Spreading predictability in complex networks

Na Zhao[2,3]◑, Jian Wang[4]◑, Yong Yu[2]◑, Jun-Yan Zhao[5], Duan-Bing Chen[1,6]*

1 School of Computer Science & Engineering, Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, People's Republic of China
2 Key Laboratory in Software Engineering of Yunnan Province, School of Software, Yunnan University, Kunming, People's Republic of China
3 Electric Power Research Institute of Yunnan Power Grid, Kunming, People's Republic of China
4College of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, People's Republic of China
5 Beijing Special Vehicle Institute, Beijing, People's Republic of China
6 Union Big Data, Chengdu, People's Republic of China

◑These authors contributed equally to this work.
* dbchen@uestc.edu.cn

Abstract

Spreading dynamics analysis is an important and interesting topic since it has many applications such as rumor or disease controlling, viral marketing and information recommending. Many state-of-the-art researches focus on predicting infection scale or threshold. Few researchers pay attention to the predicting of infection nodes from a snapshot. With developing of precision marketing, recommending and, controlling, how to predict infection nodes precisely from snapshot becomes a key issue in spreading dynamics analysis. In this paper, a probability based prediction model is presented so as to estimate the infection nodes from a snapshot of spreading. Experimental results on synthetic and real networks demonstrate that the model proposed could predict the infection nodes precisely in the sense of probability.

Introduction

Spreading dynamics is an important issue in spreading and controlling [1–3] of rumor [4–7] and disease [8–11], marketing [12], recommending [13–15], source detecting [16, 17], and many other interesting topics [18–22]. How to predict the infection probability [23], infected scale [24, 25], and even the infected nodes precisely has been gotten much attention in recent years.

Researchers have gotten many achievements on macro level of spreading such as phase transition of spreading [26] and basic reproduction number [27]. Up to now, many researches focus on estimating of infection scale. The simplest one is mean-field model, in which, the spreading coverage can be predicted by using differential equations [24]. Besides mean-field model, some more realistic models such as pair approximation [25] and permutation entropy [28] are considered to predict the infection scale or infectious disease outbreaks. The main difference between mean-field and pair approximation is that the former(latter) approximates high-order moments in term of first (second) order ones. In [28], the researchers studied the predictability of a diverse collection of outbreaks and identified a fundamental entropy barrier for disease time series forecasting through adopting permutation entropy as a model independent measure of predictability. Funk et al [29] presented a stochastic semi-mechanistic model of infectious disease dynamics that was used in real time during the 2013–2016 West African Ebola epidemic to fit the simulated trajectories in the Ebola Forecasting Challenge, and to produce forecasts that were compared to following data points. Venkatramanan et al [30] proposed a data-driven agent-based model framework for forecasting the 2014–2015 Ebola epidemic in Liberia, and subsequently used during the Ebola forecasting challenge. The data-driven approach can be refined and adapted for future epidemics, and share the lessons learned over the course of the challenge. Zhang et al [31] proposed a measurement to state the efforts of users on Twitter to get their information spreading. They found that small fraction of users with special performance on participation can gain great influence, while most other users play a role as middleware during the information propagation.

Up to now, most researches are focused on macro level of spreading prediction, but few on micro level. However, the detailed infected individuals should be known so as to contain the spread of serious infectious diseases such as SARS [32,33] and H7N7 [34,35]. Besides aspect of macro level of spreading, we should pay attention to some more details besides the general infection coverage so as to achieve fine prediction. Chen et al. did some interesting works on this area [23]. They presented an iterative algorithm to estimate the infection probability of the spreading process and then apply it to mean-field approach to predict the spreading coverage.

Combing mean-field or pair approximation models with infection probability estimating strategy [23], the number of infected nodes from a given snapshot of the propagation on network can be predicted, but can not determine which nodes will be infected. In this paper, we present a probability based prediction model to estimate the infection probability of a node, further, to determine the nodes being infected in the future.

