Detecting Communities in Social Networks through Modularity Maximization

Samridhi Khurana
MBA Candidate, IIM Bangalore
B313, Hostel Blocks,
IIM Bangalore - 560076, India

ABSTRACT
Community structure in a network plays an important role in understanding its characteristics and functioning. In a social network, community structures represent closely knit groups of people, and are vital to understand and analyze the network as a whole. The network is described by a graph with nodes representing the entities and the edges representing connections between these entities. Very recent of community detection algorithms, is a method that relies on optimization of a parameter called modularity [1], which is an indication of the partition of a network into communities. Another significant article in this regard is [2] by Santo Fortunato and Marc Barth´elemy, which brings out that optimizing modularity on large networks fails to resolve small communities, even when they are well defined. In the present article, irregularities in the mathematical formulation of modularity are addressed and the author proposes an improvised procedure for community detection. The approach suggested is based on Modularity maximization but modified in the sense that the algorithm is applied in a recursive manner on the network until all sub-communities within the communities are identified. The improvised algorithm results in a better community structure with all distinct community structure clearly spelt out.

General Terms
Community Detection, Modularity, Resolution Limit, Social Network, Node

Keywords
Probability Estimate, Recursive BGLL, Q_{single} ,Q_{pair}

1. INTRODUCTION
Community detection helps in understanding the properties of a network. Identifying community structure in a social network is of particular importance as it helps in analyzing the behavioral patterns among different communities. Community structure is not only confined to social networks and has been found in various other large-sized complex networks ([3]-[5]).

Modularity is a famous criterion employed for community detection. It essentially compares the fraction of links within a module with the expected value of the same in a random graph with the same degree distribution (Equation 1). Mathematically, the expression of Modularity takes the form

$$Q = \frac{1}{2m}\sum_{i,j}(A_{ij} - P_{ij})\,\delta(c_i, c_j) \quad (1)$$

where m is the total number of links in the graph, A_{ij} is the adjacency matrix of the graph indicating the actual number of edges between nodes i and j in the graph, and P_{ij} is the expected number of edges between nodes i and j in the graph. $\boldsymbol{\delta(c_i, c_j)}$ is set to 1 when nodes i and j fall in the same module, else is set to 0. With certain assumptions, the Probability Estimate term P_{ij}, as proposed by Newman & Girvan, is

$$P_{ij} = \frac{k_i k_j}{2m} \quad (2)$$

where each node $\boldsymbol{x}$ has a node degree $\boldsymbol{k_x}$.

The problem of identifying communities in a network is basically equivalent to optimizing modularity. Although modularity maximization seems to be an effective measure for identifying communities, it suffers from certain drawbacks; the most serious being the resolution limit of modularity [2]. Modularity was found to contain an intrinsic scale dependency and modules smaller than that scale could be resolved into separate communities. Modules identified with modularity optimization could be a single community, or a cluster of various communities merged together. This result thus introduced some caveats in the use of modularity to detect community structure.

In the present article, the author explores the notion of modularity in a greater detail, and augments to the problems associated with its mathematical formulation. The probability estimate term P_{ij} which relies on the Configuration Model is identified as the error term and an example of ring of cliques is undertaken to understand the problems further. In the last section, the author proposes an algorithm, which is a variant of the traditional Modularity optimization algorithm and relies on the recursive application of the standard BGLL algorithm [6] until all communities having size beyond a threshold are identified.

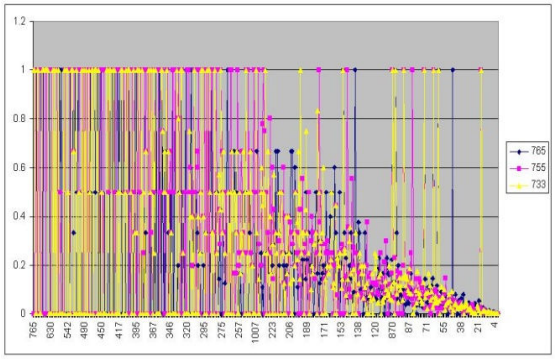
2. ANALYSIS OF THE MODULARITY EXPRESSION

