Community detection in the asymmetric stochastic block model

YEP XIV, Eindhoven

September 7, 2017

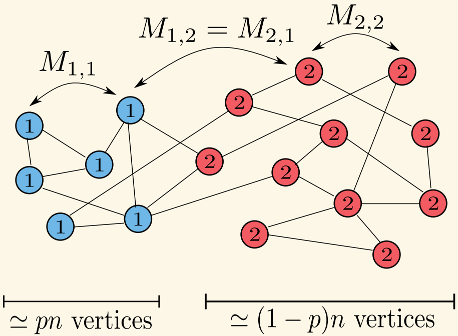
Francesco Caltagirone, Marc Lelarge & Léo Miolane

Community detection

The Stochastic Block Model (SBM)

$\mathbf{G}$ is generated as follows:

- n vertices: $1, \ldots, n$.
- Each vertex i has a label $X_i \in \{1, 2\}$ where $(X_k)_k \overset{\text{i.i.d.}}{\sim} 1 + \mathrm{Ber}(1-p)$.
- Two vertices i, j are then connected with probability M_{X_i, X_j}.

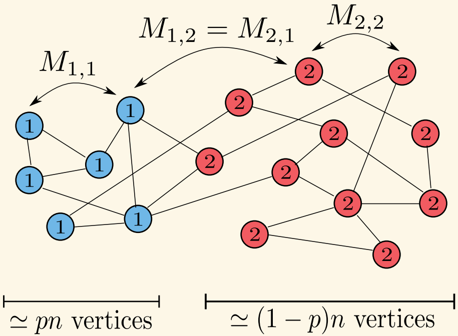


Community detection

The Stochastic Block Model (SBM)

$\mathbf{G}$ is generated as follows:

- n vertices: $1, \ldots, n$.
- Each vertex i has a label $X_i \in \{1, 2\}$ where $(X_k)_k \overset{\text{i.i.d.}}{\sim} 1 + \text{Ber}(1-p)$.
- Two vertices i, j are then connected with probability M_{X_i, X_j}.



- **Goal:** given the graph $\mathbf{G}$ we want to recover the labels $\mathbf{X}$.
- **Weak Reconstruction:** Estimate $\mathbf{X}$ better than a “random guess”.

Setting

- The connectivity matrix will be of the form:

$$\mathbf{M} = \frac{d}{n} \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

$a, c > b$ and $pa + (1-p)b = pb + (1-p)c = 1$.

Setting

- The connectivity matrix will be of the form:

$$\mathbf{M} = \frac{d}{n} \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

  $a, c > b$ and $pa + (1-p)b = pb + (1-p)c = 1$.

- **Important quantity:** the signal-to-noise ratio

$$\lambda = d(1-b)^2$$

Setting

- The connectivity matrix will be of the form:

$$\mathbf{M} = \frac{d}{n}\begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

$a, c > b$ and $pa + (1-p)b = pb + (1-p)c = 1$.

- **Important quantity:** the signal-to-noise ratio

$$\lambda = d(1-b)^2$$

Mossel et al., 2015, Massoulié, 2014, Mossel et al., 2013

In the case of two symmetric communities ($p = 1/2$), when $d > 1$ is fixed and $n \to \infty$,

- if $\lambda \leq 1$ it is not possible to recover the partition $\mathbf{X}$ better that a "random guess".
- if $\lambda > 1$ it is possible to recover the labels better than chance.

