Online Resourse 3

immediate

March 2, 2019

1 Quality assessment

1.1 Time complexity

Assume that n is the number of the vertices, d the average degree of the graph and k the number of communities in the graph before the refinement step. We assume that the graph is stored in the edge adjacency list form. This means, given a node u, we can find whether v is a neighbor of u or not in at most $O(d_u)$ time and in $O(d)$ time on average. Then $\rho(u, v)$ for a pair (u, v) can be computed in $O(d^2)$ time and maximum $\rho(u, v)$ for a fixed u or v in $O(d^3)$ time. Thus, the procedure CommGenerate would take $O(nd^3)$ time. On the other hand the procedure CommRefine would take $O(nk)$ time. Hence, the average complexity of NPCD algorithm is $O(nd^3) + O(nk) = O(nd^3)$ or $O(nk)$ as the case may be. To understand the actual speed of NPCD we run it on random networks with order (number of nodes) ranging from 10^4 to 10^5. We generate these networks with the command `sample_gnp(n, p = 5/n)` of *igraph* package on R platform, with n in the above range. The actual running time in seconds is plotted in Fig. 1. It can be seen that NPCD runs super-linearly on this kind of networks.

1.2 Existing overlapping modularity

Employability of an algorithm rests on the fact how qualitative covers it produces. There are a number of ways to assess the quality of the covers of an algorithm, of which modularity kind have been quite popular in the last decade. The modularity of a partition compares the number of edges falling in a community with the expected number of such edges. Originally, modularity was defined for undirected and unweighted networks[4]. Subsequently, many versions of modularity were proposed such as for directed networks with disjoint communities[2], directed networks with overlapping communities[5] and for weighted networks with overlapping communities[3]. Here, we consider the overlapping modularity Q_{ov} by Nicosia et. al. for assessing the quality of modules as it is widely popular among researchers for overlapping community assessment.

We present a brief overview of Q_{ov}. Since it is, in general, defined for directed networks, we recall that a node v has an in-degree d_v^{in} – the number of incoming links and an out-degree, d_v^{out} – the number of outgoing links. Apart from in-degree and out-degree of a node, there are two other factors that are crucial for the formulation of Q_{ov}.

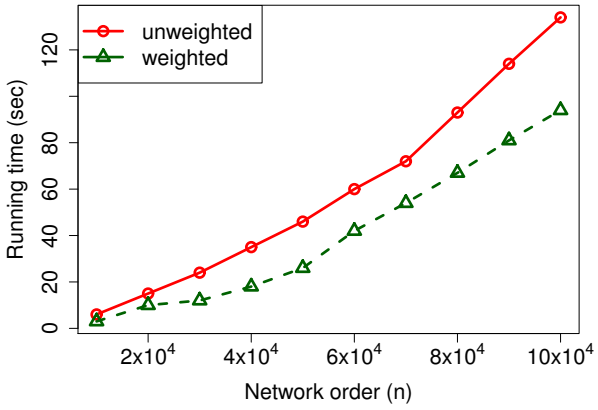


Fig. 1: Actual running time of NCPD on random networks

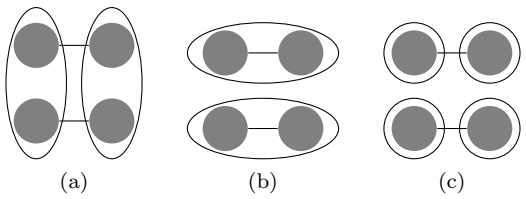


Fig. 2: Communities as rings or union of rings, where each filled circle represents a ring.

One is the *node belonging factor* of a node v to a community C, denoted by $\alpha_{v,C}$, which takes a real value lying between 0 and 1, both inclusive. The factor $\alpha_{v,C}$ essentially represents how strongly v belongs to C. Another is the *edge belonging factor* of a directed edge $e(u,v)$, denoted by $\beta_{e(u,v),C}$. The authors do not specify any particular form for it. Instead, they take $\beta_{e(u,v),C} = \mathcal{F}(\alpha_{u,C}, \alpha_{v,C})$. Then comes the choice of null model. A null model should be chosen in such a way that the in-degree and out-degree of each of its node is the same as in the original graph. This creates two variants of the edge belonging factor, namely $\beta^{in}_{e(u,v),C}$ and $\beta^{out}_{e(u,v),C}$, defined respectively as

$$\beta^{in}_{e(u,v),C} = \frac{\sum_{u \in V} \mathcal{F}(\alpha_{u,C}, \alpha_{v,C})}{|V|}$$

$$\beta^{out}_{e(u,v),C} = \frac{\sum_{v \in V} \mathcal{F}(\alpha_{u,C}, \alpha_{v,C})}{|V|}$$