2.1 Attempt to Identify the error term
The Probability Estimate term in the Modularity expression, P_{ij}, relies on Configuration Model, which is based on certain assumptions. As proposed by Newman & Girvan, keeping the degree distribution same, the probability of two edges being connected in a random graph is $\frac{k_i * k_j}{m}$. But this term does not go in accordance with some of the basic constraints. For a graph with one single edge and two nodes, the expected number of full edges between two nodes i and j in a random graph should be 1, while according to the proposed expression, it comes out to be ½. Secondly, in a graph of two nodes i and j with node degrees k_i and k_j, under the constraint $m = k_i + k_j - 1$, the nodes i and j have to connected, but the proposed expression does not set P_{ij} to 1. Thirdly, it seems true intuitively that with the rise in the number of nodes in a network (n), the probability of i and j being connected should not increase, but the proposed expression does not take care of this constraint as well.

Therefore, this indicates that there could be some imperfections existing in the Probability estimate term as proposed by Newman-Girwan.

To verify the same, marginal distribution curves for various data sets are plotted as shown in Figure1 and Figure2. Figure1 shows the variation of the value of P_{ij} vs k_j for a fixed value of k_i for the Epinions social network data set [7]. Figure 1(a) is plotted for large values of k_i while Figure 1(b) is for comparatively small values of k_i. Likewise, Figure 2 shows the plot for the Wikipedia Vote network[8].

(a)



(b)

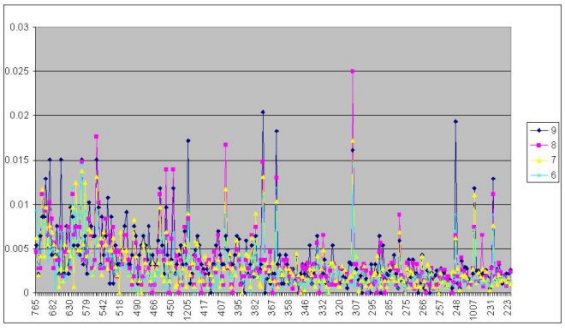


Figure 1. Epinions Data Set Plot of Pij vs kj for different fixed values of ki

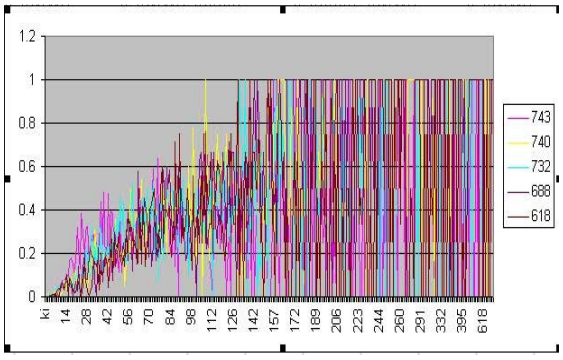


Figure 2. Wiki Vote Data Set Plot of Pij vs kj for different fixed values of ki

The above patterns show that the Probability values largely deviate from the linear model and establish the fact that unlike what the Probability estimate expression suggests, P_{ij} does not follow a linear relationship with k_j, for a given fixed value of k_i. Therefore, in order to correctly implement the modularity approach, a better estimate of expected number of edges in a randomized graph needs to be formulated.

2.2 Relaxing the Assumptions

In this section, the probability of a connection between any two nodes in a random graph is calculated using a different approach of a suitable concept.

In the community detection problem, degree of each node is known. It would be therefore right to say that the number of connections every person is willing to make is known. The aim is to find out the probability with which the nodes *i* and *j* are connected.

Let the degree of node *i* be k_i. In order to calculate the probability of *i* being connected to a node *j* which has a degree of k_j, different cases are considered, and the final probability will be as follows:

$$P = p_j^1 + (1-p_j^1)p_j^2 + (1-p_j^1)(1-p_j^2)p_j^3 + \cdots + (1-p_j^1)..\left(1-p_j^{k_i-1}\right)p_j^{k_i} \quad (3)$$

where,

P is the total probability of *i* and *j* being connected

p_j^f is the probability of a connection being formed between *i* and *j* in the f^{th} attempt

The given graph consists of *n* nodes and *m* edges. Calling the nodes as $i,j,l_1,l_2,\ldots,l_{n-2}$, p_j^f is calculated as follows.