Materials and methods

For a given snapshot, a susceptible node can be infected by a probability in the future. Denoting by $P_u(t)$ the score of node u at time t, we have,

$$P_u(t) = 1 - \prod_{v \in \Gamma_u} (1 - \mu P_v(t-1)), \tag{1}$$

where Γ_u is the neighbors of node u and infected probability μ is estimated by IAIP model (Iterative Algorithm for estimating the Infection Probability) [23]. Since an infected node always attempts to infect its susceptible neighbor once time and a recovered node doesn't infect any of its susceptible neighbor, so, in Eq. (1), for node v, it is reasonable to assume that $P_v(t) = 1$ for infected node and $P_v(t) = 0$ for recovered node. For susceptible node u, the probability to be infected at time t is $P_u(t)$. Obviously, the initial condition is,

$$P_u(0) = \begin{cases} 0 & \text{if node } u \text{ is susceptible or recovered} \\ 1 & \text{if node } u \text{ is infected} \end{cases}, \tag{2}$$

In Eq. (1), the score $P_u(t)$ for susceptible node u will be converged to a unique steady state denoted by $P_u(t_c)$, where t_c is the convergence time. The final score $P_u = P_u(t_c)$ is the probability to be infected of susceptible node while spreading achieves steady state.

Fig. 1 is a toy network with 24 nodes. The snapshot includes 5 recovered nodes and 1 infected node, as shown in Fig. 1(a). A certain spreading simulation result, average result on 10000 simulations, and result of probability prediction model from snapshot are shown in Figs. 1(b), (c) and (d) respectively. From this toy network, it can be seen that the result obtained by the probability prediction model is coincident with that by the average over 10000 simulations very well, that is, nodes 7, 8, and 19 have high probability to be infected, nodes 2 and 9 have middle probability to be infected, while other nodes have relatively low probability to be infected, as shown in Fig. 1(c) and (d). At the same time, Figs. 1(b) and (c) reflect the correlation between a certain spreading simulation and average over 10000 simulations. In order to describe how well a certain spreading simulation relative to average over 10000 simulations and result obtained by probability prediction model relative to the result of average over 10000 simulations, we use predictability χ and Pearson correlation ρ to evaluate our model. These two metrics can be calculated by:

$$\chi = \frac{1}{N}\sum_{l=1}^{N} cosin(\overrightarrow{p}_r^l, \overrightarrow{p}_r), \tag{3}$$

$$\rho = Pearson(\overrightarrow{p}_r, \overrightarrow{p}_e), \tag{4}$$

where $\overrightarrow{p}_r = \frac{1}{N}\sum_{l=1}^{N} \overrightarrow{p}_r^l$, $\overrightarrow{p}_r^l$ is the vector of infected frequency of nodes on the l^{th} simulation and $\overrightarrow{p}_e$ is the vector of infected probability of nodes obtained by probability prediction model. The element $\overrightarrow{p}_r^l(u)$ of $\overrightarrow{p}_r^l$ is determined by:

$$\overrightarrow{p}_r^l(u) = \begin{cases} \frac{1}{Q_l} & \text{if node } u \text{ is infected in the } l^{th} \text{ simulation from snapshot} \\ 0 & \text{otherwise} \end{cases}, \tag{5}$$

where Q_l is the number of infected nodes of the l^{th} simulation from snapshot.