Asymmetric communities

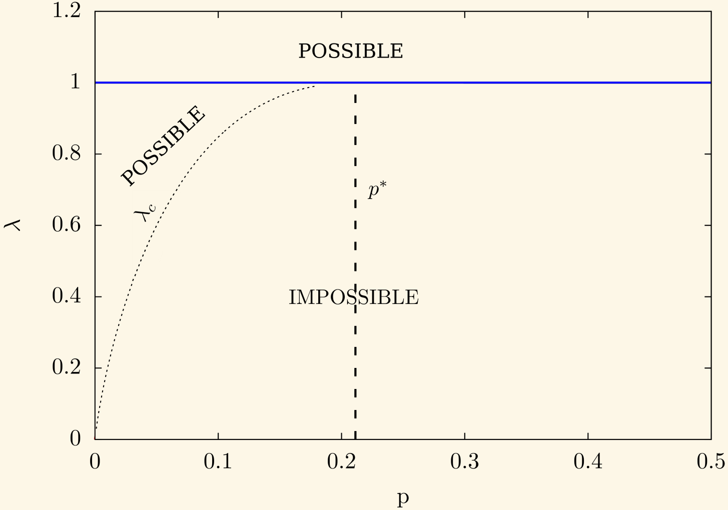
The main picture

- Does this phase transition at $\lambda = 1$ still hold when $p < 1/2$?
- The physicist's conjecture for the large degree limit ($d \to \infty$):

Asymmetric communities

The main picture

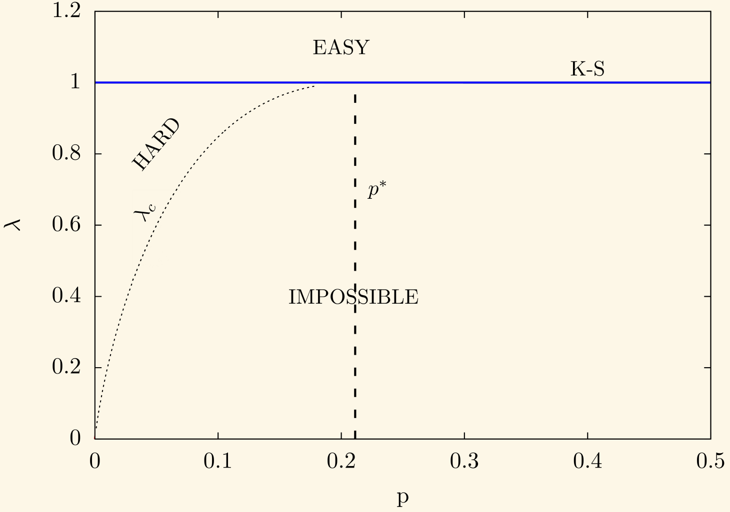
- Does this phase transition at $\lambda = 1$ still hold when $p < 1/2$?
- The physicist's conjecture for the large degree limit ($d \to \infty$):



Asymmetric communities

The main picture

- Does this phase transition at $\lambda = 1$ still hold when $p < 1/2$?
- The physicist's conjecture for the large degree limit ($d \to \infty$):

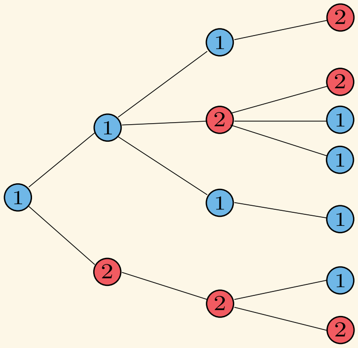


Part 1.

Local weak convergence of the SBM

Local weak convergence of the SBM

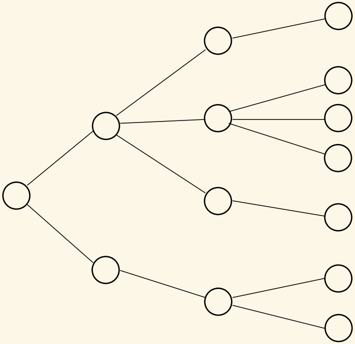
The Stochastic Block Model converges locally weakly to a "Labeled Poison Galton-Watson tree".



- Offspring distribution: Pois(d).
- The labels "propagate" from the root according to the transition matrix $\begin{pmatrix} pa & (1-p)b \\ pb & (1-p)c \end{pmatrix}$

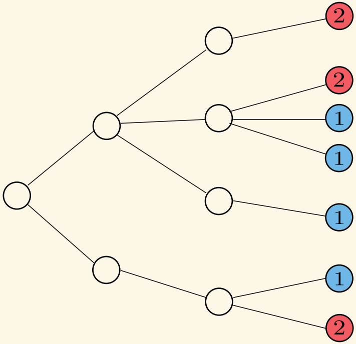
Reconstruction on trees

- **An issue:** the Galton-Watson tree, without the labels, does not give any information about the label of the root!



