Resolution limit revisited: community detection using generalized modularity density.

Jiahao Guo[1,2], Pramesh Singh[1,2], Kevin E. Bassler[1,2,3]

[1] Department of Physics, University of Houston, Houston, Texas 77204, USA.
[2] Texas Center for Superconductivity, University of Houston, Houston 77204, Texas, USA.
[3] Department of Mathematics, University of Houston, Houston, Texas 77204, USA.

E-mail: bassler@uh.edu

Abstract. Various attempts have been made in recent years to solve the Resolution Limit (RL) problem in community detection by considering variants of the modularity metric in the detection algorithms. These metrics purportedly largely mitigate the RL problem and are preferable to modularity in many realistic scenarios. However, they are not generally suitable for analyzing weighted networks or for detecting hierarchical community structure. Resolution limit problems can be complicated, though, and in particular it can be unclear when it should be considered as problem. In this paper, we introduce a metric that we call generalized modularity density Q_g that eliminates the RL problem at any desired resolution and is easily extendable to study weighted and hierarchical networks. We also propose a benchmark test to quantify the resolution limit problem, examine various modularity-like metrics to show that the new metric Q_g performs best, and show that Q_g can identify modular structure in real-world and artificial networks that is otherwise hidden.

1. Introduction

Networks are excellent tools for describing complex biological, social, and infrastructural systems [1–4]. Most real-world examples of complex networks are far from being random and have a community or modular structure within them [5–7]. Detecting this structure is crucial in understanding the function and dynamics of a complex network. Although there is no universally accepted definition of a community structure [8, 9], it is often characterized by dense connectivity within groups and sparser connectivity between different groups. Modularity, Q, is a widely used metric to quantify the presence of this type of structure [5, 10–15]. For a partition of the nodes of an unweighted network, $C = \{c_1, c_2, c_3, ..\}$, it is defined as

$$Q = \frac{1}{2m} \sum_{c \in C} \left(2m_c - \frac{K_c^2}{2m} \right) \tag{1}$$

where m_c is the number of links in community c, K_c is the sum of degrees of nodes in c, and m is the total number of links in the network. Q measures the difference between the fraction of links within communities and the expected fraction if the links were randomly placed. The partition that maximizes the metric Q identifies the community structure of the network. Despite its intuitively appealing definition, there is a fundamental problem with using Q to find community structure. Namely, communities smaller than a certain size in large network may not be detected. This *Resolution Limit* (RL) problem [16, 17] reduces the domain of applicability of Q and is often a significant issue when analyzing empirical networks.

Alternate metrics have been proposed in recent years [18–26] to mitigate the RL problem. Some of these metrics [22–26], known as modularity density metrics, weights that are functions of the internal link density of communities are applied to the two terms in Eq. 1. In this paper we propose a new metric of this form, which we call *generalized modularity density* Q_g. Q_g is an extension of Q, as it reduces to Q in a limit. The main reasons for introducing this new metric are as follows. First, it has an adjustable parameter χ that controls the resolution density of the communities that are detected. Second, Q_g can be extended to detect communities in weighted networks in a way that has a clear interpretation and is independent of the scale of the link weights.

The RL problem can be seen in the simple example of cliques arranged in a ring connected to one another in series by single links [16]. The expectation in this case is that the cliques should be detected as separate communities. Unfortunately, with some metrics, pairs of cliques are merged into the same community. Of course, if all possible cross-links between two cliques are present, then it is sensible to merge them into one community as they simply form a clique of larger size. However, when cliques are connected by an intermediate number of links or when the network is weighted, it is unclear whether the cliques should be merged or separated [27]. Intuitively, it makes sense to merge two cliques at sufficiently high density of cross-links. Generally, methods of community detection that use different metrics have a different critical value for

this density. The answer may also depend on the specific application being considered. Thus, it is useful to have some flexibility in allowing the communities to be separated or merged. Q_g achieves this goal by varying a parameter χ. We will show that for a properly chosen value of χ, the partition that maximizes Q_g separates two cliques at any desired strength of inter-connectivity. This tunability of our metric is extremely useful for analyzing networks that exhibit hierarchical community structure [5], which is found in many real-world networks. A common way to investigate these hierarchical structures is to iteratively perform community detection within detected communities [28]. Using our approach, one can simply vary χ.

Finally, we compare the performance of our metric against other modularity density metrics by using them to find the structure in a more complex benchmark network than a simple ring of cliques. Our analysis indicates that Q_g performs better than all other metrics considered. We then use Q_g to find structure in a variety of empirical and artificial networks to demonstrate its ability to detect hidden community structure. We find that it eliminates the resolution limit problem that we consider and that it is applicable to a wider range of problems than other metrics. In addition, the network partition that maximizes Q_g can be efficiently and accurately found using the recently introduced Reduced Network Extremal Ensemble Learning (RenEEL) scheme [15].

2. Methods

2.1. Generalized Modularity Density

We define the *Generalized Modularity Density* of a node partition of unweighted network as

$$Q_g = \frac{1}{2m}\sum_c (2m_c - \frac{K_c^2}{2m})\rho_c^\chi \tag{2}$$

where m is the number of total links of the network, m_c is the number of links within a community c, K_c is the sum of degrees of all nodes in community c, ρ_c is the link density of community c, the exponent χ is a control parameter. Here we assume that χ is a non-negative real number. The link density of a community is the ratio of the number of links that exist in c to the number of possible links that can exist

$$\rho_c = \frac{2m_c}{n_c(n_c-1)}\ , \tag{3}$$

where n_c is the number of nodes in c. Q_g is an extension of modularity, i.e. at $\chi = 0$, $Q_g = Q$.

The metric Q_g, like the Modularity metric Q (Eq. 1), can be easily extended to weighted networks. For Q this is done by simply replacing the number of links with the sum of link weights in m, m_c and K_c [29, 30]. Extending the definition of modularity density metrics to weighted networks is complicated by the fact that they depend on link density, and link density can be problematic to use with weighted networks. One way to deal with these problems is to simply ignore the link weights and calculate

the link density as if the network was unweighted [22, 23]. Unfortunately, this loses the information contained in the link weights. The correct way is to use a normalized definition of link density, where the sum of the weight of all internal links divided by the maximum value that sum would have if the community were fully connected with links of weight equal to the maximum weight of any link in the network,

$$\rho_c = \frac{2m_c}{n_c(n_c-1)w_{\max}} \tag{4}$$

where m_c is the sum of the weights within community c, n_c is the number of nodes in c, and $w_{\max}$ is the maximum weight of any link in the network. This definition of ρ_c is consistent with the definition for unweighted networks, but it can be problematic because it involves the global variable $w_{\max}$. The community structure found using some metrics, such as those proposed in Refs. 22, 25, 26, can be very sensitive to the value of $w_{\max}$. This makes their use potentially troublesome, especially in empirical studies where the value of $w_{\max}$ can be difficult to accurately measure. Additionally, if there is a wide distribution of link weights and $w_{\max} \to \infty$, then $\rho_c \to 0$ for all communities and the algorithms for finding the partition that maximizes the modularity density metric become numerically unstable.

Generalized Modularity Density, unlike other modularity density metrics, does not have problems with $w_{\max}$. Both terms in Eq. 2 are weighted by the same function of $w_{\max}$, which can factored out and simply modifies the value of Q_g for every possible partition by the same constant factor. It is, thus, irrelevant for determining the partition that maximizes Q_g. So, instead of the absolute link density, Eq. 4, a relative link density, given by Eq. 3 with m_c being the sum of the weight of links in c, can be used in the metric Q_g without affecting results. The community partitions found with Generalized Modularity Density are also independent of the scale of the link weights. As it is with Modularity, multiplying all link weights by a common factor does not affect the results obtained with Q_g. This important property is needed for preserving the information in the link weights.

2.2. Resolution Density

The RL problem can be viewed as a problem with a metric, when using it yields a partition that merges two "well separated" communities. A resolution-limit-free metric is expected to resolve these communities. Conversely, a metric should also avoid splitting two groups of nodes that are "well connected" to each other. The RL problem is clear at these two extremes. However, more generally, the notion of well separated/connected communities is not well defined. It is unclear whether two partially connected communities should be merged or not.

Consider the benchmark network shown in Fig. 1. This network consists of three parts: two cliques and an external arbitrary component to which the cliques are weakly connected. As the cliques are fully connected, they have no internal community structure. Assume clique 1 has n_1 nodes, clique 2 has n_2 nodes, and both n_1 and $n_2 \geq 3$.

