The impact of textual content on link prediction in online social networks

Manuel Dileo[1], Cheick Tidiane Ba[1][0000−0002−4035−7464], Matteo Zignani[1][0000−0002−4808−4106], and Sabrina Gaito[1][0000−0003−3779−2809]

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

Context: In network science, link prediction is one of the most powerful tools, successfully applied in different settings, such as predicting protein-to-protein interactions or network evolution in online social networks. When it comes down to the latter, current works have successfully leveraged structural features. 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 [1], 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. This work focuses on these aspects by leveraging 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 $\mathcal{G} = (V, E, T, X)$ 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 $\mathcal{G}_{[t_0,t_1]}$ 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 $\mathcal{G}_{[t_0,t_1]}$, the purpose of link prediction is to predict which edges will appear at a successive interval snapshot $\mathcal{G}_{[t_1,t_2]}$. 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) [3] enriched with textual attributes. GNNs can work directly on graph-structured data. Their objective is to obtain an embedding for each node by taking into consideration both the structure of the network and the attributes of the nodes, avoiding manually feature engineering the structural information. 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. Hyperparameter optimization has been conducted, testing *SAGE*, *GAT*, and *GCN* as graph convolutional layers and grids for other classic deep learning hyperparameters such as learning rate or weight decay. Regarding textual information, we compute several text-based statistics on the corpus formed by user's posts, comments, and tags, in the considered time interval. All the documents have been preprocessed through the following operations: removal of HTML tags, punctuation, numbers, stopwords, words shorter than 3 characters, stemming. 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. For performance evaluation, we rely on the experimental setting for temporal link prediction presented in [2]. We create the training set with links in $\mathcal{G}_{[t_0,t_1]}$, and predict their status in the following time interval $[t_1, t_2]$. Whereas for the test set, we extract links in $\mathcal{G}_{[t_0,t_2]}$ 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 [4], 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 average number of tags per document equal to 1.88.

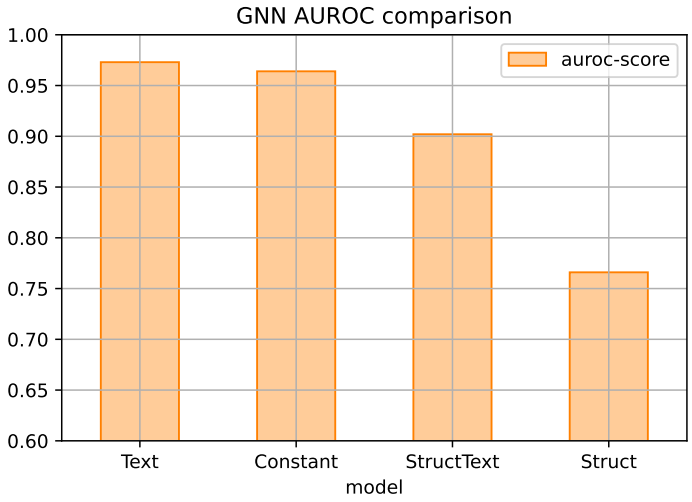

Fig. 1: AUROC score of GNNs using no features (Constant), structural features (Struct), text features (Text), structural and text features (StructText), for link prediction on the test set. AUROC scores for Text and Constant are 0.973 and 0.964. The use of text features leads to an increase in AUROC score but the structure of the net is crucial. Structural feature augmentation makes the performance worse.

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 ones obtained using traditional supervised models [2]. The use of textual 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. The best configuration of hyperparameters has four hidden neurons GCN layer and Adam as optimizer, with learning rate and weight decay respectively equal to 0.025 and $5 \cdot 10^{-5}$. Another interesting point is the addition of manually engineered structural features as node attributes. Specifically, PageRank, in, out, and average neighbor degree have been considered. 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.

	$\mathcal{G}_{[t_0,t_1]}$	$\mathcal{G}_{[t_0,t_2]}$
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 ($[t_0, t_2]$).

References

1. Z. Wang, J. Liang, and R. Li, "Exploiting user-to-user topic inclusion degree for link prediction in social-information networks," *Expert Syst. Appl.*, vol. 108, pp. 143–158, 2018.
2. Q. Liu, S. Tang, X. Zhang, X. Zhao, B. Y. Zhao, and H. Zheng, "Network growth and link prediction through an empirical lens," *Proceedings of the 2016 Internet Measurement Conference*, 2016.
3. Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2021.
4. B. Guidi, "An overview of blockchain online social media from the technical point of view," *Applied Sciences*, vol. 11, no. 21, p. 9880, 2021.