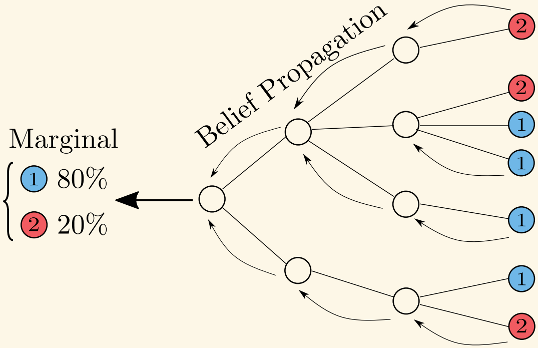
Reconstruction on trees

- **An issue:** the Galton-Watson tree, without the labels, does not give any information about the label of the root!
- We thus suppose that the labels at depth r are revealed. Can we infer the label of the root as $r \to \infty$?



Reconstruction on trees

- **An issue:** the Galton-Watson tree, without the labels, does not give any information about the label of the root!
- We thus suppose that the labels at depth r are revealed. Can we infer the label of the root as $r \to \infty$?



- Belief-Propagation gives the marginal distribution of the root given $\mathbf{G}$ and the labels at depth r.

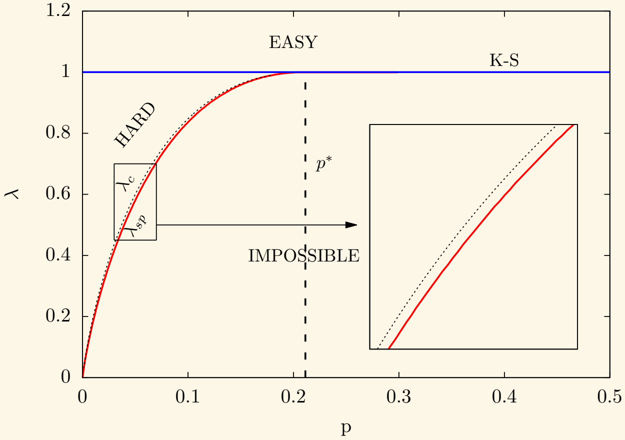
An impossibility result

- Studying the "BP recursion" one see that when $\lambda < \lambda_{\mathrm{sp}}$, the marginal does not contain any information about the true label.

An impossibility result

- Studying the "BP recursion" one see that when $\lambda < \lambda_{\rm sp}$, the marginal does not contain any information about the true label.

We thus obtain the "impossibility curve" $\lambda_{sp}(p)$ below:



Part 2.

Low-rank matrix estimation

Low-rank matrix estimation

From Bernoulli to Gaussian noise

$$A_{i,j} \sim \mathsf{Ber}\left(\frac{d}{n} + \frac{\sqrt{d}\sqrt{\lambda}}{n}\tilde{X}_i\tilde{X}_j\right) \tag{1}$$

where $$\tilde{X}_k = \begin{cases} \sqrt{(1-p)/p} & \text{if } X_k = 1 \\ -\sqrt{p/(1-p)} & \text{if } X_k = 2 \end{cases}.$$

[1] Yash Deshpande and Emmanuel Abbe (2016). "Asymptotic mutual information for the balanced binary stochastic block model". In: *Information and Inference*, iaw017.

Low-rank matrix estimation

From Bernoulli to Gaussian noise

$$A_{i,j} \sim \mathsf{Ber}\left(\frac{d}{n} + \frac{\sqrt{d}\sqrt{\lambda}}{n}\tilde{X}_i\tilde{X}_j\right) \tag{1}$$

where $$\tilde{X}_k = \begin{cases} \sqrt{(1-p)/p} & \text{if } X_k = 1 \\ -\sqrt{p/(1-p)} & \text{if } X_k = 2 \end{cases}.$$

The Bernoulli noise model (1) is "equivalent" to the Gaussian noise model (when $n, d \to \infty$)[1]:

$$A'_{i,j} = \frac{d}{n} + \frac{\sqrt{d}\sqrt{\lambda}}{n}\tilde{X}_i\tilde{X}_j + \sqrt{\frac{d}{n}}Z_{i,j} \tag{2}$$

where $Z_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$,

[1] Yash Deshpande and Emmanuel Abbe (2016). "Asymptotic mutual information for the balanced binary stochastic block model". In: *Information and Inference*, iaw017.

