Effects of Network Aggregation in Simple Diffusion Simulations *

Thomas Magelinski[1][0000-0002-6369-0753] and
Kathleen M. Carley[1][0000-0002-6356-0238]

CASOS, Institute for Software Research, Carnegie Mellon University, Pittsburgh, PA 15217, USA http://www.casos.cs.cmu.edu

Abstract. Diffusion models are at the heart of many pressing issues such as disease, fake news, and computer viruses. Epidemiologists and others have recognized the importance of link structure in static networks for diffusion simulations. At the same time, network scientists have recognized the importance of temporal structure in complex systems. In this work, we bring the two together to study the effect of condensing a temporal network into a static network in the context of diffusion. Using the College Message Dataset and a simple infected-susceptible simulation model, we study the differences in diffusion given three networks: the original temporal network, the static binarized network, and the static thresholded network. We find that the static networks exhibit significantly different diffusion properties to the original temporal network. In general, diffusion occurred more quickly and reached more of the nodes given the binarized network as opposed to the temporal network. The opposite was true for the tresholded network. Additionally, the choice of seed node lead to drastically different results in the temporal network, while there was no difference in the static networks. Lastly, we found that low transmissibility simulations on the temporal network further increased the disparity between seed nodes, showing promise as a method of finding temporally-central noes. These findings are in line with the theory of network cascades.

Keywords: Temporal Networks · Network Diffusion · Network Aggregation

1 Introduction

Diffusion is a central question in the study of complex systems due to its many applications. Diffusion models have been used to answer questions about the

* This material is based upon work supported by the Office of Naval Research Multidisciplinary University Research Initiative (MURI) under award number N00014-17-1-2675. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Office of Naval Research. Additionally, Thomas Magelinski was supported by an ARCS foundation scholarship.

spread of disease in communities, fake news on social media, and viruses on computer networks. Due to the complexity of the environment for all of these cases, network-based simulation is the prevailing modeling technique.

Given a set of entities (people, computers, etc.), a network defines all the connections between entities as links. These connections are then the basis of diffusion. For example, a "contact network" can be constructed where nodes are people in a community, and links connect people who come in contact with each other. Since diseases spread through contact, the contact network describes the environment in which the disease diffuses. Due to the complexity of these networks, it is often intractable to answer basic questions about diffusion analytically. Instead, models of diffusion are defined as a set of rules given a network, and simulations are run to understand the diffusion process.

Traditionally, networks are considered static, i.e. the nodes and links do not change with time. However, relationships in complex systems may be able to change. Further, network data is often collected over a time period, introducing a temporal component to the data itself. Still, this is often ignored in practice. Temporal networks are often aggregated and then converted into static networks in one of two ways. Aggregation is done by summing the instances that a link starts between nodes, or summing the length of time nodes are connected for, giving a static weighted network. Then, this network can be binarized, which sets all link weights above zero to one. Alternatively, a threshold can be defined which link weights must be above to be retained. Often, the mean link weight is used as a threshold. Thresholding sparsifies the network, allowing practitioners to see only strong links.

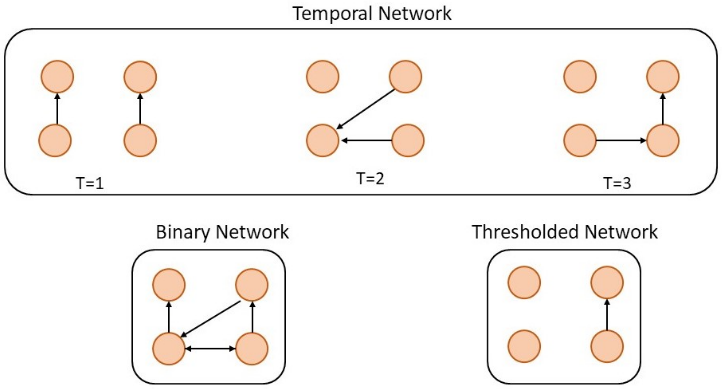


Fig. 1: Illustration of a 4-node 3-time-step temporal network (top) along with the binarized version (left) and thresholded version (right). In this case, the threshold for a link to not be dropped is 2.

