An Enhanced Local Modularity Measure

Fariborz Nahvi
Department of Computer
University of Sheikhbahaei
Faculty of Engineering
Isfahan, Iran
fanahvi@yahoo.com

Mohammad Reza Khayyambashi
Department of Computer
University of Isfahan
Faculty of Engineering
Isfahan, Iran
M.RKhayyambashi@eng.ui.ac.ir

Abstract

Recently, detection of community structure in networks has drawn a lot of attention. In this case, most of the developed methods need global knowledge of the network which is not applicable to real world graphs, since, they are are too large or evolve too quickly .Besides, we may be interested in the community structures of some given nodes, not all nodes .So, detecting the community of a specific node is more appropriate .several local modularity measures have been developed. Amongst, local modularity R works well in case of performance and simple agglomeration mechanism. But it has low recall information retrieval measure due to its predetermined number of agglomerated nodes.

In this paper, we have changed its stopping criteria to multiple regression analysis. Hence, its recall parameter is improved leading to more accurate measure. Moreover, its performance is optimized, because, this new stopping criteria and agglomeration process works simultaneously leading to lower execution time. We validate our method on two real-world networks whose community structures are known .The result shows that our method can achieve higher recall as well as better performance.

Keywords: local community detection, regression analysis stopping criteria, local modularity measure, greedy optimization stopping criteria.

1- Introduction

The rapidly growing interests in complex networks, such as the World Wide Web, citation networks, online social networks, and metabolic networks revealed a common feature of these complex networks i.e. community structure. Community is defined as a group of nodes having higher edges within group comparing to edges between groups. Many community detection algorithms have been developed using greedy optimization of a modularity function .However, it is too hard to get knowledge about the whole graph, due to its evolving nature or being too big, like the Internet .Besides, most of users need knowledge about a specific domain of network. For example, in co-purchasing networks like amazon, somebody may need the knowledge of some books related to specific subject; on the contrary, his friend prefers gaining knowledge about movies of specific genres. Hence, local community is more appropriate in these situations. To address this problem, we have modified R local measure introduced by Clauset[2] and improved its accuracy.

This paper is organized as follows. In section 2, we precede to review some state-of-the-art local community detection methods and measures. In section 3, some well-known information retrieval measures are proposed. In section 4, different stopping criteria are discussed. Our proposed algorithm is explained in section 5. Finally, the proposed algorithm is tested on two real-world networks. We conclude the paper in section 7.

2- Local community identification methods and measures

Local community detection methods [1,2,3,4,7,8] provide a means to alleviate scalability challenges by focusing on a portion of the network. In practice, such methods start a network exploration process from a seed vertex or vertex set and agglomerate adjacent nodes to the community until local community quality measure indicated as fitness function reaches its local maxima. Accordingly, the network around a seed vertex is divided into five sets:

- B: border vertices that are adjacent to at least one vertex of S.
- C: core community members that have no connection to vertices outside the community
- D: union of B and C
- S: vertices that are adjacent to at least one vertex of B
- U: unknown portion of network

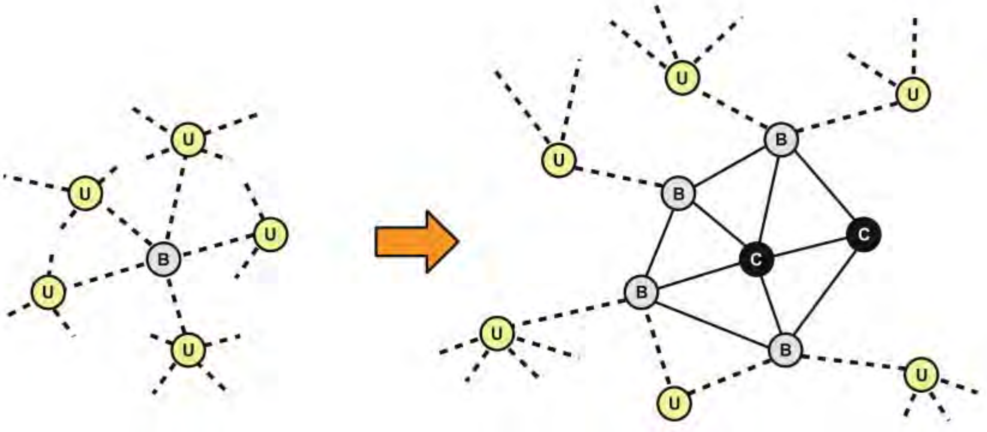


Fig 1 Illustration of core set (C) , Border Set (B) and Unvisited (U) [5]

