Minimizing the Social Influence from a Topic Modeling Perspective

Qipeng Yao[1,2(✉)] and Li Guo[1]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
yaoqipeng0706@gmail.com, guoli@iie.ac.cn
[2] School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract. In this paper, we address the problem of minimizing the negative influence of undesirable things in a network by blocking a limited number of nodes from a topic modeling perspective. When undesirable thing such as a rumor or an infection emerges in a social network and part of users have already been infected, our goal is to minimize the size of ultimately infected users by blocking k nodes outside the infected set. We first employ the HDP-LDA and KL divergence to analysis the influence and relevance from a topic modeling perspective. Then two topic-aware heuristics based on betweenness and out-degree for finding approximate solutions to this problem are proposed. Using two real networks, we demonstrate experimentally the high performance of the proposed models and learning schemes.

Keywords: Influence minimization · Blocking nodes · Social networks

1 Introduction

In the past decade, the online social networks are providing convenient platforms for information dissemination and marketing campaign, allowing ideas and behaviors to flow along the social relationships in the effective word-of-mouth manner [1,2]. From the functional point of perspective, networks can mediate diffusion including not only positive information such as innovations, hot topics, and novel ideas, but also negative information like malicious rumors and disinformation [3]. Take the rumor for example, even with a small number of its initial adopters, the quantity of the ultimately infected users can be large due to triggering a word-of-mouth cascade in the network. Therefore, it is an urgent research issue to design effective strategies for reducing the influence coverage of the negative information and minimizing the spread of the undesirable things.

This problem has received a good deal of attention by the data mining research community in the last decade [4,5], but quite surprisingly, the characteristics of the item being the subject of the influence minimization has been left out of the picture.

C. Zhang et al. (Eds.): ICDS 2015, LNCS 9208, pp. 6–15, 2015.
DOI: 10.1007/978-3-319-24474-7_2

In this paper, we aim to minimize the spread of an existing undesirable thing by blocking a limited number of nodes in a network from a topic modeling perspective. More specifically, when some undesirable thing starts with some initial nodes and diffuses through the network under the topic-aware independent cascade (TIC) model, we consider finding a set of k nodes such that the resulting network by blocking those nodes can minimize the expected contamination area of the undesirable thing, where k is a given positive integer. We refer to this combinatorial optimization problem as the *influence minimization problem*. For this problem, we first employ the HDP-LDA and KL divergence to analysis the authoritativeness, influence and relevance from a topic modeling perspective. Then we propose two topic-aware heuristics based on betweenness and out-degree for finding approximate solutions to the problem. With two large real networks including Sina microblog and Facebook, we experimentally demonstrate that the proposed topic-aware node-removal heuristics outperform the well-studied notions of centrality measures.

2 Related Works

The research on finding influential nodes that are effective for the spread of information through a social network, namely Influence Maximization Problem, has attracted remarkable attention recently due to its novel idea of leveraging some social network users to propagate the awareness of products [2,6]. To improve the efficiency of seed selection, many heuristics and optimized greedy algorithms have been proposed, *e.g.*, DegreeDiscount [2], MIA [7], DAG [8], SIMPATH [9], ShortestPath [10], SPIN [11], CELF [12], CELF++ [13] and UBLF [14–16]. Besides, Guo et al. [17] investigated the influence maximization problem from the item-based data. Rodriguez et al. [18] studied the influence maximization problem in continuous time diffusion networks. Goyal et al. [19] proposed an alternative approach to influence maximization which, instead of assuming influence probabilities are given as input, directly uses the past available data. In the works [20,21] the authors discussed the integral influence maximization problem when repeated activations are involved. Zhou and Guo [22] established a constraint influence maximization framework for special targeted users. As a reverse problem, the source detection in a social network was discussed by Zang et al. [23,24]. However, the problem of minimizing the negative influence of undesirable things gets less attention, although it is an important research issue.

