Published in "Journal of Statistical Mechanics:
Theory and Experiment 2013(11): P11010, 2013"
which should be cited to refer to this work.

# Inferring network topology via the propagation process

**An Zeng**

Department of Physics, University of Fribourg, Chemin du Musée 3,
CH-1700 Fribourg, Switzerland
E-mail: an.zeng@unifr.ch

http://doc.rero.ch

**Abstract.** Inferring the network topology from the dynamics is a fundamental problem, with wide applications in geology, biology, and even counter-terrorism. Based on the propagation process, we present a simple method to uncover the network topology. A numerical simulation on artificial networks shows that our method enjoys a high accuracy in inferring the network topology. We find that the infection rate in the propagation process significantly influences the accuracy, and that each network corresponds to an optimal infection rate. Moreover, the method generally works better in large networks. These finding are confirmed in both real social and nonsocial networks. Finally, the method is extended to directed networks, and a similarity measure specific for directed networks is designed.

**Keywords:** network dynamics, random graphs, networks, network reconstruction

http://doc.rero.ch

## Contents



## 1. Introduction

Spreading processes widely exist in various fields, including physics, chemistry, medical science, biology, and sociology [1]. For example, reaction diffusion processes [2], pandemics [3], cascading failures in electric power grids [4], and information dissemination [5] can be naturally described by the framework of spreading. In the past decade, spreading on complex networks has been intensively studied. Studies have revealed that the spreading results are strongly influenced by the network topologies [6]–[9]. With this understanding, some network manipulating methods are designed to hinder spreading in the case of diseases or accelerate spreading in the case of information dissemination [10].

Recently, increasing attention has been paid to the microscopic level when studying the spreading process on networks [11]. Since the local structure around each node can be very different, the final spreading coverage varies from several nodes to the entire network when the propagation originates from distinct nodes. So far, many methods, such as the $k$-shell [12] and the leaderrank [13], have been proposed to rank the spreading ability of the nodes (i.e., how many nodes will finally be reached when the spreading originates from this single node).

A fundamental problem related to the spreading process is how to infer the network topology from the observation of the spreading results. If this question is answered, we could, for instance, have a better understanding of the organization of groups of terrorists (social networks) and the structure of some biology systems (metabolic networks). Since building the relation between the dynamics and the network structure is a crucial problem, much effort has been made in this direction [14]. In [15], the authors design a method to reconstruct the network based on the observation of some oscillation taking place on

it. Moreover, noise is found to lead to a general one-to-one correspondence between the dynamical correlation and the network connections [16]. Very recently, the oscillation has also been used to predict the missing nodes in a network [17]. Even though spreading processes exist widely in many real systems, so far little has been reported in the literature about inferring the network topology based on the spreading. The closest studies are [18, 19], where the spreading results are used to identify the initial spreader of a certain disease or information.

In this paper, we proposed a simple method to uncover the network topology. The basic idea is that the similarity between nodes can be estimated based on the spreading results. We test our method in two well-known artificial network models. The results shows that our method has a high accuracy in inferring the network topology. Moreover, the infection rate of the spreading is found to significantly influence the inferring accuracy, and each network has an optimal infection rate. We also validate our method in both real social and nonsocial networks. Finally, we design a new similarity measure and extend our method to directed networks. The new similarity measure is shown to remarkably improve the inferring accuracy compared to existing similarity measures.

http://doc.rero.ch

## 2. Model

We consider a network with $N$ nodes and $E$ links. The network is represented by an adjacency matrix $A$, where $a_{ij} = 1$ if there is a link between node $i$ and $j$, and $a_{ij} = 0$ otherwise. To simulate the spreading process on networks, we employ the SIR model [1]. Actually, this model has been used to simulate many different propagation process. Without losing any generality, we consider online information spreading as an example in this paper. We assume that each user has probability $f$ to submit some news. As such, there will be $f \times N$ items of news propagating in the network. After some news/story $\alpha$ is submitted (or received) by a user, it will infect each of this user's susceptible neighbors with probability $\beta$. After infecting neighbors, the user will immediately become recovered. All the users who received (or got infected by) $\alpha$ at the end will be recorded. For each user $i$, the set of news/stories that he/she received is denoted as $\Gamma(i)$.

## 3. Methods and metric

### 3.1. Methods

In the following, we will describe the method we used to infer the network topology based on the news propagation process. The basic idea is that the news/stories received by users can be used to estimate the similarity between them (nodes). We assume that the nodes with higher similarity are more likely to be connected in networks. Therefore, the obtained similarity $s_{ij}$ can be regarded as the likelihood score $L_{ij}$ for two nodes to have a link, i.e., $L_{ij} = s_{ij}$.

