COMMUNITY DETECTION IN STOCHASTIC BLOCK MODELS VIA SPECTRAL METHODS

Laurent Massoulié (MSR-Inria Joint Centre, Inria)

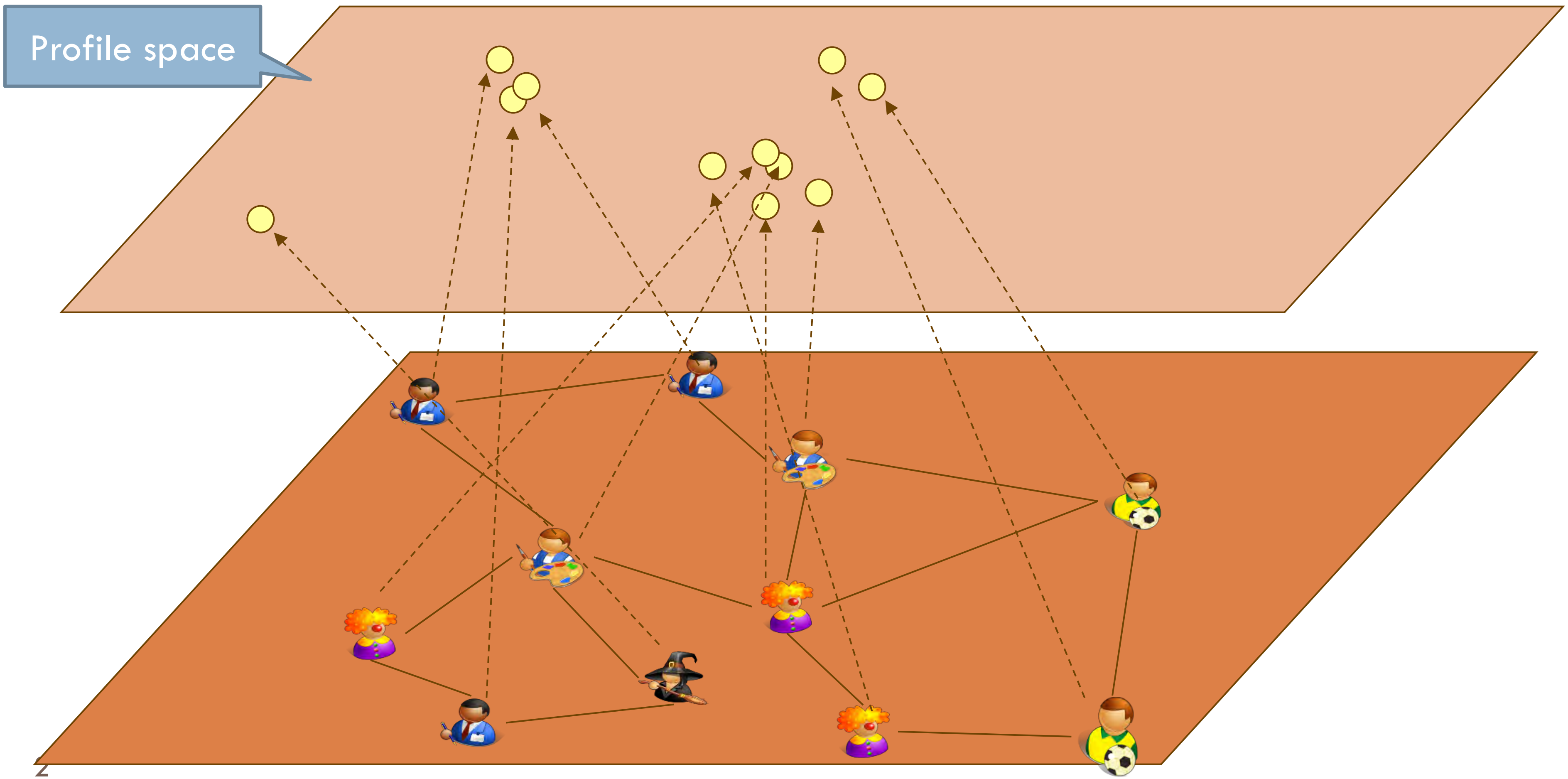
based on joint work with:

Dan Tomozei (EPFL), Marc Lelarge (Inria), Jiaming Xu (UIUC)

Community Detection

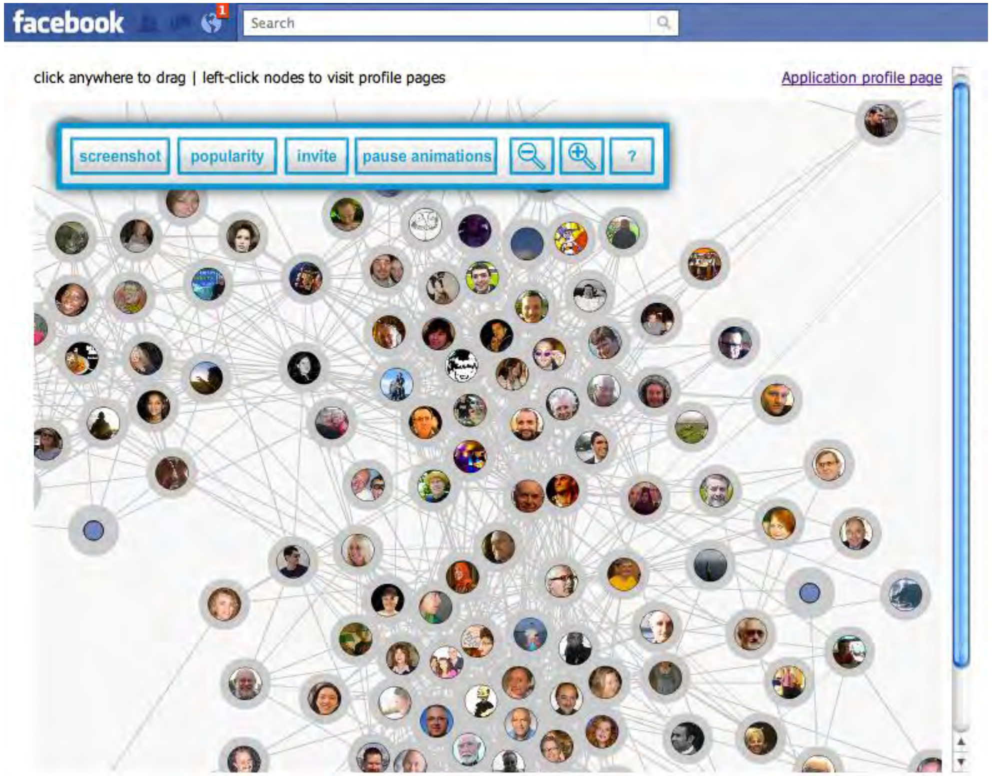
→ Identification of groups of similar objects within overall population