Some related research work has been made on minimizing the influence of negative information by removing nodes or links from a network [25,26]. It has been shown in particular that the strategies of removing nodes in decreasing order of out-degree can often be effective [5,27,28]. Kimura et al. proposed a links blocking method to minimize the expected contamination area of the network [4]. However, the fact of part nodes infected is not considered. Yu et al. addressed the problem of finding spread blockers are simply those nodes with high degree [29]. Budak et al. investigated the problem of influence limitation

where a bad campaign starts propagation from a certain node in the network and use the notion of limiting campaigns to counteract the effect of misinformation [3]. Different from previous work, our research cares more about a specific contamination scenario in the social network, and how to minimize the negative influence by blocking a small set of nodes from a topic modeling perspective.

3 Problem Formulation

To model the topic-aware social influence, we adopt the *Topic-aware Independent Cascade (TIC) Model* [30], where the user-to-user influence probabilities depend on the topic. Therefore, for each arc $(v, u) \in E$ and each topic $z \in [1, K]$ we are given a probability $p^z_{v,u}$, representing the strength of the influence exerted by user v on user u on topic z. Moreover for each item $i \in \mathcal{I}$ that propagates in the network, we have a distribution over the topics, that is for each topic $z \in [1, K]$ we are given $\gamma^z_i = P(Z = z|i)$, with $\sum_{z=1}^{K} \gamma^z_i = 1$. In this model a propagation happens like in the IC model: when a node v first becomes active on item i, has one chance of influencing each inactive neighbor u, independently of the history thus far. The tentative succeeds with a probability that is the weighted average of the link probability w.r.t. the topic distribution of the item i:

$$p^i_{v,u} = \sum_{z=1}^{K} \gamma^z_i p^z_{v,u}. \tag{1}$$

Under the directed graph $G = (V, E)$, the *influence spread* of the initially infected set S, which is the ultimately expected number of infected nodes, is denoted as $\sigma(S|V)$.

Now we present a mathematical definition for the *influence minimization problem*. Assume the negative information spreads in the network $G = (V, E)$ with initially infected nodes $S \subseteq V$, our goal here is to minimize the number of ultimately infected nodes by blocking k nodes (or vertices) of set $D \in V$, where k ($\ll |V|$) is a given const. It can be formulated as the following optimization problem:

$$D^* = \arg \min_{D \subseteq V, |D| \leq k} \sigma\big(S\big|V \backslash D\big) \tag{2}$$

where $\sigma(S|V\backslash D)$ denotes the influence (number of ultimately infected nodes) of S when the node set D is blocked.

4 Topic Model Analysis

Before we solve this problem above, we should introduce the Latent Dirichlet Allocation based on Hierarchical Dirichlet Process (HDP-LDA) method first.

In the first step, we adopt the hierarchical Dirichlet processes to learn the topic distribution $\theta_{e_{u,v}}$ for each link $e_{u,v}$. HDP-LDA is non-parametric topic model which can automatically determine the proper number of topic K based on

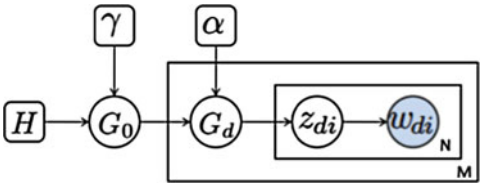

Fig. 1. Graphical Model for HDP. γ,α_0 and H are hyper parameters. G_d denotes random measure at the document level while G_0 at the corpus level. z_{di} denotes the topic of word w_{di} while w_{di} denotes the ith word in document d.

the data at hand. It has been proved that HDP outperforms other unsupervised topic models, e.g. LDA [31] and LSI [32], on modeling large scale web texts.

This step contains three sub-steps. First, we collect all the messages on the links, which forms a document set $D = \{d_{e_{u,v},i} | e_{u,v} \in E, i = 1, \cdots, N_{e_{u,v}}\}$, where $N_{e_{u,v}}$ is the number of messages on link $e_{u,v}$. Second, we adopt HDP-LDA to learn the number of topic K and the topic distribution $\theta_{e_{u,v},i}$ for each message. Third, the topic distribution for link $\theta_{e_{u,v}}$ is calculated by averaging the topic distribution $\theta_{e_{u,v},i}$ and the topic distribution of the target information $\theta_{d'}$ is predicted (Fig. 1).