Actually, the similarity $s_{ij}$ is subject to different definitions. Here, we consider some well-known similarity definitions, as follows.

(i) *Common neighbors (CN).* By common sense, two nodes, $i$ and $j$, are more likely to have a link if they received many identical news/stories. The simplest measure of this

neighborhood overlap is the directed count, namely

$$s_{ij} = |\Gamma(i) \cap \Gamma(j)|. \tag{1}$$

(ii) *Salton index (SI).* The Salton index [20] is defined as

$$s_{ij} = \frac{|\Gamma(i) \cap \Gamma(j)|}{\sqrt{|\Gamma(i)| \times |\Gamma(j)|}}, \tag{2}$$

where $|\Gamma(i)|$ the number of items of news received by user $i$.

(iii) *Jaccard index (JI).* This index was proposed by Jaccard over 100 years ago [21], and is defined as

$$s_{ij} = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|}. \tag{3}$$

(iv) *Sorensen index (SSI).* This index is used mainly for ecological community data [22], and is defined as

$$s_{ij} = \frac{2 \times |\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i)| + |\Gamma(j)|}. \tag{4}$$

(v) *Hub promoted index (HPI).* This index is proposed for quantifying the topological overlap of pairs of substrates in metabolic networks [23], and is defined as

$$s_{ij} = \frac{|\Gamma(i) \cap \Gamma(j)|}{\min\{|\Gamma(i)|, |\Gamma(j)|\}}. \tag{5}$$

(vi) *Hub depressed index (HDI).* There is a measure with the opposite effect on hubs, which is

$$s_{ij} = \frac{|\Gamma(i) \cap \Gamma(j)|}{\max\{|\Gamma(i)|, |\Gamma(j)|\}}. \tag{6}$$

(vii) *Leicht–Holme–Newman index (LHN).* This index assigns high similarity to node pairs that have many common neighbors compared to the expected number of such neighbors [24]. It is defined as

$$s_{ij} = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i)| \times |\Gamma(j)|}. \tag{7}$$

(viii) *Resource allocation index (RA).* The similarity between $i$ and $j$ is defined as the amount of resource $j$ received from $i$ [25], which is

$$s_{ij} = \sum_{\alpha \in \Gamma(i) \cap \Gamma(j)} \frac{1}{m_\alpha}, \tag{8}$$

where $m_\alpha$ is the number of users who finally received news $\alpha$.

As a benchmark, we compare the similarity-based method with the well-known preferential attachment (PA) process. The mechanism of preferential attachment has been used to generate evolving scale-free networks, where the probability that a new link is connected to the node $i$ is proportional to $k(i)$ [27]. Based on this network growing mechanism, the likelihood score for two nodes to have a link can be calculated as $L_{ij} = |\Gamma(i)| \times |\Gamma(j)|$.

http://doc.rero.ch

### 3.2. Metric

To measure the accuracy of the method in inferring the network topology, we use the standard metric of the area under the receiver operating characteristic curve (AUC) [28]. In the network topology inference problem, there are four possible outcomes from the prediction. A true positive (TP) is the prediction of a link that exists in the real network, and if the link does not exist in the real network then it is called a false positive (FP). Conversely, a true negative (TN) means that a link that does not exist in the real network is not predicted, and a false negative (FN) is the lack of prediction of a link that actually exists in the real network.

To draw the receiver operating characteristic curve (ROC) curve, only the true positive rate (TPR) and the false positive rate (FPR) are needed. The TPR defines how many TPs occur among all TP and FN samples available during the test. On the other hand, FPR defines how many FPs occur among all FP and TN samples available during the test. The ROC curve is created by plotting the TPR versus the FPR at various threshold settings. When using normalized units, the area under the ROC curve (AUC) is equal to the probability that a true link has a higher score than a nonexisting link.

In this paper, we use a simple way to calculate the AUC. We pick a true link and a nonexisting link in the network and compare their scores. If, among $n$ pairs, the real link has a higher likelihood score $L_{ij}$ than the nonexisting link $n1$ times and equal score $n2$ times, the AUC value is as follows: $\mathrm{AUC} = (n1 + 0.5 * n2)/n$. Note that, if links were ranked at random, the AUC value would be equal to 0.5. By reanalyzing the following results with another accuracy measure, we verify that the performance of the methods is not strongly influenced by the accuracy measure we used. Therefore, we only present the results of AUC in section 4.