→ Closely related objectives: clustering and embedding



Application 1: contact recommendation in online social networks

Supporting data: e.g. OSN's friendship graph



→recommend members of user's implicit community

Application 2: content recommendation to users of Netflix-like system

Supporting data: user-content ratings matrix

User / Movie	f_1	f_2	…	f_m
u_1	?	**		***
u_2	***	?		?
…				
u_n	*****	**		**

Use content communities to support recommendation
"users who liked this also liked…"

Outline

- The Stochastic Block Model
  - With labels
  - With general types
- Performance of Spectral Methods
  - “rich signal” case
- The weak signal case: sparse observations
  - Phase transition on detectability
  - A modified spectral method

Outline

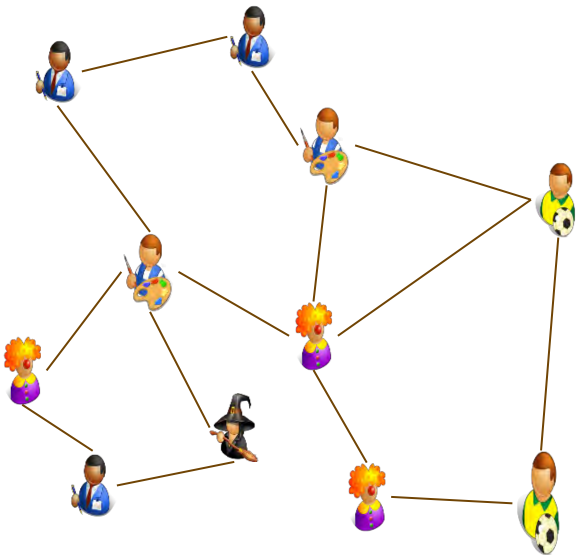
- The Stochastic Block Model
  - With labels
  - With general types
- Performance of Spectral Methods
  - “rich signal” case
- The weak signal case: sparse observations
  - Phase transition on detectability
  - A modified spectral method

The Stochastic Block Model [Holland-Laskey-Leinhardt'83]

- **n** "nodes" partitioned into **K** categories
- Category σ: $\alpha_\sigma n$ nodes
- Edge between nodes **u,v** present

with probability $b_{\sigma(u)\sigma(v)}\, s/n$

s:"signal strength"

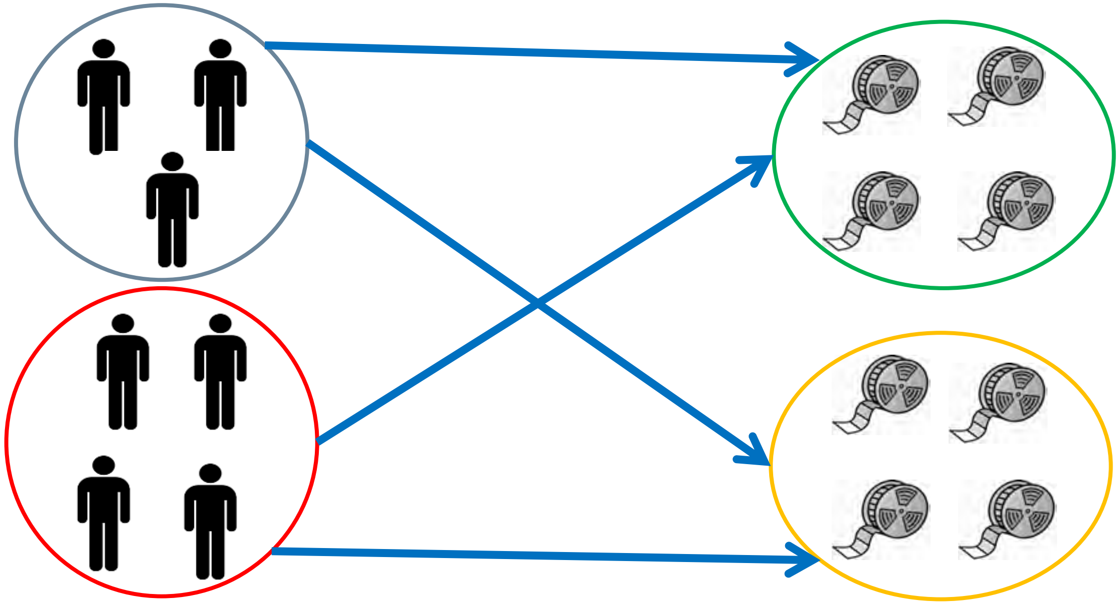


→ Observation: adjacency matrix **A**

A =  + Noise matrix

The Labeled Stochastic Block Model

- Edges (u-v) labeled by $L_{uv} \in L$ (finite set)
- Drawn from distribution $\mu_{\sigma(u)\sigma(v)}$
- Netflix case: labels 1-5 stars

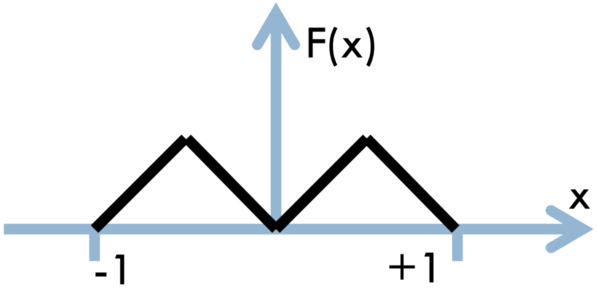


The SBM with general types [Aldous'81; Lovász'12]

- User type $\sigma(u)$ i.i.d. $\sim P$ in general set (e.g. uniform on [0,1])
- Edge (u-v) present w.p. $b_{\sigma(u)\sigma(v)} \, s/n$ for "kernel" b

e.g.

$$b_{x,y} = F(x-y)$$



- Edges (u-v) labeled by $L_{uv} \in L$ (finite set)
- Drawn from distribution $\mu_{\sigma(u)\sigma(v)}$

- Technical assumptions: compact type set and continuity of symmetric functions b and μ

Outline

- The Stochastic Block Model
  - With labels
  - With general types
- Performance of Spectral Methods
  - “rich signal” case
- The weak signal case: sparse observations
  - Phase transition on detectability
  - A modified spectral method

