Community Detection in Networks

Boning Yang

Florida State University

November 17, 2017

Framework

- **1. Introduction**
- **2. Properties and Definitions of Community**
- **3. Validation**
- **4. Method**

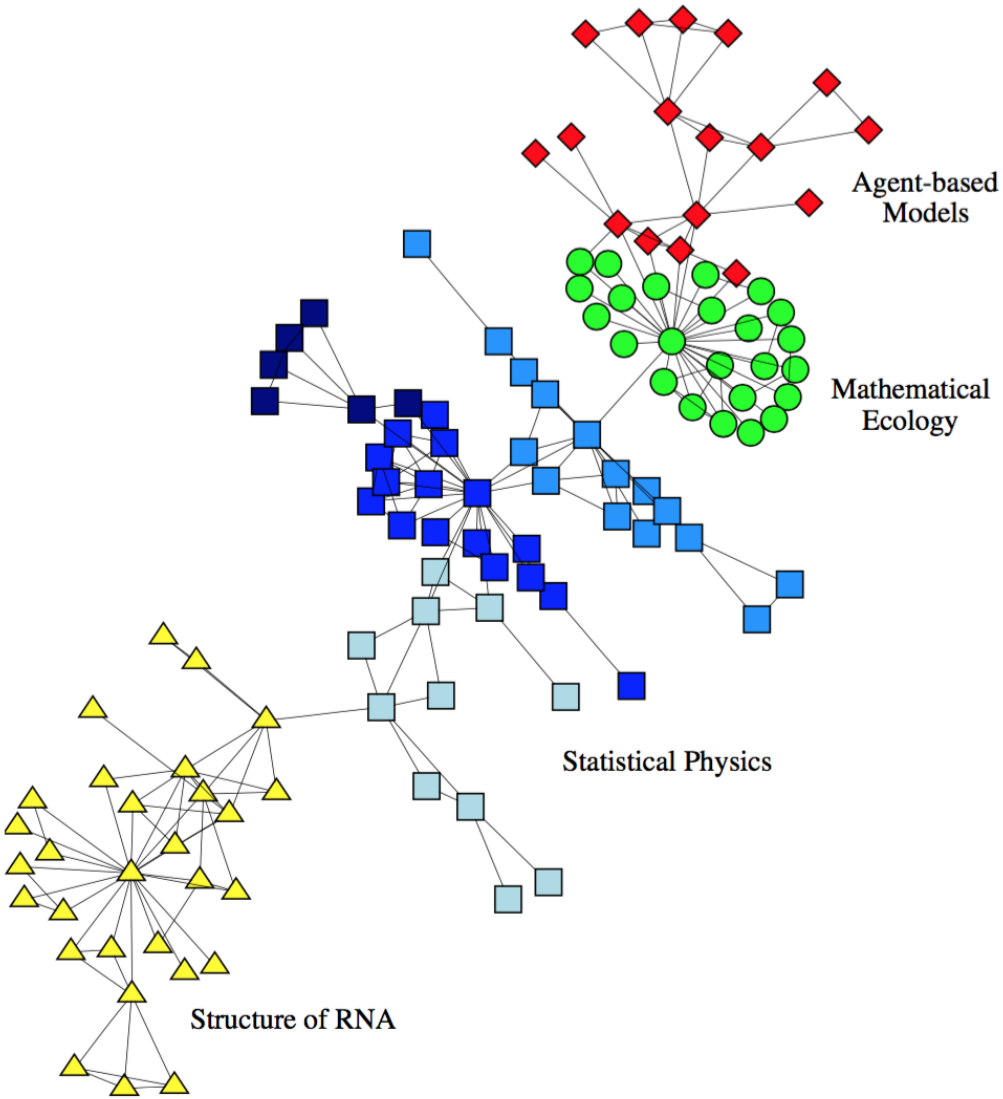
A ill-defined problem!

- No universal definition of the goal that one should look for!
- Advantage: A variety of approaches to the problem.
- Disadvantage: Ambiguity slows down progress.

Convenient assumptions

- No self-loops
- Undirected graph
- Equally-weighted graph

An example


Agent-based Models
Mathematical Ecology
Statistical Physics
Structure of RNA

My intuition of the goal

- Looking for communities that satisfy:
  - The internal part of each community should be cohesive.
  - Different communities should be loosely connected.

Some variables

- n: Number of vertices in graph.
- m: Number of edges in graph.
- n_C: Number of vertices in subgraph C.
- m_C: Number of edges in subgraph C.
- Internal degree k_i^{int} with respect to C: The number of edges connecting vertex i to vertices of C.
- External degree k_i^{ext} with respect to C: The number of edges connecting vertex i to the rest of the graph.

