Social Network Analysis

Jeorval Cano
Universidad Politecnica de
Yucatan
Social Network Analysis
Merida Yucatan
st1809028

Abstract

In this project, I am going to analyze my own Facebook network, with the help of different techniques, we will be able to get the data to make the visualization of my network, including another level to the diameter, by looking toward my friends' networks (specifically mutual friends), in the paper we will see the specification of the network to implement the different techniques such as link prediction, and community detection. Besides the complex methods I will be providing different metrics at different stages of the analysis, with that we will be able to have a better understanding of what is going on within the network. The main purpose of the paper is to show how community detection, and link prediction are made. I will show how the graph and changes in case we have a post prediction interpretation of the graph, and we will discuss the main characteristics of each of the stages so that it could be well understood.

Keywords. Social Networks, Facebook Network, Link Prediction, Community Detection.

I. Objective

For this project there are going to be some objectives that must be covered to understand the sense of all this project, there are not going to be much explanation about the basic characteristics of the network, because this paper is supposed to be about advance topics such as link prediction and community detection.

The set of main objectives for this project will be the following:

- Through metrics, we will discuss the different opportunities that the data can share with us in case we wanted to mine the data.
- Get the data from an API or web scrapping, the decision must be explained as well as the process that followed the web scrapping.
- Understand the importance of the Link Prediction, as well as its uses
- Make the implementation of the Link Prediction, and show a case of study to see how it works in a practical example

By accomplishing the different objectives, the reader should have an idea of how a project like this work and what are the main uses and cases where we can apply the different techniques to the data.

II. Introduction

In social networks, many studies have emerged from its popularity, and the wide range of topics related to graph theory and social network analysis that in the past were not exploited at all, this might be because of the lack of resources available to apply all the different theories, and algorithms, or maybe because there was not activity from where we could benefit from, but with the help of the different platforms, and the way they try to get you on their app and improve the performance of your visit the time you spend in their platforms, these algorithms become much more popular, and it makes that they get much more support and that new applications become developed because of this. Computing all this is quite expensive, because, to have the information of millions of people makes greater the complexity of the problem, nevertheless, by making a local analysis visible the different processes that can be taken into consideration whenever we are making a social network analysis, and specifically when we want to make a recommendations system with the help of data, and these techniques. At the end of the day, we can see this in our daily life, whenever we get a notification of facebook suggesting us to add someone we might know, or even when twitter asks us to follow a profile that some of our friends also follow. Like these applications there are many that we can get from Social Network Analysis, and in particular from the Link Prediction technique, and the community detection, making them valuable algorithms that can be the engine to develop powerful applications.

Therefore, to apply all the prior mentioned this project used webscrapping to get the social network of a given person in which it gets all their friends and then for each friend it gets the mutual friends, the webscrapping for this project was made in that way because getting the friends of friends it takes a lot of time and the websracpping could go over 6 hours as just getting mutual friends it took over 4 hours and also getting friends of friends was not our objective since the goal wanted it was to observe the characterisicti of a network of mutual friends given a base user.

III. Mapping Process

For the mapping of the network, there were needed a lot of resources to extract the information to get closer to have a relevant mapping, so to accomplish the goal, it was needed to use other ways to get the data because the API of Facebook does not let you get the mapping of the network as long as you do not get the necessary permissions which was not possible due to the short time to make this analysis, however, there are some cases where it is possible to get enough information To get all the nodes and edges, via API, but sometimes it costs, and the tools for developers are a bit restaged by the enterprises and make difficult to make all this process, however, there is a way to avoid this which is what was used.

The technique is web scraping, and it is a very powerful tool whenever we want to get information from the web because lets you automate the process of extracting valuable information from the user profile, and with all the “credentials” that a normal user has, this makes it beneficial for the data analysis because we can have a great volume of data, in exchange of some processing time, the reason why this process is quite long is because of the way you get the data, with web scrapping we can find ways of finding specific information with the help of the HTML notations, and also with some REGEX structures. With web scraping, you can handle the DOM as you will, and this library lets you have this interactivity between the computer, the UI, and the information that is placed in the DOM.

With this technique I was able to get all the data that we are going to be analyzing all along with this paper, and from here you can know how powerful this is, and it turns very useful whenever we want to analyze a webpage that does not provide us the necessary tools in their API to make the analysis.

