



- Review of community detection
- Community extraction
- Simulation study
- Real data analysis
- Asymptotic consistency
- Future work

Data: links between nodes

- Social and friendship networks, citation networks
- Marketing, recommender systems
- Computer, mobile, sensor networks
- World Wide Web
- Gene regulatory networks, food webs

Given a network  $N = (V, E)$

- $V$  is the set of nodes,  $E$  is the set of edges.
- $N$  is represented by its adjacency matrix  $A$ :

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge from node } i \text{ to node } j, \\ 0 & \text{otherwise.} \end{cases}$$

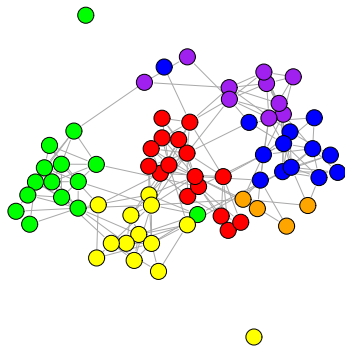
- $A$  can be **symmetric** (undirected network) or asymmetric (directed network).

# Community detection

- Communities: many links within and few links between
- Community detection is typically formulated as finding a **partition**  $V = V_1 \cup \dots \cup V_K$  which gives “tight” communities in some suitable sense.
- For simplicity, give criteria for partitioning into two communities  $V_1$  and  $V_2$ .

# Example: a school friendship network

Colors represent grades



# Graph cuts

- **Min-cut:** minimize

$$R = \sum_{i \in V_1, j \in V_2} A_{ij} .$$

Trivial solution of  $V_1 = V$  or  $V_2 = V$ .

# Graph cuts

- **Min-cut**: minimize

$$R = \sum_{i \in V_1, j \in V_2} A_{ij} .$$

Trivial solution of  $V_1 = V$  or  $V_2 = V$ .

- **Ratio cut** (Wei and Cheng, 1989): minimize

$$\frac{R}{|V_1| \cdot |V_2|},$$

where  $|V_1|$  and  $|V_2|$  are the sizes of the two communities.



- **Min-cut**: minimize

$$R = \sum_{i \in V_1, j \in V_2} A_{ij} .$$

Trivial solution of  $V_1 = V$  or  $V_2 = V$ .

- **Ratio cut** (Wei and Cheng, 1989): minimize

$$\frac{R}{|V_1| \cdot |V_2|},$$

where  $|V_1|$  and  $|V_2|$  are the sizes of the two communities.

- **Normalized cut** (Shi and Malik, 2000): minimize

$$\frac{R}{D_1} + \frac{R}{D_2},$$

where  $D_k = \sum_{i \in V_k, j \in V} A_{ij}$  is the total number of edges from nodes in  $V_k$ .



$$Q = \sum_k \left[ \frac{O_{kk}}{L} - \left( \frac{D_k}{L} \right)^2 \right]$$

- $Q$  is the sum of **observed - expected** under the **configuration model**: probability of edge between nodes with degrees  $d_i, d_j$  is  $d_i d_j / L$ .
- Typically solved by an **eigenvalue method** via relaxing  $\max_{s_j = \pm 1} \mathbf{s}^T \mathbf{M} \mathbf{s}$  to  $\max_{\|\mathbf{s}\|=1} \mathbf{s}^T \mathbf{M} \mathbf{s}$ .

# Limitation of partition methods

- Many real-world networks contain nodes with few links that may not belong to any community (“background”).
- The “strength” of a community depends on links between nodes not related to the community.
- Determining the number of communities is difficult.

- ✓ Review of community detection
  - Community extraction
  - Simulation study
  - Real data analysis
  - Asymptotic consistency
  - Future work

# Community extraction

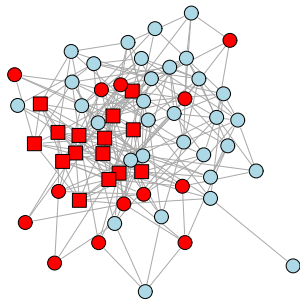
- Allow for **background** nodes that only have sparse links to other nodes.
- Extract communities **sequentially**: at each step look for a set with a large number of links within and a small number of links to the rest of the network.
- Stop when no more meaningful communities exist.

# Toy example

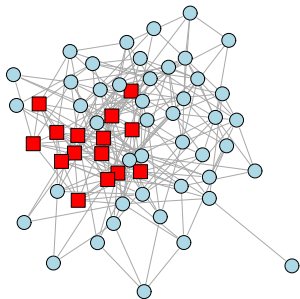
- One community with 15 nodes, total 60 nodes.
- Links between community members form independently with probability 0.5.
- Links between community members and other nodes form independently with probability 0.1.
- Links between other nodes form independently with probability 0.1.
- Compare **partition into two communities** (via modularity) to **extraction of a single community**.

Shapes represent the truth, colors represent results.

