The Network Analysis of Wikipedia Vote Network

COMP0123 Complex Network and Web
Coursework (2018-2019)
03/01/2019

QIQI ZENG
Student Number: 18070639
MSc Web Science and Big Data Analytics

Abstract:

Wikipedia is a multilingual, web-based and free online encyclopedia, which is created and edited by volunteers around the world. To maintain the order of the website and prevent certain persons from making disruptive edits, administrators will be elected via a public discussion and vote.

In order to explore the voting mechanism and better understand the current structure of voting distribution, this paper will do literature surveys and numerical researches on a network-related dataset. The research result reveals that there are different groups of voter-communities in the Wikipedia Request for Adminship (RfA) process, and it is possible for a candidate to vote for his opponents under a certain condition. These results may provide some practical insight for Wikipedia to adjust and update their voting mechanism in the future.

Content List:

1.Introduction:

Wikipedia is a large-scale, distributed and collaborative project, and it is changing how we think about the nature of work (Welser et al., 2011). To provide a large amount of information, Wikipedia applies an open model and allows the ordinary Wikipedia users to edit it. In the beginning, elite users contributed most of the work, however, beginning in 2004, the distribution of work shift from the elite to the common users (Kittur et al., 2006). By this model, Wikipedia improves its openness and get more opportunities to acquire various knowledges form different users. However, it also leads to a higher risk of being disruptive edited. According to Jankowski-Lorek et al. (2013), due to the lack of centralized supervision, it may result in an edit war when two editors have different opinion and each contributor tries to enforce their own version. Therefore, in order to avoid similar problems and maintain the order of the website, Wikipedia forms a group of Wikipedia admins by issuing a Request for amidships (RfA). The Wikipedia community via a public vote decides who to promote to amidships.

To gain an insight of this voting mechanism and obtain the structure of the voting distribution, this paper will analysis this voting network based on a view of network science. By doing some literature research and investigate the voting data from the inception of Wikipedia, this paper will answer the following questions:

- **Q1: How is the voting distribution looks like? How many candidates is likely to be chose by a single voter? How likely is a candidate to win the trust of the major voters?**

  This question is designed to help give a blueprint of the voting structure. To specify this question, how many candidates does a single voter usually pick? If a voter has only voted for several candidates, that would be logical because voter does not vote randomly. Individuals usually have their own preference, even though it sometimes changes. However, if it is found that most of the voters are connected to a huge number of candidates, then the voter’s choices would be meaningless, and there might be some hidden problem in the PfA process.

  Additionally, it should also figure out how is the votes distributed. If the votes are average distributed, then every candidate has the equal chance to be elected, that would definitely increase the difficulty of choose the final winner as everyone gets similar votes. Nevertheless, if only several candidates win most of the votes, then it would simplify the decision process. In addition, the final election result would be much more compelling, since this result meet the vote of a lager ratio of voter.

Particularly, the less potential successful candidate (those who get a large number of votes), the larger proportion of the voter picking their ideal administer successfully. Accordingly, the final result would be more compelling to the major of ordinary users.

- **Q2: Will the voters form different voting communities? If it is yes, how many communities is?**
  This question is designed to examines the diversity of the voters. Moreover, it checks if there are some groups are able to dominate the voting result. The research result for this question might provide some insights to prevent innocent voting manipulations. If a group is considerably larger than others, then the result would be determined only by that group. In addition, they may discuss in advance to vote a certain candidate so that they can ensure the success of that candidate. It can be unfair to other participants.

- **Q3: How likely is a potential winner vote to their opponents?**
  In our perception, it is unlikely for a candidate to vote his opponents, especially those who has a larger number of votes. By contrast, they are more likely to vote themselves to increase the probability of election victory. However, if it is found that candidates in RfA process would vote for their opponents, then it might be proved that this process is honest and fair, since participants tend to vote the most qualified one instead of the one who maximize their own profit.

The dataset used in this research was collected from the SNAP website. It records the administrator elections and vote history data, including 7115 nodes and 103689 edges. Nodes in the network represent Wikipedia users and a directed edge from node *i* to node *j* represents that user *i* voted on user *j*. Some network analysis tool, including Excel, Gephi and NetworkX, were applied to help analyze and visualize the network.

The result of the research suggests that it is common that a voter only votes on few candidates, and a candidate only wins few votes. Only few candidates are able to win most of the votes from voters. The research result also shows that the voters of the RfA process is well diversified, and there are around 30 communities with 4 main communities. Another interesting finding indicates that it is possible for candidates to vote on their opponents, but it is under a certain condition. i.e., when the number of votes for a candidate is quite large but not large enough, a candidates' own vote becomes crucial, so he tend to not vote on his opponents. From these results, it may be concluded that the RfA process is running a quite healthy voting system. Some points in the results might provide some useful insight and advice to Wikipedia team.