The web scrapping code works given a Facebook profile; email, password and username, the bot log in your account and then it gets to your friends and obtain their Facebook usernames, something important to mention is that this code it does not save the profiles with default usernames, i.e.- Facebook friends who never edited their usernames and appears like “profile.php?id={SOME ID NUMBER}” because it bugged the code and if we keep them later it will cause more problems. After collecting the friends of the given account(in this case mine) the bot goes to each of the friends profile and the collects the mutual friends, the reason why it does not collect all the their friends is because then to clean the data it gets really difficult because of the way the web scrapping works as it also gets pages they like and groups they are in, this happens due to the regex as it cannot differ from usernames and pages, so to avoid that the bot only gets the mutual friends and another reason is because I only wanted to have a network of my friends and the mutuals friendship among them as I also wanted to do a community analysis, and by doing friends of friends there were a lot of nodes with degree one having some communities without sense.

After getting my friend list and the mutual friends of each friend, the network can be created as each friend is a node and the mutual friendship represents the edges among them. Finally, I manually check that all the nodes are actually friends because in my case and the end there are four nodes that are not friends, but pages so I eliminated them, then I also clean the mutual friends of each friend and this is easier since are mutual there only can be people that also my friends, therefore, for each friend the code eliminate the mutual friends that are not in my list of friends that all the times are pages or groups that the regex identified as usernames.

IV. Basic Characteristics

To have a better understanding of the data we have gotten thanks to the usage of the techniques previously explained, it is necessary to measure our Social Network, as it is one of the main aspects, we need to know to proceed with the analysis, and this will help us to have a hypothesis and also have a better understanding of what is going on with the network. Besides all that, we might use these metrics to compare them with link prediction broadcast.

In terms of the nodes that are in my social network, there are 280, which as I said were extracted from my Facebook that number of nodes I did expect that number even though I have 350 friends in Facebook as I latter mentioned I have several friends with a default username, despite it seems to be a low number of nodes, there are many edges, this is beneficial for the study because we will have more predicted edges thanks to the characteristics of this network, where are all more susceptible to be linked. Besides that, I only took into consideration my friends and common friends as edges, because it is better for the study and the community detection that is going to explain later.

In terms of the average shortest path, as expected is nearby 2, this because the diameter of the graph is 2, and it is necessary to understand that this is because we only took into consideration my friends and mutual friends, which makes it more susceptible to be 2 this result was expected since the largest path is when two nodes do not know each other and my node or another works as the connection for those two nodes, meaning that the network is basically of two units depth as it is a network that in the worst case my node units them if they are not unit by their own

Other metrics such as the Density of the graph, have a low number because many groups are not connected between them, and that makes to have a low density.

Now according to the nodes in the periphery, we found that there are 279 nodes, which is curious because all except 1 are considered peripheral nodes, but this is expected as my node works as the center and the reaming must be the periphery the only case where there are more centers is in which there is a node with the same friends as I because as long as we are the link between communities it is understood that we are the node in the center.

Here is a summary of the basic characteristics of the network:

- Number of nodes: 280
- Number of edges: 2981
- Average shortest path: 1.923
- Density: 0.076
- Diameter: 2
- Eccentricity: 1
- 279 nodes are the periphery
- 1 nodes are the center
- Cluster coefficient: 0.533
- Average degree: 21

Once we have a clear understanding of the basic characteristics of the graph it is needed to know more about the centrality distribution.

In this case, we can see that the values are approximated to be below .4 and just one with a degree centrality of 1, this is a reflection of what it was said, and in the plot below it is possible to have a view of the distribution of the degree centrality of all nodes.

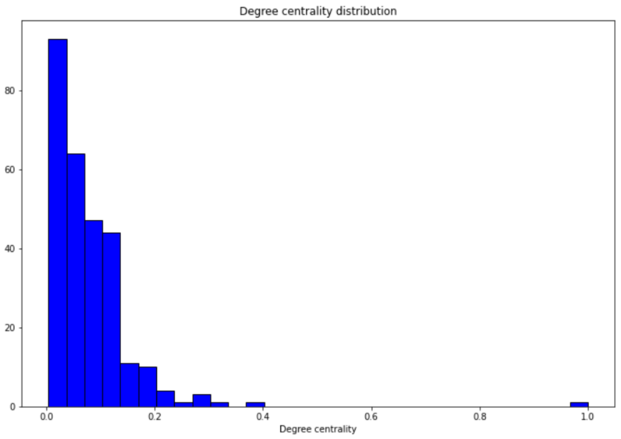

Figure 1. Degree centality.