Spectral Clustering

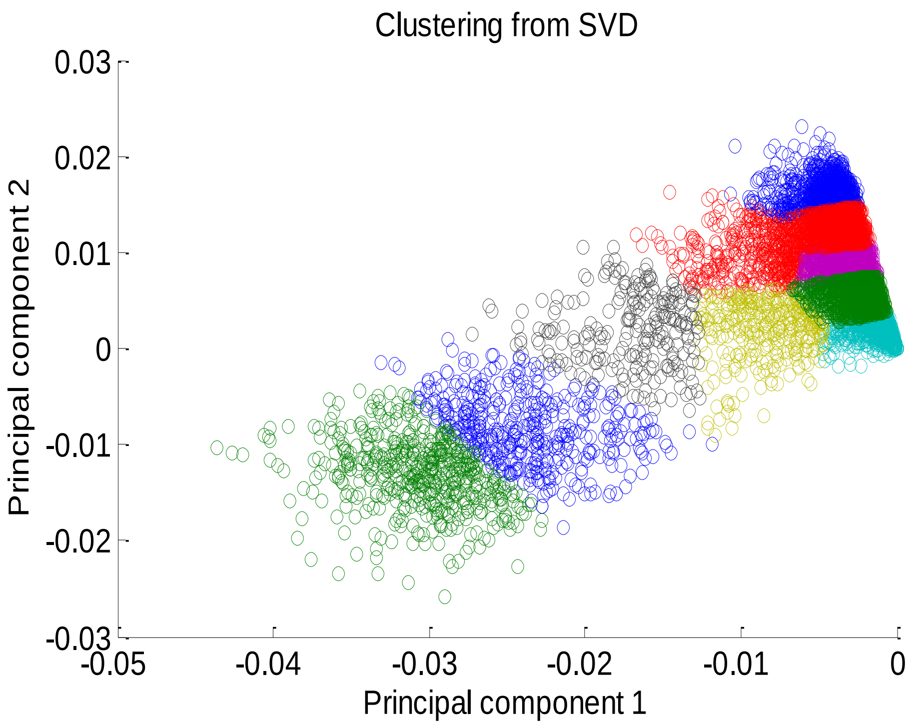
- From Matrix A extract R normalized eigenvectors x_i corresponding to R largest eigenvalues $|\lambda_1| \geq \cdots \geq |\lambda_R|$

- Form R-dimensional node representatives

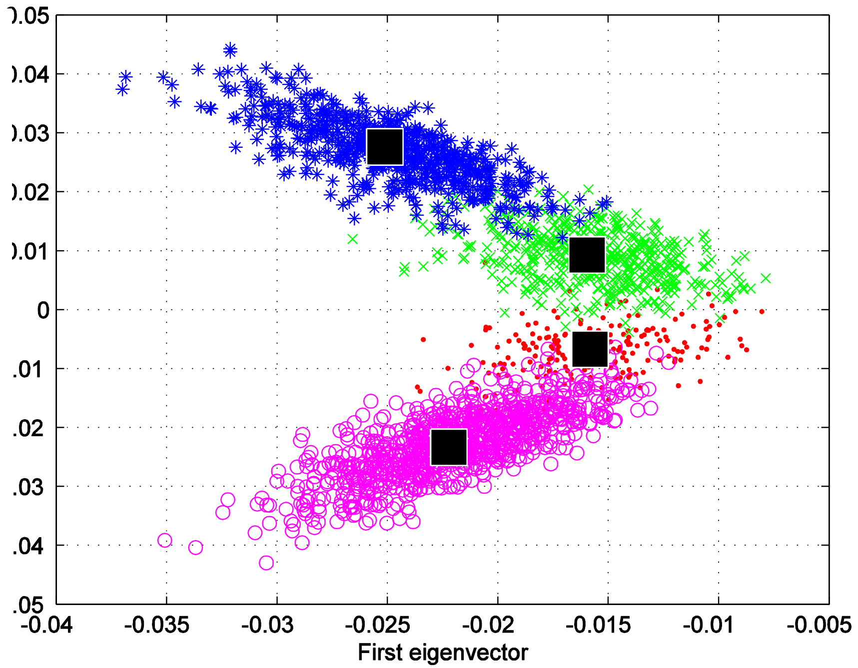
$$y_u = \sqrt{n}(x_i(u))_{\ i=1\ldots R}$$

- Group nodes u according to proximity of spectral representatives y_u

Illustration for R=2



Netflix dataset



SBM with K=4

Result for "logarithmic" signal strength s

Assume s=Ω(log(n)) and clusters are distinguishable, i.e.

$$\forall \sigma \neq \sigma' \ \ \exists \tau \text{ such that } b_{\sigma\tau} \neq b_{\sigma'\tau}$$

→Then spectrum of A consists of

- R eigenvalues λ_i of order Ω(s) (R ≤ K) and
- n-R eigenvalues λ_i of order $O(\sqrt{s})$

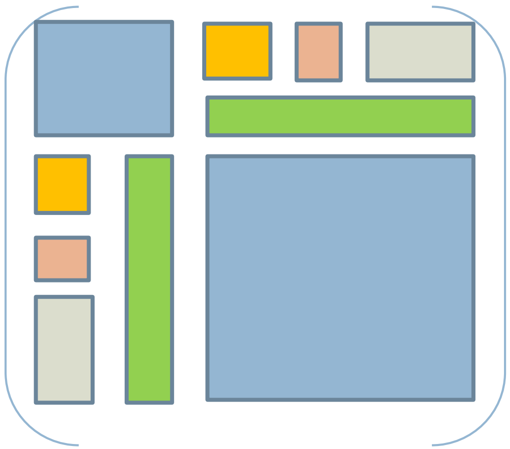
Node representatives y_u based on top R eigenvectors x_i :

Cluster according to underlying "blocks" except for negligible fraction of nodes

Proof arguments

Control spectral radius of noise matrix

+ perturbation of matrix eigen-elements

A =  + random "noise" matrix

Block matrix

non-zero eigenvalues: $\Theta(s)$

spectral separation properties “à la Ramanujan”

s-regular graph Ramanujan if

$$\lambda := \max(|\lambda_2|, |\lambda_n|) \leq 2\sqrt{s-1}$$

[Lubotzky-Phillips-Sarnak’88]

[Friedman’08]: random s-regular graph verifies whp

$$\lambda = 2\sqrt{s-1} + o(1)$$

[Feige-Ofek’05]: for Erdős-Rényi graph $G(n, s/n)$ and $s = \Omega(\log n)$, then whp $\lambda = O(\sqrt{s})$

