Appendix (Community Detection Models)

Community detection includes a broad class of methods that attempt to find structure in networks, by identifying groups of nodes that are more densely or tightly connected to each other than they are to other nodes in the network (Newman 2010). These methods do not require the number and size of groups (known as communities) to be identified in advance, in contrast to graph partitioning approaches. Each of the community detection methods implemented in TreeScaper employ a quantity known as the Hamiltonian (*H*). Roughly analogous to the use of this term in quantum mechanics, the Hamiltonian represents the energy imposed by a given community structure. The structure with the minimum energy represents the most natural division of nodes. Below we provide definitions of *H* for each of the methods in TreeScaper, as well as some explanation of how these definitions influence the detected communities, in order to help users efficiently explore parameter space and properly interpret model output. We also direct those interested to more in-depth explanations of these methods in papers by Fortunato (2010), Reichardt and Bornholdt (2006), Raghavan et al. (2007), and Traag et al. (2011).

No Null Model

In the first case, which we term No Null Model (NNM, also known as the label propagation method), *H* is defined for the set of all communities, $\{\sigma\}$, and is given by

$$H(\{\sigma\}) = -\sum_{ij} A_{ij}\delta(\sigma_i, \sigma_j)$$

where the sum is over all *i* and *j*, A_{ij} is the adjacency between nodes *i* and *j* (i.e., is there an edge connecting *i* and *j*?), and σ_i is the community to which bipartition *i* belongs. $\delta(\sigma_i, \sigma_j)$ is defined as 1 when *i* and *j* are in the same community, and 0 otherwise. The NNM contains no tunable parameters and is generally of the least interest, since its Hamiltonian has only one global optimum with all nodes in a single community. However, local optima could be of some interest.

One way to refine our approach to community detection beyond the NNM involves first defining an expectation for structure based on a stochastic model of network construction. In these cases, the existence of communities can be revealed by comparison between the actual density of edges in a subgraph and the density one would expect to have in the null subgraph without community structure. The expected edge density depends on the chosen null model. The two following methods are based on different choices of null model.

Erdős-Rényi Model

For the Erdős-Rényi Model (ERM), H is given by

$$H(\{\sigma\}) = -\sum_{ij}\left[A_{ij} - c^2\left(p_{ij}^{+}\lambda^{+} - p_{ij}^{-}\lambda^{-}\right)\right]\delta\left(\sigma_i, \sigma_j\right)$$

where c is the size of the community in which i and j are placed, p_{ij} is the probability of a positive (p_{ij}^{+}) or negative edge (p_{ij}^{-}) between i and j in the null case, while λ^{+} and λ^{-} are tunable parameters. All other definitions are as above for the NNM. For Erdős and Rényi's random graph model, the probability of occurrence for any particular positive edge (p_{ij}^{+}) is m^{+}/n^2, while the corresponding probability for any particular negative edge (p_{ij}^{-}) is m^{-}/n^2, where m^{+} is the sum of the positive edge weights, m^{-} is the sum of the absolute value of the negative edge weights, and n is the number of nodes.

If community detection is performed on a bipartition covariance network and there are no polytomies in the tree set, then for each bipartition, the sum of its covariances with all other bipartitions equals zero. Correspondingly, the sum of the absolute value of the positive covariances is equal to the sum of the absolute value of the negative covariances. In this case, the above equation simplifies to

$$H(\{\sigma\}) = -\sum_{ij}\left[A_{ij} - c^2 p_{ij}^{+}(\lambda^{+} - \lambda^{-})\right]\delta\left(\sigma_i, \sigma_j\right)$$

For this model, changing the λ tuning parameters affects the preferred community size by adjusting the reward and penalty for including positive and negative edges, respectively, in a community. In the simplified equation, if $\lambda^{+} > \lambda^{-}$, then small communities are preferred. If $\lambda^{-} > \lambda^{+}$, then large communities are preferred.

Configuration Null Model

For the Configuration Null Model (CNM), H is given by:

$$H(\{\sigma\}) = -\sum_{ij}\left[A_{ij} - \left(\frac{k_i^{+}k_j^{+}}{m^{+}}\lambda^{+} - \frac{k_i^{-}k_j^{-}}{m^{-}}\lambda^{-}\right)\right]\delta\left(\sigma_i, \sigma_j\right)$$

where k_i^{+} is the sum of all the positive edges connecting to node i, k_i^{-} is the sum of the absolute value of all negative edges connecting to node i, m^{+} is the sum of the positive edges in community σ, and m^{-} is the sum of the absolute value of all negative edges in community σ. All other terms are as above.

If bipartition-covariance community detection is performed and there are no polytomies in the tree set the above equation simplifies to:

$$H(\{\sigma\}) = -\sum_{ij}\left[A_{ij} - \frac{k_i^+ k_j^+}{m^+}(\lambda^+ - \lambda^-)\right]\delta(\sigma_i, \sigma_j)$$

As with ERM, changing the λ tuning parameters affects the preferred community size by adjusting the reward and penalty for including positive and negative edges, respectively. In the simplified equation, if $\lambda^+ > \lambda^-$, then communities are more likely to include negative edges. If $\lambda^- > \lambda^+$, then communities are less tolerant of negative edges and more strongly favor only positive edges.

Constant Potts Model

For the Constant Potts Model (CPM), H is given by

$$H(\{\sigma\}) = -\sum_{ij}\left[A_{ij} - c^2(\lambda^+ - \lambda^-)\right]\delta(\sigma_i, \sigma_j)$$

where all terms are as above. As with the other models, changing the values of the λ tuning parameters affects the preferred community size. If $\lambda^+ > \lambda^-$, small communities are preferred. If $\lambda^- > \lambda^+$, large communities are preferred.

References

Andrade RF, Rocha-Neto IC, Santos LBL, de Santana CN, Diniz MVC, Lobão TP, Goés-Neto A, Pinho STR, El-Hani CN. (2011) Detecting network communities: an application to phylogenetic analysis. *PLoS Comp. Biol*. **7**:e1001131.

Fortunato S. (2010) Community detection in graphs. *Physics Report* **486**:75-174

Gori K, Suchan T, Alvarez N, Goldman N, Dessimoz C. (2015) Clustering genes of common evolutionary history. *Mol. Biol. Evol.* **33**: 1590-1605.

Hillis DM, Heath TA, St. John K. (2005) Analysis and visualization of tree space. *Syst. Biol*. **54**:471–482.

Lewitus E, Morlon H. (2016) Characterizing and comparing phylogenies from their Laplacian spectrum. *Syst. Biol*. **65**:495-507.

Raghavan UN, Albert R, Kumara S. (2007) Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E* **76**:036106.

Reichardt J, Bornholdt S. (2006) Statistical mechanics of community detection. *Phys. Rev. E* **74**:016110.

Traag VA, Van Dooren P, Nesterov Y. (2011) Narrow scope for resolution-limit-free community detection. *Phys. Rev. E* **84**:016114.