

On Inconsistencies in Quantifying Strength of Community Structures

Wen Haw Chong

DSO National Laboratories
20 Science Park Drive Singapore 118230

cwenhaw@dso.org.sg

Abstract

Complex network analysis involves the study of the properties of various real world networks. In this broad field, research on community structures forms an important sub area. The strength of community structure is typically quantified by the modularity measure. The measure is based on summing the differences in actual and expected fraction of edges per community (across all communities in the network), whereby the latter is computed based on randomizing the edges subjected to certain constraints. In this paper, we investigate the differences between two commonly used definitions of modularity and highlight one of them as inadequate for quantifying the strength of community structures. We first show this by mathematical proving. We then investigate the empirical differences by developing and testing two variants of a community detection algorithm whereby the variants differ based on their modularity definitions. We observe varying differences in detection accuracy when applying the variants on artificially generated networks. For networks with strong community structures, we show that sensible results are still obtainable with the inadequate measure, which explains why this issue did not come to light previously.

Keywords: community structure, networks, modularity.

1 Introduction

In the real world, a variety of networks exist and their statistical, mechanical and temporal properties are the subject of much research, known broadly as complex network analysis. An important sub-area in the study of complex networks is that of community detection. In various real world networks, communities are of practical interest. For example, communities may be indicative of social groups or cliques in a network of relationships. They may also be indicative of topics of common interest as in the case of the world-wide web (Kleinberg, and Lawrence 2001). In the case of biochemical networks, they may correspond to functional units (Ravasz, et al. 2002). Other examples of community structures can be found in collaboration networks, computer networks and food webs, etc.

Formally, a community is defined as a group which has more than the expected number of links between its members, whereby ‘expected’ means by random chance. Hence connections within communities are expected to be much denser than between the communities. To quantify the strength of community structures, a measure known as modularity was first proposed by Newman and Girvan (2004). This remains a successful and widely used measure. The modularity measure quantifies the idea that community structures should result in fewer than expected number of inter-community edges and conversely, higher than expected number of intra-community edges in the network. Many recently developed community detection algorithms thus focus on deriving groupings that maximizes modularity. The general consensus was that the best algorithm returns the maximum modularity on real world networks. The problem itself is an NP hard problem with ongoing active research in more efficient and accurate algorithms.

In the next section, we shall describe the commonly used definitions of modularity. Section 3 analyse one of the definitions mathematically. We derive the algorithm optimizing each definition in section 4. Section 5 presents the experimental measure used while section 6 presents the experiments. Finally we conclude in section 7.

2 Modularity Definitions

Given a network with n nodes and m edges (or links), let \mathbf{A} be the corresponding adjacency matrix, whose element A_{ij} is the edge weight between nodes i and j . This is either 1 or 0 for unweighted networks and may be of any other values for weighted networks. Let k_i be the degree of node i . We further denote the community containing node i as its parent community c_i .

We bring the reader’s attention to two definitions of modularity commonly occurring in the literature, which we denote as $Q1$ and $Q2$. To avoid any ambiguity, we also list all related and alternative forms that are equivalent for each definition. The first form (Clauset, et al. 2004; Newman, 2006a, 2006b; Fortunato, and Barthélemy 2007) assumes that the expected number of edges between nodes i and j if edges are placed at random is $k_i k_j / 2m$. Modularity is then quantified by the difference between the actual and expected number of edges:

$$Q1 = \frac{1}{2m} \sum_{ij} [A_{ij} - k_i k_j / 2m] \delta(c_i, c_j) \quad (1)$$

where c_i and c_j are the parent communities of nodes i and j respectively and $\delta(c_i, c_j)$ is 1 if c_i equals c_j and is 0 otherwise. If the network has been divided into p communities, then a symmetric $p \times p$ matrix \mathbf{e} can be

defined whose element e_{vw} is the fraction of all edges that link nodes in community v to nodes in community w . The diagonal elements e_{vv} is then simply the fraction of edges falling within each community. $Q1$ has been shown (Clauset, et al. 2004) to be equivalent to

$$\sum_v (e_{vv} - (d_v / 2m)^2) \quad (2)$$

where d_v is the sum of degrees of nodes in community v . Hence $Q1$ can be alternatively computed based on summation of terms over communities, instead of considering node pairs. This is very similar in form to the other modularity definition to be introduced.

