ISIT Tutorial
Information theory and machine learning
Part II

Emmanuel Abbe
Princeton University

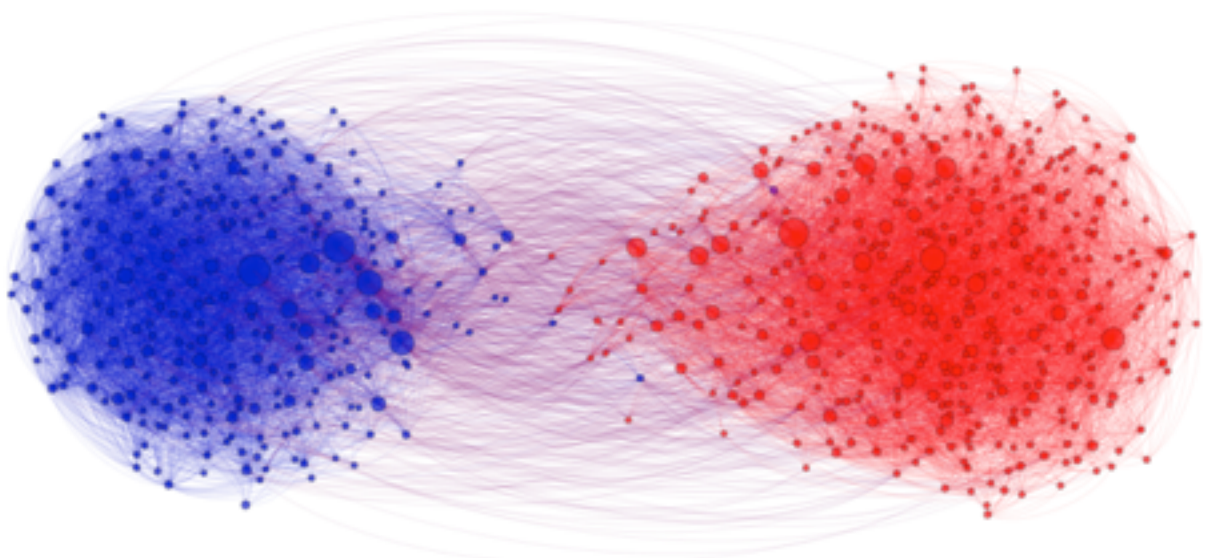
Martin Wainwright
UC Berkeley

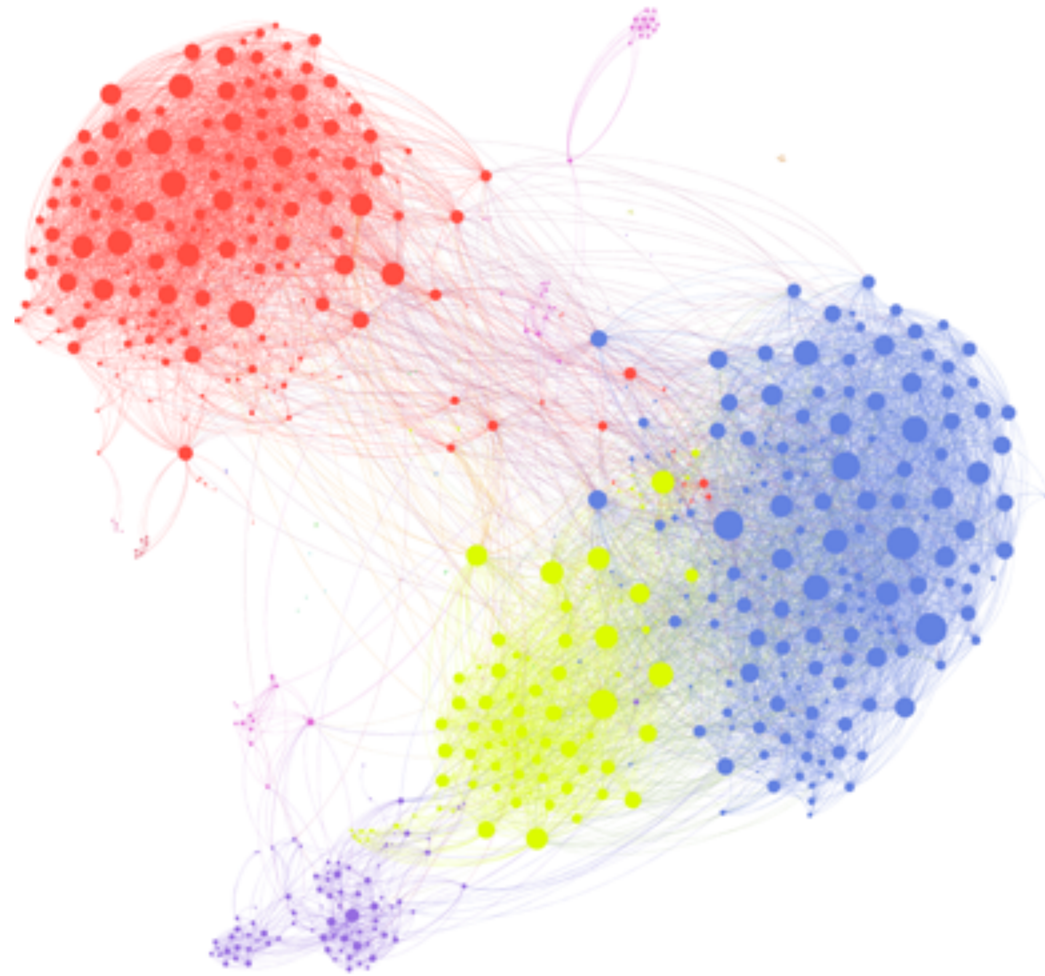
Inverse problems on graphs

A large variety of machine learning and data-mining problems are about inferring global properties on a collection of agents by observing local noisy interactions of these agents