2.Background

2.1 Modularity

Modularity is a measure of how well a network is partitioned into communities. When a graph G is portioned into S groups, the modularity, Q, is defined as:

$$Q(G,S) = \frac{1}{2m}\Sigma_{s\in S}\Sigma_{i\in S}\Sigma_{j\in S}\left(A_{ij} - \frac{k_i k_j}{2m}\right)$$

Where m is the total number of edges, K_i and k_j are degrees of node i and node j, and the adjacency matrix element A_{ij} = 1, if there is an actual edge between i and j, otherwise it is 0. The modularity lies in the range [-1,1], and -0.3<Q<7 means significant community structure.

2.2 Rich club

The rich-club phenomenon in complex networks is that the nodes with high degrees are on average more intensely interconnected than the nodes with smaller degrees (McAuley et al., 2007). Rich-club coefficient quantitatively measures the density of interconnectivity among a group of the richest nodes in a network. For each degree k, the rich-club coefficient is the ratio of the number of actual to the number of potential edges for nodes with degree greater than k:

$$\phi(r) = \frac{2E_k}{N_k(N_k - 1)}$$

N_k is the number of nodes with degrees lager than k, and E_k is the actual number of links among the N_k nodes.

2.3 Degree Assortativity

Assortativity measures the similarity of connections in the graph with respect to the node degree. A positive degree assotativity coefficient means the nodes tend to connect with nodes of similar degrees, while a negative coefficient indicates that high-degree nodes tend to connect with low-degree nodes. The calculation formula for assortativity coefficient is showed below:

$$r = \frac{\sum_i j_i k_i - M^{-1}\sum_i j_i \sum_{i'} k_{i'}}{\sqrt{\left[\sum_i j_i^2 - M^{-1}\left(\sum_i j_i\right)^2\right]\left[\sum_i k_i^2 - M^{-1}\left(\sum_i k_i\right)^2\right]}},$$

where j_i and k_i are the excess in-degree and out-degree of the vertices that the ith edge leads into and out of respectively, and M is the number of edges (Newman, 2003).

3.Literature Survey

The RfA process of Wikipedia has always been an interesting topic for scholars and there are few related works that focus on similar topic.

The works of Cabunducan et al. (2011) has studied the factors that influence different stages of the voting process, and they considered both support and oppose votes. In the research, they pointed out that a group of influential supporters (or opposers) can skew an election in favor (or against) a certain candidate. They clarified that may be because an influential node might dominate the outcome of the voting process. When a leading voter gets a voting preference, it is possible for him to guide others to vote the same with him. That probably indicate the answer to Q1: the votes might not be average distributed since some group of voters can be influenced by a single decision of an influential voter. Additionally, Cabunducan et al. (2011) also found out that the participants of friends and communication between user and candidate (the latter weighing more heavily) can simulate the participation of ordinary users, and users are motivated to support candidates whom they are acquaintances with. Relatively, his result may suggest the answer of Q2 that there would be several voting communities as some users are introduced to this PfA process by a certain candidate, therefore, they tend to vote for that certain candidate. However, this work mainly focuses on the individual behavior and give specific explanation. This paper will focus more on the macro-relationship among the participants of the voting process and verify the guesses above.

Another work produced by Kordzadeh and Kreider (2015) analyzed which characters of candidates are more likely to lead to the election success. It is demonstrated that more senior and those who have tried a smaller number of times to become administrators are more likely to be elected in the RFA process. Their later work in 2016 reveals that total contribution of a candidate can substantially influence one's success in the RfA process, nevertheless, those who contributed more frequently to the User Talk pages were less likely to be successful in the promotion process. They explained that might be because community members are more likely to trust users who have a significant demonstrated history of project contribution, but users who spend a large amount of their effort in personal communication instead of the core goal contribution of projects may waste precious community resources (Kordzadeh & Kreider, 2016). Therefore, those who focus too much on communication may win less trust from the public. Their research results seem to have no direct relation to the questions appeared in this paper, however, they provide a good insight to the voting mechanism, and they may give some supplementary explanation to the research results of the research in this paper later.

Another quite similar work found is the research of Lorek et al. (2013). Their work numerically analyzed that the mean number of votes in a single RfA increased gradually from 20 to 88 during 2005 and 2010. They also revealed that the ratio of accepted candidates, those who successfully get the major votes and win finally, is between 57% and 70% during 2005 and 2008, while from 2009 to 2010, this value reduced below 50%. To some extend, they statistically introduced the voting data and basically described the voting structure. However, they only analysis it from a numeric angle, and they ignored the relationship among participants. Later, this paper will analysis it from an angle of network science. Furthermore, the work of Lorek et al. (2013) gave two remarkable hypotheses, which provides an insight to Q2 and explained why there exists several groups of voters (It already gave the positive answer that there are several voting groups):

- Hypothesis A: new admins are elected on the basis of acquaintance. Voters are more likely to choose the candidates that they have already known.
- Hypothesis B: new admins are elected on the basis of similarity of experiences in editing of articles on various topics.