Three kinds of measures

- Internal connectedness
- External connectedness
- Hybrid of the previous two

Internal connectedness

- Internal degree $k_C^{int} = \sum_{i \in C} k_i^{int}$.
- Average internal degree $k_C^{avg-int} = \frac{k_C^{int}}{n_C}$.
- Internal edge density $\delta_C^{int} = \frac{k_C^{int}}{n_C(n_C-1)}$.

External connectedness

- External degree $k_C^{ext} = \sum_{i \in C} k_i^{ext}$.
- Average external degree $k_C^{avg-ext} = \frac{k_C^{ext}}{n_C}$.
- External edge density $\delta_C^{ext} = \frac{k_C^{ext}}{n_C(n-n_C)}$.

Hybrid measure

- Involved with k_C^{int} and k_C^{ext}.
- Total degree $k_C = k_C^{int} + k_C^{ext}$.
- Conductance $C_C = \frac{k_C^{ext}}{k_C}$.

Extension to weighted graph

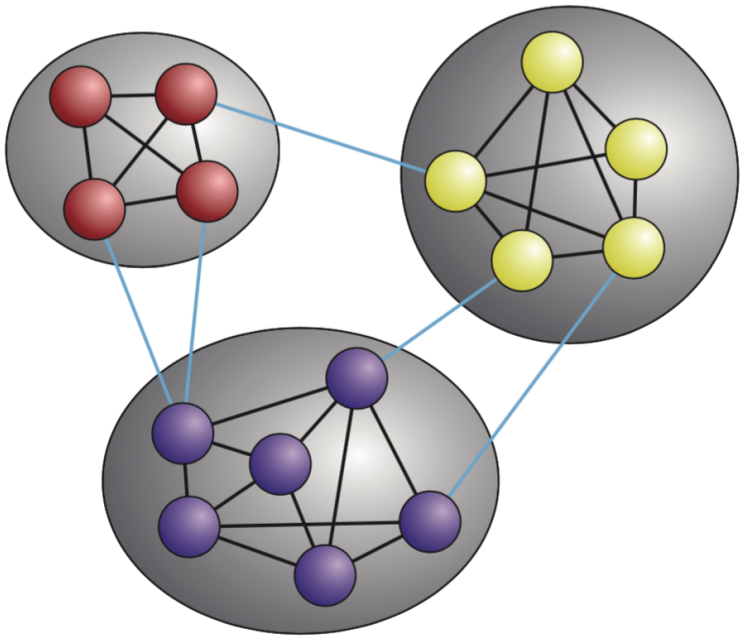
- In the simple case, define most terms with the elements A_{ij} in adjacency matrix A.
- Adjacency matrix A can be viewed as a special case of the weighted matrix W.
- Replacing A_{ij} by W_{ij} in the definition of the previous terms to make a generalization.

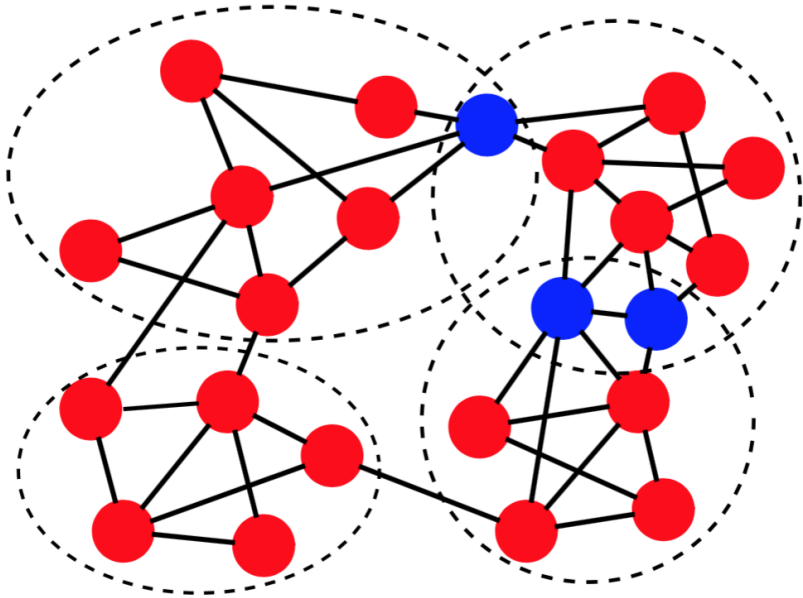
My guess of another measure

- How about random walk view?
- The expected steps to get out of this subgraph?
- I should minus or divide something at the same time.

Classical/Modern View of community

- Disjoint VS Overlap or Partition VS Cover





Left: Classical; Right: Modern.

Classical/Modern View of community

- Traditional definitions of community rely on counting edges in various ways.
- It may be better to focus on the probability that vertices share edges.

