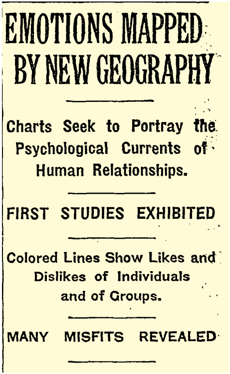
Community Structure in Networks: Practice and Significance

Elizabeth Leicht

Research Fellow CABDyN Complexity Centre

7 January 2011

Learning from Networks



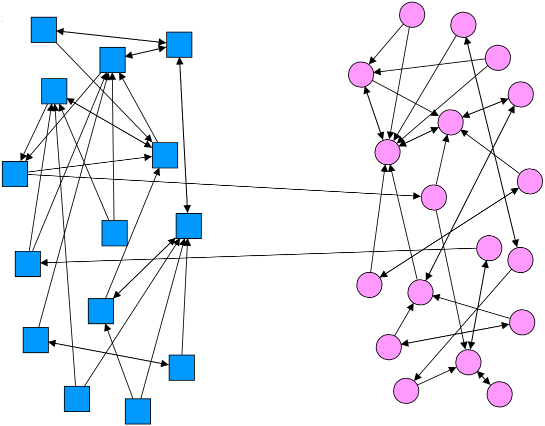
EMOTIONS MAPPED BY NEW GEOGRAPHY

Charts Seek to Portray the Psychological Currents of Human Relationships.

FIRST STUDIES EXHIBITED

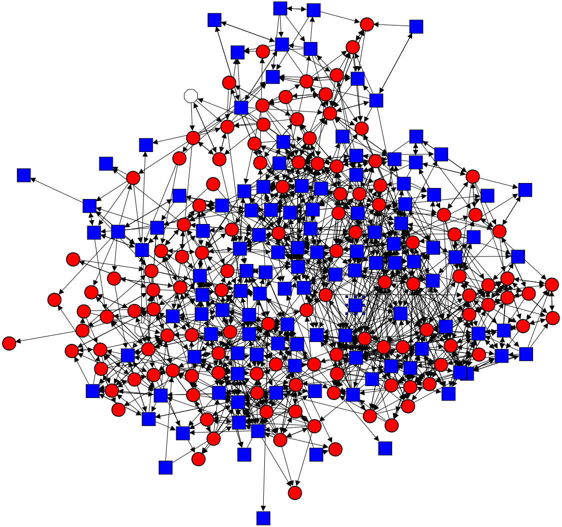
Colored Lines Show Likes and Dislikes of Individuals and of Groups.

MANY MISFITS REVEALED

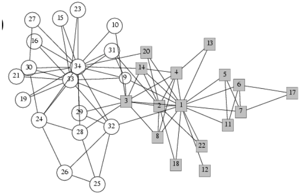


Friendship map of students in a 7th grade class–adapted from *Who Shall Survive*, Jacob Moreno, 1934.

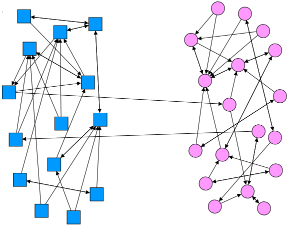
Dealing with large networks



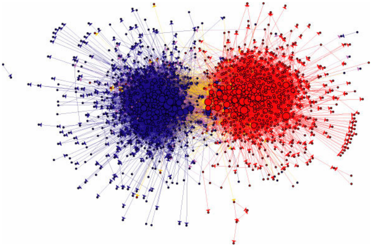
Detecting communities in networks



Girvan & Newman



J. Moreno



Adamic & Glance

Calculating modularity

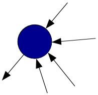
M. E. J. Newman *PNAS* **103**, 8577 (2006).

E. A. Leicht and M. E. J. Newman *Phys. Rev. Lett.* **100**, 118703, (2008).

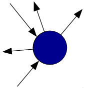
$$Q = \frac{1}{m} \sum_{i,j=1}^{n} \left[A_{ij} - P_{ij} \right] \delta_{c_i, c_j}$$

- $A_{ij} = \begin{cases} 1, & \text{if there is an edge from } j \text{ to } i \\ 0, & \text{otherwise} \end{cases}.$
- $P_{ij} =$ the expected number of edges from j to i.
- $c_i =$ the community to which i belongs.