Static networks are used because there is a vast tool set for analyzing them. However, removing temporal structure in a network can drastically change the dynamics that occur *on* networks, such as diffusion. To illustrate the effect of network aggregation, a simple temporal network is shown in Figure 1. In the binary network, it seems that something can be transferred from the bottom right node, to the bottom left, and finally to the top left. However, looking at the ground-truth temporal network, we see that that this path is impossible. By the time diffusion has occurred from bottom right to bottom left, the link from bottom left to top left has already disappeared. In the thresholded network, information about when diffusion from bottom right to top right can occur is lost, as are the other connections.

Given that a very basic network can have drastically different diffusion properties after aggregation, we seek to answer the question: **How does network aggregation effect diffusion?**

To answer this question, we use a very simplistic diffusion model and a real network dataset, the College Message Dataset, which contains private messages between college students over 6 months. Many complex diffusion simulations have been created, and validated. However, a simplistic simulation allows us to control for network aggregation, and better interpret our results. In the following sections we discuss prior work in diffusion and temporal networks, give details about the dataset, introduce our experimental design, present results, and discuss limitations.

2 Prior Work

Much of the work in network-based diffusion simulations has been done in epidemiology in order to understand the spread of disease. One of the most famous, though basic, examples is the SIR model [4]. In this model, agents are either susceptible, infected, or recovered. Susceptible agents will become infected based on a transmission probability parameter p and the contacts they have with infected agents. This model has been adapted as a system dynamics model as well as an agent-based model requiring a network.

More recently, work has been done to understand the role that the underlying network plays in diffusion simulation. May and Alun ran simulations on scale-free networks, which are thought to be reasonable artificial network models for large-scale systems [5]. Moreno et al have looked further to understand the effects of heterogeneous networks in diffusion, and found that scale-free networks lack an epidemic threshold found in other network models [7]. Meyers has recognized that many diffusion processes are asymmetric, and as such should be modeled with a directed network [6].

At the same time, the importance of temporal structure in networks has gained recognition [3]. Work by Kempe, Kleinberg, and Kumar shows that traditional path theorems may fail for *time-respecting paths*, which are essential for diffusion (again, see example in Introduction regarding Figure 1). Whitbeck et al. followed up on this work exploring temporal reachability and its importance for communication on dynamic networks [12].

Additionally, the benefits of thresholding networks in diffusion has been studied by Valente, who found that thresholding allows researchers to study diffusion properties at different granularities, suggesting that multiple thresholds should be used simultaneously [10]. Further, link aggregation has been shown to enhance/obscure community properties depending on their granularity [9].

So, while network structure has proved to be an important aspect of network diffusion, and temporal structure is an important aspect of networks in general, little work has been done to show how temporal structure effects network diffusion directly. Further, comparison between network transformations has been done, but only for static networks. Volz and Meyers have studied SIR models dynamic contact networks, but did not contrast with static networks to understand the importance of the temporal structure [11]. Here, we seek to bridge that gap by constructing three networks from the College Message Dataset: the temporal network, the binarized network, and the mean-thresholded network. We then run susceptible-infected, or SI, simulations on each network to study the effects of network construction directly.

It is important to note that much more complex diffusion models have been developed and validated [1, 13]. These works show significant departures from results from the SIR model, but come with more computational and conceptual overhead. Here, we seek to limit the complexity of diffusion as much as possible to isolate the differences in experiments to the network construction.

3 The Dataset

In this study we analyze the College Messeging Dataset [8]. The dataset was constructed by collecting private messages sent on an online social media platform at the University of California, Irvine, The dataset consists of source nodes (message senders), target nodes (message receivers), and time-stamps. Basic statistics can be seen in Table 1.

Feature	**Value**
Nodes	1899
Messages	59,835
Time-span	193 Days

Table 1: Basic statistics of the College Message Dataset

4 Experimental Setup

4.1 Simulation Conditions

Since the dataset being analyzed is a communication network, we define a simple model for *idea* diffusion. This model is a susceptible-infected model. That is, at each time-step, infected users infect their susceptible neighbors with probability, p_t, also referred to as the *transmission probability*. Initially, "seed node(s)" must be decided, which are users who start with the idea in the beginning of the simulation.