Partition



Extraction





# Extraction criterion

Maximize

$$W(S) = \frac{O(S)}{|S|^2} - \frac{B(S)}{|S| \cdot |S^c|},$$

where

$$O(S) = \sum_{i,j \in S} A_{ij}, \quad B(S) = \sum_{i \in S, j \in S^c} A_{ij}.$$

The links **within the complement** of set  $S$  do not matter.

# Adjusted criterion

- In **sparse** networks, tends to pick small disconnected components first.
- To avoid small communities, can use

Maximize

$$W_a(S) = |S| \cdot |S^c| \left( \frac{O(S)}{|S|^2} - \frac{B(S)}{|S| \cdot |S^c|} \right).$$

The factor  $|S| \cdot |S^c|$  encourages more balanced solutions.

- **Tabu Search** (Glover, 1986; Glover and Laguna, 1997): a local optimization technique based on **label switching**.
- Switch labels to improve the value of the criterion but each node has to keep its label for at least  $T$  iterations.
- Run the algorithm for many randomly ordered nodes.

- ✓ Review of community detection
- ✓ Community extraction
  - Simulation study
  - Real data analysis
  - Asymptotic consistency
  - Future work

# Numerical evaluation

- $S$  is the extracted community.
- $C_S$  is the true community that matches  $S$  best.

## PPV and NPV

$$\text{PPV} = \frac{|C_S \cap S|}{|S|}$$

Purity

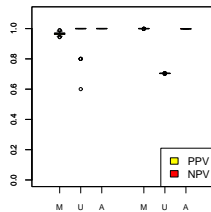
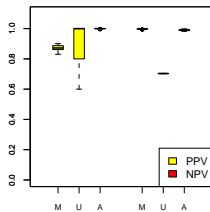
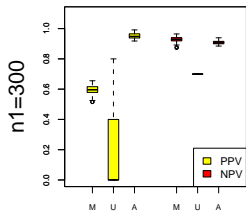
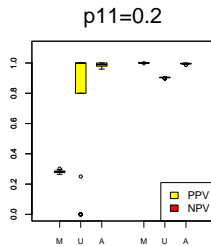
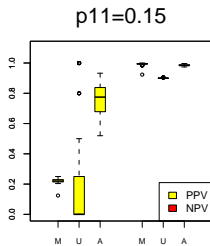
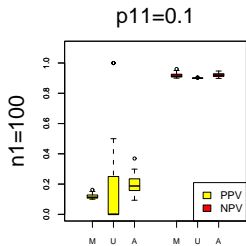
$$\text{NPV} = 1 - \frac{|C_S \cap S^c|}{|S^c|}$$

Completeness

# Simulation I

- One community with background
- $n = 1000$
- $n_1 = 100, 200, 300$
- $p_{12} = 0.05, p_{22} = 0.05$
- $p_{11} = 0.1, 0.15, 0.2$

# Results of simulation I



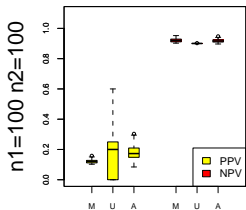
# Simulation II

- Two communities plus background
- $n = 1000$
- $n_1 = 100, 300, n_2 = 100, 300$
- $p_{12} = p_{23} = p_{13} = p_{33} = 0.05$
- $p_{11} = 0.1, 0.15, 0.2$
- $p_{22} = 0.08, 0.12, 0.16$

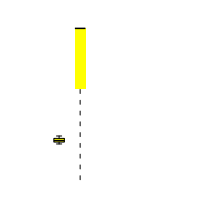
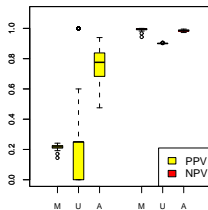


# Results for simulation II

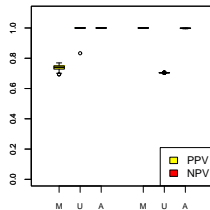
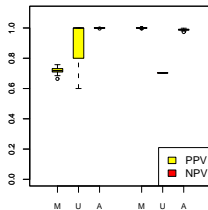
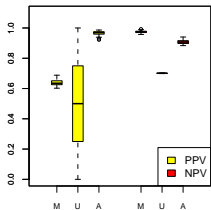
p11=0.1 p22=0.08



p11=0.15 p22=0.12



n1=300 n2=300



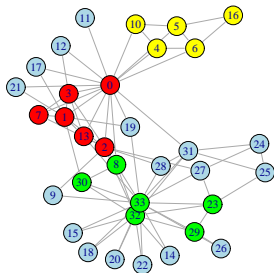
- ✓ Review of community detection
- ✓ Community extraction
- ✓ Simulation study
- Real data analysis
-

# Karate club network

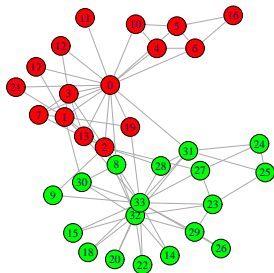
- Friendships between 34 members of a karate club (Zachary, 1977).
- This club has subsequently split into two parts following a disagreement between an instructor (node 0) and an administrator (node 33).