4.1 Model Description

HDP defines a set of random measures G_d, one for each document, and a global random measure G_0. G_d models the topic distributions at the document level while G_0 at the corpus level. Each word w_{di} is associated with a topic z_{di} sampled from G_d. To share the topics across documents, the document-specific random measures G_d are drawn from the global measure with Dirichlet process $DP(\alpha, G_0)$, where α is a concentration factor. The global measure G_0 is also sampled from a corpus-level DP with a concentration parameter γ and a base probability measure H. In summary, we define the generative process of HDP as follows.

$$\begin{aligned} G_0|\gamma, H \sim DP(\gamma, H), \quad G_d|\alpha, G_0 \sim DP(\alpha, G_0) \\ z_{di}|G_d \sim G_d, \quad w_{di}|z_{di} \sim F(z_{di}) \end{aligned} \tag{3}$$

HDP can be constructed with the Chinese Restaurant Franchise processes (CRF). In the metaphor of CRF, a restaurant franchise corresponds to a corpus, and each restaurant corresponds to a document. A global menu of dishes in the restaurant corresponds to a topic $\phi_1, \cdots, \phi_K$. A customer corresponds to a word in a document. And the process of a customer picking a table corresponds to generating a word with a topic. In particular, we need to maintain the counts of customers and tables. Here, n_{dbk} denotes the number of customers in the restaurant d at table b eating dish k and m_{dk} denotes the number of tables in the restaurant d serving dish k. In this paper, marginal counts are represented with dots. Thus, $n_{db\cdot}$ represents the number of customers in the restaurant d

at table b, and so on. In metaphor of CRF, for a word w_{id}, the conditional distribution for the word's topic selection z_{di} given $z_{d1}, \cdots, z_{d,i-1}$ and G_0 as in Eq. (4), where G_d is integrated out.

$$z_{di}|z_{d,1:i-1}, \alpha, G_0 \sim \sum_{b=1}^{m_{d\cdot}} \frac{n_{db\cdot}}{i-1+\alpha}\delta_{\psi_{db}} + \frac{\alpha}{i-1+\alpha}G_0 \tag{4}$$

And the conditional distribution of $\psi_{db^{new}}$ is given in Eq. (5).

$$\psi_{db^{new}}|\psi_{1:d-1,\cdot}, \psi_{d,1:m_{d\cdot}-1}, \gamma, H \sim \sum_{k\in K} \frac{m_{\cdot k}}{m_{\cdot\cdot}+\gamma}\delta_{\phi_k} + \frac{\gamma}{m_{\cdot\cdot}+\gamma}H \tag{5}$$

Equations (3), (4) and (5) together describe the CRF construction of HDP.

4.2 Model Inference

We adopted the Gibbs sampling algorithm to infer the latent state of HDP. In Gibbs sampling scheme [33], the state of one variable is sampled with all the other states fixed. We sample the latent variables in sequence until convergence. In HDP, the latent variables of interests are the corpus-level topic distribution β, the topic for each word z_{di}, and the number of tables for each topic in document m_{kj}.

- **Sampling G_0.** Given CRF construction of HDP, the corpus-level topic distribution G_0 can be instantiated as $G_0 = \sum_k \beta_k \delta_{\phi_k} + \beta_u H$. And it is distributed as in Eq. (6):

$$\beta = (\beta_1, ..., \beta_K, \beta_u)|m_{\cdot,P\cdot}, \gamma \sim Dir(m_{\cdot 1}, ..., m_{\cdot K}, \gamma) \tag{6}$$

- **Sampling z_{ji}.** Given CRF construction of HDP, It can be realized by grouping together terms associated with each k.

$$p(z_{ji} = k|z^{-ji}, m, \beta) = \begin{cases} (n_{j\cdot k}^{-ji} + \alpha_0\beta_k) f_k^{-x_{ji}}(x_{ji}, w_{ji}) & \text{for existing } k\ , \\ \alpha_0\beta_u f_{k^{new}}^{-x_{ji}}(x_{ji}) & \text{for new topic } k = k^{new}. \end{cases} \tag{7}$$

- **Sampling m.** Given the CRF construction of HDP, the number of tables is determined by the scaling factors as well as the number of words in the documents. Antoniak(1974) [34] has shown that m_{jk} is distributed as in Eq. (8):

$$p(m_{jk} = m|z, m^{-jk} = k, \beta) = \frac{\Gamma(\alpha_0\beta_k)}{\Gamma(\alpha_0\beta_k + n_{j\cdot k})} s(n_{j\cdot k}, m)\alpha_0{\beta_k}^m \tag{8}$$

where $s(n, m)$ are unsigned Stirling number of the first kind.