Since knowledge is spreading, rather than something like disease, there is no concept of "recovered." Once you have heard the idea, you never forget it. Also, the interpretation of p_t is slightly different. In a traditional diffusion model, the contact is definite and p_t describes the probability that the contact results in infection. Here, the p_t describes the probability that the contact contains the information. For example, if the idea spreading is a news article, p_t is the probability that a person with knowledge of the news is actually talking about the news in a particular message. This is a natural interpretation, as even the most prevalent ideas do not consume 100% of our communication.

The output of this simulation we are interested in is the fraction of nodes infected. For each simulation a time-series of this quantity is outputted.

4.2 Network Construction

Three types of networks are considered: temporal, binarized, and thresholded. The temporal network is analyzed as an ordered list of links. If the source node (message sender) has the idea, it spreads it to the target node (message receiver) with probability p_t. The links are iterated through in temporal order, following this procedure. Thus, simulation using the temporal network occurs over the real time line. This is a major distinction from that of static networks, as will now be discussed.

To create a static network, the instances of messages between users are summed to create a weighted, directed network between users. Then, the binary network is created by removing the link weights (setting them to one). The thresholded network first removes links below the mean link weight, then binarizes the remaining links. The procedure to simulate on the two networks is the same and is as follows. Nodes in the network spread the information to their direct neighbors (obeying link direction), with probability p_t. This process is then repeated until convergence. Without repetitions, seed nodes can only spread to their direct neighbors, defeating the purpose of analyzing diffusion with simulations (you could simply analyze neighbors instead.) As mentioned earlier, these "repetitions" no longer have any mapping to the real timeline.

Because of this, we cannot directly quantitatively compare temporal and static networks as the simulation progresses. Instead, we qualitatively assess their progression through simulation, and quantitatively assess their differences after convergence. This assessment is discussed in detail in Section 5.

4.3 The Virtual Experiment

Four choices must be made before a simulation can take place: which network to use, what the transmissibility rate is, which nodes are seed nodes, and when the seed nodes get the information. We have already discussed that simulations will take place on the three networks described above. Three transmissibility rates are tested: 1, 0.5, and 0.1. Note that transmissibility of 1 shows the connectivity of the network, removing all stochasticity.

For simplicity, only one seed node is selected per simulation and the seed occurs at the beginning of the time line. We ran experiments selecting top-6 nodes in terms of out degree (total messages sent) as seed nodes, as these are the most likely to reach a large fraction of users.

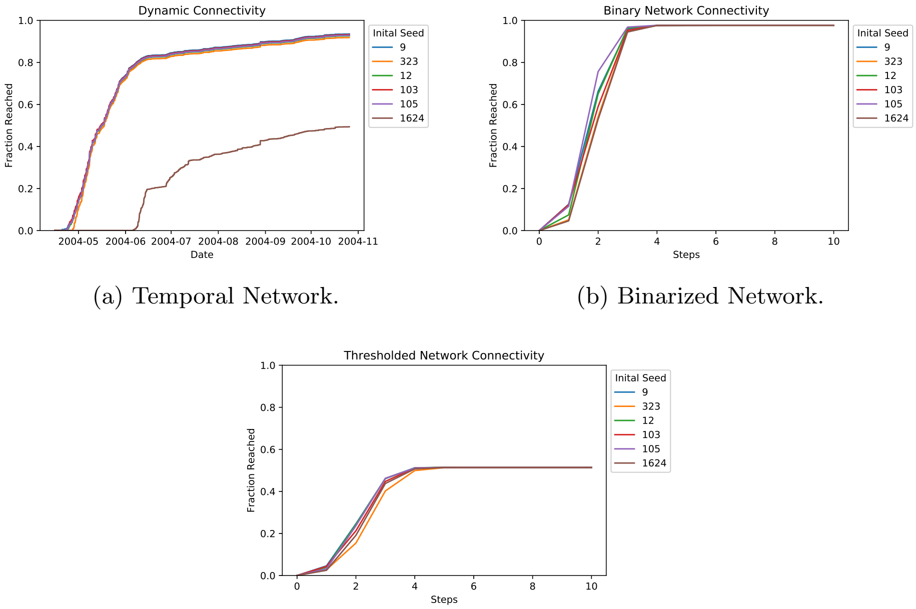
Variable	**Cases**	**Implications**
Network Type	Temporal, Binary, Threshold	Diffusion environment
Infection Rate	1, 0.5, 0.1	Likelihood of spreading the idea
Initial Seed	9, 323, 12, 103, 105, 1624	Who starts with the idea
Initial Seed Time	0	When the idea originates
Number of Seeds	1	How many people start with the idea

Table 2: Summary of the virtual experiment setup. Each condition was replicated 25 times, except for the deterministic case where transmissibility is 1, leading to 918 total trails.