## 4. Results

### 4.1. Artificial networks

We first test our method in two artificial network models: (i) Watts–Strogatz (WS) networks [29], (ii) Barabasi-Albert (BA) networks [27]. When implementing our method, we select the Jaccard similarity definition as an example here. Figure 1 shows the AUC in the parameter space ($\beta$, $f$) for both WS and BA networks. Actually, both $\beta$ and $f$ control the amount of data we can obtain from the spreading process. If $\beta$ is too small, the news can only propagate several steps, and the data for similarity calculation will be limited. If $f$ is small, only a few items of news are propagating in the network and the obtained similarity matrix will be sparse as well. The first crucial observation in figure 1 is that the surface of the AUC has a pronounced maximum around $\beta = 0.15$ in WS networks and $\beta = 0.1$ in BA networks for all values of $f$. As discussed above, a small $\beta$ will result in a sparse similarity matrix and eventually lead to a poor AUC. In the case of large $\beta$ values, the spreading will cover almost all the network. Consequently, the information of the local network structure cannot embed in the spreading results. The optimal $\beta$ is somehow close to the critical infection rate for the spreading coverage [1]. Compared to $\beta$, the influences of $f$ on the AUC is smaller. Even though the AUC keeps increasing with $f$, the increasing speed becomes significantly slower once $f$ is larger than 0.3.

http://doc.rero.ch

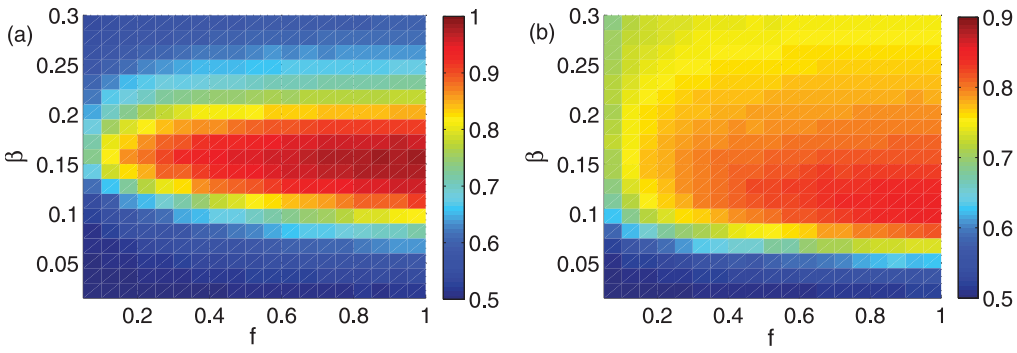


**Figure 1.** The AUC in the parameter space ($\beta$, $f$) for (a) WS networks ($N = 500$, $p = 0.1$, $\langle k \rangle = 10$) and (b) BA networks ($N = 500$, $\langle k \rangle = 10$). The results are averaged over 10 independent realizations.

Next, we move to investigate how the network structure properties influence the inferring accuracy. From figure 1, we can already see that the AUC in BA networks is lower than that in WS networks, which indicates that it is generally easier to infer a network with homogeneous degree distribution. Furthermore, we study the effect of average degree on the inferring accuracy in detail, with results reported in figure 2. Figures 2(a) and (b) show that, as the average degree $\langle k \rangle$ increases, the AUC shifts to the left in both networks. In figures 2(c) and (d), we can see that both the optimal $\beta^*$ and maximum AUC$^*$ decrease with increasing $\langle k \rangle$. Interestingly, $\beta^*$ is very stable under different $f$. In WS networks, $\beta^*$ stays almost unchanged when changing $f$. In BA networks, $\beta^*$ slightly decreases as $f$ increases.

We further apply our method on artificial networks with different size. We present the maximum AUC$^*$ (with respect to optimal $\beta^*$) against $N$ under different $\beta$ in figures 3(a) and (b). Interestingly, the inferring accuracy constantly increases with the network size. The curve with the optimal $\beta^*$ enjoys the largest slope ($\beta^* = 0.15$ in WS networks and $\beta^* = 0.1$ in BA networks). However, the slope slowly becomes smaller as $N$ increases. In figures 3(c) and (d), we report the maximum AUC$^*$ against $N$ under different $f$. The results show that $f$ can always improve AUC$^*$.

In reality, the infection rate might not be the same in different spreading processes. For example, some items of news are interesting and thus propagate wider than other news. Besides this, the spreading may only originate from a small region in the network. In the following, we investigate the non-uniform spreading parameters and localized initial condition in the SW and BA models.

In order to model the non-uniform spreading parameters, we modify the spreading process above. Specifically, the infection rate is no longer a constant. After a node is randomly selected as the initial spreader, an infection rate will be set as a random value in the range $[\beta-\epsilon, \beta+\epsilon]$. $\beta$ is the average infection rate and $\epsilon$ is the error magnitude. When $\epsilon = 0$, the spreading process reduces to the SIR model we considered before. Once $\epsilon > 0$, the infection rate will be different in each spreading process (i.e., each initial spreader selection is corresponding to a different infection rate setting).

