Layer Prediction on Multilayer Social Networks and Applications

Xie He

University of North Carolina, Chapel Hill

1 Introduction

Multiplex networks have been studied extensively over the past few years. The structure of a multiplex network has proven to be very useful in determining a lot of different structures, especially social networks.

The idea of multiplex network is to use graph structure to represent a dataset. For example, given a number of agents, and their interaction with each other, can we model these agents and their interactions using a simple graph model. The answer is we could, we would let the agents be the nodes of a graph and their interactions be the edges of this graph. Thus we constructed a network from a dataset and we could use a lof of properties of networks to study this dataset.

Multiplex network is a extended idea upon the idea of network. In general, we have different types of interactions between our agents, and all we want to do is to study the different interactions between the same group of individuals, thus when we visualize a graph, it will have different layers in general.

Social networks have always been an important part of multilayer network study. There are a lot of different applications that multilayer networks could be used to study social groups, but few have been made on how two different types of interactions could be similar to each other.

Namely, take a simple student group as example. They have a leader in this group; have persons they will go to if they have good news and persons they will go to if they have bad news; they also have close friends. All these different relationships are not necessarily the same in terms of social study, but how do these interactions interact with each other and to what extant are they similar to each other could be a very interesting topic to look into in this case.

To study this, we introduce the idea of layer similarity. That is, how does the behavior in one layer of a multiplex network represent the behavior of another layer of the same network. To study the similarity between different layers can be very useful in social study because if two types of interactions between people are extremely same then when designing the questionnaire, we could save a lot of time and ask as few questions as we could, which might result in higher respond rate and thus give us more useful information than simply ask a lot of redundant questions.[2]

2 Algorithms

We know there are a lot of different algorithms to test the similarity between two layers. Here we simply employ an existing method that has been doing a very good job in predicting the similarity between two different layers, [1].

The authors use a expectation maximization to compute the likelihood of an edge existing in a particular layer.

$$M_{ij}^{\alpha} = \sum_{k,l=1}^{K} u_{ik} v_{jl} w_{kl}^{\alpha}$$

In their paper,M_{ij}^{α} represents the likelihood of existence of an edge in layer α, whereas they give each node i two membership vectors, u_i and v_i , which determine how i forms outgoing and incoming links respectively. In the not directed network, we just

have $u = v$. Also, note that w^{α} is the affinity matrix for layer α, while K is how many communities there are in this multiplex network.

By communities, we basically mean clusters. Individuals within same social group are considered a community and the modeled require pre knowledge about how many communities there are in a social network before being able to predict the similarity and the existence of edges in a particular layer.

We note also that in [1], they did not specify the way the chose this K, but just decides to go with the K that gives the best AUC level. We take the same procedure here in this project. We also have different plots with different K values just to make sure that the same pattern are the same within the same dataset.

We describe our method in details in the next section.

[1] The idea of

3 Results

We test our methods on two different public datasets for the simplicity to reproduce the result.

One is a Stanford dorm dataset that has eight different layers and the other one is an indian villiage dataset with twelve different layers. Both dataset are public online and the Stanford dorm dataset is uploaded as a test example due to its small size on the github of this report.

These datasets are both social network datasets with layers such as "bad news", "influence". The indian villiage dataset is a little bit different in that it also has "borrow money" layer.

The Stanford dorm datasets overall has 200 nodes per layer and the indian villiage dataset has around 400 nodes per layer.

The two datasets are both directed multiplex networks. The directions is considered as one person nominating the other person when being asked relevant person. For example, if node A points to node B in "bad news" layer, this simply implies that "A will talk to B when A heard a bad news."

We describe in details the test we have done on each dataset: for each layer, we remove 20 per cent of information from this layer. Note it is very important to notice that it is 20 percent of the the information instead of 20 percent of the edges. Though removing 20 percent of the information from the adjacency matrix on average gives a 20 percent of removed edges, we made this distinction because the [1] original paper made it clear that this is what the authors in [1] have chosen to do and we kept their methods to be careful.