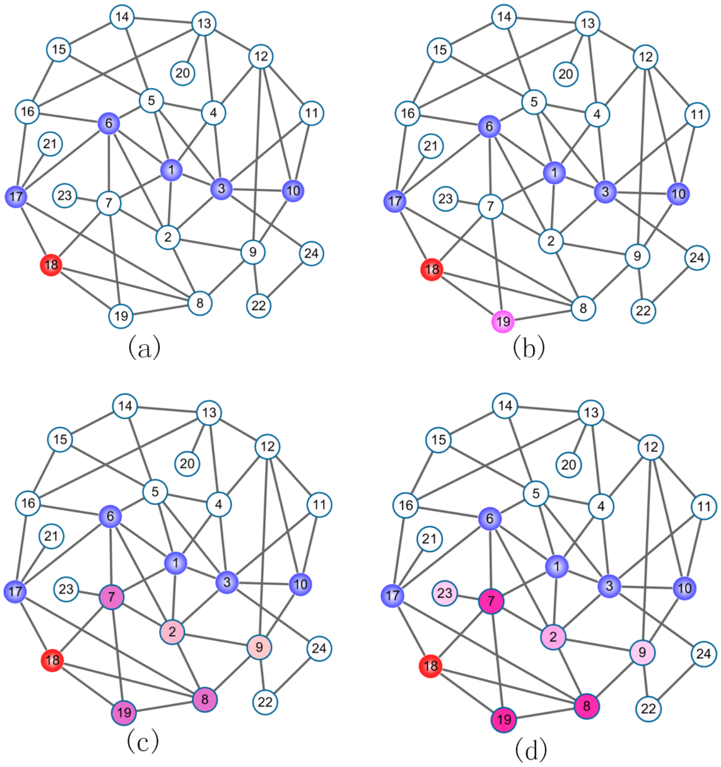
Results and Discussion

To simulate the spreading process on networks, we employ the Susceptible-Infected-Removed (SIR) model [36]. In a network, we randomly select one node as the initial spreader. The information from this node will infect each of this node's susceptible neighbors with probability μ, namely the infection probability. After infecting neighbors, the node will immediately become recovered (i.e., the recovering probability is 1). The new infected nodes in next step will infect their neighbors as the initial node. If it is not specially stated, we take the snapshot after five steps of spreading from the initial node as the known information.

We test our method on synthetic and real networks. Synthetic networks are Wattes-Strogatz (WS) networks [37], Barabási-Albert (BA) networks [38] and Given-Newman (GN) community networks [39]. Each synthetic network has 4000 nodes and each GN community network has 40 communities. We will discuss our model on three aspects: (1) the effect of infected probability μ, (2) the effect of structure of networks, and (3) the effect of stage of snapshot.

The effect of infected probability

Fig. 2 shows the predictability χ and correlation ρ under different infected probability μ on WS, BA and GN networks. Generally, the predictability and correlation get larger with μ getting larger. For very large μ, e.g., $\mu = 0.3$, the predictability and correlation approach to 1 since most of nodes will be infected. From Fig. 2, it can be seen that there exists a transition point, in detail, the transition point at $\mu = 0.15$ for WS network and at $\mu = 0.1$ for GN network. This can be explained as follows: the



(a) (b)

(c) (d)

Fig 1. (Color online)A toy network with 24 nodes. (a) The snapshot includes 5 recovered nodes, i.e., 1, 3, 6, 10, 17, and 1 infected node, i.e., node 18, (b) a certain spreading simulation result from snapshot, only node 19 is infected when spreading achieves steady state, (c) average result on 10000 simulations from snapshot, and (d) result of probability prediction model from snapshot. In (c) and (d), the shades of nodes indicate the probability to be infected when spreading achieves steady state.

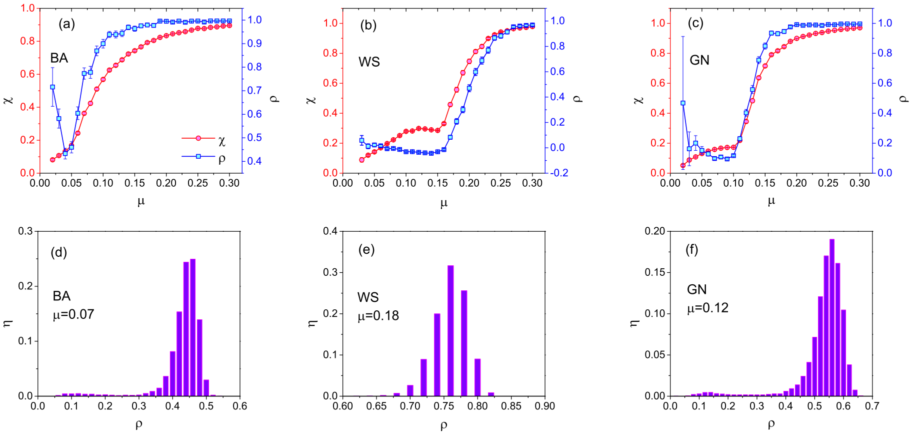


Fig 2. The predictability χ and correlation ρ under different infected probability μ on (a) BA, (b) WS and (c) GN networks. The distribution of Pearson coefficient in (d) BA, (e) WS and (f) GN are shown. The network parameters are $N = 4000, \langle k \rangle = 10, p = 0.1$ for WS network, $N = 4000, \langle k \rangle = 10$ for BA network, and $N = 4000, \langle k \rangle = 10, \langle k_{in} \rangle = 7$ for GN network. The error bar in (a-c) and the distribution of correlation ρ in (d-f) are obtained by the results under 200 snapshots