A short history of the definition of community

- A k-plex is a maximal subgraph in which each vertex is adjacent to all other vertices of the subgraph except at most k of them.
- A relaxation of complete subgraphs.
- Simply focus on the internal connectedness.

A short history of the definition of community

- A subgraph C is a community if $k_i^{ext} > k_i^{int}$ for all i in C.
- Idea: The number of internal edges should be larger than the number of external edges.
- But treat the rest of the subgraph as a single object.

A short history of the definition of community

- A subgraph C is a strong community if the internal degree of any vertex within C exceeds the number of edges joining the vertex to any other subgraph.
- Treat other communities separately rather than a whole.
- There is an extension of weak community.

A short history of the definition of community

- A strong community is a subgraph each of whose vertices has a higher probability to be linked to every vertex of the subgraph than to any other vertex of the graph.
- But may be difficult to measure the probability in advance.
- There is also an extension of weak community.

Is the definition for community a necessity?

- No, many algorithms don't need to specify the definition of community.
- But also yes, we need to evaluate the algorithms by artificial benchmark(test set).

Why also yes?

- There are not enough real networks to serve as test set, artificial test set is needed.
- In an element of test set, the input is a network, the output is the the set of communities within the network.
- If we want to use artificial test set, we should generate communities according to the definition of community.

Stochastic block model

- This is a famous model to generate artificial test set.
- Suppose we have q communities. The community of vertex i is indicated with the integer label $g_i = 1, 2, \ldots, q$.
- The probability vertices i and j are connected is $P(i \leftrightarrow j) = p_{g_i g_j}$, depending only through the communities they are in.

Stochastic block model

- If $p_{q_1 q_2} = p$, we say it is a random graph.
- Random graph should be included in the test set. Because some algorithms may 'detect' non-existent community.

Partition similarity measure

- The accuracy of community detection techniques depends on their ability to detect the communities networks where the community structures are known.
- So a similarity measure of partitions, or the "loss function" is needed.
- Most similarity measures have their extension of in terms of covers

Variables

- $X = \{X_1, X_2, \ldots, X_n\}$ and $Y = \{Y_1, Y_2, \ldots, Y_n\}$ are partitions.
- n_i^X and n_j^Y: the number of vertices in clusters X_i and Y_j.
- $n_{ij} = |X_i \cap Y_j|$.

Three types of similarity measures

- Pair counting
- Cluster matching
- Information theory.

Pair Counting

- a_{11}: Number of pairs of vertices which are in the same community in both partitions.
- a_{10}: Number of pairs of elements which are in the same community in X and in different communities in Y.
- a_{01}: Number of pairs of elements which are in the same community in Y and in different communities in X.
- $J(X, Y) = \frac{a_{11}}{a_{11}+a_{01}+a_{10}}$

Cluster matching

- $H(X, Y) = \min_{match} \frac{1}{n} \sum_{k=match(k')} n_{kk'}$.
- *match* is a permutation function.

Information Theory

- Most methods are based on the following three terms, x and y are indices of subgraphs in partitions X and Y.
- $P(x, y) = \frac{n_{xy}}{n}$.
- $P(x) = \frac{n_x^X}{n}$.
- $P(y) = \frac{n_y^Y}{n}$.

The influence of sparsity on detectability

- Most network of interest are sparse—the average degree of the graph is much less than the number of vertices.
- Many algorithms can only work on sparsity networks.
- Small noise can change the structure community.

  Algorithms often discover non-existent communities.

Spectral methods

- The eigenvalue spectrum of several graph matrices (e. g., the adjacency matrix, the Laplacian, etc.) typically consists of a dense bulk of closely spaced eigenvalues, plus some outlying eigenvalues separated from the bulk by a significant gap.

Modularity

- $Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) \delta (C_i, C_j)$.
- P_{ij} is chosen according to what property we want to achieve, a standard choice is $P_{ij} = \frac{k_i k_j}{2m}$.
- The δ is the Kronecker Product.

When $P_{ij} = \frac{k_i k_j}{2m}$

- $Q = \sum_C [f_1(C) - f_2(C)]$.
- $f_1(C)$ is the fraction of edges of the graph falling within community C.
- $f_2(C)$ is the expected fraction of edges that would fall inside C if the graph is a random graph preserving the degree of each vertex of the original network.

Other methods

- $L_1(G|g) = \sum_{r,s=1}^{q} e_{rs} \log\left(\frac{e_{rs}}{n_r n_s}\right)$.
- $L_2(G|g) = \sum_{r,s=1}^{q} e_{rs} \log\left(\frac{e_{rs}}{e_r e_s}\right)$.
- e_{rs}: Number of edges running from group r to group s.
- n_r (n_s): Number of vertices in r (s).
- e_r (e_s): Sum of the degrees of the vertices in r (s).
- Need to specify the number of groups beforehand.