CLUSTERING 2 ALTERNATIVES AND IMPROVEMENTS TO MODULARITY

Analysis of Large Scale Social Networks

Bart Thijs

WHAT?

Each cluster approach has to solve common issues:

1. *How to define similarity or sharing of similar properties between objects and between groups?*
2. *Which criterion is used to assign two objects to the same cluster or to divide them from each other?*
3. *How many clusters or groups are required to represent the structure in the data?*
4. *Can items or objects belong to only one group or is multiple assignment possible?*
5. *When is a clustering of items a good clustering?*

MODULARITY BASED APPROACHES

General idea for good clustering:

More edges inside a cluster than outside.

Shift in focus from edges between nodes with lowest similarity to nodes with highest betweenness.

Modularity = fraction of edges within the clusters minus the expected fraction if edges were distributed randomly

See: https://en.wikipedia.org/wiki/Modularity_(networks)

MODULARITY BASED APPROACHES

Two problems:

Resolution limit

Degeneracy problem (modularity plateau)

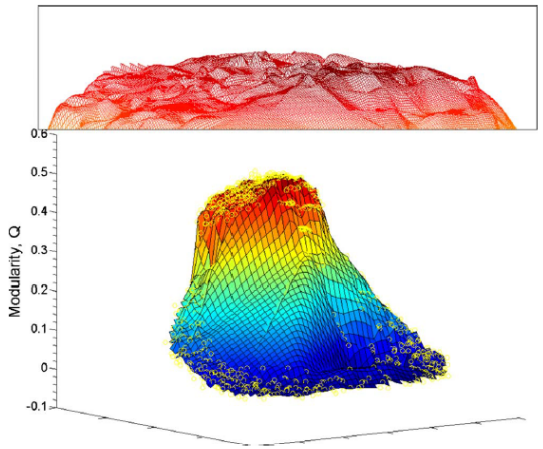


FIG. 3. (Color online) The modularity function of a hierarchical random graph model [47], with $n=256$ nodes arranged in a balanced hierarchy with assortative modules (see Appendix E), reconstructed from 1199 sampled partitions (circles), and its rugged high-modularity region (inset).

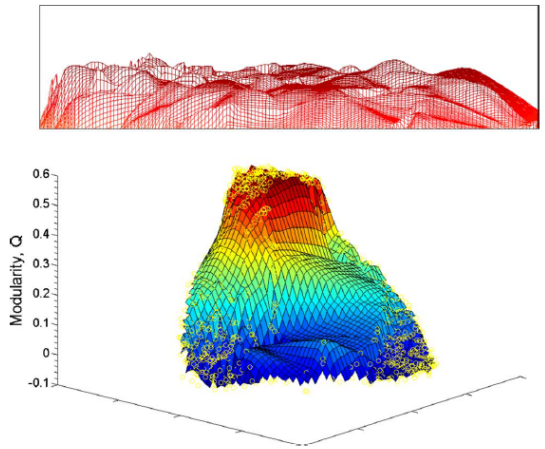


FIG. 4. (Color online) The modularity function for the metabolic network of the spirochaete *Treponema pallidum* with $n=482$ nodes (the largest component) and 1199 sampled partitions, showing qualitatively the same structure as we observed for hierarchical networks. The inset shows the rugged high-modularity region.

MODULARITY BASED APPROACHES

Improvements:

Add resolution parameter

$$\mathcal{H} = \frac{1}{2m}\sum_c \left[e_c - \gamma \frac{K_c^2}{2m} \right],$$

Other Quality Function Constant Potts Model (CPM)

$$\mathcal{H} = \sum_c \left[e_c - \gamma \binom{n_c}{2} \right],$$

INFORMATION BASED APPROACH

Alternative idea for good clustering:

> *The description length of a random walk across nodes and modules is lower when using an optimal partitioning of the network*

Map Equation

See: http://www.mapequation.org/

Lecture based on:
Ludvig Bohlin, Daniel Edler, Andrea Lancichinetti, and Martin Rosvall (2014) *Community detection and visualization of networks with the map equation framework.*

http://www.mapequation.org/assets/publications/mapequationtutorial.pdf

Map Equation

$$L(\mathsf{M}) = q_{\curvearrowleft} H(\mathcal{Q}) + \sum_{i=1}^{m} p_{i\circlearrowleft} H(\mathcal{P}_i)$$

- This function expresses the theoretical minimum length of the description of random walk in a network with partitioning ***M***
- The objective is to minimize this function
- The first term refers to transfers from one module to another
- The second term refers to description of the walk inside the ***m*** different modules

