Link prediction in blockchain online social networks with textual information

Manuel Dileo[1], Cheick Tidiane Ba[1], Matteo Zignani[1], and Sabrina Gaito[1]

Computer Science Department, Universitá degli Studi di Milano, Milan, IT
{manuel.dileo, cheick.ba, matteo.zignani, sabrina.gaito}@unimi.it

Context: Link prediction in online social networks is one of the most widely used applications; in particular, good performances have been obtained by leveraging structural features only, and considering coarse-grain temporal resolutions. However, we still have a limited understanding of *a*) the impact of content-based similarity on link formation, *b*) whether it can enhance link prediction, and *c*) the importance of the structure compared to node attributes inferred from the produced textual content. Some works, such as [4], have improved prediction performance by fusing structural and textual information taken from networked data. However, these approaches have been tested only on static networks. Moreover, there is limited understanding of which text-based features should be used. In this study, we focus on the former issues by evaluating the impact of content-based similarity on link formation, and by highlighting the role of node attributes inferred from the produced textual content. Specifically, we apply state-of-art graph neural networks on a high-resolution temporal dataset gathered from a growing online social network along with attributes derived from textual content created by the users.

Methodology: "Follow" links and text information have been modeled as an attributed temporal directed graph $\mathscr{G} = (V, E, T, X)$ where V is the set of users, while links $(u, v, t) \in E$ denote a directed "follow" link from user u to user v at time t, and X is a $|V| \times f$ matrix of node attributes, with f the dimension of attribute vectors. Given a time interval $[t_0, t_1]$, the snapshot graph $\mathscr{G}_{[t_0,t_1]}$ represents the directed graph where for each link $e = (u, v, t) \in E$, we have that $t \in [t_0, t_1]$. Given a graph interval snapshot $\mathscr{G}_{[t_0,t_1]}$, the purpose of link prediction is to predict which edges will appear at a successive interval snapshot $\mathscr{G}_{[t_1,t_2]}$. It can be treated as a binary classification task, where we assign label 1 if the link is predicted to form in the following time interval, 0 otherwise [2]. To this end, we use Graph Neural Networks (GNNs) [5] enriched with textual attributes. We use a two-layer GCN to obtain node embeddings, and then a dot product followed by a sigmoid function to perform the link prediction task. We have tested GCN, GAT, and SAGE as graph convolutional operators. Regarding textual information, for each user, we consider two types of features derived from text: *i) text-based statistics* and *ii) user interest*. Text-based statistics are computed on the corpus formed by users' posts, comments, and tags, in the considered time interval. Specifically, we compute the number of posts and comments, the number of tags, the average and standard deviation of the length of the content produced. Whereas for user interest, we rely on topic modeling with Latent Dirichlet Allocation (LDA), as in [3]. Therefore, given an author and a document, we compute a topic vector, that is a probability distribution over a fixed number of topics and it represents how much the user talks about those topics; then, to represent

a user interest, we average all of its topic vectors. For performance evaluation, we rely on the experimental setting for temporal link prediction presented in [2]. We create the training set with links in $\mathscr{G}_{[t_0,t_1]}$, and predict their status in the following time interval $[t_1,t_2]$. Whereas for the test set, we extract links in $\mathscr{G}_{[t_0,t_2]}$ and predict their status in the following interval, $[t_2,t_3]$. The performance evaluation is performed with the area under the receiver operating characteristic curve (AUROC).

Objective and Data: Our goal is to study the impact of textual information compared to the structural's one on link formation in online social networks, using GNNs. To this aim, we rely on Steemit, a blockchain-based online social network [1], that allows the retrieval of high-resolution temporal information for the construction of an attributed temporal network, describing "follow" relationships between users and text content produced by users. We collected data from June 3, 2016, up to January 21, 2021. The starting date is the day the "follow" operation has been made available on Steemit. We have gathered two types of information: *a*) the "follow" relationships, available in the `custom_json` transactions; and *b*) posts, comments and their tags, available in the `comment` transactions. In all, we extracted $369,326$ "follow" operations and $1,083,390$ comment operations. We consider two intervals: period 1 from June 3, 2016, to August 2, 2016 ($[t_0,t_1]$), period 2 from June 3, 2016, to September 2, 2016 ($[t_0,t_2]$); the time interval $[t_2,t_3]$ refers to period from September 3, 2016, to October 2, 2016. The main properties of the resulting graphs are summarized in Table 1. As for textual information, overall, we obtain $327,151$ posts, $756,239$ comments and an average number of tags per document equal to 1.88.

