9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Investigating the Effectiveness of Using Ego-Network Structures for Identifying Influencers on Social Media

social network analysis, influencer, ego network, graph neural networks, machine learning

Extended Abstract

Identifying influential users, which are so-called influencers, has been an important research problem in the social network analysis research community. The most common approach for identifying influencers is using centrality measures of nodes in a social network representing relationships among social media users. In another line of studies, model-based influence maximization algorithms have been also used for identifying influencers. More recently, machine-learning-based approaches for identifying influencers using multiple centrality measures have been proposed.

While most existing methods for identifying influencers require the entire structure of a social network, it is difficult to obtain the entire structure, especially in social media, because the networks are very large, while the access to network data is limited [1, 2]. When the knowledge on a network is severely limited, the methods for identifying influencers have shown to degrade their effectiveness [1, 2]. However, how to effectively identify influencers with limited knowledge on a social network has still been an open issue.

We tackle the problem of predicting influencers only from their ego networks [3]. An ego network of a user is an induced subgraph of a social network consisting of the target user and his/her followers (Fig. 1). By definition, obtaining an ego network of a user is much easier than obtaining the complete structure of a social network. Thus, predicting whether a given user is an influencer or not only from his/her ego network is expected to be beneficial in practice.

Since there exist several definitions of influencers [4], we examine the effectiveness of using ego networks of social media users for two types of influencer identification tasks using three Twitter datasets (Higgs [5], Non-Topic [6], Nepal [7] datasets). The two types of influencers are those based on total influence and those based on indirect influence. An influencer based on total influence is a user whose tweets are seen from many users. In contrast, an influencer based on indirect influence is a user whose tweets are retweeted from many other users who do not directly follow the user. For each dataset, we extracted top-1% users based on total influence and indirect influence as influencers.

We construct models for predicting whether the given user is an influencer or not using a decision-tree-based model of LightGBM [8] and graph neural networks (GNNs) [9]. As features obtained from an ego network of each node, we use in-degree centrality (number of followers), out-degree centrality (number of followees), the number of outgoing links from the ego network, the number of incoming links from outside the ego network, betweenness centrality, closeness centrality, eigenvector centrality, and ego-node flag representing whether the given node is an ego node or not. For each dataset, 75% of users were randomly selected as the training data, and 25% of users were selected as the test data. As heuristic baselines for identifying influencers, we use in-degree centrality (i.e., the number of followers) and collective influence (CI) with $l = 1$.

Tables 1, 2, and 3 show the prediction accuracy of influencer identification in the Higgs, Non-topic, and Nepal datasets, respectively. These results show that the GNN and LightGBM models achieve higher or comparable accuracy than heuristic baselines (i.e., follower and CI). In particular, for the Non-topic dataset, the GNN and LightGBM models achieve much higher F1 scores than the heuristic baselines in the two influencer identification tasks. For the Nepal dataset, these models can identify influencers based on indirect influence more accurately than the baselines. This suggests the potential benefits of using ego-network structures for the influencer identification tasks.

In contrast, the effectiveness of the GNN and LightGBM models is suggested to be dependent on the datasets and tasks. For instance, these models only achieve comparable F1 scores with baselines for the Higgs dataset.

To clarify the factors affecting the effectiveness of using ego-network structures, more efforts will be needed in future research. We should also note that finding the top-1% influencers is a difficult task, and therefore, the accuracy scores of the models are not so high. However, the accuracy of the models should be further improved for practical use.

In summary, this study empirically demonstrates the potential powers of ego-network structures for identifying influencers on social media.

References

[1] Sho Tsugawa and Kazuma Kimura. Identifying influencers from sampled social networks. *Physica A: Statistical Mechanics and its Applications*, 507:294–303, 2018.

[2] Sho Tsugawa and Hiroyuki Ohsaki. Benefits of bias in Crawl-Based network sampling for identifying key node set. *IEEE Access*, 8:75370–75380, 2020.

[3] Linton C Freeman. Centered graphs and the structure of ego networks. *Mathematical Social Sciences*, 3(3):291–304, 1982.

[4] Fabián Riquelme and Pablo González-Cantergiani. Measuring user influence on Twitter: A survey. *Information Processing & Management*, 52(5):949–975, 2016.

[5] M De Domenico, A Lima, P Mougel, and M Musolesi. The anatomy of a scientific rumor. *Scientific Reports*, 3:2980, October 2013.