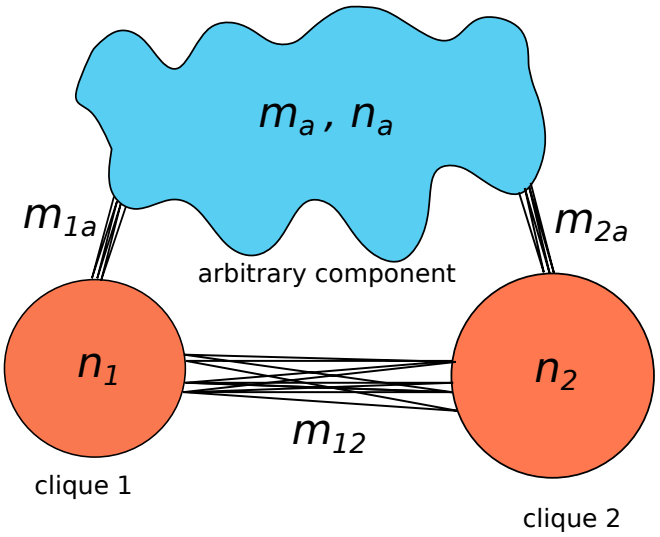

Figure 1: **Benchmark network for studying the resolution limit problem.** The network consists of two cliques of sizes n_1 and n_2 and an arbitrary component with n_a nodes and m_a links. The two cliques share m_{1a} and m_{2a} links with the arbitrary component, respectively, and have m_{12} links between. The links of the network can be weighted, in which case, m_a, m_{1a}, m_{2a} and m_{12} are the sums of link weights.

Without loss of generality, we assume $n_2 \geq n_1$. Let m_{12} be the sum of weights of links between the two cliques, and let m_{1a} and m_{2a} be the sum of the weights of links that connect each clique with the arbitrary component. n_a and m_a are number of nodes and the sum of weights of links within the arbitrary component, respectively. Without loss of generality, assume $n_1 \leq n_2$. Also, assume that $m_{1a} \ll n_1^2 w_{\max}$ and $m_{2a} \ll n_2^2 w_{\max}$, so that the cliques are only weakly connected to the arbitrary component. The RL question concerning this network is whether or not the two cliques should be merged or split, and whether or not using a given metric will meet this expectation. This choice of network gives greater flexibility to explore the RL problem than a simple ring of cliques, since the external component can have an arbitrary structure and the strength of inter-connectivity between the two cliques can be varied. Generally, there is a threshold, or critical, value of m_{12} below which the cliques are separated and above which they are merged. We impose an arbitrary *expected critical value* $m_{\exp}$ such that the cliques should be merged if $m_{12} \geq m_{\exp}$ and separated if $m_{12} < m_{\exp}$. Instead of using the values of m_{12} and $m_{\exp}$, it is convenient to use normalized inter-clique link density

$$d = \frac{m_{12}}{n_1 n_2 w_{\max}} \tag{5}$$

and normalized expected critical resolution link density

$$\delta_{\exp} = \frac{m_{\exp}}{n_1 n_2 w_{\max}} \, . \tag{6}$$

For unweighted networks $w_{\max} = 1$.

Given a metric, we can examine the RL question in the benchmark network. For a given set of network parameters, the two cliques are either merged or split. Accordingly, the parameter space can be divided into Merged (M) and Split (S) phases. The value of the link density at the boundary of the two phases is δ. At the same time, there is an expected result, corresponding a specific understanding of the problem, given by $\delta_{\rm exp}$. The metric can then be evaluated by comparing the results obtained by using it with the expected results. Specifically, we define a resolution-limit-free metric as one for which $\delta \geq \delta_{\rm exp}$ for all parameters of the benchmark network. Then, the metric is resolution-limit-free with respect to the expected resolution density.

3. Results

3.1. Benchmark Test

We now analytically study the extent to which the RL exists in benchmark network of Fig. 1 when Q_g is used as the metric. We also compare the results to that obtained when using other metrics. Whether the use of the metric Q_g will split the cliques or not is determined by the sign of $\Delta Q_g = Q_g^{merge} - Q_g^{split}$, where Q_g^{merged} and Q_g^{split} are the values of Q_g if the cliques are merged or split, respectively. Let us define the variables

$$p = \frac{n_1}{n_2} \tag{7}$$

and

$$t = \frac{m_a}{n_1 n_2 w_{\max}} \ . \tag{8}$$

$p \in (0, 1]$ is the ratio size of the cliques. $t \in [0, \infty)$ measures the external influence on them. Then,

$$\Delta Q_g \sim \Big(\frac{r+2d}{r+2}\Big)^{\chi}\Big(2d + r - \frac{(r+2d)^2}{r+2d+2t}\Big) - \Big(r - \frac{r^2 - 2 + 2d^2 + 2dr}{r+2d+2t}\Big) \tag{9}$$

where $r = p + 1/p$. If $\Delta Q_g < 0$, splitting is preferred, and if $\Delta Q_g > 0$, merging is preferred. Eq. 9 determines whether the use of the metric Q_g, for a given value of χ, will lead to M or S phase as a function of the variables (p, d, t). The value of d at which the phase boundary separating the M and S phases occurs is δ_{Q_g}.

In the limit of large external influence parameter t, which is often the situation encountered in empirical studies where RL problems are considered problematic, the value of δ_{Q_g} for a given value of χ is

$$\lim_{t\to\infty} \delta_{Q_g} = \frac{r}{2}\left[\left(1 + \frac{2}{r}\right)^{\frac{\chi}{\chi+1}} - 1\right] \ . \tag{10}$$

This limit increases from $\delta_{Q_g} = 0$, when $\chi = 0$, to $\delta_{Q_g} = 1$, when $\chi \to \infty$, for all values of p. At intermediate values of χ the result is only weakly dependent on p, being just slightly larger at small p, as can be seen in Fig. 2. The figure shows shows the phase

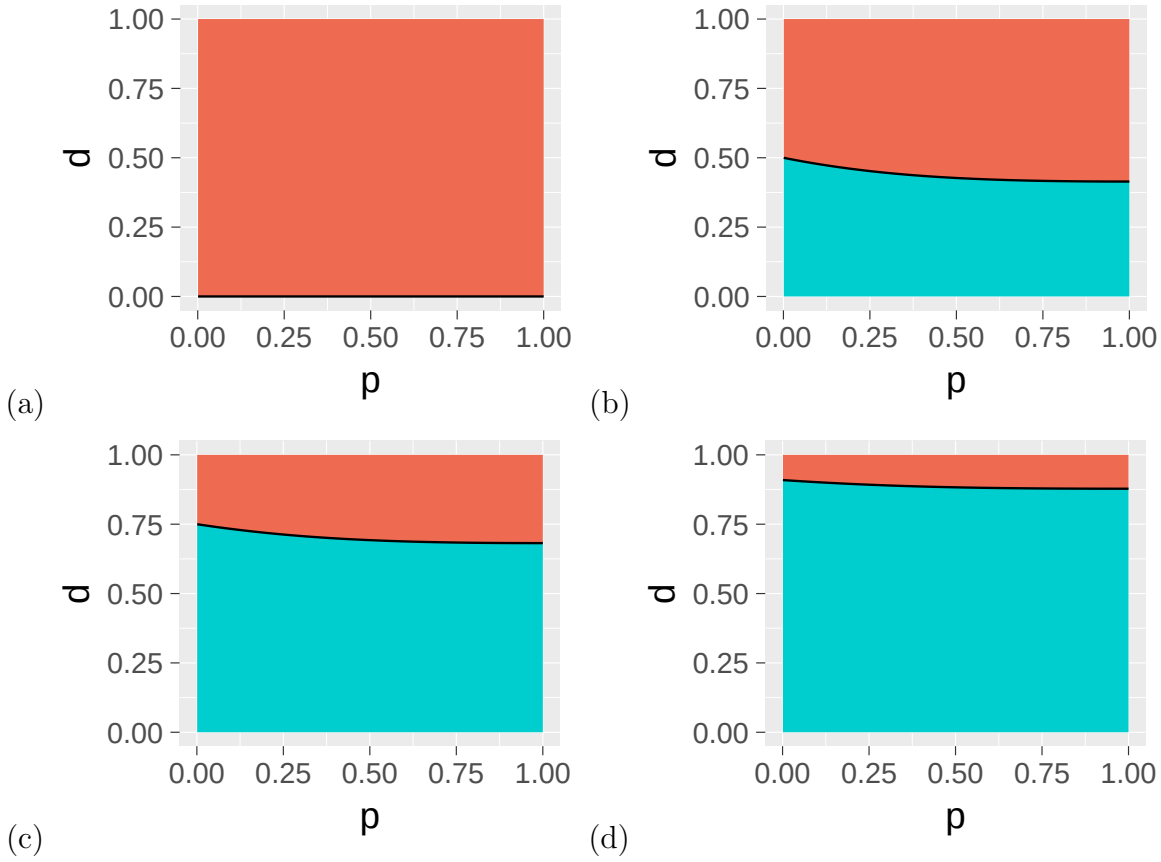

Figure 2: **Phase diagram of clique splitting with generalized modularity density at large external influence as χ is varied**. The values of clique size ratio p and link density d where the M phase occurs is shown in orange and where the S phase occurs is shown in blue. Results are for different values of the control parameter χ: (a) $\chi = 0$, (b) $\chi = 1$, (c) $\chi = 3$, (d) $\chi = 10$. The external influence parameter is $t = 10^6$

diagram as a function of p and d at various values χ for large t. For $\chi = 0$, when $Q_g = Q$, the cliques are merged at all values of p and d as shown in Fig. 2(a). For $\chi > 0$ at smaller values of d the cliques separate and are, thus, resolved. As χ increases, δ_{Q_g} also increases and approaches 1 in the limit of large χ, Figs. 2(b)-(d), meaning that at large χ the cliques are always resolved.