Examples:

- community detection in social networks



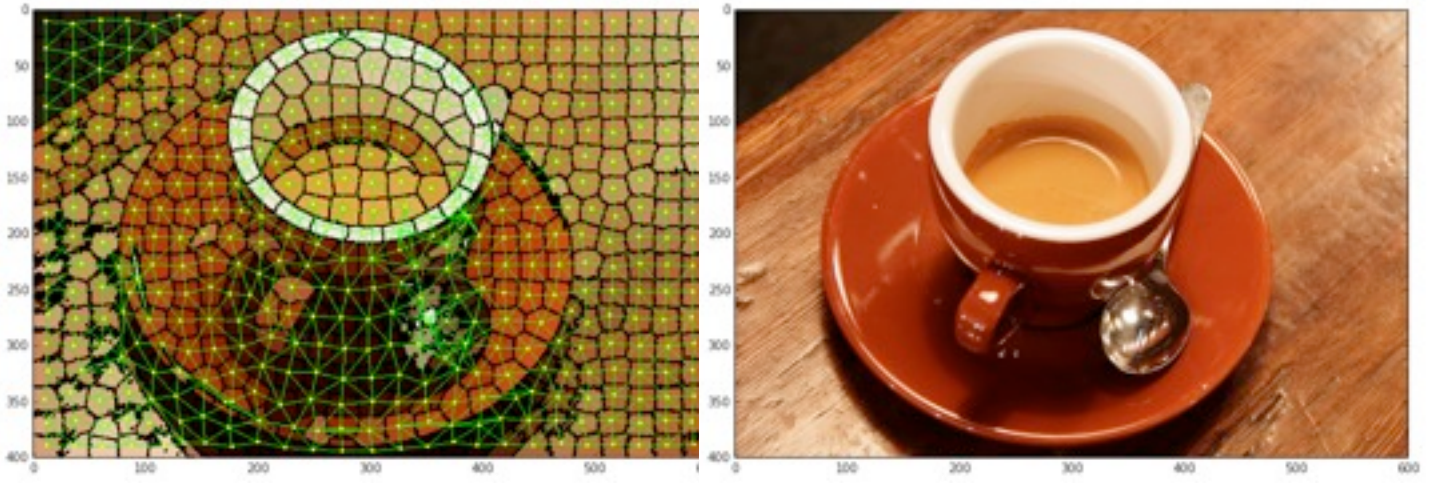


Inverse problems on graphs

A large variety of machine learning and data-mining problems
are about inferring global properties on a collection of agents
by observing local noisy interactions of these agents

Examples:

- community detection in social networks
- image segmentation

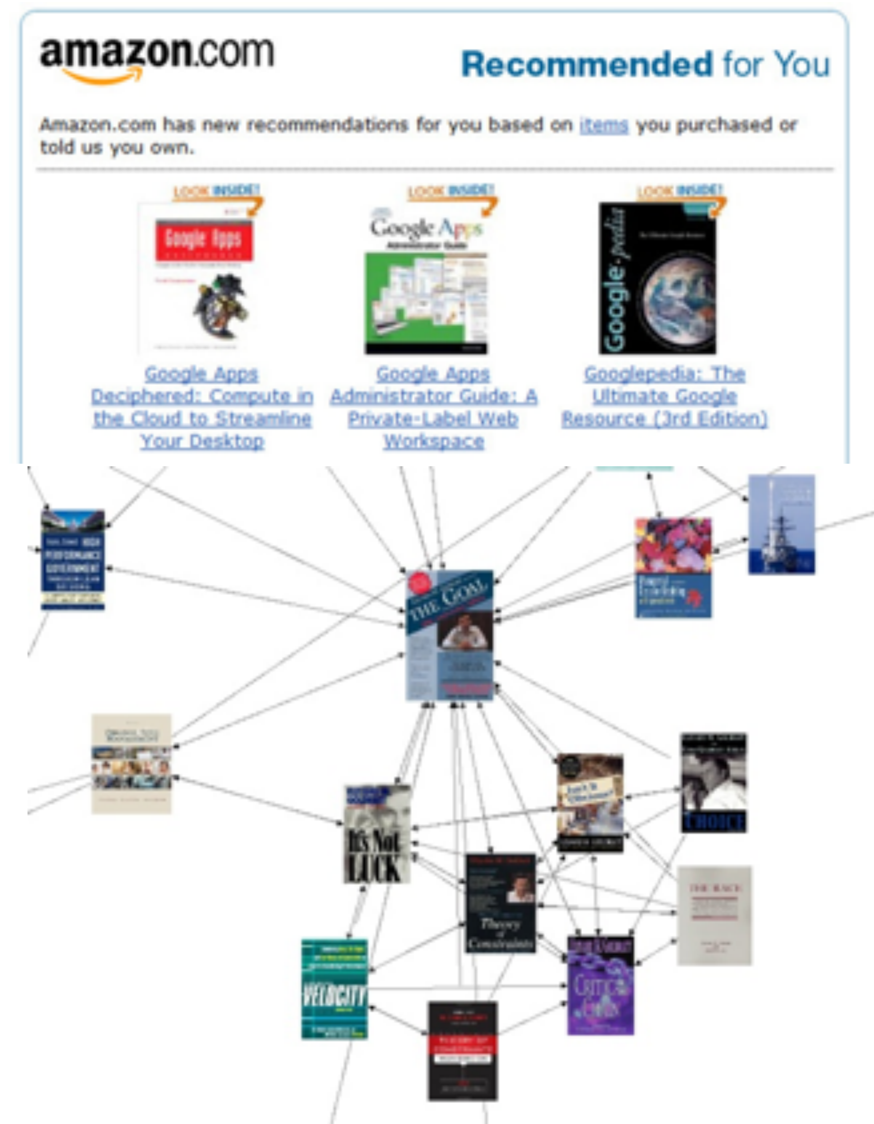


Inverse problems on graphs

A large variety of machine learning and data-mining problems are about inferring global properties on a collection of agents by observing local noisy interactions of these agents

Examples:

- community detection in social networks
- image segmentation
- data classification and information retrieval



Inverse problems on graphs

A large variety of machine learning and data-mining problems are about inferring global properties on a collection of agents by observing local noisy interactions of these agents

Examples:

- community detection in social networks
- image segmentation
- data classification and information retrieval

- object matching, synchronization
- page sorting
- protein-to-protein interactions
- haplotype assembly
- ...