We also model the localized initial condition. Instead of selecting the initial spreader from all the nodes in the network, we now consider only the nodes in one specific region as the initial spreader candidates. In practice, we randomly select a node as the seed

http://doc.rero.ch

http://doc.rero.ch

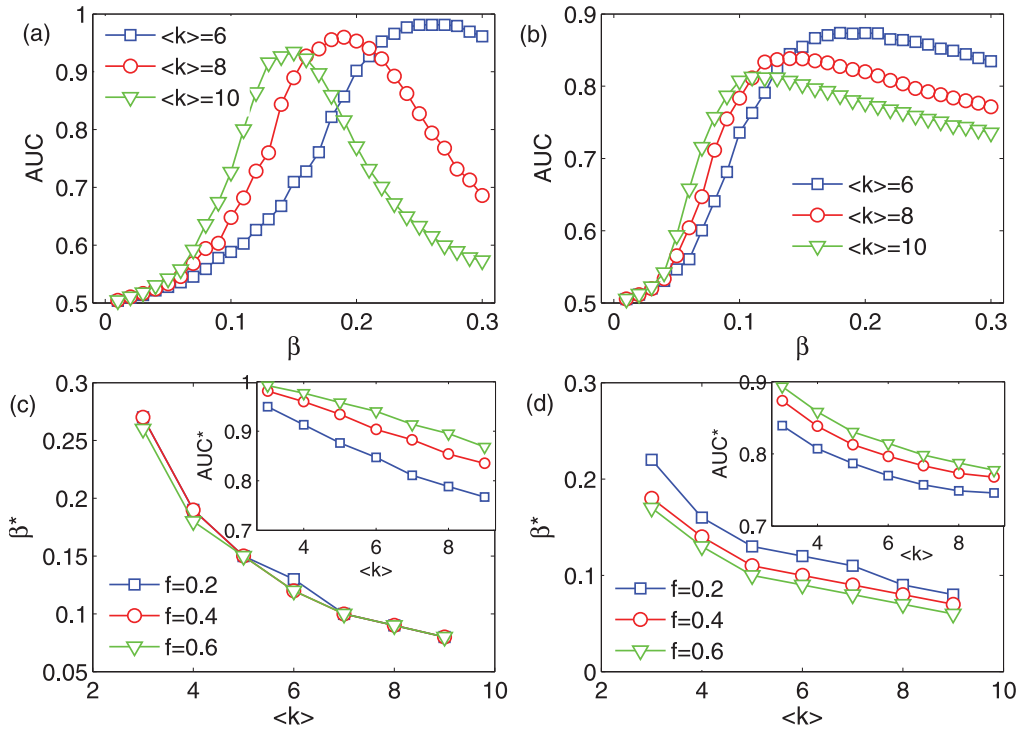


**Figure 2.** The dependence of the AUC on $\beta$ under different $\langle k \rangle$ in (a) WS networks ($N = 500$, $p = 0.1$) and (b) BA networks ($N = 500$), respectively. (c) and (d) show the relation between the optimal $\beta^*$ and $\langle k \rangle$ under different $f$ in WS networks and BA networks, respectively. The insets in (c) and (d) are the relation between the maximum AUC$^*$ and $\langle k \rangle$ under different $f$. The results are averaged over 10 independent realizations.

and calculate the shortest path length from the seed to all the other nodes. The $\eta * N$ nodes with the smallest shortest path length to the seed will form the region for the initial spreader candidates. Clearly, the region is as large as the whole network when $\eta = 1$. Once $\eta < 1$, the spreading can only originate from a part of the network.

Figures 4(a) and (b) show the effect of the non-uniform spreading parameters on the inference accuracy. We have already discussed that neither small nor large $\beta$ is good for inferring the network topology. This is because the similarity matrix is too sparse under small $\beta$ while the similarity between nodes cannot be accurately estimated under large $\beta$, since the viruses cover almost the whole network. The non-uniform spreading parameter setting can increase/decrease some infection rates in spreading. This makes both the small $\beta$ case and large $\beta$ case have some spreading processes with infection rate close to the optimal $\beta^*$, which leads to an improvement in the AUC under these values of $\beta$. However, the non-uniform spreading parameter setting may significantly lower the maximum AUC, and the optimal $\beta^*$ will be shifted to a smaller value.

