Identifying Structural Hole Spanners in Online Social Networks Using Machine Learning

Qingyuan Gong, Jiayun Zhang, Xin Wang, Yang Chen
School of Computer Science, Fudan University, China
{gongqingyuan,jiayunzhang15,xinw,chenyang}@fudan.edn.cn

ABSTRACT

Online social networks play an important role in our daily activities. As an important concept in social network analytics, the structural hole theory shows that the positions in social networks that can bridge different user groups will get benefits. Existing solutions for identifying structural hole spanners normally require the knowledge of the entire social graph. In this paper, we propose a novel solution to uncover structural hole spanners according to the users' profiles and user-generated contents (UGCs), instead of referring to the entire social graph. We propose a machine learning-based model to implement the identification. We further leverage the ego networks and the cross-site linking function to enhance the identification. A real-world dataset collected from Foursquare and Twitter is used to evaluate the identification performance of our model. The results show that our model can achieve a high F1-score of 0.857.

CCS CONCEPTS

• **Human-centered computing → Social networking sites**.

KEYWORDS

Online Social Networks, Structural Hole Spanner Detection, Machine Learning, Cross-Site Linking, Ego Networks

ACM Reference Format:
Qingyuan Gong, Jiayun Zhang, Xin Wang, Yang Chen. 2019. Identifying Structural Hole Spanners in Online Social Networks Using Machine Learning. In *SIGCOMM '19: ACM SIGCOMM 2019 Conference (SIGCOMM Posters and Demos '19), August 19–23, 2019, Beijing, China.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3342280.3342319

1 INTRODUCTION

Attracting billions of users, online social networks (OSNs), such as Facebook and Twitter, have become a leading Internet service around the world. Structure hole (SH) is an important concept in social network analytics [3]. The structural hole theory shows that users will get advantages from filling the "holes" between different users or user groups that are otherwise disconnected. As shown in Fig. 1, the user in the middle, known as an SH spanner, plays a critical role in the information dissemination. The structural hole theory has been used in different social network-related applications, such as requirements identification in open-source software development [2], specific user group identification [13], online self-disclosure analytics [9], and the evolution of new business development performance [4].

Normally, the social graph is needed for uncovering SH spanners, and several social graph-based algorithms such as HIS [10], MaxD [10] and HAM [8] are proposed. Unfortunately, identifying SH spanners in OSNs is not easy. First, in some OSNs such as Facebook, users are allowed to hide their social connections, making third-party social application providers difficult to obtain the entire social graph. Second, for a newly registered OSN user, her social connections are still under development. Applying graph-based algorithms directly might fail to discover some potential SH spanners. Last but not least, due to the large sizes of OSNs, some existing algorithms, such as HAM, suffer from the scalability problem when applying to a large social graph [12].

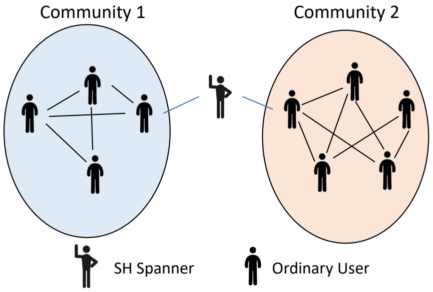

Figure 1: Structural Hole Spanners (SH Spanners) in Social Networks

In this paper, we formulate and explore the problem of identifying SH spanners in OSNs, without referring to the entire social graph. Our model leverages a machine learning-based framework to make use of a user's profile and UGCs. The design of our model is able to get rid of the disadvantages of the graph-based SH spanner detection algorithms. It avoids the requirement of using the entire social graph which might be inaccessible, and is still useful for the newly registered users. To further improve the identification performance, our model refers to a user's ego network [1] according to the structural hole theory. In addition, our model adopts the cross-site linking function [7], to make use of a user's information on another OSN. Based on the real-world user data collected from Foursquare and Twitter, we evaluate the identification performance of our model. Evaluation shows that our model can achieve an F1-score of 0.857.

