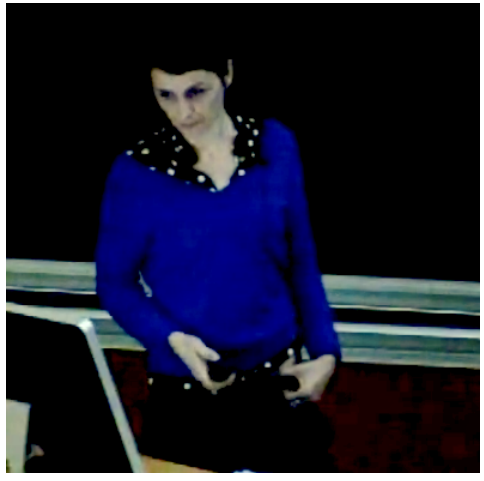
Robustness of Community Detection to Random Geometric Perturbations



Sandrine Péché

Univ. Paris Diderot

Vianney Perchet

ENSAE Paris
Criteo AI Lab

December 2020

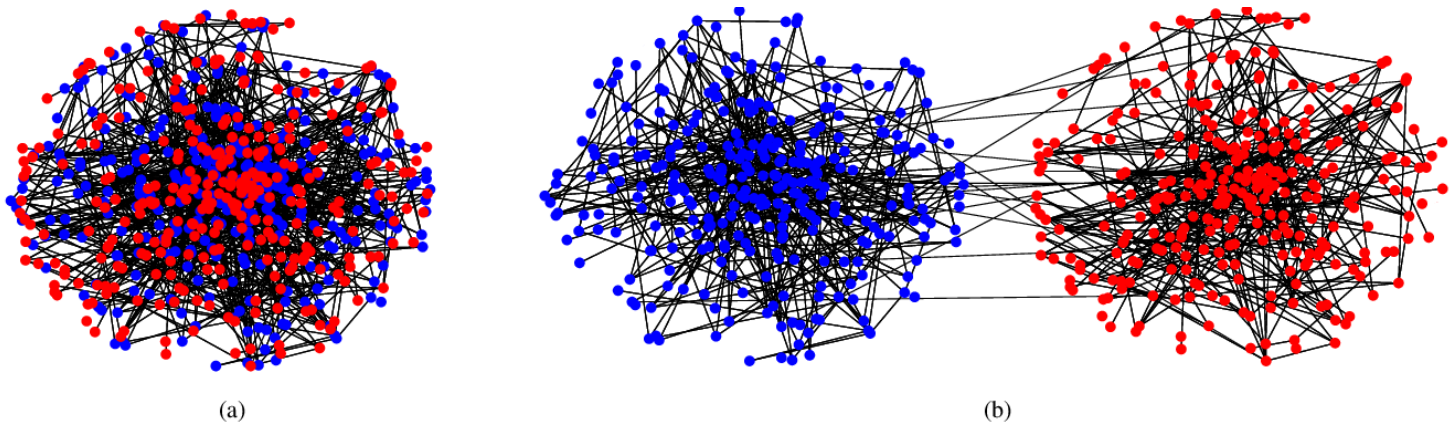
Stochastic Block Model

Social network Users are **nodes** connected by **edges**

Communities **2** (for simplicity) communities

- "**high**" proba. connexion **within** communities
- "**low**" proba. connexion **across** communities

Objectives Based on connexions graph, **recover** communities



Example of a SBM. from Abbe et al [2016]

Spectral Methods

Probas $p \in [0,1]$ within comm. and q across (with $p > q$)

Balance N users total, $N/2$ in each comm.

Adjacency $\tilde{A}$ of expectation $\mathrm{A} = \begin{pmatrix} P & Q \\ Q & P \end{pmatrix}$ where

- P is $N/2 \times N/2$ all-coordinates equal p
- Same thing for Q

Spectrum A is of rank 2 and

- $\lambda_1 = \frac{N(p+q)}{2}$ associated to $v_1 = (1, \ldots, 1)/\sqrt{N}$
- $\lambda_2 = \frac{N(p-q)}{2}$ ass. $v_2 = (+1, .., +1, -1, .., -1)/\sqrt{N}$

Spectral Methods: $\tilde{v}_2$ (of $\tilde{A}$) close enough to v_2 for recovery

