Comparing Different Community Detection and Data Analysis

Kevin Liu
Ontario Tech University
yenhung.liu@ontariotechu.net

Yujie Dong
Ontario Tech University
yujie.dong@ontariotechu.net

ABSTRACT

Community detection in the network is a world-class problem. In this article, we utilized five methods to separate the system into different communities and evaluate the final results. Before this step, we also analyze the data set. The data set includes American football, politics, karate, and Facebook. American football is about the situation that teams play with others. Politics is about the situation that the same customer buys different politics books. Karate relates to the relationships among the manager and instructor of the club and the members. Degree, betweenness, modularity, modularity with betweenness are utilized to analyze the data sets and understand the data sets with real meaning and find different combinations that fit different data sets. GN is executed to do community detection and methods Louvain, Clauset et al. SemiSyn-LPA, and Asyn-LPA are also executed. By comparison, modularity can be a criterion to assess community situation. Still, compared to Label propagation Algorithms(LPA)based algorithms, it is excellent enough in the performance of detecting overlapping communities and execution time.

KEYWORDS

Social Network Analysis, Community Detection, Overlapping Community, Modularity, Label Propagation Algorithm

ACM Reference Format:
Kevin Liu and Yujie Dong. 2020. Comparing Different Community Detection and Data Analysis. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In data science, the dataset is usually a vast data, with a large volume of data. Meanwhile, there is a connection existing among different data, such as social networks. When social networks are under analysis, how to discover different communities in social networks is always a problem. From the previous study, the various researchers have put forward different methods to categorize different communities and judge them by different means, such as the value of modularity.

From the previous study, different research usually utilizes different datasets. Hence, our objective is to detect the community in the same datasets and try to understand the disadvantages and advantages brought by various methods(betweenness and closeness), which can make a detailed analysis. As for analyzing community detection methods, modularity and other possible assessment methods can be applied.

The rest of the paper is organized as follows. Section II describes the related research in the correlated area, and the paper we collected in the very beginning. Section III is about analyzing the five different datasets which contain three small datasets and one large-scaled dataset. Those datasets' community structures are known in the previous research. We use Gephi to calculate the attributes, such as betweenness centrality and closeness centrality and using that information to present the understandable insight. Section IV describes the evaluation of different community detection algorithms. We compare the predicted number of communities with the ground-truth datasets and evaluate the performance on detecting overlapping communities and execution time. Section V is the discussion for five different community detection algorithms and the possible improvement. Section VI is the conclusion of the whole article.

2 RELATED WORK

This section will give a brief introduction to the eight papers we collected at the beginning. Each paper has a different topic, but it is mainly related to community detection. In addition to the brief introduction, we will also discuss the ideas we got from these papers.

In the paper[9], they propose a method based on the most common community-structure function, Modularity. Community detection is one of the famous research areas in social network analysis. The previous work about community detection can be divided into two perspectives, first is graph partitioning, second is community structure detection. In some points, these two methods are addressing the same problem. Still, graph partitioning mainly works on the small dataset or known dataset, and community structure detection primarily works on the large dataset or unknown datasets, such as social networks or web network. However, the method proposed by this paper is a combination of graph partitioning and community structure detection. It transforms modularity into an eigenvector-based matrix and uses a spectral algorithm to detect the community. In the experiment stage, they test their community detection method with the famous dataset from the previous research, such as Karate club, Jazz musicians, etc. They evaluate their performance by using the modularity value for each result. Modularity is a numerical indicator used to assess specific network areas how good a particular division is. The larger the modularity value, the stronger the community structure in the network. The performance of the proposed method is better than the base-line way, Girvan and Newman and Clauset et al., and similar to the Duch and Arenas. The speed of the computation time also outperforms the base-line method. The result also shows up the better result, especially on the vast scale network. Community detection in large

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

scale networks has become a popular research area because of Big Data. Although the graph partitioning and community structure detection are the different directions in the previous research, the combination of these two areas can create a powerful and effective method. The author also points out that the way they proposed can be the best in large scale networks.

In the paper[10], they explain the challenge in previous research on finding the overlapping community and also proposed a novel community detection method based on k-clique to detect the overlapping community. When we talk about the community in the network, we can think as a few well-connected groups of nodes, and each group has less connection between each other. In the previous research, the existing community detection method can competently perform the detection and divide the network in the separate part. However, if the community contains more small communities inside, the previous method is hard to detect those overlapping communities. For example, if we analyze the ego network from a scientist, we may be able to use the community detection method to divide his system into several parts, such as family-based groups, schoolmates-based groups, and science-based groups. Nonetheless, in the science-based group, there are overlapping communities that are based on different research areas. The previous community detection is hard to detect those overlapping networks. The proposed method in this paper combines the technique called 'k-clique.' K-clique is a fully connected subnetwork, and the size is k. In the experiment stage, we observe that the k value will be different depending on the dataset. Therefore, when analyzing the network, we need to have a certain degree of understanding of the datasets and fine-tune the k value according to different datasets. The experiment process also makes us aware that it is challenging to have a uniform and standardized community detection method.