The effect of varying t at fixed χ on the (p, d) phase diagram are shown in Fig. 3. As shown in Fig. 3(a), at $t = 0$ when there is no influence by the external component on the two cliques, the S phase occupies the entire space and $\delta_{Q_g} = 1$ for all p. In this case, the cliques are always separated unless they are fully connected to each other. For $t > 0$, when there is some influence from an external component, the cliques are merged and, thus, not resolved for large values of d. As t increases, shown in Figs. 3(b)-(d), the M phase occupies an increasing area and δ_{Q_g} decreases until reaching the limiting value given by Eq. 10.

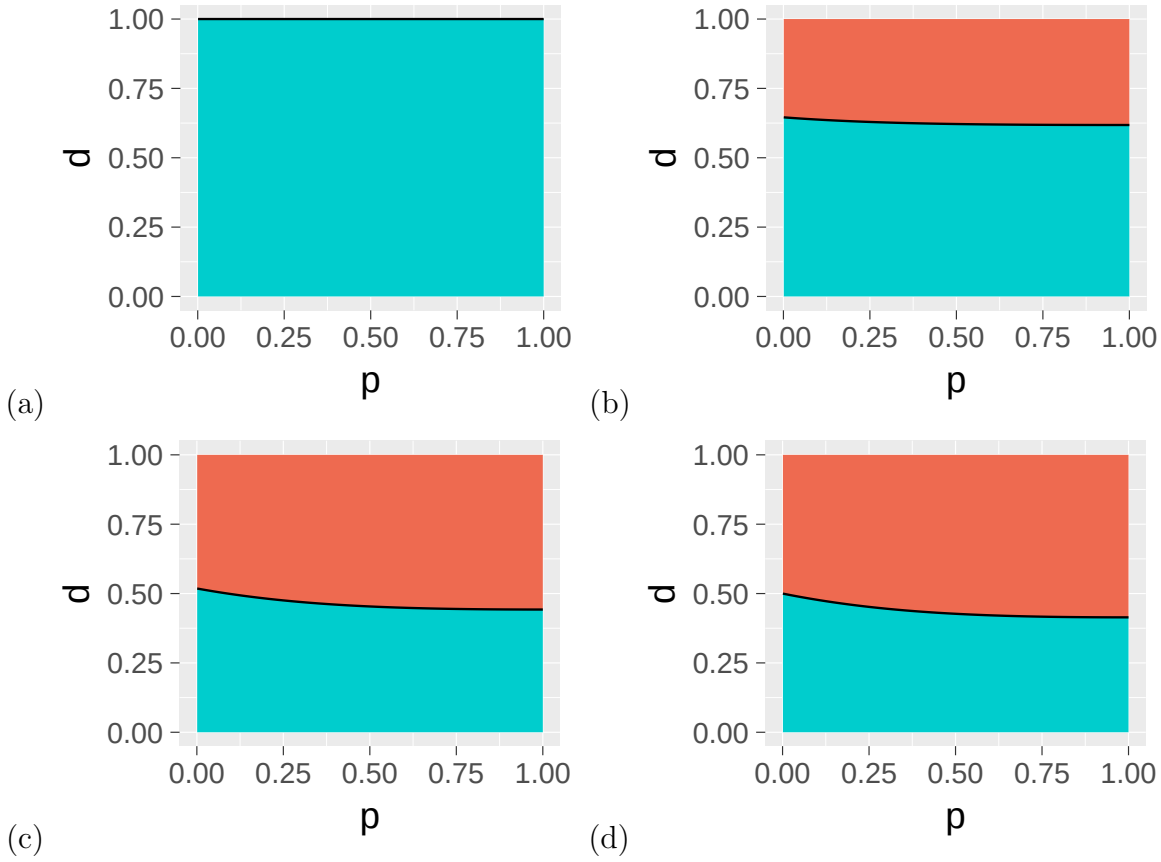

Figure 3: **Phase diagram of clique splitting with generalized modularity density at fixed χ as the external influence is varied**. The values of clique size ratio p and link density d where the M phase occurs is shown in orange and where the S phase occurs is shown in blue. Results are for $\chi = 1$ and different choices of the external influence parameter t: (a) $t = 0$, (b) $t = 1$, (c) $t = 10$, (d) $t = 10^6$.

These results show that, as the control exponent χ is varied, a wide range of δ_{Q_g} results. The range increases with t and varies from 0 to 1, the complete possible range, in the limit of large t. This freedom gives leeway in applications to choose χ so that δ_{Q_g} matches the expected critical resolution link density δ_{exp}.

In general, as χ increases the number of communities found also increases, but gives stable results for a range of χ. (See the example discussed in Sec. 3.2.2.) Increasing χ thus tends to result in smaller communities being detected. The appropriate, or best, choice of χ depends on the problem. If there is some "ground truth" knowledge about the community structure in the network, or in similar networks, that knowledge can be used to select a χ that results in communities that match the ground truth. If there is no ground truth knowledge, then a default choice of $\chi = 1$ may be appropriate. That choice results in a critical resolution density of $\delta_{Q_g} = 1/2$ in the limit of large t and r (Eq. 10). Thus, an advantage of Q_g is that even for the extreme values of t and r, the metric has a positive lower bound of δ_{Q_g} that can be controlled by χ.

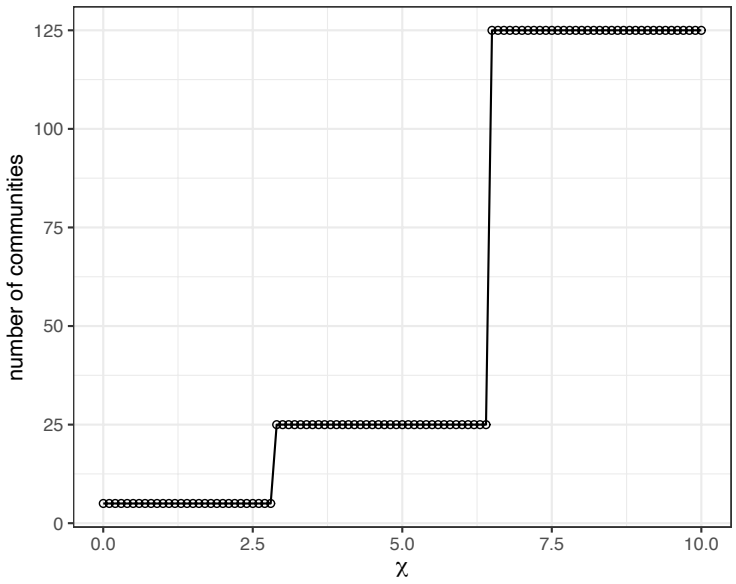

Figure 4: **Number of communities found using Q_g with different χ.** Communities in each level (from the largest to the smallest) in the hierarchy are revealed as χ is varied.

In contrast to Q [10], Q_{ds} [22], Q_x [25] and Q_{AFG} [19] (see Supplemental Information S1), Q_g has a finite non-zero lower limit of δ, which implies that for d smaller than this value, the two cliques of the benchmark network are guaranteed to be split for all possible values of (r, t). Thus, Q_g can successfully avoid resolution limit problem in these extreme cases (See last paragraph in Section 2.2). While the metric Q_w (see Supplemental Information S1) also shows this lower limit (Table 1), the advantage of Q_g is that the lower limit of δ_{Q_g} can be adjusted by tuning the parameter χ for any desired resolution density. Table 1 summarizes the kind of resolution problems with Q, Q_{ds}, Q_x, Q_w and Q_{AFG} that would be encountered when tested on the benchmark network (See Supplemental Information Section S3 for details).

In principle, a reasonable δ_{exp} is always in $[0, 1]$ but a given metric can still have a δ that is out of this range. Since δ_{exp} is strictly positive (no matter how small), if it is possible to construct a network for which $\delta \to 0$ then that metric presents a resolution limit problem. Even worse, if $\delta < 0$, it would result in merging of disconnected communities. On the other hand $\delta \to 1$ does not pose a resolution problem as long as $\delta \leq 1$ and $\delta \geq \delta_{\text{exp}}$ is satisfied. However, higher δ_{exp} imposes a stricter criterion for merging. But if $\delta > 1$, it will have the unwanted consequence of cliques being subdivided. Thus, a metric is problematic if it can not avoid $\delta \to 0$, $\delta < 0$ or $\delta > 1$.