Considering the case that nodes *i* and *j* are connected in the first attempt. To calculate p_j^f, the total number of connections available (i.e. $2m/2$) are considered, and the connections coming from node *i* are ignored as the basic assumption of no self loops is considered to be true. The probability expression thus takes the form

$$p_j^1 = \frac{k_j}{2m - k_i} \quad (4)$$

Considering the next case that node *i* and *j* are not connected in the first time and a connection is formed between them in the second attempt, p_j^f is calculated like before, with the additional removal of the degree of the nodes with which *i* got connected to in the first attempt.

$$p_j^2 = \sum_{\alpha_1=1}^{n-2} p_{l_{\alpha_1}}^1 \frac{k_j}{2m - k_i - k_{l_{\alpha_1}}} \quad (5)$$

The next case will be of nodes *i* and *j* forming a connection in the third attempt.

$$p_j^3 = \sum_{\alpha_2=1}^{n-2} p_{l_{\alpha_2}}^2 \sum_{\alpha_1=1}^{n-2} p_{l_{\alpha_1}}^1 \frac{k_j}{2m - k_i - k_{l_{\alpha_1}} - k_{l_{\alpha_2}}} \quad (6)$$

$$\boldsymbol{\alpha_1 \neq \alpha_2}$$

In a similar manner, the probability of the connection being formed between Node *i* and Node *j* in the t^{th} attempt, given that Node *i* and Node *j* were not connected in any of the prior attempts can be given as follows.

$$\sum_{\alpha_1=1}^{n-2}\sum_{\alpha_2=1}^{n-2}\cdots\sum_{\alpha_{t-1}=1}^{n-2} p^1_{l_{\alpha_1}} p^2_{l_{\alpha_2}} \dots p^{t-1}_{l_{\alpha_{t-1}}} \left\{\frac{k_j}{2m - k_i - k_{l_{\alpha_1}} - k_{l_{\alpha_2}} \dots - k_{l_{\alpha_{t-1}}}}\right\}$$

$$\alpha_1 \neq \alpha_2 \neq \alpha_3 \dots \neq \alpha_{t-1} \text{ `} \quad (7)$$

The above calculated values (Equation 4-7) can be put up back in equation 3 and P (defined in Equation 3) can be calculated, which can then be used in the Modularity expression in place of the Probability Estimate term P_{ij}.

This analysis brings forward the point that when the probability of two nodes being connected is calculated by a method other than the one that relies on Configuration Model, then the probability obtained is a very complex one, and the Probability Estimate term does not provide a good approximation to it. This, along with the non-linear marginal distribution curves plotted for various data sets, brings out the irregularities and the inconsistencies possibly lying in the Probability Estimate term P_{ij}. Though the above calculated probability is not relying on any assumptions, it cannot serve as a fit in the Modularity expression because of its complex structure that would make the evaluation computationally heavy. This analysis may not have provided a better estimate to P_{ij} in actual terms, but has led to the conclusion that P_{ij} term is based on certain unreasonable assumptions and may be the cause of the Resolution Limit of Modularity.

3. DETAILED ANALYSIS OF THE RING OF CLIQUE PROBLEM

One interesting example pointed out in the Resolution limit paper is the Ring of Clique problem. In a network made of identical cliques, connected to each other by a single link, ideally, all cliques should be identified as separate community, no matter what community detection algorithm is used. But due to resolution limit, as the number of cliques go higher than √L, where L are the total number of links in the network, modularity optimization methods identify two cliques combined as one separate community.

Modularity optimization is unable to identify the natural partitions in the network and identifies a pair of cliques as one single community. The analysis is Section 2 points out to the Probability Estimate term as the error term which could be the reason for Resolution limit. In other words, the Probability estimate term P_{ij} could potentially be the reason for the tendency of Modularity to merge individually distinct communities together and identify them as a single cluster.

This section analyzes the Ring of Cliques network from a different perspective. Instead of calculating P_{ij} as proposed by Newman-Girvan, the probability of two nodes being connected is calculated mathematically.

3.1 Description of Ring of Cliques Network

A clique or a complete graph is one in which there is a connection between every pair of nodes. A clique in a network signifies a group of persons who interact with each other more regularly and intensely than others in the same setting. In this network, let each clique consist of *x* nodes. This means that there are ${}^xC_2$ links inside each clique. Figure 3 shows a schematic representation of a clique with x = 6.