A motivation for perturbations

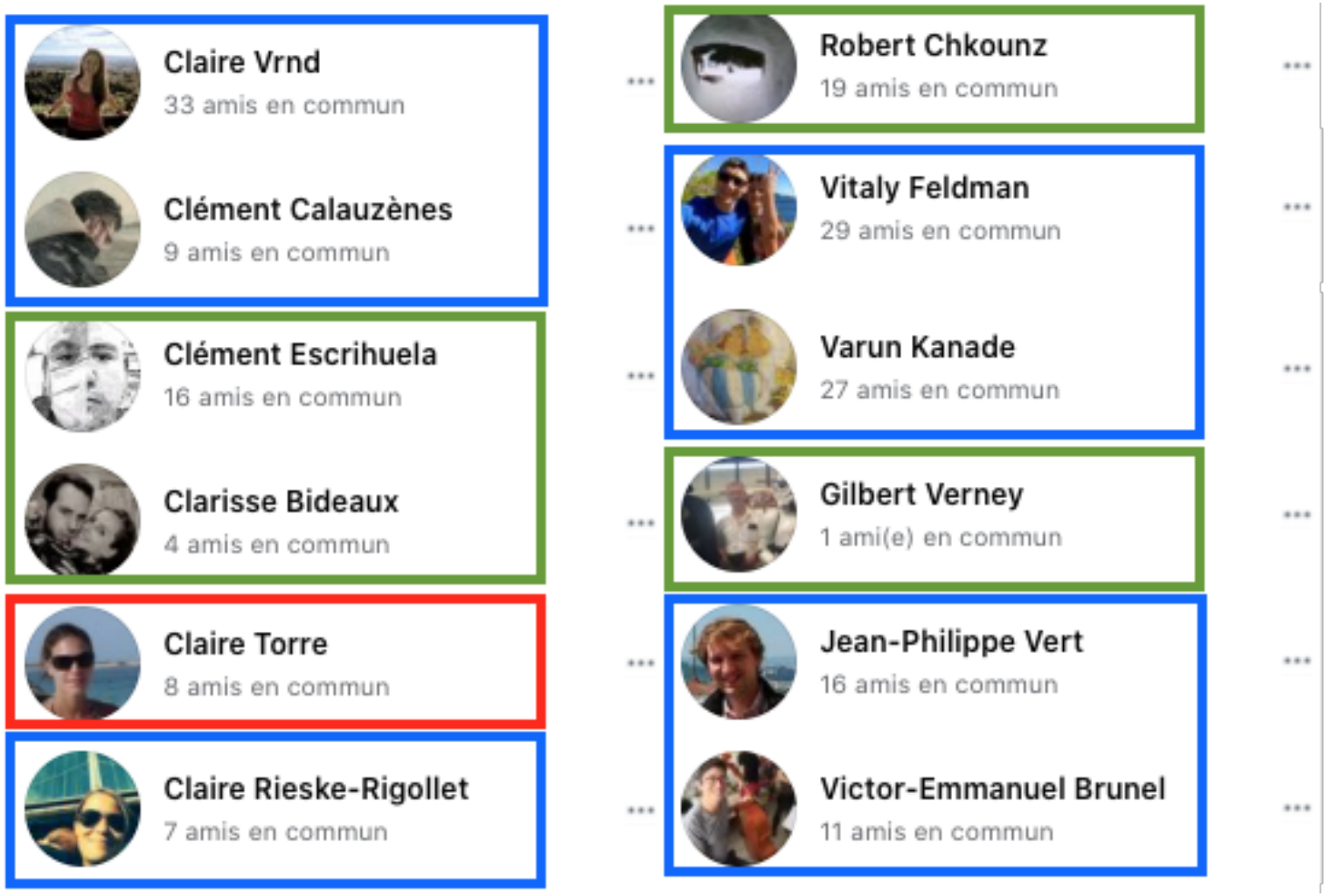
- Social networks, 2 types of connections between users
  1. **exogenous** similarity. Belong to the same community
  2. **endogenous** similarity. Crossed paths at some point

A motivation for perturbations

- Social networks, 2 types of connections between users
  1. **exogenous** similarity. Belong to the same community
  2. **endogenous** similarity. Crossed paths at some point

- Look at my own interesting facebook life

A motivation for perturbations



A pic of my list of friends from academia, Handball, Random

A motivation for perturbations

- Social networks, 2 types of connections between users
  1. **exogenous** similarity. Belong to the same community
  2. **endogenous** similarity. Crossed paths at some point

- Look at my own interesting facebook life
  1. I belong to the academic/research in ML **community**
  2. I belong to the handball **community**
  3. I randomly met people

A motivation for perturbations

- Social networks, 2 types of connections between users
  1. **exogenous** similarity. Belong to the same community
  2. **endogenous** similarity. Crossed paths at some point

- Look at my own interesting facebook life
  1. I belong to the academic/research in ML **community**
  2. I belong to the handball **community**
  3. I randomly met people

Spectral methods recover comm. with (**random**) perturbations ?

"agnostically" from the perturbation model.

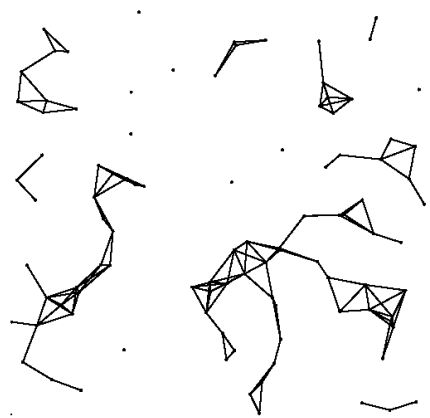
Perturbations via Geometric Graphs

Model $X_i \in \mathbb{R}^d$, connection with proba $K(X_i, X_j) \in [0, 1]$

- $X_i \sim \mathcal{N}(0, I_d)$ i.i.d.
- $d = 2$ and $K(X_i, X_j) = \exp(-\gamma \|X_i - X_j\|^2)$
- **not observed**

Ideas Users are connected if **close enough**

Generalize easily (heavier computations)



Example of a Geometric Graph. from Mitsche et al [2017]

Model and Objectives

Model SBM(p, q) perturbed by a geom $K(\cdot) = \exp(-\gamma \| \cdot \|^2)$

- $X_i \sim X_j$ w.p. $p + \kappa K(X_i, X_j)$ within
- $X_i \sim X_j$ w.p. $q + \kappa K(X_i, X_j)$ across

Objectives **Recovery**: find $v \in \mathbb{R}^N$ with $\|v\| = 1$ s.t.