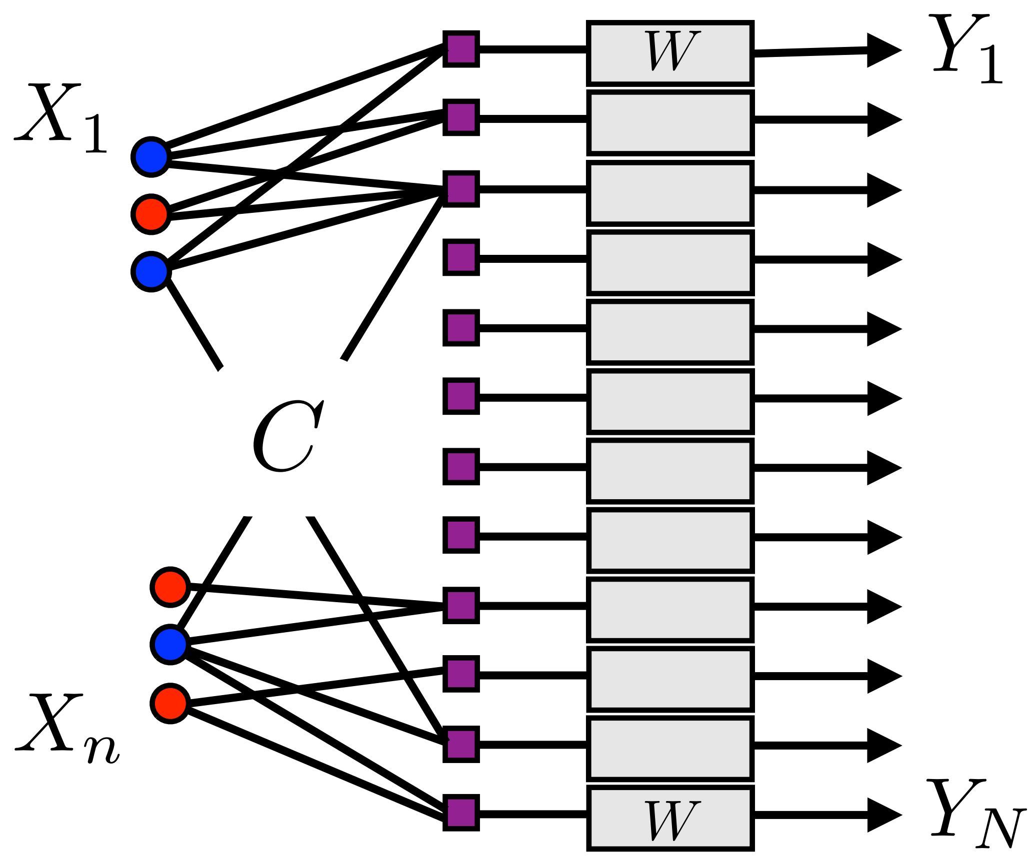
Inverse problems on graphs

A large variety of machine learning and data-mining problems are about inferring global properties on a collection of agents by observing local noisy interactions of these agents

In each case: observe information on the edges of a network that has been generated from hidden attributes on the nodes, and try to infer back these attributes

Dual to the graphical model learning problem (previous part)

What about graph-based codes?



Different: the code is a design parameter and takes typically specific non-local interactions of the bits (e.g., random, LDPC, polar codes).

(1) What are the relevant types of "codes" and "channels" behind machine learning problems?
(2) What are the fundamental limits for these?

Outline of the talk

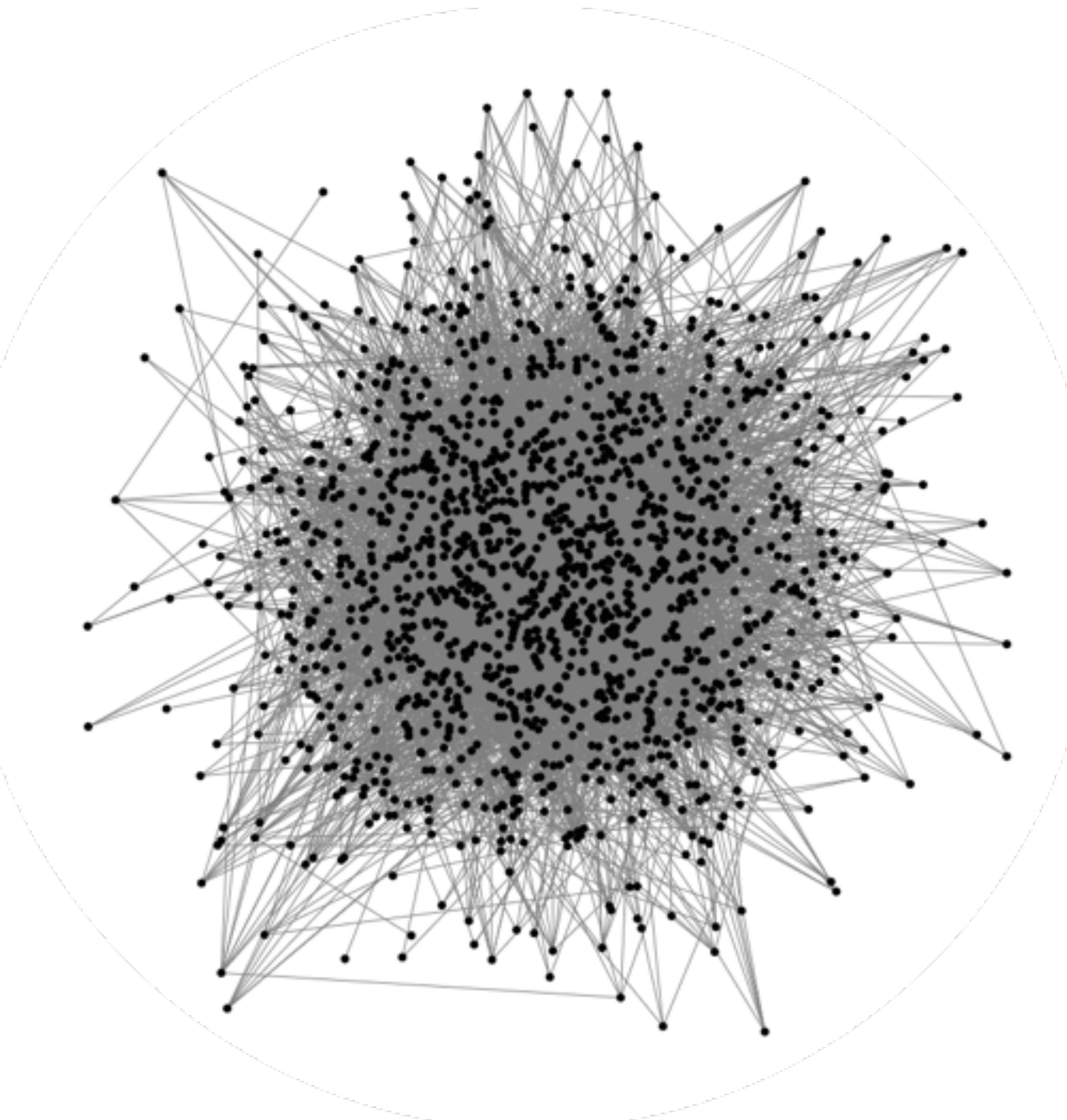
1. Community detection and clustering
2. Stochastic block models :
   fundamental limits and capacity-achieving algorithms
3. Open problems
4. Graphical channels and low-rank matrix recovery