Now the formula for Q_{ov} is given in Eq. (1).

$$Q_{\text{ov}} = \frac{1}{|E|} \sum_{C \in \mathcal{K}} \sum_{u,v \in V} \left[\beta_{e(u,v),C}\, a_{uv} - \frac{\beta^{out}_{e(u,v),C}\, d^{out}_u\, \beta^{in}_{e(u,v),C}\, d^{in}_v}{|E|} \right] \tag{1}$$

Note that here a_{uv} assumes only the values 0 and 1. To make the formula in Eq. (1) computable, an appropriate form for $\mathcal{F}$ needs to be chosen. This is done via the following choice:

$$\mathcal{F}(\alpha_{u,C}, \alpha_{v,C}) = \frac{1}{\left(1 + e^{-f(\alpha_{u,C})}\right)\left(1 + e^{-f(\alpha_{v,C})}\right)}$$

where we take $f(x) = 60x - 30$.

The overlapping modularity Q_{ov} fundamentally satisfies the same requirements as the modularity of Newman does, and hence it suffers with the same resolution problems. To be specific, it does not punish the existence of "giant" communities – the communities that contain disproportionately large number of vertices. Additionally it is limited to unweighted networks. To resolve these issues, we need to introduce a new modularity measure.

1.3 Weighted overlapping modularity

Here we develop a new quality measure, called *weighted overlapping modularity*, for assessing overlapping communities in weighted networks. We shall formulate this notion in a number of steps.

Let $\mathcal{K}$ be a cover and $C \in \mathcal{K}$. Then if C is a good community (either in strong or weak sense), then the ratio $d^{\text{int}}(C)/d(C)$ is high i.e., close to 1. However, its converse is not true, i.e., a high score of $d^{\text{int}}(C)/d(C)$ does not always lead to the conclusion that C is a good community. To illustrate it let us consider the community structure as shown in Fig. 2, where filled circles represent rings(cycles) of the same size, say n. Then each of the communities in Fig. 2(a) and (c) has the same ratio $d^{\text{int}}(C)/d(C) = n/(n+1)$, which is nearly 1, when n is large. In Fig. 2(b), $d^{\text{int}}(C)/d(C) = 1$, for each community C. However, the communities in Fig. 2(a) are not even connected. Checking whether each of the communities is connected or not is a computationally expensive task. However, one can observe that the community structure in Fig. 2(c) has **communities more** than the community structures in Fig. 2(a) and (b) have, hence the former must be assigned a **higher** quality score than the latter two. Thus a good quality measure should be *generous in assigning higher scores to the covers that contain large number of communities.*

Another issue to deal with is how to incorporate the overlap. A close review of the existing overlapping community quality measures reveals that they tend to assign high scores to the covers that contain minimum overlap and highest to the ones that have no overlap at all. We also adopt the same requirement, although it seems

to be unrealistic. Note that $|m_v|$ is the number of communities v belongs to. Then, for a community C, we have

$$|C| \leq \sum_{v \in C} |m_v| \leq |C| \,.\, |\mathcal{K}| \,,$$

which can be rewritten as

$$\frac{1}{|\mathcal{K}|} \leq \frac{\sum_{v \in C} |m_v|}{|C| \,.\, |\mathcal{K}|} \leq 1$$

Now consider the quantity $1 - \frac{\sum_{v \in C} |m_v|}{|C| . |\mathcal{K}|}$. It gets high values when the overlap is minimum or when the number of communities is high. Thus it also includes our first requirement. Now let us consider the following quantity.

$$\left(1 - \frac{\sum_{v \in C} |m_v|}{|C| \,.\, |\mathcal{K}|}\right) \frac{d^{\text{int}}(C)}{d(C)}$$

