Notes on community detection in TreeScaper

August 11, 2022

1 Community detection

A network is said to have community structure if the nodes of the network can be easily grouped into sets of nodes such that each set of nodes is densely connected internally and sparsely connected between sets. Community detection (CD) problem of a given network is a mathematical formulation of the problem that seeks a grouping of the nodes, denoted as *community assignment*, that accesses the best community structure on the given network. A community assignment can be expressed as a function

$$\delta : N \times N \to \{0, 1\}$$

where N is the set of nodes and $\delta(i, j) = 1$ if node i, j belong the same community, otherwise, $\delta(i, j) = 0$.

In general, CD problem can be formulated on any graph and allows communities overlap with each other. TreeScaper solves a particular class of CD problem on weighted undirected network that seeks for non-overlaped community assignment.

Conceptually, TreeScaper seeks for community assignment that appoints nodes of the graph to different clusters such that internal connections within cluster is relatively stronger than the inter-clusters connections.

Modularity is a typical way to quantify the goodness of a community structure regrading to a *reference network*. We will first introduce statisical null models that describe the average network with **no** community structure under respective statisical setting. Then, the modularity with respect to the null model can be defined as the quantity that measures how far the given network locally is from the average network. With the modularity of null model well defined, the community assignment that access the largest modularity is interpreted as the optimal community assignment that gives clusters with strongest internal connections with respect to the null model. Finally, we point out that the definition of modularity can be generalized to arbitrary reference network other than the average network from a statisical null model, which is the commonly used definition of modularity.

TreeScaper essentially implements two types of γ-parameterized reference network, the ***constant Potts model(CPM)*** which is a rescaled Erdős–Rényi null model(ERNM) and the ***rescaled configuration null model(CNM)***. The options, ERNM and no-null model(NNM), provided in TreeScaper are special choices of γ in CPM.

1.1 Average network of null model.

The null models of the community detection problem on network G considers a set of networks $\mathcal{G}$ that includes G as sample space. Equip $\mathcal{G}$ with uniform distribution and the null model is conceptually stated as followed.

The average network $\overline{\mathcal{G}}$ has no community structure.

Note that the average here referred to the network consists of edges with edge-wise expected weight among $\mathcal{G}$ as edge weight.

TreeScaper implements the Erdős–Rényi null model(ERNM) and the configuration null model(CNM).

1. **Erdős–Rényi null model.**

   For a given weighted undirected graph G, let $W(G)$ be the sum of all edge weights, referred as edge energy. ERNM considers all graphs that has the same edge energy for given W,

   $$\mathcal{G}^{\text{ERNM}} := \{G : W(G) = W\} \tag{1}$$

   Let n be the number of the common set of nodes in ERNM, then edge weight between node i and node j in the average network of ERNM, $\overline{\mathcal{G}^{\text{ERNM}}}$, is given by

   $$\overline{w_{ij}^{\text{ERNM}}} = \frac{W}{\begin{pmatrix} n \\ 2 \end{pmatrix}}. \tag{2}$$

2. **Configuration null model.**

   For a given weighted undirected graph G with n nodes. Let $\mathbf{d} := \{d_i\}_{i=1}^{n}$ be the sequence of degree of node i, $i = 1, \cdots, n$, where the degree of node is the sum of edge weights that attached to that node. CNM considers all graphs that has the same degree sequence $\mathbf{d}$:

   $$\mathcal{G}^{\text{CNM}} := \left\{ G : \sum_{j=1}^{n} w_{ij} = d_i, \forall i \right\}. \tag{3}$$

   The average network of CNM, G^{CNM}, has edge weight

   $$\overline{w_{ij}^{\text{CNM}}} = \frac{d_i d_j}{2W} = \frac{d_i d_j}{2\sum_{k=1}^{n} d_k} \tag{4}$$

From ERNM it is straightforward to conclude that $\overline{\mathcal{G}^{\text{ERNM}}}$ has no community structure in common sense, where all nodes are connected to any node with identical edge weight. The CNM is more subtle but necessary to understand.

Compared to ERNM that distributes the overall edge energy evenly, CNM restricts the idea to distributing the edge energy in every node evenly, as $\forall \hat{G} \in \mathcal{G}^{\text{CNM}}$, any node i in $\hat{G}$ has the same degree with itself in G.