In the paper[7], they focus on discovering the social circle in the social network. Our social network is complex and will grow over time, and there is currently no unified way to unify this type of network. We can discover the social circles in the social network by different attributes, such as family or the same school. Using social circles can help users filter the content and protect their privacy. Previous methods for discovering the social circle mainly relied on manuals. The method proposed by the author based on machine learning can automatically detect social circles. It considers not only the node features but also graph structure and is unsupervised learning. The social networks they choose are from Twitter, Google+, and Facebook and are hand-labelled ground truth data to evaluate the performance of their methods. During the experiment stage, their method performs better than the base-line methods on Balance Error Rate and F1 Score. Comparing the performance on three different datasets, most of the methods perform better on Facebook data. The reason can be Facebook dataset are fully labeled, and compared to other dataset facebook supplies more comprehensive information about the user. The different attributes of the network can also be the reason. The edges in the Facebook are undirected and in Google+ and Twitter are directed because of the follower relationships. From this paper, we can see that the social circle obtained on Facebook can be more informative. Social networks can affect many perspectives; that's why many advertisements will actively buy ads on Facebook and how the well-known Cambridge Analytica used social network analysis to influence the US presidential election.

In the paper[12], they define 260 ground-truth datasets to evaluate the community detection methods. They also propose a new community detection method, which can detect all the members in the community and the communities belonging to a single node. Evaluation of the community detection function is a difficult problem because of the lack of ground-truth datasets. Thus, the authors define the ground-truth datasets from multiple areas, such as interest-based groups, scientific collaboration networks, etc. The community detection methods they evaluated in this paper based on 13 structural definitions, such as Conductance, Triad Participation Ratio, etc. In the experiment stage, they observe that if the network has a less overlapping community, the Conductance-based method performs the best. If the network has a more overlapping community, the Triad Participation Ration based method performs the best. Thus, we can understand that to get a better result. We need to understand the dataset first and then choose proper community detection methods. Based on that 13 structural definitions, they also proposed the new community detection, which can detect the community from a single node. The method contains the PageRank-Nibble algorithm to give the surrounding node values and calculate the community scoring function according to the collected nodes. When the scoring function reaches the local optima, it means they found a community. The more local optima, the more communities there are. However, the performance for their method only performs well on detect the node belongs to one community if the node belongs to multiple communities the performance drop.

In the paper[15], the node features can be utilized in detecting communities in Networks. There are many community detection methods used in the network. The joint community detection criterion is based on the node belonging to the two communities so that these two communities can be connected more easily. Different labels are assigned here, and the evaluation of detection criterion and estimation are based on previous labels. We can utilize fixed weight refreshing the label"i" and reckoning the whole standard to optimize over label assignment. It also showed the influence from the parameter in the process of estimation. Fixed label assignments have been utilized to optimize over weights.The JCDC standard is not based on model, but it is slowly contained consistency. This user can hope the standardized node characteristics and discard some special units to reckon the relevant similarity. A simulation research has been carried out to check the result. Different methods have been applied to measure different functionalities.

These methods can be applied to data collection related to world trade network. We can categorize the world into three characteristics: African continental, the countries which can have structure situation. Therefore, we can also compare all the methods with each other. The second data collection we chose is from the New England Company consisting of 71 lawyers. It has provided seven functions of the nodes: status, gender, the position of the office room, years with the firm, age, practice, and law school attended. When we applied the fixed weight into the testing standard, it will improve the testing efficiency and also utilize the adjacent matrix or the elements of nodes. The data collection can be categorized by all the details we have mentioned here. So here, we can see different node features are embodied in the picture.

In the paper[13], node attributes can also be applied into community detection in the network. The algorithm is the basis we can discover the principle of the organization. When we detect the community, there are two possibilities we can utilize we can use the source of information: the network features, and the features and attributes of nodes. Sometimes even around the nodes which have the same edges and the same characteristics, the algorithm also only focuses on these two kinds of data. In this paper, they have developed the structure of the edge and attribute of nodes to detect the overlapping, extendable community with node attribute. To discover the network community can be treated as a group of nodes concentrating in the community, and one node can belong to multiple communities. Because the nodes in the communities have many shared features and many relationships among them, two sources of data can be used in executing the communities related tasks. Until now, how to detect these attributes has been solved. This paper has introduced the structure of edges and the node attributes. It is based on the attributes of the model that has the network of the nodes. Firstly, our model can detect the node to duplicate the qualification of membership so that we can avoid no connection erected among several communities. Secondly, compared with the previous work, we can build a reliable relationship between the network and the attribute. Thirdly, for adapting this model and finding the communities, this paper has introduced the adapting model. Therefore they can find the communities, and develop the method called block-coordinate ascent method. CESNA is the first community detection way to build the model based on two shared nodes, membership identities, communities and attributes. However, CESNA can detect duplication, non-duplication, hierarchical network in the community, two-node attribute, and the structure of the graph. The paper has quantized the accuracy of CESNA through six live social, information, and content sharing networkS: Facebook, Google+, Twitter, Wikipeida and Flicker. LDA is the means to find some similar "subject" attribute and the node between the links. After these analysis, we can see the model found has several characteristics. The first is that the node belongs to the same communities should connect. The communities may be duplicated, because every node can belong to several communities. If two nodes belong to several communities, they can be connected and is situated in a collective community. The nodes in the same communities may share some attributes.