3.2. Applications

3.2.1. American college football network We use the Q_g metric to detect communities in the network of American college football games between Division IA colleges during

Metric	Resolution limit problem
Q	$\delta \to 0$ when $t \to \infty$
Q_{ds}	$\delta < 0$ when p is small
Q_x	$\delta < 0$ when p and ρ are small
Q_w	$\delta_{min} = 0.236$ when $t \to \infty$ and $p = 1$
Q_{AFG}	$\delta < 0$ when $s < 0$ and p is small $\delta > 1$ when $s > 0$ and p is small

Table 1: **Resolution limit problems of different metrics.** Q, Q_{ds}, Q_x, Q_w and Q_{AFG} have different resolution limits problems. ρ, which appears in Q_x, is the global link density. s is used in the metric Q_{AFG} as a weight to every node (equivalent to adding a self-loop to every node) and thereby modifying the strength of a community. Q_{AFG} reduces to modularity at $s = 0$, and by controlling s substructures ($s > 0$) or superstructures ($s < 0$) can be explored.

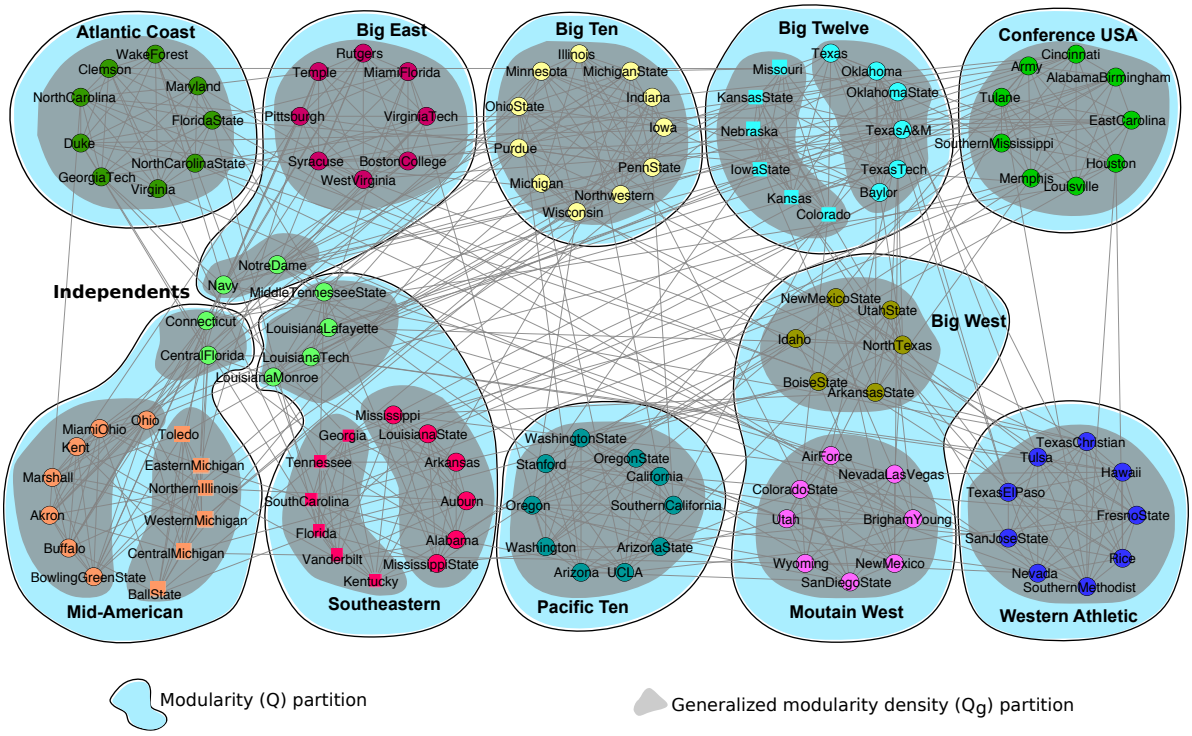

Figure 5: **Communities found in American college football network.** Blue blobs show the communities detected by modularity and gray blobs show the communities found by generalized modularity density.

regular season of Fall 2000 [10, 31]. A link between two colleges is present if they played a game against each other. Colleges play games within the same conference more frequently, thus, a community detection algorithm should be able to recover these conferences from the network data. First, we show the result of using modularity (Q) that are indicated by light blue blobs in Fig 5. It matches the conference memberships (distinguished by node color) well except Independents, which are absorbed by three communities and that it groups Big West and Mountain West in the same community. Using $Q_g(\chi = 3)$ in this network we find communities that are shown by gray blobs. There are some key differences between the Q and Q_g partitions. First, the Q_g partition does not merge the Independents with other conferences. Instead, it divides them into three disjoint communities. Second, it successfully identifies the Big West and Mountain West as two different groups. But more interestingly, unlike modularity, it divides each of the Mid-American, Southeastern, and Big Twelve conferences into two communities. This apparent deviation from ground truth actually turns out to be a major advantage of using Q_g. Each of these three conferences have subdivisions within them that are in perfect agreement (considering their membership as of year 2000) with the partition found by Q_g. Mid-American conference has East Division and West Division, Southeastern also has Eastern Division and Western Division, whereas Big-Twelve conference has Northern Division and Southern Division. These subdivisions are indicated by different node shapes (circles and squares) in Fig. 5.

3.2.2. Artificial network with hierarchical community structure To demonstrate the ability of Q_g for detecting the community structure at different resolution densities, we construct a hierarchical network. Similar constructions have been used as a model for hierarchical network structure [19]. We consider a structure shown in Fig. 6 that includes four levels of hierarchy, although it can be extended to include any number of levels. The elementary level (level 1) is a clique formed by fully connecting five nodes with links weighted α_1. To construct a level 2 network, we use the clique network from level 1 as a *generalized node* to form a clique of size 5 with links weighted α_2. Link between two generalized nodes is achieved by connecting all the internal nodes from one generalized node to those in another generalized node. Similarly, level k network is constructed by using level $k-1$ network as a generalized node to form a clique of size 5 with links weighted α_k. Here we keep $\alpha_1 > \alpha_2 > ... > \alpha_k$ so that the hierarchy of structure is preserved.

We use the metric Q_g on a level 4 network with $\alpha_k = 5 - k$ and show that it successfully detects the planted hierarchical communities at every level. The level 4 network consists of 125 level 1 cliques, and 625 nodes in total. The results obtained by maximizing Q_g is shown in Fig. 4. We observe that the 5 level 3 cliques are detected when $\chi < 2.8$, the 25 level 2 cliques are detected when $2.9 < \chi < 6.4$, the 125 level 1 cliques are detected when $\chi > 6.5$. There are 3 stages, corresponding to 3 levels of construction. There should not be a single "best" choice of χ by the nature of the problem. The choice of χ or desired resolution density should be based on specific requirement and the background

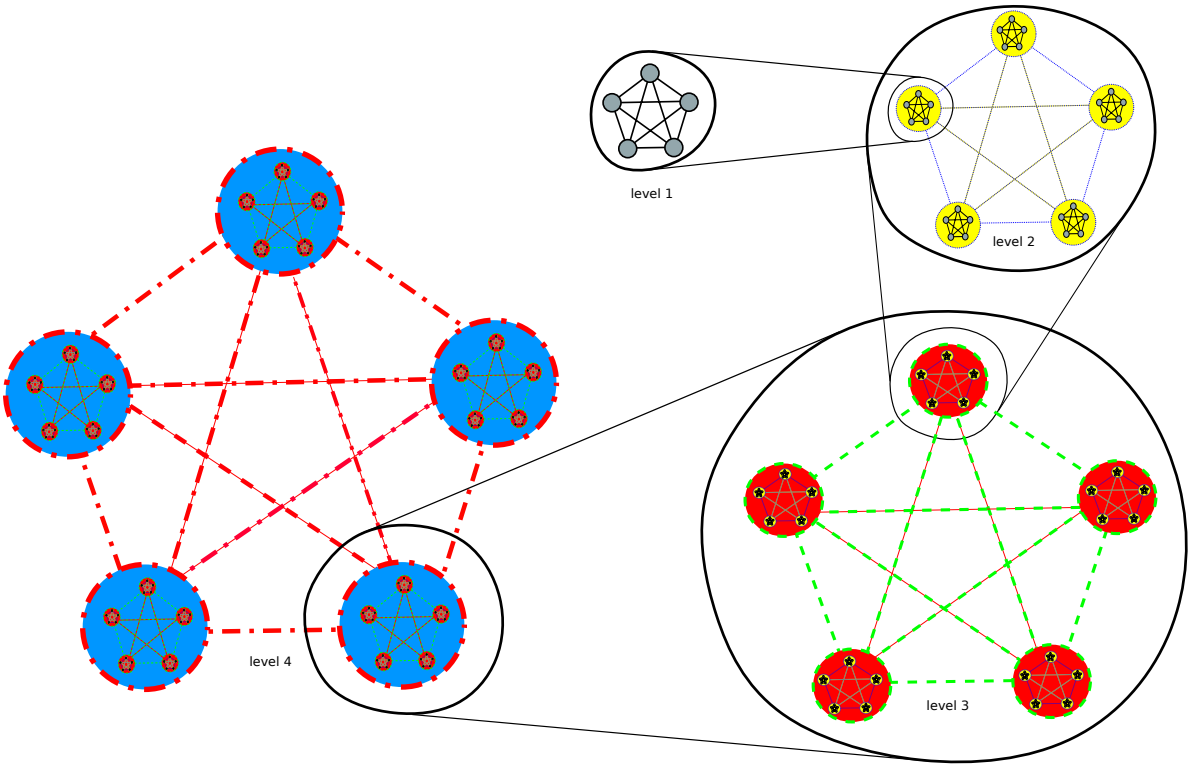

Figure 6: **Example hierarchical network.** Level 1: A clique of five nodes. Level 2: A clique of five level 1 cliques. Level 3: A clique of five level 2 cliques. Level 4: A clique of five level 3 cliques.

information of the particular problem.