Figures 4(c) and (d) show the effect of a localized initial condition on the inference accuracy. Actually, the localized initial condition mainly influences the results under small $\beta$. When $\beta$ is very large, the spreading covers almost the whole network and the spreading results will be independent of the original spreaders. In SW networks, the localized initial condition will lower the accuracy under small $\beta$. This is because a large part of the

http://doc.rero.ch

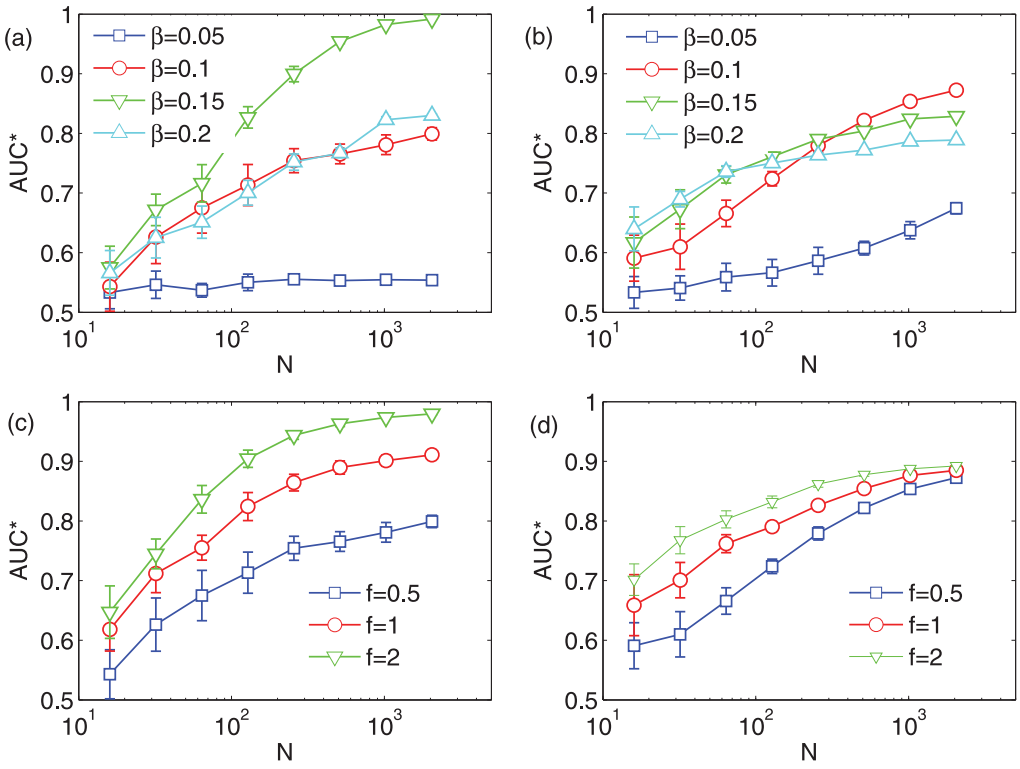


**Figure 3.** The maximum AUC* (with respect to optimal $\beta^*$) against $N$ under different $\beta$ in (a) WS networks ($\langle k\rangle = 5$, $p = 0.1$) and (b) BA networks ($\langle k\rangle = 5$), respectively. (c) and (d) show the maximum AUC* against $N$ under different $f$ in WS and BA networks, respectively. The results are averaged over 10 independent realizations.

network has no spreading record to calculate the similarity matrix. Interestingly, the localized initial condition seems to improve the accuracy under small $\beta$ in BA networks. BA networks have some hub nodes which connect to almost all the other nodes in the network and these hub nodes can effectively enhance the local spreading to a global level (so the similarity matrix will not be too sparse). In the local region where the initial spreaders are chosen, the inference accuracy becomes better, since more spreading information is available for calculating the similarity.

### 4.2. Real undirected networks

We will validate our method in real undirected networks, and all the similarity definitions discussed above will be compared. Both social and nonsocial networks are selected.

The social networks are as follows: Dolphins (friendship network with 62 nodes and 159 links) [30], Jazz (musical collaboration network with 198 nodes and 2742 links) [31], Netsci (collaboration network of network scientists with 379 nodes and 914 links) [32], and Email (email communication network with 1133 nodes and 5451 links) [33].