INFORMATION BASED APPROACH

INFORMATION BASED APPROACH

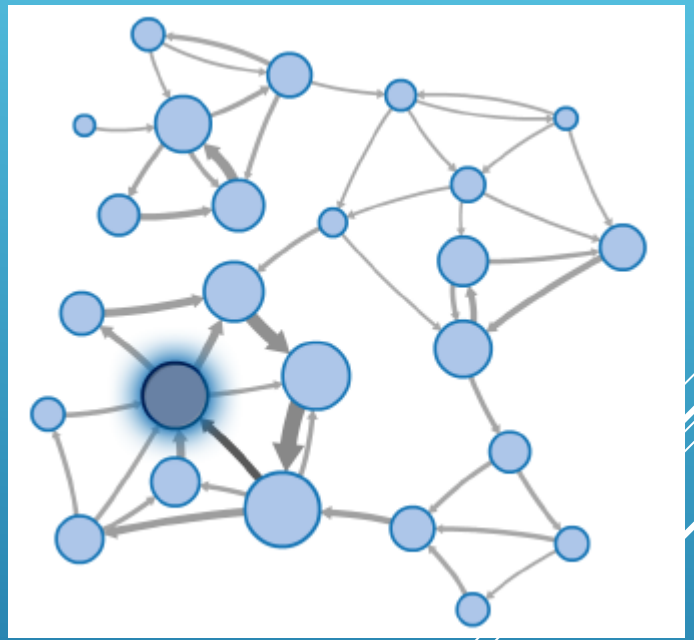
Theoretical minimum length

Shannon's source coding theorem is used to express this minimum length as the entropy of the probability distribution

$$L(\mathcal{P}) = H(\mathcal{P}) \equiv -\sum_i p_i \log p_i,$$

Probability distributions can be calculated for the transition from one module to another and for nodes within a module

The partitioning into modules allows the use to different codebooks (Index & Module)



INFORMATION BASED APPROACH

Probability distributions using random walks

- Probability of step from a to b

$$p_{\alpha\to\beta} = \frac{W_{\alpha\to\beta}}{\sum_\beta W_{\alpha\to\beta}}.$$

- Probability of being at node a

$$p_\alpha = \sum_\beta p_\beta p_{\beta\to\alpha}.$$

- Adding a teleportation parameter proportional to the total weights of links from the node

$$p^*_\alpha = (1-\tau)\sum_\beta p^*_\beta p_{\beta\to\alpha} + \tau\frac{\sum_\beta W_{\alpha\to\beta}}{\sum_{\alpha,\beta} W_{\beta\to\alpha}}.$$

- Probabilities for links and nodes

$$q_{\beta\to\alpha} = p^*_\beta p_{\beta\to\alpha}$$

$$p_\alpha = \sum_\beta q_{\beta\to\alpha}.$$

INFORMATION BASED APPROACH

Probability distributions of modules

- Probability entering or exiting a module

$$q_{i\curvearrowleft} = \sum_{\alpha \in j \neq i, \beta \in i} q_{\alpha \to \beta}$$

$$q_{i\curvearrowright} = \sum_{\alpha \in i, \beta \in j \neq i} q_{\alpha \to \beta}.$$

- Probability of using *Index Codebook*

$$q_{\curvearrowleft} = \sum_{i=1}^{m} q_{i\curvearrowleft}$$

- Probability of using *Module Codebook* of module *i*

$$p_{i\cup} = \sum_{\alpha \in i} p_{\alpha} + q_{i\curvearrowright}$$

Calculating Entropy

- Average length of Index Codebook

$$H(\mathcal{Q}) = -\sum_{i=1}^{m} (q_{i\frown}/q_{\frown}) \log(q_{i\frown}/q_{\frown})$$

- Average length in Module Codebook

$$H(\mathcal{P}^i) = -(q_{i\frown}/p_{i\cup}) \log(q_{i\frown}/p_{i\cup}) - \sum_{\alpha \in i} (p_\alpha/p_{i\cup}) \log(p_\alpha/p_{i\cup})$$

Results in

$$L(\mathsf{M}) = q_{\frown} H(\mathcal{Q}) + \sum_{i=1}^{m} p_{i\cup} H(\mathcal{P}_i)$$

INFORMATION BASED APPROACH

INFORMATION BASED APPROACH

Linkage

- Two-level algorithm closely related to Louvain:
  - Bottom-up
  - Each node is moved to neighbouring node/module
  - Largest decrease in map equation is retained.
  - Additional recursive steps
    - Submodule movements: each module is treated as separate network
    - Single node movements