information almost do not diffusion if μ is small ($\mu < 0.15$ for WS networks and $\mu < 0.1$ for GN network), and the infected nodes are highly random for different simulations. It is noted that in BA network, it almost do not exist transition point. It can be explained as follows: since the heterogenous of its topological structure, regardless the location of initial infected node, the information will easily reach to the node with large degree, eventually, reach to other nodes. Interestingly, if μ is very small (e.g., $\mu = 0.02$), the correlation is getting large in BA network, as shown in Fig. 2(a). Actually, for very small μ, just only a few snapshots have infected nodes, the results have no statistical significance. Besides, the distribution of correlation ρ under the results of 200 independent runs are listed in Figs. 2(d-f). From these three subfigures, it can be seen that the distributions of correlation ρ of BA and GN networks are similar, while that of WS network are generally large comparing with BA and GN networks.

The effect of structure of networks

Fig. 3 shows the predictability and correlation for three types of networks with different structural parameters. For WS network, we study the effect of the rewiring parameter p on predictability and correlation. For BA network, we consider a variant of it in which each new node u connects to the existing node v with probability $p_u = (k_u + B)/\sum_v (k_v + B)$ [40, 41]. This modified model allows a selection of the exponent of the power-law scaling in the degree distribution $p(k) \sim k^{-\gamma}$ with $\gamma = 3 + B/m$ in the thermodynamic limit where m is the number of nodes should be connected when a new node is added and B is tunable parameter. With this network, we study the effect of B on predictability and correlation. For GN network, we study the effect of $\langle k_{in} \rangle$ on predictability and correlation, where $\langle k_{in} \rangle$ is the average internal degree of nodes in community. For a node u in community C, its internal degree k_u^{in} can de written as:

$$k_u^{in} = \sum_{u,v} \delta_{u,v}, \tag{6}$$

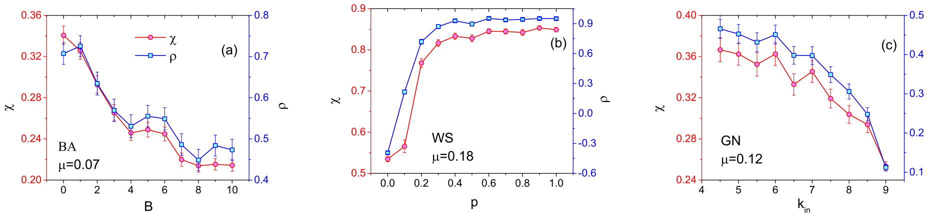


Fig 3. The predictability χ and correlation ρ for three types of networks with different structural parameters. In (a), B is a tunable parameter while generating network, (b) p is rewiring probability, and (c) $\langle k_{in} \rangle$ is average internal degree.

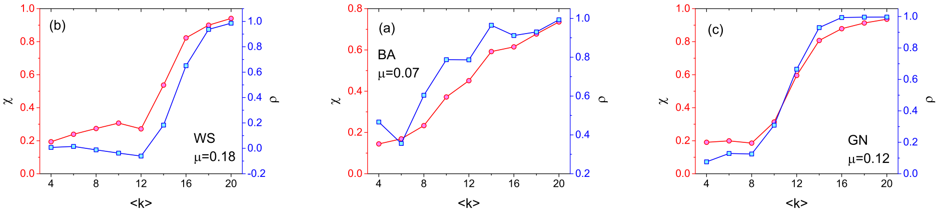


Fig 4. The effect of average node degree $\langle k \rangle$ on predictability χ and correlation ρ.

$\delta_{u,v} = 1$ if v is also in community C, otherwise $\delta_{u,v} = 0$. For standard BA network, i.e., $B = 0$, there are a few nodes with extremely large degree, the information can be spread out easily so long as it reaches to a node with large degree. So, it is relatively easy to predict which node will be infected in the future. As the B increases, the network evolves to random, a node getting infected or not will be hard to predict relatively, so the predictability and correlation decrease when B increases, as shown in Fig. 3(a). If rewiring probability $p < 0.2$, the information is hard to diffusion to other nodes since the WS network is almost regular, so it is hard to predict the infected nodes. When rewiring probability $p > 0.2$, the network has relatively strong random, the information reaches to other nodes easily, consequently, it is easy to predict the infected nodes, as shown in Fig. 3(b). In GN network, if average internal degree $\langle k_{in} \rangle$ is larger, the community structure is clearer, correspondingly, the information is hard to escape the community boundary, and the predictability and correlation will getting worse, as shown in Fig. 3(c).