Community detection and clustering

Networks provide local interactions among agents

social networks:
“friendship”

call graphs:
“calls”

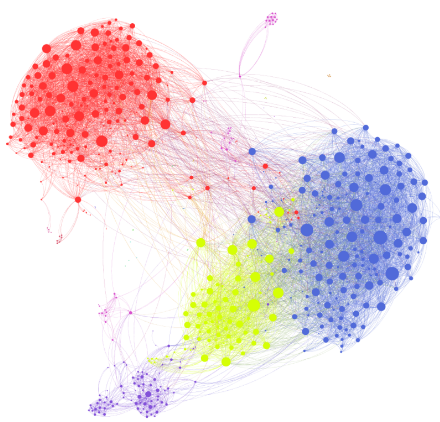


biological networks:
“protein interactions”

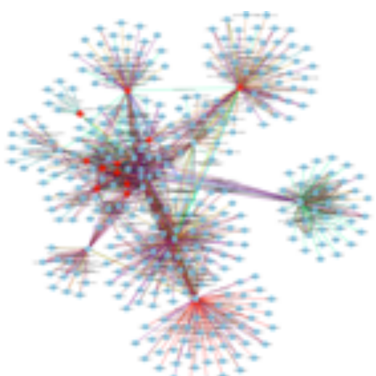
genome HiC networks:
“DNA contacts”