2-1- **M measure**

In this algorithm, which is also known as LWP, all the nodes of the community having within links are considered in the numerator of the fraction, in contrast to R measure. The denominator is D_{out} stating outward links of community, similar to B_{out} [3].

$$M=\frac{Din}{Dout}$$

Fig 2 M Measure [3]

2-2- **L measure**

In this measure, instead of the number of the internal and external edges, their density is considered. In the following formulas, |D| shows size of the local community of D and |B| is the size of border community nodes B. This method won't agglomerate new nodes into local community just because of increasing in L. otherwise, it decides based on both new L_{in} and L_{ex} values [4]. This prevents agglomerating wrong or outlier nodes to be incorporated into community.

$$L_{in} = \frac{\sum_{i \in D} |\Gamma(i) \cap D|}{|D|}$$

Fig 3 Average internal degree of Nodes in D [4]

$$L_{ex} = \frac{\sum_{i \in B} |\Gamma(i) \cap S|}{|B|}$$

Fig 4 Average external degree of Nodes in B [4]

$$L = \frac{L_{in}}{L_{ex}}$$

Fig 5 Calculating L according to L_{in} and L_{ex} [4]

2-3- **T measure**

Ngonmag [5] proposed T measure in 2012 based on the reforms on the L measure. In addition to reforms in the quality function, he also changed the process of adding new nodes to community and proposed BMEl algorithm. Unlike L algorithm expansion phase, which adds only one node to the local community at each stage, this algorithm adds all of the nodes which increase the quality function as a whole. After finishing expansion phase, algorithm enters to optimization phase and removes nodes which cause negative fitness. He used $(l + d_i)$ implicative factor to avoid detouring searching space if starting node was a bridge node. So, this factor prefers the internal links closer to the starting node. Similarly, it penalizes the nodes which are far from the starting node.

$$T_{in} = \frac{\sum_{i \in D} \frac{|\Gamma(i) \cap D|}{(1+d_i)}}{D}$$

Fig 6 Internal average degree of nodes in D [5]

$$T_{ex} = \frac{\sum_{i \in D} |\Gamma(i) \cap S|(1+d_i)}{D}$$

Fig7 External average degree of nodes in D [5]

$$T = \frac{T_{in}}{T_{ex}}$$

Fig 8 L Measure according to T_{in} and T_{ex} [5]

2-4- **Outwardness measure**

This measure was introduced by Bagrow. Despite its simple mathematic calculation, its algorithm has stopping problem. The fewer outwardness measures, the more optimal is the node to be agglomerated into community. As it is illustrated in the following figure, the local community is surrounded by border nodes. As an example, the value of parameter Ω for node i is 2/3 and for node j it is -1. Thus, node j is chosen to be added to the local community as it will make a denser community structure. In the formula below, K_v is indicative of node v degree, K_v^{out} is the output degree of node v in relation to local community and K_v^{in} is the internal degree of the node [6].

$$\begin{aligned}\Omega_v(C) &= \frac{1}{k_v} \sum_{i \in n(v)} \Big([i \notin C] - [i \in C] \Big) \\ &= \frac{1}{k_v} \left(k_v^{\text{out}} - k_v^{\text{in}} \right)\end{aligned}$$

Fig 9 Outwardness measure of community C over node v[6]

3- Information Retrieval Measures

There is different information retrieval measures used for evaluating accuracy of algorithms when the label of reference dataset is known. Amongst, recall, precision and F-measure are more popular. Precision, measures what percent of positive predicted items, are classified as positive.

$$precision = \frac{tp}{\Pr ediction(+)} = \frac{tp}{tp+fp}$$

Fig 10. Precision measure

Recall, measures the percentage of items which are classified as positive, though, they were positive in reference dataset. This is known as sensitivity measure in some references especially in binary classification.

$$recall = \frac{tp}{Truth(+)} = \frac{tp}{tp+fn}$$

Fig 11. Recall measure

Finally, F-measure is a harmonic measure based on two considered measures above.