Also: $\rho(A - \bar{A}) = O(\sqrt{s})$

spectral separation properties "à la Ramanujan"

Corollary: in SBM with $s = \Omega(\log n)$, whp $\rho(A - \bar{A}) = O(\sqrt{s})$ → A's leading eigen-elements close to those of $\bar{A}$

For $s = \Theta(1)$, $\rho(A - \bar{A}) \sim C\sqrt{\frac{\log n}{\log \log n}}$

→spectral separation is lost

Result for "logarithmic" signal strength s – Labeled SBM

Random projection method: transform categorical labels into numerical data

For each label l generate W(l) i.i.d. uniform on [0,1]

Perform Spectral clustering on matrix $\{A_{ij}W(L_{ij})\}$

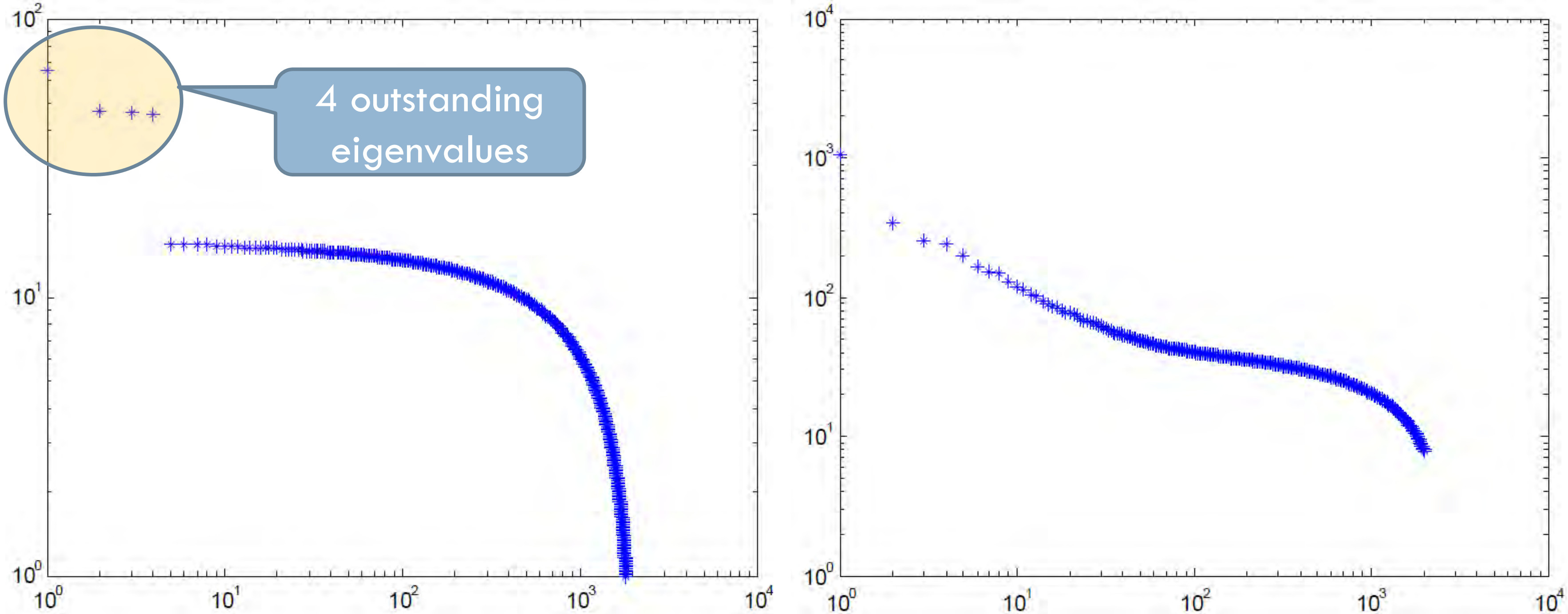
→ Under modified distinguishability condition

$$\forall \sigma \neq \sigma', \exists \tau, \ell \text{ such that } b_{\sigma\tau}\nu_{\sigma\tau}(\ell) \neq b_{\sigma'\tau}\nu_{\sigma'\tau}(\ell)$$

Same result holds as in unlabeled scenario

Discrepancy between SBM with small K and Netflix

Eigenvalue distributions



SBM with K=4

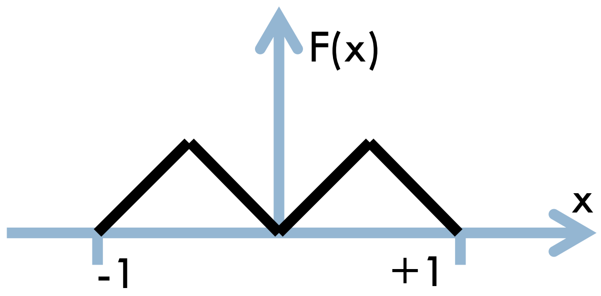
Netflix (subset)

→motivates consideration of SBM with general types

SBM with general types

- User types σ(u) i.i.d. $\sim P$ from general set (e.g. uniform on [0,1])
- Edge (u-v) present w.p. $b_{\sigma(u)\sigma(v)}\, s/n$ for "kernel" b

  e.g. $b_{x,y} = F(x-y)$



- Edges (u-v) labeled by $L_{uv} \in L$ (finite set)
- Drawn from distribution $\mu_{\sigma(u)\sigma(v)}$

→Form matrix $\{A_{ij}W(L_{ij})\}$ from random projections $W(l)$ of labels

SBM with general types: Spectral properties for logarithmic s

Define kernel $K(x,y) := \sum_l W(l)\mu_{xy}(l)$ and integral operator $Tf(x) := \int K(x,y)f(y)P(dy)$

→spectrum of $s^{-1}\{A_{ij}W(L_{ij})\} \approx$ spectrum of T

- Eigenvalue convergence: $s^{-1}\lambda_i^{(n)} \to \lambda_i$
- Eigenvector convergence: $x_i(u) \to \varphi_i(k_u)$