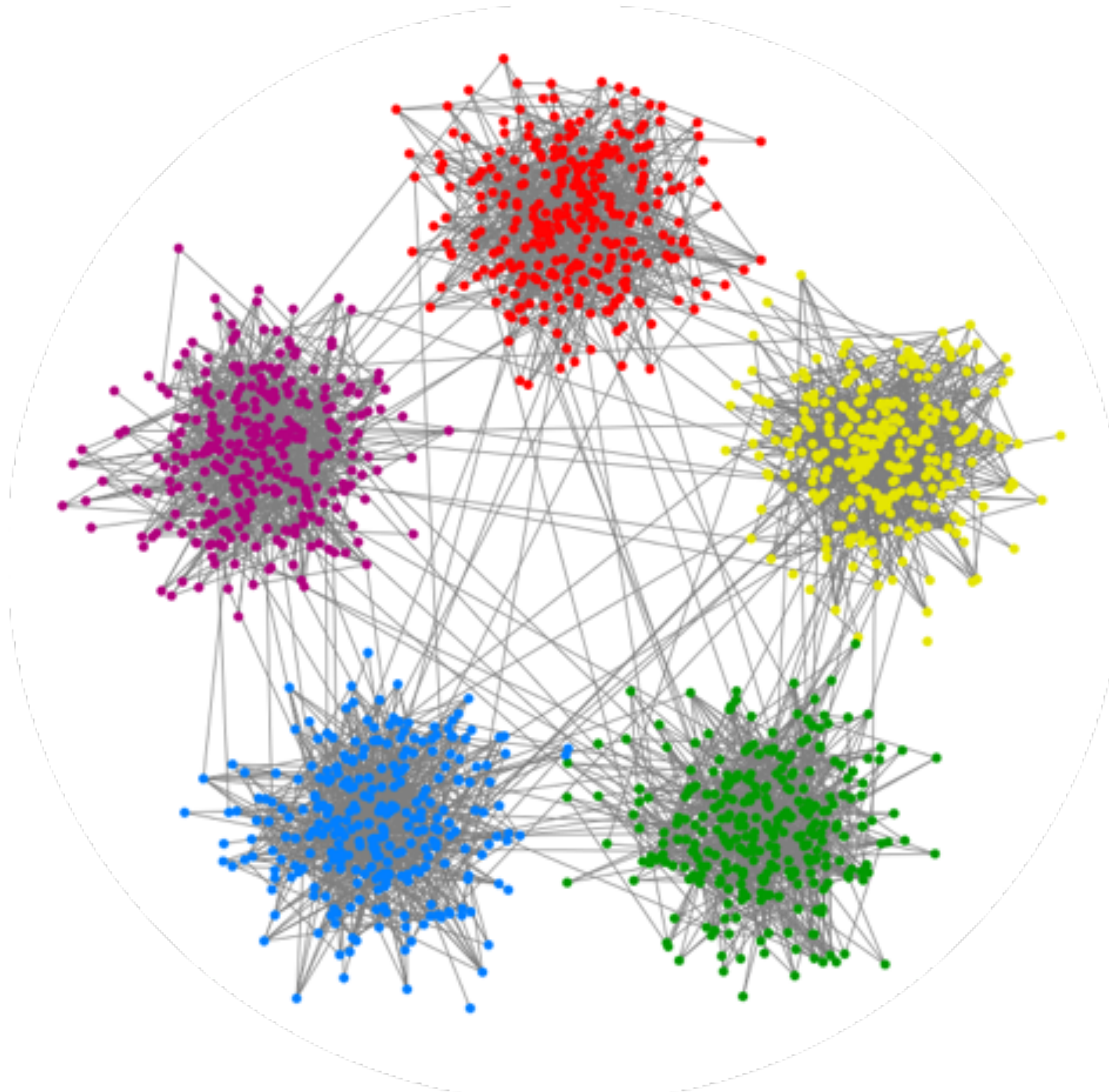
Networks provide local interactions among agents
one often wants to infer global similarity classes

social networks:
"friendship"



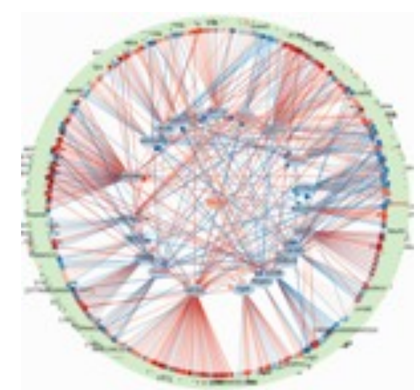
biological networks:
"protein interactions"





call graphs:
"calls"

genome HiC networks:
"DNA contacts"



The challenges of community detection

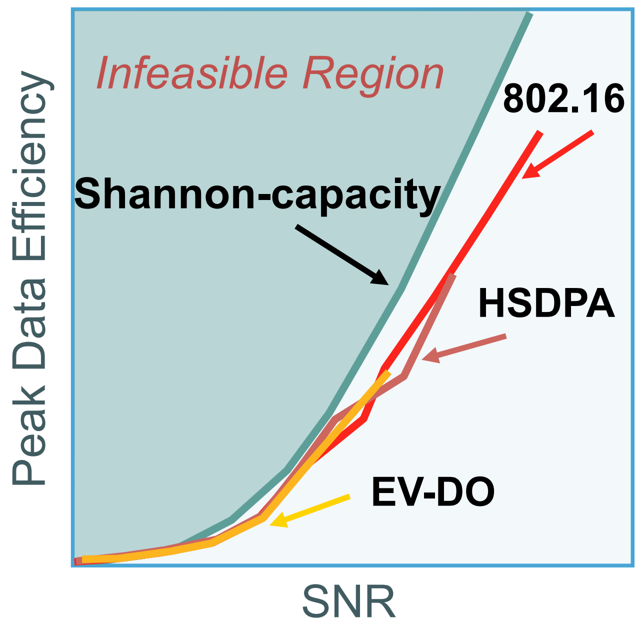
A long-studied and notoriously hard problem

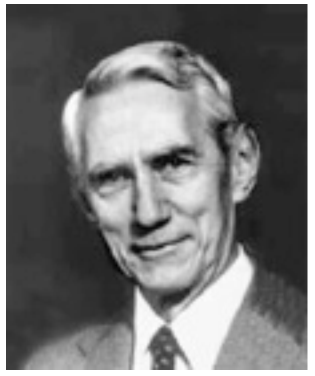
what is a good clustering?
assort. and disassort. relations
→ work with models

how to get a good clustering?
computationally hard
→ many heuristics...