The final result of their work describes a positive validation of hypothesis B. The voters, as the other active admins, their experience is increasing over time, and their thresholds of accepting a candidate are likely to increase. It gave negative hypothesis to hypothesis A. If this hypothesis is true, then it would be harmful to the sustainable development of Wikipedia since the communities is becoming increasingly closed.

4.Methodology

The data was collected from the SNAP website, which is produced by Leskovec, J. from Stanford University. The dataset was created using the complete dump of Wikipedia page edit history and all the administrator elections and vote history data are included. There are 2,794 elections, 103,663 total votes and 7,066 participants, comprising 7115 nodes and 103689 edges. To maintain the integrality and consistency of data, all of the available nodes and edges are applied to the quantitative analysis in this research. The network contains all the Wikipedia voting data from the inception of Wikipedia till January 2008. In this dataset, around half of the votes are by existing admins, while the other half comes from ordinary Wikipedia users.

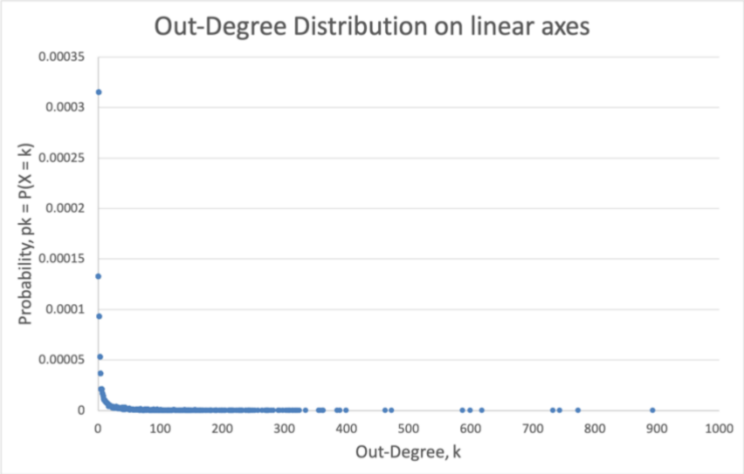
During the quantitative analysis process, the degree distribution was plotted to explore how the vote assigned. In addition to the distribution on the linear axes, the one based on log-log axes was also produced to show the distribution more clearly. In order to find out which number of communities is optimal for this network, and to measure how well it is partitioned, the modularity was used as an assistant tool. Moreover, to test if there is someone vote on his opponents, the algorithm of assortativity and rich club was applied. When examining this problem, the candidates with top votes were mainly considered, since they are those who have the probability to win and directly influence the voting result. If the degree assortativity coefficient is positive, and the rich club coefficient increases with the degree, it might indicate that the candidates with top votes frequently vote on each other (As they are competitors, this positively validate that candidate may vote for their competitors).

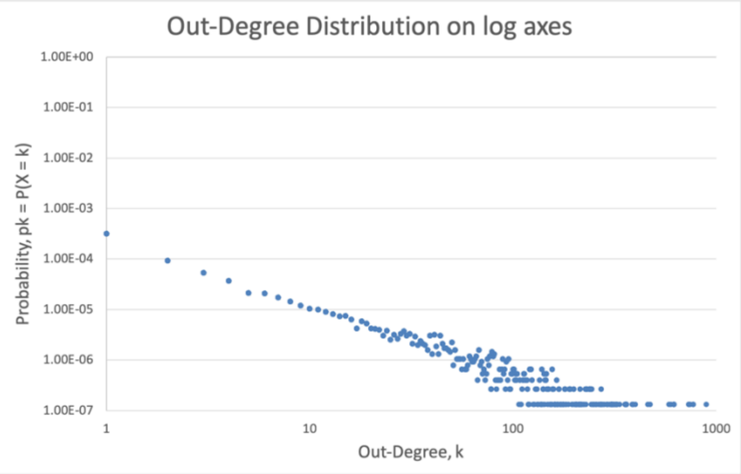
Few network analysis tools were used in this research. The most basic tool, Excel, was used to plot the in-degree and out-degree distributions. A professional visualization and exploration software 'Gephi' was used to help visualize the network and partition the network into communities. In addition, the networkx library and matplotlib library in Python3 facilitated the analysis of mixing pattern and rich-club coefficient.