Associated eigen-functionn

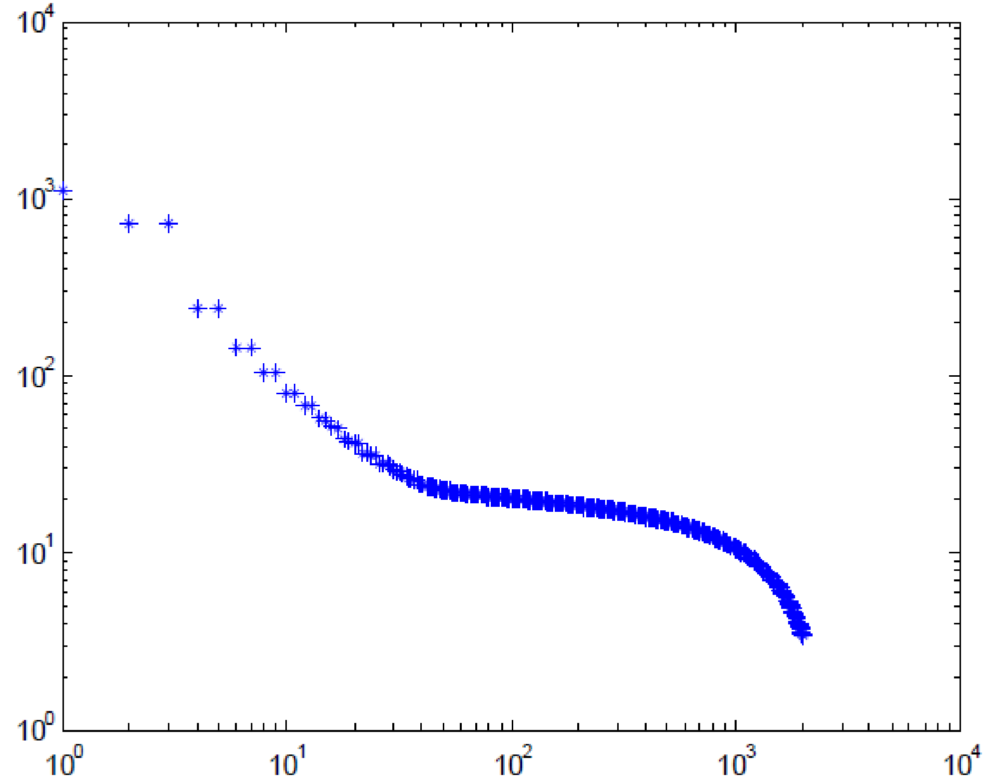
Type of node u

SBM with general types: Spectral properties for logarithmic s

→Flexible model

-power-law spectra (convolution operator + Fourier analysis)

-better matches to Netflix data



SBM with general types: estimation for logarithmic s

- For fixed R form R-dimensional node representatives

$$y_u = \sqrt{n}\left\{\frac{\lambda_k}{\lambda_1} x_k(u)\right\}_{k=1\ldots R}$$

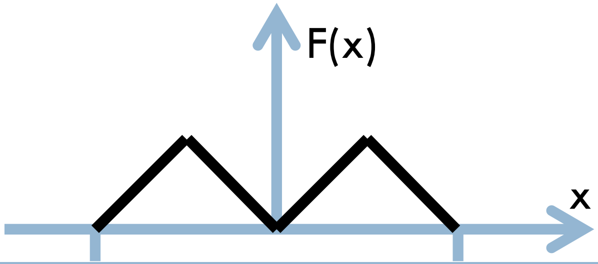
→Embeds nodes according to pseudo-distance d_R that "captures geometry" of hidden node types $\sigma(u)$ with embedding accuracy controlled by "residual energy" $\varepsilon_R := \sum_{k>R} {\lambda_k}^2$ of operator's spectrum

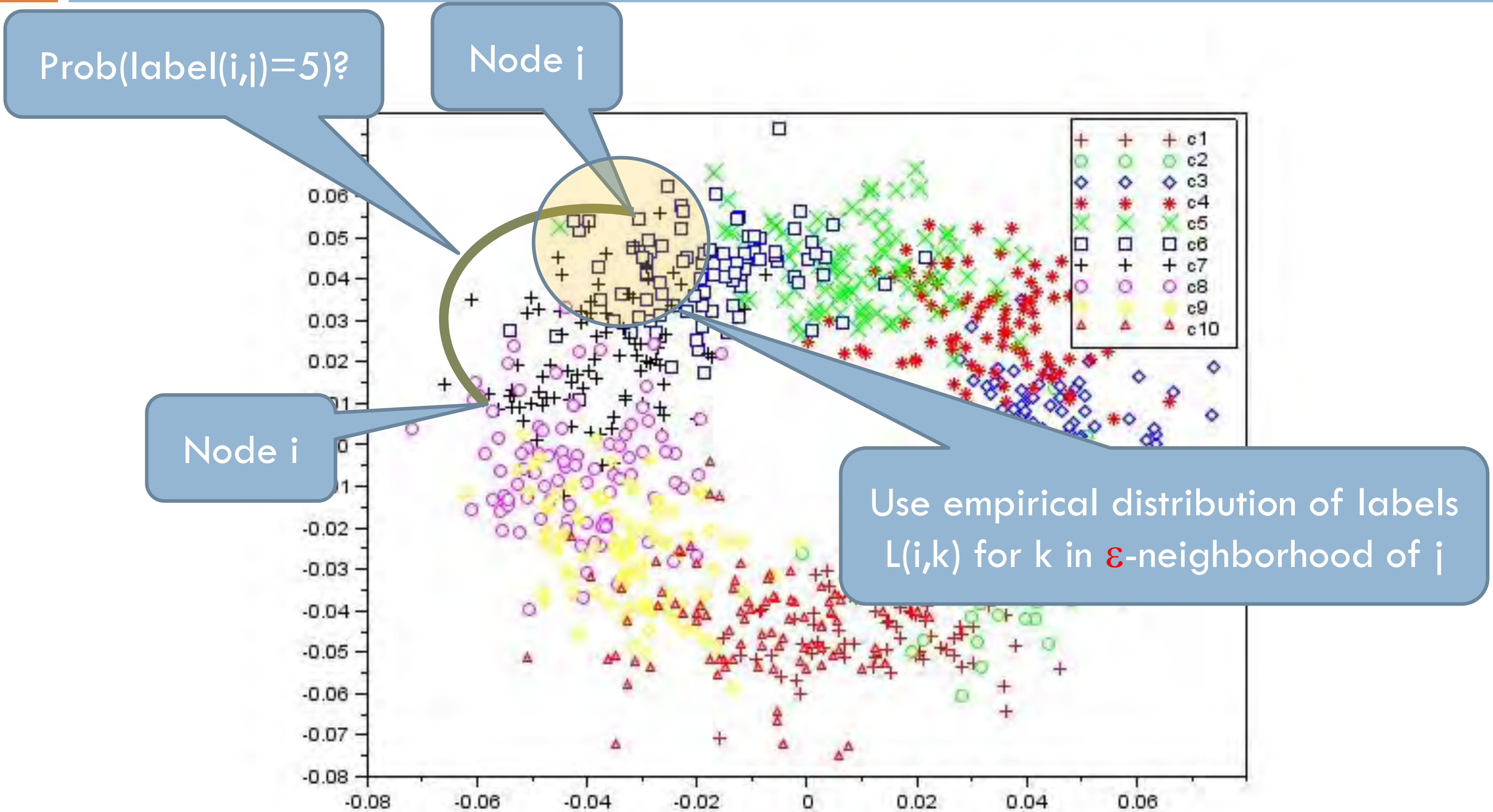
SBM with general types: estimation for logarithmic s