In the paper[6], the study about the complex network is a critical twork to technology network, from social networks to information networks. The connected nodes connect the complex network With the increase of the size of the network. This paper has shown that one of the most critical questions to check the information of the communities is to detect the distribution among different subjects. This paper utilizes the analysis of quantity through complex websites, which has contained all the related articles about the web of science. Here, the authors have used cite space to do some visual investigation. So, they have confirmed the most influential, the most important and the most activated node, which uses scientific quantitive analysis. Modularity is one of the most popularized technologies, but how to optimize the modularity is a NP-hard problem. The density of intra-community links has been introduced to solve the structure of communities about network. The abbreviation of the frequency is D. If D has a higher value, it has a better distribution of the area. There are also some other traditional community detection technology that has been invented. The first method is called graph partitioning, and the center of the graph is much more concentrating than others around it. The second method is hierarchical clustering. These graphs can contain many small clusters, which are situated on different levels. Under this circumstance, hierarchical skills can be applied to using recognize hierarchical skills. It can be divided into two different algorithms:(1) agglomerative algorithms (2) divisive algorithms. The third method is called a spectral clustering. The fourth method is a decisive method, which is based on low similarity to eliminate the edge between different groups to seperate one community with the other. What have mentioned above is the traditional ways. There are also some community detection technologies based on the optimization of the network community. The first is a called greedy technique, which is composed of three different branches in algorithms. The first is Greedy method of Newman. The second is Fast Greedy algorithm. The third is Blondal's Louvain algorithm. After applying the greedy technologies, the second step is to simulated anneling. The third is extremal optimization. The extremal ooptimization is to focus on the optimization of part of the variables. It can assign the adaptability for every node. Adaptation is to acquire the degree of modularity through grabbing part of the nodes. The fourth is spectral optimization. The fifth algorithm is an evolutionary algorithm. The evolutionary algorithm is one kind based on artificial inspiration. How to detect the overlapping community detection technology is a current task that cannot be solved by traditional technology. Some methods can solve the problem related to dynamic communities. The first is "Potts model". The second method is a random walk, which we can choose a moving-assistant machine entering this community from one node and walk into the next node with a random choice because its high density of node in the community and multiple routes. The third method is the diffusion community. The fourth method is synchronization. Synchronization is a new phenomenon and has received interest from a different area.

In the paper[8], community structure is still a very significant property to distinguish the essence of the complex network. Though it is substantial, it is still an unsolved problem. Sixteen different types of the network have been presented in this paper. Meanwhile, it also compares our distribution way with Infomap, LPA, Fastgreedy, and Walktrap. The structure of a community can reveal the vast amount of concealed messages related to complex networks. The world does not understand the sysgtem in the real, and colossal research has been carried out to reveal these structures in the real world network to be more capable of managing, maintaining, updating, and changing the community. Now, most of the methods are focused on detecting the node community. The way that has been popularized is based on optimizing the modularity. Some means are to enforce every node to be assigned to every community. This method does not always reflect the network in the real world and exists with the duplicated community. For example, in the social network, one can have a family relationship community, working for community friends, community , and hobbies communities. Recently, the detection of linked communities has been sent. The concept of connected community is useful for the overlapping community, because more edges mean unique the identity is while

the node also has multiple characters. The result is that this paper has tested the property of our methods in the synthetic and realistic network. We have analyzed 16 different types of networks. There are four networks we need to analyze, which contain synthetic network, real-world network, biological network. As for the synthetic network based on LFR based network, 128 nodes, 597 edges, eight communities and ten overlapping nodes have been contained here. As for the real-world networks, the Zachary karate club is the famous experienced network, and the other widely used system is American college football network. The Facebook network is direction-fixed users. As for the biological network, the E. Coli transcriptional regulatory consists of 29 sub-network with broken connection and five isolated sub-network. Compared with other methods, we have tested our functional purposes. We will compare this distribution with four different popular algorithms: Infomap, LPA, Pastgreedy, and Walktrap. When we have a discussion, we can find the many algorithms owning their limitations. This topology has provided a refreshed and simple frame to detect the network of the communities.

All in all, community detection has always been a challenging problem in social network analysis. Because of the difference in data set size and attributes, different community detection methods may not always provide the best solution. Especially in large scale networks, there are often smaller and overlapping communities within large communities. It is more difficult to detect these overlapping small networks than to detect large populations. Besides, how to evaluate the pros and cons of a community detection method has always lacked a standard practice. After all, community detection is a prediction. Still, in the paper[12] they proposed to use ground-truth datasets to evaluate the performance of community detection, which provides a reasonable evaluation method. However, according to the experimental results of the above paper, we can also see that when community detection is required, understanding the content and attributes of a data set, using suitable algorithms and fine-tuning are the best methods for community analysis.

