EPiC Series in Computing

Volume 91, 2023, Pages 138–145

Proceedings of 38th International Conference
on Computers and Their Applications

Overlapping Communities Discovery and Improvement

Rong Yang
Western Kentucky University
Bowling Green, KY 42101
Email: Rong.Yang@wku.edu

Abstract

Decomposing a network into communities is one of the most used techniques in network science. Modularity is typically used to measure the goodness of such a decomposition. In this paper we develop a method which allows us to begin with a crisp decomposition (no overlaps) and move to an overlapping decomposition while increasing the modularity. We also show that the same technique can be used to improve existing overlapping decompositions.

1 Introduction

The notion of decomposing a network into communities (or clusters, modules) in order to better understand its structural makeup has been an active part of network science since its inception. Indeed, in their seminal work [20], Wasserman and Faust spend two full chapters on subgroups, the predecessors of what we now call communities. Since in a real life network one generally has no real idea of how many communities there are or how big they are, discovering these groupings is a primary focus. Much of the considerable effort put toward community detection has been in finding crisp (or hard) decompositions, meaning that the vertex set is broken into a partition so that each vertex belongs to one and only one of the blocks of the partition (community). The widely used IGraph package for R includes no fewer than nine algorithms for crisp community detection [11, 4, 17, 18, 15, 10, 2, 3, 16, 19, 14].

More recently, with the recognition that naturally occurring communities are often not disjoint, attention has been focused on less restrictive decompositions having either overlapping communities or fuzzy communities (or both). A decomposition with overlapping communities corresponds to what is mathematically called a cover of the vertex set and in this case a given vertex may belong to several communities. There are several competing ideas about what fuzzy decomposition ought to be, but one common definition is that each vertex has a certain probability of belonging to a given community [9, 6].

The aim of this work is to develop a method whereby we can take the output from one of the many available crisp community detection algorithms and use it to discover associated overlapping decompositions. These overlapping decompositions should represent in some sense (specifically using a notion of modularity) an improvement over the crisp decomposition which gave rise to them. The same technique is also shown to be useful for improving existing overlapping decompositions.

2 Notations And Terminology

Now we introduce the notations and terminology which we will use throughout the discussion. The networks considered here are unweighted and undirected graphs $G = (V, E)$ with $|V| = n$ vertices and $|E| = m$ edges. It is assumed that they have no loops and no multiple edges. $A = (a_{v,w})$ is the adjacency matrix for G and $\|A\| = \sum\limits_{i,j=1\ldots n} a_{i,j}$. The degree sequence for G is d$_G$= $(d_{v_1}, \ldots, d_{v_n})$. Let $C = (C_1, \ldots, C_r)$ be a cover of V, meaning that $\bigcup_{k=1,k} C_k = V$. Then n_k denotes the number of vertices in C_k, m_k denotes the number of edges which C_k contains (when considered as the subgraph induced by its set of vertices), and s_v denotes the number of communities to which v belongs.

3 Two Types Of Modularity

The idea of modularity was originally created by Newman and Girvin in [11] as a function for quantifying the quality of a community decomposition. This formulation only applies to crisp decompositions, and it basically compares the density of edges within a community with the expected density with respect to a chosen null model, usually the ErdÃ¶s-RÃ©nyi random graph. This NG-modularity varies between -1 and 1 and the higher the value, the better the decomposition. Since Newman and Girvin's work, a number of alternate definitions or extensions of modularity have been put forward, many of them designed to handle overlapping and fuzzy communities as well as crisp ones. For the purposes of this paper, we are particularly interested in two of them.

LAV Modularity

This version of modularity was introduced by Lazar, Abel, and Vicsek in [7] (hence the acronym in its name). To summarize their approach (with some notational changes),

$$\Delta_{v,k} = \frac{\sum\limits_{w \in C_k} a_{v,w} - \sum\limits_{w \notin C_k} a_{v,w}}{d_v \cdot s_v}$$

is interpreted as the contribution of v to the modularity of community C_k. Note that the first summation in the numerator is the number of connections v has with members of C_k while the second is the number of connections it has outside of C_k. This provides a kind of measure for the intuitive notion that a valid community should have many internal connections and few external ones. Dividing by the degree of the vertex serves to normalize the measure and dividing by the number of communities to which v belongs is intended to prevent situations where v belongs to many almost identical communities. The overall contribution of community C_k is then given by

$$M_k = \frac{1}{n_k} \sum_{v \in C_k} \Delta_{vk} \cdot \frac{m_k}{\binom{n_k}{2}}$$