Besides the network parameter listed above, the density of network, i.e., average node degree $\langle k \rangle$, also affects the predictability and correlation, as shown in Fig. 4. From Fig. 4, it can be seen that the *predictability* and correlation are small for small average node degree $\langle k \rangle$. Especially in WS and GN networks, for a large scope of average node degree ($\langle k \rangle < 12$ in WS and $\langle k \rangle < 8$ in GN), the predictability and correlation are extremely small, there exists an obvious transition points, as shown in Fig. 4(a) and (c).

The effect of stage of snapshot

We further analyze the predictability χ and correlation ρ under different stage of snapshot, as shown in Fig. 5. In Fig. 5, T is the spreading time of snapshot. Generally, it is difficult to estimate the infected rate precisely if just the snapshot in the early stage is given since there is little usable information, so, it is hard to predict the infected nodes. As T increases, more information could be used, the predictability χ and correlation ρ are getting better. In the late stage, many nodes of snapshot are

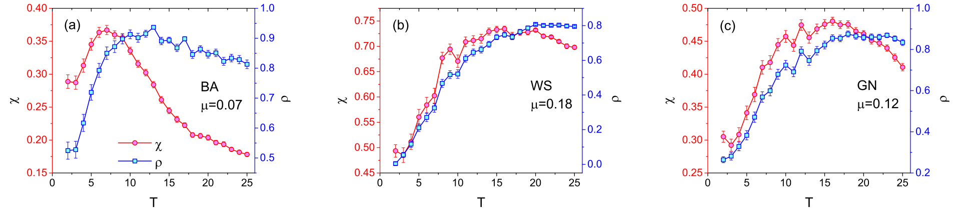


Fig 5. The predictability χ and correlation ρ under different stage of snapshot. Smaller T indicates earlier stage and larger T indicates latter stage.

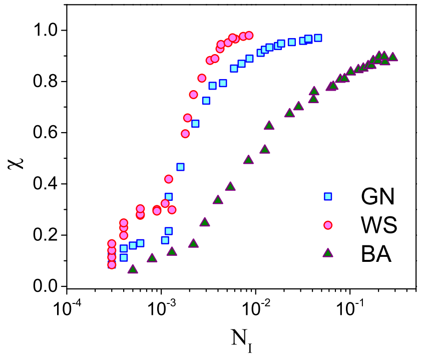


Fig 6. The correlation between the number of infected nodes N_I and the predictability χ.

infected or recovered, the left nodes are hard to be infected in reality, so the predictability χ and correlation ρ are getting worse, especially in BA network since most of all nodes are recovered.

Actually, if we analyze the number of infected nodes of snapshot in Figs. 2-5, we can find an interesting phenomena, as shown in Fig. 6. There is an obviously positive correlation between the number of infected nodes of snapshot and predictability χ. At the same time, it can be seen that the WS network has the strongest positive correlation while BA network has the weakest positive correlation under same number of infected nodes of snapshot. This might be universal, more infected nodes exist in snapshot, the information will be diffused easier, and so, it is more easy to predict the infected nodes in the future, correspondingly, the predictability χ will getting better.

Besides synthetic networks, we also analyze the predictability χ and correlation ρ on 11 real networks. The properties and analysis results on these real networks are shown in Table 1. From Table 1, it can be seen that the results are rather good, especial for the case of large N_I, this is consistent with the results in Fig. 6. For networks Y2H and power, the predictability χ and correlation ρ are extremely low since N_I is very small. Actually, in these cases, there are few infected nodes in snapshot of spreading. Furthermore, the networks are very sparse, so, it is hard to predict the nodes being infected from snapshot in the future.

Conclusion

Up to now, most of researches mainly focus on the infection scale or threshold when they study the spreading dynamics in complex networks. However, following questions may be more important and interesting: Which nodes will be infected in the future

Table 1. The properties and analysis results on 11 real networks. The infected probability $\mu = 0.15$.

