Extracting Information from Negative Interactions in Multiplex Networks using Mutual Information

Alireza Hajibagheri[1], Gita Sukthankar[1], and Kiran Lakkaraju[2]

[1]University of Central Florida, Orlando, Florida
[2]Sandia National Labs, Albuquerque, New Mexico
alireza@eecs.ucf.edu, gitars@eecs.ucf.edu, klakkara@sandia.gov

Abstract. Many interesting real-world systems are represented as complex networks with multiple types of interactions and complicated dependency structures between layers. These interactions can be encoded as having a valence with positive links marking interactions such as trust and friendship and negative links denoting distrust or hostility. Extracting information from these negative interactions is challenging since standard topological metrics are often poor predictors of negative link formation, particularly across network layers. In this paper, we introduce a method based on mutual information which enables us to predict both negative and positive relationships. Our experiments show that SMLP (Signed Multiplex Link Prediction) can leverage negative relationship layers in multiplex networks to improve link prediction performance.

Keywords: multiplex link prediction; complex networks; mutual information

1 Introduction

While both positive and negative relationships clearly exist in many social network settings, the vast majority of research has only considered positive relationships. On social media platforms, people form links to indicate friendship, trust, or approval, but they also link to signify disapproval of opinions or products. In this paper, we address the problem of predicting future user interactions from other layers of the network where a layer could represent either negative or positive user interactions. To do this, it is crucial to determine the correlation between layers. Two network layers of opposite valence are likely to have negatively correlated link formation processes. To capture the interdependencies between different layers, we use mutual information to determine the sign of the correlation. Mutual information expresses the reduction in uncertainty due to another variable; we demonstrate that the average value of a layer's mutual information can be used to calculate the correlation of link formation processes across network layers with unknown valences.

2 Proposed Method

In previous work, we introduced MLP [1], a hybrid architecture that utilizes multiple components to address different aspects of the link prediction task. While MLP utilizes information from all layers of a network to improve link prediction in a target layer, it is not able to capture negative correlations among different layers of a network. For example, let us consider α and β as layers of a network N where α and β represent positive (trades) and negative (raids) relationships between users in a game respectively. Now, if our goal is to predict future links of layer α using information from β, MLP fails to capture the negative effect of a link between two nodes on β. In order to model such relationships, we propose a model based on *mutual information* that can capture the correlation sign between different layers in order to modify the MLP weighting procedure. In this section, we first provide a short description on the concept of mutual information and describe how it has been previously used for the link prediction task in single layer networks. Finally, we introduce our new signed multiplex link prediction model.

2.1 Using Mutual Information for Link Prediction

Considering a random variable X associated with outcome x_k with probability $p(x_k)$, its self-information $I(x_k)$ can be denoted as $I(x_k) = \log \frac{1}{p(x_k)} = -\log p(x_k)$ [2]. The higher the self-information is, the less likely the outcome x_k occurs. On the other hand, the mutual information of two random variables can be denoted as:

$$I(x_k; y_j) = \log \frac{p(x|y)}{p(x)} = -\log p(x_k) - (-\log p(x_k|y_j)) = I(x_k) - I(x_k|y_j) \quad (1)$$

The mutual information is the reduction in uncertainty due to another variable. Thus, it is a measure of the dependence between two variables. It is equal to zero if and only if two variables are independent. Tan et al. [3] proposed the following link prediction model based on mutual information. Let $\Gamma(x)$ represent node x's neighbors, then for the node pair (x, y), the set of their common neighbors is denoted as $O_{xy} = \Gamma(x) \cap \Gamma(y)$. Given a disconnected node pair (x, y), if the set of their common neighbors O_{xy} is available, the likelihood score of node pair (x, y) is defined as:

$$s_{xy}^{MI} = -I(L_{xy}^1 | O_{xy}) \quad (2)$$

where $I(L_{xy}^1 | O_{xy})$ is the conditional self-information of the existence of a link between node pair (x, y) when their common neighbors are known (refer to [3] for more details).

2.2 Signed Multiplex Link Prediction

Various experiments on multilayer link prediction have indicated that using neighborhood information from different layers of a network in a multiplex environment can improve the performance of link prediction. Hristova et al. [4] proposed the concept of a multilayer neighborhood where a link that exists on more than one layer in a multiplex network is called a *multiplex link*. Following the definition of a multilayer network, the ego network of a node can be redefined as the multilayer neighborhood. While the simple node neighborhood is the collection of nodes one hop away from it, the multilayer global neighborhood (denoted by GN) of a node i can be derived by the total number of unique neighbors across layers:

$$\mathit{\Gamma}_{GNi} = \{j \in V^{\mathcal{M}} : e_{i,j} \in E^{\alpha \cup \beta}\} \tag{3}$$