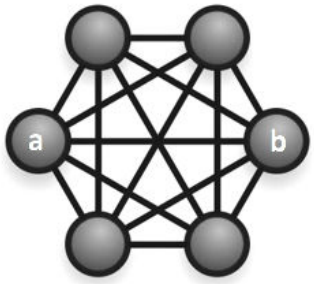


Figure 3: Clique with six nodes: x = 6 and m = ${}^6C_2$. Such cliques are connected in a ring like manner via two nodes, say a and b. Node a and Node b have degree 6 while the other four nodes have degree 5.

To form the network of Ring of Cliques, n such identical cliques are connected together in a ring like manner. It is assumed that the network consists of even number of cliques, i.e. n is an even natural number. Also, to simplify the mathematical formulation of the network, it is assumed that two different nodes of the cliques are involved in the ring connections. In Figure 3, for example, two different nodes, say Node a and Node b, participate in the ring formation. This assumption makes the calculation of required probabilities easier and more concrete. The network is formed by joining n such cliques together, and therefore in every clique, two different nodes are involved in ring connections. Under the given setting, Node a and Node b are two distinct nodes, and therefore, the number of nodes inside each clique has to be greater than two.

In a clique of x nodes, there is an edge between all possible pairs of nodes which implies that each node in the clique K_x has a degree of x-1. Of these x nodes, there are two such nodes, Node a and Node b, which also participate in ring connections, which makes their degree x instead of x-1. Therefore, every clique K_x in the network will consist of x-2 nodes of degree x-1 and 2 nodes of degree x. This can be understood more clearly by the example in Figure 3. All the nodes of the clique K_6 are connected to each other, except for themselves (no self-links). This makes the degree of each node as 5. Now n such cliques are connected together in such a manner that Node a connects this clique to another identical clique by a single link, and likewise, Node b connects this clique to another identical clique, again by a single link. This ring connection increases the degree of Node a and Node b by one. Therefore, Nodes a and b have a degree of 6, while the rest of the nodes have a degree of 5. The description of the clique, K_x as discussed above, is tabulated in Table 1.

Table 1: Description of the Clique K_x

Number of vertices in each clique	x (x>2)
Number of edges inside each clique (m_c)	${}^xC_2$
Degree distribution of the vertices	x-2 nodes with degree x-1 2 nodes with degree x

The network of Cliques consists of n identical cliques, K_x (as described in Table 1), attached together in a ring like manner such that two different nodes of every clique participate in the ring connections. The network as a whole thus consists of *n*

additional edges, other than the edges inside the *n* cliques, which link the *n* identical cliques together. As each clique can be viewed as a subgraph with strong interactions, the network therefore is a realization of *n* communities attached in a ring like manner. Figure 4 gives a pictorial representation of the Ring of Cliques Network. Table 2 summarizes the description of the network.

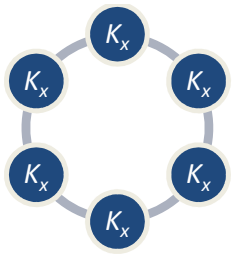


Figure 4: Ring of Cliques Network: n cliques (representative of a community) connected in a ring like manner. Each clique is a well-knit structure, and the best partition of the network into communities should be the one in which every clique is realized as a distinct community

Table 2: Description of the Ring of Cliques network

Number of cliques in the network	N	
Number of connections in the network (m)	Intra-Clique Edges* (m_c)	${}^{x}C_2$ n
	Inter-Clique Edges* (m_r)	
	Total edges (m)	n ${}^{x}C_2$ + n
Degree distribution of the vertices	n(x-2) nodes with degree x-1 2n nodes with degree x	

* Refer to Appendix I

Figure 5 shows the network of 6 cliques each having 6 nodes. The degree distribution of vertices as mentioned in Table 2 can be verified for Figure 5.

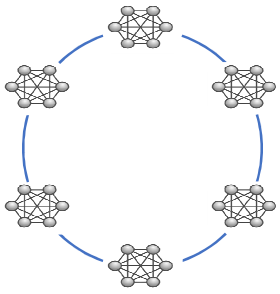


Figure 5: A closer look to the connections in the Ring of Cliques Network

3.2 Probability Calculations

This section calculates the probability of connection between two nodes *i* and *j* in each of the clique. As discussed in the previous section (Refer to Table 1), the nodes in the network have either a degree of *x* or *x-1*. To calculate the probability of two nodes being connected, the connections can be categorized into three cases:

- Both the nodes forming a link have a degree of *x-1*
- Both the nodes forming a link have a degree of *x*
- One of the nodes has a degree of *x-1* while the other has a degree of *x*

Figure 5 depicts the Ring of Cliques network. A closer look to the edge connections is shown in Figure 6. Figure 6a considers the nodes having degree *x-1* and highlights the edges falling between all such nodes. Figure 6b highlights the edges falling between pairs of nodes with degree *x*, and Figure 6c accentuates the edges that lie between nodes wherein one node has a degree of *x* and the other has a degree of *x-1*.