- Exact $|\langle v, v_2 \rangle| = 1$, w.p. $\to 1$, $v \in \{\pm 1/\sqrt{N}\}^N$
- ε-Weak $|\langle v, v_2 \rangle| \geq \varepsilon$, w.p $\to 1$, $v \in \{\pm 1/\sqrt{N}\}^N$
- ε-Soft $|\langle v, v_2 \rangle| \geq \varepsilon$, w.p $\to 1$

Param. Interesting regimes

- $p \sim q \sim p - q$ and $p \sim \frac{f(N)}{N}$, with $f(N) \gg \log(N)$
- $\frac{1}{\gamma} \sim p$ (similar degree in SBM and geom)

Our results and techniques

Techniques **Spectral** methods

1. Compute spectrum of geom. graph
2. Compare it with SBM spectrum
3. Conditions s.t. eigenvector $\tilde{v}_2$ correlated to v_2

Results Exact/Weal/Soft recovery possible

- Small perturbations: **exact** recovery
- λ_1 known, **weak** recovery if $\frac{p-q}{2} \geq \frac{2\kappa}{\gamma}(1+\varepsilon)$
- General case, **soft** recovery if $\frac{p+q}{2} \geq \frac{p-q}{2} + \frac{\kappa}{2\gamma}$

Spectral methods **robust**/agnostic to geometric perturbations

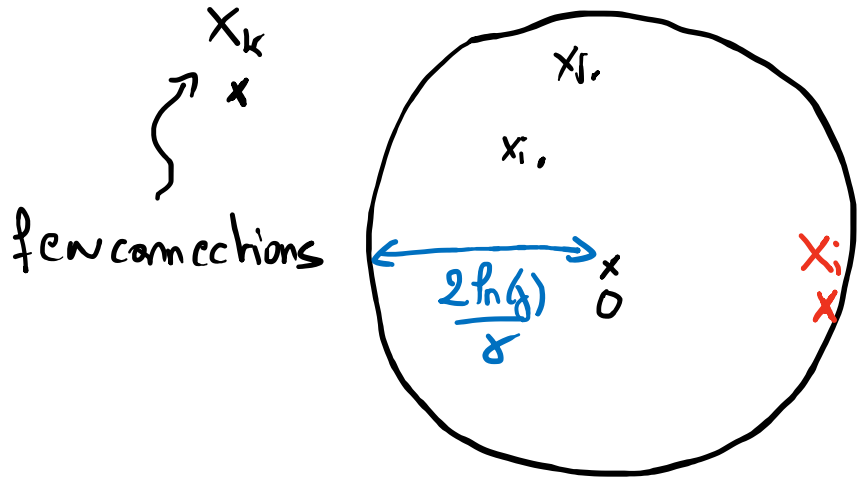
Spectrum of Geometric Graphs

- $\mathcal{G} = \left(\exp(-\gamma \|X_i - X_j\|^2) \right)_{i,j}$: expected (cond. to X_i) adjacency

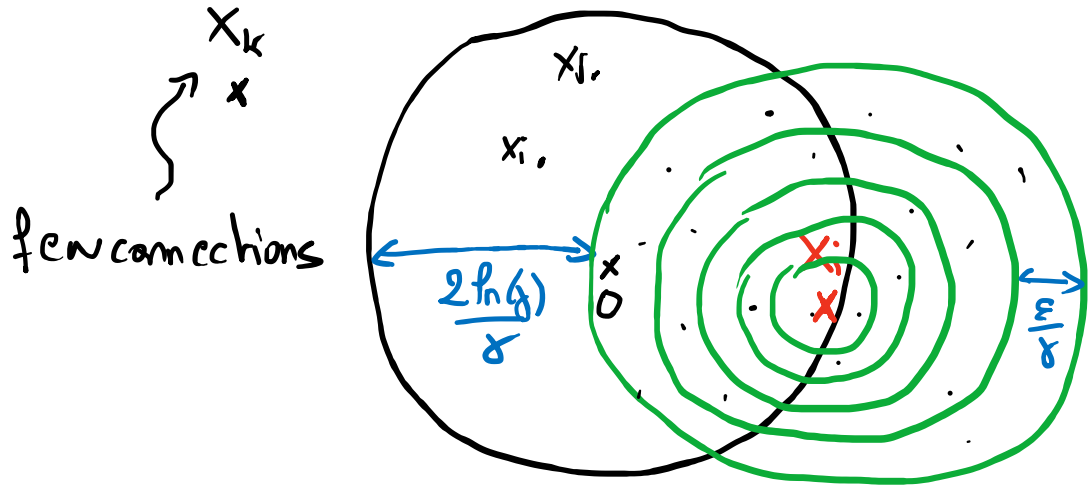
> Th. $\frac{N}{2\gamma} \leq \text{spec\_rad}(\mathcal{G}) \leq \frac{N}{2\gamma}(1 + o(1))$ with proba $\to 1$