This restriction yields a simple implication, those nodes with large degree in $\mathcal{G}^{\text{CNM}}$ has relatively stronger connections to all nodes in $\overline{\mathcal{G}^{\text{CNM}}}$ and vice versa, as any node's degree are distributed evenly with respect to $\mathcal{G}^{\text{CNM}}$ in CNM average network. The following example illustrates the difference between ERNM and CNM on a node that is weakly connected to others in G.

Example 1. Consider a set of 20 points in $\mathbb{R}^2$ distributed as Fig. 1.

Compute the Eucldiean distance d_{ij} between node i and j. Then use the reciprocal of distance to define edge weights,

$$w_{ij} = 1 - \frac{d_{ij}}{\max_{k,l \in N} d_{kl}},$$

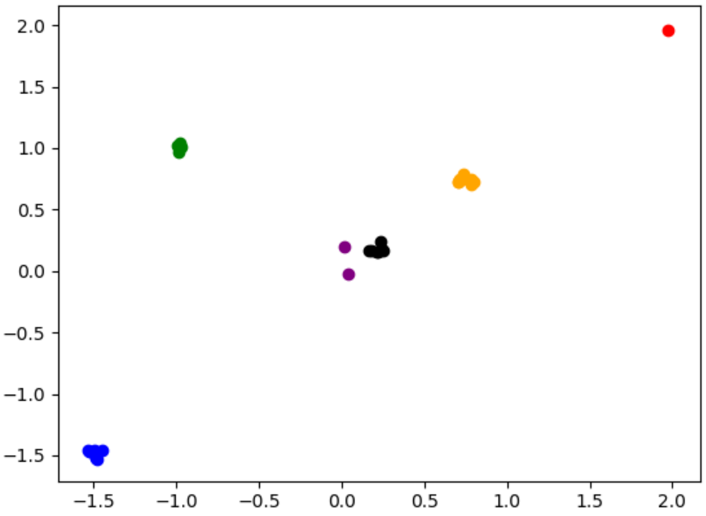

Figure 1: Points on $\mathbb{R}^2$

denoted as the affinity of node i, j. The larger w_{ij} is, the closer node i, j are in $\mathbb{R}^2$. The affinity network and its ERNM and CNM reference networks are given in Fig. 2.

The following figure illustrates different average network by plotting the internal edges in communities $\{\text{blue}\}$, $\{\text{green}\}$ and $\{\text{black}, \text{purple}, \text{gold}, \text{red}\}$, using thickness to represent edge weights. The inter-community edges are omitted to highlight the difference associated to the red node.

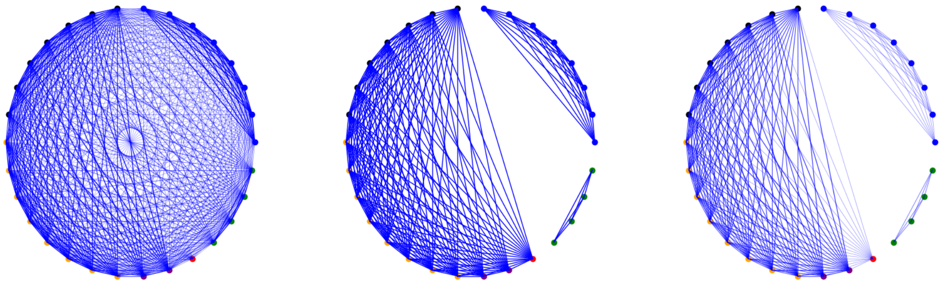

Figure 2: Affinity network G, internal edges of $\overline{\mathcal{G}^{\text{ERNM}}}$ and internal edges of $\overline{\mathcal{G}^{\text{CNM}}}$

Notice that for the red node in the top right, has significantly larger distance to all other nodes, which makes the corresponding degrees in affinity network significanltly smaller. In $\overline{\mathcal{G}^{\text{ERNM}}}$, the red node is expected to joined to other nodes with a relatively stronger edge, compared to the corresponding edges in G. In $\overline{\mathcal{G}^{\text{CNM}}}$, the model address the issue of the red node being weakly connected and therefore it is expected to have relatively weaker connection in the average network.

Both null models make sense but they produce different average network and therefore different notion of community structure. One should choose the null model based on their knowledge of the sample space $\mathcal{G}$ and/or their knowledge of the given G.

1.2 Modularity

We have introduced the null model that specifies network without community structure as the average network of some sample spaces of networks. However, the given G being different from the average network $\overline{\mathcal{G}}$ is only necessary but not sufficent condition for G to have community structure.