Note that the two factors in the quantity above are not independent. For example when $|\mathcal{K}|$ increases more edges start falling between the communities and hence $d^{int}(C)$ decreases for all $C \in \mathcal{K}$. Taking the average of the quantity above over all the communities C in $\mathcal{K}$, our weighted overlapping measure takes the form,

$$Q_{\text{wo}}(\mathcal{K}) = \frac{1}{|\mathcal{K}|} \sum_{C \in \mathcal{K}} \left(1 - \frac{\sum_{v \in C} |m_v|}{|C| \,.\, |\mathcal{K}|}\right) \frac{d^{\text{int}}(C)}{d(C)}, \tag{2}$$

Note that $\sum_{v \in C} |m_v| \geq |C|$. Therefore, in case of disjoint communities, Eq. (2) reduces to

$$Q_{\text{wo}}(\mathcal{K}) = \frac{1}{|\mathcal{K}|} \left(1 - \frac{1}{|\mathcal{K}|}\right) \sum_{C \in \mathcal{K}} \frac{d^{\text{int}}(C)}{d(C)} \tag{3}$$

From Eq. (3) it is clear that, for a cover $\mathcal{K}$, the upper bound for Q_{wo} is $1 - 1/|\mathcal{K}|$. That means Q_{wo} never achieves 1. Even the upper bound is achievable only when each community is a connected component.

1.4 Comparison between Q_{wo} and Q_{ov}

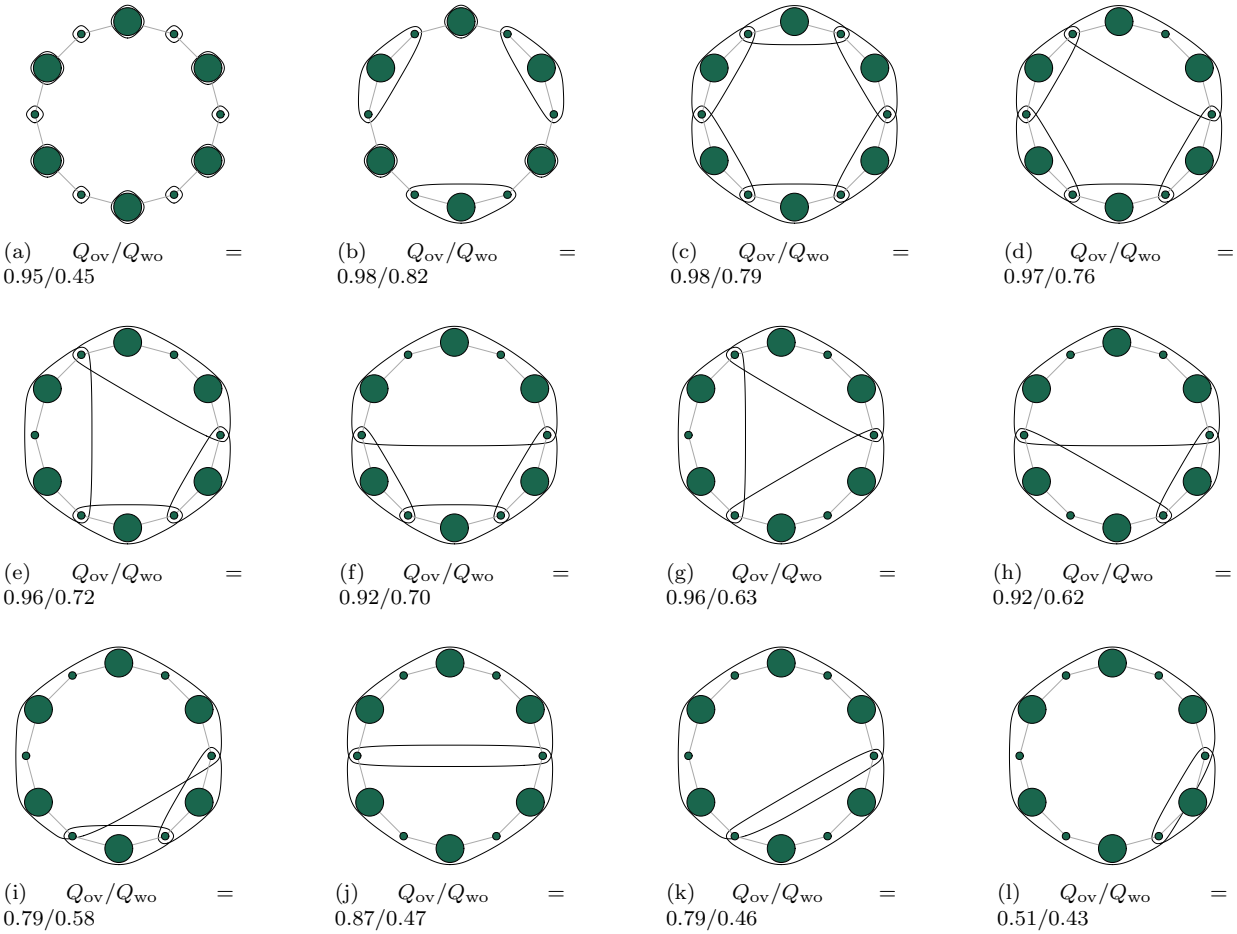
First we shall compare the effectiveness of Q_{wo} over Q_{ov} on model unweighted networks. For this we recall the ring network where the nodes are cliques of fixed size[1]. We alter its structure a bit by inserting "mediator" nodes in between the adjacent cliques. We refer it to as *node mediated ring network*. The model still being simple and highly modular, now incorporates the possibility of overlap among the communities. Such a model with 6 cliques and 6 mediator nodes is shown in Fig. 3(a). The possible community structures in this network, ignoring the alike or identical ones, are shown in Fig. 3 (a)-(l). For the sake of computation of Q_{ov} and Q_{wo}, we assume the size of each clique to be 10.