Now that we saw the degree centrality distribution, we can move forwards to see the closeness centrality distribution, as you may see, the values are .5 and greater, which makes it possible because of the closeness between nodes, and as long as the diameter is of length 2 in the closeness thanks to the node in the center become relevant for the understanding of the plot of closeness centrality distribution.

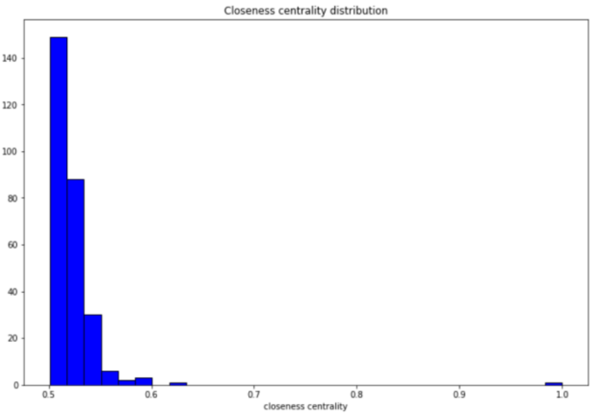

Figure 2. Closeness centrality.

Finally the model degree distribution, it is possible to observe that we have an aglomeration in lower values, what might represent aproximatly 1/6 of the nodes that are in the network, and just one connected to all of them, in the plot bellow the distribuiution parcial and it was expected thanks to the communities found along the network. The reason why there are nodes with degree one, exactly 23 nodes, is because those nodes are actully friends in which we have no common friends ,and also friends that either their account has been deactived or they have deactived to other people to see their friend list, including mutual friends.

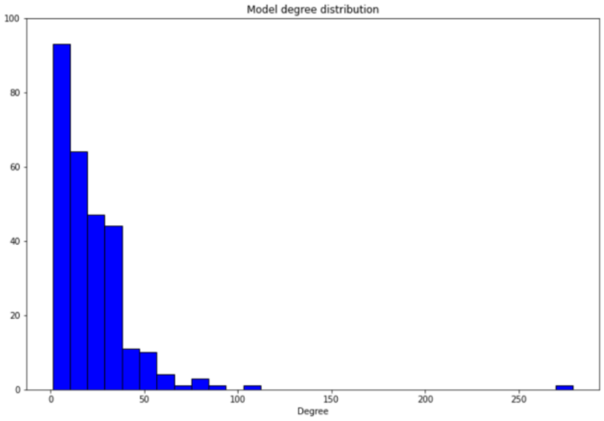

Figure 3. Degree distribution.

In the following section, we will discuss the graphical representation of the network, as well as the different techniques employed to have a better visualization of the network, for it was necessary to make a community detection and a segmentation of the colores in the proposed plot.

V. VISUALIZATION

Once we have understood how is the graph according to the characteristics, and all the peculiarities that it has, the next step it was the plotting and as we could see there are nodes that are just connected to the central node, these nodes are going to be quite difficult to evaluate because of the lack of edges to where they are connected, and besides that, they might be assigned to a community that is not precisely their own, despite this we can see that in the plot there are some agglomerations that advise us remarkable communities, nevertheless, it is necessary to make the community detection to be sure that it is that way.

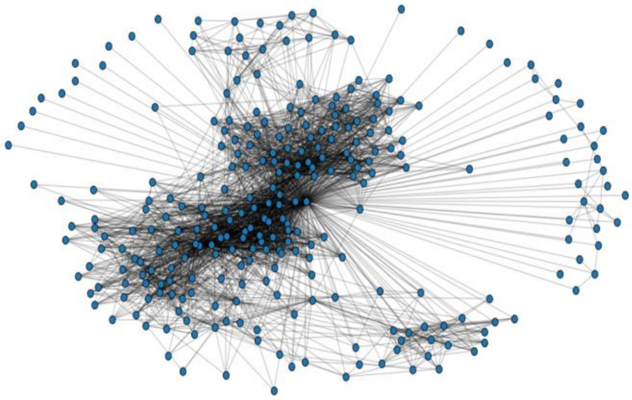

Figure 4. Social Network Graph.