5. Results

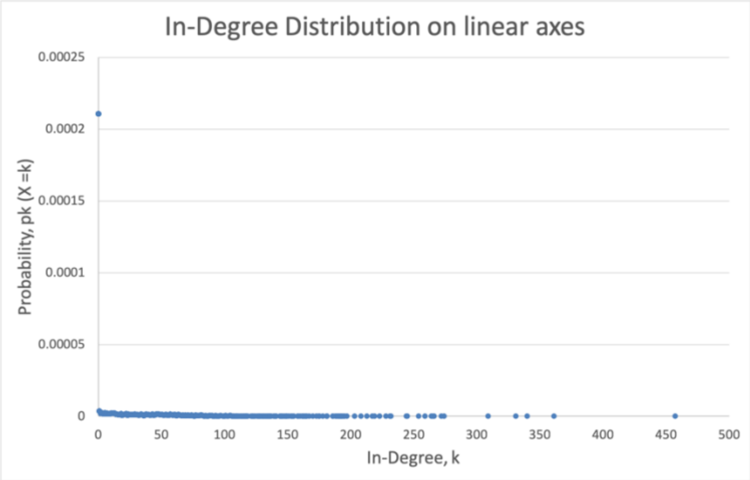
5.1 The Voting Distribution

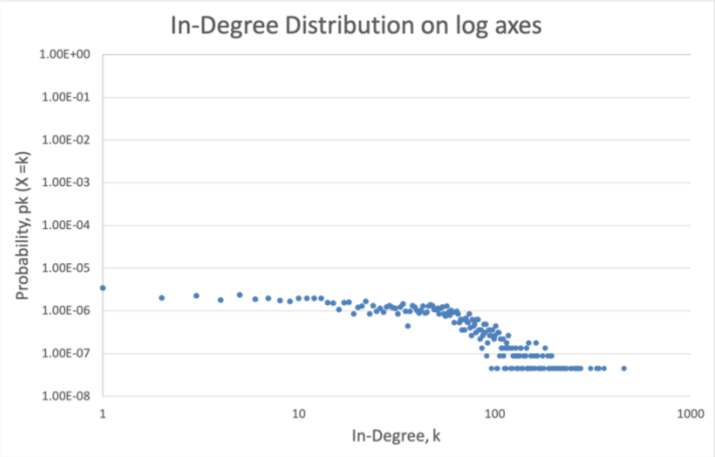
To investigate the voting distribution, the out-degree distribution and in-degree distribution has been visualized, which are showed below.





The out-degree of a node here represent the number of candidates that a single voter pick. The left chart describes the out-degree distribution on linear axes. From this chart, it can only tell that it is more possible for a voter to vote on only few candidates rather than a larger number of candidates, but due to the density of points, no more information can be provided. Therefore, a distribution on log axes was produced to solve this problem. From the right chart, it is clear that the present a decreasing trend, and it follows a power law degree distribution. That means, an ordinary user is more likely to choose only one candidate, and the probability of picking more than one candidate is decreasing significantly as the increase of candidate number. That is quite reasonable. The communication between participants is limited, so the number of candidate that a voter know is also limited. Moreover, as Cabunducan et al. (2011) stated, some voters were introduced to this RfA event by a certain candidate, so they may not know most of the other candidates. It is extremely rare that a single voter has chosen more than 100 candidates, though several voters have done. As its most extreme, there is one having chosen around 1000 candidates. Since the probability of this is very small, its effect can be negligible.





The in-degree of a node means the number of votes for a candidate. The in-degree distribution describes how the votes distributed. It is noticeable that there is an outlier in the left linear-based chart, which shows a much higher probability than all the other in-degree. However, due to the point density, it is not intuitive which specific value it is. From the left log-based chart, it illustrates that it is most likely that a candidate only has one vote. Another very interesting phenomenon showed in the right chart is that the vote almost equally distributed among the degree from 1 to 80, which means a candidate have similar probability of getting votes with a quantity from 1 to 80. However, as the in-degree continue to rise, the probability reduces significantly, following a power-law degree distribution. This chart illustrates that it is normal for ordinary candidates get the votes under 80, however, the probability of getting more votes is small. Only few candidates are able to get more than 100 or even 500 votes.

Comparing the out-degree and in-degree distribution, it is found that both of them show downward trend as the node degree rise. The difference is that the out-degree distribution strictly follows a power law distribution while the in-degree distributed equally for degrees from 1 to 80. In addition, it is worth to mention that the largest out-degree is around 900 while the largest in-degree is only around 460. In the most extreme case, a voter voted about 900 candidates, while a candidate won only about 460 votes at most. That is interesting and might conflict with our common sense. Regularly, voters are much more than candidates. Since the total votes is fixed, a common sense is that the votes of the champion candidate should be lager than the vote that a voter give. However, the data showed here does not meet this common sense. Even though the reason has not been studied in this research, it is noticeable for future scholars.