Low-rank matrix estimation

From Bernoulli to Gaussian noise

$$A_{i,j} \sim \text{Ber}\left(\frac{d}{n} + \frac{\sqrt{d}\sqrt{\lambda}}{n}\tilde{X}_i\tilde{X}_j\right) \tag{1}$$

where $$\tilde{X}_k = \begin{cases} \sqrt{(1-p)/p} & \text{if } X_k = 1 \\ -\sqrt{p/(1-p)} & \text{if } X_k = 2 \end{cases}.$$

The Bernoulli noise model (1) is "equivalent" to the Gaussian noise model (when $n, d \to \infty$)[1]:

$$A'_{i,j} = \frac{d}{n} + \frac{\sqrt{d}\sqrt{\lambda}}{n}\tilde{X}_i\tilde{X}_j + \sqrt{\frac{d}{n}}Z_{i,j} \tag{2}$$

where $Z_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, and thus to

$$Y_{i,j} = \sqrt{\frac{\lambda}{n}}\tilde{X}_i\tilde{X}_j + Z_{i,j}$$

[1] Yash Deshpande and Emmanuel Abbe (2016). "Asymptotic mutual information for the balanced binary stochastic block model". In: *Information and Inference*, iaw017.

Low-rank matrix estimation

The new statistical model

"Spiked Wigner" model

$$\underbrace{\mathbf{Y}}_{\text{observations}} = \sqrt{\frac{\lambda}{n}}\,\underbrace{\mathbf{X}\mathbf{X}^\intercal}_{\text{signal}} + \underbrace{\mathbf{Z}}_{\text{noise}}$$

- $\mathbf{X}$: vector of dimension n with entries $X_i \overset{\text{i.i.d.}}{\sim} P_0$. $\mathbb{E}X_1 = 0$, $\mathbb{E}X_1^2 = 1$.
- $Z_{i,j} = Z_{j,i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$.
- λ: signal-to-noise ratio.
- λ and P_0 are known by the statistician.

Goal: recover the low-rank matrix $\mathbf{X}\mathbf{X}^\intercal$ from $\mathbf{Y}$.

Principal component analysis (PCA)

B.B.P. phase transition

Spectral estimator:
Estimate $\mathbf{X}$ using the eigenvector $\hat{\mathbf{x}}_n$ associated with the largest eigenvalue μ_n of $\mathbf{Y}/\sqrt{n}$.

Principal component analysis (PCA)

Spectral estimator:
Estimate $\mathbf{X}$ using the eigenvector $\hat{\mathbf{x}}_n$ associated with the largest eigenvalue μ_n of $\mathbf{Y}/\sqrt{n}$.

B.B.P. phase transition

- if $\lambda \leq 1$ $\begin{cases} \mu_n & \xrightarrow[n\to\infty]{\text{a.s.}} 2 \\ \mathbf{X} \cdot \hat{\mathbf{x}}_n & \xrightarrow[n\to\infty]{\text{a.s.}} 0 \end{cases}$
- if $\lambda > 1$ $\begin{cases} \mu_n & \xrightarrow[n\to\infty]{\text{a.s.}} \sqrt{\lambda} + \frac{1}{\sqrt{\lambda}} > 2 \\ |\mathbf{X} \cdot \hat{\mathbf{x}}_n| & \xrightarrow[n\to\infty]{\text{a.s.}} \sqrt{1 - 1/\lambda} > 0 \end{cases}$

Baik et al., 2005; Benaych-Georges and Nadakuditi, 2011

Minimal Mean Square Error (MMSE)