4- Different Stopping Criterions

In this section, three different stopping criteria are declared [6].

4-1- **Strong to not**

C is a strong community when every node in it has more neighbors inside (k^{in}_{i}) than outside (K^{out}_{i}) [10]. This criterion works acceptable for fully disjoint separated communities, but fails as the border contrast decreases. In another words, a strong community is too rigorous. If C never becomes strong, the algorithm continues forever, indicating a lack of community structure in the explored part of network.

$$k_i^{\mathrm{in}}(C) > k_i^{\mathrm{out}}(C), \quad \forall i \in C$$

Fig 12. c is a strong community

4-2- **p-Strong**

Community C is p-strong if strong condition applies partly, not for all nodes. This could be taken into account as a generalized notion of "Strong to not" measure. An additional advantage of this measure is that multiple values of p can be used simultaneously, since a community that is p1-strong is also p2 -strong (p1 > p2). This is due to separation of agglomerating and stopping phases letting both processes to be executed simultaneously [6].

$$\sum_{i \in C} \left[k_i^{\mathrm{in}}(C) > k_i^{\mathrm{out}}(C) \right] \geq p\,|C|$$

Fig 13. P-strong measure definition [6]

4-3- **Regression analysis**

In this method, the stopping criterion predicts LMF, with a regression function. Suppose n nodes have already been agglomerated and local modularity function is predicted using y polynomial of order two.

$$y = ax^2+bx+c$$

Fig 14. regression function of LMF

We conclude that local modularity around a specific node is found, if all the situations below are satisfied simultaneously. Hence, agglomerating should stop and final three nodes removed.

- $a < 0$
- $LMF(i) > y(i)$, $i = n, n - 1, n – 2$.
- $n - 3 > -b/2a$.
- $LMF (n) \geq LMF (n - 1) \geq LMF (n - 2)$.

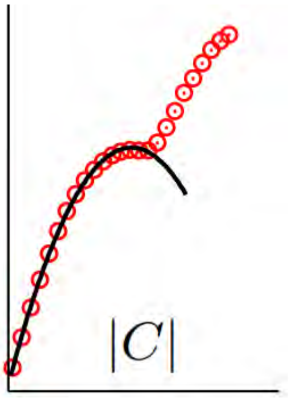


Fig 15. LMF vs regression function

As shown in figure above, when you pass the border of the community, LMF shown as red small circles, will start to increase, while the parabola, unaware of the next three values, continues downward.

5- Proposed Algorithm

Clauset [2] proposed an idea that the border nodes of the community should have more links with their own local community than the neighbor nodes.

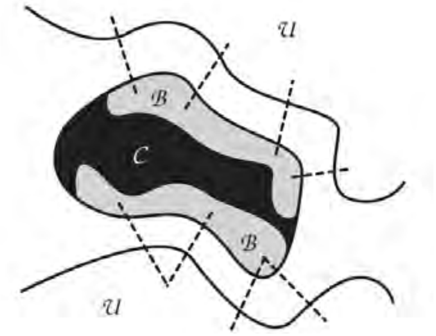


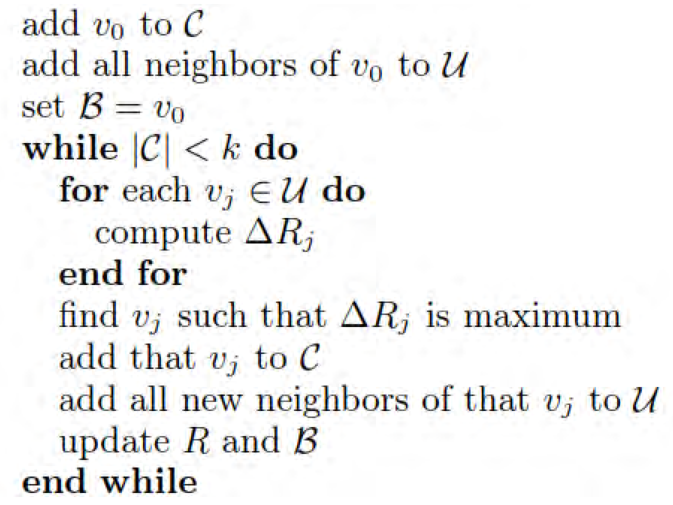
Fig 16. different parts of local community C