Networks	#Nodes	#Edges	χ	ρ	N_I
cond-mat	39577	175693	0.7612	0.9430	0.0152
astro-Ph	16046	121251	0.8358	0.9426	0.0575
email	1133	5451	0.7854	0.9860	0.0628
c.elegens	453	2025	0.5577	0.9900	0.1143
ecoli	230	695	0.6110	0.9558	0.0509
internet	22963	48436	0.4525	0.9541	0.0625
PGP	10680	24316	0.6292	0.8074	0.0069
TAP	1373	6833	0.5998	0.5897	0.0101
HEP	7610	15751	0.4396	0.5975	0.0016
Y2H	1846	2203	0.2618	0.3214	0.0016
power	4941	6594	0.2375	0.2762	0.0003

and how to predict these nodes precisely? In this paper, we focused on this topic and presented a probability based prediction model to predict the infection nodes. Three synthetic and eleven real networks are used to evaluate the proposed model. Experimental results demonstrate that the model proposed could predict the infection nodes precisely in the sense of probability. In this paper, we just discuss the prediction model on static networks. The analyzing will get more difficult if the networks are evolving [42–44]. Furthermore, we analyze the effect of structure of networks, but we don't consider the moving or self-protecting of individuals while disease outbreaks. Actually, as the diseases information makes individuals alert and take measures to prevent the diseases, the effective protection is more striking in small community [45]. We will study these more comprehensive cases deeply in the future.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China with Grant No 61673085, by the Science Strength Promotion Programme of UESTC under Grant No. Y03111023901014006 and by the National Key Research and Development Program of China under Grant No. 2018YFB2100100.

References

1. Ruan Z, Tang M, Liu Z. Epidemic spreading with information-driven vaccination. Phys. Rev. E. 2012; 86: 036117.

2. Iribarren JL, Moro E. Impact of human activity patterns on the dynamics of information diffusion. Phys. Rev. Lett. 2009; 103: 038702.

3. Arruda GF, Petri G, Rodrigues FA, Moreno Y. Impact of the distribution of recovery rates on disease spreading in complex networks. Phys. Rev. Research. 2020; 2: 013046.

4. Zhang Y, Zhou S, Zhang Z, Guan J, Zhou S. Rumor evolution in social networks. Phys. Rev. E. 2013; 87: 032133.

5. Doerr B, Friedrich T, Sauerwald T. Quasirandom Rumor Spreading. ACM Trans. Alg. 2014; 11: 9.

6. Ma J, Li D, Tian Z. Rumor spreading in online social networks by considering the bipolar social reinforcement. Physica A. 2016; 447: 108–115.

7. Kwon S, Cha M, Jung K. Rumor detection over varying time windows. PLoS ONE. 2017; 12(1): e0168344.

8. Meloni S, Perra N, Arenas A, Gómez S, Moreno Y, Vespignani A. Modeling human mobility responses to the large-scale spreading of infectious diseases. Sci. Rep. 2011; 1: 62.

9. Goltsev AV, Dorogovtsev SN, Oliveira JG, Mendes JFF. Localization and spreading of diseases in complex networks. Phys. Rev. Lett. 2012; 109: 128702.

10. Granell C, Gómez S, Arenas A. Competing spreading processes on multiplex networks: Awareness and epidemics. Phys. Rev. E. 2014; 90: 012808.

11. Leventhal GE, Hill AL, Nowak MA, Bonhoeffer S. Evolution and emergence of infectious diseases in theoretical and real-world networks. Nat. Commu. 2015; 8: 6101.

12. Miquel-Romero M-J, Adame-Sánchez C. Viral marketing through e-mail: the link consumer-company. Management Decision. 2013; 51: 1970–1982.

13. Lü L, Medo M, Yeung CH, Zhang Y-C, Zhang Z-K, Zhou T. Recommender systems. Phys. Rep. 2013; 519: 1-49.

14. Ren X, Lü L, Liu R, Zhang J. Avoiding congestion in recommender systems. New J. Phys. 2014; 16: 063057.

15. Chen D-B, Zeng A, Cimini G, Zhang Y-C. Adaptive social recommendation in a multiple category landscape. Eur. Phys. J. B. 2013; 86: 61.

16. Pinto PC, Thiran P, Vetterli M. Locating the source of diffusion in large-scale networks. Phys. Rev. Lett. 2012; 109: 068702.

17. Shen Z, Cao S, Wang W-X, Di Z, Stanley HE. Locating the source of diffusion in complex networks by time reversal backward spreading. Phys. Rev. E. 2016; 93: 032301.

