Improving the Reliability of Network Analysis Using Link Prediction

Daniele Bellutta[0000−0002−6131−9846] and Kathleen M. Carley[0000−0002−6356−0238]

Carnegie Mellon University, Pittsburgh, PA 15213, USA
{dbellutt,kathleen.carley}@cs.cmu.edu

Abstract. In an effort to mitigate the common problem of missing information when conducting network analysis on samples of social media data, this study introduces a new link prediction algorithm for dynamic heterogeneous networks and evaluates its ability to improve the reliability of node centrality measures computed on a streamed sample of tweets. Though count-based and propagation-based centrality measures did not exhibit significant changes in reliability after applying link prediction to the streamed sample, path-based measures such as betweenness and closeness centrality were shown to markedly benefit from this strategy.

Keywords: Link prediction · Covert networks · Network analysis

1 Introduction

Recent concerns over the rise of harmful messaging on social media have spurred research to detect and analyze such content [13]. However, these efforts often depend upon online platforms making their data available, often forcing researchers to rely on samples representing 1% or less of all activity on a platform [15]. Analyses conducted on these samples therefore frequently face the problem of missing data, which can lead analysts to draw incorrect conclusions [2, 5].

This study explores the notion that *link prediction*, the process of identifying unseen or unformed connections in a network [16], may be an effective strategy for making subsequent analysis more reliable. First, this work introduces and validates a new link prediction algorithm for *dynamic heterogeneous* networks, which both vary over time and contain multiple types of nodes and links. Second, this study assesses whether adding algorithmically predicted connections to a streamed Twitter sample improves the reliability of centrality measures computed on that sample when compared to more complete user timelines. Though most measures were not significantly impacted, link prediction did improve the identification of highly central users with betweenness or closeness centrality.

2 Related Work

Though some past findings point to certain centrality measures being relatively robust to the sampling of a network [3], others have instead concluded that this

holds only when the amount of error in a network is small [2]. Others have also observed that topology impacts the reliability of network centralities, with scale-free and small-world networks exhibiting particular vulnerability to node removal and node addition errors, respectively [5]. Given that missing information can therefore lead centrality metrics to give inaccurate results, it is unfortunate that link prediction has not been extensively studied for improving their reliability. Past findings have indeed shown link prediction to improve the quality of protein interaction networks [6], so this is an idea worth exploring.

Furthermore, few studies in the otherwise large body of link prediction research [8] have sought to predict links in the dynamic heterogeneous networks available from social media data. An efficient iterative algorithm exists for streamed networks [1], but it can get stuck in suboptimal configurations over the long term. Another method instead uses several node similarity metrics as input features for a neural network [12], but its performance depends on the choice of features. A more recent algorithm learns node and link embeddings from dynamic multirelational networks [17] but does not consider multiple kinds of nodes. Though a few other methods also exist [9, 18], their high computational complexity can make them difficult to apply to large data sets. Notably, all these methods were only intended to predict future links, not to fill in missing information.

3 Methods

To evaluate whether link prediction can make network analysis more reliable, this research used two sets of pro- and anti-France tweets posted in response to the 2020 Nice stabbings. First, a sample of more than 600,000 tweets collected using Twitter's streaming API was filtered for pro- and anti-France messages using hundreds of manually labeled hashtags. This constituted the "observed" data on these discussions. Next, the academic API was used to collect the timelines of the users who had authored these pro- and anti-France tweets. Though not all their tweets were still available, this second set provided a more complete record of these users' messages. Following this process, the 1,000 users who tweeted most frequently were isolated for further analysis. The weighted user communication network (comprising retweets, quotes, replies, and mentions) was then extracted from each data set, yielding extremely sparse networks of density 0.0022 for the observed sample and 0.0030 for the timelines. Additionally, the shared-URL network, which connected users who tweeted the same URL, was extracted from the observed tweets. This produced a network of density 0.0038 that could provide multidimensional link prediction algorithms with more information.

To exploit the richness of social media data, a new method was developed for predicting links in dynamic heterogeneous networks. Inspired by other tensor-based algorithms [4], the proposed Tensor Perturbation Method (TPM) learns the fundamental structure of a network that has been encoded as a multidimensional tensor, such as by stacking the adjacencies for multiple layers or time periods. By extending the matrix-based NMFP algorithm developed by Wang et al. [16] for use on tensors, TPM attempts to fill in missing entries by repeat-

edly perturbing the input tensor and averaging together those reconstructions. In each perturbation, noise is added to the network by randomly removing connections, and the modified tensor is then compressed via low-rank non-negative Tucker decomposition [7] to remove that noise and learn the tensor's structure. After repeating these steps a certain number of times, the decompressed reconstructions are averaged together to attain an approximation of the true network.

Like in NMFP [16], the Colibri-S algorithm [14] is used to select the best multilinear rank for the compression step. However, since TPM operates on tensors, the input must first be unfolded in each dimension before using Colibri-S to determine a suitable rank for that dimension. Since the noise added to the input data should not impact the optimal number of latent factors for representing the unseen true network, this process is only executed once on the raw input tensor.