To make the community detection, I used the greedy algorithm. This algorithm consists of getting the maximum modularity, by first spreading different communities to each of the nodes, and then by joining them and comparing either if the modularity increases or decreases the community starts forming and making the maximization of itself every iteration. In the case of the function proposed by NetworkX, they use the Clauset-Newman-Moore greedy modularity maximization function proposed in 2004. [1]

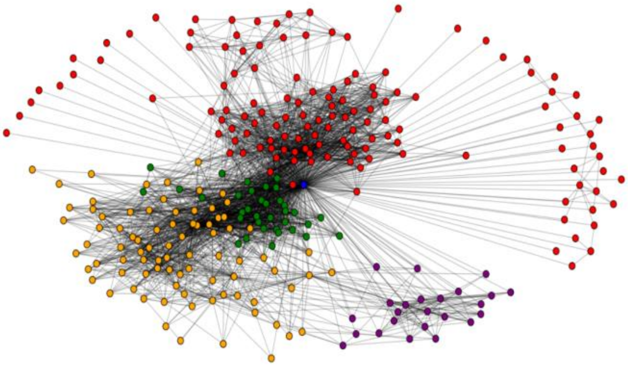

Figure 5. Communties of the graph.

As we can see we were able to get 4 different communities among the graph and the blue node in the center is my node and I left it as blue to differ from the other nodes as I consider it belongs to all the communities as it is connected to all the nodes so for the link prediction my node it would not be tested as theire is not node which is not connected to mine, my forecast was to have 3 or 4 different communities because my friend list is most conformed by three groups one is middle school friends, high school friends and university friends but I forgot about the friends I met playing xbox those are the pruple ones, nevertheless, once the community detection was done we can clearly, thanks to the different colores, where are the communities located. Once we have the plot, we can start making the link prediction.

For the further plots I will be using the colors to make the segmentation of the communities, because it can be easier to understand the concept of link prediction and, and how would it work in a real-life example like the one we are analyzing.

VI. Link prediction

The link prediction in a task was given a set of nodes and edges, we can predict, according to some metrics, and the characteristics of the network, a link between two nodes. This makes it an important approach, because most of the time we cannot expect that our network is complete, nevertheless with link prediction, we can understand that a bigger scope can be predicted from the graph, and that among them, some characteristics make likely to this connection to happen, independently from the scheme of the data.

Link prediction has been a task that has gained popularity thanks to the wide range of applications that can be provided, this because it is close to social networks, where people interact, and we can have behavior on people, besides we can make use of this task in bioinformatics, to predict the protein-protein interactions. And in geographical data, where we can have an idea of where the terrorist groups are.

As you might see, link prediction is a wide topic, with different measurements, and characteristics, nevertheless it is important to mention that in this project we will approach our social network data, from here we can expect to have a link prediction system able to determine the linkage between the nodes and to be corresponding to the measurements and characteristics. Therefore, the link prediction in this network is how likely the pairs of my friends who has not friendship become friends.

For the first instance I used the resource allocation index algorithm for the graph, I chose this algorithm since in consider the nodes with a high degree to influence less in the final result, in this case, we were also able to consider the community detection in the graph since we used the version of the algorithm that considers communities which is resource allocation index using community information, by making so we got an increase in the number of edges, originally, there were originally 2981, after the link prediction, we had to have a threshold, so .3 was defined as the threshold for this project as a test because is not a large threshold nor small, I expected that around 500 pairs would pass the threshold considering the sizes of the communities and the degree distribution, and after doing the link prediction, it was obtained that 623 pairs pass the threshold meaning that it can become new edges. In the image below you can observe those 623 new edges in the graph without the original edges.

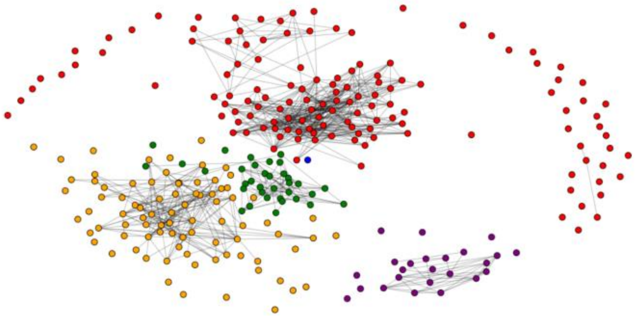

Figure 6. Link prediction using algorithm.