The alternative and also commonly used definition for modularity (Newman, and Girvan 2004; Newman 2004; Duch, and Arenas 2005; Ruan, and Zhang 2006) is:

$$Q2 = \sum_v (e_{vv} - a_v^2) \quad (3)$$

where $a_v = \sum_w e_{vw}$ is the fraction of edges which has at least one end in community v . In matrix notation, $Q2$ can be written as $Tr(e) - \|e\|^2$ where $Tr(\cdot)$ is the trace of the matrix and $\|x\|$ indicates the sum of the elements of matrix x . It has generally been assumed that this definition is roughly equivalent to the first, with the same null model, i.e. if there is no community structure, the number of expected links between two nodes are proportional to the product of their degrees. However on closer examination, it is found that the formula does not follow this assumption.

3 Analysis

In this section, we analyse the measure $Q2$ mathematically for an unweighted network. For node i in community c_i , we can divide its edges into those that link to other nodes in the same community, i.e. the inner degrees $k_{i,in}$ and those that link to nodes outside the community, i.e. the outer degrees $k_{i,out}$. It is also obvious that $k_i = k_{i,in} + k_{i,out}$. From inspection, given any community (indexed by v), we can write the number of edges with 1 or both ends in it as

$$\sum_i (k_i - 0.5k_{i,in}) \delta(c_i, v). \quad (4)$$

This is equivalent to summing the degrees of all nodes in the community with the inner degrees weighted by half. Correspondingly,

$$a_v = \frac{1}{m} \sum_i (k_i - 0.5k_{i,in}) \delta(c_i, v) \quad (5)$$

Substituting equation (5) into the definition for $Q2$, we have

$$\begin{aligned} \sum_v (e_{vv} - a_v^2) &= \sum_v \left[\frac{1}{2m} \sum_{ij} A_{ij} \delta(c_i, v) \delta(c_j, v) \right. \\ &\quad \left. - \frac{1}{m} \sum_i (k_i - 0.5k_{i,in}) \delta(c_i, v) \frac{1}{m} \sum_j (k_j - 0.5k_{j,in}) \delta(c_j, v) \right] \\ &= \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{2}{m} (k_i - 0.5k_{i,in})(k_j - 0.5k_{j,in}) \right] \delta(c_i, c_j) \end{aligned} \quad (6)$$

where we have made use of the fact that $\sum_v \delta(c_i, v) \delta(c_j, v) = \delta(c_i, c_j)$.

Hence the second term that is being subtracted and which supposedly corresponds to the null model has undesired terms dependent on the community structures, specifically the inner degrees of each community. This departs from the null model. Nonetheless, if the network of application has strong community structures, we shall see that the communities may still be detected satisfactorily by maximizing $Q2$. This is probably why the inadequacies of $Q2$ were not detected and highlighted earlier in the literature. Another reason is that community detection algorithms are typically applied on real world networks and compared on the basis of their modularity values according to one definition. For the issue to manifest itself, this will require comparing two versions of the same algorithm, one optimizing each modularity definition. In our experiments, we adopt this approach.

4 Algorithm

We select a current community detection algorithm that also proves easy to customize for each modularity definition. Blondel et al. (2008) has proposed an algorithm that they have shown to be extremely fast and feasible for application on networks of up to a billion nodes, constrained only by the amount of memory available for computation. The algorithm is readily applicable on both weighted and unweighted networks. We derive two variants of the algorithm corresponding to $Q1$ and $Q2$.

The algorithm starts with each node defined as a community. The steps are as follows:

1. For each node i , move it out of its own community to its neighbor's community where modularity gain is positive and maximum. If no positive gain is possible, node i remains in its community.
2. Repeat step 1 over all nodes multiple times until modularity has converged to its maximum value.
3. Construct a new network whose nodes now represent the communities. The links between the new nodes are calculated as the sum of the link weights between nodes in the corresponding communities.

Repeat the whole process until there are no possible node movements (in step 1) which will increase modularity. At this stage, the algorithm terminates.