18. Seebens H, Schwartz N, Schupp PJ, Blasius B. Predicting the spread of marine species introduced by global shipping. Proc. Natl. Acad. Sci. USA. 2016; 113(20): 108–115.

19. Chen D-B. Empirical study on structural properties in temporal networks under different time scales. Eur. Phys. J. B. 2015; 88: 320.

20. Lü L, Chen D-B, Zhou T. The small world yields the most effective information spreading. New J Phys. 2011; 13: 123005.

21. Cimini G, Chen D-B, Medo M, Lü L, Zhang Y-C. Enhancing topology adaptation in information sharing social networks. Phys. Rev. E. 2012; 85: 046108.

22. Centola D. The spread of behavior in an online social network experiment. Science. 2010; 329: 1174–1197.

23. Chen D-B, Xiao R, Zeng A. Predicting the evolution of spreading on complex networks. Sci. Rep. 2014; 4: 6108.

24. Pastor-Satorras R, Vespignani A. Epidemic spreading in scale-free networks. Phys. Rev. Lett. 2001; 86: 3200.

25. Mata AS, Ferreira RS, Ferreira SC. Heterogeneous pair-approximation for the contact process on complex networks. New J. Phys. 2014; 16: 053006.

26. Döbereiner H-G, Dubin-Thaler B, Giannone G, Xenias HS, PSheetz M. Dynamic phase transitions in cell spreading. Phys. Rev. Lett. 2004; 93: 108105.

27. Rodrigues HS, Monteiro MTT, Torres DFM, Zinober A. Dengue disease, basic reproduction number and control. Int. J. Comput. Math. 2012; 89: 334–346.

28. Scarpino SV, Petri G. On the predictability of infectious disease outbreaks. Nat. Commun. 2019; 10: 898.

29. Funk S, Camacho A, Kucharski, AJ, Eggo RM, Edmunds WJ. Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model. Epidemics. 2018; 22: 56–61.

30. Venkatramanana S, Lewis B, Chen J, Higdon D, Vullikanti A, Marathe M. Using data-driven agent-based models for forecasting emerging infectious diseases. Epidemics. 2018; 22: 43–49.

31. Zhang X, Han D-D, Yang R, Zhang Z. Users' participation and social influence during information spreading on Twitter. PLoS One. 2017; 12(9): e0183290.

32. Ksiazek TG, Erdman D, Goldsmith CS, et al. A novel coronavirus associated with severe acute respiratory syndrome. N Engl J Med. 2003; 348:1953–1966.

33. Rota PA, Oberste MS, Stephan S. Monroe SS, et al. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. Science. 2003; 300(5624):1394–1399.

34. Fouchier RAM, Schneeberger PM, Frans W. Rozendaal FW, et al. Avian influenza A virus (H7N7) associated with human conjunctivitis and a fatal case of acute respiratory distress syndrome. Proc Natl Acad Sci U S A. 2004; 101(5): 1356–1361.

35. Koopmans M, Wilbrink B, Conyn M, et al. Transmission of H7N7 avian influenza A virus to human beings during a large outbreak in commercial poultry farms in the Netherlands. Lancet. 2004; 363(9409): 582–583.

36. Anderson RM, May RM, Anderson B. Infectious diseases of humans:dynamics and control. Boston: Oxford Univ. Press. 1992.

37. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature. 1998; 393: 440–442.

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

39. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys. Rev. E. 2004; 69(2): 026113.

40. Albert R, Barabási A-L. Statistical mechanics of complex networks. Rev. Mod. Phys. 2002; 74: 47–97.

41. Dorogovtsev SN, Mendes JFF. Evolution of networks. Adv. Phys. 2002; 51: 1079–1187.

42. Holme P, Saramäki J. Temporal networks. Phys. Rep. 2012; 519: 97–125.

43. Darbon A, Colombi D, Valdano E, Savini L, Giovannini A, Colizza V. Disease persistence on temporal contact networks accounting for heterogeneous infectious periods. R. Soc. Open Sci. 2019; 6: 181404.

44. Zino L, Rizzo A, Porfiri M. Analysis and control of epidemics in temporal networks with self-excitement and behavioral changes. Eur. J Control. 2020; In press.

45. Liu T, Li P, Chen Y, Zhang J. Community size effects on epidemic spreading in multiplex social networks. PLoS One. 2016; 11(3): e0152021.