Recommendation System Based on Statistically Validation of Bipartite Structure Networks on IMDB Database

Onur Poyraz,Department of Computational Science and Engineering, Bogazici University

***Abstract*—Complex networks have several types but in today's applications, many of the complex networks have an intrinsic bipartite structure. Therefore there is two set of nodes which are not connected among themselves instead they are connected with the nodes which belong to another set of nodes. In these networks, element of one set is hugely different than element of others.(for example animality world and food world). This types of networks usually investigated by constructing projected network on one of the sets. This means to analyze such system we remove one of the set of nodes and we project edges to the remaining nodes. Actually these method implies the common interests of one set of nodes on the other one. However in complex structures this does not gives too much information about the structure of the network. Therefore there should be further investigations and applications. We have to reduce the edges in the projected network. Otherwise it just shows the systems heterogeneity.[1] Tuminello et all describe an unsupervised method to statistically validate each link of a projected network against a null hypothesis that takes into account system heterogeneity. They test their algorithm on different complex structures and their method is able to detect network structures. In this structure I used statistically validation algorithm to discover the common interests of the users over movies of the IMDB Movie Lens dataset.**

***Keywords*—Statistically Validated Networks, Bipartite Complex Networks, Null Hypothesis, Heterogeneity, Preferential Link**

I. Introduction

In recent years, there are a lot of investigetion on complex network structures. This is because with the growing of internet, there is need to analyze the relationships between the datas to extract relatively important information from the unnecessary ones. These types of algorithms and systems are hugely demanded from the companies and research people.

When one investigated previous works, there is some common features almost all complex networks have. One type of these networks is bipartite network. I will analyze bipartite networks with starting one-mode projection. This will reduce the dimensions and shapes of the structures and left the important nodes, since for example I am not interested in the movies instead I interested in users choice. In this paper we analyze bipartite structured networks. However this method is not sufficient for the huge networks since in that types of networks,nodes are connected to each other in anyway. So I will use preferential link[1] to extract informative connections. This seperates the potentially noncoincidence edges from the potentially coincidence edges. To analyze that I use statistically validation of given node method.
One of the other important thing is heterogeneity. This term represents the non-uniform structures of the networks. For example, author-paper bipartite structure has a heterogeneus structure since almost all authors participate in different works and different number of works.

II. Methodology

Here I will introduce my algorithm that I follow for this paper. Here I started with the One-Mode Projection. Than I get weighted undirected projection graph.After taking weighted graphs I use statistically validation algorithm to find most powerful weights in the network. For that purpose I use null-hypothesis of random connectivity between elements to specify the relationship between nodes. Than I test every node of the network against null-hypothesis. For a reference point I assign a p-value to check each of the node's statistical validation.[2] For a final step I check statistical significance level of each edges between nodes and I create statistically validated network.[3]

A. Creation of Statistically Validated Network

In this part I give details about my algorithm. Here firstly, one have to decide the bipartite system. Once we determine bipartite system S (IMDB Movie Lens Dataset in this report) I divide it into subsets A and B. In our applied dataset A represents the Users and B represents the Movies. I define the edges between nodes from users ratings for movies. The dataset contains users rating from 0 to 5. But for simplicity I convert it to binary method such that it just contain like and dislike. Since I am interested in the similarities between users I projected my network on the user subdomain (A). Corresponding adjacency projected network is obtained by linking vertices of A which share at least a common first neighbour element of B in the bipartite system.[4] After that I statistically validate each link against null hypothesis of random co-occurance of common neighbors.[1]. Here I have to take into account heterogeneity of elements in both sets. To make that we divide system to subsystems. In this report I just use 2 degree subsystem because for the dataset that I use the other ones are not highly necessary to complete statistical validation. By that way I reduce the cost of calculations. So my system S includes N_B number of elements in set B. On the other hand its calculate N_i and N_j which are number of connections that the nodes made from A. Lastly, $N_{i,j}$ is the number of common connections that node i and node j made.