Note that in this algorithm, there is no merging of communities, only inter-community movements of nodes. Communities which end up with no nodes are then dropped from subsequent processing. This differs from earlier agglomerative merging approaches (Clauset, Newman and Moore 2004). The two variants described in this paper differ from each other only in their modularity gain formula in step 1. Both our variants also differ slightly from the original algorithm¹ described by Blondel

¹ It was not clear which modularity definition was optimized. The authors have indicated clarification of their formula in subsequent versions of the paper.

et al. (2008) in the sense that the modularity gain formula in step 1 is different.

4.1 Variant 1

Variant 1 optimizes $Q1$. We show the derivations for unweighted networks. It can easily be generalized to weighted networks. In calculating the modularity gain formula, we need to consider modularity changes both when a node leaves its community and when it joins another community. Let Q_{leave} be the modularity change for a node leaving its community and Q_{join} be the modularity change when the same node joins another community. Then for a single node movement, $\Delta Q1 = Q_{join} - Q_{leave}$.

Consider the case when a node i leaves its current community w to join another community v . In the joining process, new summation terms arise in the equation for $Q1$ to constitute the change in modularity. This can be calculated as

$$\begin{aligned} Q_{join} &= 2 \times \frac{1}{2m} \sum_j (A_{ij} - k_i k_j / 2m) \delta(c_j, v) \\ &= \frac{1}{m} \sum_j A_{ij} \delta(c_j, v) - \frac{2}{(2m)^2} \sum_j k_i k_j \delta(c_j, v) \\ &= \frac{k_{i,in}}{m} - \frac{2k_i}{(2m)^2} \left(\sum_j k_{j,in} \delta(c_j, v) + \sum_j k_{j,out} \delta(c_j, v) \right) \\ &= \frac{k_{i,in}}{m} - \frac{k_i \sum_{in}}{m^2} - \frac{2k_i \sum_{out}}{(2m)^2} \end{aligned} \quad (7)$$

where \sum_{in} is the sum of weights of links inside v and \sum_{out} is the sum of weights of links incident to nodes in v . Similarly, during the leaving process, certain terms drop off from the summation for $Q1$ as follows.

$$Q_{leave} = \frac{k_{i,in'}}{m} - \frac{2k_i}{(2m)^2} \sum_{j \neq i} k_j \delta(c_j, w) \quad (8)$$

where $k_{i,in'}$ is the inner degree of node i in community w .

4.2 Variant 2

For variant 2, $Q2$ is maximized. For the same leaving and joining scenario, we simply need to update quantities corresponding to the two affected communities, i.e. e_{vv} , a_v , e_{ww} and a_w . Let the updated quantities after the node movement be tagged with a subscript $*$. Then the modularity gain can be easily calculated as

$$\begin{aligned} \Delta Q2 &= e_{vv*} - a_{v*} + e_{ww*} - a_{w*} \\ &\quad - (e_{vv} - a_v + e_{ww} - a_w) \end{aligned} \quad (9)$$

5 Distance Measure

We use an entropy based measure, the Variation of Information (VI) to quantify the distances between two sets of partitions, which in this case are the sets of actual and detected communities. This measure is a true metric on the space of community assignments with all the properties of a proper distance measure. The variation of information has also been advocated by Karrer et al. (2008) for measuring the distance between two community sets.

Formally, given two different partitioning of the network, C with p communities and C' with p'

communities, let there be n_v nodes in the v th community of partition C and n'_w nodes in the w th community of C' . Denote the size of the intersection, i.e. the number of nodes that are common to both communities as n_{vw} . The VI distance is then defined as

$$d'_{VI}(C, C') = H(C) + H(C') - 2I(C, C') \quad (10)$$

where the entropy of community set C is:

$$H(C) = - \sum_{v=1}^p \frac{n_v}{n} \log \frac{n_v}{n} \quad (11)$$

and the mutual information between the two sets of communities is:

$$I(C, C') = \sum_{v=1}^p \sum_{w'=1}^{p'} \frac{n_{vw'}}{n} \log \frac{n_{vw'}}{n} \frac{n}{n_v} \frac{n}{n'_{w'}} \quad (12)$$

The maximum VI distance achievable between two community sets is $\log n$ which happens when one assignment places all nodes in one community and the other places each node individually in a community of its own. Accordingly the VI distance can be normalized to between 0 and 1 via the following:

$$d_{VI}(C, C') = d'_{VI}(C, C') / \log n \quad (13)$$

In all our experiments, we have tabulated the $d_{VI}(C, C')$ values. The smaller $d_{VI}(C, C')$ is, the closer the match between the detected and actual communities, and the more accurate the community detection algorithm is. For identical match, $d_{VI}(C, C') = 0$.