SIGCOMM Posters and Demos '19, August 19–23, 2019, Beijing, China
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6886-5/19/08…$15.00
https://doi.org/10.1145/3342280.3342319

2 DESIGN AND IMPLEMENTATION

Fig. 2 shows the overall workflow of our model. Our machine learning-based model takes a user's profile and UGCs as the input, and output the identification result that whether she is an SH spanner or an ordinary user. We use the subsets of *descriptive features*, *ego network features*, and *cross-site features* to characterize each user. A supervised machine learning-based classifier is used to implement the identification.

The *descriptive features* consist of the user's demographic information and the statistics of her UGCs in the OSN. Note that our model can work without the cross-site features and ego network features, but including them will achieve a better identification performance.

When the users' friend lists are available, the additional analysis on the user's ego network can enhance the identification performance. Comparing with collecting the entire social graph, crawling an individual user's ego network[1] is much easier. In particular, we adopt four classic metrics, i.e., effective size, efficiency, constraint and hierarchy. These metrics characterize the ego network of one user from different aspects according to the structural hole theory [3], and we denote them as *ego network features*.

We make use of the cross-site linking function, which is widely supported by OSNs [7], to get more information of a user. We not only exploit a user's information from the targeted OSN, but also her generated data on an external OSN linked by her. The user information on the external OSN are extracted as *cross-site features*. In our data-driven evaluations by only referring to her information on an external OSN, these features are shown to be useful for the prediction of whether a new user will become an SH spanner.

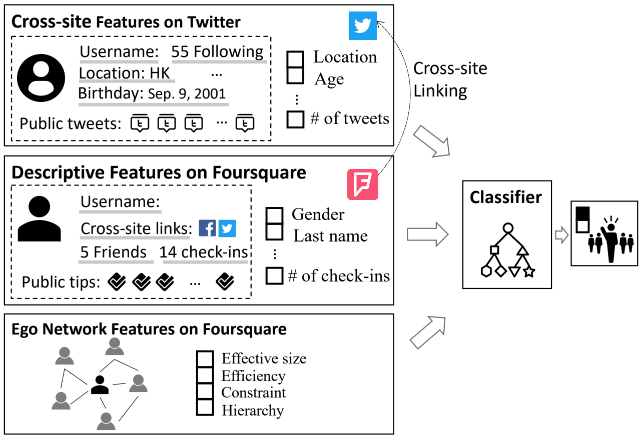

Figure 2: The SH Spanner Identification Model

3 PERFORMANCE EVALUATION

In our study, we use Foursquare, a representative location-based social network, as the targeted OSN to study the identification of SH spanners. We used Breadth First Search (BFS) to crawl a subset of 9,299,408 Foursquare users during Apr. 28, 2019 - May 20, 2019. In our dataset, 35.31% of Foursquare users have linked their Twitter accounts, and we focus on these users. We further crawl these users' profiles and tweets on Twitter. To obtain the ground-truth of SH spanners and ordinary users, we use the HIS algorithm [10] to find SH spanners, and 2,373 users are selected. Accordingly, we randomly select 2,373 users from the rest of users to represent the ordinary users. The proportion between the numbers of users in the training and validation subset and the test subset is 1:1. To evaluate the identification performance, we use a set of classic metrics, i.e., precision, recall, F1-score and AUC [5]. Precision denotes the fraction of identified SH spanners who are really SH spanners. Recall is the fraction of SH spanners who are correctly identified. F1-score is the harmonic mean of precision and recall. AUC (area under the ROC curve) means the probability that this model would rank a randomly selected SH spanner higher than a randomly chosen ordinary user. The classifier is implemented by CatBoost [11], a gradient boosting library. The parameters in the classifier are determined through a parameter tuning phase, using the user instances in the training and validation subset. We apply a grid search to sweep the parameters space. Once a set of parameters is given, we can get an F1-score using 5-fold cross-validation. We choose the set of the parameters that could help the classifier achieve the highest F1-score. After that, the classifier is able to judge whether a user in the test subset is an SH spanner.