Case 1: Probability that two nodes of degree *x-1* are connected in the clique K_x ($P_{x\text{-}1,x\text{-}1}$)

This can be considered equivalent to the problem of calculating the probability of a connection in a clique where both the nodes have a degree of *x-1*. The network consists a total of *n(x-2)* nodes of degree *x-1*, which means ${}^{n(x-2)}C_2$ edges are possible between them. But of these total possibilities, only a few edges are actually present in the network. Considering an individual clique, the edges which are incident on vertices having degree *x-1* are the ones formed between nodes of degree *x-1* (*x-2* such nodes are there), as exhibited by Figure 6a. The number of such connections inside a clique is thus, ${}^{x-2}C_2$, and there are a total of *n* cliques. In a clique, the probability of edge connections which are incident on vertices of degree *x-1*, $P_{x-1,x-1}$ is given by equation 8a.

$$P_{x-1,x-1} = \frac{{}^{x-2}C_2}{{}^{n(x-2)}C_2} = \frac{x-3}{n\{n(x-2)-1\}} \quad (8a)$$

Case 2: Probability that two nodes of degree *x* are connected in the clique K_x ($P_{x,x}$)

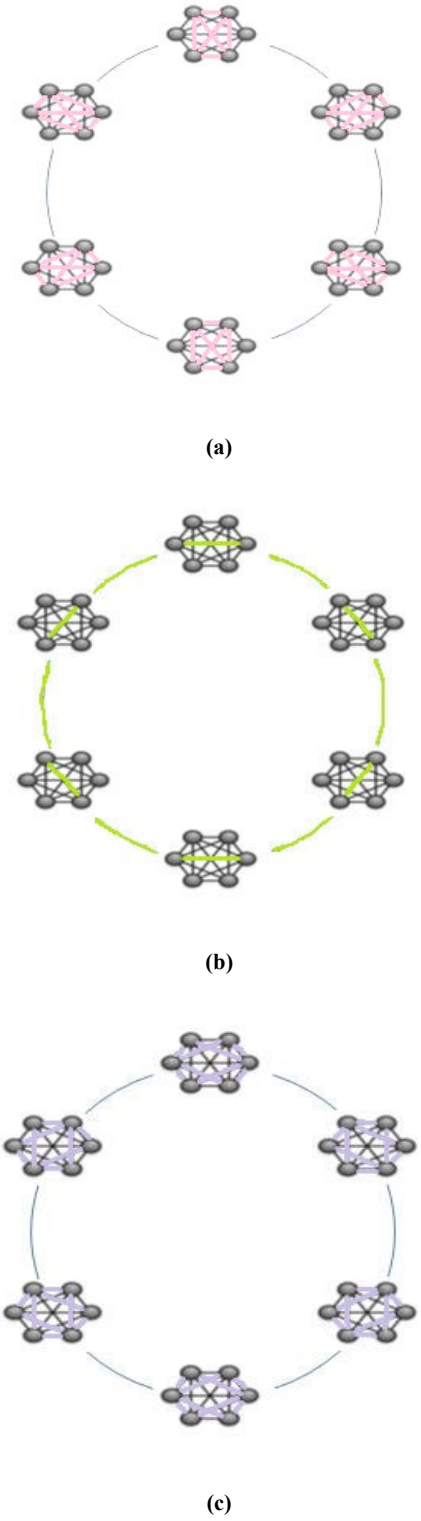
This can be considered equivalent to the problem of calculating the probability of two nodes being connected given that both the nodes have a degree of *x*. The network consists a total of *2n* nodes of degree *x*, which means ${}^{2n}C_2$ edges are possible between them. The edges in the clique which are incident on vertices having degree *x* are *n*, as exhibited by Figure 6b. The probability of edge connections in the clique which are incident on vertices of degree *x*, $P_{x,x}$ is given by equation 8b.