This criterion recognized as local modularity R and is defined as below. In its formula, B_{in} is the number of links between B and C and B_{out} is the number of links between B and S.

$$R = \frac{Bin}{Bin + Bout}$$

Fig 17 R Measure [2]

According to the comparison [9], R measure has a good performance but it has low recall parameter due to its predetermined number of nodes k as illustrated in algorithm below.



```
add $v_0$ to $\mathcal{C}$
add all neighbors of $v_0$ to $\mathcal{U}$
set $\mathcal{B} = v_0$
while $|\mathcal{C}| < k$ do
   for each $v_j \in \mathcal{U}$ do
      compute $\Delta R_j$
   end for
   find $v_j$ such that $\Delta R_j$ is maximum
   add that $v_j$ to $\mathcal{C}$
   add all new neighbors of that $v_j$ to $\mathcal{U}$
   update $R$ and $\mathcal{B}$
end while
```

Fig 18 Greedy optimization algorithm of R [2]

This measure, stops in either of these two conditions:

- If number of agglomerated nodes exceeds parameter k.
- The entire enclosing component discovered resulting that R reaches one as maximum possible value.

As stated by Wu et al [9], input parameter k as stopping condition results in low recall of this measure. One possibility to improve this measure is to change its stopping condition. So, in this paper, we use method suggested in [6] which estimates modularity measure by linear regression.

5-1- **Proposed Stopping criteria**

According to regression analysis method described beforehand [6], R measure is estimated based on multiple regression with two predictor variable as shown below.

$$Rk = \alpha Bin + \beta Bout + ek$$

Fig 19 multiple regression of R measure

For finding the best value of R, SSE method is used to minimize difference between the calculated R and the estimate R_k. When their difference becomes minimal, the best value of R is specified implying that the community is found.

6- Evaluation

In this section, proposed algorithm has been applied on two real world ground-truth networks and the recall improvement is evaluated.

6-1- **American College Football Team**

In this medium-scale network of 115 teams identified as NCAA, two teams are connected through an edge if they compete to each other. Teams are divided into twelve groups. Two teams are more probable to hold a competition comparing to teams of different groups. Hence, every group could be identified as a cluster. This network is kind of ground-truth one which could be used for measuring accuracy of the proposed algorithm. The proposed measure is tested on this dataset revealing results shown in diagram below. As you could see, recall parameter of modified R measure is improved comparing to original one.

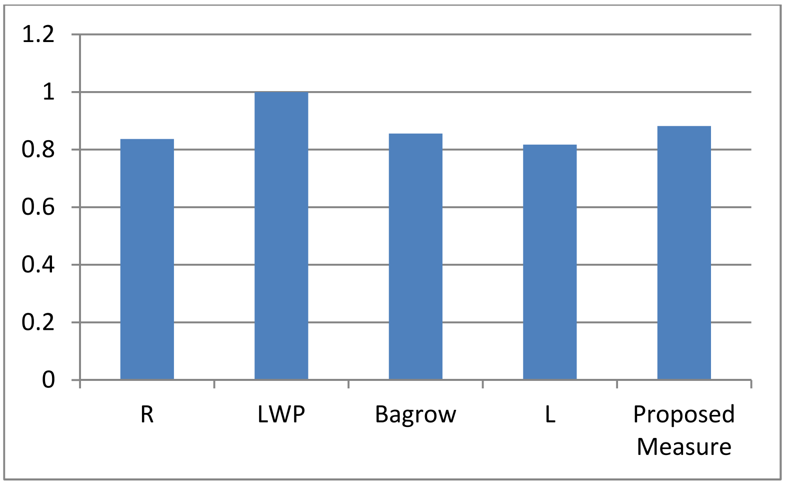


Fig 20. Comparing recall parameter of local algorithms on NCAA

6-2- **Zachary Karate Club network**

This network is constructed according to studies by Zachary sociologist. This club is divided to two parts according to argue between administrator and instructor. This network has 34 nodes and 78 edges. As shown in diagrams below, although L measure has good performance, its rigorous policy on controlling density of community resulted in low recall parameter.

