SCIENCE CHINA
Information Sciences

• HIGHLIGHT •

February 2018, Vol. 61 028102:1–028102:3
https://doi.org/10.1007/s11432-017-9154-1

Network topology inference from incomplete observation data

Peng DOU, Guojie SONG* & Tong ZHAO

Key Laboratory of Machine Perception (MOE), Peking University, Beijing 100871*, China*

Received 23 May 2017/Accepted 29 June 2017/Published online 5 December 2017

Citation Dou P, Song G J, Zhao T. Network topology inference from incomplete observation data. Sci China Inf Sci, 2018, 61(2): 028102, https://doi.org/10.1007/s11432-017-9154-1

Introduction and related work. Network inference problem covers lots of fields in real life, such as viral-marketing, stopping rumors and controlling the spread of diseases. It aims to infer the hidden network topology with only the infection time stamps of nodes when various messages diffuse among the network. Most classical methods assume that the infection time of each infected node is fully observed [1, 2]. However, in real world scenarios, we usually only obtain incomplete cascades in which some infection time of activated node is missing. For example, microblog users may delete the published microblogs, making the release time hard to follow. Several researches focus on incomplete scenario recently. Refs. [3, 4] tackle with different scenario of incomplete as they consider snapshot data. The work of [5] distinguishes potential short edges that contain missing nodes and then adjusts the network structure inferred by other models. Refs. [6, 7] can recover the network edge weights but need topology as input. Our work is to efficiently infer the network structure from incomplete cascades without knowledge of topology.

Our main challenge is to recover the missing time stamp, in case that those nodes are incorrectly regarded as failing to transmit and then mislead the transmission probability. We propose a Greedy-NIIC (network inference on incomplete cascades) algorithm to recover the cascades via Monte-Carlo simulation diffusion process and greedily select the edge that can maximize the marginal gain. Experiments on both synthetic and real-world data reveal that NIIC can accurately recover the network structure from incomplete cascades comparing with existing methods.

Model. Given a hidden directed network $G(V, E)$ that contains n nodes, the information is diffused over the network and leaves a trace or cascade $\boldsymbol{t} = \{t_1, t_2, \ldots, t_n\}$, which is a n-dimensional vector that records the time when each node gets infected. Observations are recorded as a set C of cascades $\{\boldsymbol{t}^1, \boldsymbol{t}^2, ..., \boldsymbol{t}^c\}$.

However, the infection time of some nodes may be missing. In each cascade, set $\widetilde{V}$ contains uninfected nodes or missing nodes whose infection time has not been observed. Then, the incomplete cascades set is denoted by $\widetilde{C} = \{\widetilde{\boldsymbol{t}^1}, \widetilde{\boldsymbol{t}^2}, ..., \widetilde{\boldsymbol{t}^c}\}$.

Our method is based on Continuous-Time Independent Cascade Model [8]. The diffusion process is controlled by activation probability γ_{uv} and the pairwise transmission likelihood $P(u, v)$. Former researches have pointed out that $P(u, v)$ usually follows Exponential, Power-Law or Rayleigh distribution. To simplify the problem, we assume that $P(u, v)$ follows Exponential law. The diffusion trace starting from root node and spreading to the others forms a directed transmission tree T. Many possible Ts can lead to a given cascade $\boldsymbol{t}$. The probability that $\boldsymbol{t}$ spreads among network G, denoted as $f(\boldsymbol{t}; G)$, is approximated by only considering the most likely transmission tree according to NetInf. The likelihood $f(\boldsymbol{t}; T)$ that a cas-

* Corresponding author (email: gjsong@pku.edu.cn)

cade spreads along a specific directed transmission tree is the joint likelihood of pairwise transmission likelihood $P(u, v)$. More details can be found in Appendix A.

Network inference on incomplete cascades aims to infer the structure of the network (with no more than m edges) underlying these incomplete cascades set $\widetilde{C}$, which is

$$\begin{aligned} G &= \text{argmax}_{|G|\leqslant m} f(\widetilde{C}; G) \\ &= \text{argmax}_{|G|\leqslant m} \sum_{\widetilde{\boldsymbol{t}}\in\widetilde{C}} f(\widetilde{\boldsymbol{t}}; G). \end{aligned} \tag{1}$$

The submodularity of this problem is proved in Appendix B. Maximizing submodular function has been proved to be **NP**-hard [9]. Greedy algorithm is commonly applied in such circumstance which can acquire at least $(1 - \frac{1}{e})$ of the optimal value. Starting from an empty graph, we iteratively add a new edge e_i at step i, which maximizes the marginal gain:

$$\begin{aligned} e_i =& \text{argmax}_{e\in G\setminus G_{i-1}} \sum_{\widetilde{\boldsymbol{t}}\in\widetilde{C}} f(\widetilde{\boldsymbol{t}}; G_{i-1} \cup e) \\ &- f(\widetilde{\boldsymbol{t}}; G_{i-1}), \end{aligned} \tag{2}$$

and stops when there have been already m edges in graph G.