What is the *expected* number of edges between two nodes?



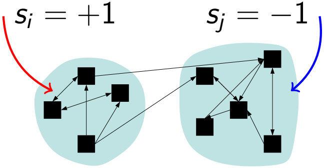
k_i^{in}



k_j^{out}

$$P_{ij} = \frac{k_i^{\mathrm{in}} k_j^{\mathrm{out}}}{m}$$

Division of a network into two communities



$$Q = \frac{1}{2m} \sum_{ij}^{n} \left[A_{ij} - \frac{k_i^{\mathrm{in}} k_j^{\mathrm{out}}}{m} \right] (s_i s_j + 1)$$

$$\delta_{ij} = \frac{1}{2}(s_i s_j + 1)$$

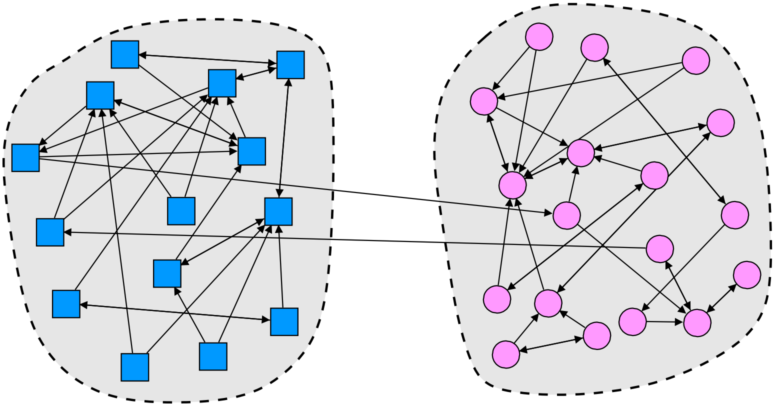
Let $B_{ij} = \left[A_{ij} - \frac{k_i^{\mathrm{in}} k_j^{\mathrm{out}}}{m} \right]$ be an element of the *modularity matrix*.

$$Q = \frac{1}{2m}\mathbf{s}^{\mathrm{T}}\mathbf{B}\mathbf{s} = \frac{1}{2m}\mathbf{s}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{s} = \frac{1}{4m}\mathbf{s}^{\mathrm{T}}\left[\mathbf{B} + \mathbf{B}^{\mathrm{T}}\right]\mathbf{s}$$

Approximate group ID by the sign of the entry for the node in the leading eigenvector, $\mathbf{v}^{(1)}$.

$$s_i = \begin{cases} +1, & \text{if } v_i^{(1)} > 0 \\ -1, & \text{if } v_i^{(1)} < 0 \end{cases}$$