Define Distance $d^2(x,y) = \int [K(x,z) - K(y,z)]^2 P(dz)$

- captures model structure
- Verifies $0 \le d_R \le d$
- And $\iint \left[d^2(x,y) - {d_R}^2(x,y)\right] P(dx) P(dy) = \varepsilon_R$

Illustration with [0,1] types





Embedding allows consistent estimation of label distributions

Consistency result for logarithmic s

Inference of label distribution based on

- R-dimensional embedding
- Empirical measures on ε-neighborhoods

For fraction of $1 - \sqrt{\varepsilon_R}$ node pairs, estimation error verifies

$$\lim_{\varepsilon \to 0}\left(\lim_{\varepsilon_R \to 0} \mathrm{Error}\right) = 0$$

Outline

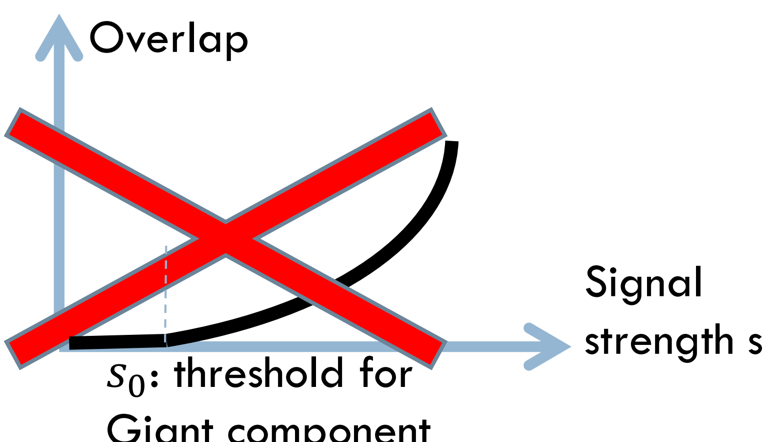
- The Stochastic Block Model
  - With labels
  - With general types
- Performance of Spectral Methods
  - "rich signal" case
- The weak signal case: sparse observations
  - Phase transition on detectability
  - A modified spectral method

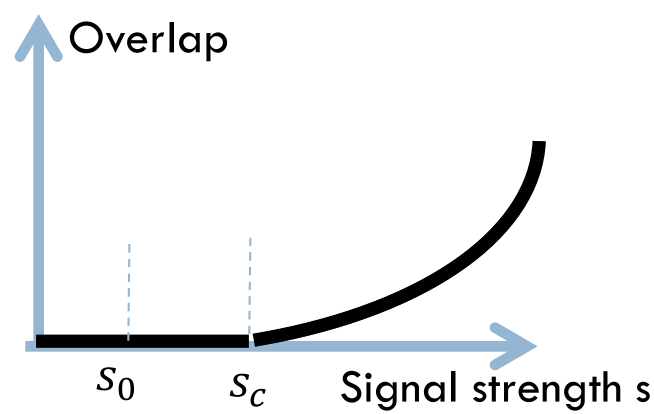
Weak signal strength: $s = \Theta(1)$

- Correct classification of all but negligible fraction of nodes impossible (isolated nodes…)

→ Assess performance of clustering $\hat{\sigma}$ by *overlap* metric:

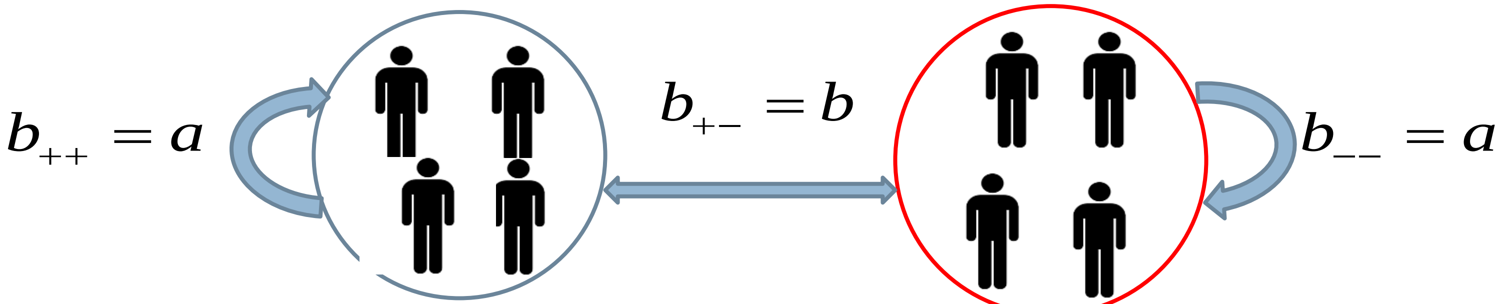
$$\mathrm{ov}(\hat{\sigma}) = \frac{1}{n}\sum_{u=1}^{n} 1\{\sigma_u = \hat{\sigma}_u\} - \max_k(\alpha_k)$$





Weak signal strength : s=1

Symmetric two-communities scenario: $\alpha_+ = \alpha_- = \frac{1}{2}$



Conjecture ([Decelle-Krzakala-Moore-Zdeborova 2011]:

- For $\tau := \frac{(a-b)^2}{2(a+b)} < 1$, overlap tends to zero for any $\hat{\sigma}$

→Proven by [Mossel-Neeman-Sly 2012]

- For $\tau > 1$, positive overlap can be achieved

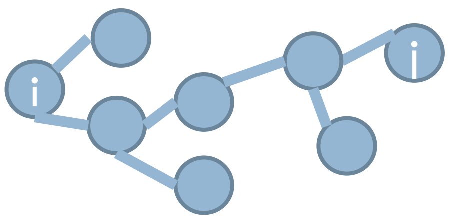
(by Belief Propagation [DKMZ 2011]; by "spectral redemption" [KMMNSZ-Zhang 2013])