In Fig 3(a), each clique as well as each mediator node represents a distinct community. The values of Q_{ov} and Q_{wo}, for the cover in Fig. 3(a) are 0.95 and 0.45, respectively. Q_{wo} takes considerably lower value because it treats the singleton but non-isolated nodes as bad communities. In fact, it assigns lower values to the covers that have a large number of their communities with proportionally poor intra-connections. For the community structure in Fig. 3(b), both Q_{ov} and Q_{wo} achieve their highest values. From Fig. 3(c)-(h), as the number of communities reduces to 3, both Q_{ov} and Q_{wo} decrease, although the latter one decreases rapidly. Fig. 3(i) contains a giant community, yet Q_{ov} assigns it a value as high as 0.79 while Q_{wo} – being able to detect that –assigns it a value of 0. The covers in Fig. 3(j)-(k) containing only two communities, none of them being giant, receive higher values by Q_{ov} but lower by Q_{wo}. The cover in Fig. 3(l), again contains a giant community, thereby receiving 0 by Q_{wo}, but unnoticeably 0.51 by Q_{ov}.

References

[1] Santo Fortunato and Marc Barthlemy. Resolution limit in community detection. *PNAS*, 104(1):36–41, 2007. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0605965104.

[2] E. A. Leicht and M. E. J. Newman. Community Structure in Directed Networks. *Physical Review Letters*, 100 (11):118703, March 2008. doi: 10.1103/PhysRevLett.100.118703. URL `https://link.aps.org/doi/10.1103/PhysRevLett.100.118703`.



(a) Q_{ov}/Q_{wo} = 0.95/0.45
(b) Q_{ov}/Q_{wo} = 0.98/0.82
(c) Q_{ov}/Q_{wo} = 0.98/0.79
(d) Q_{ov}/Q_{wo} = 0.97/0.76
(e) Q_{ov}/Q_{wo} = 0.96/0.72
(f) Q_{ov}/Q_{wo} = 0.92/0.70
(g) Q_{ov}/Q_{wo} = 0.96/0.63
(h) Q_{ov}/Q_{wo} = 0.92/0.62
(i) Q_{ov}/Q_{wo} = 0.79/0.58
(j) Q_{ov}/Q_{wo} = 0.87/0.47
(k) Q_{ov}/Q_{wo} = 0.79/0.46
(l) Q_{ov}/Q_{wo} = 0.51/0.43

Fig. 3: The node mediated ring graph

[3] Z. Lu, X. Sun, Y. Wen, G. Cao, and T. L. Porta. Algorithms and Applications for Community Detection in Weighted Networks. *IEEE Transactions on Parallel and Distributed Systems*, 26(11):2916–2926, November 2015. ISSN 1045-9219. doi: 10.1109/TPDS.2014.2370031.

[4] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69 (2):026113, 2004. doi: 10.1103/PhysRevE.69.026113.

[5] V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri. Extending the definition of modularity to directed graphs with overlapping communities. *J. Stat. Mech.*, 2009(03):P03024, 2009. doi: 10.1088/1742-5468/2009/03/P03024.