$$P_{x,x} = \frac{1}{{}^{2n}C_2} = \frac{1}{n\{2n-1\}} \quad (8b)$$

Case 3: Probability that nodes of degree *x* and *x-1* are connected in the clique K_x ($P_{x,x\text{-}1}$)

This can be considered equivalent to the problem of calculating the probability of two nodes being connected given that one of the nodes has a degree of *x* and the other has a degree of *x-1*. The network consists of *n(x-2)* nodes of degree *x-1*, and *2n* nodes of degree *x*. Number of links between node of degree *x* and node of degree *x-1* will thus be ${}^{n(x-2)}C_1$ ${}^{2n}C_1$ edges are possible between them. The edges between nodes of degree *x* and *x-1* actually present in the network are ${}^{2}C_1$ ${}^{x-2}C_1$ in each of the *n* cliques. This can be verified for Figure 6c. The probability of edge connections which are incident on vertices of degree *x* and *x-1* in each of the clique, $P_{x,x-1}$ is given by equation 8c.

$$P_{x,x-1} = \frac{{}^{x-2}C_1\ {}^{2}C_1}{{}^{n(x-2)}C_1\ {}^{2n}C_1} = \frac{1}{n^2} \quad (8c)$$



(a)

(b)

(c)

Figure 6: Inter-ring and Intra-ring edge connections in the Ring of Cliques Network. (a) Intra-ring edges incident on vertices with degree x-1are highlighted. (b) Edges incident on vertices with degree x are highlighted. (c) Intra-ring edges incident on vertices with degree x-1 and x are highlighted

3.2 Modularity of Partitions

Let us now calculate Modularity when each clique is considered a separate community and Modularity when two clique together are considered as one community.

Q_{single}: Modularity Considering each clique as a distinct partition

Revisiting Equation 1, Modularity is calculated for the above described Ring of Cliques network. *Pij* is calculated by considering the above described cases and the set of equation 8 is used to compute Modularity.

$$Q_{single} = \frac{1}{2m}\sum_{i,j}(A_{ij} - P_{ij})\,\delta(c_i, c_j)$$

$$Q_{single} = \frac{1}{2m}\{\sum_{i,j} A_{ij}\delta(c_i, c_j) - \sum_{i,j} P_{ij}\,\delta(c_i, c_j)\}$$

$$Q_{single} = \frac{1}{2m}[\ 2\,n\,{}^{x}C_2 - 2\,\{\ n\,P_{x-1,x-1} + n\,P_{x,x} + n\,P_{x,x-1}\}\ ]$$

$$Q_{single} = \frac{1}{2m}[\ 2\,n\,{}^{x}C_2 - 2\,\{\ \frac{x-3}{n(x-2)-1} + \frac{1}{2n-1} + \frac{1}{n}\}\ ]$$

$$Q_{single} = \frac{1}{m}\left[\ n\,{}^{x}C_2 - \left\{\ \frac{x-3}{n(x-2)-1} + \frac{1}{2n-1} + \frac{1}{n}\right\}\ \right] \quad (9)$$

$\mathbf{Q}_{pair}$: Modularity Considering a pair of cliques as a distinct community

Considering a cluster of two cliques together as a separate community, Q_{pair} is calculated. When two clusters are considered as a single community, the difference effectively comes only in the P_{ij} term under the case $P_{x,x}$, where the contribution of inter-clique edge also becomes relevant.

$$Q_{pair} = \frac{1}{2m}\sum_{i,j}(A_{ij} - P_{ij})\,\delta(c_i, c_j)$$

$$Q_{pair} = \frac{1}{2m}\{\sum_{i,j} A_{ij}\delta(c_i, c_j) - \sum_{i,j} P_{ij}\,\delta(c_i, c_j)\}$$

$$Q_{pair} = \frac{1}{2m}[\ 2\,n\,{}^{x}C_2 - 2\,\{\ \frac{2\,{}^{x-2}C_2}{n(x-2)C_2}\frac{n}{2} + \frac{3}{{}^{2n}C_2}\frac{n}{2} + \frac{2\,{}^{x-2}C_1\,{}^{2}C_1}{n(x-2)C_1\,{}^{2n}C_1}\frac{n}{2}\}\ ]$$

$$Q_{pair} = \frac{1}{m}[\ n\,{}^{x}C_2 - \{\ \frac{x-3}{n(x-2)-1} + \frac{3}{2(2n-1)} + \frac{1}{n}\}\ ]$$