Here the final factor gives the density of community C_k and the entire expression is normalized by dividing by the number of vertices in C_k. Finally, the overall LAV modularity is defined to be the average of the contributions of the communities:

$$M\left(G,C\right)=\frac{1}{r}\sum_{k=1\ldots r}M_k$$

This does not claim to be an extension of Newman-Girvin modularity to include overlapping communities (and indeed it is not such an extension). Nevertheless, it is a reasonable and intuitively appealing measure of the goodness of a decomposition. To put this in a concrete light, we will use a TriTail graph which is very simple, yet has a characteristic of a node that belongs to overlapping communities as an example. Consider the following small TriTail graph with a crisp community decomposition shown in Figure 1, it is easy to see that

$$\Delta=\begin{bmatrix}0 & 0\\ 0 & 0\\ 1/3 & -1/3\\ -1 & 1\end{bmatrix}$$

and we then have $M_1=0$, $M_2=\frac{1}{3}$, and $M=\frac{1}{6}=0.1666667$.

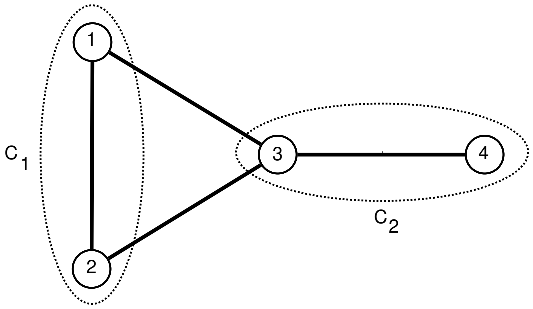

Figure 1: The TriTail Graph With A Crisp Decomposition

Soft Modularity

Soft modularity was invented by Havens, Bezdek, Leckie, Ramamohanarao, and Palaniswami in [8]. This approach essentially follows from the observation that classical Newman-Girvin modularity can be defined in a matrix formulation which can in fact be applied to any decomposition - crisp, overlapping, or fuzzy. It thus represents a true generalization of Newman-Girvin modularity and can be usefully employed to compare modularities across all types of decompositions. It is thus particularly good for comparing modularities across different decompositions of whatever variety since its connection with classical modularity is so clear. Our ultimate goal is to make changes which will improve the soft modularity. The details (specialized to our situation) are as follows. Let

$$B=A-\frac{\left(d_G\right)^T\cdot d_G}{\|A\|}$$

where the exponent T denotes the transpose. We represent the communities by an $n\times r$ matrix $U=(u_{v,k})$ where $u_{v,k}=1$ if vertex v belongs to community C_k and $u_{v,k}=0$ otherwise. Then this extended or soft modularity is given by

$$Q = \frac{trace\left(U^T \cdot B \cdot U\right)}{\|A\|}$$

In the case of the TriTail graph and its crisp decomposition shown above, we have

$$B = \begin{pmatrix} -1/2 & 1/2 & 1/4 & -1/4 \\ 1/2 & -1/2 & 1/4 & -1/4 \\ 1/4 & 1/4 & -9/8 & 5/8 \\ -1/4 & -1/4 & 5/8 & -1/8 \end{pmatrix}$$

and

$$U^T \cdot B \cdot U = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

so that Q = 0.

An Empirical Comparison

LAV modularity and soft modularity are mathematically quite distinct. As we have indicated, soft modularity is a genuine extension of Newman-Girvin modularity, while LAV modularity represents an independent heuristic approach. Thus it would be difficult to establish a direct relationship between the two. However, using a selection of real-world networks which commonly appear in network science, we compared the two modularities as shown in Figure 2. The Pearson correlation between the two sets of modularity values is 0.5495717, so while not the same, they are clearly correlated.