Definition

$$\begin{aligned}
\mathrm{MMSE}_n &= \min_{\hat{\theta}} \frac{1}{n^2} \mathbb{E} \left\| \mathbf{X}\mathbf{X}^\intercal - \hat{\theta}(\mathbf{Y}) \right\|^2 \\
&= \frac{1}{n^2} \sum_{1 \le i,j \le n} \left(X_i X_j - \mathbb{E}[X_i X_j | \mathbf{Y}] \right)^2
\end{aligned}$$

Minimal Mean Square Error (MMSE)

Definition

$$
\begin{aligned}
\mathrm{MMSE}_n &= \min_{\hat{\theta}} \frac{1}{n^2} \mathbb{E} \left\| \mathbf{X}\mathbf{X}^\intercal - \hat{\theta}(\mathbf{Y}) \right\|^2 \\
&= \frac{1}{n^2} \sum_{1 \le i,j \le n} \left(X_i X_j - \mathbb{E}[X_i X_j | \mathbf{Y}] \right)^2
\end{aligned}
$$

We have to study the posterior distribution of the signal $\mathbf{X}$ given $\mathbf{Y}$!

Connection with statistical physics

A planted spin glass model

- Posterior distribution: $\mathbb{P}(\mathbf{X} = \mathbf{x} \,|\, \mathbf{Y}) = \dfrac{1}{Z_n} P_0(\mathbf{x}) e^{H_n(\mathbf{x})}$ where

Connection with statistical physics

A planted spin glass model

- Posterior distribution: $\mathbb{P}(\mathbf{X} = \mathbf{x} \,|\, \mathbf{Y}) = \frac{1}{Z_n} P_0(\mathbf{x}) e^{H_n(\mathbf{x})}$ where

$$H_n(\mathbf{x}) = \sum_{i<j} \sqrt{\frac{\lambda}{n}} Y_{i,j} x_i x_j - \frac{\lambda}{2n} x_i^2 x_j^2$$

Connection with statistical physics

A planted spin glass model

- Posterior distribution: $\mathbb{P}(\mathbf{X} = \mathbf{x} \,|\, \mathbf{Y}) = \frac{1}{Z_n} P_0(\mathbf{x}) e^{H_n(\mathbf{x})}$ where

$$H_n(\mathbf{x}) = \sum_{i<j} \sqrt{\frac{\lambda}{n}} Y_{i,j} x_i x_j - \frac{\lambda}{2n} x_i^2 x_j^2$$

- Free energy: $F_n = \frac{1}{n} \mathbb{E} \log Z_n$

Connection with statistical physics

A planted spin glass model

- Posterior distribution: $\mathbb{P}(\mathbf{X} = \mathbf{x} \,|\, \mathbf{Y}) = \frac{1}{Z_n} P_0(\mathbf{x}) e^{H_n(\mathbf{x})}$ where

$$H_n(\mathbf{x}) = \sum_{i<j} \sqrt{\frac{\lambda}{n}} Y_{i,j} x_i x_j - \frac{\lambda}{2n} x_i^2 x_j^2$$

- Free energy: $F_n = \frac{1}{n} \mathbb{E} \log Z_n$
- In physics $\frac{\partial}{\partial \lambda} F_n = \mathbb{E} \langle U(\mathbf{x}) \rangle$.

Connection with statistical physics

A planted spin glass model

- Posterior distribution: $\mathbb{P}(\mathbf{X} = \mathbf{x} \,|\, \mathbf{Y}) = \dfrac{1}{Z_n} P_0(\mathbf{x}) e^{H_n(\mathbf{x})}$ where

$$H_n(\mathbf{x}) = \sum_{i<j} \sqrt{\frac{\lambda}{n}} Y_{i,j} x_i x_j - \frac{\lambda}{2n} x_i^2 x_j^2$$