A community structure in G requires not only having edge weights different from the expected edge weights in $\overline{\mathcal{G}}$ but also having those stronger connections being clustered inside communities and those weaker connections being placed in between communities for some community assignment.

Consider ERNM, for example, with a community assignment δ on G that has perfectly identified community structure in the following sense,

$$\sum_{\delta(i,j)=1} w_{i,j} \geq \sum_{\delta(i,j)=1} \overline{w_{i,j}} \tag{5a}$$

$$\sum_{\delta(i,j)=0} w_{i,j} \leq \sum_{\delta(i,j)=0} \overline{w_{i,j}} \tag{5b}$$

One may have noticed that for ERNM, the same total edge weights

$$W = \sum_{i,j\in N} w_{i,j} = \sum_{\delta(i,j)=1} w_{i,j} + \sum_{\delta(i,j)=0} w_{i,j} = \sum_{i,j\in N} \overline{w_{i,j}} = \sum_{\delta(i,j)=1} \overline{w_{i,j}} + \sum_{\delta(i,j)=0} \overline{w_{i,j}}$$

implies that (5a) is equivalent with (5b) as followed.

$$\begin{aligned} \sum_{\delta(i,j)=1} w_{i,j} + \sum_{\delta(i,j)=0} w_{i,j} &= \sum_{\delta(i,j)=1} \overline{w_{i,j}} + \sum_{\delta(i,j)=0} \overline{w_{i,j}} \\ \Rightarrow \sum_{\delta(i,j)=1} w_{i,j} \geq \sum_{\delta(i,j)=1} \overline{w_{i,j}} &\Leftrightarrow \sum_{\delta(i,j)=0} w_{i,j} \leq \sum_{\delta(i,j)=0} \overline{w_{i,j}}. \end{aligned}$$

Similar argument is applicable to the CNM but it is not necessarily applicable to all null models. In TreeScaper, it is sufficent to solely consider (5a) for either ERNM or CNM, which leads to the following quantity, known as *modularity*, for accessing the goodness of community assignment δ in G with respect to a null model.

$$\rho(\delta, G, \overline{\mathcal{G}}) := \sum_{\delta(i,j)=1} w_{ij} - \overline{w_{ij}}. \tag{6}$$

By maximizing (6) over all possible community assignment, the community structure identified in the resulting optimal community assignment is said to be the optimal community structure under the modularity of the null model. The optimization algorithm for maximizing (6) implemented in TreeScaper is described in **?**.

Note that the role of null model in the definition of modularity is to define the average network $\overline{\mathcal{G}}$ as a reference of connection strength. The edge in G with weight larger than the expected edge weight in $\overline{\mathcal{G}}$ is connection strong enough and is better to be an internal connection in a community. Therefore, modularity of a null model can be further generlized to the case where the reference network is free to be specified as followed.

Definition 1.1. For a given network G and a given reference network $G^{\mathrm{Ref.}}$ that shares the same set of nodes, the modularity of a community assignment δ is defined as:

$$\rho(\delta, G, G^{\mathrm{Ref.}}) := \sum_{\delta(i,j)=1} w_{ij} - w_{ij}^{\mathrm{Ref.}}. \tag{7}$$

An equivalent algebraic definition of (7) in quadratic form is given in **?** as:

$$\rho(\delta, G, G^{\mathrm{Ref.}}) := \mathrm{tr}(X_\delta^T (A - A^{\mathrm{Ref.}}) X_\delta) \tag{8}$$

where $A, A^{\text{Ref.}}$ are respective adjacency matrices with edge weights on entries and X_δ is a $0,1$-valued node-to-community matrix. Suppose δ specifies m communities $\{C_j\}_{j=1}^m$ of n nodes $N = \{1, 2, \cdots, n\}$, then X_δ is $n \times m$ matrix that can be given by

$$[X_\delta]_{ij}^{n\times m} = \begin{cases} 1, & i \in C_j \\ 0, & i \notin C_j. \end{cases}$$

Definition 1.2. The optimal community structure of G of modularity with respect to the reference network $G^{\text{Ref.}}$ is given by the maximal δ_* solves from the maximization problem

$$\delta_* := \arg\max_\delta \rho(\delta, G, G^{\text{Ref.}}). \tag{9}$$

1.3 Rescaled average network as reference network.

We have introduced the modularity to arbitrary reference network in (7), it is important to specifiy some rules of the reference network to avoid bad reference network that generate obscure and even useless optimal community structure.