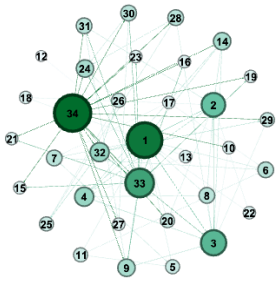
3 DATASET ANALYSIS

3.1 Karate[14]

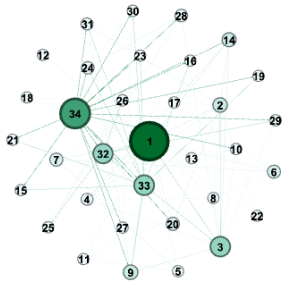
At an American university in the 1970s, 34 members of a karate club build a social network like this and generate related data collection. Because a disagreement happened among the administrators and coaches and lead to some members are taken by these coaches, Zachary has utilized various ways to assess the strength of connection among members and apply some algorithms to stabilize the relationship. Node one and node 34 are the instructor and the manager of the club.

Attribute	Data(All)
Node	34
Edge	78
AvgDegree	2.294
GraphDensity	0.139
Diameter	5
AvgPathLength	2.408
ClusteringCoefficient	0.588

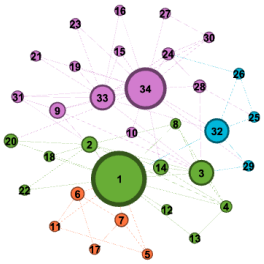
Ranked by degree, we can see the analysis from the picture below.



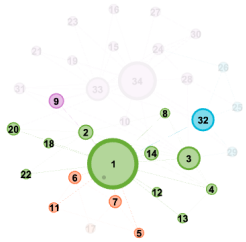
Ranked by betweenness centrality, we can see the analysis from the picture below.

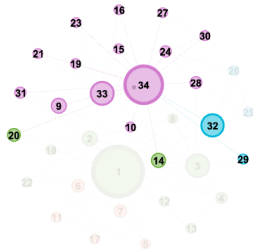


After this data set was modularized, we can see more clearly about how node one and node 34 connect with other groups in the club members. The club members have been categorized into 4 groups.



To find how many groups node1 and node34 are connected to.



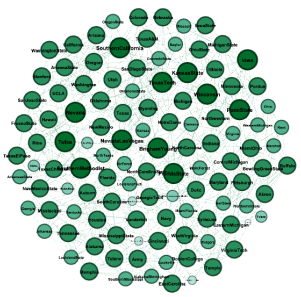


Node 34 has less connection with other groups compared to node 1. It is suggested that node 34 should be connected with node 1, which means the manager should be connected to the administrator. Meanwhile, because node one and node 34 disagree with each other, through node 32, which are both friends with node one and node 34, we can try to let node one and node 34 become friends so that it can make the relationships among members more stable.

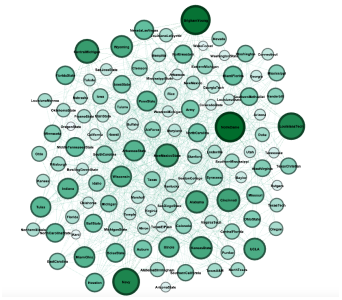
3.2 American football[5]

This data set is the Network of American football games between Divison IA colleges during regular season Fall 2000. We can see table 1: Football. Any information else can be seen in table 1.

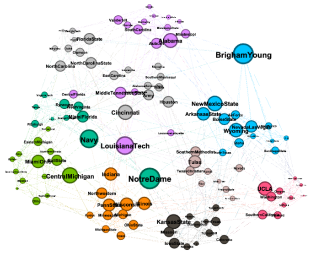
Attribute	Data(All)
Node	115
Edge	613
AvgDegree	10.661
GraphDensity	0.094
Diameter	4
AvgPathLength	2.508
ClusteringCoefficient	0.113



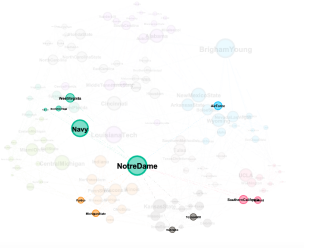
Ranked by degree, we can see the analysis result from the picture above. Many teams can play a similar amount of game, and the range of the amount of number is 7-12.



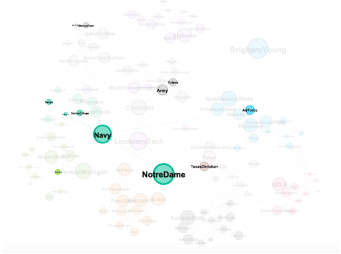
Ranked by betweenness centrality, we can see the analysis result from the picture above. We can find from the size of the node that BrighamYoung,Navy, LouisianaTech and NotreDame are the biggest.



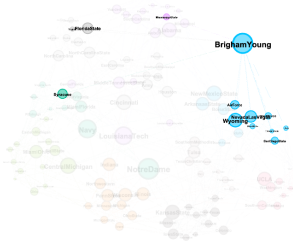
After modularity, these teams can be categorized into seven conferences.