Similarly, the core neighborhood (denoted by CN) of a node i across layers of the multilayer network is defined as:

$$\mathit{\Gamma}_{CNi} = \{j \in V^{\mathcal{M}} : e_{i,j} \in E^{\alpha \cap \beta}\} \tag{4}$$

The goal is to determine the type of correlation (negative or positive) between two layers of a multiplex network. To this end, we calculate the average mutual information value for a target layer based on predictor layers. There are two assumptions: 1) Using the core neighborhood, if there is a negative correlation between two layers, the value of average mutual information would decrease substantially but would not change significantly if there is a positive correlation between the two layers. 2) If the global neighborhood is used to calculate the value of mutual information for a node pair, this would increase the value of average mutual information if there is a positive correlation between two layers and would not change significantly otherwise. After the sign of the relationship between the target layer and other predictor layers has been determined, MLP is used for predicting future links with the minor addition that layer weights can have both negative and positive values. Given a disconnected node pair (x, y), the mutual information of link existence (assuming that the set of their common neighbors O_{xy} is available) can be derived as:

$$I(L^1_{xy}; O_{xy}) = I(L^1_{xy}) - I(L^1_{xy}|O_{xy}) \tag{5}$$

For a multiplex network, O_{xy} can be redefined to use the global and core neighborhood of the two nodes. As a result, the average mutual information of a target layer would be defined as:

$$MI^{\alpha} = \frac{1}{|E^{\alpha}|} \sum_{x,y \in E^{\alpha} \& x \neq y} I(L^1_{xy}; O_{xy}) \tag{6}$$

The value of MI^{α} is calculated using information from a predictor layer β. It can be used to indicate the sign (negative or positive) of the correlation between the link formation in two layers. This sign is then used within the second phase of link prediction task (MLP) to assign weights to all predictor layers and hence improve the performance of link prediction task for the target layer α.

3 Experimental Study

This paper evaluates the SMLP framework on networks extracted from two real-world datasets, Travian and Cannes2013. Not only do we compare our results with two other approaches for fusing cross-layer information, but we also consider scores generated by mutual information paired with core and global neighborhood as distinct methods.

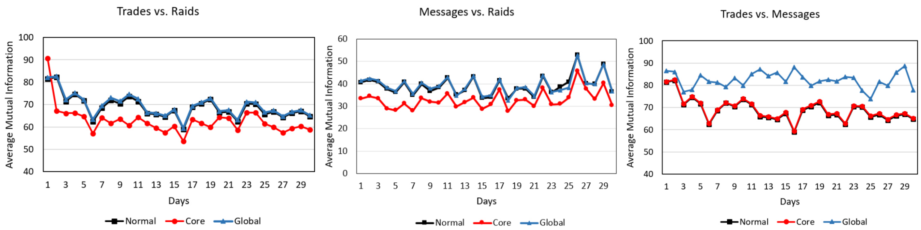

Fig. 1: Average mutual information values over time calculated using normal and core neighborhood definitions.

3.1 Datasets

We use two real-world dynamic multiplex networks to demonstrate the performance of our proposed algorithm. These networks are considerably disparate in structure and were selected from different domains. Negative links (raids) exist between users of our MMOG dataset to evaluate the performance of our method in predicting negative links as well as examining these layers on positive ones.

- **Travian MMOG** [5] Travian is a browser-based, real-time strategy game. Trades, messages, and raids networks were extracted from Travian for this research.
- **Twitter Interactions** [6] This dataset consists of Twitter activity before, during, and after an "exceptional" event as characterized by the volume of communications. The Cannes2013 dataset was created from tweets about the Cannes film festival that occurred between May 6,2013 to June 3, 2013.

3.2 Evaluation Metrics

For the evaluation, we measure receiver operating characteristic (ROC) curves for the different approaches. The ROC curve is a plot of the *true positive rate (tpr)* against the *false positive rate (fpr)*. We report area under the ROC curve (AUROC), the scalar measure of the performance over all thresholds.

3.3 Analysis of Multilayer Neighborhood

As mentioned before, our assumption is that both core and global neighborhood definitions would enable us to study correlations between different layers of a network. Figure 1 shows average mutual information values for Travian network for each day of the 30 day period. There are negative and positive correlations between trades (or messages)-raids and trades-messages respectively. As shown in the figure, at every timestep of the network, the average value decreases when the core neighborhood is used in the case of negative correlation (Figure1 (a) and (b)) and does not change significantly when global neighborhood is used to calculate the average value. On the other hand, as shown in Figure 1 (c), when dealing with positive correlations (trades-messages), the value of average mutual information does not change significantly using

Table 1: AUROC performances for a target layer averaged over all snapshots and ten runs.