For single layer cases, we than use the rest 80 percent of the information from that layer to predict the likelihood of an edge's existence in that layer. After this we calculate the AUC value on the removed 20 percent of the entries of the adjacency matrix to see how well our performance are.

On average the Stanford dorm gives a around 0.6 AUC value for the single layer cases and the indian villiage gives a 0.5 AUC value for the single layer cases.

Now, we are not satisfied with the single layer cases for sure. The most important experiment we did are for the double layer prediction. Namely, we removed 20 percent of information from layer A than we provide the model with the rest 80 percent of information of layer A and the entire information of layer B as supportive layer to predict the likelihood of an edge's existence in that layer.

This explains precisely which layers are more similar to each other. In particular we plot out the heat map we produced for the Stanford dorm cases 1. From the heat map of the Stanford dorm, it is very clear that the "feel positive" and the "spend time" are very different from all the other layers. When using the similar layers to predict each other, they did a very good job while if you use not similar layers to predict them, they did a bad job.

This conclusion is clearer if we also draw the ROC plots of one of each group, without loss of generosity we choose "bad news" 2 and "feel positive" 3.

It is clear from the plots that the similar layers did a very good job with the AUC around and above 0.9 while nonsimilar layers only give around 0.7 AUC.

Thus we could safely say that if we could apply this information to different communities numbers, which is a prerequisite information from the model, than we are able to say that the social network from the Stan-

ford dorm data could definitely be condensed into two different layers almost tellingn the same information.

We go ahead the iterate through different numbers of communities, where we later call K and input these different K into the data to see what we would get.

4, 5, 6 gives the three different scatter plots with different K, but all the scatter plots of the Stanford Dorm dataset tells the same story: we have a strong connectivity between the similar layers and a different pattern between not similar layers.

We also produce the same experiment of the Indian Villiage dataset, due to the page limit of the report, we only shown one of the scatter plot in this report, more scatter plots could be found on github. 7

Not suprisingly, the result is not as good as the Stanford dorm dataset. Three reasons could lead to this scenario: first, this scatter plot is averaged out over 6 different villiages, thus the result could be very different if we are looking at a single village; second, this dataset is much more spare in terms of the edge density, almost having 400 nodes, it only have as many edges and thus could result in a not giant connected component, which will make it very difficult to detect any communities in this network and thus result in a worse performance; the third reason is that the layers in this dataset is inherently different from the Stanford dorm dataset. It has many financial related interpersonal relationship layers like "borrow money" while the Stanford dorm dataset mainly focuses on the interpersonal social relationship.

All these could lead to interesting potential future work which we would describe in detail in the future work section.

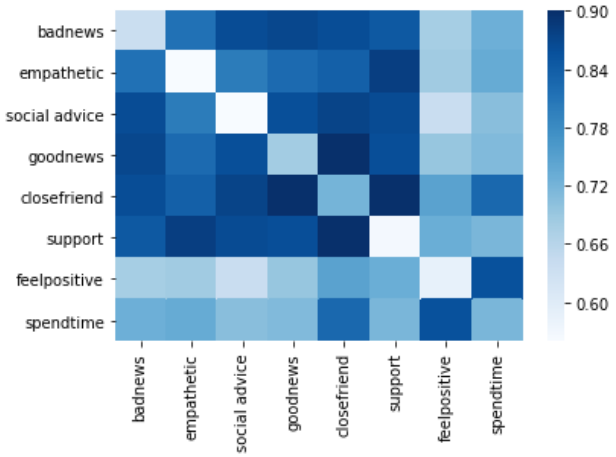

Figure 1: Heat map of Stanford Dorm Dataset

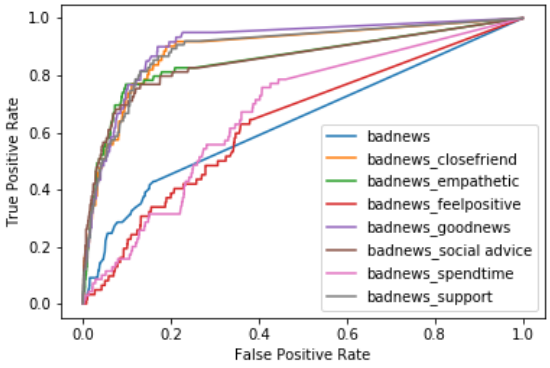

Figure 2: ROC plots for "bad news" layer in the Stanford Dorm dataset

4 Conclusion

We have seen that our method works perfectly on one data we have chosen and not so good on the other one. We have figured out that the similarity between layers could lead to simpler questionnaire in social study thus save us time and efforts in study different social structures. We have also noticed that the social network and the financial interpersonal network, though both belong to interpersonal relationship, are very different and thus could not be used to predict each other and are not similar to each other.