5.2 The voting Communities

Gephi was applied to find the communities. Using the modularity function in the statistics module, the directed graph can be automatically divided to different group. There is a parameter (resolution) required to decide the number of groups. A lower resolution leads to more communities with an average smaller size of each community, and vice versa. Therefore, different values of resolution were applied to find the optimal number of communities. The auxiliary measurement parameter modularity Q reaches the peak at resolution = 1.00 (The test result is showed below). Therefore, the resolution was set to 1, giving the modularity a value of 0.425.

Resolution	Modularity, Q
0.50	0.375
0.75	0.422
1.00	0.425
1.25	0.414
1.50	0.412
1.75	0.408
2.00	0.311

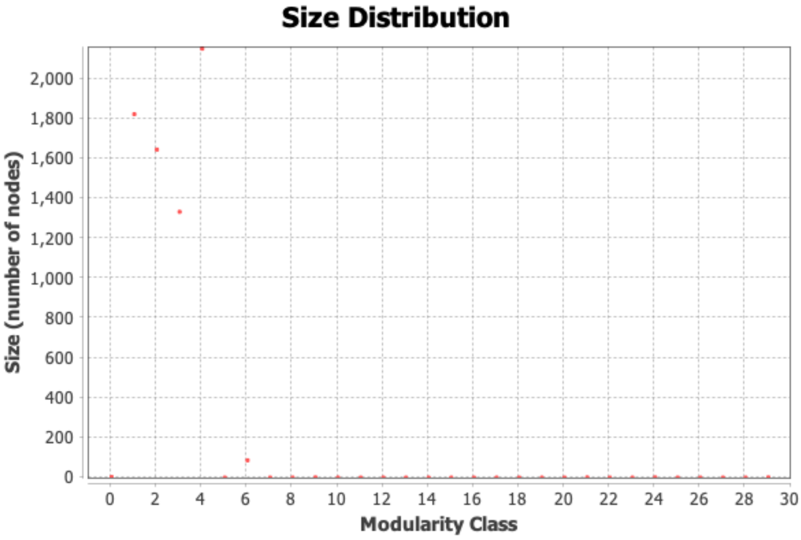
Since the modularity 0.425>0.3, the voters in the RfA process is proved to be able to divided into different communities. As the result below demonstrated, there are 30 communities in total, in which four of them occupies the major proportion: community 4 accounts for 30.30%, community 1 accounts for 25.68%, community 2 accounts for 23.18%, and community 3 accounts for 18.79%. Community 6 makes up a relatively high proportion (1.25%) among the rest groups, and the occupation of all the other communities is less than 1%, having very limited impact on the RfA process. (Appendix A lists the grouping detail.)

Parameters:

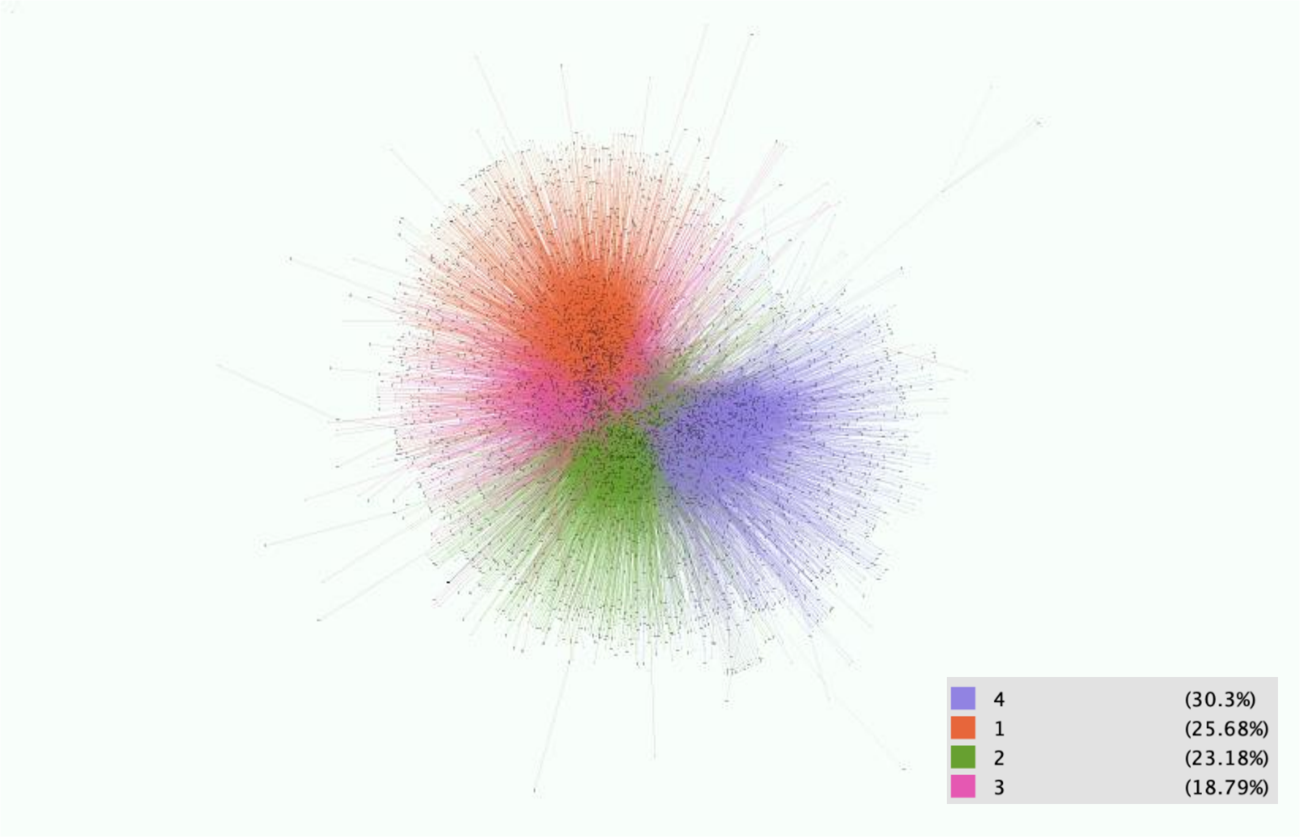
Randomize: On
Use edge weights: On
Resolution: 1.0