	$\mathscr{G}_{[t_0,t_1]}$	$\mathscr{G}_{[t_0,t_2]}$
Number of Nodes	$7,400$	$20,849$
Number of Edges	$33,920$	$323,228$
Density	0.0006	0.007
Avg Degree	9.17	31.01
Std Degree	25.90	206.43
Largest SCC	$2,313$	$12,505$
New links in the next period	$74,228$	$138,604$

Table 1. Main properties of the Steemit "follow" graph for period 1, from June 3, 2016, to August 2, 2016, ($[t_0,t_1]$) and the following incremental period 2 from June 3, 2016, to September 2, 2016 ($[t_0,t_2]$).

Results: Figure 1 shows the AUROC of GNNs with different combinations of features on the test set. Overall, the results of Graph Neural Network models on link prediction tasks outperform those obtained using traditional supervised models [2]. The use of text-based statistics as node features leads to an increase in performance compared to not using features. However, the performance gain is low so the structure is crucial to understand the network evolution. Note also that the addition of user interest vectors as node features does not enhance the performance; hence, not every addiction of textual features leads to an increase in performance. Another interesting point is the addition of

manually engineered structural features as node attributes. Structural feature augmentation makes the performance worse. This problem arises because the periods considered, as shown in Table 1, see a rapid network evolution; hence, centrality measures are not able to summarize in an effective way the structural information.

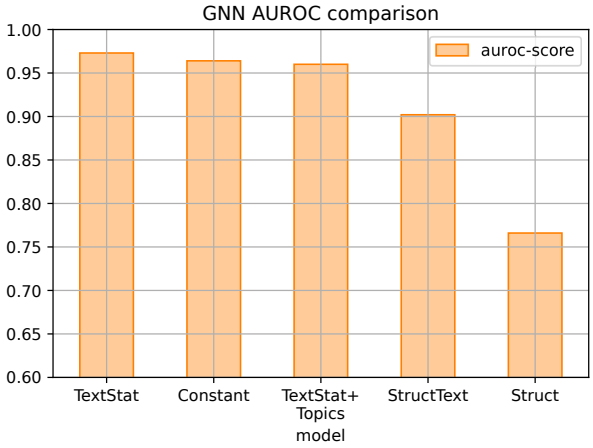

Fig. 1. AUROC score of GNNs using no features (Constant), structural features (Struct), text-based statistics features (TextStat), text-stat and user interest features (TextStat+Topics), structural and text features (StructText), for link prediction on the test set. AUROC scores for TextStat, Constant, and TextStat+Topics are 0.973, 0.964, and 0.962. The use of text features can lead to an increase in AUROC score but the structure of the network is crucial. Structural feature augmentation makes the performance worse.

References

1. Guidi, B.: An overview of blockchain online social media from the technical point of view. Applied Sciences 11(21), 9880 (2021)
2. Liu, Q., Tang, S., Zhang, X., Zhao, X., Zhao, B.Y., Zheng, H.: Network growth and link prediction through an empirical lens. Proceedings of the 2016 Internet Measurement Conference (2016)
3. Parimi, R., Caragea, D.: Predicting friendship links in social networks using a topic modeling approach. In: PAKDD (2011)
4. Wang, Z., Liang, J., Li, R.: Exploiting user-to-user topic inclusion degree for link prediction in social-information networks. Expert Syst. Appl. 108, 143–158 (2018)
5. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P.S.: A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems 32(1), 4–24 (2021)