(10)

3.4 Results

A closer look to Equation 9 and Equation 10 establishes the following relation between Q_{single} and Q_{pair}.

$$Q_{single} - Q_{pair} = \frac{1}{m}\frac{1}{2(2n-1)} \quad (11)$$

In Equation 11, the right-hand term is positive, which implies that Q_{single} is greater than Q_{pair}. Ideally, the best partition of the

network of Ring of Cliques should identify *n* communities, each cluster being a separate community. This in itself justifies the usage of Modularity optimization to detect community structure. The results and the methodology also indicate towards problems lying in the Probability Estimate term based on the Configuration model.

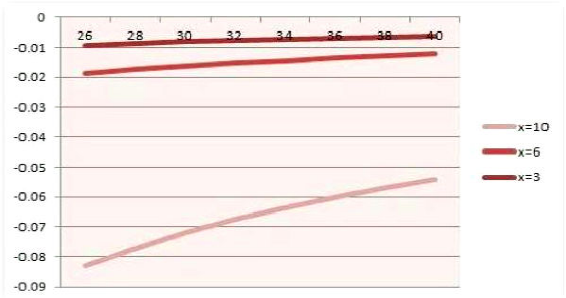


Figure 7: Plot of Qpairs-Qsingle vs n, for x = 10, x = 6 and x = 3.

The proposed approach to probability calculations is able to wave off the resolution limit concerns associated with Modularity measure. The Probability Estimate term used by Newman-Girvan in the Modularity expression could be the cause of Resolution Limit in community detection. The use of a different approach here leads to a better community detection and identifies smaller-sized distinct communities as well separate units, instead of identifying them as a single unit.

4. DETECTING SUB-COMMUNITIES WITHIN COMMUNITIES

Modularity optimization may result in merging of various small-sized communities into one large sized community. The community structure identified by algorithms based on modularity maximization, thus, may consist of various clusters merged together as one single module. This calls for a better community detection method which identifies every distinct community as a separate one. The author proposes to recursively apply modularity maximization algorithm on the communities identifies in a network until all sub-communities are detected. As a first step, BGLL Algorithm [6] based on Modularity maximization gives the community structure. The above discussion points to the fact that certain communities may be merged as one community. Therefore, in the second step, communities identified which are large sized (bigger than a set threshold) are considered as a separate network, and BGLL algorithm is applied on them individually, therefore leading to the further breakdown of these communities into the smaller ones. This procedure is repeated until no more breakdown of communities occurs, i.e. where no sub-community is big enough to be dissolved further.

This procedure ensures that all the communities are identified, specifically the smaller ones, which were probably merged with the bigger ones if the BGLL algorithm was applied only once on the network as a whole. Basically, the bigger communities identified in the first step, can be separated out from the network, and considered as a network by itself in an attempt to find communities within it. This is in accordance with the basic aim of better community detection that the author has been working on so far.

4.1 Recursively applying BGLL Algorithm

The recursive procedure is computationally carried out as explained below:

1. BGLL algorithm is applied on the network data set and the result conveys which nodes falls into which community Id.
2. The next step is to identify the communities that are relatively large-sized and need to be again analyzed for further sub-communities. For this, a parameter is set to be given by the user to specify the threshold to the size of communities. Communities which are larger than the threshold will be analyzed further.
3. For every community identified in Step 2, a file which contains Node Id of the nodes falling into that community is maintained.
4. The nodes falling into that identified community (the big ones) along with the given original network are used to find the sub-graph induced by these nodes on the original network
5. As the next step, BGLL Algorithm is applied, considering the induced sub-graph as the original network.

4.2 Results of Recursive Application

The above recursive procedure is implemented on Wiki-Vote data set [8], and the results are as follows:

Given data set: Wiki-Vote data set (Wikipedia who-votes-on-whom network): 7,115 nodes, 103,689 edges

Step 1: Run BGLL code on the given data set, and obtain the community structure, i.e. the number of communities identified and the list of nodes falling into every community.

Output:

Number of levels - 3

Level 0: 8298 nodes (Total nodes- 7115)

Level 1: 1237 nodes

Level 2: 1212 nodes

Modularity: 0.427131

Step 2: Once the results of BGLL code applied on the whole network are there, the large sized communities have to be analyzed for further sub-communities. For this, a parameter is to be set which specifies the threshold size of communities. Communities detected in Step1 which are larger than the threshold will be considered as large communities.