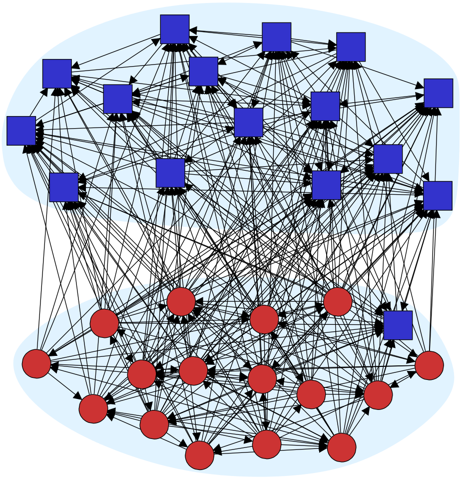
Two communities and more



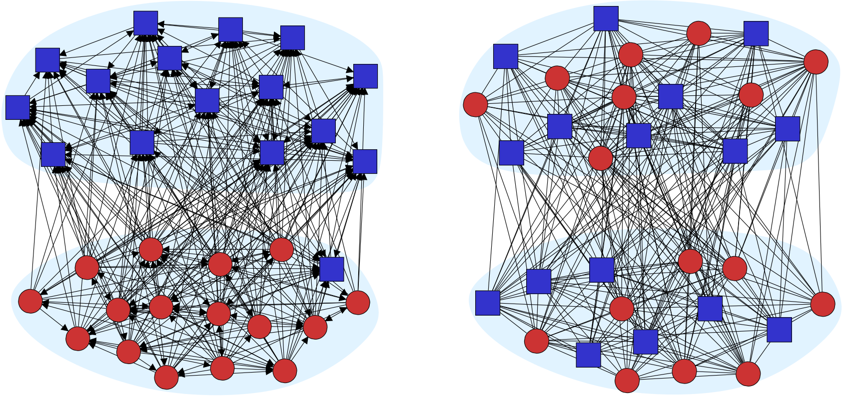
Friendship network from 7th grade class divided into two communities by method.

Communities with bias in edge direction

- Construct a network of n nodes and connect pairs of nodes with probability p.
- Allow random edge direction for *intra-community* edges.
- Bias edge direction for *inter-community* edges.



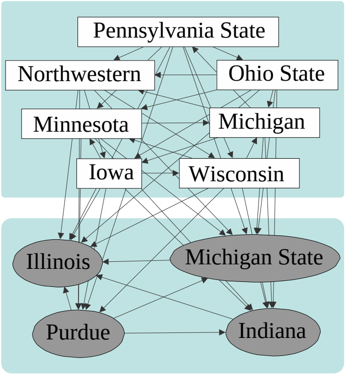
Communities with bias in edge direction



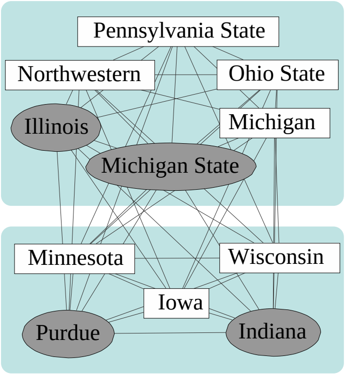
Allowing directed edges Ignoring directed edges[1]

[1] M. E. J. Newman *PNAS* **103**, 8577 (2006).

Edge direction bias in real networks



Accounting for win-loss result

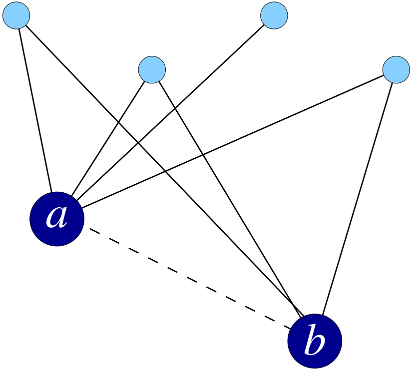


Tracking only games played

American football games among US "Big Ten" schools with directed edges from losing team to winning team.

Exploratory analysis of networks structure

M. E. J. Newman and E. A. Leicht *PNAS* **104**, 9564-9569, (2007.)



- Group identity inferred from network structure.
- A pattern for edges is not pre-determined.

Method: the data and the model

- **Data**
  - Observed: network edges, $A_{ij} \forall i, j$.
  - Missing: group identity of each node, $g_i \forall i$.
- **Model parameters**
  - θ_{ri}: probability there exists an edge from a node in (group) r to a node i.
$$\sum_{i=1}^{n} \theta_{ri} = 1$$
  - π_r: probability a randomly selected node $\in$ (group) r.
$$\sum_{i=1}^{n} \pi_i = 1$$

A likelihood problem

The likelihood of the data given the model is,

$$\Pr(A, g|\pi, \theta) = \Pr(A|g, \pi, \theta) \Pr(g|\pi, \theta)$$

where

$$\Pr(A|g, \pi, \theta) = \prod_{ij} \theta_{g_j,i}^{A_{ij}} \quad \text{and} \qquad \Pr(g|\pi, \theta) = \prod_{j} \pi_{g_j}$$