The nonsocial networks are as follows: Word (adjacency network in English text with 112 nodes and 425 links) [32], *E. coli* (metabolic network of *E. coli* with 230 nodes and 695 links) [34], USAir (Airline network of USA with 332 nodes and 2126 links) [35], TAP (yeast

http://doc.rero.ch

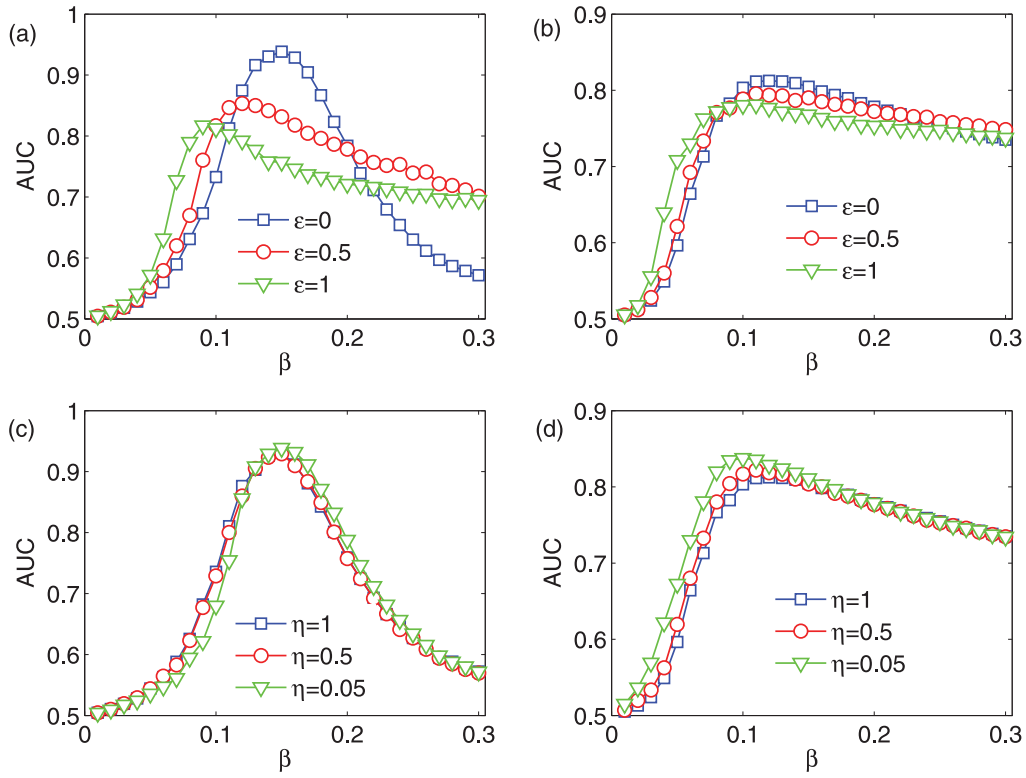


**Figure 4.** AUC versus $\beta$ under the non-uniform infection rate setting in (a) SW and (b) BA networks. AUC versus $\beta$ under the localized initial condition in (c) SW and (b) BA networks. In this figure, $f = 0.4$. The parameters are $N = 500$, $p = 0.1$, $\langle k \rangle = 10$ for the WS networks and $N = 500$, $\langle k \rangle = 10$ for the BA networks. The results are averaged over 10 independent realizations.

protein–protein binding network generated by tandem affinity purification experiments, with 1373 nodes and 6833 links) [36], and PPI (a protein–protein interaction networks with 2375 nodes and 11 693 links) [37].

The results in table 1 show that the similarity-based network inferring method can achieve significant higher accuracy than the preferential attachment method. Among the similarity measures we considered, the SI, JI, and RA generally perform best, and they are very robust in their performance. We also examine the performance of different similarity metrics in these networks with the non-uniform spreading parameters and localized initial condition. The results show that the SI, JI, and RA metrics still generally perform best, and the AUC is not significantly influenced.

### 4.3. Real directed networks

Actually, our method can be easily extended to directed networks. However, all the similarity measures discussed above are symmetric (i.e., $s_{ij} = s_{ji}$). This implies that, if a direction link exists, the link in the other direction will exist as well. This will largely lower the accuracy. To solve the problem, we propose an asymmetric similarity (AS) measure for inferring the network topology in directed networks. Mathematically, it can

**Table 1.** AUC of different similarity definitions in real undirected networks. The parameters are set as $\beta = 1/\langle k \rangle$ and $f = 0.5$. The similarity with best performance in each network is highlighted in bold font.