- **Ideas** of proof.
  1. Slice $\mathbb{R}^d$ into shells
     - ⋆ X_i far from 0 have few connections. Few of them.
     - ⋆ Intermediate X_i have expected degrees of order $\frac{N}{2\gamma}$
     - ⋆ Prove this by slicing even more
  2. Consequence of concentration inequalities.
     - ⋆ Requires $\gamma \to \infty$ and $\gamma \frac{\log(N)}{N} \to 0$
  3. Conclude with Perron Frobenius

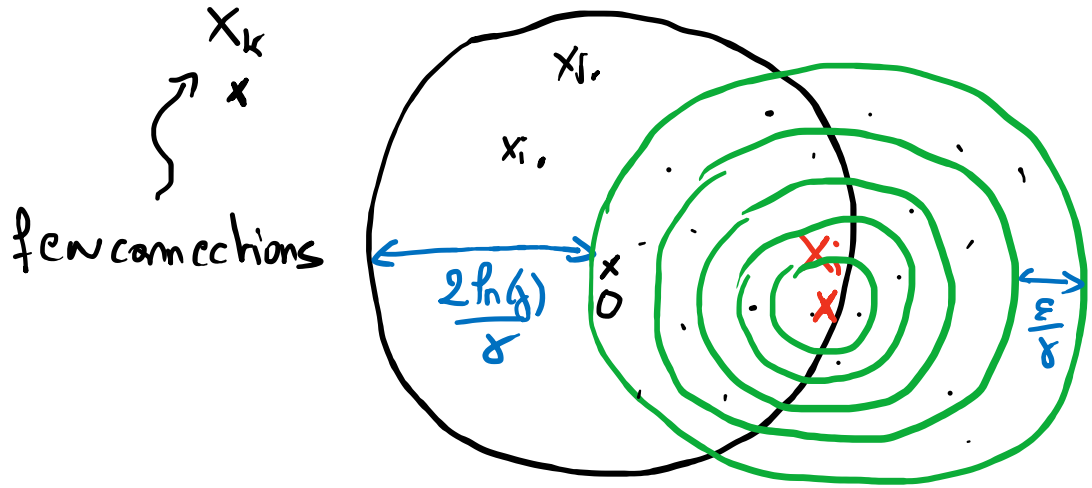
Let's try to be a bit more precise



Let's try to be a bit more precise

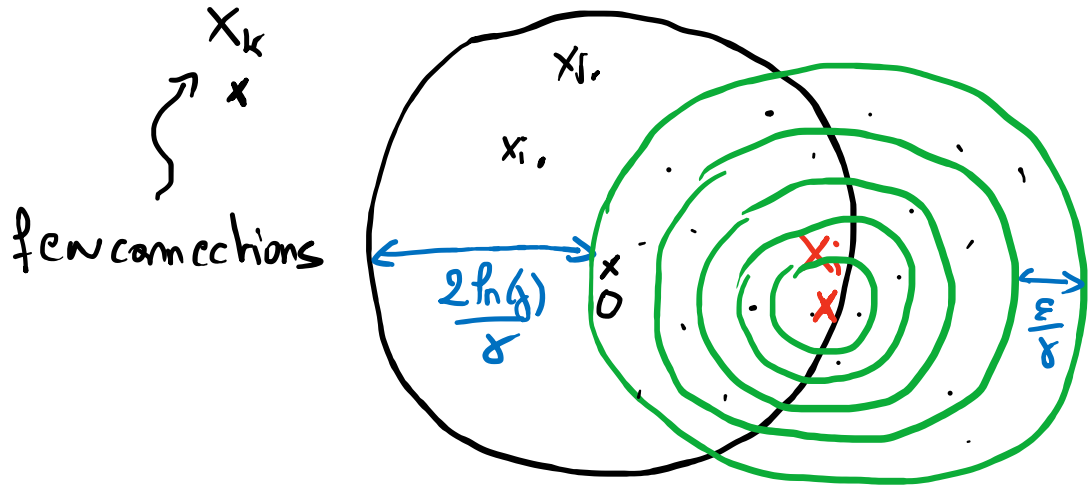


Let's try to be a bit more precise



$$e^{-\frac{\|x\|^2}{2}} \frac{N\varepsilon}{2\gamma}(1+o(1)) \leq E[n_k] \leq \frac{N\varepsilon}{2\gamma}$$