Frequently, one works not with the likelihood itself, but with the log-likelihood,

$$\mathcal{L} = \ln \Pr(A, g|\pi, \theta) = \sum_{j} \left[\ln \pi_{g_j} + \prod_{i} \theta_{g_j,i}^{A_{ij}} \right]$$

Dealing with missing data

- We cannot directly observe g.
- We can calculate an expected value for the log-likelihood over all possible values of g.

$$\overline{\mathcal{L}} = \sum_{g_1=1}^{c} \ldots \sum_{g_n=1}^{c} \Pr(g|A,\pi,\theta) \sum_i \Big[\ln \pi_{g_i} + \sum_j A_{ij} \ln \theta_{g_i,j} \Big]$$
$$= \sum_{ir} q_{ir} \Big[\ln \pi_r + \sum_j A_{ij} \ln \theta_{rj} \Big]$$

where

$$q_{ir} = \Pr(g_i = r|A,\pi,\theta) = \frac{\Pr(A, g_i = r|\pi,\theta)}{\Pr(A|\pi,\theta)} = \frac{\pi_r \prod_j \theta_{rj}^{A_{ij}}}{\sum_s \pi_s \prod_j \theta_{sj}^{A_{ij}}}$$

An iterative method–the EM algorithm

- Initialize model parameters (θ, π) with random values.

- Find the probability a given node i is a member of group r (E-step).

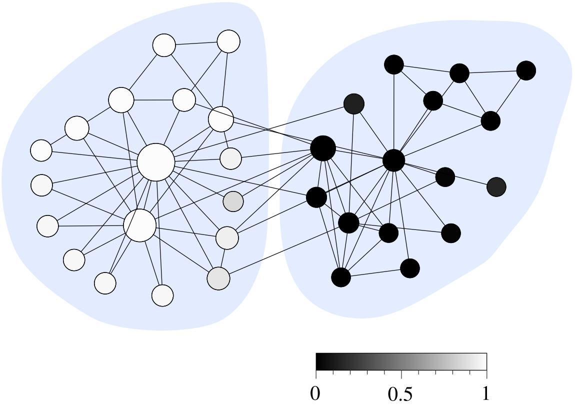
$$q_{ir} = \frac{\pi_r \prod_j \theta_{rj}^{A_{ij}}}{\sum_s \pi_s \prod_j \theta_{sj}^{A_{ij}}}.$$

- Maximize the model parameter (M-step)

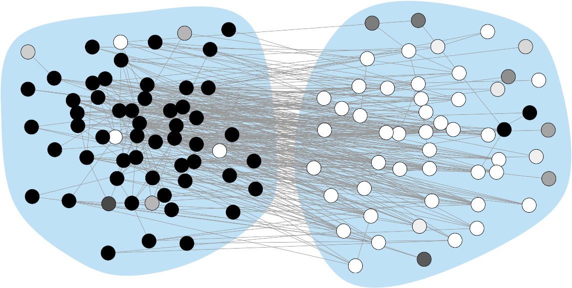
$$\pi_r = \frac{1}{n} \sum_i q_{ir}, \qquad \theta_{rj} = \frac{\sum_i A_{ij} q_{ir}}{\sum_i k_i q_{ir}},$$

- Iterate until convergence.

