

# Asymptotic results for the stochastic block model

Jan van Waaij

In collaboration with

Bas Kleijn



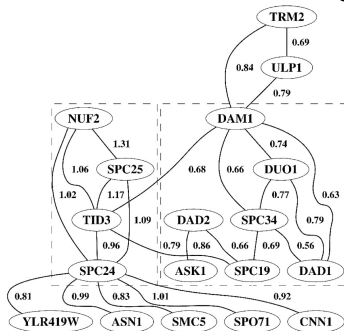
<https://github.com/Jan-van-Waaij/Berlin2018>

# Community detection on networks

- ▶ More present than ever... social networks, protein-protein networks...
- ▶ A lot of attention... Abbe (2017), "Community detection and stochastic block models: recent developments"

## Example

Protein-protein networks. Protein's interact. Goal: find communities of interacting protein's, Chen and Yuan (2006).



# The stochastic block model

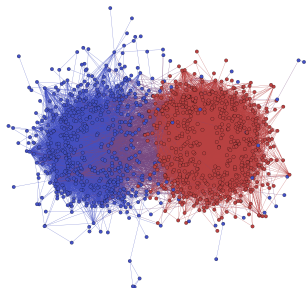
Consider a graph of  $2n$  nodes, partitioned in two sets of size  $n$ .

$$\Pr(\text{edge between } i \text{ and } j) = \begin{cases} p_n & i \text{ and } j \text{ same partition,} \\ q_n & i \text{ and } j \text{ different partition.} \end{cases}$$

**Parameter space** Set of labels  $\Theta \ni \theta = (\theta_1, \dots, \theta_{2n})$ ,  
 $\theta_i \in \{0, 1\}$ ,  $\sum_i \theta_i = n$ .

$\theta$  and  $\neg\theta = (1 - \theta_1, \dots, 1 - \theta_{2n})$  give rise to the same partition.

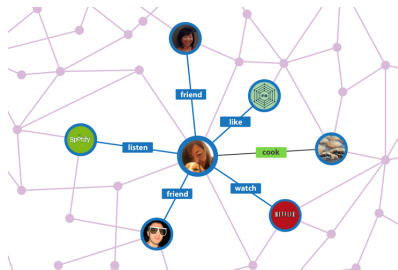
Define  $\Theta = \Theta / \sim$ ,  $\theta \sim \eta :\Leftrightarrow \theta = \eta$  or  $\theta = \neg\eta$ .



# Statistical problem

- ▶ Given a graph of  $2n$  nodes and its random edges, can we determine the communities?
- ▶ Of course impossible when  $p_n = q_n$ .
- ▶ When  $p_n - q_n > c$  is constant, then it's easy.
- ▶ E.g. Facebook, giant graph  $> 2$  billion, average number of friends  $155 \approx 7.19 * \log(2'271'000'000)$ ,

$$\mathbb{P}(\text{probability of being friends}) \approx \frac{7.19 * \log n}{n}.$$



# Exact recovery and detection

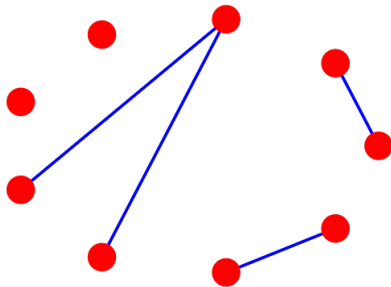
## Definition

An estimator  $\hat{\theta}$  exactly recovers  $\theta_0$  when  $\mathbb{P}_{\theta_0}(\hat{\theta} = \theta_0) \rightarrow 1$ .

## Definition

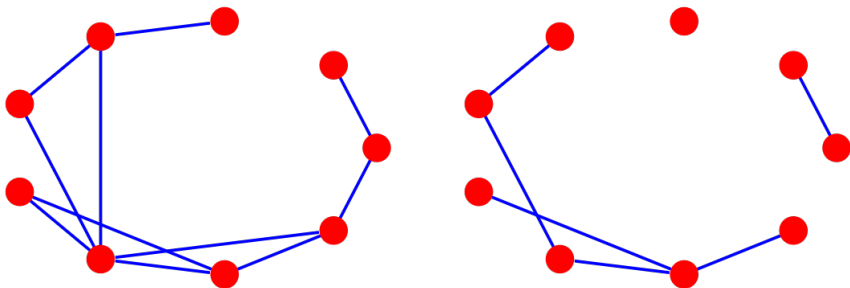
An estimator  $\hat{\theta}$  detects  $\theta_0$  when

$$\mathbb{P}(\text{fraction of mismatched labels} \rightarrow 0) = 1.$$



## Theorem

- ▶ If  $p_n = \frac{a \log n}{n}$ ,  $q_n = \frac{b \log n}{n}$ , then the graph is connected with probability converging to one, if and only if  $a, b > 0$  and  $a + b > 1$ .
- ▶ If  $p_n = a_n/n$ ,  $q_n = b_n/n$ , then the graph has a connected component of  $\mathcal{O}(n)$  with probability converging to one, if and only if  $a, b > 0$  and  $a + b > 1$ .