To solve the problem, we first have to recover sufficient number of possible cascades for each incomplete cascade, so as to approximate the case of the hidden truth one. We use Monte-Carlo simulation to generate a set of possible cascades $D(\widetilde{\boldsymbol{t}})$. For each incomplete cascade $\widetilde{\boldsymbol{t}}$, it has a set $\widetilde{V}$ of nodes whose infection time is missing or who have not been affected. In each iteration of Greedy-NIIC algorithm, the Monte-Carlo is performed M times for every incomplete cascade based on current network structure. During iteration i of Greedy-NIIC, current network G_{i-1} contains $i-1$ edges. The simulation starts at a node u from $V\setminus\widetilde{V} = \{u_1, u_2, \ \ldots, u_{|V\setminus\widetilde{V}|}\}$ whose infection time is observed. u tries to activate all of its neighbors in $G_{i-1} \cup e$ (e is the current under-test edge) who have no infection time with predefined probability γ_{uv}. For simplicity, we assume γ is the same for every edge. If a neighbor v is activated, we firstly generate a random number U from $(0, 1)$ on uniform distribution, and then sample the t_v from Exponential distribution. After v has become active, v tries to activate its neighbors in G_{i-1} that have not got infection time. The diffusion process stops when no new node gets infected and it naturally forms a directed transmission tree $T(u)$ whose root is u. For each observed node like u from $V \setminus \widetilde{V}$, we perform the above process and thus get $|V \setminus \widetilde{V}|$ spanning trees: $T(u_1)$, $T(u_2)$,..., $T(u_{|V\setminus\widetilde{V}|})$. For each incomplete cascade $\widetilde{\boldsymbol{t}}$, we perform the process described above for M times. Each time produces one possible complete cascade of incomplete $\widetilde{\boldsymbol{t}}$. Those possible complete cascades should be different, in both the time stamps and topology of spanning trees. As M grows larger, the simulation results can cover more possible cases. The same process should be done for each incomplete cascade in $\widetilde{C}$. Some detail situations that may occur in the simulation process are introduced in Appendix C.

Then, we are able to approximate the likelihood of incomplete cascade $f(\widetilde{\boldsymbol{t}}; G)$ with the average likelihood of possible cascades in $D(\widetilde{\boldsymbol{t}})$ when M is sufficiently large. Finding the most likely transmission tree is required for computing the likelihood of each recovered possible cascade, in which the spanning trees in simulation process can help. One simulation of one cascade leaves us with $|V \setminus \widetilde{V}|$ spanning trees: $T(u_1), T(u_2), \ldots, T(u_{|V\setminus\widetilde{V}|})$. We find the parents of all the root nodes $u_1, u_2, \ldots, u_{|V\setminus\widetilde{V}|}$ except source s and connect all spanning trees into one new tree, namely maximum combination tree $\tilde{T}$. The parent of each observed node u is the node that maximizes the pairwise transmission likelihood $P(\text{par}(u), u)$. And we prove that the maximum combination tree is the most likely transmission tree in Appendix D.

Here are some methods for improving the efficiency of Greedy-NIIC algorithm.

(1) We can let the simulation time M decrease with the rise of edge number in the inferred network;

(2) If the marginal gain of adding edge e into network is less than threshold (e.g., the threshold is set as 10 in our experiments), this edge will be abandoned and will not be added in the following iterations.

Besides, applying greedy process on incomplete input dataset brings some controversial issues. Detailed greedy iteration rules are defined in order to make full use of the available cascade information, as shown in Appendix E.