Output:

Threshold Size= 355 (5% of Community Size)

Number of Communities with size greater than 355= 4

Community Id of the large communities: 423, 395, 209, 59

(See Table 1).

Table 1: Communities Identified in Wiki-Vote Network which are larger than the Threshold Size

Community Id	Number of nodes falling in Community
423	1189
395	2623
209	1113
59	2009

Step 3: Step1 and Step2 are recursively applied on the large communities identified in Step2.

Output:

Within community Id 423, Community Id 3766 is the big one

Within community Id 395, Community Id 5252, 5104, 4414 are the big ones.

Within community Id 209, Community Id 2453 is the big one

Within community Id 59, Community Id 324, 203, 8 are the big ones.

This procedure clearly brings out the sub-community structure present within the communities, and therefore enables a better community identification method. The recursive application of BGLL algorithm leads to the disintegration of the communities into distinct small sized communities which had been merged together due to the Resolution limit of Modularity.

5. CONCLUSION

This paper revisits the notion of modularity and provides an in-depth analysis of the problems lying in its mathematical formulation. The ring of clique problem also puts forward the theoretical problems lying in the Modularity framework, therefore putting a question mark on the existing Modularity optimization techniques used to identify communities. The author suggests the recursive application of the modularity maximization algorithm with the aim of identifying sub-communities within the communities. The recursive application of BGLL Algorithm is an alternative to the standard BGLL algorithm, and provides a better community structure and therefore, can be extremely useful in the fields of social market analysis, and network analysis in general, where understanding the community structure is important to understand the network as a whole.

We have discussed at length how Recursive BGLL algorithm can be used to identify community patterns in networks. However, the threshold limits need to be defined in the recursive BGLL Algorithm. One of the areas for further improvement is to identify the number of steps up to which the recursion algorithms should be carried out. The current mechanism sets a threshold limit and applies recursion until the sub-community size is greater than the specified threshold. Further work is required to study the threshold limits as applied in the recursive application.

5. REFERENCES

[1] B M. E. J. Newman and M. Girvan, Phys. Rev. E 69, 026113 (2004)

[2] Santo Fortunato and Marc Barthelemy (2007). "Resolution limit in community detection". Proceedings of the National Academy of Sciences of the United States of America

[3] G. W. Flake, S. Lawrence, C. Lee Giles and F. M. Coetzee, IEEE Computer 35(3), 66-71 (2002)

[4] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. 99, 7821-7826 (2002)

[5] K. Eriksen, I. Simonsen, S. Maslov and K. Sneppen, Phys. Rev. Lett. 90, 148701 (2003)

[6] Blondel, V.; Guillaume, J.; Lambiotte, R; Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, IOP Publishing,* 2008

[7] Stanford Large Network Dataset Collection: Epinions Social Network

[8] Stanford Large Network Dataset Collection: Wikipedia vote network

[9] van der Hofstad, Remco (2013). "Chapter 7". Random Graphs and Complex Networks (PDF)

APPENDIX I

'Ring of Cliques' Network

Intra-Clique Edges:
Edges which are incident on nodes, both of which belong to the same clique

Inter-Clique Edges:
Edges which are incident on nodes, which belong to the adjacent cliques

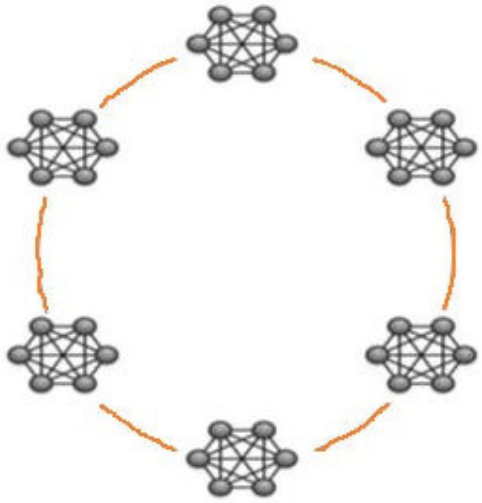


Figure: Ring of Cliques Network with six inter-clique edges shown in orange, and ninety intra-clique edges shown in black