6 Experiments

It is not straight forward to compare the variants on real world networks. Due to the absence of a ground truth partition for comparing the derived communities against, it is difficult to gauge the quality of the partition found. An easier approach is to apply the variants on artificial networks where we know the actual partition.

We apply both algorithm variants on artificial networks, such that we can systematically vary the strength of the community structure. We generate unweighted networks of 200 nodes with 4 communities of 50 nodes each. Edges are placed probabilistically between pairs of nodes by considering whether they are in the same or different communities.

Let the probability of connection between two nodes in the same community be pin . For nodes in different communities, let the connection probability be $pout$. For community structures to be evident, pin should be higher than $pout$ to obtain more intra-community edges than inter-community edges. If the difference is large, then there is a strong community structure. Also pin can be varied to obtain weakly or strongly connected communities.

We have tested pin values of 0.2 to 0.6 in steps of 0.1. For each pin value, we let $pout = factor \times pin$ and vary $factor$ from 0.1 to 0.5 in steps of 0.1. This provides a systematic way to vary the strength of community structures from strong to weak.

For each combination of pin and $pout$ values, we generate 5 networks for trials. We have observed that for the range of pin values tested, detection accuracy behaves in the same manner and generally decreases as $pout$ is

increased. Hence it suffices to examine the boundary cases corresponding to the strongest and weakest connected communities. In table 1, we display the mean and standard deviations of the normalized VI distances between the actual and derived communities (over 5 trials for each combination) for the cases of $pin = 0.2$ and $pin = 0.6$. The means are plotted in figure 1 for easy comparison. For comparison purpose, we have also included the results from the leading eigenvector (EV) method (with no fine tuning) for community detection. This algorithm was proposed by Newman (2006b) and maximizes $Q1$.

$pin = 0.2$			
$pout$	Variant 1	Variant 2	EV method
0.02	0.048 (0.028)	0.047 (0.029)	0.145 (0.090)
0.04	0.070 (0.047)	0.288 (0.035)	0.155 (0.061)
0.06	0.191 (0.047)	0.567 (0.041)	0.271 (0.014)
0.08	0.262 (0)	0.643 (0.030)	0.298 (0.021)
0.1	0.262 (0)	0.704 (0.031)	0.336 (0.018)
$pin = 0.6$			
$pout$	Variant 1	Variant 2	EV method
0.06	0 (0)	0 (0)	0.010 (0.015)
0.12	0 (0)	0.14 (0.099)	0.055 (0.042)
0.18	0 (0)	0.63 (0.016)	0.093 (0.070)
0.24	0 (0)	0.709 (0.016)	0.207 (0.047)
0.3	0.026 (0.023)	0.738 (0.001)	0.188 (0.029)

Table 1: Average VI distances (normalized) between actual and derived communities. Each value is the average over 5 trials. Standard deviations are bracketed.

As expected, for all algorithms, detection accuracy declines as the strength of community structure decreases. Variant 2 proves to be the least robust of all and the VI distances between its detected communities and the actual communities increase rapidly with $pout$. This simply means that detection accuracy is not robust. However when there are strong community structures, e.g. when $pout = 0.1 \times pin$, corresponding to the left most points on the graphs, it performs identically to variant 1 in detecting the exact communities. Variant 1 is the most accurate in returning the actual communities and slightly outperforms the leading eigenvector method. Results from the latter can be improved with fine tuning node memberships of the individual communities although this will be extremely expensive for large networks. For $pin = 0.6$ where communities are strongly connected, community detection is perfect in most cases for variant 1, as reflected by the zero VI distances.