4. Conclusion

Community detection in networks is commonly performed by finding the partition of the network nodes that maximizes an objective function. Such a partition can sometimes yield unexpected community structure. Resolution limit, for example, is an unwanted but inevitable consequence of modularity maximization. Other such metrics, namely modularity density measures, which attempt to fix this problem also differ in the community structure that they obtain and can also violate our general expectation. While at what number of cross links between two strongly connected groups of nodes should be called a single community remains mostly subjective and vague, our metric Q_g provides a quantifiable notion and solves the resolution limit problem. In particular, with a free parameter χ, one can control this threshold of merging two cliques. It is quite appealing to have a metric that can be adjusted to meet the specific requirement

set by the user because the idea of a community may vary from one application to another and may be specific to the network under consideration. At the same time, due to its ability to detect communities at many resolution densities it also useful in uncovering the hierarchical community structure, an inherent characteristic observed in many complex networks. The existing benchmarks, e.g. the ring of cliques, are too restrictive to evaluate and compare the performance of different metrics with respect to solving the specific resolution limit problem. In this paper we consider a more general yet simple network structure, which can be used to quantitatively examine the limits of metrics such as modularity. Using this general framework we demonstrated that our metric Q_g eliminates resolution limit problem at a desired resolution density, shows better performance, and is straightforward to extend for studying weighted and directed networks. Among other important problems, finding communities at high resolution is particularly useful in analyzing gene regulatory networks where the goal of functional annotation of genes is to find very specific gene functions [32].

Acknowledgments

This work was supported by the NSF through grants DMR-1507371 and IOS-1546858.

References

[1] Newman M E J 2010 *Networks, An Introduction*. Oxford University Press.
[2] Chauhan R, Ravi J, Datta P, Chen T, Schnappinger D, Bassler K E, Balazsi G and Gennaro M L 2016 Reconstruction and topological characterization of the sigma factor regulatory network of Mycobacterium tuberculosis *Nature communications*, **7** ncomms11062
[3] Treviño S III, Sun Y, Cooper T F and Bassler K E 2012 Robust detection of hierarchical communities from Escherichia coli gene expression data *PLoS computational biology*, **8**(2):e1002391
[4] Bhavnani S K, Bellala G, Victor S, Bassler K E and Visweswaran S 2012 complementary bipartite visual analytical representations in the analysis of SNPs: a case study in ancestral informative markers *J. Am. Med. Inform. Assoc.*, **19**(e1):e5-e12
[5] Newman M E J 2003 The Structure and Function of Complex Networks *SIAM Rev.*, **45**(2):167-256
[6] Danon L, Diaz-Guilera A, Duch J and Arenas A 2005 Comparing community structure identification *J. Stat. Mech.*, P09008
[7] Fortunato S 2010 Community detection in graphs *Physics reports*, **486**(3):75-174
[8] Schaub M T, Delvenne J, Rosvall M and Lambiotte R 2017 The many facets of community detection in complex networks *Applied Network Science*, **2**(1):4
[9] Peel L, Larremore D B and Clauset A 2017 The ground truth about metadata and community detection in networks *Science Advances*, **3**(5):e1602548
[10] Girvan M and Newman M E J 2002 Community structure in social and biological networks *Proc. Natl. Acad. Sci.*, **99**:8271-8276
[11] Newman M E J and Girvan M 2004 Finding and evaluating community structure in networks *Phys. Rev. E*, **69**:026113
[12] Newman M E J 2004 Detecting community structure in networks *Eur. Phys. J. B*, **38**(2):321-330
[13] Sun Y, Danila B, Josic K and Bassler K E 2009 Improved community structure detection using a modified fine-tuning strategy *Europhysics Letters*, **86**(2):28004

[14] Treviño S III, Nyberg A, del Genio C I and Bassler K E Fast and accurate determination of modularity and its effect size *J. Stat. Mech.*, P02003

[15] Guo J Singh P and Kevin E. Bassler K E 2019 Reduced network extremal ensemble learning (RenEEL) scheme for community detection in complex networks *Sci Rep*,**9**:14234

[16] Fortunato S and Barthelemy M 2007 Resolution limit in community detection *Proc. Natl. Acad. Sci.*, **104**(1):26-41

[17] V. A. Traag, P. Van Dooren and Y. Nesterov. Narrow scope for resolution-limit-free community detection *Phys. Rev. E*, **84**, 016114 (2011).

[18] Ronhovde P and Nussinov Z 2010 Local resolution-limit-free Potts model for community detection *Phys. Rev. E*, **81**, 046114

[19] Arenas A, Fernández A and Gomez S 2008 Analysis of the structure of complex networks at different resolution levels *New J. of Physics*, **10**,(5):053039

[20] Granell C, Gomez S and Arenas A 2012 Hierarchical multiresolution method to overcome the resolution limit in complex networks *International Journal of Bifurcation and Chaos*, **22**(07):1250171

[21] Aldecoa R and Marin I 2011 Deciphering Network Community Structure by Surprise *PloS one* **6**(9):e24195

[22] Chen M, Nguyen T and Szymanski B K 2013 A New Metric for Quality of Network Community Structure *ASE Hum. J.*, **2**(4):226-240

[23] Chen M, Kuzmin K and Szymanski B K 2014 Community Detection via Maximization of Modularity and Its Variants *IEEE Transactions on Computational Social Systems*, **1**(1):46-65

[24] Botta F and del Genio C I 2016 Finding network communities using modularity density *J. Stat. Mech.*, 123402

[25] Chen T, Singh P and Bassler K E 2018 Network community detection using modularity density measures *J. Stat. Mech.*, 053406

[26] Haq N F, Moradi M and Wang Z J 2019 Community structure detection from networks with weighted modularity *Pattern Recognition Letters*,**122**:14-22

[27] Lancichinetti A and Fortunato S 2011 Limits of modularity maximization in community detection 2011 *Phys. Rev. E.*, **84**(6):066122

[28] Park J, Wood I B, Jing E, Nematzadeh A, Ghosh S, Conover M D, Ahn Y Y 2019 Global labor flow network reveals the hierarchical organization and dynamics of geo-industrial clusters *Nature Communications*, **10**(1):3449

[29] Newman M E 2004 Analysis of weighted networks *Phys. Rev. E.*, **70**(5):056131

[30] Leicht, E. A. and Newman, M. E. J 2008 Community Structure in Directed Networks *Phys. Rev. Lett.*, **100**(11):118703

[31] Evans T S 2010 Clique graphs and overlapping communities *J. Stat. Mech.*, P12037

[32] Mentzen W I and Wurtele E S 2008 Regulon organization of Arabidopsis *BMC plant biology*, **8**(1):99

Supplemental Information

S1. Other metrics

Besides the metric Q_g, we test the performance of the following metrics. Each variable has the same meaning as Eq. 2 (in the main text) unless otherwise noted.
Modularity [10]:

$$Q = \frac{1}{2m}\sum_c \left(2m_c - \frac{K_c^2}{2m}\right) \tag{S1}$$

Weighted Modularity [26]:

$$Q_w = \frac{1}{2m}\sum_c \left(2m_c - \frac{K_c^2}{2m}\right)(\rho_c + 1) \tag{S2}$$

Excess Modularity Density [25]:

$$Q_x = \frac{1}{2m}\sum_c \left[2m_c(\rho_c - \rho) - \frac{K_c^2(\rho_c - \rho)^2}{2m}\right] \tag{S3}$$

Here $\rho = 2m/[n(n-1)]$ is the global link density.
Modularity Density introduced in Ref. [22] has a term that corresponds to Split Penalty. But as discussed in Ref. [25], this term may be problematic. Therefore, here we analyze a modified version of modularity density without the Split Penalty term:

$$Q_{ds} = \frac{1}{2m}\sum_c \left(2m_c\rho_c - \frac{K_c^2\rho_c^2}{2m}\right) \tag{S4}$$

AFG method of modularity Q_{AFG} in Ref. [19] can have different resolution densities by assigning self-loop weighted s to each node and tuning the value of s. It still finds the partition by maximizing modularity after assigning the self-loops.

$$Q_{AFG} = \frac{1}{2m+2Ns}\sum_c \left[(2m_c + 2n_c s) - \frac{(K_c + 2n_c s)^2}{2m+2Ns}\right] \tag{S5}$$

where N is total number of nodes, n_c is the number of nodes of community c.