In the graph below we can observe in different edges once they are added in the original graph, it is important to remember that there are only going to be added the ones that pass the threshold, we can see that the connections are between the different communities, and it is difficult to find anyone that mixes the communities.

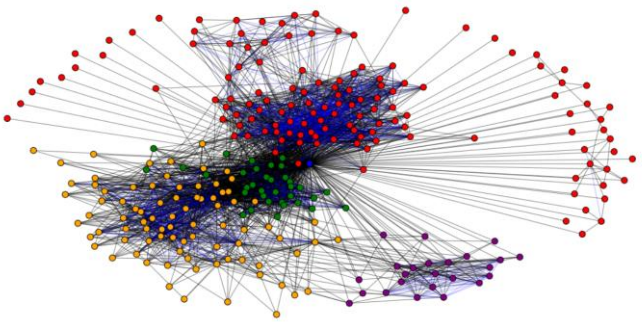

Figure 7. Graph after adding new edges.

The metrics like the density of the graph, and the average shortest path stay similar but the only improvement is in the clustering coefficient, this because after making the lin prediction and adding the edges the Cluster coefficient was 0.64, before making the link prediction and adding the edges, we had .533, which is more than 0.1, it was expected to have some improvement in this area because there are more links, and those new links are supposed to be in the places where the communities are, so it made a lot of sense when having a higher Cluster Coefficient. And another metrics that got larger a little was the average degree that was 27 meaning that the some nodes have more mutual friends.

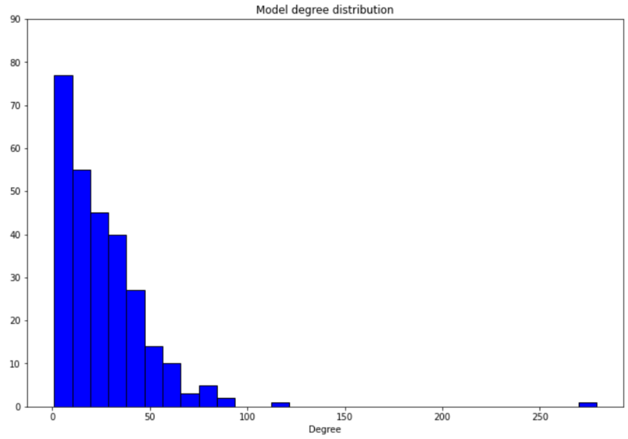

Figure 8.New degree distribution.

What most changes was the degree distribution as the nodes with a small number of mutual friends got a higher degree as the nodes with a degree lesser than 50 got more mutual friends and the number of nodes with only having a mutual friends quantity around 0 decreases.

For the second instance of the Link prediction, I decided to try it out the same algorithm (resource index allocation) only without including the community detection and also use the same treshold, as a result, I expected to have more edges because there was not going to be any limitation from the side of the communities, and I also expected that there would be links among nodes of different communities, something that almost did not happened in the prior algorithm, after running the prediction, it was obtained that 994 pairs passed the treshold meaning 994 pairs are likely to become friends, therefore; in total, in this case the graph has a total of 3975 edges, which is more than the one we did with the community detection.

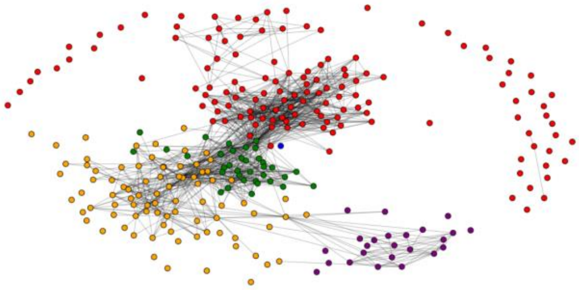

Figure 9. Graph of only new edges.

As you can see in image above, there are a lot edges among nodes of different communites, this is the biggest difference between an algorith, that itdoes not consider communities and an algorithm that it does.

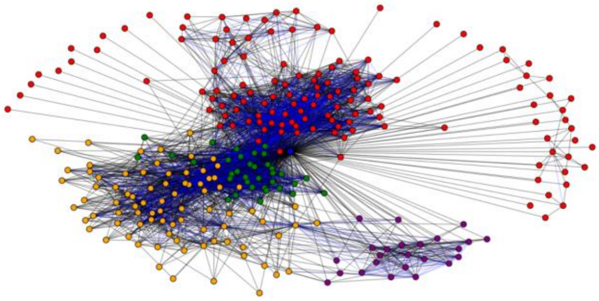

Figure 10. Graph after adding new edges.

