding realizations of the differences $(D_{pi}, D_{Gi})$ for $i = 1, \ldots, n$, then we can compute their correlation by the Pearson correlation coefficient [44, p. 26],

$$\rho(D_p, D_G) \triangleq \frac{\sum_{i=1}^{n} (D_{pi} - \overline{D_p})(D_{Gi} - \overline{D_G})}{\sqrt{\sum_{i=1}^{n} (D_{pi} - \overline{D_p})^2} \sqrt{\sum_{i=1}^{n} (D_{Gi} - \overline{D_G})^2}}.$$  \hspace{1cm} (4)

Only if $\rho(D_p, D_G)$ 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.
<table>
<thead>
<tr>
<th>$N$</th>
<th>Network size (the number of nodes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E[D]$</td>
<td>Average degree</td>
</tr>
<tr>
<td>$E[D^2]$</td>
<td>Second moment of degree</td>
</tr>
<tr>
<td>$d_{\text{max}}$</td>
<td>Largest degree</td>
</tr>
<tr>
<td>$E[H]$</td>
<td>Average shortest path length (the average hop-count)</td>
</tr>
<tr>
<td>$E[1/H]$</td>
<td>Global efficiency</td>
</tr>
<tr>
<td>$\lambda_1$</td>
<td>Spectral radius (the largest eigenvalue of the adjacency matrix)</td>
</tr>
<tr>
<td>$\mu_{N-1}$</td>
<td>Algebraic connectivity (the second smallest eigenvalue of the Laplacian matrix)</td>
</tr>
<tr>
<td>$\rho_D$</td>
<td>Assortativity</td>
</tr>
<tr>
<td>$C_G$</td>
<td>Average clustering coefficient</td>
</tr>
</tbody>
</table>

Table 1: Graph metrics

### 3.2 Evaluated network metrics

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

The assortativity $\rho_D$, which is the degree correlation between connected nodes [45], can be calculated as

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

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

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

where $\triangle_i$ 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 $\tau_c^{\text{HMF}}$ derived from the heterogeneous mean-field (HMF) approach [1] is

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

where $D$ is the degree of a randomly selected node and the epidemic threshold $\tau_c^{(1)}$ derived from NIMFA [46] is

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

Many graph metrics can also be bounded. For example, the average degree follows $E[D] \leq \lambda_1$ in connected graphs [47] and the largest eigenvalue of the Laplacian matrix $\mu_1 \geq \frac{N}{N-1}d_{\text{max}}$, while the algebraic connectivity is $\mu_{N-1} \leq d_{\text{min}}$.  

---

4
between the difference of the prevalence coefficient is moment of degree E coefficients ρ metric as D infected initially. For each network and initial state, a corresponding time series of the prevalence is y.

For any pair of networks, the prevalence difference D is relatively weak correlations between the difference of the prevalence D are calculated based on (2) and (3). For each network metric, we calculate the correlations via (4) between a set of network sizes N network size D metric differences based on (2) and (3). For each network metric, we calculate the correlations via (4) between a set of table.

<table>
<thead>
<tr>
<th>ρ(Dp(y0), DG(metric))</th>
<th>DG(E[D])</th>
<th>DG(E[D^2])</th>
<th>DG(λ1)</th>
<th>DG(E[H])</th>
<th>DG(E[1/H])</th>
</tr>
</thead>
<tbody>
<tr>
<td>ER graphs, Dp(y0 = 0.2)</td>
<td>0.941 0.856 0.940 0.953 0.939</td>
<td>0.939</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WS graphs, Dp(y0 = 0.2)</td>
<td>0.877 0.826 0.921 0.952 0.958</td>
<td>0.958</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BA graphs, Dp(y0 = 0.2)</td>
<td>0.940 0.838 0.871 0.952 0.945</td>
<td>0.945</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF graphs, Dp(y0 = 0.2)</td>
<td>0.944 0.612 0.561 0.861 0.823</td>
<td>0.823</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ER graphs, Dp(y0 = 1.0)</td>
<td>0.947 0.866 0.944 0.948 0.932</td>
<td>0.932</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WS graphs, Dp(y0 = 1.0)</td>
<td>0.905 0.818 0.927 0.952 0.954</td>
<td>0.954</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BA graphs, Dp(y0 = 1.0)</td>
<td>0.945 0.856 0.908 0.954 0.948</td>
<td>0.948</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF graphs, Dp(y0 = 1.0)</td>
<td>0.948 0.631 0.459 0.792 0.783</td>
<td>0.783</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Matrics with strong positive correlations

3.3 Correlation analysis

For any pair of networks, the prevalence difference DP and the metric difference DG can be calculated based on (2) and (3). For each network metric, we calculate the correlations via (4) between a set of metric differences DG and their corresponding prevalence differences DP on four network models: the SF graphs [34], the BA graphs [35], the ER random graphs [36] and the WS small-world graphs [37]. Specifically, the SF graphs are generated by the configuration model [34, 48] 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 τ = 3.0, which is above the epidemic threshold of every network realization. Two kinds of initial state are chosen: y0 = 0.2 or y0 = 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 DP under initial condition y0 as Dp(y0). We further denote the metric difference DG for one specific metric as DG(metric). All metrics shown in Section 3.2 are considered and the Pearson correlation coefficients ρ(Dp(y0), DG(metric)) are calculated by Eq. (4). The sample size of each correlation coefficient is (100 choose 2) = 4950. Table 2 and Tab. 3 indicate that there are generally strong correlations between the difference of the prevalence DP 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 λ1. 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. However, there are relatively weak correlations between the difference of the prevalence DP and the differences of the network size N, the largest degree dmax, the algebraic connectivity µN−1, the assortativity ρD and the average clustering coefficient CG. 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 DG are highly correlated to the difference of the prevalence DP, is possible. A straightforward method is randomly generating the network realizations by the corresponding network
Table 3: Matrics with weak positive correlations