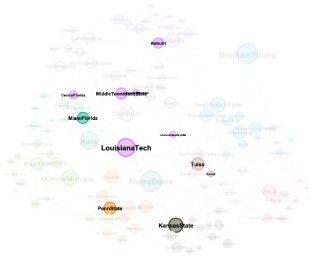
This team1 NotreDame is connected with four other conferences.



This team2 Navy is connected with four other conferences.



This team BrighamYoung is connected with three other conferences.



This team LouisianaTech is connected with 4 other conferences from a different conference.

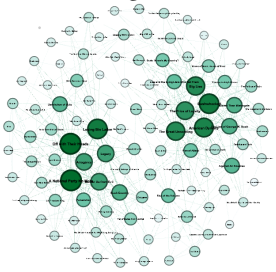
In conclusion, these teams with the higher score in betweenness means they have a higher connection with more communities. Actually, these teams are in different conferences so they will have more chance to be the "bridge" with other organizations.

3.3 Politics[1]

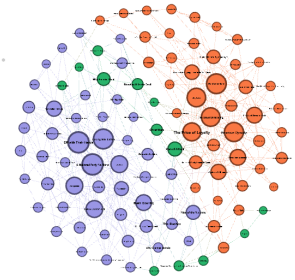
This paper is about politics books. This data set is about books related to America politics. The edges between books show that the same buyer often buys them all. The books can be classified into liberal books, conservative books, and neural books.

Attribute	Data(All)
Node	105
Edge	441
AvgDegree	8.4
GraphDensity	0.081
Diameter	7
AvgPathLength	3.079
ClusteringCoefficient	0.488

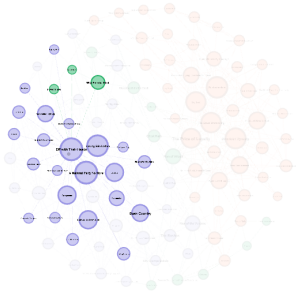
Ranked by degree, we can see the analysis result from the picture below. From this picture, we can find some books which are bought most.



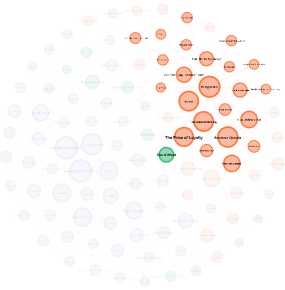
Ranked by political value and degree, we can see the analysis result from the image below. The value represents three labels including conservative, liberal, and neural.



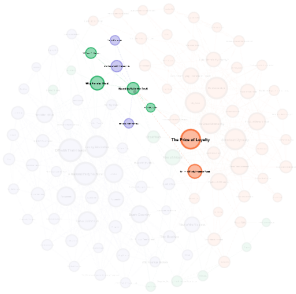
Finding the most popular book in conservative books, we can see that the people who read the most popular conservative would like to learn more other conservative books and have the possibility to understand the neural book.

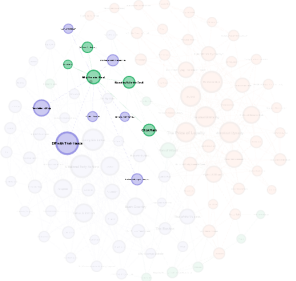


Finding the most popular book in liberal books, we can see that the people who read the most popular liberal book would like to learn more other liberal books and seldom read the neural book.



Finding the most popular book in neural books, we can see the analysis result from the picture below. We can find that the reader who read the neural books would like to read the books from the other two different parts.





3.4 Facebook Ego Network[7]

Finally, we introduce our largest dataset, Facebook Ego Network. This dataset is from paper[7]. Community detection on large networks has always been a challenge. In addition to requiring a large amount of computation time, the lack of large-scale networks with ground-truth data is also a significant challenge for community detection. Therefore, to evaluate the performance of community detection on large-scale networks, they created a dataset with a ground-truth label. They manually label the social circle in the network with the 26 categories, such as school, clubs, hometowns, etc. These datasets constitute ten ego-network which means they collected ten people's Facebook social network, the total number of users is 4039, and the number of different social circles is 183.

Attribute	Facebook(All)	Individual 1	Individual 2
Node	4039	333	1034
Edge	88234	5038	53498
AvgDegree	43.691	30.258	103.478
GraphDensity	0.011	0.091	0.1
Diameter	8	11	9
AvgPathLength	3.693	3.752	2.952
ClusteringCoefficient	0.617	0.557	0.534

The table above shows the network information based on the total 10 Ego Networks and two different individual's ego networks. If we take a look at Facebook(All), although our network has large numbers of nodes and edges, the diameter of the graph is 8, and the average path length is 3.693. The clustering coefficient is 0.617, which indicates people have a higher chance of meeting people if they have common friends. Combining our diameter and clustering coefficient value, we can know that our data set perfectly fits the small-world model perfectly fit the small-world model, which is high clustering coefficient and low diameter. The degree distribution shown in Figure 1 is a power-law distribution, which means that a few nodes occupy a large number of edges. We think the result is reasonable, because in real life a few people have the most relationships, such as celebrities, so Facebook is an online version of social network, I think it is reasonable to have power-law distribution. In individual 2's network, the average degree is 103.478 which is way higher than the other networks, so the average path length is lower than other networks. Thus, we can assume that people in individual 2 network have more connections with each other.