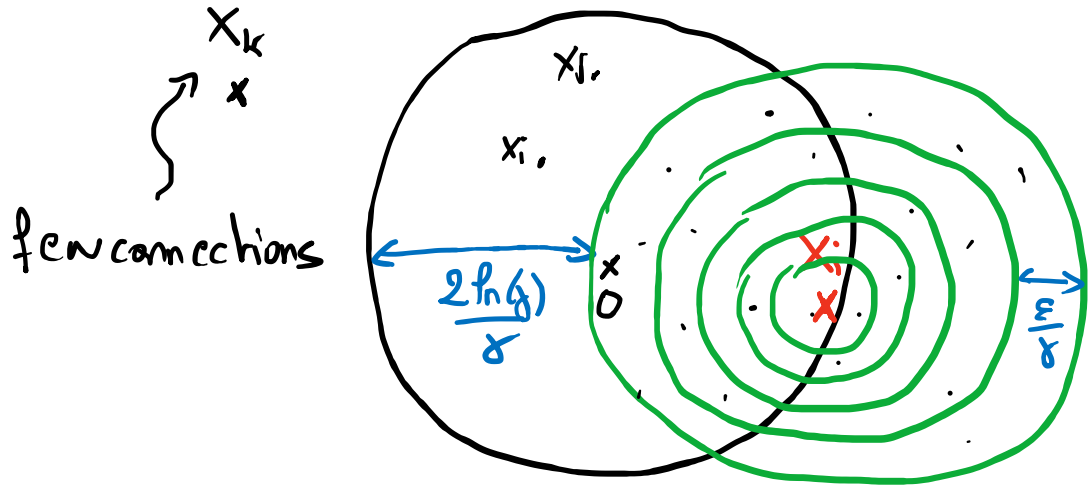
Let's try to be a bit more precise



$$e^{-\frac{\|x\|^2}{2}} \frac{N\varepsilon}{2\gamma}(1+o(1)) \leq E[n_k] \leq \frac{N\varepsilon}{2\gamma}$$

$$\Rightarrow \text{Degree} \approx \sum e^{-k\varepsilon} n_k \approx \frac{N}{2\gamma}$$

Let's try to be a bit more precise



$$e^{-\frac{\|x\|^2}{2}} \frac{N\varepsilon}{2\gamma}(1+o(1)) \leq E[n_k] \leq \frac{N\varepsilon}{2\gamma}$$

$$\Rightarrow \text{Degree} \approx \sum e^{-k\varepsilon} n_k \approx \frac{N}{2\gamma}$$

concentration inequalities

Separations and Recovery

Recovery depends on range of eigenvalues $\frac{N(p+q)}{2}$, $\frac{N(p-q)}{2}$ and $\frac{N\kappa}{2\gamma}$

Trivial $\frac{N(p-q)}{2} \gg \sqrt{\frac{N(p+q)}{2}} + \frac{N\kappa}{2\gamma}$ **exact** reco. (negligible noise)

Easy $\frac{N\kappa}{2\gamma} \ll \sqrt{\frac{N(p+q)}{2}}$ standard SBM (negligible perturbations)

Interesting $\frac{N\kappa}{2\gamma} \sim \frac{N(p-q)}{2} \gg \sqrt{\frac{N(p+q)}{2}}$

$p+q$ is known

Th: ε-**Weak** reco is possible if $\frac{N(p-q)}{2} \geq 4\frac{N\kappa}{2\gamma}(1+\varepsilon)$

Proof:

- **David-Kahan** $\sin(\theta)$ to $(\mathrm{Sbm} - \lambda_1 v_1 v_1^\top)$ and $(\widetilde{A} - \lambda_1 v_1 v_1^\top)$
  - $\mathrm{Sbm} = \begin{pmatrix} P & Q \\ Q & P \end{pmatrix} = \lambda_1 v_1 v_1^\top + \lambda_2 v_2 v_2^\top$
  - $\widetilde{A} = \mathrm{Sbm} + \kappa\mathcal{G} + \mathcal{E} = \lambda_1 v_1 v_1^\top + \lambda_2 v_2 v_2^\top + \kappa\mathcal{G} + \mathcal{E}$
- $w_2 = \mathrm{sign}(\text{highest eigenvector of } \widetilde{A} - \lambda_1 v_1 v_1^\top)/\sqrt{N}$

$$\frac{1}{N} d_H(v_2, w_2) \leq \|v_2 - w_2\|^2 \leq \frac{8}{\lambda_1^2}\|\kappa\mathcal{G} + \mathcal{E}\|^2 \leq \frac{8}{\lambda_1^2}\mu_1^2(1+o(1))$$