S2. Derivation of equation of phase for modularity

We assess the performance of modularity Q using the benchmark test described in the main text. In the form shown in Eq. S1, modularity is the sum of the quantity within

parenthesis over each community. Thus, two partitions of splitting or merging the two cliques yield the following values of modularity Q

$$Q_{split} = Q_1 + Q_2 + Q_{ex} \tag{S6}$$

$$Q_{merge} = Q_{(1+2)} + Q_{ex} \tag{S7}$$

where Q_1, Q_2 are the two terms corresponding to clique 1 and 2 as separate communities, $Q_{(1+2)}$ is the corresponding term when the two cliques are merged. Q_{ex} is the sum over the remaining communities in the external component, which do not change in the two partitions. The difference between the two modularity values is given by

$$\Delta Q = Q_{merge} - Q_{split} = Q_{(1+2)} - Q_1 - Q_2. \tag{S8}$$

Using Eq. S1, we have:

$$\Delta Q = \frac{1}{2m}\left[(2m_{(1+2)} - \frac{K_{(1+2)}^2}{2m}) - (2m_1 - \frac{K_1^2}{2m} + 2m_2 - \frac{K_2^2}{2m})\right] \tag{S9}$$

According to the construction of the example network, we can rewrite Eq. S9 with $(n_1, n_2, m_{12}, m_a, n_a, m_{1a}, m_{2a})$. To simplify the expression and capture the principle features, we take $n_1, n_2 >> 1$. Recall the construction of separation of the two cliques from the external component, we also have $n_1^2 >> m_{1a}, n_2^2 >> m_{2a}$. Plugging all in ΔQ, we obtain:

$$\Delta Q = \frac{1}{2m}\left(2m_{12} - \frac{2(n_1^2 n_2^2 + (n_1^2 + n_2^2)m_{12} + m_{12}^2)}{n_1^2 + n_2^2 + 2m_a + 2m_{12}}\right) \tag{S10}$$

Eq. S10 can be rewritten more concisely by omitting the normalization factors and using variables (d, r, t) defined in Section 3.1

$$\Delta Q \sim 2d - \frac{2(1 + rd + d^2)}{r + 2d + 2t} \tag{S11}$$

The space is reduced to three principal dimensions (d, r, t), where $0 \leq d \leq 1$, $r \geq 2$ and $t \geq 0$. Eq. S11 is the equation of phase that is used to plot the phase diagram of Fig. S1. We obtain δ_Q, which determines the phase boundary as the value of d for which $\Delta Q = 0$,

$$\delta_Q = \sqrt{t^2 + 1} - t \tag{S12}$$

S3. Benchmark test

By carrying out similar mathematical analyses as in the last section, we can obtain an equation of phase for each metric and we can identify possible RL problems of each metric. Some advantages of using this benchmark test includes that it covers a wider range of cases, it can be used by working on the formula without any guess or speculation of specific network, and it provides a clear view of metric performance including all RL problems previously reported. The difference between values of a metric between merge

and split cases, for other metrics can be written as follows.
For weighted modularity Q_w

$$\Delta Q_w = \frac{2r+2d+2}{r+2}(2d+r-\frac{(r+2d)^2}{r+2d+2t}) - 2(r-\frac{r^2-2+2d^2+2dr}{r+2d+2t}). \tag{S13}$$

For excess modularity density Q_x

$$\begin{aligned}\Delta Q_x = (2d+r)(\frac{2d+r}{r+2}-\rho) - \frac{(r+2d)^2}{r+2d+2t}(\frac{r+2d}{r+2}-\rho)^2 \\ - \left(r(1-\rho) - \frac{r^2-2+2d^2+2dr}{r+2d+2t}(1-\rho)^2\right).\end{aligned} \tag{S14}$$

For modified (without the split penalty term) modularity density Q_{ds}

$$\Delta Q_{ds} = \frac{(2d+r)^2}{r+2} - \frac{(2d+r)^4}{(r+2d+2t)(r+2)^2} - (r-\frac{r^2-2+2d^2+2dr}{r+2d+2t}) \tag{S15}$$

For Q_x (Eq. S3), in addition to (d, r, t), the phase space consists of an extra principal variable ρ, which is the global link density and its maximum value ρ_{max} is obtained when n_a is smallest as other variables (n_1, n_2, m_{12}, m_a) are fixed.

The phase diagrams of Q, Q_{ds}, Q_w, $Q_x(\rho = \rho_{max})$, $Q_{AFG}(n_a = \sqrt{2m_a})$ and $Q_g(\chi = 1)$ are shown respectively in Fig. S1, Fig. S2, Fig. S3, Fig. S4, Fig S5 and Fig. 3 (in the main text). As shown in the figures, the behavior varies a lot across different metrics and the particular choice of other variables. In the following, we will observe some general characteristics of all phase diagrams. Then, we will examine each one in more detail and demonstrate that Q_g performs better than other metric.

First, there are two phases (M (red) and S (blue) phase) in the phase diagram and as expected and M phase is above S phase implying that nearly all metrics tend to merge the two cliques when d is close to 1 and to split when d is close to 0. This meets the common expectation in extreme cases. But different metrics disagree when d is in intermediate range. Other variables such as (t, ρ) also dictate the performance in this range. Fig. S1 shows the RL problem of modularity with a much clear view. We know that $\delta = \sqrt{t^2+1} - t$, as $t \to \infty$, we have $\delta \to 0$. This trend is also shown in the Fig. S1. Therefore, given any value $d > 0$, we can construct a network with large enough t so that $d > \delta$, which means the two cliques, as long as they are connected, they will be merged into one community if the external component has enough links. This is the RL problem of modularity. However, if $d = 0$, there is no RL because for any $t \geq 0$, $\delta > 0$ is always true and modularity maximization would not merge two disconnected cliques. More generally it can be shown that if two subgroups of the network are disconnected, they are guaranteed to be split.
As shown in Fig. S2, Q_{ds} depends strongly on p. A different type of RL problem can be seen in the figures. If p is small enough, $\delta = 0$ can always be true whatever d is. It means that if the sizes of two cliques are different enough, they will be merged even if

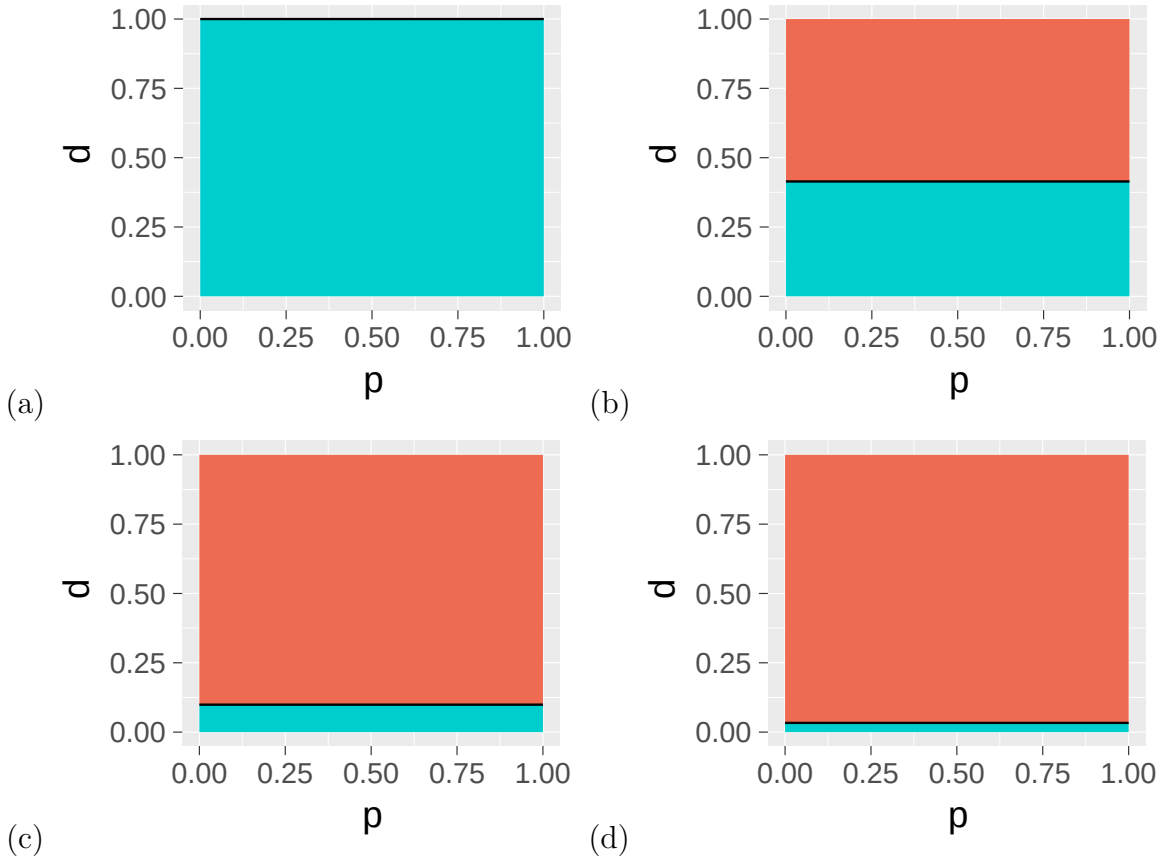

Figure S1: **Phase diagram of clique splitting with modularity Q as the external influence is varied.** The values of clique size ratio p and link density d where the M phase occurs is shown in orange and where the S phase occurs is shown in blue. Results are for different choices of the external influence parameter t: (a) t=0 (b) t=1 (c) t=5 (d) t=15.