4 EVALUATION OF COMMUNITY DETECTION

In this section, we will use five common community detection algorithms, which are well-known Girvan-Newman(GN)[5], two modularity optimization based methods and two different label propagation algorithms to detect the community structure in the real datasets. We did not choose a computer-generated network for the dataset because it cannot represent real-world conditions. On the contrary, we used the common datasets in the previous research for analysis. The structure of these datasets is known and has been verified so that it can be used as our ground-truth datasets.

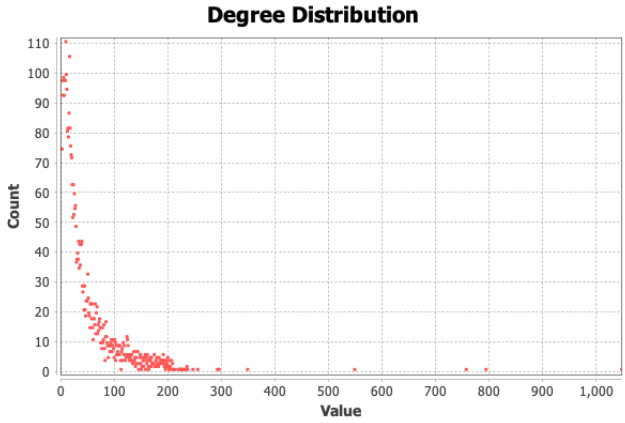


Figure 1: Facebook Ego networks(All)

4.1 Community Detection Algorithm

Girvan-Newman algorithm(GN)[5] is a community detection method based on the betweenness centrality. The general idea is first to calculate the betweenness value of each edge, and gradually remove the edge from large to small. Because an edge with a high betweenness value is usually a bridge connecting two different communities, by removing these edges, it can help us divide the network into different communities. But there are some drawbacks to this method. First, it can only use on the undirected and unweighted graph. Second, the computation time is too slow if we use the large-scale network.

One effective community detection method can be modularity optimization. The general idea is that modularity is an indicator for analyzing community detection. Generally, the higher the value of modularity, the stronger the community structure in the network. We used two methods based on modularity optimization, one is greedy algorithm from Clauset et al.[3], second is from Louvain algorithm[2]. Both methods outperform other modularity optimization algorithms in the speed of computation even on the large-scale dataset. However, modularity based method does not effectively detect overlapping communities

Label propagation Algorithms(LPA) is a community detection method based on the diffusion of information. The general idea is that each node will be assigned a label at the beginning, and for each iteration, the label of each node will be changed to a new label, depending on the labels of the most neighbours around. The LPAs we use in this paper is the semi-synchronous based method from [4] and asynchronous based method from [11].

4.2 Zachary's karate club

The first dataset we used is the well-known Karate club[14]. We used five different community detection algorithms we talked above to calculate the modularity and the number of detected communities. Table 1 shows the experiment result for our algorithms. If we use modularity as a measure of the success of community detection analysis, all algorithms detect the existence of at least three communities. However, if we compare the real community structure, there are only two communities in this network. The result means that

Table 1: Evaluation on Zachary's karate club

Method	Modualrity Q	Community
GN	0.36	2
GN	0.40	5
Louvain	0.37	2
Louvain	0.42	4
Clauset et al	0.38	3
SemiSyn-LPA	0.33	3
Asyn-LPA	0.39	5

even if the modularity is large, the detected communities may not necessarily be the same as the real community structure. We still have to understand other attributes of the dataset, such as Node degree. According to Node degree, Node 1 and Node 34 have the largest Node degree, respectively, so for other people, these two people have the greatest influence on the entire network. Then we can study the properties of its members again and understand that these two people are instructors. Therefore, although the modularity value will drop, we fine-tune our algorithm GN and Louvain algorithm to only detect two communities. The detected community members by Louvain algorithm predict one node wrong, which is node 10, and GN only predicts one node wrong, which is node 3. The other algorithms predict more than two communities, so we combine those communities into two communities and then test the accuracy. The Clauset et al. algorithm and SemiSyn-LPA only predict one node wrong, which is node 3, and the Asyn-LPA's prediction perfectly fits the ground-truth dataset.

4.3 American College football

The second dataset we used is American college football[5]. A node represents the team in the American colleges. An edge represents they have a game with each other. Thus, the graph for our network is undirected and unweighted. Table 2 shows the performance of our five different community detection methods. Louvain algorithm, Clauset et al. , SemiSyn-LPA and Asyn-LPA perform pretty well on modularity values, and the number of detected communities is 23. However, if we compare the actual community structure, there are only 12 communities in the network. This means that even a high modularity value does not necessarily represent the accuracy of predicting the community structure. This result is similar to the result of the Karate Club dataset. Then we observed the performance of the GN algorithm. We found that the highest modularity value under the GN algorithm is 0.599, and the number of detected communities is only 10. I think this result is because the GN algorithm relies on deleting the Edges with the highest Betweenness value. However, such a method requires fewer links between communities to show sparse between each community. However, in the football dataset, the number of edges of each node is very close, and there are many connections between communities, because there will be competition between conference and conference, so the performance of the GN algorithm is not good.