Given the samples, the posterior of topic distribution of message j can be calculated as in Eq. (9)

$$\theta_j = (\theta_{j1}, \theta_{j2}, \cdots, \theta_{jK}) \sim Dir(n_{j\cdot 1} + \alpha_0\beta_1, n_{j\cdot 2} + \alpha_0\beta_2, \cdots, n_{j\cdot K} + \alpha_0\beta_K) \tag{9}$$

And the distributions of the link can be computed by averaging the distribution of messages on that link:

$$\theta_{e_{u,v}} = \frac{\sum_{i\in d_{e_{u,v}}} \theta_i}{N_{e_{u,v}}} \tag{10}$$

4.3 Prediction

We have trained the model on a fully observed data of social network $G = (V, E)$ at hand, and get the word distribution for each topic denoted as ϕ_k, where $k = 1, 2, ..., K$ and K is the number of topics. We will use the ϕ_k to predict the topic distribution $\theta_{d'}$ for the new message d' with EM algorithm. In the E-step, given fixed ϕ_k and random topic distribution $\theta_{d'}$, we can compute the topic of every word z_{ji}. And in the M-step, we will compute the new $\theta_{d'}$ with the result from E step. The E-step and M-step is conducted iteratively until convergence.

5 Analysis and Solution for Influence Minimization

The problem of learning the parameters of the TIC models takes in input the social graph $G = (V, E)$, a log of past propagations $\mathbb{D}$, and an integer K, which can be learnt by the Latent Dirichlet Allocation based on Hierarchical Dirichlet Process (HDP-LDA) method. The propagation log is a relation $(\texttt{User}, \texttt{Item}, \texttt{Time})$ where a tuple $(u, i, t) \in \mathbb{D}$ indicates that user u adopted item i at time t. The output of the learning problem is the set of all parameters of the TIC propagation model, which we denote Θ: these are γ_i^z and $p_{v,u}^z$ for all $i \in \mathcal{I}$, $(v, u) \in E$, and $z \in [1, K]$. Assuming that each propagation trace is independent from the others, the likelihood of the data given the model parameters Θ, can be expressed as: $\mathcal{L}(\Theta; D) = \sum_{i \in \mathcal{I}} \log \mathcal{L}(\Theta; D_i)$. We then adopt the standard EM inference of parameters Θ for TIC. We calculate the topic distributions of each uninfected node w and negative information i via HDP-LDA, then calculate the KL divergences $d(w, i)$ between node w and information i from the topic perspective.

Now we are back to the optimal problem (2), any straightforward method for exact solution suffers from combinatorial explosion for a large network. Therefore, we consider approximately solving the problem, while a natural idea is to block the nodes in the neighborhood of infected set. Specifically, given the initially infected set S and the negative information i, define the neighborhood set $N(S)$ like

$$N(S) := \big\{v \in V \backslash S : \ \exists u \in S, \ s.\ t.\ (u, v) \in E\big\}.$$

We want to block k susceptible nodes in the set $N(S)$ to minimize the negative influence. Since the set $N(S)$ is usually very large (i.e. $|N(S)| \gg k$), a natural question arises, *how to select k susceptible nodes from the set $N(S)$ to block in order to make the ultimate influence as small as possible*? In this paper, given the negative information $i \in \mathcal{I}$, we introduce two scoring methods for the nodes in $N(S)$, and then select k nodes with the highest scores as the objectives to block.