# Karate club network

Community extraction



Modularity



# Political books network

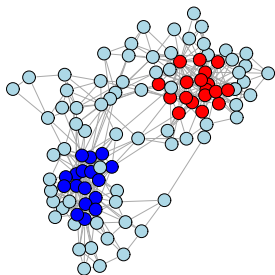
Links in the political books network (Newman, 2006) represent pairs of books frequently bought together on amazon.com.

Blue: liberal

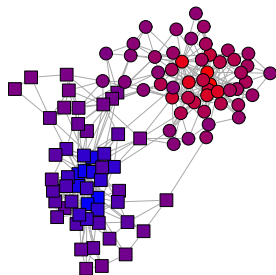
Red: conservative

# Political books network

Community extraction



Modularity



# School friendship network

The school friendship network is compiled from the National Longitudinal Study of Adolescent Health (AddHealth).

Grade 7: red

Grade 8: blue

Grade 9: green

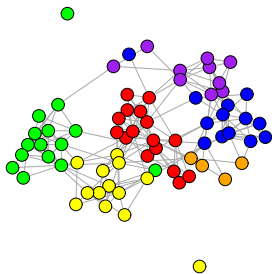
Grade 10: yellow

Grade 11: purple

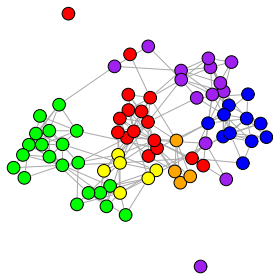
Grade 12: orange

# School friendship network

Grades



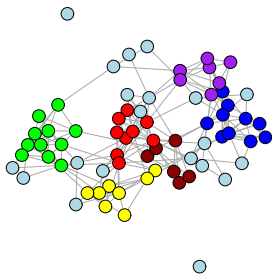
Modularity with 6 communities



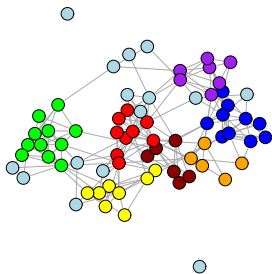


# School friendship network

Extracting 6 communities



Extracting 7 communities



- ✓ Review of community detection
- ✓ Community extraction
- ✓ Simulation study
- ✓ Real data analysis
  - **Asymptotic consistency**
  - Future work

# Block models

One of the simplest random graph models for communities

- Each node is assigned to a block independently of other nodes, with probability  $p_k$  for block  $k$ ,  $\sum_{k=1}^K p_k = 1$ .
- Given that node  $i$  belongs to block  $a$  and node  $j$  belongs to block  $b$ ,  $P[A_{ij} = 1] = p_{ab}$ , and all edges are independent.
- Parametrized as  $P_n = \frac{1}{n} P$ , where  $p_{ab} = P_n[A_{ij} = 1] \rightarrow 0$ .
- Expected node degree  $\bar{d}_n = n \sum_{k=1}^K p_k^2$
- Can stipulate background: assume  $p_{aK} < p_{bb}$  for all  $a = 1, \dots, K$ , and all  $b = 1, \dots, K - 1$ .

# Asymptotic consistency result

- For simplicity, assume one community and background ( $K = 2$  with parameters  $p_{11}, p_{12}, p_{22}$  ).
- Let  $\mathbf{c}$  be the true labels,  $\hat{\mathbf{c}}^{(n)}$  the estimated labels.

## Theorem

For any  $0 < \epsilon < 1$ , if  $p_{11} > p_{12}$ ,  $p_{11} > p_{22}$  and  $p_{11} + p_{22} > 2p_{12}$ ,  $\frac{n}{\log n} \rightarrow \infty$ , the maximizer  $\hat{\mathbf{c}}^{(n)}$  of both *unadjusted* and *adjusted* criteria satisfies

$$P[\hat{\mathbf{c}}^{(n)} = \mathbf{c}] \rightarrow 1 \quad \text{as } n \rightarrow \infty.$$

- Holds for  $p_{12} = p_{22} = p < p_{11}$
- Proof: apply Bickel and Chen (PNAS, 2009)

# Bickel & Chen consistency framework

- Assume a block model with known  $K$
- Given a proposed label assignment  $\mathbf{s}$ , true labels  $\mathbf{c}$ , let  $R$  be the **confusion matrix** with

$$R_{ab}(\mathbf{s}, \mathbf{c}) = \frac{1}{n} \sum_{i=1}^n I(s_i = a, c_i = b) .$$

- Many criteria, including ours, can be written as a function of the confusion matrix.
- **Key condition**: the population version of the criterion is maximized by the “correct” confusion matrix  $diag(1, \dots, k)$ .

- Eigenvalue method
- Determining the number of communities
- Adjusted criterion

$$W_a(S) = (|S| \cdot |S^c|) \left( \frac{O(S)}{|S|^2} - \frac{B(S)}{|S| \cdot |S^c|} \right)$$