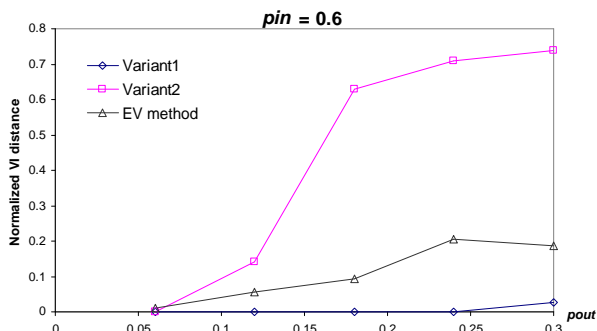
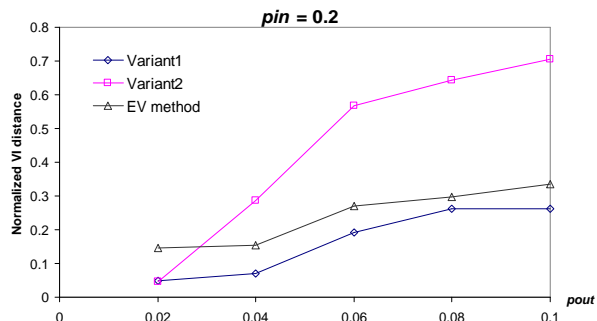


Figure 1: Average VI distances (normalized) from table 1.

This simple experiment is sufficient to demonstrate that $Q2$ is inadequate as a maximization criterion, although it may perform satisfactorily in some cases. This is also evident when we apply both variants on real world networks. We have tested them on several real world networks, namely the “karate club” network from Zachary (1977), the dolphin network from Lusseau (2003) and a network of books about politics (Krebs). It was observed that both variants can result in very similar detected sets of communities. The differences between them can once again be quantified with VI distance.

Networks	Hierarchy levels		
	1	2	3
Karate	0.095	0.070	0
Dolphin	0.105	0.058	0
Book	0.211	0.020	0

Table 2: VI distances (normalized) between community sets of both variants for each hierarchy level and for each network.

The VI distances between the detected communities from both variants are illustrated in table 2 for the mentioned real world networks. For each network tested here, both variants return hierarchies with three levels. We denote the lowest hierarchy level as level 1 and the highest level as level 3. The latter corresponds to the case where all nodes are considered as one community. Hence the hierarchies have a one to one correspondence in levels and can be compared in a straightforward manner. At level 1, detected communities appear significantly different for Krebs’ book network, although at level 2, both variants again return very similar results. For the

other networks at various levels, detected communities differ only slightly. This simply means that although Q_2 is not an appropriate measure, it may not be very obvious empirically.

For a rough idea of what the figures entail, figure 2 illustrates the karate network and resulting partitions from both variants. It can be seen that partitions of both variants differ only slightly at each hierarchy level. At level 1, although variant 1 indicates seven communities and variant 2 indicates six, the general community structures are quite similar. Two of the communities from variant 1 are essentially “merged” as one in variant 2. At level 2, both variants return three communities with the only difference arising in the membership of one node (numbered 10).

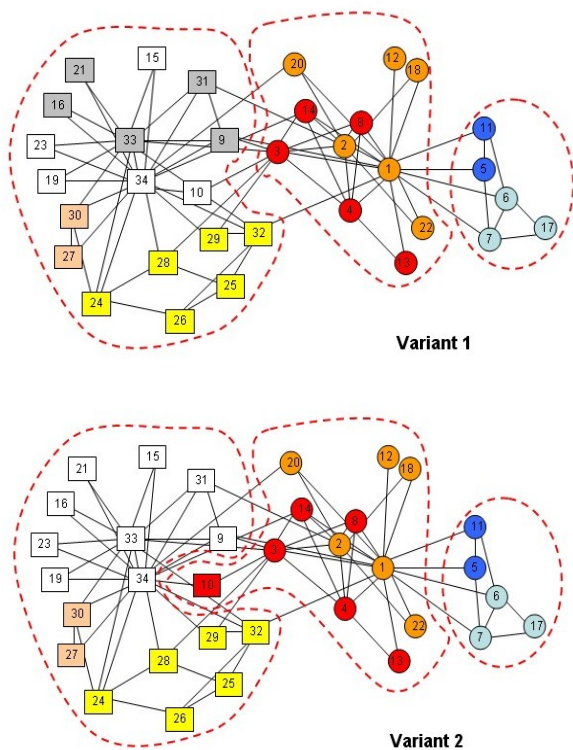


Figure 2: The karate network and partitions returned by the 2 variants. Numbered vertices represent club members while edges represent friendships. The squares and circles represent the actual two factions of the club. Level 1 partitions are colour coded while level 2 partitions are circled.