Table 2: Evaluation on American College football

Method	Modualrity Q	Community
GN	0.599	10
GN	0.597	12
Louvain	0.60	12
Louvain	0.92	23
Clauset et al	0.90	23
SemiSyn-LPA	0.92	23
Asyn-LPA	0.92	23

Table 3: Evaluation on US Politics Books network

Method	Modualrity Q	Community
GN	0.48	3
GN	0.51	5
Louvain	0.49	3
Louvain	0.52	4
Clauset et al	0.50	4
SemiSyn-LPA	0.48	4
Asyn-LPA	0.48	4

4.4 US Politics Books network

The third dataset we used is the US Politics Books network[1]. A node represents books about American politics. A edge represents that customers who bought this book will also buy other books. Thus, the graph for our network is undirected and unweighted. Those data have been marked as three different labels, liberal, neutral, or conservative. We apply the five community detection algorithms we talked about above to detect the community. Table 3 shows the performance of our five different community detection methods. Louvain algorithm and GN perform the best on modularity value. However, the number of communities detected by each algorithm exceeds the number of known communities. Thus, we do the fine-tuning on our GN and Louvain algorithm to make them divide the network into only three communities. We compare the prediction result with the ground-truth data. In the Louvain algorithm, there are 13 wrong predictions in a total of 105 nodes. The accuracy rate is 0.81. However, the prediction result in GN is pretty bad compared to the ground-truth dataset. We can not even calculate the accuracy rate because we can not identify the politics community with the ground-truth dataset.

4.5 Facebook ego network

The datasets described above are all small networks, so almost all previous community detection methods can perform well. But detecting the community structure in a large network is always a problem, so in the end, we use the largest dataset of all our datasets, Facebook Ego Network[7]. Node represents the user. Edge represents the friendship connection in Facebook. Thus, the graph of this network is undirected and unweighted. Table 4 shows the performance of the five algorithms in the small Facebook individual ego network. Table 5 shows the performance and execution time of

Table 4: Evaluation on Facebook ego network(Individual 1)

Method	Modualrity Q	Community
GN	0.27	23
GN	0.416	36
Louvain	0.415	23
Louvain	0.46	16
Clauset et al	0.44	12
SemiSyn-LPA	0.401	21
Asyn-LPA	0.394	19

Table 5: Evaluation on Facebook ego network(All)

Method	Modualrity Q	Community	Excution Time
GN	-	-	-
Louvain	0.415	183	-
Louvain	0.83	16	5.6s
Clauset et al	0.77	13	21s
SemiSyn-LPA	0.73	44	0.00036s
Asyn-LPA	0.81	79	3.9s

the five algorithms in a large Facebook ego network (all) consisting of 10 people.

Let's start with the individual Facebook network. Although this network only has 333 nodes, it has a total of 23 communities(social circles) according to the ground-truth dataset. If you look at performance from the perspective of modularity optimization, the Louvain algorithm and Clauset et al. have the highest modularity value, 0.46 and 0.44 but the number of communities they detect is 16 and 12, which is quite different from the number of our real communities. Next, we observe the performance of SemiSyn-LPA and Asyn-LPA. Although the modularity value is not the highest, the number of communities they detect is similar to the number of real communities. Evaluating from the performance of these five algorithms, we may be able to assume that the Louvain algorithm and Clauset et al's way of dividing communities is based on common community definition which is that the community should have dense links inside, and fewer connections to other communities. And SemiSyn-LPA and Asyn-LPA may be able to detect overlapping communities more effectively than other algorithms. The overlapping communities can be described as two different communities that have high interaction with each other.

Next, let's discuss the large-scale network. We add up the number of communities in all personal Facebook networks to get a total of 183 communities. Then the detection community is based on five community detection algorithms, but GN is not suitable for our large-scale network. The reason is that GN gradually removes the edge with high betweenness value and preserves the community structure with the highest modularity value. However, the disadvantage of this algorithm is that the computation time for large-scale networks is too slow. According to our test, we ran the algorithm for two hours and did not get results. Then we evaluate our community detection method according to modularity. The Louvain algorithm and Asyn-LPA have the highest modularity value, 0.83 and 0.81, but the number of communities they detect has huge differences, 16 and 79, respectively. According to our experimental results, even if the dataset becomes larger, the number of communities detected by the Louvain algorithm and GN is similar to an individual Facebook network. However, Asyn-LPA and SemiSyn-LPA have detected a larger number of communities in response to the larger data sets. But even though the number of communities that can be detected has increased, there is still a gap with the number of real communities. We can think that even if Asyn-LPA and SemiSyn-LPA can effectively detect more communities, there still seems to be a quantity limit.