	Trade	Message	Raids	Retweet	Mention	Reply
SMLP	**0.871±0.013**	**0.843±0.031**	**0.793±0.007**	**0.812±0.002**	**0.834±0.003**	**0.839±0.002**
MLP	0.821±0.001	0.803±0.002	0.758±0.001	0.812±0.002	0.834±0.003	0.839±0.002
MI (CN)	0.740±0.016	0.753±0.011	0.744±0.013	0.774±0.009	0.759±0.012	0.782±0.005
MI (GN)	0.737±0.010	0.746±0.011	0.747±0.009	0.771±0.011	0.767±0.012	0.773±0.009
MI (N)	0.716±0.012	0.727±0.006	0.703±0.012	0.731±0.007	0.725±0.013	0.742±0.014
AA	0.744±0.030	0.752±0.020	0.658±0.017	0.740±0.003	0.737±0.011	0.761±0.003
EA	0.731±0.004	0.763±0.020	0.661±0.009	0.749±0.003	0.758±0.031	0.744±0.002

the core neighborhood but increases drastically using the global neighborhood definition (the average difference is less than one). Hence, this justifies our assumption that the average mutual information value of a certain layer can be used to determine the sign of its role in target layer link prediction.

3.4 Performance of Signed Multiplex Link Prediction (SMLP)

Table 1 shows the results of different algorithms on the Travian and Cannes2013 datasets. With 30 days of data from Travian and 27 days for Cannes2013, we were able to extensively compare the performance of the proposed methods and the impact of using different elements. AUROC performances for a target layer averaged over all snapshots are calculated, and our proposed framework is shown at the top of the table, followed by variants of mutual information (MI) based link prediction models using different definitions of neighborhood (N which stands for Normal proposed by Tan et al. [3], CN stands for Core Neighborhood and GN stands for Global Neighborhood). The algorithms shown in the bottom half of the table (Average Aggregation (AA) and Entropy Aggregation (EA) [7]) are techniques for multiplex networks proposed by other research groups. The settings given in [1] were used for MLP. Bold numbers indicate the best results on each target layer considered. As expected, SMLP is the best performing algorithm in all cases since not only it utilizes both historical and cross-layer information, but also mutual information enables SMLP to capture negative correlations between different layers. As a result, node pairs that are connected by raids in Travian, are penalized for this connection and eventually receive a lower score compared to node pairs that are only connected on messages and trades. This holds true when raids is the target layer and two nodes are connected on either messages or trades layers which are negatively correlated with raids. On the other hand, it is evident that methods specifically designed for multiplex link prediction outperform Mutual Information (N) which is unable to leverage cross-layer information. Also, Average Aggregation and Entropy Aggregation are able to achieve higher AUROC scores compared with Mutual Information based methods since they collect more information from the network using different similarity metrics such as common neighbors, Adamic/Adar, etc. Finally, for Twitter layers, SMLP and MLP achieve similar results since there are no negative layers to modify the sign of the weights associated with different layers of the network.

4 Conclusion and Future Work

In this paper, we introduce a new link prediction framework, SMLP (Signed Multiplex Link Prediction), that employs a holistic approach to accurately predict links in dynamic multiplex networks by incorporating negative relationships between users. Our analysis on real-world networks created by a variety of social processes suggests that SMLP effectively models multiplex network coevolution in many domains and is also able to capture negative correlation between layers in order to improve link prediction performance.

In future work, we are planning to extend our results to other multilayer networks that contain negative interactions; we are particularly interested in studying datasets that contain multiple negative interaction layers to see if they result in positive correlations. Another promising line of inquiry is studying the usage of multiple features (beyond common neighbors) to calculate mutual information, since different structural features highlight different aspects of network formation.

Acknowledgments

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. The Travian dataset was provided by Drs. Rolf T. Wigand and Nitin Agarwal (University of Arkansas at Little Rock, Department of Information Science).

References

1. Hajibagheri, A., Sukthankar, G., Lakkaraju, K.: A holistic approach for link prediction in multiplex networks. In: International Conference on Social Informatics. (2016)
2. Shannon, C.E.: A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review **5**(1) (2001) 3–55
3. Tan, F., Xia, Y., Zhu, B.: Link prediction in complex networks: a mutual information perspective. PloS One **9**(9) (2014) e107056
4. Hristova, D., Noulas, A., Brown, C., Musolesi, M., Mascolo, C.: A multilayer approach to multiplexity and link prediction in online geo-social networks. arXiv preprint arXiv:1508.07876 (2015)
5. Hajibagheri, A., Lakkaraju, K., Sukthankar, G., Wigand, R.T., Agarwal, N.: Conflict and communication in massively-multiplayer online games. In: Social Computing, Behavioral-Cultural Modeling, and Prediction. Springer (2015) 65–74
6. Omodei, E., De Domenico, M., Arenas, A.: Characterizing interactions in online social networks during exceptional events. arXiv preprint arXiv:1506.09115 (2015)
7. Pujari, M., Kanawati, R.: Link prediction in multiplex networks. Networks and Heterogeneous Media **10**(1) (2015) 17–35