Topic-aware Betweenness Scoring Method. Given the initially infected nodes S, the betweenness score $b(w)$ of a node $w \in N(S)$ is defined as follows:

$$b(w) = \sum_{u \in S, v \in V \backslash S} \frac{n(w; u, v)}{N(u, v)} \tag{11}$$

where $N(u, v)$ denotes the number of the shortest paths from node u to node v in G, and $n(w; u, v)$ denotes the number of those paths that pass w. Here we set $n(w; u, v)/N(u, v) = 0$ if $N(u, v) = 0$. We defined the topic-aware betweenness as

$$tb(w) = \frac{b(w)}{d(w, i)}. \tag{12}$$

Topic-aware Out-degree Scoring Method. Previous work has shown that simply removing nodes in order of decreasing out-degrees works well for preventing the spread of contamination in most real networks [5]. Here we focus on the contaminated set S and the corresponding $i \in \mathcal{I}$. We define the out-degree score $o(w)$ of node $w \in N(S)$ as the number of non-contaminative nodes around w. We defined the topic-aware out-degree as

$$to(w) = \frac{o(w)}{d(w, i)}. \tag{13}$$

Equations (12) and (13) are reasonable, since we can find that the smaller $d(w, i)$ is, the more susceptible the node w is; and the bigger $b(w)$ or $o(w)$ is, the more pivotal the node w is. Hence blocking the nodes with the highest topic-aware betweenness and outdegree score should be effective for preventing the spread of contamination in the network.

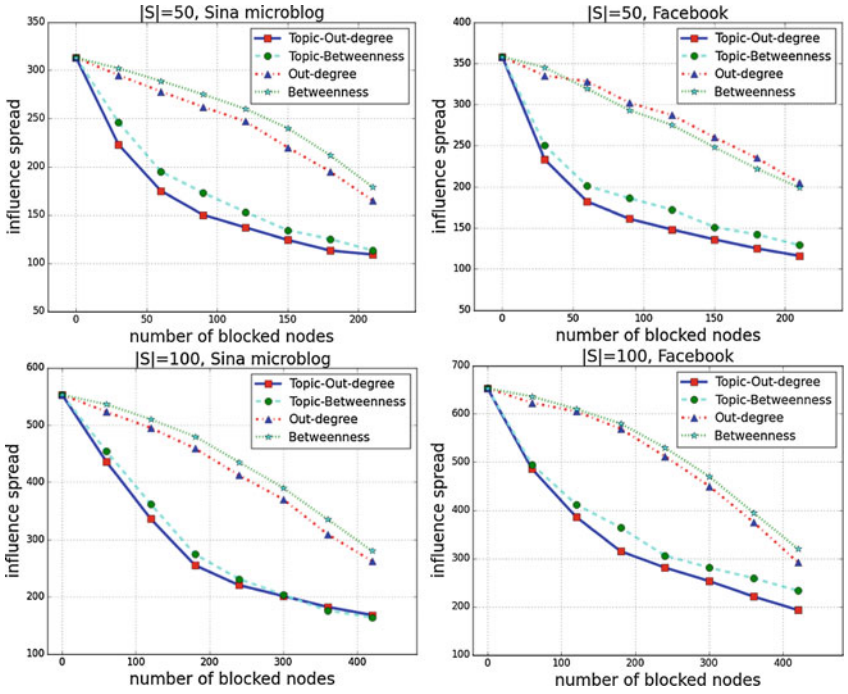

Fig. 2. Experiment result on two data sets.

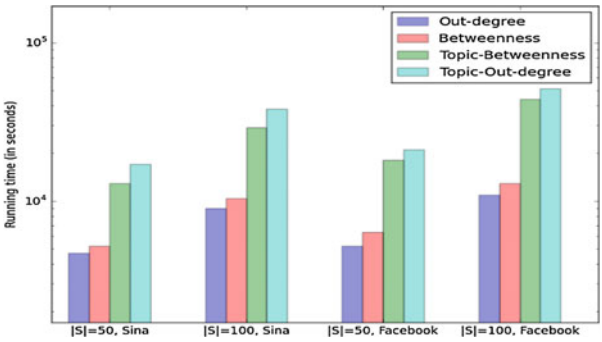

Fig. 3. The time comparison among the four methods.