According to the evaluation on the test subset, our model can achieve an F1-score of 0.857 and an AUC value of 0.856. We could also use our model to predict whether a newly registered user will become a future SH spanner. By excluding the users' descriptive features, as well as the ego network features on the targeted OSN (Foursquare), we only consider the cross-site features from the external OSN (Twitter). The resulted F1-score is 0.775 and the AUC value is 0.786, signifying that the model can work well in "cold start" scenarios [6]. In addition, we explore the identification performance for the users who have not linked their profiles to external OSNs. Under such a situation, we only take the descriptive features and ego network features into consideration. Both the resulted F1-score and AUC value equal to 0.787. The above results validate that our model can achieve a good performance to identify the SH spanners, without analyzing the entire social graph.

4 CONCLUSION

In this paper, we study the problem of identifying SH spanners in OSNs without referring to the entire social graph. Our solution introduces a new angle of uncovering SH spanners, and reveals the relationship between SH spanners and the users' profiles and UGCs. The cross-site linking function and the ego networks can be further leveraged to enhance the identification performance. Our solution is very helpful for third-party social application providers to find SH spanners accurately.

5 ACKNOWLEDGEMENT

The authors are grateful to Chenfan Zhuang for her help in data crawling. This work is sponsored by National Natural Science Foundation of China (No. 61602122, No. 71731004). Yang Chen is the corresponding author.

[1] A user's ego network is the sub-network formed by the user (ego) and all her connected neighbors (alters) in a selected OSN.

REFERENCES

[1] Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Robin I. M. Dunbar. 2017. Online Social Networks and information diffusion: The role of ego networks. *Online Social Networks and Media* 1 (2017), 44–55.

[2] Tanmay Bhowmik, Nan Niu, Prachi Singhania, and Wentao Wang. 2015. On the Role of Structural Holes in Requirements Identification: An Exploratory Study on Open-Source Software Development. *ACM Transactions on Management Information Systems* 6, 3 (2015), Article 10.

[3] Ronald S. Burt. 2009. *Structural Holes: The Social Structure of Competition*. Harvard University Press.

[4] Eunyoung Choi and Kun Chang Lee. 2016. Relationship between social network structure dynamics and innovation: Micro-level analyses of virtual cross-functional teams in a multinational B2B firm. *Computers in Human Behavior* 65 (2016), 151–162.

[5] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27, 8 (2006), 861–874.

[6] Qingyuan Gong, Yang Chen, Xinlei He, Fei Li, Yu Xiao, Pan Hui, Xin Wang, and Xiaoming Fu. 2018. Identification of Influential Users in Emerging Online Social Networks Using Cross-site Linking. In *Proc. of ChineseCSCW*. Springer, 331–341.

[7] Qingyuan Gong, Yang Chen, Jiyao Hu, Qiang Cao, Pan Hui, and Xin Wang. 2018. Understanding Cross-site Linking in Online Social Networks. *ACM Transactions on the Web* 12, 4 (2018), 25:1–25:29.

[8] Lifang He, Chun-Ta Lu, Jiaqi Ma, Jianping Cao, Linlin Shen, and Philip S. Yu. 2016. Joint Community and Structural Hole Spanner Detectionvia Harmonic Modularity. In *Proc. of ACM KDD*.

[9] Young D. Kwon, Reza Hadi Mogavi, Ehsan Ul Haq, Youngjin Kwon, Xiaojuan Ma, , and Pan Hui. 2019. Effects of Ego Networks and Communities on Self-Disclosure in an Online Social Network. In *Proc. of IEEE/ACM ASONAM*.

[10] Tiancheng Lou and Jie Tang. 2013. Mining Structural Hole Spanners Through Information Diffusion in Social Networks. In *Proc. of WWW*.

[11] Liudmila Ostroumova Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. In *Proc. of NeurIPS*. 6639–6649.

[12] Wenzheng Xu, Mojtaba Rezvani, Weifa Liang, Jeffrey Xu Yu, and Chengfei Liu. 2017. Efficient Algorithms for the Identification of Top-k Structural Hole Spanners in Large Social Networks. *IEEE Transactions on Knowledge and Data Engineering* 29, 5 (2017), 1017–1030.

[13] Qiu Fang Ying, Dah Ming Chiu, and Xiaopeng Zhang. 2018. Diversity of a User's Friend Circle in OSNs and Its Use for Profiling. In *Proc. of International Conference on Social Informatics*. Springer, 471–486.