7 Conclusion

In this paper, we have identified the inconsistencies of a modularity definition that was used frequently to characterize community structures. Although it is fairly straightforward to prove this mathematically, the issue surprisingly has not been highlighted. We also demonstrate this empirically by testing variants of the same algorithm specially customized for each modularity definition. The take away message is that an imperfect measure may still give rise to sensible results on certain

data sets and caution is necessary in the process of defining a measure.

Incidentally, the variant 1 algorithm has performed very accurately in the experiment here and it may be worthwhile to investigate it in experiments with more extensive scenarios, such as using artificial communities of different sizes or incorporating hierarchies. More community detection algorithms in addition to the leading eigenvector algorithm can be compared with it.

While we have shown Q_1 to be the better modularity definition, it has its limitation in the form of the recently discovered “resolution limit” issue, which was highlighted by Fortunato and Barthélemy (2007). Essentially, algorithms which maximise Q_1 may have less power to detect smaller communities in large networks. Under certain conditions, it can be shown that groupings which aggregate the smaller communities into larger one(s) can result in larger modularity values than if they were detected individually. It remains for a measure to be discovered that can satisfactorily quantify the strength of community structures without the resolution limit effects. However the resolution limit issue can be easily circumvented by clever algorithm design. For example, the algorithm utilized here provides the user with a hierarchy of community structures at different resolution levels, as was also pointed out by Blondel et al. (2008). This is perhaps a more preferable approach given that many real world networks contain hierarchical community structures.

8 References

- Blondel, V., Guillaume, J-L., Lambiotte, R. and Lefebvre, E. (2008): Fast unfolding of community hierarchies in large networks. arXiv:0803.0476.
- Clauset, A., Newman, M. and Moore, C. (2004): Finding community structure in very large networks. *Phys. Review E* **70**:066111.
- Duch, J. and Arenas, A. (2005): Community detection in complex networks using Extremal Optimization. *Phys. Review E* **72**:027104.
- Fortunato, S. and Barthélemy, M. (2007): Resolution limit in community detection, *Proc. Natl. Acad. Sci. USA* **104**(1):36-41.
- Karrer, B., Levina, E. and Newman, M. (2008): Robustness of community structure in networks. *Phys. Review E* **77**:046119.
- Kleinberg, J. and Lawrence S. (2001): The structure of the web. *Science* **294**(5548):1849-1850.
- Krebs, V.: <http://www.orgnet.com>.
- Lusseau, D. (2003): The emergent properties of a dolphin social network. *Proc. R. Soc. London B (suppl.)* **270**, S186-S188.
- Meila, M. (2007): Comparing clusterings – an information based distance. *Journal of Multivariate Analysis* **98**, 873-895.
- Newman, M. (2004): Fast algorithm for detecting community structure in networks. *Phys. Review E* **69**:066133.

- Newman, M. (2006a): Modularity and community structure in networks. *Proc. Natl. Acad. Sci.*, vol. **4**:8577-8582.
- Newman, M. (2006b): Finding community structure in networks using the eigenvectors of matrices. *Phys. Review E* **74**:036104.
- Newman, M. and Girvan, M. (2004): Finding and evaluating community structure in networks. *Phys. Review E* **69**:026113.
- Ravasz, E., Somera A., Mongru, D. and Barabási A.-L. (2002): Hierarchical Organization of Modularity in Metabolic Networks. *Science* **297**(5586):1551-1555.
- Ruan, J. and Zhang, W. (2006): Identification and evaluation of weak community structures in networks. *Proc. of National Conference on Artificial Intelligence*, Boston, MA, July 2006, AAAI-06.
- Zachary, W. (1977): An information flow model for conflict and fission in small groups. *Journal of Anthropological Research* **33**, 452-473.