Results:

Modularity: 0.425
Modularity with resolution: 0.425
Number of Communities: 30

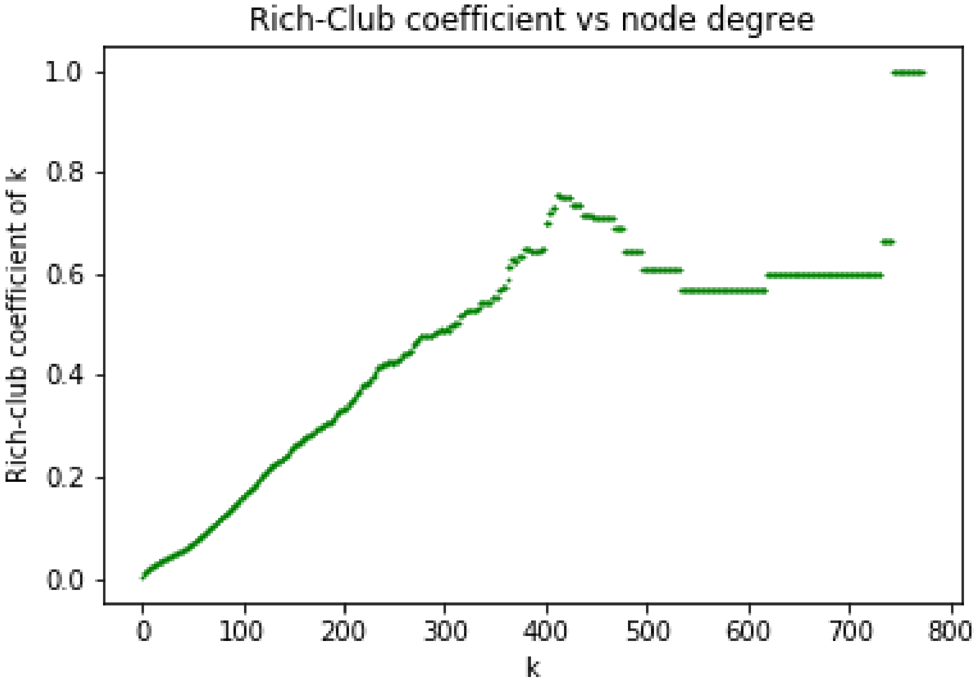


Lorek et al. (2013) has verified that new admins are elected on the basis of similarity of experiences in editing of articles on various topics. Thus, it is understandable that there are different groups of communities. The main communities consist of four groups, and there is no group get a size that significantly lager than others. Therefore, the voter in the RfA process are well diversified and the final election result would not be determined by a single group of voters. The visualization of the voting graph is showed below:



5.3 The number of votes on opponents

By calculation using networkX, the degree assortative coefficient gets a negative value: -0.08305248270016026. Thus, it can be concluded that the high-degree nodes tend to connect with low-degree nodes more frequently, which might imply that the votes of those top candidates are mainly from the ordinary user instead of the other candidate with many votes. However, this only suggests rich node are more likely to connect with poor node, it does not indicate the top candidates do not vote for each other.