Finally, in terms of execution time, modularity based methods require more computation time than LPA based methods. I think the possible reason can be the difference between the algorithm stop mechanism. After all, most of the modularity based methods stop when it reaches the highest modularity value, but LPA based methods have more stop mechanisms, such as stop when the label of all nodes is not changing.

5 DISCUSSION

Based on the experiment results of the previous section, we can get some insight. We will discuss them below.

First is about using modularity as a criterion for judging community structure. Usually, the higher the modularity value, the better the community structure of this network. But in our experiment, when the modularity value is high, the number of detected communities is not the same as the number of real communities. So in addition to modularity, I think we should look for other judgment criteria because obviously, modularity does not necessarily apply to all situations

Second is about overlapping communities. According to the experiment results of the above section, the performance of the Louvain algorithm and GN to detect overlapping communities is not good, especially GN. The reason may be that the way they divide the network is based on the Sparsity between each community. Therefore, if the two communities have high interaction, it is difficult for the Louvain algorithm and GN to separate this network. However, in LPA based methods, the number of detected communities will also increase with the increased size of the dataset. In the experiment of our individual Facebook network, the number of detected communities is very similar to the real data. However, in the Facebook network (All), although the number of detected communities has increased, it is still a bit different from the ground-truth dataset. Thus, we assume that, in addition to analyzing the relationship between the node and the edge of the network, maybe we can incorporate other attributes, such as user's information, and convert it into a dimension space and then compare the similarity. Perhaps it is a possible solution to increase the performance of detecting overlapping communities.

Third is about execution time. Execution time is very important for any algorithm. A good algorithm can still have a good execution time, even with a large amount of data. In our experiments, we did not measure the execution time for small networks, because the time is swift. But in large-scale networks, there is a gap in execution time. According to experimental data, the operation time of the five

algorithms from fast to slow is SemiSyn-LPA, Asyn-LPA, Louvain algorithm, Clauset et al. and GN. Overall, LPA based methods are faster than modularity optimization-based methods.

6 CONCLUSION

In this article, we use the three well-known small datasets from previous research and one large-scale dataset as our ground-truth datasets. We start from performing the data analysis to get the insights from the ground-truth dataset. And then, we use five different community detection algorithms to detect the community structure and evaluate the performance with the ground-truth dataset. According to the experiment result, we understand that modularity can be a criterion to assess community structure, but it's not comprehensive enough for every situation. Label propagation Algorithms(LPA) based algorithms are better than modularity optimization-based algorithms in the performance of detecting overlapping communities and execution time. We also discuss the possible solution to improve the performance of the previous methods. However, all the network graphs we use in this paper are unweighted and undirected. Thus, future work can evaluate the different graph styles, such as directed and weighted graphs, and also different types of networks, such as biological networks. We hope that our research can contribute to this field and inspire more people to invest in relevant research.

REFERENCES

[1] Lada A Adamic and Natalie Glance. 2005. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*. 36–43.

[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

[3] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. 2004. Finding community structure in very large networks. *Physical review E* 70, 6 (2004), 066111.

[4] Gennaro Cordasco and Luisa Gargano. 2010. Community detection via semi-synchronous label propagation algorithms. In *2010 IEEE International Workshop on: Business Applications of Social Network Analysis (BASNA)*. IEEE, 1–8.

[5] Michelle Girvan and Mark EJ Newman. 2002. Community structure in social and biological networks. *Proceedings of the national academy of sciences* 99, 12 (2002), 7821–7826.

[6] Bisma S Khan and Muaz A Niazi. 2017. Network community detection: A review and visual survey. *arXiv preprint arXiv:1708.00977* (2017).

[7] Jure Leskovec and Julian J Mcauley. 2012. Learning to discover social circles in ego networks. In *Advances in neural information processing systems*. 539–547.

[8] Wei Liu, Matteo Pellegrini, and Xiaofan Wang. 2014. Detecting communities based on network topology. *Scientific reports* 4 (2014), 5739.

[9] Mark EJ Newman. 2006. Modularity and community structure in networks. *Proceedings of the national academy of sciences* 103, 23 (2006), 8577–8582.

[10] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *nature* 435, 7043 (2005), 814–818.

[11] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. 2007. Near linear time algorithm to detect community structures in large-scale networks. *Physical review E* 76, 3 (2007), 036106.

[12] Jaewon Yang and Jure Leskovec. 2015. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems* 42, 1 (2015), 181–213.

[13] Jaewon Yang, Julian McAuley, and Jure Leskovec. 2013. Community detection in networks with node attributes. In *2013 IEEE 13th International Conference on Data Mining*. IEEE, 1151–1156.

[14] Wayne W Zachary. 1977. An information flow model for conflict and fission in small groups. *Journal of anthropological research* 33, 4 (1977), 452–473.

[15] Yuan Zhang, Elizaveta Levina, Ji Zhu, et al. 2016. Community detection in networks with node features. *Electronic Journal of Statistics* 10, 2 (2016), 3153–3178.