Before attempting to improve the reliability of network centrality metrics, TPM's performance was compared to that of seven other link predictors, beginning with classics like common neighbors (CN), Adamic-Adar (AA) and resource allocation (RA) [10]. Two methods based on matrix decomposition, NMFP [16] and SPM [10], and two multidimensional methods, CPTD[4] and MLRW [11], were also tested. Each algorithm was executed on the observed communication network, and the predicted links were compared to the additional links present in the timeline communication network. Specifically, this performance was quantified using three metrics common to many link prediction studies: average precision [8], top precision (precision-at-k with k equal to the number of additional links in the timeline network) [16], and the area under the receiver operating characteristic curve (AUROC) [16]. Algorithm parameters were optimized using five-fold cross-validation by trying various parameter configurations on random samples of 85% of the observed communication network and evaluating the predictions against the remaining 15%. Lastly, the three multidimensional methods (including TPM) were also given the observed shared-URL network as input.

Once TPM's effectiveness had been established, its predictions were evaluated for improving the reliability of seven centrality measures: authority, betweenness, closeness, degree, eigenvector, hub, and Katz centrality. Each user's centrality was calculated in three networks: (1) the observed communication network, (2) the timeline communication network, and (3) the augmented communication network, which consisted of the observed network plus the top-scoring links predicted by TPM. For the augmented network, the number of top-scoring links added was 10% of the number of links in the observed network (i.e., 224).

To measure the reliability of centralities computed on the observed data, the Pearson (r) and Kendall rank (τ) correlation coefficients were calculated between the centralities in the observed network and those in the timeline network. To understand whether applying TPM increased these correlations, the same coefficients were then calculated between the centralities computed on the augmented network and those computed on the timeline network. Additionally, each centrality measure was used to rank the nodes in each of those three networks. For each centrality, the top 10% and top 1% of nodes identified in the observed and augmented networks were compared to the most central nodes in the timeline

Table 1. The link prediction performance results for each algorithm. Note: n_c represents the number of columns in the matrix given as input to the Colibri-S algorithm.

Predictor	Avg. prec.	Top prec.	AUROC	Parameters
AA	0.0016	0.0070	0.5758	Symmetrized input
CN	0.0013	0	0.5763	Symmetrized input
CPTD	**0.0033**	0.0084	0.5941	Equally averaged both layers
MLRW	0.0012	0.0111	0.5197	
NMFP	0.0018	0	0.6496	Perturbed 10% of links; 20 iterations Colibri-S: $c = 50n_c$, ϵ = 1e-8
RA	0.0020	0.0125	0.5749	Symmetrized input
SPM	0.0010	0.0042	0.5406	Perturbed 10% of links; 20 iterations
TPM	**0.0033**	**0.0139**	**0.7836**	Perturbed 10% of links; 20 iterations Colibri-S: $c = 50n_c$, ϵ = 1e-6

network. Following Borgatti et al. [2], the true positive rate (TPR) of each centrality metric (i.e., the proportion of central nodes that were identified correctly) was calculated for the observed and augmented networks at each threshold.

4 Results

Table 1 reports the performance of each link prediction method. Though TPM and CPTD tied for the highest average precision, TPM achieved greater top precision and AUROC, meaning it was better at discovering missing connections in the streamed sample of tweets. However, none of the algorithms was able to achieve particularly strong performance on such a sparse network.

Table 2 shows the measures of centrality reliability for the observed and augmented networks when compared to the timeline network. Even without link prediction, eigenvector centrality and the related measures of hub and authority score had remarkably high Pearson correlations and true positive rates (TPRs) at identifying the most central nodes. In contrast, betweenness, closeness, and degree centrality showed worryingly low TPRs at pinpointing highly central nodes. Katz centrality fared somewhat better at ranking the top 10% of nodes but similarly dropped to 50% accuracy at finding the top 1% of nodes.

Only betweenness and closeness centrality exhibited appreciable changes in reliability after applying link prediction to the observed network. Closeness centrality saw a sizeable increase in Pearson correlation but a large decrease in its ability to correctly identify the most central 10% of nodes. However, closeness centrality also saw an increase of 10% in its TPR at pinpointing the most central 1% of nodes, which is substantial given that it did not correctly identify any of the top 1% of nodes without the help of link prediction. Betweenness centrality exhibited even greater benefits from link prediction, with an increase of 3% in its TPR at identifying the top 10% of nodes and a doubling of its TPR at identifying the top 1% of nodes from 20% to 40%. Given the small changes in Kendall rank correlation, these results indicate that link prediction changed these centrality

Table 2. The results of the centrality reliability analysis when comparing the observed network and augmented network, respectively, to the timeline network.