Experimental evaluation. We evaluate our method on both synthetic and real world incomplete dataset. We take NetInf [1], NetRate [2] and PSE [5] as baselines and consider the F1-score and precision as evaluation metrics. In synthetic experiments, we use Kronecker Graph model to build the diffusion networks and generate synthetic incomplete cascades. We consider 3 different types of miss mode: Random Miss, Edge Miss and Block

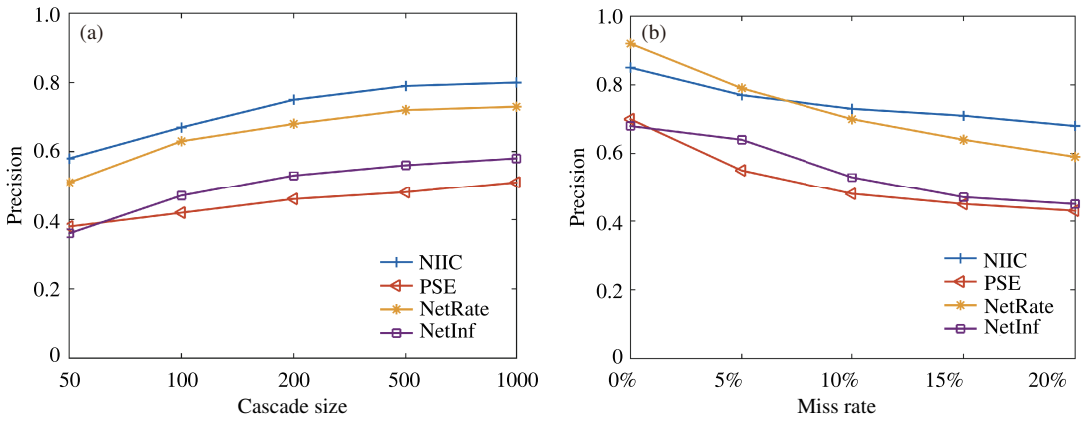

Figure 1 (Color online) Experiment results on real world data. (a) Performance on different cascade sizes; (b) performance on different miss rates.

Miss. We examine our model against different miss mode, network size, cascade size, miss rate and network structure. Experiment results in Appendix F show the excellent performance of our method in all circumstances. We also make experiments on real world dataset, namely the MemeTracker dataset with 500 active nodes. As shown in Figure 1, the precision of all four methods becomes higher with the increase of cascade number. Our method can achieve nearly 80% accuracy while the accuracy of NetInf is no more than 60%. As the miss rate increases, NIIC outperforms NetRate gradually. When the miss rate is 20%, our method can achieve over 65%, while the accuracy computed by the other three methods is below 60%.

Conclusion. In this article, we propose the problem of NIIC and develop an efficient method to conduct Monte-Carlo simulation on each greedy iteration, which shows excellent performance on accuracy in both synthetic and real world datasets.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 61572041), Beijing Natural Science Foundation (Grant No. 4152023), National High Technology Research and Development Program of China (863 Program) (Grant No. 2014AA015103).

Supporting information Appendixes A–F. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Rodriguez M G, Leskovec J, Krause A. Inferring networks of diffusion and influence. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington DC, 2010. 1019–1028
2 Rodriguez M G, Balduzzi D, Schölkopf B. Uncovering the temporal dynamics of diffusion networks. In: Proceedings of the 28th International Conference on Machine Learning, Bellevue, 2011. 561–568
3 Amin K, Heidari H, Kearns M. Learning from contagion (without timestamps). In: Proceedings of the 31st International Conference on Machine Learning, Beijing, 2014. 1845–1853
4 Sefer E, Kingsford C. Convex risk minimization to infer networks from probabilistic diffusion data at multiple scales. In: Proceedings of 2015 IEEE 31st International Conference, Seoul, 2015. 663–674
5 Dou P, Du S Z, Song G J. Inferring diffusion network on incomplete cascade data. In: Proceedings of the 17th International Conference on Web-Age Information Management, Nanchang, 2016. 325–337
6 Zong B, Wu Y H, Singh A K, et al. Inferring the underlying structure of information cascades. In: Proceedings of 2013 IEEE 13th International Conference on Data Mining, Brussels, 2012. 1218–1223
7 Lokhov A Y. Reconstructing parameters of spreading models from partial observations. In: Proceedings of the 29th Conference on Neural Information Processing Systems, Barcelona, 2016. 3459–3467
8 Kempe D, Kleinberg J, Tardos E. Maximizing the spread of influence through a social network. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, 2003. 137–146
9 Khuller S, Moss A, Naor J S. The budgeted maximum coverage problem. Inf Process Lett, 1999, 70: 39–45