Each of the stochastic experiments were repeated 25 times. More repetitions could be performed, but simulations showed convergence after 25 trials. This gives a total of $3*2*6*25+3*1*6*1 = 918$ simulation runs. The experimental setup is summarized in Table 2. Again, the output of this simulation we are interested in is the fraction of nodes infected. For each simulation a time-series of this quantity is outputted. The resulting time-series are qualitatively compared to find differences in diffusion through time. Then, ANOVA is preformed at the final value of “fraction reached,” to find quantitative differences between models.

5 Results

First, the connectivity of the 6 tested seed nodes was tested by simulating with $p_t = 1$. The results for this experiment are shown in Figure 2. The first major finding is that the temporal node set has high connectivity for 5 of the six nodes, while the sixth node only has potential to reach about half of the nodeset. This disparity is mostly due to the fact that the sixth user started sending messages months after the others. The other individuals have extremely similar plots, indicating that they messaged each other early on. The differences in seeds are not seen in the static networks, since they are unable to capture sixth user’s delay.



(a) Temporal Network.

(b) Binarized Network.

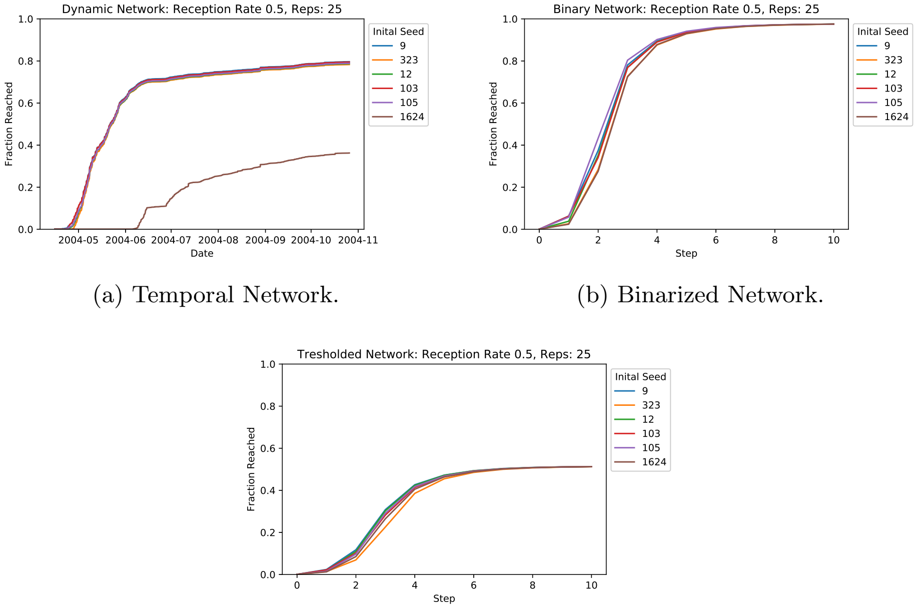
(c) Thresholded Network.

Fig. 2: Connectivity plots of each network type. These are achieved by running the simulation with $p_t = 1$. Trials were not repeated as they are deterministic.

The second major result from the connectivity study is that the binarized network is the easiest to travel on, followed by the dynamic network (for high-degree nodes), and then the thresholded network. Even before stochastic simulation, these plots show the harm that static network construction can do to a temporal network in terms of diffusion. Binarizing a network opens all the pathways at once, overstating the ease of travel, while thresholding cuts pathways, understating the ease of travel. Both can remove disparities between initial seeds.

Next, diffusion experiments were run with $p_t = 0.5$ and $p_t = 0.1$. The mean value over repetitions for the experiments are shown in Figures 3 and 4, respectively. It seems that decreasing p_t, increases the extent to which the network construction effects diffusion. Again, the binary network is the easiest to diffuse on, and thresholded is the hardest. However, the differences in final fraction reached are larger between networks as p_t decreases. Also, the effect of seed node is more pronounced at lower values of p_t, as shown by the larger spread of diffusion patterns in Figure 4a.