In short, we have showed an application on multiplex social network that the link prediction method could show similarity between different social interaction and thus study the similar and non similar pattern in a social network.

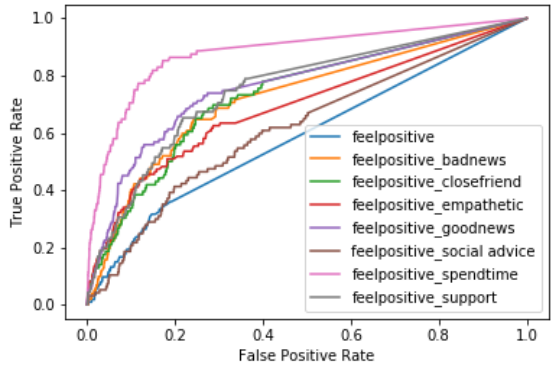

Figure 3: ROC plots for "feel positive" layer in the Stanford Dorm dataset

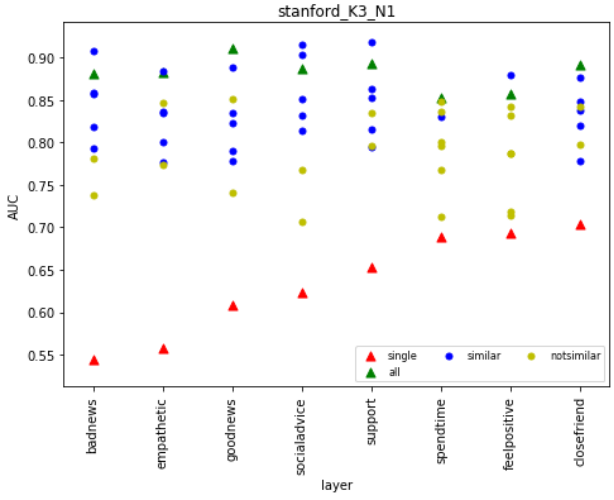

Figure 4: Stanford Dorm Dataset with 3 different communities averaged out over 5-fold cross validation. Along the x-axis are different layers that is being predicted; red triangle represents using 80 percent of its own information; green triangle represents using 80 percent of its own information and all other layers; blue dots means using a similar layer as reference; yellow dots means using a non-similar layers as reference.

5 Future Work

The idea of how many communities could affect the outcome could be very interesting. The basic steps

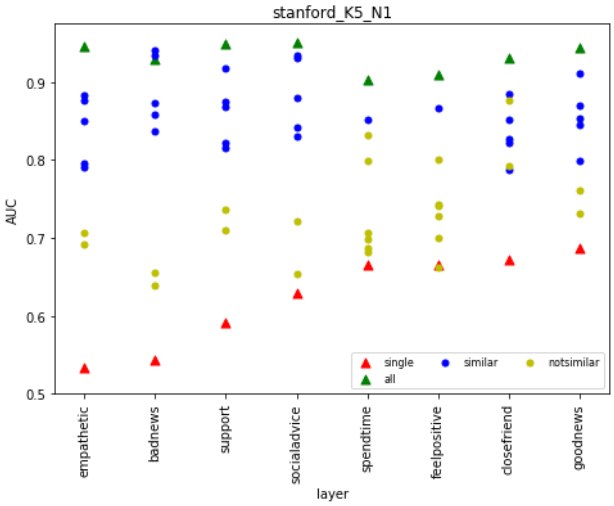

Figure 5: Stanford Dorm Dataset with 5 different communities averaged out over 5-fold cross validation. Along the x-axis are different layers that is being predicted; red triangle represents using 80 percent of its own information; green triangle represents using 80 percent of its own information and all other layers; blue dots means using a similar layer as reference; yellow dots means using a non-similar layers as reference.

to proceed is to find the best K for each pair of relationships and run 5-fold cross validation to see if we can find a better explanation for this behavior.