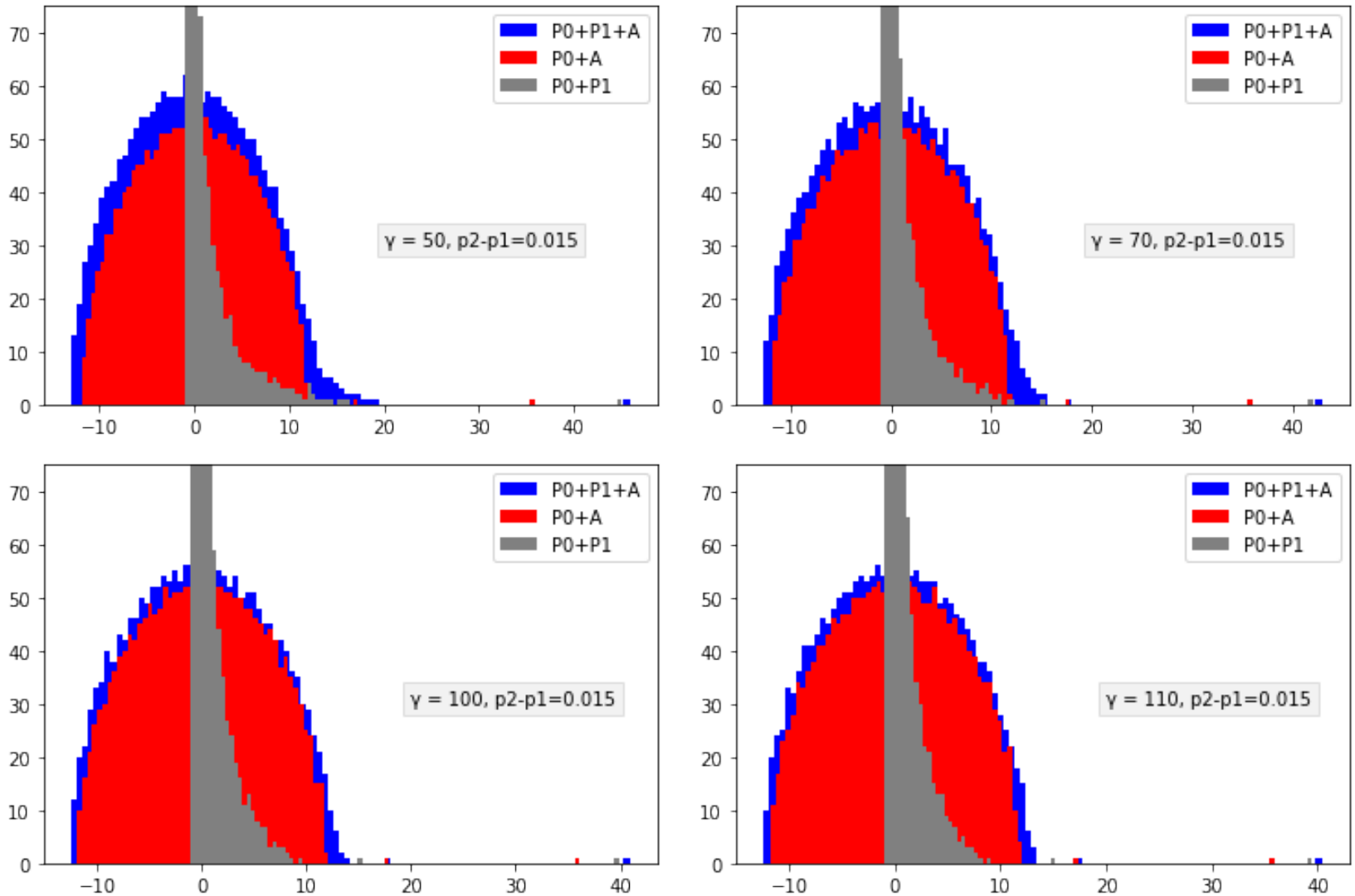
(weak reco: lhs smaller than $1/2$)

$p+q$ is not known

Th: **Soft** recovery if also $\frac{N(p+q)}{2} \geq \frac{N\kappa}{2\gamma} + \frac{N(p-q)}{2}$