6 Experiment Results

We experimentally evaluate the performance of our proposed approaches on two networks. One is crawled from Sina microblog containing 2,000 nodes, 14,426 edges and the propagation log. The other is Facebook data acquired from Stanford Network Analysis Project containing 4,039 nodes and 88,234 edges, where the topic probability for each user is created by the HDP-LDA model. We use the Gibbs sampling method to estimate the hyper parameters γ, α_0 and H in HDP-LDA. We employ the Monte-Carlo simulation of TIC model to estimate the influence spread.

From the results in Fig. 2, we can observe that the ultimate influence spreads by Topic-aware heuristics are significantly reduced compared to that by Out-degree and Betweenness centralities, especially in the early stage. For the infected set S with $|S| = 50$ on Sina microblog, we can observe that the proposed method can reduce the negative spread from 320 to 180 by blocking 60 nodes. Here the blocked 60 nodes only accounts to 15 % of the nodes that are connected to infected nodes. From the running results in Fig. 3, we can draw a conclusion that, although the performance is improved greatly, the time cost of topic-aware heuristics are still in the same magnitude with centrality measures.

7 Conclusion

In this paper we investigate the problem of minimizing the spread of negative things by blocking nodes in social networks from a topic modeling perspective. We use the HDP-LDA and KL divergence to analysis the influence and relevance, then two topic-aware heuristics based on betweenness and out-degree for finding approximate solutions are proposed. Using two real networks Sina Microblog and Facebook, we demonstrate experimentally the high performance of the proposed algorithms.

There are several interesting future directions. First, how to extend it to a dynamic network when the network structure changes over time is an interesting

question [35]. Second, how to minimize the negative influence with the real cascade data is also a practical problem.

Acknowledgements. This work was supported by the 973 project (No. 2013CB 329606), and the Strategic Leading Science and Technology Projects of Chinese Academy of Sciences (No. XDA06030200), Australia ARC Discovery Project (DP1402206).

References

1. Domingos, P., Richardson, M.: Mining the network value of customers. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 57–66. ACM (2001)
2. Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 199–208. ACM (2009)
3. Budak, C., Agrawal, D., El Abbadi, A.: Limiting the spread of misinformation in social networks. In: Proceedings of the WWW 2011, pp. 665–674. ACM (2011)
4. Kimura, M., Saito, K., Motoda, H.: Minimizing the spread of contamination by blocking links in a network. In: AAAI, vol. 8, pp. 1175–1180 (2008)
5. Wang, S., Zhao, X., Chen, Y., Li, Z., Zhang, K., Xia, J.: Negative influence minimizing by blocking nodes in social networks. In: AAAI (Late-Breaking Developments) (2013)
6. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 137–146. ACM (2003)
7. Chen, W., Wang, C., Wang, Y.: Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1029–1038. ACM (2010)
8. Chen, W., Yuan, Y., Zhang, L.: Scalable influence maximization in social networks under the linear threshold model. In: ICDM 2010 (2010)
9. Goyal, A., Lu, W., Lakshmanan, L.V.: Simpath: an efficient algorithm for influence maximization under the linear threshold model. In: IEEE 11th International Conference on Data Mining (ICDM), pp. 211–220. IEEE (2011)
10. Kimura, M., Saito, K.: Tractable models for information diffusion in social networks. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) PKDD 2006. LNCS (LNAI), vol. 4213, pp. 259–271. Springer, Heidelberg (2006)
11. Narayanam, R., Narahari, Y.: A shapley value-based approach to discover influential nodes in social networks. IEEE Trans. Autom. Sci. Eng. **99**, 1–18 (2010)
12. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: KDD 2007 (2007)
13. Goyal, A., Lu, W., Lakshmanan, L.V.: Celf++: optimizing the greedy algorithm for influence maximization in social networks. In: WWW 2011 (2011)
14. Zhou, C., Zhang, P., Guo, J., Zhu, X., Guo, L.: Ublf: an upper bound based approach to discover influential nodes in social networks. In: IEEE 13th International Conference on Data Mining (ICDM), pp. 907–916. IEEE (2013)
15. Zhou, C., Zhang, P., Guo, J., Guo, L.: An upper bound based greedy algorithm for mining top-k influential nodes in social networks. In: 23rd International World Wide Web Conference (WWW), pp. 421–422. ACM (2014)