More intuitively, the figure above illustrates the relationship between rich-club coefficient and the degree k. From the degree 0 to about 400, the coefficient increases steadily, giving the rich-club coefficient a value of about 0.8. That might be because the some low-degree participants join this RfA process only for voting on acquaintances, then they connect with no more other participants with high degree, therefore, these low-degree nodes get low rich-club coefficient. Again, only a half of the votes in the dataset comes from ordinary Wikipedia users, the other half are by existing admins. They have more degrees and they also vote on some "rich" candidates, so they tend to have lager rich-club coefficient. However, between degree 400 to 500 the coefficient reduced. Since candidate's votes become larger, the probability of their requests to be accepted gets increasingly larger. They need to be very careful at this time, and their votes might determine his or her own success. Thus, they tend to not vote for their opponents, especially those who have more votes than themselves, leading to a downwards trend for rich club coefficient at this stage, and this number only increase slightly between the degree from 500 to 750.

It is very interesting that the rich-club coefficient suddenly increases to infinity close to 1 when the node degree is close to 800. That might be because when the votes of a candidate get large enough, candidates do not have to worry about election failure. Thus, those who have the most top votes are free to choose any admin. In addition, as Burke and Kraut (2008) suggested, the RfA success was strongly influenced by the diverse experience and contributions to the development of WikiProjects. Therefore, it can be drawn that the top candidates usually have more experience in managing this system, so they tend to know this system better than others, so they are more likely to pick out the most qualified admins. Therefore, it forms a fully connected network among the most top candidates.

6. Discussion

In conclusion, the voting distribution of the election follows power-law distribution, so it might verify that the RfA process runs an quite healthy system. It is common that a candidate only wins few votes, and only few candidates are able to win the most trust of voters. Similarly, a voter normally only votes on few candidates, it is really rare for a voter to have voted on more than 100 candidates. Even though there are few, they might be negligible. However, for further study, it is worth to investigate why the largest out-degree is much greater than the largest in-degree. The result might give insight to Wikipedia team to tell if there is any problem in the RfA process.

The research result also shows that the voters of the RfA process are well diversified, and the voters can be divided to around 30 communities with 4 main groups. Thus, no groups can dominate the result and it is hard for anyone to innocent manipulate the votes. In this system, all of the votes of a single voter are possible to influence the final result.

Another finding is that it is possible for candidates to vote on their opponents, even though this is conditional. If the number of the votes gives a candidate the hope of winning, but the number does not promise his or her success, this candidate is likely to be very careful and tend not to vote in his or her opponents. However, if the number of the votes of a candidate is large enough, which guarantee the success of the candidate, or it is small enough, giving the candidates no hope at all, the vote of this candidate would not influence his own result. Thus, he or her would vote freely. In addition, this paper supposed that top candidates with the most votes know the Wikipedia editing system better than others, so they are more likely to pick out the most qualified admins than others. Therefore, the future study can test if the performance of the RfA process would be improved when we increase the weight of top candidates' votes.

There is a limitation of this study. The dataset used in this research includes the voting data of multiple years, while the edges of this network is not weighted. Specifically, if Voter A votes the same candidate this year and last year while Voter B votes two different candidates, voter A would only have one link while voter B have two. That would affect the test accuracy when doing the degree distribution analysis.

7. Reference

Burke, M. and Kraut, R. (2008) "Taking up the mop: identifying future wikipedia administrators" *Florence, Italy*, April 05 - 10, 2008, The ACM Digital Library [online]. Available from: https://dl.acm.org/citation.cfm?id=1358871 (Accessed: 3rd January 2019).

Cabunducan, G., Castillo, R., and Lee, J. B. (2011) "Voting Behavior Analysis in the Election of Wikipedia Admins", 2011 *International Conference on Advances in Social Networks Analysis and Mining*, IEEE [online]. Available from: https://ieeexplore.ieee.org/abstract/document/5992657/references (Accessed: 3rd January 2019).

Jankowski-Lorek, M., Ostrowski, L., Turek, P. et al. (2013) "Modeling Wikipedia admin elections using multidimensional behavioral social networks", *Soc. Netw. Anal. Min.*, 3: 787., SpringerLink [online]. Available from: https://link.springer.com/article/10.1007%2Fs13278-012-0092-6 (Accessed: 3rd January 2019).

Kittur, A., Chi, E., Pendleton, B., Suh, B. and Mytkowicz, T. (2006) "Power of the Few vs. Wisdom of the Crowd: Wikipedia and the Rise of the Bourgeoisie" *Word Wide Web*, 1(2), November, ResearchGate [online]. Available from: https://www.researchgate.net/publication/200772541_Power_of_the_Few_vs_Wisdom_of_the_Crowd_Wikipedia_and_the_Rise_of_the_Bourgeoisie (Accessed: 3rd January 2019).