A common practice is to uniformly rescale the average network from some null models by a parameter, denoted as $\gamma \geq 0$. This is usually used to exploit different community structure on different level. The null model set the relative relation between edges in reference networks.

1. The modularity of rescaled ERNM is widely known as constant Potts model(CPM) and it has the following formula.

$$\rho(\delta, G, G^{\text{CPM},\gamma}) := \sum_{\delta(i,j)=1} w_{ij} - \gamma. \tag{10}$$

2. The modularity of rescaled CNM is given by

$$\rho(\delta, G, G^{\text{CNM},\gamma}) := \sum_{\delta(i,j)=1} w_{ij} - \gamma \frac{d_i d_j}{2W}. \tag{11}$$

As we discussed in previous section, the rescaled ERNM has edges relatively the same in the reference network and the rescaled CNM has edges associated to low degree nodes relatively weaker than other edges. Then the parameter γ adjust the absolute level of the reference network. The usage of this one-parameter family of reference networks replies on the observation.

Proposition 1.3. *Let $G^{Ref.}$ be the reference network of a network G.*

1. *If $w_{ij} \geq w_{ij}^{Ref.}$, the community assignment, δ_0, that put all nodes in one community, i.e.,*

$$\delta_0(i, j) = 1, \forall i, j \in N. \tag{12}$$

obtains an optimal community structure.

2. *If $w_{ij} \leq w_{ij}^{Ref.}$, the community assignment, δ_0, that put each node in a respectively one-node-community, i.e.,*

$$\delta_1(i, j) = 0, \forall i \neq j \in N. \tag{13}$$

obtains an optimal community structure.

Proof. The proof is straight-forward. If $w_{ij} \geq w_{ij}^{\text{Ref.}}$, any edge is stronger than the reference network and therefore all edges should be included as internal edge, which implies δ_0 being the optimal community assignment that includes all edges as internal edge of a big community.

To see it algebraically, consider any community assignment δ different from δ_0, i.e., $\{(i,j) : \delta(i,j) = 0\}$ is not empty, such that $\delta(k,l) = 0$. Then

$$\begin{aligned}
\rho(\delta_0, G, G^{\text{Ref.}}) &= \sum_{i,j} w_{ij} - w_{ij}^{\text{Ref.}} \\
&= \sum_{\delta(i,j)=1} w_{ij} - w_{ij}^{\text{Ref.}} + \sum_{\delta(i,j)=0} w_{ij} - w_{ij}^{\text{Ref.}} \\
&\geq \sum_{\delta(i,j)=1} w_{ij} - w_{ij}^{\text{Ref.}} \\
&= \rho(\delta, G, G^{\text{Ref.}}).
\end{aligned}$$

The equality holds if and only if $w_{ij} - w_{ij}^{\text{Ref.}}$ for any i, j that satisfies $\delta(i,j) = 0$.

Similar proof is easy to generate for the other case. □

Corollary 1.4. *For one-parameter rescaled network given by adjacency matrix $A^{Ref.,\gamma} := \gamma \cdot A^{Ref.,1}$ with $w_{i,j}^{Ref.,1} > 0 \forall i,j \in N$, the following statements hold.*

1. *δ_0 obtains an optimal community structure of G with respect to the reference network $G^{Ref.,0}$ which has 0 adjacency matrix.*

2. *$\exists \gamma_{\max} > 0$ such that $\forall \gamma > \gamma_{\max}$, δ_1 obtains an optimal community structure of G with respect to the reference network $G^{Ref.,\gamma}$. In addition, $\gamma_{\max} = 1$ for CPM.*

According to coro, increasing δ from 0 to some large enough value, the optimal community structure change from one big community to all nodes being separated from each other, the intermediate optimal community structure during the scan on γ is then expected to exploit different level of community strcture. The following Example 2 illustrates the idea of scaning over γ to exploit different structure.

Example 2. Consider a weighted network G given in Fig. 3. Let $\{\{1,2,3,4\},\{5\}\}$ be the set of communities determined in δ_1 and let $\{\{1,2,3\},\{4\},\{5\}\}$ be the set of communities determined in δ_2. Then the internal edges of δ_1 is given in G_1 and the internal edges of δ_2 is given in G_2.