To access the results quantitatively, a two-way ANOVA was performed on each random seed. For every seed, both the type of network and the transmissibility was significant at the $p < 0.05$ level. Interaction between the two variables was not found to be significant. These results are in line with our qualitative



(a) Temporal Network.

(b) Binarized Network.

(c) Thresholded Network.

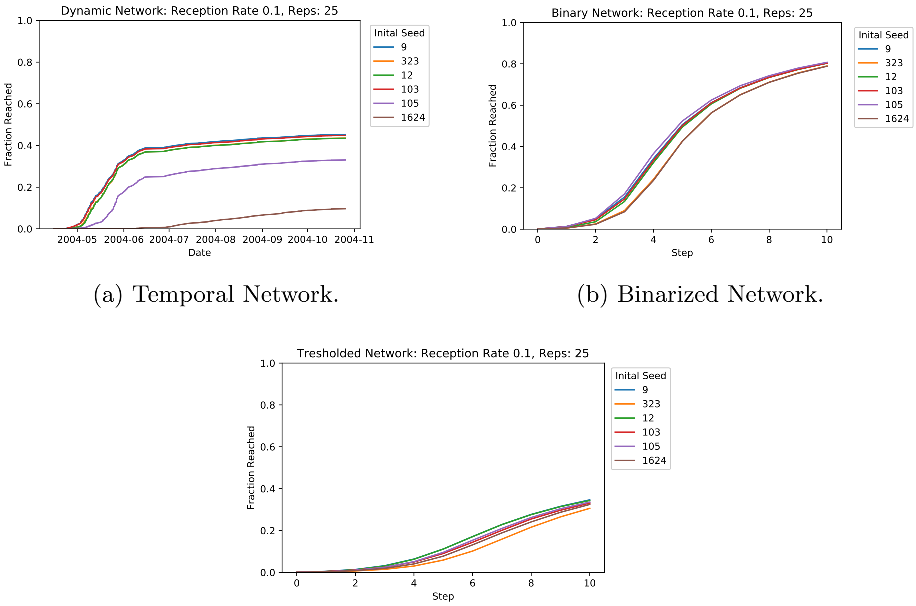
Fig. 3: Diffusion plots of each network type with $p_t = 0.5$. Each experiment was repeated 25 times. The average over the repetitions are shown.

analysis: network construction has an effect on diffusion, as does transmissibility. The fact that significance was found for each seed further bolsters our conclusions.

In a temporal network, a user with one well targeted link early on can reach a large fraction of the nodeset. However, if the transmissibility is low, there is a high chance that nodes miss their opportunity and do not reach the larger nodeset. Nodes consistently reaching much of the nodeset despite low p_t, then, are the truly influential nodes. This result is closely related to cascading phenomena [2].

6 Concerns and Future Work

Validation for this work is considered to be "face validity." The aim of this study was to test the effect of two different static network construction procedures on diffusion in a real dataset. It is true that the diffusion simulation was simplistic, and does not completely capture real-word diffusion phenomena. This is deliberate, so that the differences in results could be easily interpreted and attributed to the network construction methods, rather than an artifact of a complex diffusion model. Now that the network effects on simple diffusion have been studied, more complex diffusion models is a reasonable next step in future research.



(a) Temporal Network.

(b) Binarized Network.

(c) Thresholded Network.

Fig. 4: Diffusion plots of each network type with $p_t = 0.1$. Each experiment was repeated 25 times. The average over the repetitions are shown.

True validation could be challenging, as it would require two things: empirical data of something spreading on a network, assurance that all diffusion is done on the network itself. Future work could include collecting such a dataset on social media, where the conversation around news and major events can be tracked, though it would be extremely challenging. The difficulty lies in the second requirement, assurance that diffusion occurs only inside the network. Even if the conversation was restricted to social media, there are many popular but disconnected platforms that users frequent.

Lastly, this work points to one major question. We see that aggregating temporal data into a static network loses key information. However, often times temporal aspects of datasets are lost or not collected. Then, what can be done? First steps from our work to answer this question could be done by looking at semi-static networks. Here, we considered aggregating the whole timeline or none of it. But, what if the data was aggregated by month? This seems to be a realistic data scenario. When nothing is aggregated, the real time line is simulated on, when everything is aggregated, abstract "time-step" space is simulated on. If there is only some aggregation, do you need repetitions, and if so, how many? It is likely that the answer to this question relies on the temporal structure of the dataset. Using this dataset and others to guide researchers on simulate practices in the absence of temporal granularity to minimize the effects we've seen in our study would prove extremely useful.