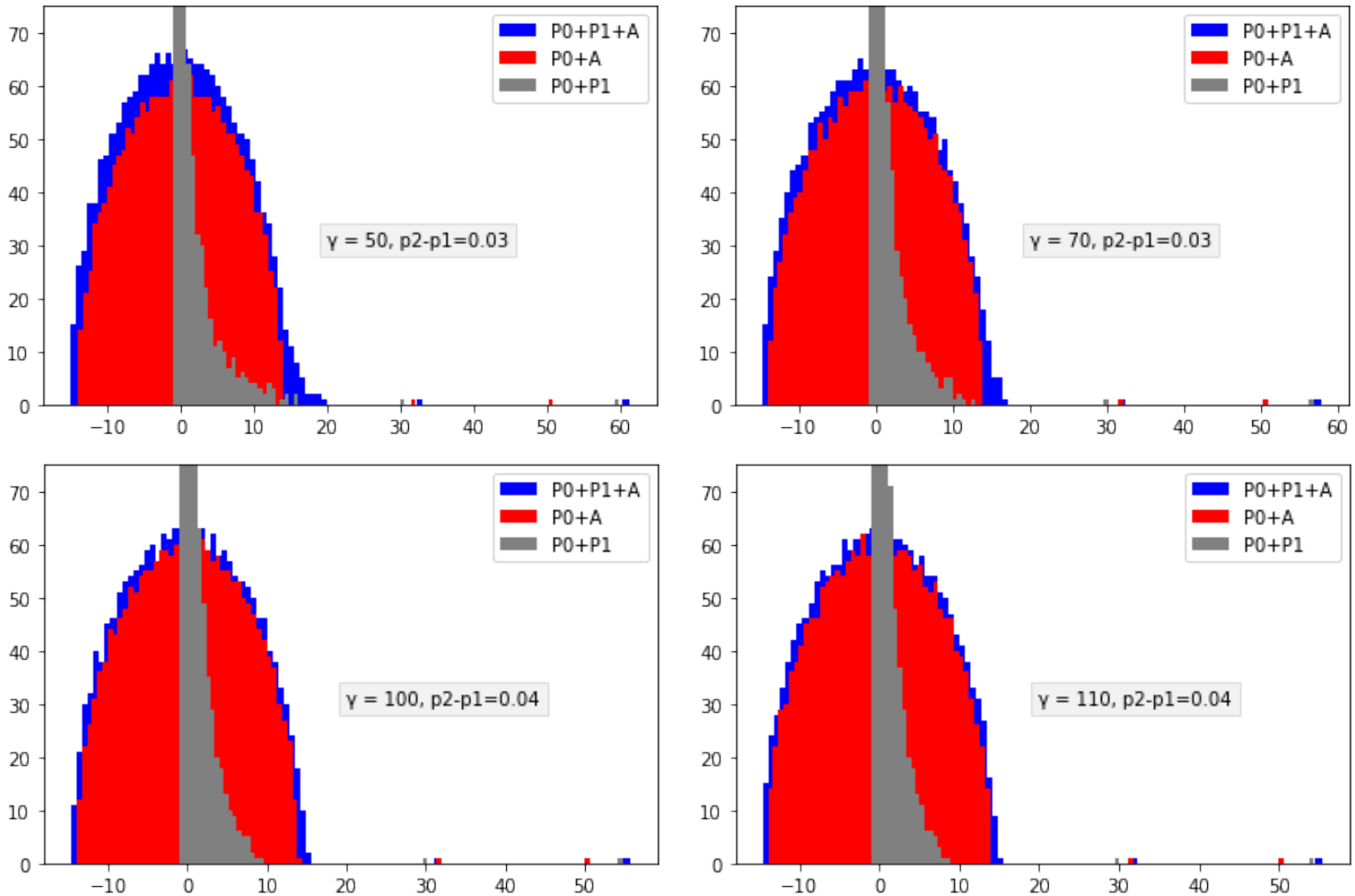
- If $\frac{p-q}{2} \gg \frac{\kappa}{2\gamma}$, **weak** reco (and even exact at limit)
  - $\widetilde{v}_2^\top v_2 = 1 - \frac{p+q}{2p}(\sqrt{\frac{p+q}{p-q}} - 1)\sqrt{\frac{\frac{\kappa}{2\gamma}}{\frac{p-q}{2}}} + \mathcal{O}(\frac{\frac{\kappa}{2\gamma}}{\frac{p-q}{2}})$
- Elements of **proof**
  - "Perturbation analysis"
  - (at least) 2 eigenvalues separate from spectrum of $\kappa\mathcal{G}$
    because $\lambda_1 \tilde{v}_1^\top v_1 \not\to 0$
  - Control the correlation of $\tilde{v}_2$ with v_1, then with v_2

Experiment 1/4 (reco for $\gamma \sim \frac{\kappa}{p-q} = 66$?)



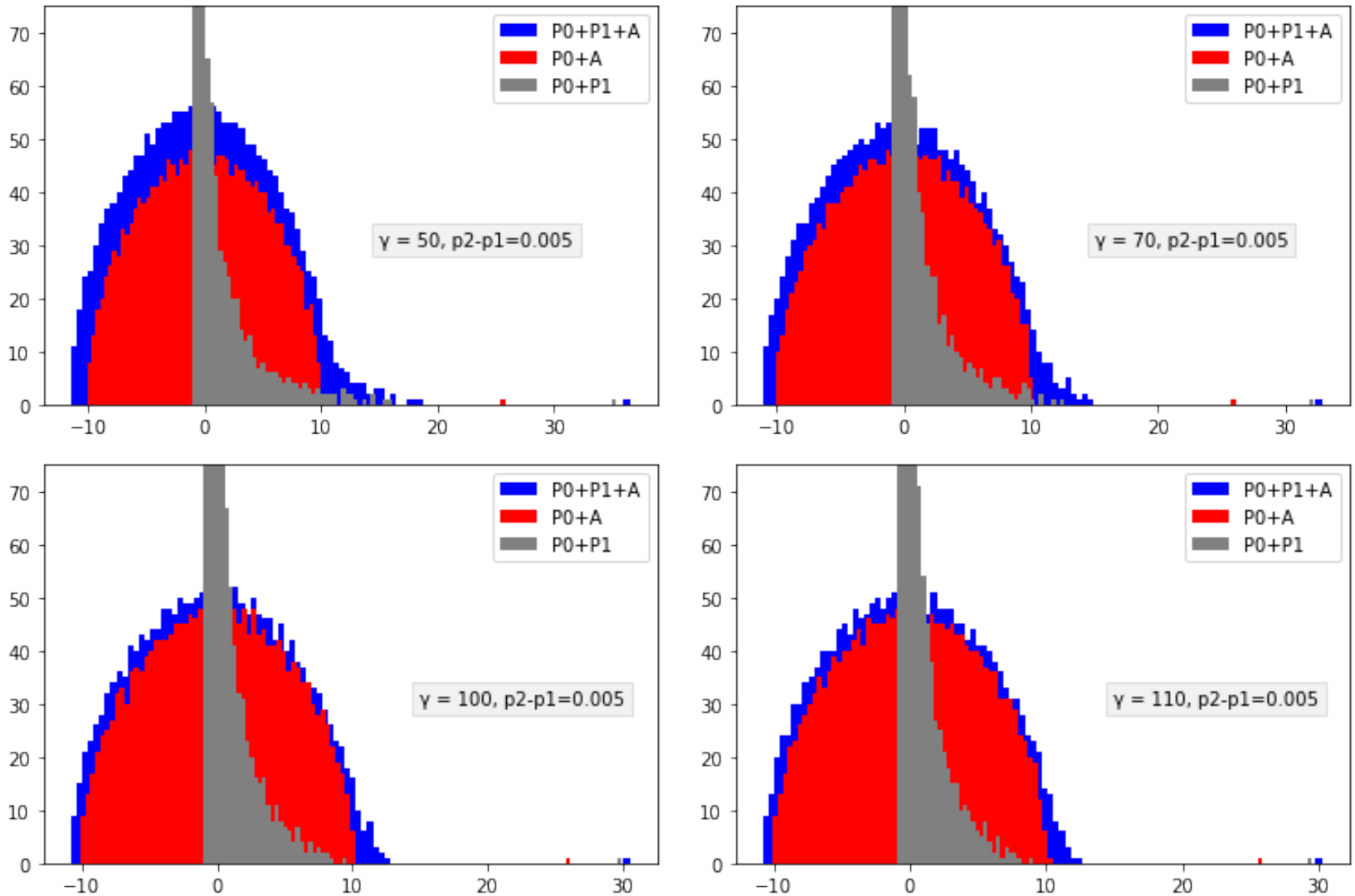
Spectrum for $\gamma \in \{50, 70, 100, 110\}$, $p - q = 1.5\%$, $N = 2.000$

