On Minimum Spanning Trees and the Inference of Message Cascades

Bernhard C. Geiger

Know-Center GmbH, Graz 8010, Austria,
geiger@ieee.org,
WWW home page: www.know-center.at

1 Introduction and Problem Statement

We consider the problem of inferring how a message m of user u_0 spreads in a given directed, feed-based social network $\mathscr{G} = (V,E)$ where each node $v \in V$ corresponds to a user and where the edge $(u,v) \in E \subseteq V^2$ indicates that user v sees on her feed what user u posted or forwarded. In addition to $\mathscr{G}$, we have access to a partially ordered set of tuples $M = \{(u_0,t_0),\dots,(u_N,t_N)\}$, where (u_i,t_i) indicates that user u_i forwarded message m at time $t_i \geq t_{i-1}$; the original author u_0 posted the message at time t_0. Given $\mathscr{G}$ and M, we aim to determine the most likely paths this message has taken in the network, i.e., we wish to infer the *message cascade* of m. The problem has attracted a lot of attention and spawned literature regarding the inference [1–3], analysis [4–6], [7, Sec. 6.2], and prediction of such cascades [8, 9]. The more general problem of inferring the graph $\mathscr{G}$ from several sets M was considered in [10, 11].

Under assumptions similar to the independent cascade model [12], the most likely message cascade coincides with a minimum spanning tree for the directed network, where the weight of an edge is given by the log-probability of the event that a message is forwarded along this edge. We show that if the probability that a user forwards a message is independent of its sender, then the minimum spanning tree problem can be solved even without knowledge of the respective probabilities.

2 Message Cascades as Minimum Spanning Trees

We assume the independent cascade model for message forwarding. Specifically, let $p_{(u_i,u_j)}$ denote the probability that user u_j forwards a message from her feed that she received via user u_i; it may depend on the time u_i forwarded the message, the message content, the relationship between users u_i and u_j, the original author u_0, the message creation time t_0, the local time of user u_j, or on the time $|t_i - t_j|$ that elapsed since u_i forwarded the message. Mathematically, $p_{(u_i,u_j)} = p_{(u_i,u_j)}(u_0,t_0,t_i,t_j,m)$.

It can be shown that the most likely message cascade coincides with a minimum spanning tree of a subgraph of $\mathscr{G}$ that is compatible with M. To this end, let $\mathscr{G}_M = (V_M,E_M)$, where $V_M = \{u_0,u_1,\dots,u_N\}$ and where an edge $(u_i,u_j) \in E_M$ if and only if $(u_i,u_j) \in E$ and $t_j \geq t_i$. Let $\mathscr{T}_M$ denote the set of directed spanning trees of $\mathscr{G}_M$ rooted at u_0. Depending on the behavior of the feed, further edges may need to be removed from E_M; e.g., if the feed of user u_j only shows the first forwarded instance of m [8, Sec. 4].

All trees $T \in \mathscr{T}_M$ are valid message cascades, i.e., compatible with $\mathscr{G}$ and M. It remains to determine the most likely message cascade. Under the assumed probabilistic model, the log-likelihood of T can be computed as (e.g., [11, p. 485])

$$\mathsf{LL}(T) = \sum_{(u_i,u_j)\in\mathsf{edges}(T)} \log p_{(u_i,u_j)}(u_0,t_0,t_i,t_j,m). \tag{1}$$

Thus, inferring the most likely message cascade can be achieved by determining the minimum spanning tree of $\mathscr{G}_M$ rooted at u_0, with the weight of edge (u_i,u_j) chosen as $-\log p_{(u_i,u_j)}(u_0,t_0,t_i,t_j,m)$. This can be done in $\mathscr{O}(|E_M|+(N+1)\log(N+1))$ [13].

Learning or modeling the probabilities $p_{(u_i,u_j)}(u_0,t_0,t_i,t_j,m)$, which are required to determine the most likely cascade T, is non-trivial [12]. Under certain simplifying assumptions, however, the problem becomes tractable. Namely, suppose that the probability that user u_j forwards a message does not depend on the user u_i from which it was received, nor at the time t_i at which it was received. In other words, we have $p_{(u_i,u_j)}(u_0,t_0,t_i,t_j,m) = p_{(u_i,u_j)}(u_0,t_0,t_j,m) = p_{(u_k,u_j)}(u_0,t_0,t_j,m) := p_{u_j}(u_0,t_0,t_j,m)$, where the second equality holds for all k such that $(u_k,u_j) \in E$. To see how this simplifies the problem of maximizing (1) over $\mathscr{T}_M$, note that any directed spanning tree in $\mathscr{T}_M$ has N edges and each node u_i, $i = 1,\ldots,N$ has in-degree one (u_0 has no incident edges). It follows that (1) evaluates to

$$\mathsf{LL}(T) = \sum_{j=1}^{N} \log p_{u_j}(u_0,t_0,t_j,m) \tag{2}$$

for every $T \in \mathscr{T}_M$. Therefore, under this assumption, every spanning tree of $\mathscr{G}_M$ rooted at u_0 is minimum and all corresponding message cascades are equally likely.