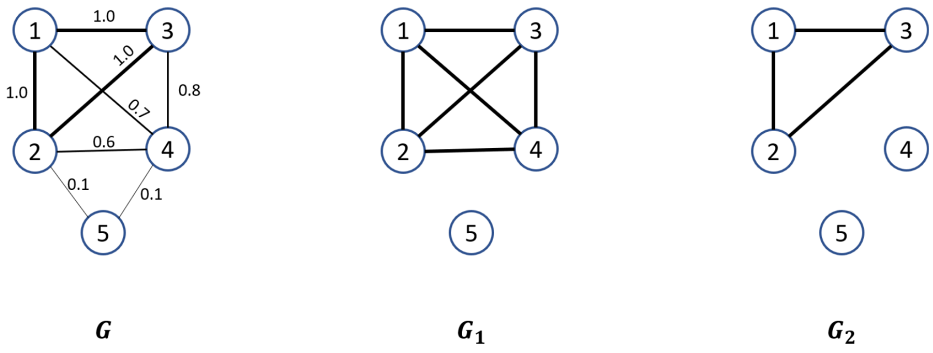

Figure 3: weighted dense network and reference networks

The modularity of δ_1 is given by

$$\begin{aligned}\rho(\delta_1, G^{\gamma,\text{CPM}}) &= (1-\gamma)_{\text{node 1, 2}} + (1-\gamma)_{\text{node 1, 3}} + (0.7-\gamma)_{\text{node 1, 4}} \\ &+ (1-\gamma)_{\text{node 2, 3}} + (0.6-\gamma)_{\text{node 2, 4}} + (0.8-\gamma)_{\text{node 3, 4}} \\ &= 5.1 - 6\gamma.\end{aligned}$$

and

$$\begin{aligned}\rho(\delta_2, G^{\gamma,\text{CPM}}) &= (1-\gamma)_{\text{node 1, 2}} + (1-\gamma)_{\text{node 1, 3}} + (1-\gamma)_{\text{node 2, 3}} \\ &= 3 - 3\gamma.\end{aligned}$$

When $\gamma < 0.7$, P_1 is better than P_2 and when $\gamma > 0.7$, P_2 is better than P_1. Recall that the γ in the reference network represents the reference strenght of the internal connections. When $\gamma < 0.7$, on average, all existing edges are considered contributing to grouping nodes and therefore all nodes should be in a same group. When $\gamma > 0.7$, those edges with 4 are too weak compared to the reference network and therefore they no longer make credits in grouping 4 to other nodes.

The Example 3 continued from Example 1 illustrates the implications of the choice of null model, or more precisely the choice of $G^{\text{Ref.},1}$.

Example 3. In CPM, the nodes are joined with idenical edge weights γ, and by increasing γ, those groups that are relatively far away from the center are picked up sooner than rescaled CFM, as CFM takes degrees of nodes into consideration and has relative weaker standard for those far-away-points.

Figures in Fig. 4 are weight difference $w_{ij} - w_{ij}^{\text{Ref.},\gamma}$ of internal edges and from left to right the figures have increasing γ. The thickness represents the the absolute difference and the red color indicates negaive difference.

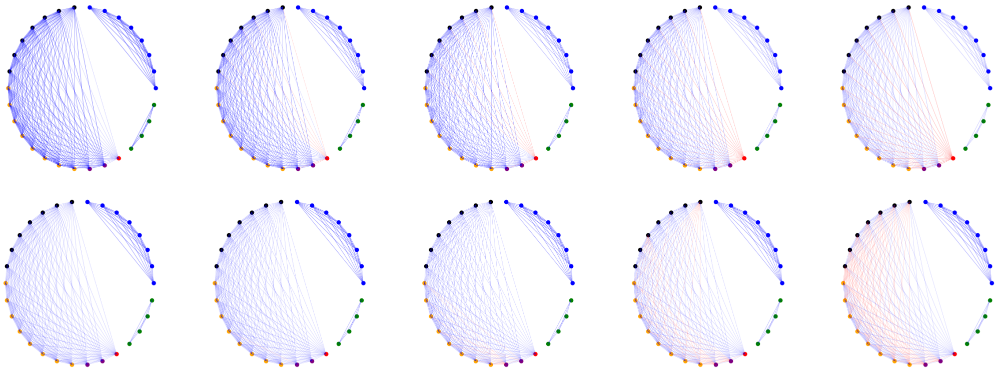

Figure 4: Edge differences in CPM (on top) and rescaled CNM (below) with γ increasing from left to right.