16. Zhou, C., Zhang, P., Zang, W., Guo, L.: On the upper bounds of spread for greedy algorithms in social network influence maximization. IEEE Trans. Knowl. Data Eng
17. Guo, J., Zhang, P., Zhou, C., Cao, Y., Guo, L.: Item-based top-k influential user discovery in social networks. In: IEEE 13th International Conference on Data Mining Workshops (ICDMW), pp. 780–787. IEEE (2013)
18. Rodriguez, M.G., Schölkopf, B.: Influence maximization in continuous time diffusion networks, arXiv preprint arXiv:1205.1682
19. Goyal, A., Bonchi, F., Lakshmanan, L.V.: A data-based approach to social influence maximization. Proc. VLDB Endowment **5**(1), 73–84 (2011)
20. Zhou, C., Zhang, P., Zang, W., Guo, L.: Maximizing the long-term integral influence in social networks under the voter model. In: 23rd International World Wide Web Conference (WWW), pp. 423–424. ACM (2014)
21. Zhou, C., Zhang, P., Zang, W., Guo, L.: Maximizing the cumulative influence through a social network when repeat activation exists. In: ICCS 2014 (2014)
22. Zhou, C., Guo, L.: A note on influence maximization in social networks from local to global and beyond. Procedia Comput. Sci. **30**, 81–87 (2014)
23. Zang, W., Zhang, P., Zhou, C., Guo, L.: Discovering multiple diffusion source nodes in social networks. Procedia Comput. Sci. **29**, 443–452 (2014)
24. Zang, W., Wang, P., Zhou, C., Guo, L.: Topic-aware source locating in social networks. In: 24th International World Wide Web Conference. ACM (2015)
25. Yao, Q., Zhou, C., Xiang, L., Cao, Y., Guo, L.: Minimizing the negative influence by blocking links in social networks. In: 2014 International Standard Conference on Trustworthy Computing and Services (2014)
26. Yao, Q., Zhou, C., Shi, R., Wang, P., Guo, L.: Topic-aware social influence minimization. In: 24th International World Wide Web Conference. ACM (2015)
27. Albert, R., Jeong, H., Barabási, A.-L.: Error and attack tolerance of complex networks. Nature **406**(6794), 378–382 (2000)
28. Newman, M.E., Forrest, S., Balthrop, J.: Email networks and the spread of computer viruses. Phys. Rev. E **66**(3), 035101 (2002)
29. Habiba, Yu, Y., Berger-Wolf, T.Y., Saia, J.: Finding spread blockers in dynamic networks. In: Giles, L., Smith, M., Yen, J., Zhang, H. (eds.) SNAKDD 2008. LNCS, vol. 5498, pp. 55–76. Springer, Heidelberg (2010)
30. Barbieri, N., Bonchi, F., Manco, G.: Topic-aware social influence propagation models. In: Proceedings of the ICDM 2012, pp. 81–90. IEEE Computer Society (2012)
31. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
32. Dumais, S.T.: Latent semantic analysis. Ann. Rev. Inf. Sci. Technol. **38**(1), 188–230 (2004)
33. Casella, G., George, E.I.: Explaining the gibbs sampler. Am. Stat. **46**(3), 167–174 (1992)
34. Antoniak, C.E.: Mixtures of dirichlet processes with applications to bayesian nonparametric problems. Ann. Stat. **2**, 1152–1174 (1974)
35. Zhang, P., Zhou, C., Wang, P., Gao, B.J., Zhu, X., Guo, L.: E-tree: an efficient indexing structure for ensemble models on data streams. IEEE Trans. Knowl. Data Eng. **27**(2), 461–474 (2015)

http://www.springer.com/978-3-319-24473-0

Data Science
Second International Conference, ICDS 2015, Sydney, Australia, August 8-9, 2015, Proceedings
Zhang, C.; Huang, W.; Shi, Y.; Yu, P.S.; Zhu, Y.; Tian, Y.; Zhang, P.; He, J. (Eds.)
2015, X, 194 p. 60 illus. in color., Softcover
ISBN: 978-3-319-24473-0