| | CN | SI | JI | SSI | HPI | HDI | LHN | RA | PA |
|---|---|---|---|---|---|---|---|---|---|
| Dolphins | 0.8098 | 0.8088 | **0.8351** | 0.8164 | 0.7836 | 0.8110 | 0.7989 | 0.8200 | 0.6678 |
| Word | 0.8082 | 0.8109 | 0.8041 | 0.8044 | 0.7774 | 0.7921 | 0.6747 | **0.8192** | 0.7674 |
| Jazz | 0.7918 | 0.7933 | 0.7891 | 0.8007 | 0.7370 | 0.7925 | 0.6876 | **0.8041** | 0.7552 |
| *E. coli* | 0.8712 | **0.9022** | 0.8944 | 0.8943 | 0.8345 | 0.8918 | 0.7689 | 0.8900 | 0.8302 |
| USAir | 0.9086 | **0.9145** | 0.9074 | 0.9066 | 0.8510 | 0.8999 | 0.6524 | 0.9132 | 0.8984 |
| Netsci | 0.8998 | **0.9186** | 0.9183 | 0.9167 | 0.9086 | 0.9148 | 0.9071 | 0.9138 | 0.6672 |
| Email | 0.8439 | **0.8758** | 0.8676 | 0.8670 | 0.8157 | 0.8554 | 0.7276 | 0.8558 | 0.8131 |
| TAP | 0.8691 | 0.9033 | 0.9065 | **0.9082** | 0.8854 | 0.9034 | 0.8903 | 0.8942 | 0.7223 |
| PPI | 0.8937 | 0.9345 | **0.9349** | 0.9342 | 0.8613 | 0.9324 | 0.8117 | 0.9124 | 0.8404 |

**Table 2.** AUC of different similarity definitions in real directed networks. The parameters are set as $\beta = 2/\langle k_{\text{out}} \rangle$ and $f = 0.5$. The similarity with best performance in each network is highlighted in bold font.

| | CN | SI | JI | SSI | HPI | HDI | LHN | RA | AS | PA |
|---|---|---|---|---|---|---|---|---|---|---|
| Prisoners | 0.7339 | 0.8159 | 0.8133 | 0.8164 | 0.7951 | 0.7987 | 0.7559 | 0.7483 | **0.8350** | 0.6469 |
| SM FW | 0.6643 | 0.6834 | 0.6634 | 0.6543 | 0.6839 | 0.6484 | 0.6127 | 0.6774 | **0.7635** | 0.6111 |
| LR FW | 0.7046 | 0.7135 | 0.7097 | 0.7012 | 0.7052 | 0.7019 | 0.7038 | 0.7102 | **0.7308** | 0.6855 |
| Neural | 0.7083 | 0.7076 | 0.7051 | 0.7052 | 0.7049 | 0.6995 | 0.6476 | 0.7209 | **0.7658** | 0.6864 |
| Metabolic | 0.7043 | 0.7373 | 0.7239 | 0.7246 | 0.7596 | 0.7175 | 0.7027 | 0.7178 | **0.8031** | 0.6542 |
| PB | 0.8757 | 0.8784 | 0.8761 | 0.8777 | 0.8606 | 0.8722 | 0.7676 | 0.8767 | **0.8926** | 0.8677 |

be expressed as

$$s_{ij} = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i)|}. \tag{9}$$

A large $s_{ij}$ indicates that $j$ received most of the news/stories passing through $i$. Therefore, the network is more likely to have a directed link from $i$ to $j$.

We considered several real directed networks to validate our method. The networks are the following: Prisoners (friendship network between prisoners with 67 nodes and 182 links) [38], SM FW (food web network in St. Mark area with 54 nodes and 356 links) [39], LR FW (food web network in Little Rock area with 183 nodes and 2494 links) [39], Neural (the neural network of *C. elegans* with 297 nodes and 2359 links) [40], Metabolic (the metabolic network of *C. elegans* with 453 nodes and 2040 links) [39], and PB (the hyperlink between the blogs of politicians with 1222 nodes and 19090 links) [41]. Again, we observe that the inferring accuracy of the similarity-based method is higher than that of the preferential attachment method. Interestingly, the AS measure performs best among all the similarity measures. The results indicate that the asymmetric feature is crucial for inferring network topology in directed networks.

Like in undirected networks, we examine the performance of different similarity metrics in a directed network with non-uniform spreading parameters and localized initial

http://doc.rero.ch

condition. We observe that the accuracy is largely lowered. Generally speaking, it is more difficult for the virus/information to propagate in these networks due to the directionality of the links. Therefore, the virus is very likely to stay in a local region under the localized initial condition, which results in a very sparse similarity matrix for inferring the network topology and thus a much lower AUC. The phenomenon is even more serious in some acyclic networks (such as the SM food web and LR food web).

## 5. Conclusion

To summarize, we have proposed a method to infer the network topology based on the spreading process on networks. Specifically, the similarity between nodes is estimated by the information/virus that nodes received, and the nodes with the highest similarity are assumed to be connected. We tested our method in classic artificial network models and found that our method enjoys high inferring accuracy. Moreover, we found that the infection rate in the spreading process significantly affects the inferring results and that there is an optimal infection rate for each network. The findings were confirmed in many real networks. Finally, the method was extended to directed networks. We have proposed a new similarity measure, which is shown to perform better than other well-known similarity measures in directed networks.