<table>
<thead>
<tr>
<th>ρ(D_p(y_0), D_G(metric))</th>
<th>D_G(d_{max})</th>
<th>D_G(C_G)</th>
<th>D_G(\mu_{N-1})</th>
<th>D_G(\rho_D)</th>
<th>D_G(N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ER graphs, D_p(y_0 = 0.2)</td>
<td>0.821</td>
<td>0.477</td>
<td>0.490</td>
<td>-0.014</td>
<td>-0.059</td>
</tr>
<tr>
<td>WS graphs, D_p(y_0 = 0.2)</td>
<td>0.805</td>
<td>-0.036</td>
<td>-0.002</td>
<td>0.624</td>
<td>-0.012</td>
</tr>
<tr>
<td>BA graphs, D_p(y_0 = 0.2)</td>
<td>0.386</td>
<td>0.358</td>
<td>0.854</td>
<td>0.595</td>
<td>-0.031</td>
</tr>
<tr>
<td>SF graphs, D_p(y_0 = 0.2)</td>
<td>0.398</td>
<td>0.182</td>
<td>0.657</td>
<td>0.013</td>
<td>-0.038</td>
</tr>
<tr>
<td>ER graphs, D_p(y_0 = 1.0)</td>
<td>0.856</td>
<td>0.525</td>
<td>0.524</td>
<td>0.082</td>
<td>-0.018</td>
</tr>
<tr>
<td>WS graphs, D_p(y_0 = 1.0)</td>
<td>0.807</td>
<td>-0.031</td>
<td>0.081</td>
<td>0.666</td>
<td>-0.039</td>
</tr>
<tr>
<td>BA graphs, D_p(y_0 = 1.0)</td>
<td>0.284</td>
<td>0.410</td>
<td>0.813</td>
<td>0.535</td>
<td>-0.003</td>
</tr>
<tr>
<td>SF graphs, D_p(y_0 = 1.0)</td>
<td>0.247</td>
<td>0.100</td>
<td>0.659</td>
<td>0.006</td>
<td>0.034</td>
</tr>
</tbody>
</table>

3.4 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 \( D_p \) 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

\[
\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |M_{ei} - M_{bi}|
\]

and

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

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 supplementary material 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 \( D_p \), 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 $\lambda_1$, 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 $D_p$, i.e., the network size $N$, the largest degree
$d_{\text{max}}$, the algebraic connectivity $\mu_{N-1}$, the assortativity $\rho_D$ and the average clustering coefficient $C_G$,
their MAE and MSE of the treatment group are close to those of the control group.

4 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 $\tau$. 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.

4.1 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.

4.2 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 $\gamma$ ranges in the interval $\gamma \in [2.5, 3.0]$. For the SW graphs, the rewiring
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 \( V_{\text{tmp}} \), cooling rate \( 0 < r < 1 \) and step length \( S_N \)

Output: Estimated network \( G_e \), final prevalence difference \( D_p \)

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 \( N_c = \text{round}(S_N \times D_p) \) 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 \( D_2 \) to the benchmark.

9. if \( D_2 < D \) then

10. \( D \leftarrow D_2; G \leftarrow G_2; \)

11. else if \( \text{Exp}(-(D_2 - D)/V_{\text{tmp}}) > \text{random}(0, 1) \) then

12. \( D \leftarrow D_2; G \leftarrow G_2; \)

13. end

14. \( V_{\text{tmp}} = r \times V_{\text{tmp}}; \)

15. end
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 $D_{\text{RC}}$ and $D_{\text{PA}}$ 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 $D_{\text{RC}} - D_{\text{PA}}$. Figure 1 shows that these four kinds of networks can be almost exactly classified by the difference value $D_{\text{RC}} - D_{\text{PA}}$. Indeed, $D_{\text{RC}} > D_{\text{PA}}$ for almost all SF and BA graphs while $D_{\text{RC}} < D_{\text{PA}}$ 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)

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

against the False Positive Rate (FPR)

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

where $d$ is the threshold of the difference value $D_{\text{RC}} - D_{\text{PA}}$, $N_{\text{TP}}(d)$ is the number of true positives of $D_{\text{RC}} - D_{\text{PA}} > d$, $N_{\text{FP}}(d)$ is the number of false positives of $D_{\text{RC}} - D_{\text{PA}} > d$. The denominators $N_{\text{TP}}(d) + N_{\text{FN}}(d)$ and $N_{\text{FP}}(d) + N_{\text{TN}}(d)$ are the number of real positives and real negatives of $D_{\text{RC}} - D_{\text{PA}} > d$, respectively.

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

5 Estimating the topology of small networks and prevalence

5.1 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.

5.2 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.
Figure 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.
Figure 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.
Figure 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.

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.

6 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 $\lambda_1$. 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 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.

Abbreviations

SARA: Simulated Annealing Link-rewiring Algorithm; SIS: Susceptible-Infected-Susceptible; NIMFA: the N-Intertwined Mean-Field Approximation; HMF: Heterogeneous Mean-Field; SF: Scale-Free; BA: Barabási-Albert; ER: Erdős-Rényi; WS: Watts-Strogatz; RC: Randomly Connecting; PA: Preferential Attachment; ROC: Receivers Operating Characteristic; TPR: True Positive Rate; FPR: False Positive Rate; AUROC: the Area Under the ROC Curve; MAE: the mean absolute error; MSE: the mean squared error.

Declarations

Availability of data and material

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

LM and QL are supported by China scholarship council.

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.

Acknowledgements

We thank Marcus Märtens for providing simulation tools.
References