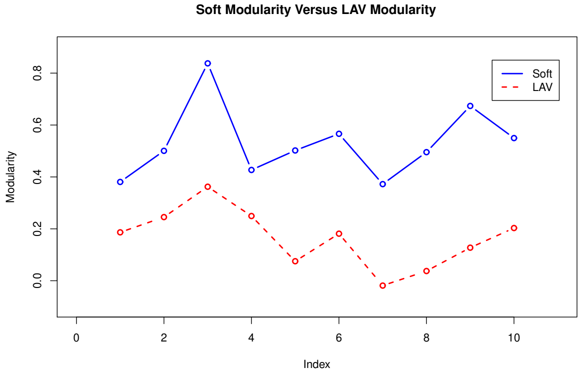

Figure 2: Comparing LAV and Soft Modularity

4 Improving A Decomposition

Having seen the correlation between LAV modularity and soft modularity, it is reasonable to expect that increasing the LAV modularity will also result in an increase for soft modularity. our approach toward moving from a crisp decomposition to an overlapping one (or to improving a given decomposition of whichever kind) is to attempt to identify vertices which would make a positive contribution to the overall modularity if they were added to some new community. For this purpose, the matrix Δ produced by the LAV modularity computation is ideal since it ascribes a modularity contribution to each vertex-community pair. Lazar, Abel, and Vicsek were primarily interested in $\Delta_{v,k}$ in the case when $v \in C_k$ and they interpreted it as a measure of how "justifiable" it is to assign v to the community C_k. Here we are looking at $\Delta_{v,k}$, whether v belongs to C_k or not, as an indicator of the contribution to overall modularity that putting v into C_k would effect.

In order to increase the LAV modularity, we examine the matrix Δ row by row. If, in row v we discover a positive entry $\Delta_{v,k}$ and v does not belong to C_k, we add v to C_k (while leaving it in any communities it already belonged to). Once a pass through Δ is complete, Δ is reevaluated and the process repeated until a pass results in no vertex movements.

For example, with the crisp decomposition for the TriTail given earlier, $\Delta_{3,1} = 1/3 > 0$ and vertex 3 does not belong to community C_1, so we add vertex 3 to community C_1, which is the only movement from pass 1 through Δ. The result is the overlapping decomposition shown in Figure 3. This increases the LAV modularity from its initial value of 0.1666667 to 0.5694444. At the same time, the change decreases the soft modularity from its initial value of 0 to -0.015625 (which shows the correlation between the two measures of modularity is not perfect).

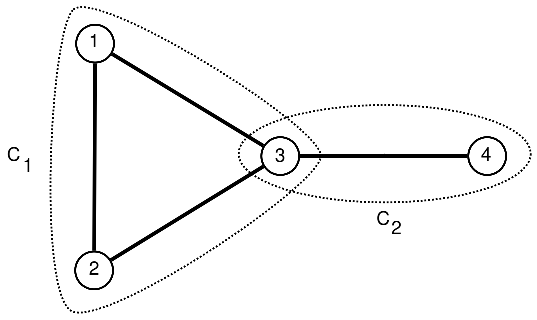

Figure 3: The TriTail Graph With An Overlapping Decomposition

For the second pass, the Δ is

$$\Delta = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1/6 & -1/6 \\ 1 & 1 \end{pmatrix}$$

where $\Delta_{4,1} > 0$ indicates that vertex 4 should be added to community C_1. When we do this, we get the decomposition shown in Figure 4, which shows that the technique can not only produce overlapping communities, but decompositions containing embedded communities as well. For this decomposition, the LAV modularity is 0.3333333 (higher than the original modularity, but lower than it was after just one pass) and the soft modularity is 0, which represents an improvement over the value at the end of pass 1 and in fact a return to the value that it had for the original crisp decomposition. A third pass yields no vertex movement. Bear in mind that the behavior in this very small artificial example is not typical - the example is intended merely to insure that the concepts involved are clear.