As we can see there are more areas linked, the blue edges are the new edges. Finally when talking about the general characteristics of the graph we got an increment of more than .1 in the Cluster coefficient, as we did in the prior Link Prediciton, the other stat stayed very likely. Even the average degree only vary two decimal, however, there are not 28.2 friends so it is rounded to 28 so is the same average degree.

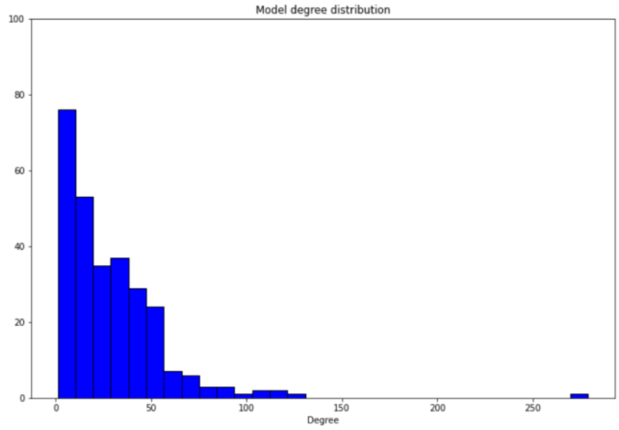

Figure 11. New degree distribution.

What is changes more comparing the two algorithms is the degree distribution, something I did not expect, the main differen is that appeared more nodes with a degree between 100 and 150, this means that nodes with a large number of mutual friends it gets even larger, and also nodes with few friends it got more as the first bin in the original graph it has more than 90 nodes and in this time it has fewer than 80.

VII. Conclusion

Along with the project I was able to put into practice different techniques which helped me to have better analysis and to be much more clearer when comparing the two different link prediction methods, by noticing the different metrics at different instances of the analysis, we can see a clear advance from the raw graph, where we did not have any of the community detection represented in the graph, neither the new edges of the graph, to then have the community detection plotted and well analyzed, to reach the point where we could do the Link Prediction either taking advantage of the community detection or not.

I believe that understand the different topics and how they can be merged into one project it is very beneficial for the study, the complexity of the study might be upgraded but in exchange, we have a whole analysis full of enrichment predictions and important insights. I was able to accomplish all the different objectives mention at the beginning of the paper, and in addition to it, the magnificent visualizations make much clearer to understand the different approaches and the objective of the analysis.

VIII. References

[1] Clauset, A., Newman, M. E., & Moore, C. “Finding community structure in very large networks.” Physical Review E 70(6), 2004.

[2] Dong, L., 2021. The Algorithm of Link Prediction on Social Network. [online] hindawi. Available at: <https://www.hindawi.com/journals/mpe/2013/125123/> [Accessed 3 August 2021].

[3] Jeremy, R. and Cooke, E., 2021. Link prediction and link detection in sequences of large social networks using temporal and local metrics.. [online] Core.ac.uk. Available at: <https://core.ac.uk/download/pdf/232195981.pdf> [Accessed 3 August 2021].

[4] JOSHI, P., 2021. Link Prediction | Link Prediction in Social Networks. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2020/01/link-prediction-how-to-predict-your-future-connections-on-facebook/> [Accessed 3 August 2021].

[5] Soundarajan, Sucheta & Hopcroft, John. (2012). Using community information to improve the precision of link prediction methods. WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web Companion. 10.1145/2187980.2188150.

[6] Daud, Nur & Ab hamid, Siti hafizah & Saadoon, Muntadher & Sahran, Firdaus & Anuar, Nor. (2020). Applications of link prediction in social networks: A review. Journal of Network and Computer Applications. 166. 102716. 10.1016/j.jnca.2020.102716.

[7] Wang, Peng & Xu, Baowen & Wu, Yurong & Zhou, Xiaoyu. (2014). Link Prediction in Social Networks: the State-of-the-Art. Science China Information Sciences. 58. 10.1007/s11432-014-5237-y.

[8] Liu, Shuxin & Ji, Xinsheng & Liu, Caixia & Bai, Yi. (2017). Extended resource allocation index for link prediction of complex network.