Another clear thing to do is to separate the current experiment's averaging out process. Since we are averaging over the same dorm data set over different runs and average over different villiages, this could lead to clear change in the outcome as well.

One possible explanation for all of these are that the social networks are inherently very different and could have many different properties. Especially when the relationship networks and financial network are not necessarily the same, thus another direction to go is clearly study on the sociological side of the picture, which is also another important application part of this project.

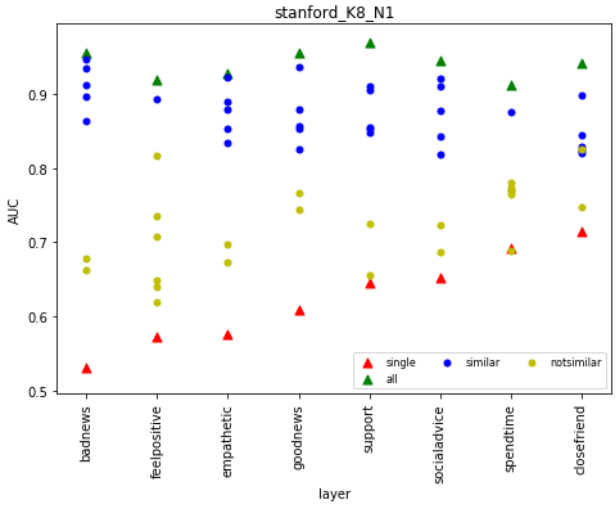

Figure 6: Stanford Dorm Dataset with 8 different communities averaged out over 5-fold cross validation. Along the x-axis are different layers that is being predicted; red triangle represents using 80 percent of its own information; green triangle represents using 80 percent of its own information and all other layers; blue dots means using a similar layer as reference; yellow dots means using a non-similar layers as reference.

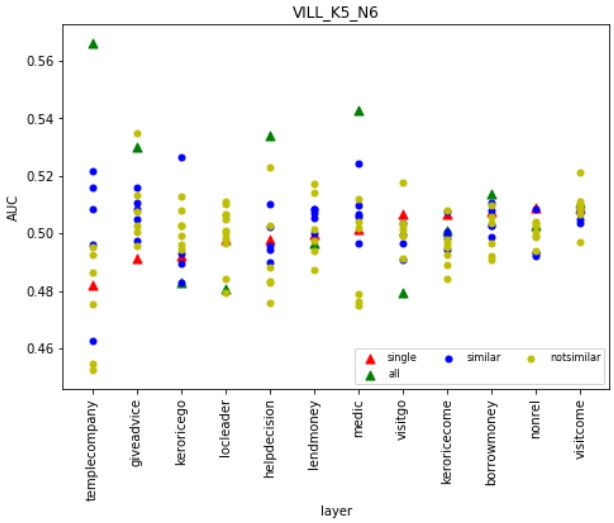

Figure 7: Indian Village Dataset with 3 different communities averaged out over 5-fold cross validation. Along the x-axis are different layers that is being predicted; red triangle represents using 80 percent of its own information; green triangle represents using 80 percent of its own information and all other layers; blue dots means using a similar layer as reference; yellow dots means using a non-similar layers as reference.

6 Acknowledgement

We thank Eun Li, Olivia Staoni, Peter Mucha for helpful discussion and for help throughout the project. Note this is an ongoing collaboration among the MURI project.

References

[1] Caterina De Bacco, Eleanor A. Power, Daniel B. Larremore, and Cristopher Moore. Community detection, link prediction, and layer interdependence in multilayer networks. *Phys. Rev. E*, 95:042317, Apr 2017.

[2] Francois Lorrain and Harrison C White. Structural equivalence of individuals in social networks. *The Journal of mathematical sociology*, 1(1):49–80, 1971.