## Different modes of estimation: exact recovery

$$\mathbb{P}_{\theta_0}(\hat{\theta} = \theta_0) \rightarrow 1$$

- ▶ Only possible if the graph is connected.
- ▶ When  $0 < c < a_n, b_n < C$  and

$$p_n = \mathbb{P}(\text{edge between vertices of the same class}) = \frac{a_n \log n}{n}$$

$$q_n = \mathbb{P}(\text{edge between vertices of different class}) = \frac{b_n \log n}{n}$$

Mossel, Neeman, and Sly (2016): Exact recovery is possible if and only if

$$(a_n + b_n - 2\sqrt{a_n b_n} - 1) \log n + \frac{1}{2} \log \log n \rightarrow \infty.$$

# Different modes of estimation: detection

$\mathbb{P}(\text{fraction of mismatched labels goes to zero}) \rightarrow 1.$

► Only possible if there is a giant component.



$$p_n = \mathbb{P}(\text{edge between vertices of the same class}) = \frac{a_n}{n}$$

$$q_n = \mathbb{P}(\text{edge between vertices of different class}) = \frac{b_n}{n}$$

Mossel, Neeman, and Sly (2016): Exact recovery is possible if and only if

$$\frac{n(a_n - b_n)^2}{a_n + b_n} \rightarrow \infty.$$



# Bayesian setup

- ▶ Recall, parameter space is set of labels  
 $\Theta \ni \theta = (\theta_1, \dots, \theta_{2n})$ ,  $\theta_i \in \{0, 1\}$ ,  $\sum_i \theta_i = n$ . Identify  $\theta$  and  $\neg\theta = (1 - \theta_1, \dots, 1 - \theta_{2n})$ ,  $\Theta = \Theta / \sim$ .
- ▶  $\Theta$  has  $\binom{2n}{n}$  elements,  $\#\Theta = \frac{1}{2} \binom{2n}{n}$ .
- ▶ Uniform prior,  $\pi(\theta) = \frac{1}{\frac{1}{2} \binom{2n}{n}}$ .
- ▶ Posterior is given by

$$\pi(\theta \mid X) = \frac{p_\theta(X) \pi(\theta)}{\sum_{\eta \in \Theta} p_\eta(X) \pi(\eta)} = \frac{p_\theta(X)}{\sum_{\eta \in \Theta} p_\eta(X)}$$

- ▶ Maximum a-posteriori (MAP) estimator same as maximum likelihood estimator (MLE).

# Goals

- ▶ Establish posterior consistency.
- ▶ Confidence sets.

# Posterior consistency

## Theorem

If  $(p_n = \frac{a_n \log n}{n})$  and  $(q_n = \frac{b_n \log n}{n})$  are such that

$$(a_n + b_n - 2\sqrt{a_n b_n} - 2) \log n \rightarrow \infty, \text{ then } \pi(\theta_0 | X) \xrightarrow{P_{\theta_0}} 1.$$

## Idea of the proof.

1. Choose  $\theta_0 \in \boldsymbol{\theta}_0$ . Let  $Z_\theta = \{i : \theta_i = 0\}$ . Let  $V_k = \{\theta : \#(Z_{\theta_0} \setminus Z_\theta) = k\}$ .  $V_k$  has  $\binom{n}{k}^2$  elements.
2. For every  $\theta \in V_k$  there is a test of power  $a_{n,k}$ ,

$$\mathbb{P}_{\theta_0} \phi(X) + \mathbb{P}_\theta (1 - \phi(X)) \leq a_{n,k}.$$

3.  $\mathbb{P}_{\theta_0} \Pi(\boldsymbol{\theta} \neq \boldsymbol{\theta}_0 | X) \leq \sum_{k=1}^{\lfloor n/2 \rfloor} \binom{n}{k}^2 a_{k,n} = \mathbf{o}(1).$

# Credible sets turn out to be confidence sets

## Definition

A  $1 - \alpha$  credible set, is any measurable set  $D(X) \subset \Theta$  with  $\Pi(D(X) \mid X) \geq 1 - \alpha$ .

## Definition

A  $1 - \alpha$  confidence set is any measurable set  $C(X) \subset \Theta$  with  $P_{\theta_0}(\theta_0 \in C(X)) \geq 1 - \alpha$ .

## Theorem

*When exact recovery holds,  $1 - \alpha$ -credible sets, are  $1 - 3\alpha$  confidence sets.*

## Proof.

See blackboard.