No method proven to achieve positive overlap until Nov'13

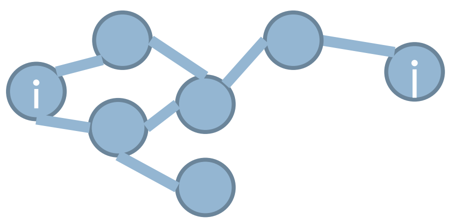
Detection by modified spectral method

Form matrix $B^{(l)}$: $B^{(l)}{}_{ij}$ = nb of self-avoiding paths of length l

Ex: for l=4



$B^{(l)}{}_{ij} = 1$



$B^{(l)}{}_{ij} = 2$

Typical case: for tree-shaped l-neighborhood of i,

$B^{(l)}{}_{ij} = 1_{\{d(i,j)=l\}}$

Main result: spectral structure of $B^{(l)}$ for $\tau > 1$ & path length $l \sim c \log(n)$,

Let $\alpha = \frac{a+b}{2}$, $\beta = \frac{a-b}{2}$ (hence $\tau = \frac{\beta^2}{\alpha}$) eigenvalues of $\begin{pmatrix} a/2 & b/2 \\ b/2 & a/2 \end{pmatrix}$

- Top eigenvalue $\sim \tilde{\Theta}(\alpha^l)$, top eigenvector y: $|\langle y, B^{(l)} e\rangle| \sim |y|\,|B^{(l)} e|$
- 2nd eigenvalue $\geq \tilde{\Omega}(\beta^l)$, 2nd eigenvector z: $|\langle z, B^{(l)} \sigma\rangle| \sim |z|\,|B^{(l)} \sigma|$
- 3rd eigenvalue $= O(n^{\varepsilon}\sqrt{\alpha^l})$ for all $\varepsilon > 0$

Spectral separation
"à la Ramanujan"

- 2nd eigenvector z of $B^{(l)}$ positively correlated with spin vector σ

→Hence positive overlap obtained by estimate $\hat{\sigma}(u) = \begin{cases} +1 \text{ if } z_u\sqrt{n} > T \\ -1 \text{ if } z_u\sqrt{n} \leq T \end{cases}$

For suitable threshold T

Proof elements 1) matrix expansion

- Expected adjacency matrix $\bar{A} = \frac{a}{n}\left[\frac{1}{2}(ee' + \sigma\sigma') - I\right] + \frac{b}{2n}(ee' - \sigma\sigma')$

- Centered simple path adjacency matrix $\Delta_{ij}^{(\ell)} := \sum_{i_0^\ell \in P_{ij}} \prod_{t=1}^{\ell} (A - \bar{A})_{i_{t-1} i_t}$

→Expansion: $B^{(\ell)} = \Delta^{(\ell)} + \sum_{m=1}^{\ell} (\Delta^{(\ell-m)} \bar{A} B^{(m-1)}) - \sum_{m=1}^{\ell} \Gamma^{\ell,m}$

"small" terms

“Smallness” of matrix coefficients

- Trace method: $\rho(M)^{2k} \leq \text{Trace}\left(M^{2k}\right)$

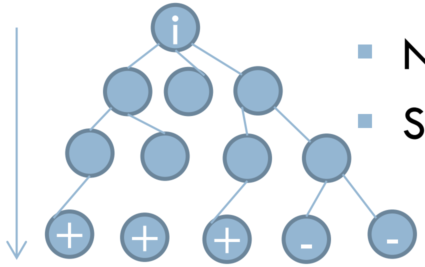
+ combinatorics (à la [Füredi-Komlós’81])

Here: count contributions of concatenations of simple paths

→Bounds on spectral radii: whp, for all $\varepsilon > 0$

$$\rho(\Delta^{(\ell)}) \leq n^{\epsilon}\alpha^{\ell/2},$$
$$\rho(\Gamma^{\ell,m}) \leq n^{\epsilon-1}\alpha^{(\ell+m)/2}, \ m = 1,\ldots,\ell.$$

Proof elements 2) Quasi-deterministic growth of node neighborhoods



- Nb of distance t neighbors: $S_t(i)$
- Sum of spins of distance t neighbors: $D_t(i)$

→then whp:

$$S_t(i) = \alpha^{t-l} S_l(i) + \tilde{O}\left(\alpha^{t/2}\right)$$

$$D_t(i) = \beta^{t-l} D_l(i) + \tilde{O}\left(\alpha^{t/2}\right)$$

Proof: Chernoff bounds on binomial random variables

Corollary: For $m \le l$, whp

Close to vectors $\{S_{m-1}(i)\}$, $\{D_{m-1}(i)\}$

$$\sup_{|x|=1, x'B^{(\ell)}e = x'B^{(\ell)}\sigma = 0} |e'B^{(m-1)}x| = \tilde{O}\left(\sqrt{n}\alpha^{(m-1)/2}\right)$$

$$\sup_{|x|=1, x'B^{(\ell)}e = x'B^{(\ell)}\sigma = 0} |\sigma'B^{(m-1)}x| = \tilde{O}\left(\sqrt{n}\alpha^{(m-1)/2}\right)$$

Weak Ramanujan property

- Previous results combined give

$$\sup_{|x|=1, x'B^{(\ell)}e = x'B^{(\ell)}\sigma = 0} |B^{(\ell)}x| \le n^{\epsilon}\alpha^{\ell/2}.$$

Use spectral radius bounds

$$B^{(\ell)} = \Delta^{(\ell)} + \sum_{m=1}^{\ell} (\Delta^{(\ell-m)} \bar{A} B^{(m-1)}) - \sum_{m=1}^{\ell} \Gamma^{\ell,m}$$

Express in terms of e, σ : $\bar{A} = \frac{a}{n}\left[\frac{1}{2}(ee' + \sigma\sigma') - I\right] + \frac{b}{2n}(ee' - \sigma\sigma')$

Use bounds from quasi-deterministic growth on

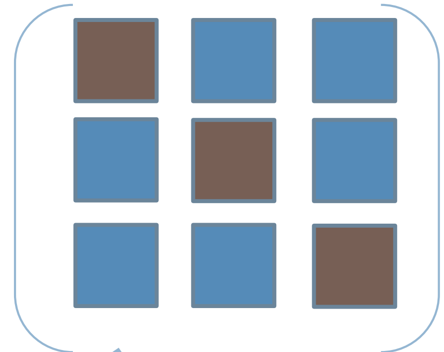
$$\sup_{|x|=1, x'B^{(\ell)}e = x'B^{(\ell)}\sigma = 0} |e'B^{(m-1)}x|$$