[6] Sho Tsugawa. Empirical analysis of the relation between community structure and cascading retweet diffusion. In *Proceedings of the 13th International AAAI Conference on Web and Social Media, (ICWSM 19)*, pages 493–504, July 2019.

[7] Ayan Kumar Bhowmick, Martin Gueuning, Jean-Charles Delvenne, Renaud Lambiotte, and Bivas Mitra. Temporal sequence of retweets help to detect influential nodes in social networks. *IEEE Transactions on Computational Social Systems*, 6(3):441–455, June 2019.

[8] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[9] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of the International Conference on Learning Representations (ICLR 17)*, pages 1–14, 2017.

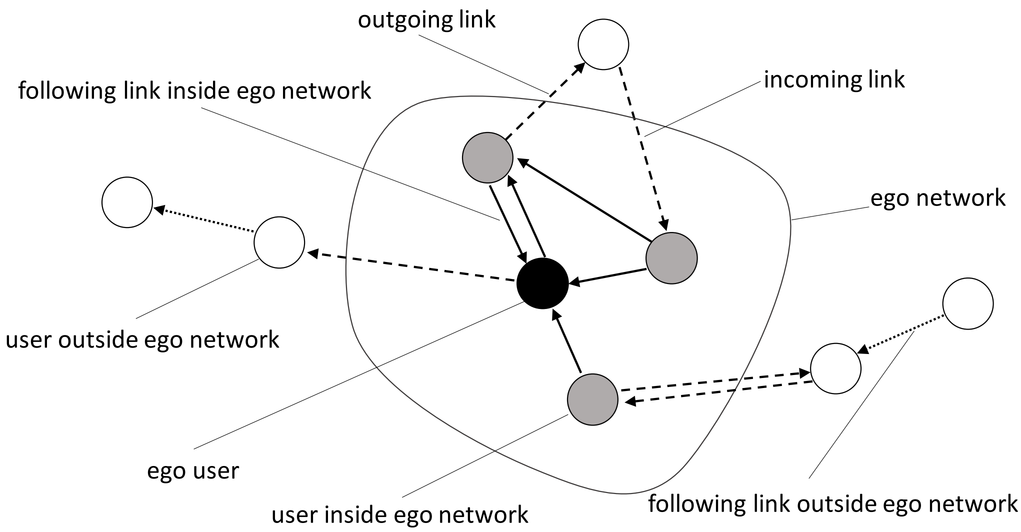

Figure 1: Ego Network. The black circle is the target user (i.e., ego node), and the gray circles are the users who follow the target user. The ego network consists of black and gray circles, and solid links between them.

Table 1: Prediction accuracy for the Higgs dataset

	Total influence			Indirect influence		
	F1	Prec.	Recall	F1	Prec.	Recall
LightGBM	0.6916	0.8592	0.5787	0.3011	0.2963	0.3061
GNN	0.6523	0.7411	0.5825	0.3003	0.3023	0.2983
follower	0.6965	0.8746	0.5787	0.3053	0.2823	0.3323
CI	0.6691	0.7718	0.5905	0.2879	0.0351	0.8822
random	0.0197	0.0101	0.3856	0.0205	0.0104	0.7144

Table 2: Prediction accuracy for the Non-topic dataset

	Total influence			Indirect influence		
	F1	Prec.	Recall	F1	Prec.	Recall
LightGBM	0.1738	0.1130	0.3759	0.2185	0.1524	0.3854
GNN	0.1572	0.1061	0.3031	0.1631	0.1126	0.2957
follower	0.1143	0.0762	0.2287	0.1107	0.0618	0.5283
CI	0.1104	0.0628	0.4545	0.1612	0.0971	0.4737
random	0.0212	0.0109	0.3582	0.0249	0.0159	0.0565

Table 3: Prediction accuracy for the Nepal dataset

	Total influence			Indirect influence		
	F1	Prec.	Recall	F1	Prec.	Recall
LightGBM	0.4586	0.9880	0.2986	0.0880	0.0698	0.1189
GNN	0.4553	0.6922	0.3392	0.0750	0.0529	0.1287
follower	0.4557	1.0000	0.2951	0.0684	0.0538	0.0938
CI	0.4282	0.7804	0.2951	0.0639	0.0452	0.1089
random	0.0201	0.0102	0.6572	0.0228	0.0119	0.2739