$$H(X|N_B, N_i, N_j) = \frac{{}^{N_i}C_X \, {}^{N_B - N_i}C_{N_j - X}}{{}^{N_B}C_{N_j}} \tag{1}$$

The above formula gives the probability that elements i and j share X neighbors in set B using hypergeometric distribution[5]

$$p(N_{i,j}) = 1 - \sum_{i=1}^{N_{i,j}-1} H(X|N_i, N_j, N_B) \tag{2}$$

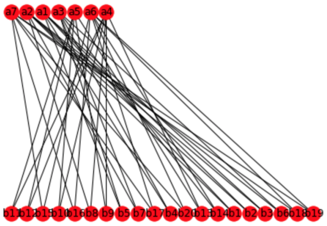
In above equation if the similarity between two nodes is relatively high than p-value approachs zero. Therefore I set a threshold (s = 0.0005) to check and assign a link validated. If p-value is under s than I validate that link. After all the resulting network is my new statistically validated network.[6]

III. Results

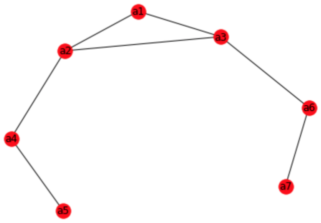
I implement my algorithm to two different dataset. One of that is used as a example data from Tumminello et all. The other one is MovieLens dataset. I will put both of the output for clearance.

A. Example Data from Tumminello

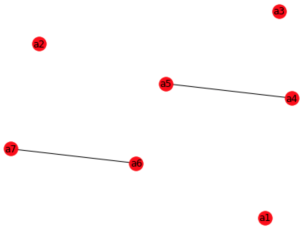
Here firstly I draw the bipartite system. This graphs shows all the nodes and edges.



After that I draw projected graph. That removes one set and project all the network into the other set of nodes.

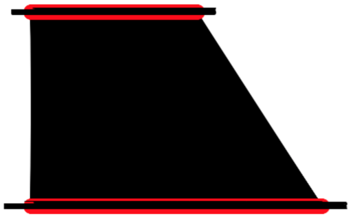


Finally, I apply the algorithm to dataset and plot the resulting statistically validated network.

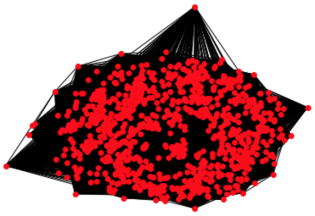


B. MovieLens Dataset

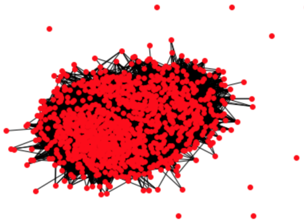
This dataset is quite larger (100.000 ratings) so the plots that I made is not meaningful. However I will put some of the meaningful plots. The first one is initial bipartite structure



The second one is corresponding projection graph.



The final one is plotted after the applied algorithm.



IV. Conclusion

In this project I implement the statistical validation of complex networks. Since I kept for all linked nodes the projected node, after the statistically validation I can compare their links again. This give me power of recommend movies to users according to their statistically validated neighbor user. From the data I can find similarities of users and user's hobbies. I try to make a recommendation system that recommend movies for just relevant users.

V. Future Works

After this work one can make good visualisation methods for this type of networks for clearence. It is important to understand what you are making. Some clustering algorithms can apply to resulting statistically validated network for better performance.

Acknowledgment

The authors would like to thank Haluk Bingol and Michele Tuminello

REFERENCES

[1] Michele Tumminello et al, Statistically Validated Networks in Bipartite Complex Systems: 2011

[2] Miller RG, (1981) Simultaneous Statistical Inference. New York: Springer- Verlag, second edition.

[3] Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B 57: 289300.

[4] Herrmann C, Species Co-Occurrence Patterns among Lyme Borreliosis Pathogens in the Tick Vector Ixodes ricinus, 2013

[5] Feller W (1968) An Introduction to Probability Theory and Its Applications, volume 1. New York: Wiley, third edition.

[6] Newman MEJ (2001) Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. Phys Rev E 64: 016132.