7 Conclusion

Diffusion models are at the heart of many pressing issues such as disease, fake news, and computer viruses. Epidemiologists and others have recognized the importance of link structure in static networks for diffusion simulations. At the same time, network scientists have recognized the importance of temporal structure in complex systems. In this work, we bring the two together to study the effect of condensing a temporal network into a static network in the context of diffusion. Using the College Message Dataset and a simple infected-susceptible simulation model, we see that static networks exhibit significantly different diffusion properties to the original temporal network. In general, diffusion occurred more quickly and reached more of the nodes with the binarized network as opposed to the temporal network. This result was even more apparent as p_t was decreased.

Additionally, the choice of seed node lead to drastically different results in the temporal network, while there was no difference in the static networks. Lastly, we found that low transmissibility simulations on the temporal network further increased the disparity between seed nodes, showing promise as a method of finding temporally-central noes. These findings are in line with the theory of network cascades.

This work highlights the importance of retaining temporal data in networks. We hope that this work will be used as a stepping stone for future work guiding practitioners for best practices in simulation in the absence of sufficient temporal granularity.

References

1. Carley, K.: A theory of group stability. American sociological review pp. 331–354 (1991)
2. Crucitti, P., Latora, V., Marchiori, M.: Model for cascading failures in complex networks. Physical Review E **69**(4), 045104 (2004)
3. Holme, P., Saramäki, J.: Temporal networks. Physics reports **519**(3), 97–125 (2012)
4. Kermack, W.O., McKendrick, A.G.: A contribution to the mathematical theory of epidemics. Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character **115**(772), 700–721 (1927)
5. May, R.M., Lloyd, A.L.: Infection dynamics on scale-free networks. Physical Review E **64**(6), 066112 (2001)
6. Meyers, L.A., Newman, M.E.J., Pourbohloul, B.: Predicting epidemics on directed contact networks. Journal of Theoretical Biology **240**(3), 400–418 (Jun 2006). https://doi.org/10.1016/j.jtbi.2005.10.004, http://www.sciencedirect.com/science/article/pii/S0022519305004418, using contact networks to predict epidemics / spread
7. Moreno, Y., Pastor-Satorras, R., Vespignani, A.: Epidemic outbreaks in complex heterogeneous networks. The European Physical Journal B-Condensed Matter and Complex Systems **26**(4), 521–529 (2002)
8. Panzarasa, P., Opsahl, T., Carley, K.M.: Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. Journal of the American Society for Information Science and Technology **60**(5), 911–932 (2009). https://doi.org/10.1002/asi.21015, https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21015
9. Taylor, D., Caceres, R.S., Mucha, P.J.: Super-resolution community detection for layer-aggregated multilayer networks. Phys. Rev. X **7**, 031056 (Sep 2017). https://doi.org/10.1103/PhysRevX.7.031056, https://link.aps.org/doi/10.1103/PhysRevX.7.031056
10. Valente, T.W.: Social network thresholds in the diffusion of innovations. Social Networks **18**(1), 69–89 (Jan 1996). https://doi.org/10.1016/0378-8733(95)00256-1, http://www.sciencedirect.com/science/article/pii/0378873395002561, uses thresholds for diffusion, shows why they are good if different scales are used at the same time
11. Volz Erik, Meyers Lauren Ancel: Susceptibleinfectedrecovered epidemics in dynamic contact networks. Proceedings of the Royal Society B: Biological Sciences **274**(1628), 2925–2934 (Dec 2007). https://doi.org/10.1098/rspb.2007.1159, https://royalsocietypublishing.org/doi/full/10.1098/rspb.2007.1159
12. Whitbeck, J., de Amorim, M.D., Conan, V., Guillaume, J.L.: Temporal Reachability Graphs. arXiv:1207.7103 [cs] (Jul 2012), http://arxiv.org/abs/1207.7103, arXiv: 1207.7103
13. Xia, C., Wang, L., Sun, S., Wang, J.: An sir model with infection delay and propagation vector in complex networks. Nonlinear Dynamics **69**(3), 927–934 (2012)