Experiment 2/4 (reco for $\gamma \sim \frac{\kappa}{p-q} = 33$?)



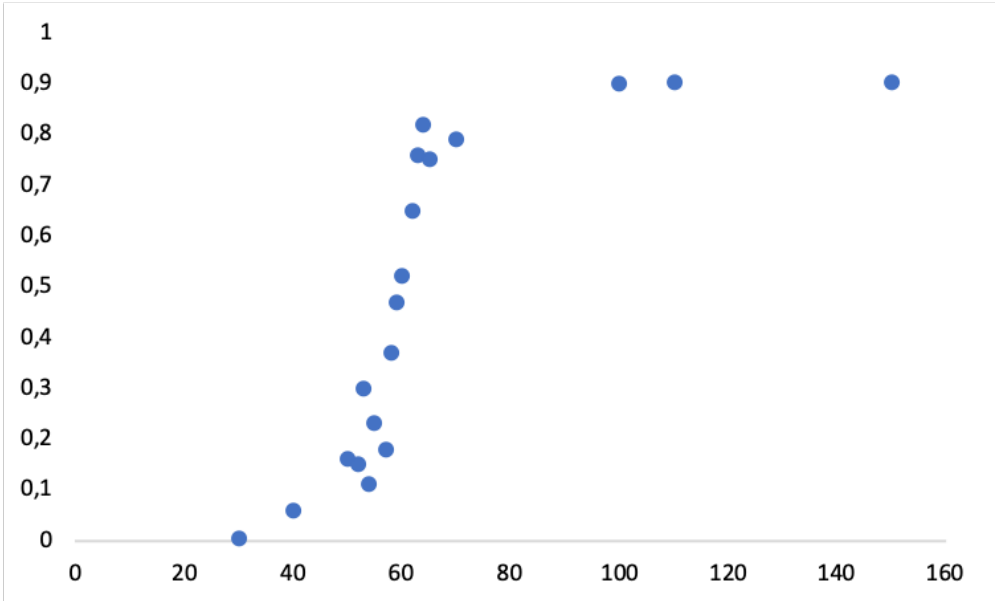
Spectrum for $\gamma \in \{50, 70, 100, 110\}$, $p - q = 3\%$, $N = 2.000$

Experiment 3/4 (reco for $\gamma \sim \frac{\kappa}{p-q} = 200$?)



Spectrum for $\gamma \in \{50, 70, 100, 110\}$, $p - q = 0.5\%$, $N = 2.000$

Experiment 4/4



Correlation between $\tilde{v}_2$ and v_2

Parameters $N = 2.000$, $p = 2.5\%$, $q = 1\%$, $\kappa = 0.97$

Theory Ranges of parameters

- $\frac{N\kappa}{2\gamma} \sim \frac{N(p-q)}{2}$ for $\gamma \sim 65$
- Soft reco as soon as $\gamma \geq \frac{\kappa}{2q} = 49$

Conclusions

- Spectral methods **robust** to **random** perturbations
  - as long as two eigenvalues separate from spectrum
  - (or λ_1 is known)
  - happens if λ_2 is 4 times bigger than spectral radius of pert.
- Algorithm **independent** of presence/absence/quanttiy of noise
- Can be generalized to other perturbations
  - Different Kernel, higher dimension, non-Gaussian features
  - Compute spectral radiius
  - Check that 2 eigenvalues separate
- Extend to more than 2 communities (and unbalanced)

Robustess in the **sparse** regime ($p = \frac{a}{N}$, $q = \frac{b}{N}$, $\frac{1}{\gamma} = \frac{c}{N}$) ?

Robustness of spectral methods for community detection, Stephan and Massoulié, COLT'19