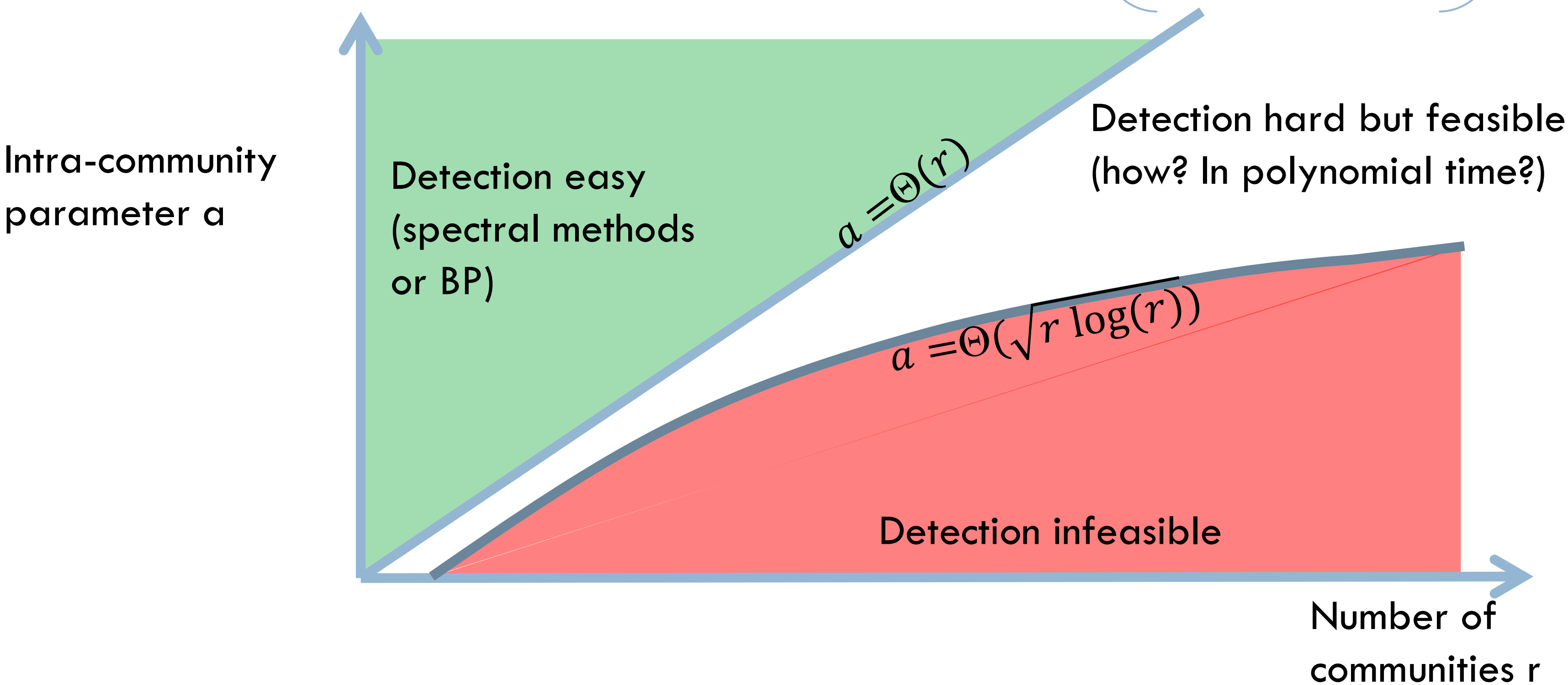
$$\sup_{|x|=1, x'B^{(\ell)}e = x'B^{(\ell)}\sigma = 0} |\sigma'B^{(m-1)}x|$$

Remaining mysteries about SBM's (1)

Conjectured "phase diagram" for more than 2 blocks

(assuming fixed inter-community parameter b)

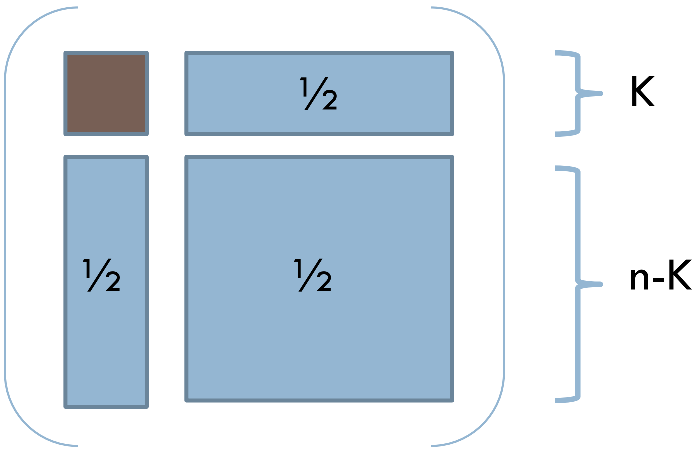




Remaining mysteries about SBM's (2)

Clique detection problem: add a size-K clique to random graph with edge-probability ½

i.e. a 2-block SBM with unbalanced block sizes:



→for $K = \Omega\left(\sqrt{n}\right)$ clique easily detectable (e.g. inspection of node degrees)

→are there polynomial-time algorithms for smaller yet large K?
(e.g. $K = \Theta\left(\sqrt[3]{n}\right)$)
A notoriously hard problem ("planted clique detection" recently proposed as a new benchmark of algorithmic hardness)

Conclusions and Outlook

- "Vanilla" spectral methods efficient for strong (logarithmic) signal strength
- Alternatives needed at low signal strength
  - Belief propagation conjectured optimal
  - Spectral approach on path-expanded matrix proven optimal down to "easy/hard" transition
- Computationally efficient methods for "hard" cases?
  - Detection in SBM = rich playground for analysis of computational complexity with methods of statistical physics
- Does SBM model correctly real-life data?
- Speed of convergence, better-than-random label projections, choice of embedding dimension…

References

- D. Tomozei, L.M., distributed user profiling via spectral methods, ACM Sigmetrics'10
- M. Lelarge, L.M., J. Xu, Reconstruction in the labelled stochastic block model, ITW'13
- J. Xu, L.M., M. Lelarge, Edge label inference in generalized SBM: from spectral theory to impossibility results, COLT'14
- L.M., Community detection thresholds and the weak Ramanujan property, ACM STOC'14