We remark that many extensions can be made in this direction. For example, the inferring accuracy can be further improved if the time information of the spreading is known (i.e., at what time the nodes receive the virus). In addition, it is interesting and important to design an more efficient method for cases where only partial information of the spreading can be obtained.

http://doc.rero.ch

## Acknowledgment

The author would like to thank an anonymous referee for comments that improved the paper. The author acknowledges support from the China Scholarship Council.

## References

[1] Dorogovtsev S N, Goltsev A V and Mendes J F F, 2008 *Rev. Mod. Phys.* **80** 1275
[2] Colizza V, Pastor-Satorras R and Vespignani A, 2007 *Nature Phys.* **3** 276
[3] Pastor-Satorras R and Vespignani A, 2001 *Phys. Rev. Lett.* **86** 3200
[4] Motter A E, 2000 *Phys. Rev. Lett.* **93** 098701
[5] Medo M, Zhang Y-C and Zhou T, 2009 *Europhys. Lett.* **88** 38005
[6] Eguiluz V M and Klemm K, 2002 *Phys. Rev. Lett.* **89** 108701
[7] Boguna M, Pastor-Satorras R and Vespignani A, 2003 *Phys. Rev. Lett.* **90** 028701
[8] Serrano M A and Boguna M, 2006 *Phys. Rev. Lett.* **97** 088701
[9] Castellano C and Pastor-Satorras R, 2010 *Phys. Rev. Lett.* **105** 218701
[10] Bishop A N and Shames I, 2011 *Europhys. Lett.* **95** 18005
[11] Zeng A and Zhang C-J, 2013 *Phys. Lett.* A **377** 1031
[12] Kitsak M *et al,* 2010 *Nature Phys.* **6** 888
[13] Lu L, Zhang Y-C, Yeung C H and Zhou T, 2011 *PLoS One* **6** e21202
[14] Boccaletti S, Latora V, Moreno Y, Chavez M and Hwang D-U, 2006 *Phys. Rep.* **424** 175
[15] Shandilya S G and Timme M, 2011 *New J. Phys.* **13** 013004
[16] Ren J, Wang W-X, Li B and Lai Y-C, 2010 *Phys. Rev. Lett.* **104** 058701
[17] Su R-Q, Wang W-X and Lai Y-C, 2012 *Phys. Rev.* E **85** 065201(R)
[18] Comin C H and da Fontoura Costa L, 2011 *Phys. Rev.* E **84** 056105

[19] Pinto P C, Thiran P and Vetterli M, 2012 *Phys. Rev. Lett.* **109** 068702
[20] Salton G and McGill M J, 1983 *Introduction to Modern Information Retrieval* (Auckland: McGraw-Hill)
[21] Jaccard P, 1901 *Bull. Soc. Vaudoise des Sci. Naturelles* **37** 547
[22] Sorensen T, 1948 *Biol. Skr.* **5** 1
[23] Ravasz E, Somera A L, Mongru D A, Oltvai Z N and Barabasi A-L, 2002 *Science* **297** 1553
[24] Leicht E A, Holme P and Newman M E J, 2006 *Phys. Rev.* E **73** 026120
[25] Zhou T, Lu L and Zhang Y-C, 2009 *Eur. Phys. J.* B **71** 623
[26] Barabasi A-L and Albert R, 1999 *Science* **286** 509
[27] Barabasi A-L and Albert R, 1999 *Science* **286** 509
[28] Hanely J A and McNeil B J, 1982 *Radiology* **143** 29
[29] Watts D J and Strogatz S H, 1998 *Nature* **393** 440
[30] Lusseau D *et al,* 2003 *Behav. Ecol. Sociobiol.* **54** 396
[31] Gleiser P M and Danon L, 2003 *Adv. Complex Syst.* **6** 565
[32] Newman M E J, 2006 *Phys. Rev.* E **74** 036104
[33] Guimera R, Danon L, Diaz-Guilera A, Giralt F and Arenas A, 2003 *Phys. Rev.* E **68** 065103
[34] Jeong H, Tombor B, Albert R, Oltvai Z N and Barabasi A, 2000 *Nature* **407** 651
[35] http://vlado.fmf.uni-lj.si/pub/networks/data/default.htm
[36] Gavin A C *et al,* 2002 *Nature* **415** 141
[37] von Mering C, Krause R, Snel B, Cornell M, Oliver S G, Fields S and Bork P, 2002 *Nature* **417** 399
[38] www.casos.cs.cmu.edu/index.php
[39] www.cosinproject.org/
[40] Duch J and Arenas A, 2005 *Phys. Rev.* E **72** 027104
[41] http://incsub.org/blogtalk/images/robertackland.pdf

http://doc.rero.ch