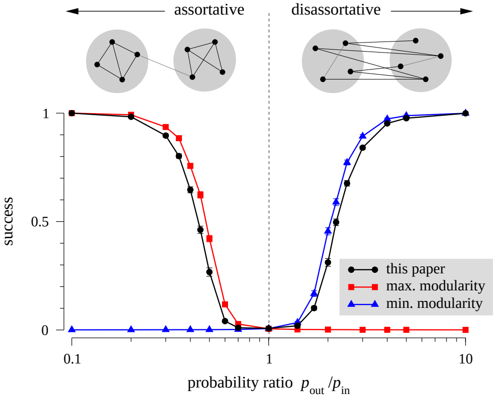
Zachary karate club



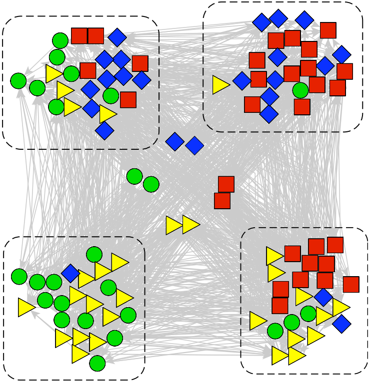
Disassortative word network



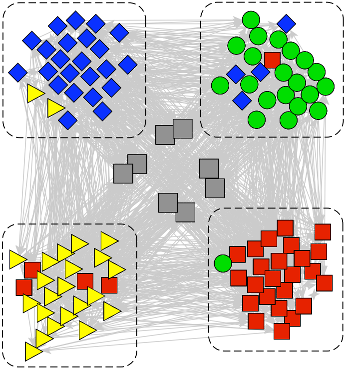
Assortative & disassortative structure



Keystone network



Modularity method



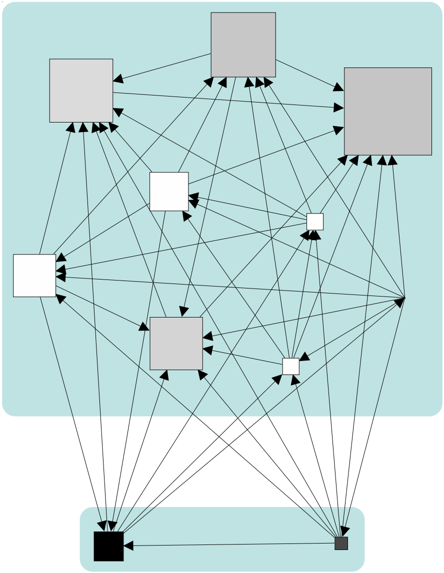
EM method

We assign nodes to groups based on the set of keystone nodes to which they are connected.

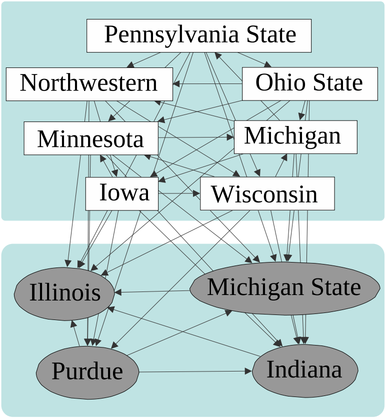
"Big Ten" results with EM approach

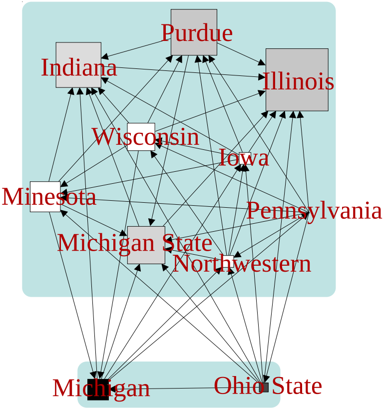
- Node size is proportional to the probability of the team losing to teams assigned to group 1.
- Node shading corresponds to the probability that the node is assigned to group 1.

$q_{j1} \longrightarrow$



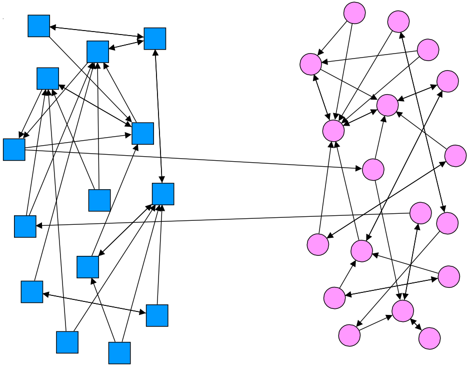
Two methods for one network





Summary

- There are many existing methods for detecting structure in complex.
- Moving forward we need to focus on improving our understanding of what these structures indicate in real networks.



Friendship map of students in a 7th grade class–adapted from *Who Shall Survive*, Jacob Moreno, 1934.