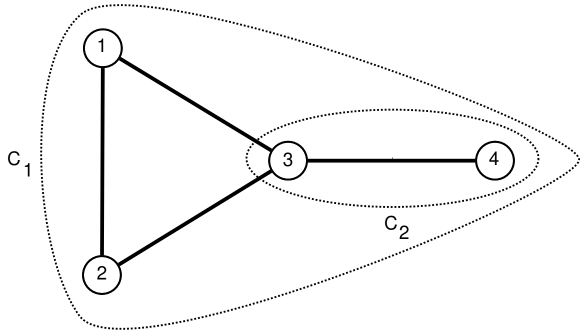

Figure 4: The TriTail Graph And Its Ultimate Decomposition

5 Applications

To test the efficacy of our approach, we employed two methods, both based on real life networks. In the first group of tests, we obtained a crisp decomposition using the IGraph library for R and its cluster_fast_greedy method. With that decomposition as a starting point, the above "optimization" technique was then used to obtain a (possibly overlappinng) decomposition. Of course, in some cases the crisp decomposition is quite appropriate to the network structure and no changes are made to that decomposition. This happened, for example, with the well known Zachary's Karate Club network [22, 12]. More typical is the C. Elegans Neural Network [21, 12]. Here our technique resulted in a 4.5% increase in soft modularity and the overlap percentage (the percent of vertices which belong to more than one community) went from 0% for the crisp fast greedy decomposition to 1.7%.

More impressive results were obtained by beginning, not with a crisp decomposition, but with a decomposition produced by CFinder [1, 13]. This freely available software uses the clique percolation method [5] to identify communities (possibly overlapping) in networks. In fact, it usually produces a number of different possible sets of communities based on the value of a parameter k which specifies the size of clique to begin with. For testing, we chose the value of k which gave rise to the highest soft modularity among the CFinder decompositions. Due to the way the CFinder algorithm works, there are, in general, some vertices which are not assigned to any community. Since our technique requires that every vertex must belong to at least one community, there are two obvious things that can be done with these "leftover" vertices. First, we can put each one of them into its own new community. Second, we can group them all together into a single new community. The second method gives better results in almost all cases, so that is what we discuss here. Using the C. Elegans Neural Network as before, the soft modularity increases from an initial value of 0.1983418 for the CFinder decomposition to 0.4047188 (a 104% increase). At the same time, the overlap percentage goes from 27% to 56%. Naturally such dramatic increases cannot always be achieved, but with the networks we studied, double digit increases in both soft modularity and overlap percentage were common.

6 Conclusions

In this paper we introduce and discuss a method that we developed to improve modularity. The method begins with a crisp decomposition (no overlaps) and moves to an overlapping decomposition while increasing the modularity. The tests we conducted using some real life networks have shown that our method improves not only soft modularity but the existing overlapping decompositions as well. In the future, we plan to test the method on more different variety of complex networks using other available crisp community detection algorithms and to use relevant different modularity measurements besides LAV and Soft that were used in this paper.

References

[1] http://www.cfinder.org/, December 2020.

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

[3] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Gorke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2):172–188, 2008.

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

[5] Imre Derényi, Gergely Palla, and Tamás Vicsek. Clique percolation in random networks. *Physical review letters*, 94(16):160202, 2005.

[6] Steve Gregory. Fuzzy overlapping communities in networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(02):P02017, 2011.

[7] Anna Lázár, Dániel Abel, and Tamás Vicsek. Modularity measure of networks with overlapping communities. *EPL (Europhysics Letters)*, 90(1):18001, 2010.

[8] Christopher Leckie and Kotagiri Ramamohanarao. A soft modularity function for detecting fuzzy communities in social networks.

[9] Tamás Nepusz, Andrea Petróczi, László Négyessy, and Fülöp Bazsó. Fuzzy communities and the concept of bridgeness in complex networks. *Physical Review E*, 77(1):016107, 2008.

[10] Mark EJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104, 2006.

[11] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[12] M.E.J. Newman Network Data Page. http://www-personal.umich.edu/ mejn/netdata/, December 2016.

[13] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435@articlederenyi2005clique, title=Clique percolation in random networks, author=Derényi, Imre and Palla, Gergely and Vicsek, Tamás, journal=Physical review letters, volume=94, number=16, pages=160202, year=2005, publisher=APS (7043):814–818, 2005.

[14] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. In *International Symposium on Computer and Information Sciences*, pages 284–293. Springer, 2005.

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

[16] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1):016110, 2006.

[17] M Rosvall and CT Bergstrom. Maps of information flow reveal community structure in complex networks. Technical report, Citeseer, 2007.

[18] Martin Rosvall, Daniel Axelsson, and Carl T Bergstrom. The map equation. *The European Physical Journal Special Topics*, 178(1):13–23, 2009.

[19] Vincent A Traag and Jeroen Bruggeman. Community detection in networks with positive and negative links. *Physical Review E*, 80(3):036115, 2009.

[20] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[21] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-world networks. *nature*, 393(6684):440–442, 1998.

[22] Wayne W Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, pages 452–473, 1977.