- Free energy: $F_n = \dfrac{1}{n} \mathbb{E} \log Z_n$
- In physics $\frac{\partial}{\partial \lambda} F_n = \mathbb{E}\langle U(\mathbf{x}) \rangle$.
- Here in statistics $F_n \simeq \frac{1}{n} I\big(\mathbf{X}; \mathbf{Y}\big)$ and

$$\frac{\partial}{\partial \lambda} F_n = \mathrm{MMSE}_n$$

Main result

Limiting formula for the MMSE

Theorem[2]

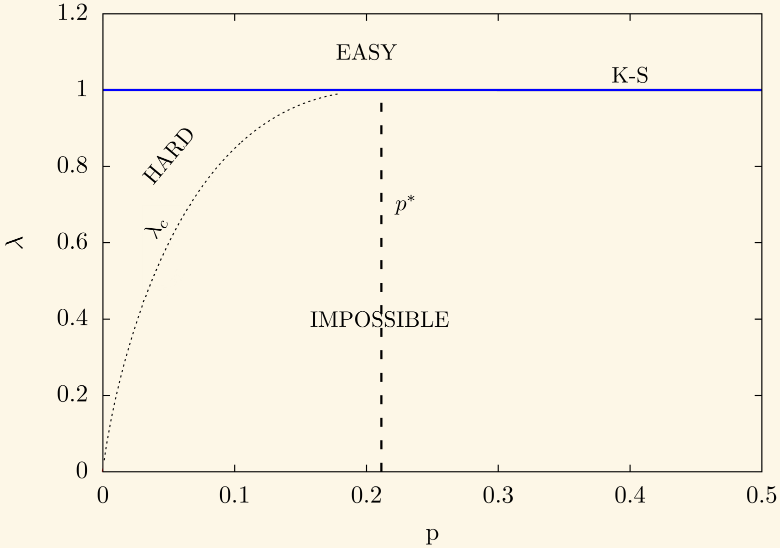
$$F_n \xrightarrow[n\to\infty]{} \max_{q\in[0,\mathbb{E}X^2]} \mathcal{F}(q)$$

$$\mathrm{MMSE}_n \xrightarrow[n\to\infty]{} \mathbb{E}_{P_0}[X^2]^2 - \ q^*(\lambda)^2$$

where

$$\mathcal{F} : q \geq 0 \mapsto \mathbb{E}_{\substack{X_0\sim P_0\\ Z_0\sim\mathcal{N}}}\left[\log\int_{x_0} dP_0(x_0) e^{\sqrt{\lambda q} Z_0 x_0 + \lambda q X_0 x_0 - \frac{\lambda q}{2} x_0^2}\right] - \frac{\lambda}{4} q^2$$

[2] Barbier et al., 2016, Lelarge and Miolane, 2016


EASY
K-S
HARD
λ_c
p^*
IMPOSSIBLE
λ
p
1.2
1
0.8
0.6
0.4
0.2
0
0.1
0.2
0.3
0.4
0.5

Thank you for your attention.

Any questions?

References I

- Baik, Jinho, Gérard Ben Arous, and Sandrine Péché (2005). "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices". In: *Annals of Probability*, pp. 1643–1697.
- Barbier, Jean et al. (2016). "Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula". In: *Advances in Neural Information Processing Systems*, pp. 424–432.
- Benaych-Georges, Florent and Raj Rao Nadakuditi (2011). "The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices". In: *Advances in Mathematics* 227.1, pp. 494–521.
- Deshpande, Yash and Emmanuel Abbe (2016). "Asymptotic mutual information for the balanced binary stochastic block model". In: *Information and Inference*, iaw017.
- Lelarge, Marc and Léo Miolane (2016). "Fundamental limits of symmetric low-rank matrix estimation". In: *arXiv preprint arXiv:1611.03888*.
- Massoulié, Laurent (2014). "Community detection thresholds and the weak Ramanujan property". In: *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*. ACM, pp. 694–703.

References II

- Mossel, Elchanan, Joe Neeman, and Allan Sly (2013). "A proof of the block model threshold conjecture". In: *arXiv preprint arXiv:1311.4115*.
- – (2015). "Reconstruction and estimation in the planted partition model". In: *Probability Theory and Related Fields* 162.3-4, pp. 431–461.