$d = 0$ [25]. It clearly violates our expectation. This problem gets alleviated as $t \to \infty$. But it always exists for arbitrary t.

For phase diagram of Q_w shown in Fig. S3, the phase boundary moves down as $t \to \infty$. But it has a lower bound which means, when d is small enough, the two cliques of example network will always be split whatever other variables are. Thus, it has no extreme cases of RL as Q and Q_{ds}. Note that M phase is reduced to a straight line $d = 1$ here in Fig. S3(a) which means the extreme case of expectation is satisfied

As for Q_x, because there is one more variable ρ, the analysis is more complicated. As we can see from Equ. S3 and Equ. S4, $Q_x(\rho \to 0) \to Q_{ds}$ which means Q_x will behave the same as Q_{ds} when global link density $\rho = 0$. Because of the arbitrary external component, it can be easily achieved. Also we should be aware of the fact that most real-world networks are sparse thus $\rho \to 0$ is a common case where Q_x will fail to solve RL as Q_{ds}. In Fig. S4, we show the phase diagram when ρ equals to its maximum. The phase boundary, starting from $d = 1$, goes down first and then rises up again. So, when

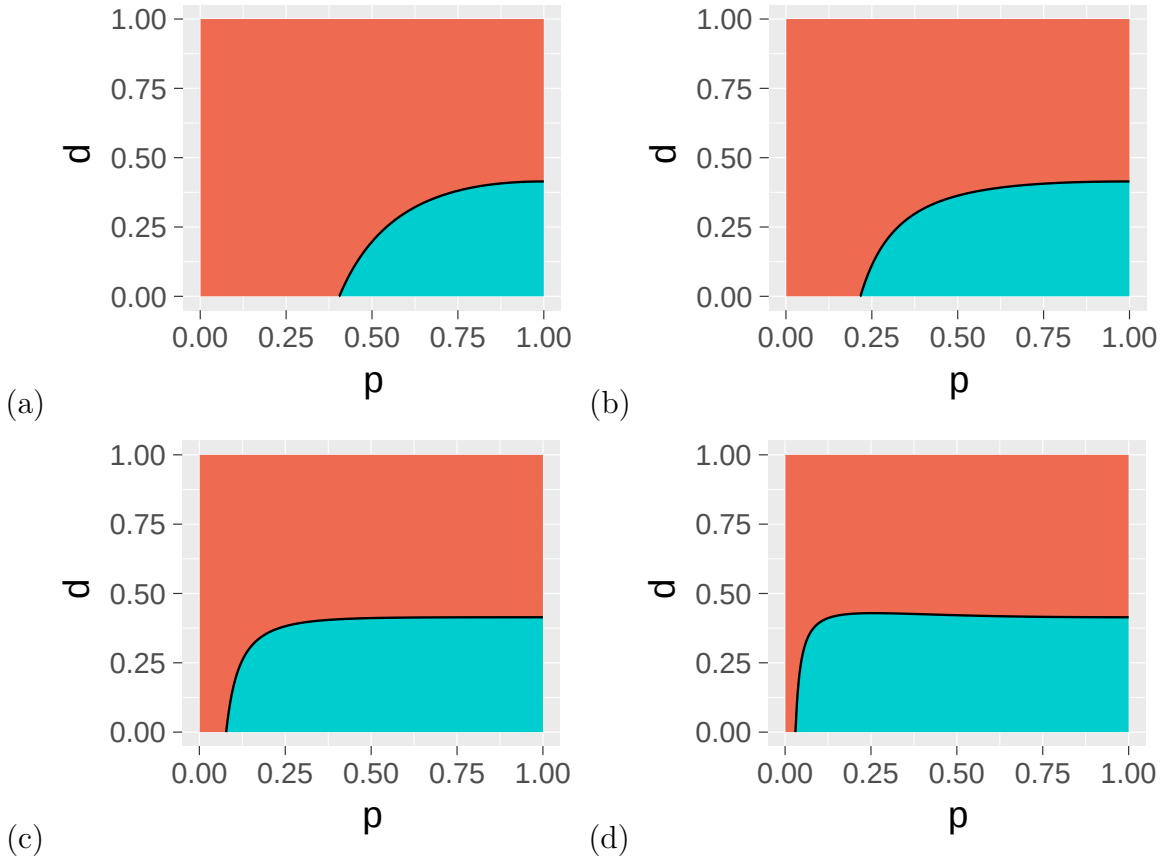

Figure S2: **Phase diagram of clique splitting with modularity density Q_{ds} as the external influence is varied.** The values of clique size ratio p and link density d where the M phase occurs is shown in orange and where the S phase occurs is shown in blue. Results are for different choices of the external influence parameter t: (a) t=0 (b) t=1 (c) t=5 (d) t=15.

$t \to \infty$, the two cliques will always be split as long as $d < 1$. But this requires both ρ, t are very large which is uncommon for most real-world networks.

We use the AFG method [19] on the benchmark network, which attempts to solve the RL problem by assigning a self loop of weight s to each node. This method allows one to explores communities at different resolution densities by controlling s. Using Q_g, this is achieved by controlling χ so the two methods are similar in spirit. However, irrespective of the choice of s, the metric Q_{AFG} will behave like modularity Q and fail to resolve clusters if n_a in the benchmark network is sufficiently large. To avoid that, we show the phase diagram (Fig. S5) for $n_a = \sqrt{2m_a}$, which is the smallest possible n_a for a fixed m_a (in the large m_a limit) and perhaps the best case scenario for Q_{AFG}. Moreover, if a specific resolution density is desired then s must be selected according to the network size, unlike Q_g, which has the lower bound that is independent of the network size. Even if s is chosen according to the network size, the phase diagram in Fig. S5 (a), (c), (d)

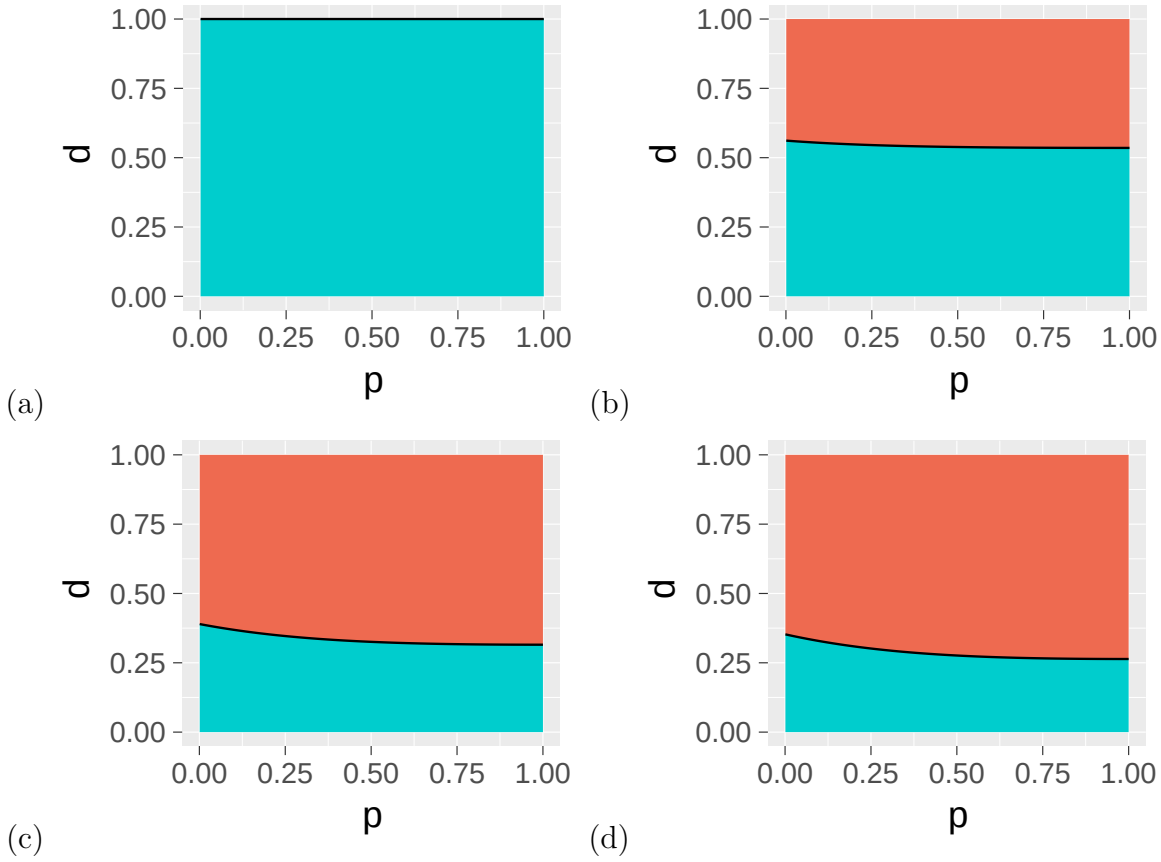

Figure S3: **Phase diagram of clique splitting with weighted modularity Q_w as the external influence is varied.** The values of clique size ratio p and link density d where the M phase occurs is shown in orange and where the S phase occurs is shown in blue. Results are for different choices of the external influence parameter t: (a) t=0 (b) t=1 (c) t=5 (d) t=15.