Under the additional assumption that the feed of user u_j only shows a single forwarded instance of m (e.g., the first [8, Sec. 4], the most recent, or even just a randomly selected one), node u_j has in-degree one in $\mathscr{G}_M$, i.e., $\mathscr{G}_M$ is already a tree.

3 Practical Implications

Eq. (2) allows to determine the most likely message cascades compatible with $\mathscr{G}$ and M by simply determining the set of spanning trees of $\mathscr{G}_M$ *without knowledge of the forwarding probabilities* $p_{u_j}(u_0,t_0,t_j,m)$.

We remain to discuss how realistic the simplifying assumptions are. The assumption that $p_{(u_i,u_j)}(u_0,t_0,t_i,t_j,m) = p_{u_j}(u_0,t_0,t_j,m)$ implies that user u_j decides whether or not to forward m exclusively based on the message content, the identity of the original author, the message creation time t_0, and on the time t_j she considers forwarding it. The user does not base her decision on i) the user u_i from whom she received the message or ii) the time t_i at which the message appeared on her feed. Instantiating i) for, e.g., the Twitter network means that user u_j retweets a message m irrespective of the user u_i who retweeted it such that it appeared on her feed. This assumption is realistic, since for retweets appearing in the Twitter feed, the identity of the retweeter u_i appears less prominently than the identity of the original author u_0. Instantiating ii) would require that $t_j - t_i$ is small enough such that the message m appears on the feed of user u_j. This assumption is unproblematic for an active user u_j with appropriate feed settings or may be enforced by eliminating edges from E_M for which $t_j - t_i$ exceeds a certain threshold.

References

1. Taxidou. I., Fischer, P.M.: Online analysis of information diffusion in Twitter. Proc. Int. Conf. on World Wide Web (WWW), 1313 – 1318 (2014)
2. Zola P., Cola G., Mazza M., Tesconi M.: Interaction Strength Analysis to Model Retweet Cascade Graphs. Applied Sciences 10(23), 8394 (2020)
3. Cogan, P., Andrews, M., Bradonjic, M., Kennedy, W.S., Sala, A., Tucci, G.: Reconstruction and analysis of Twitter conversation graphs. Proc. ACM Int. Workshop on Hot Topics on Interdisciplinary Social Networks Research (HotSocial), 25 – 31 (2012)
4. ten Thij, M., Ouboter, T., Worm, D., Litvak, N., van den Berg, H., Bhulai, S.: Modelling of Trends in Twitter Using Retweet Graph Dynamics. Proc. Int. Workshop on Algorithms and Models for the Web-Graph (WAW), 132 – 147 (2014)
5. Webberley, W., Allen, S., Whitaker, R.: Retweeting: A study of message-forwarding in Twitter. Proc. Workshop on Mobile and Online Social Networks, 13 – 18 (2011)
6. Rizoiu, M.-A., Graham, T., Zhang, R., Zhang, Y., Ackland, R., Xie, L.: #DebateNight: The Role and Influence of Socialbots on Twitter During the 1st 2016 U.S. Presidential Debate. Proc. Int. AAAI Conf. on Web and Social Media, 12(1) 300 – 309 (2018)
7. Kwak, H., Lee, C., Park, H., Moon., S.: What is Twitter, a social network or a news media? Proc. Int. Conf. on World Wide Web (WWW), 591 – 600 (2010)
8. Kupavskii, A., Ostroumova, L., Umnov, A., Usachev, S., Serdyukov, P., Gusev, G., Kustarev, A.: Prediction of retweet cascade size over time. Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM), 2335 – 2338 (2012)
9. Huang, Z., Wang, Z., Zhu, Y., Yi, C., Su, T.: Prediction of Cascade Structure and Outbreaks Recurrence in Microblogs. Proc. Chinese National Conf. on Social Media Processing (SMP), 53 – 64 (2017)
10. Gomez-Rodriguez, M., Leskovec, J., Krause, A.: Inferring Networks of Diffusion and Influence. ACM Trans. Knowl. Discov. Data 5(4), 21 (2012)
11. Xu, S., Smith, D.A.: Contrastive Training for Models of Information Cascades. Proc. AAAI Conf. on Artificial Intelligence (AAAI), 483 – 490 (2018)
12. Dickens, L., Molloy, I., Lobo, J., Cheng, P., Russo, A.: Learning Stochastic Models of Information Flow. Proc. IEEE Int. Conf. on Data Engineering (ICDE), 570 – 581 (2012)
13. Gabow, H.N., Galil, Z., Spencer, T., Tarjan, R.E.: Efficient algorithms for finding minimum spanning trees in undirected and directed graphs. Combinatorica 6(2), 109 – 122 (1986)

Acknowledgments

The author thanks Meizhu Wang for suggesting literature. The work has been supported by the HiDALGO project and has been funded by the European Commission's ICT activity of the H2020 Programme under grant agreement number 824115. The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Climate Action, Environment, Energy, Mobility, Innovation and Technology, the Austrian Federal Ministry of Digital and Economic Affairs, and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.