Tutorial motto:
Can on establish a clear line-of-sight
as in communications with
the Shannon capacity?

WAN wireless tech.





The Stochastic Block Model

The stochastic block model

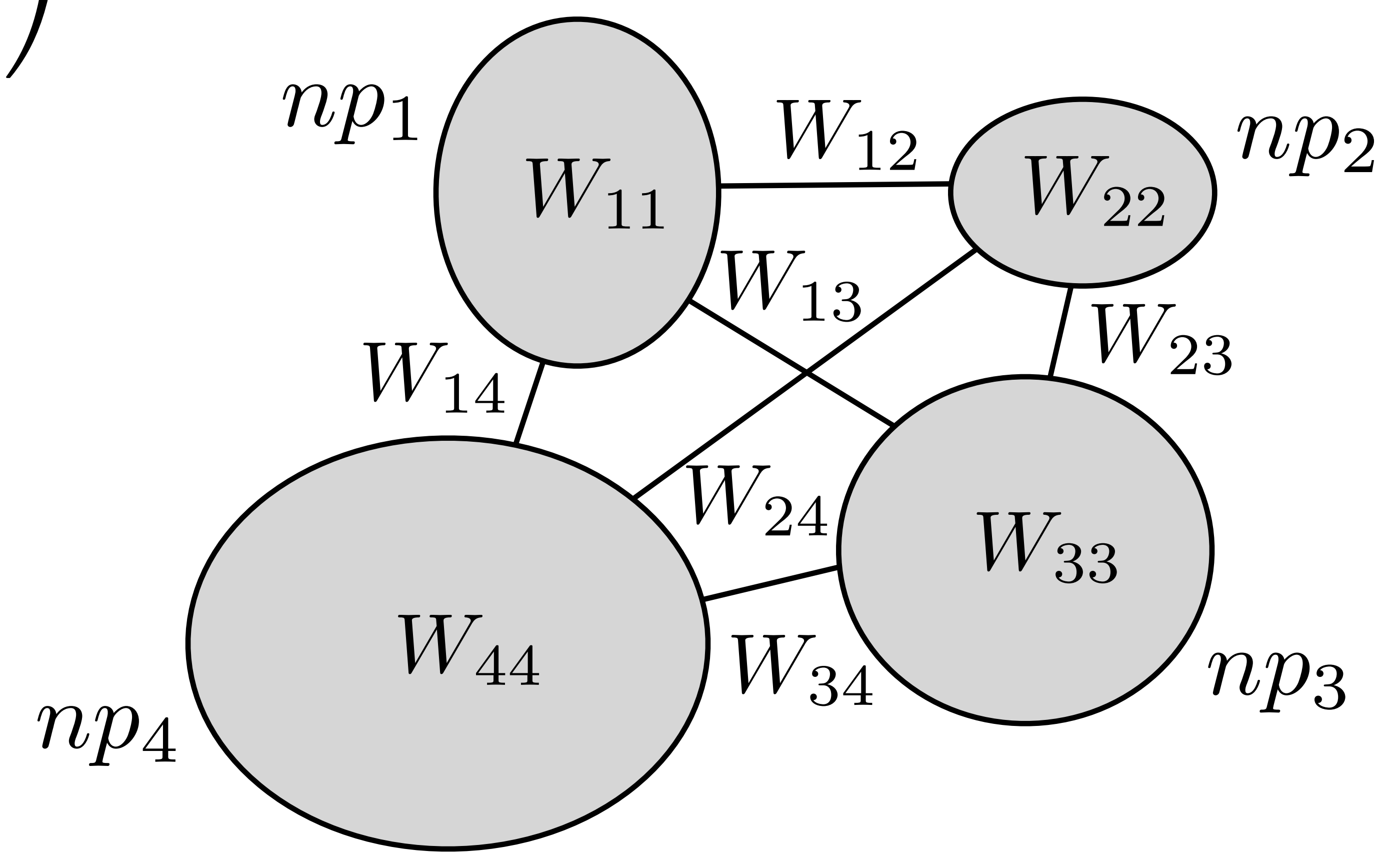
$$\mathrm{SBM}(n, p, W)$$

$$P = \mathrm{diag}(p)$$

$p = (p_1, \ldots, p_k)$ <- probability vector = relative size of the communities

$$W = \begin{pmatrix} W_{11} & \cdots & W_{1k} \\ \vdots & \ddots & \vdots \\ W_{k1} & \cdots & W_{kk} \end{pmatrix}$$

<- symmetric matrix with entries in [0,1]
= prob .of connecting among communities



$k = 4$

The DMC of clustering..?