shows that the metric Q_{AFG} will fail when the two cliques are somewhat different in size (small p). When p is small and $s \neq 0$, it either merges two disconnected cliques (Fig. S5 (a)), or splits a larger clique formed by clique 1 and clique 2 (Fig. S5 (c) and (d)). When $s = 0$ (Fig. S5 (b)), the metric Q_{AFG} is the same as modularity Q and it will have the same problems as outlined before. The phase diagram also shows that for a non-zero value of s, the resolution density varies a lot as a function of p. This implies that merging or splitting the two cliques is heavily influenced by their relative sizes. Thus, in a network with a wide range of community sizes, this method will be biased either towards merging well separated communities or splitting well connected communities, an observation also made in [27].

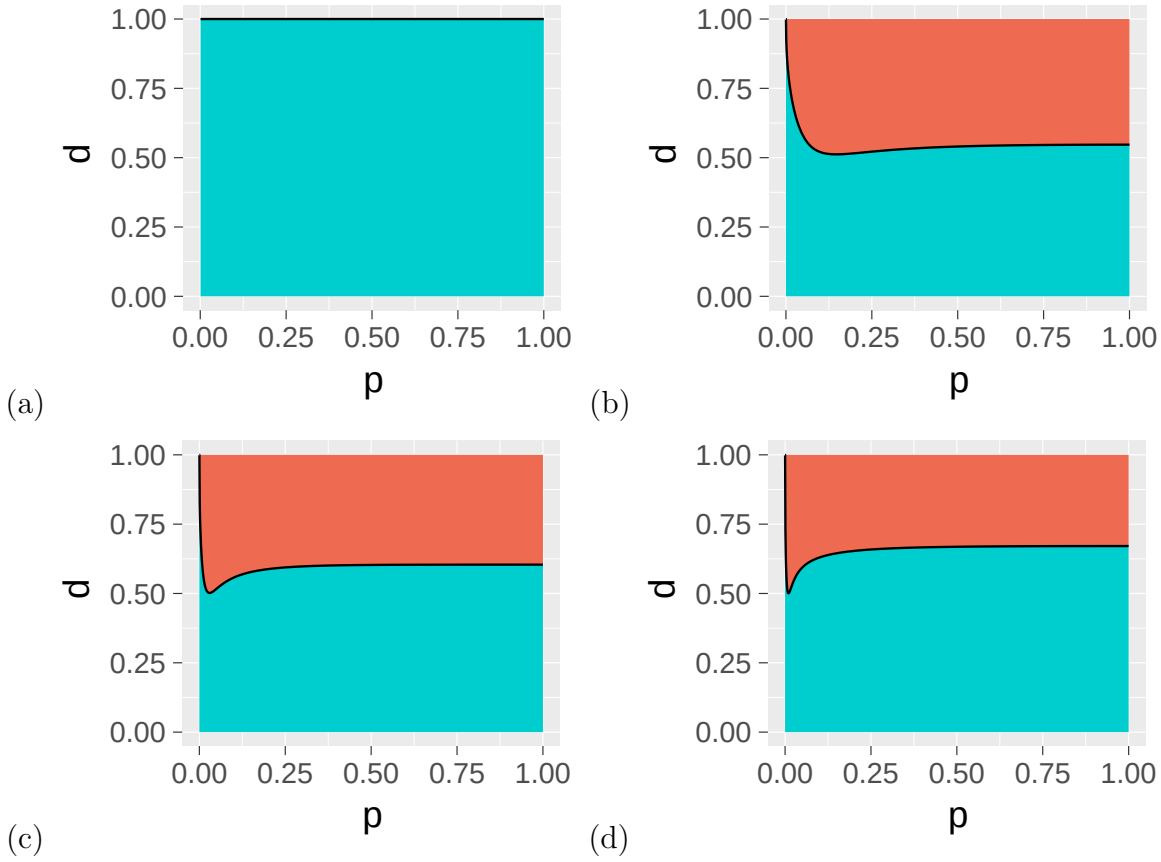

Figure S4: **Phase diagram of clique splitting with excess modularity density Q_x as the external influence is varied.** The values of clique size ratio p and link density d where the M phase occurs is shown in orange and where the S phase occurs is shown in blue. Results are for $\rho = \rho_{max}$ and different choices of the external influence parameter t: (a) t=0 (b) t=1 (c) t=5 (d) t=15.

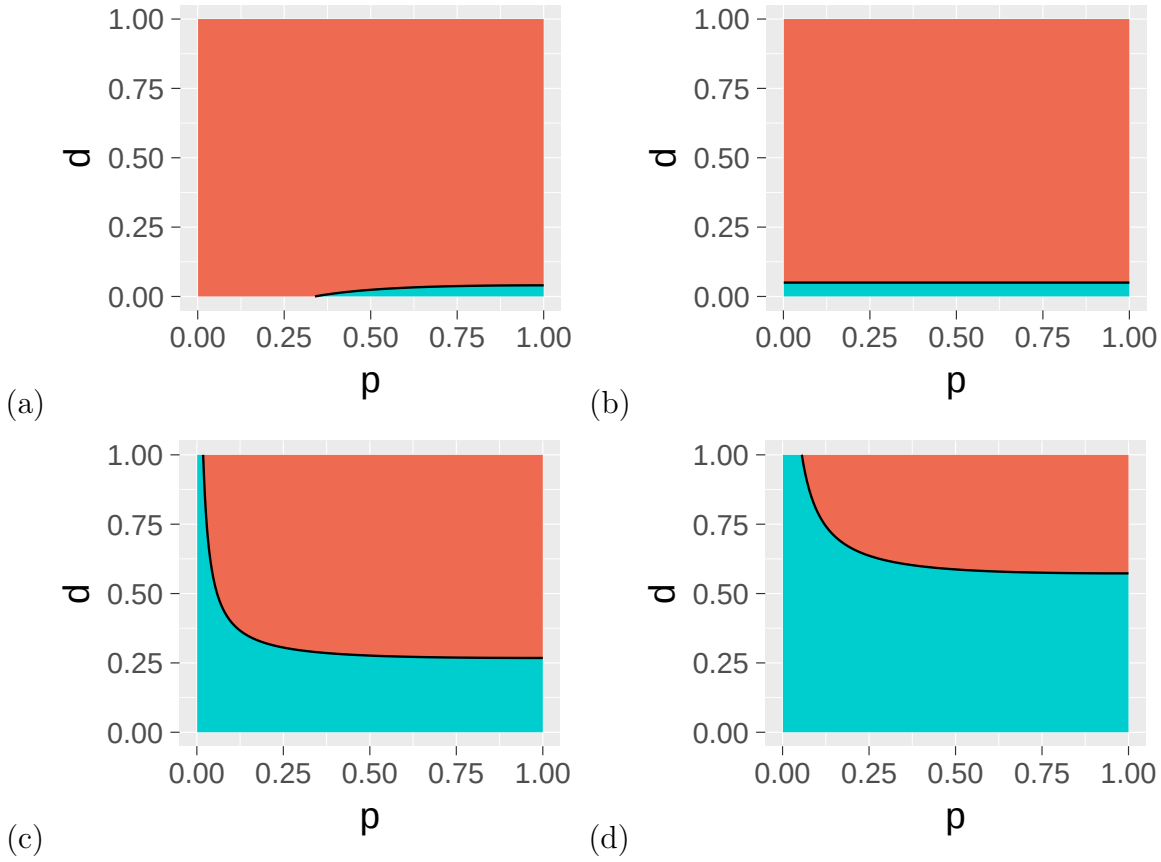

Figure S5: **Phase diagram of clique splitting with Q_{AFG} at fixed external influence t as the parameter s is varied.** The values of clique size ratio p and link density d where the M phase occurs is shown in orange and where the S phase occurs is shown in blue. Results are for $n_a = \sqrt{2m_a}$, $t = 10$ and different choices of s: (a) $s = -\frac{m}{2N}$ (b) $s = 0$ (c) $s = \frac{m}{2N}$ (d) $s = \frac{m}{N}$.