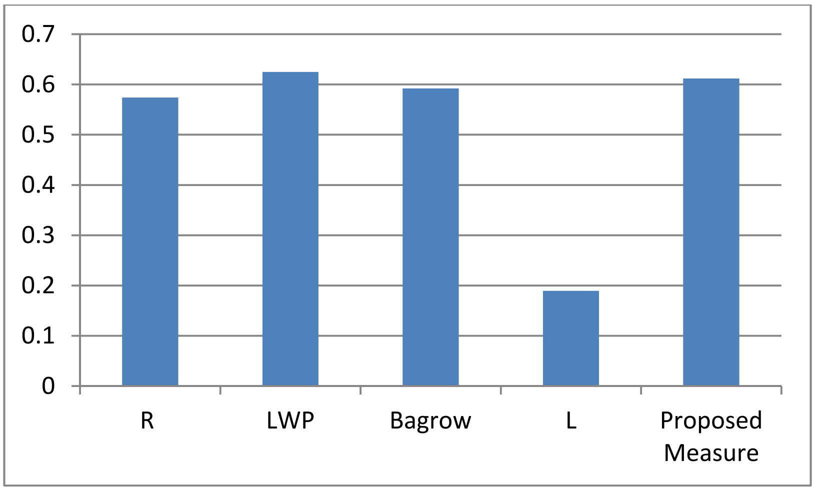


Fig 21. Comparing recall parameter of local algorithms on Zachary

7- Conclusion

In this paper, we have modified local modularity R. In contrast to good performance of this local measurer, it suffers from low recall. So, we changed stopping condition from predetermined input k to multiple regressions with two predictor variable formula. Besides reaching better recall parameter, algorithm could execute agglomeration and stopping phase simultaneously. In case of future, adding the ability of detecting outliers is suggested.

References

[1] S. Papadopoulos, A. Skusa, and N. Wagner, “Bridge Bounding : A Local Approach for Efficient Community Discovery in Complex Networks.”,ACM,2009.

[2] A. Clauset, “Finding local community structure in networks,” Physical Review E, vol. 72, no. 2, p. 026132, 2005.

[3] F. Luo, J. Wang, and E. Promislow, “Exploring Local Community Structures in Large Networks,” 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI’06), pp. 233–239, Dec. 2006.

[4] J. Chen, O. R. Zaiane, and R. Goebel, “Local Community Identification in Social Networks.” IEEE International Conference on Advances in Social Network Analysis and Mining,ASONAM09,2009,pp.237-242.

[5] B. Ngonmang, M. Tchuente, and E. Viennet, “Local Community Identification in Social Networks,” Parallel Processing Letters, vol. 22, no. 01, p. 1240004, Mar. 2012.

[6] J. P. Bagrow, “Evaluating Local Community Methods in Networks,” Journal of Statistical Mechanics: Theory and Experiment ,2008.

[7] J. Chen, O. R. Za, and S. R. Goebel, “ONDOCS : Ordering Nodes to Detect Overlapping Community Structure,” Data Mining for Social Network Data, vol. 12, pp. 125–148,2010.

[8] G. W. Flake, S. Lawrence, and C. L. Giles, “Efficient identification of Web communities,” Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’00, pp. 150–160, 2000.

[9] Y.-J. Wu, H. Huang, Z.-F. Hao, and F. Chen, “Local Community Detection Using Link Similarity,” Journal of Computer Science and Technology, vol. 27, no. 6, pp. 1261–1268, 2012.

[10] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, “Defining and identifying communities in networks.,” Proceedings of the National Academy of Sciences of the United States of America, vol. 101, no. 9, pp. 2658–63, Mar. 2004.