Kordzadeh, N. and Kreider, C. (2015) "Request for Adminship (RfA) within Wikipedia: How Do User Contributions Instill Community Trust?" *SAIS 2015 Proceedings* [online]. Available from: https://pdfs.semanticscholar.org/17f7/6cba3f7f714fd8d602651952df0eff43885c.pdf (Accessed: 3rd January 2019).

Kordzadeh, N. and Kreider, C. (2016) "Revisiting Request for Adminship (RfA) within Wikipedia: How Do User Contributions Instill Community Trust?" *Spring2016*, Volume 4, Issue 1, SAIS [online]. Available from: https://quod.lib.umich.edu/j/jsais/11880084.0004.102/--revisiting-request-for-adminship-rfa-within-wikipedia-how-do?rgn=main;view=fulltext (Accessed: 3rd January 2019).

Lescovek, J. (2008) Dataset: 'Wikipedia vote Network', Snap [online]. Available from: http://snap.stanford.edu/data/wiki-Vote.html (Accessed: 3rd January 2019).

McAuley, J.J., Costa, L.D.F. and Caetano, T.S. (2007) "The rich-club phenomenon across complex network hierarchies" Cornell University [online].

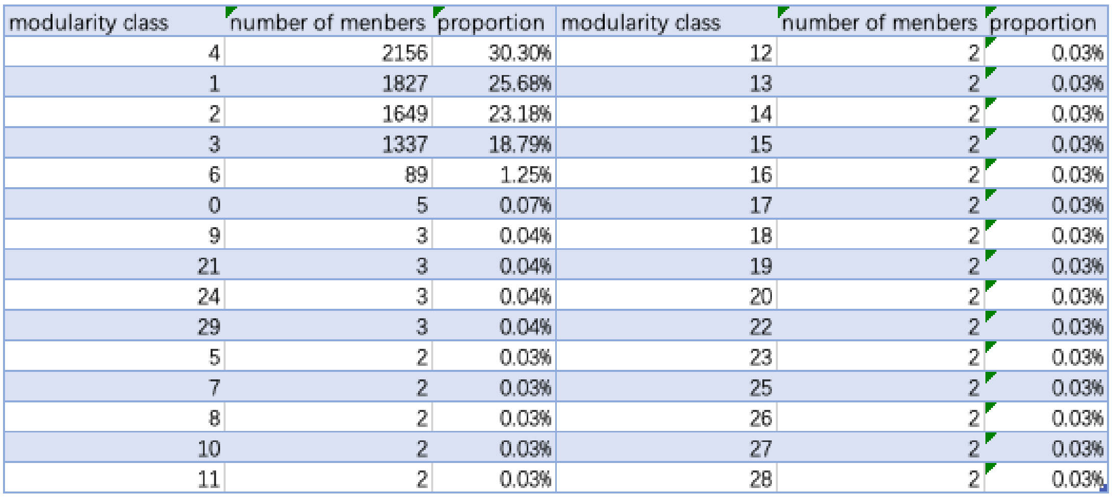
Available from: https://arxiv.org/abs/physics/0701290 (Accessed: 3rd January 2019).

Newman, M. E. J (2003) "Mixing Patterns in networks", *Phys. Rev.*,E 67, 026126(2003), Cornell University [online]. Available from: https://arxiv.org/abs/cond-mat/0209450 (Accessed: 3rd January 2019).

Welser, H.T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., and Smith, M. (2011) "Finding social roles in Wikipedia" *iConference 2011*, February 8-11, 2011, ResearchGate [online]. Available from: https://www.researchgate.net/publication/220889391_1_Finding_social_roles_in_Wikipedia_1 (Accessed: 3rd January 2019).

8. Appendix

Appendix A:



modularity class	number of menbers	proportion	modularity class	number of menbers	proportion
4	2156	30.30%	12	2	0.03%
1	1827	25.68%	13	2	0.03%
2	1649	23.18%	14	2	0.03%
3	1337	18.79%	15	2	0.03%
6	89	1.25%	16	2	0.03%
0	5	0.07%	17	2	0.03%
9	3	0.04%	18	2	0.03%
21	3	0.04%	19	2	0.03%
24	3	0.04%	20	2	0.03%
29	3	0.04%	22	2	0.03%
5	2	0.03%	23	2	0.03%
7	2	0.03%	25	2	0.03%
8	2	0.03%	26	2	0.03%
10	2	0.03%	27	2	0.03%
11	2	0.03%	28	2	0.03%