	Observed v. Timelines				Change after Link Prediction			
	Correlation		True Positive Rate		Correlation		True Positive Rate	
Centrality	r	τ	Top 10%	Top 1%	r	τ	Top 10%	Top 1%
Authority	1.000	0.748	0.82	1	−2e-6	+0.003	0	0
Betweenness	0.604	0.739	0.62	0.2	+0.035	−0.021	+0.03	+0.2
Closeness	0.719	0.776	0.46	0	+0.103	−0.015	−0.09	+0.1
Degree	0.816	0.816	0.64	0.5	+1e-4	−0.003	0	0
Eigenvector	1.000	0.795	0.97	1	+2e-6	+0.004	0	0
Hub	1.000	0.844	0.93	1	−9e-6	−0.002	0	0
Katz	0.887	0.789	0.79	0.5	+2e-4	−3e-04	−0.01	0

rankings more at the top end than elsewhere. These rank effects also appear to be more impactful than the changes in Pearson correlation, since a node's rank affects an analyst's interpretation more than its absolute centrality value.

5 Discussion

The link prediction trials point to TPM being better able to identify missing connections in the social network of Twitter users discussing the 2020 Nice stabbings. Though the comparison of TPM to other multidimensional link predictors did not take advantage of its ability to simultaneously work with multiple layers, time periods, and node types, future research should apply the algorithm to dynamic heterogeneous networks. This ability should be particularly advantageous when analyzing rich social media data, especially given the generally poor performance of link prediction on such sparse networks. Future research should also compare TPM to other link prediction methods across multiple data sets.

In the centrality analysis, several popular measures showed remarkably low agreement between the streamed Twitter sample and the more complete timelines. Analysts using measures other than eigenvector, hub, or authority centrality should be extremely cautious when generalizing conclusions drawn from samples of social media data. However, if collecting more data is not an option, analysts wishing to use path-based metrics like betweenness or closeness centrality can apply link prediction to increase the reliability of their identification of highly central users. Though more work must be done to replicate these results on other data sets, scholars studying harmful content on social media may find that applying link prediction to their collected networks can make their analysis more reliable than would normally be possible given the widespread difficulties with collecting comprehensive data from online platforms.

Acknowledgements. This work was supported in part by the Knight Foundation and U.S. Army grant W911NF20D0002. Additional support was provided by the Center for Computational Analysis of Social and Organizational Systems

and the Center for Informed Democracy and Social Cybersecurity. The views contained herein are those of the authors and should not be interpreted as representing the policies of the Knight Foundation, U.S. Army, or U.S. government.

References

1. Aggarwal, C.C., Xie, Y., Yu, P.S.: A framework for dynamic link prediction in heterogeneous networks. Statistical Analysis and Data Mining: The ASA Data Science Journal **7** (2014)
2. Borgatti, S.P., Carley, K.M., Krackhardt, D.: On the robustness of centrality measures under conditions of imperfect data. Social Networks **28**(2) (2006)
3. Costenbader, E., Valente, T.W.: The stability of centrality measures when networks are sampled. Social Networks **25**(4) (2003)
4. Dunlavy, D.M., Kolda, T.G., Acar, E.: Temporal link prediction using matrix and tensor factorizations. ACM Transactions on Knowledge Discovery from Data **5**(2) (2011)
5. Frantz, T.L., Cataldo, M., Carley, K.M.: Robustness of centrality measures under uncertainty: Examining the role of network topology. Computational and Mathematical Organization Theory **15**(4) (2009)
6. Hulovatyy, Y., Solava, R.W., Milenković, T.: Revealing missing parts of the interactome via link prediction. PLOS ONE **9**(3) (2014)
7. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Review **51**(3) (2009)
8. Kumar, A., Singh, S.S., Singh, K., Biswas, B.: Link prediction techniques, applications, and performance: A survey. Physica A: Statistical Mechanics and its Applications **553** (2020)
9. Lakshmi, T.J., Bhavani, S.D.: Link prediction in temporal heterogeneous networks. In: Intelligence and Security Informatics (2017)
10. Lü, L., Pan, L., Zhou, T., Zhang, Y.C., Stanley, H.E.: Toward link predictability of complex networks. Proceedings of the National Academy of Sciences **112**(8) (2015)
11. Nasiri, E., Berahmand, K., Li, Y.: A new link prediction in multiplex networks using topologically biased random walks. Chaos, Solitons & Fractals **151** (2021)
12. Ozcan, A., Oguducu, S.G.: Link prediction in evolving heterogeneous networks using the NARX neural networks. Knowledge and Information Systems **55**(2) (2018)
13. Phadke, S., Mitra, T.: Many faced hate: A cross platform study of content framing and information sharing by online hate groups. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (2020)
14. Tong, H., Papadimitriou, S., Sun, J., Yu, P.S., Faloutsos, C.: Colibri: Fast mining of large static and dynamic graphs. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2008)
15. Twitter: Volume streams. Twitter Developer Platform (2023), https://developer.twitter.com/en/docs/twitter-api/tweets/volume-streams
16. Wang, W., Cai, F., Jiao, P., Pan, L.: A perturbation-based framework for link prediction via non-negative matrix factorization. Scientific Reports **6** (2016)
17. Xia, T., Gu, Y., Yin, D.: Research on the link prediction model of dynamic multiplex social network based on improved graph representation learning. IEEE Access **9** (2021)
18. Xue, H., Yang, L., Jiang, W., Wei, Y., Hu, Y., Lin, Y.: Modeling dynamic heterogeneous network for link prediction using hierarchical attention with temporal RNN. In: Machine Learning and Knowledge Discovery in Databases (2021)