The (exact) recovery problem

Let $X^n = [X_1, \ldots, X_n]$ represent the community variables of the nodes (drawn under p)

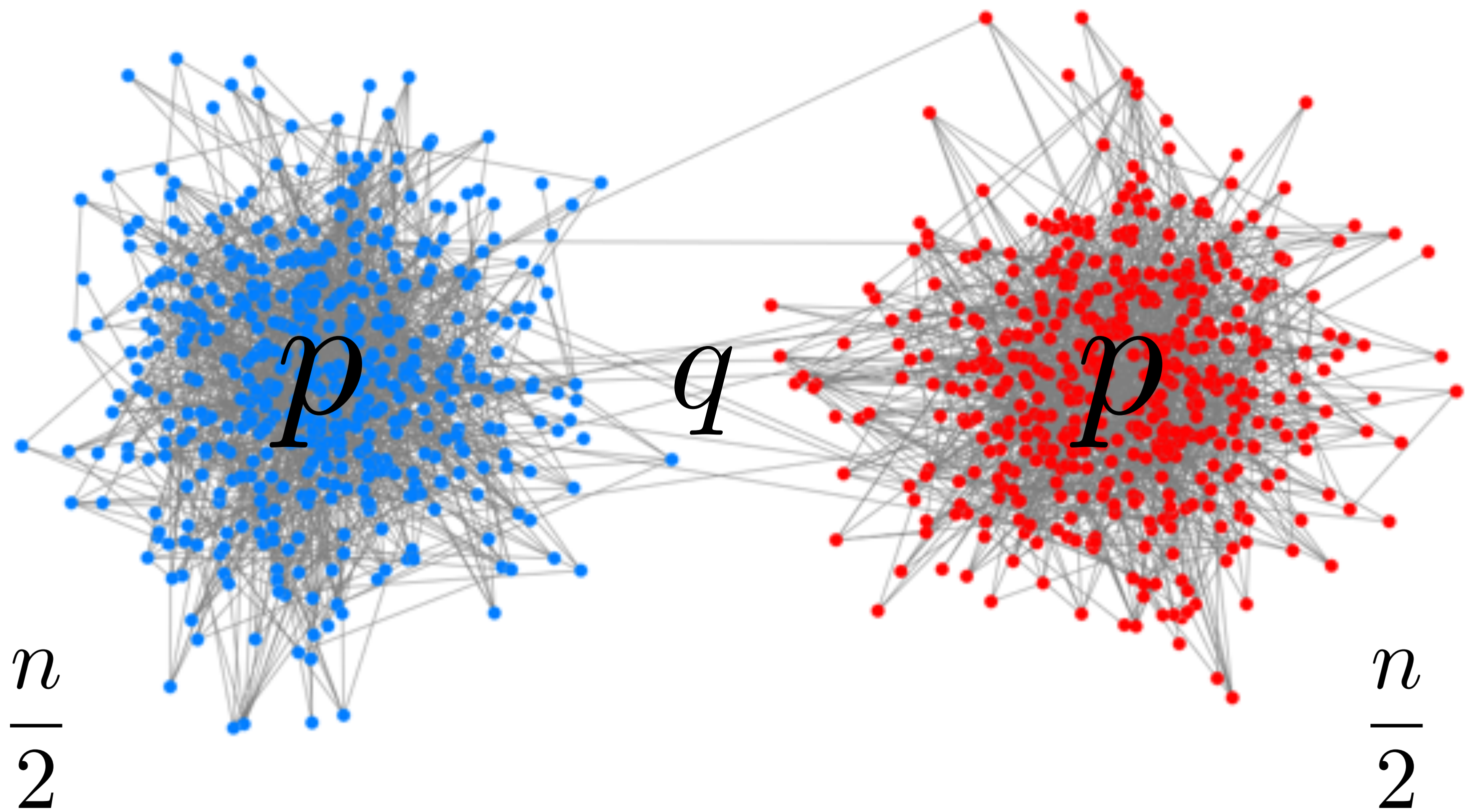
Definition. An algorithm $\hat{X}^n(\cdot)$ solves (exact) recovery in the SBM if for a random graph G under the model, $\lim_{n\to\infty} \mathbb{P}(X^n = \hat{X}^n(G)) = 1$.

We will see weaker recovery requirements later

Starting point:
progress in science often comes from understanding special cases...

SBM with 2 symmetric communities: 2-SBM

2-SBM



$$p_1 = p_2 = 1/2$$

$$W_{11} = W_{22} = p \qquad W_{12} = q$$

Some history for 2-SBM

Recovery problem

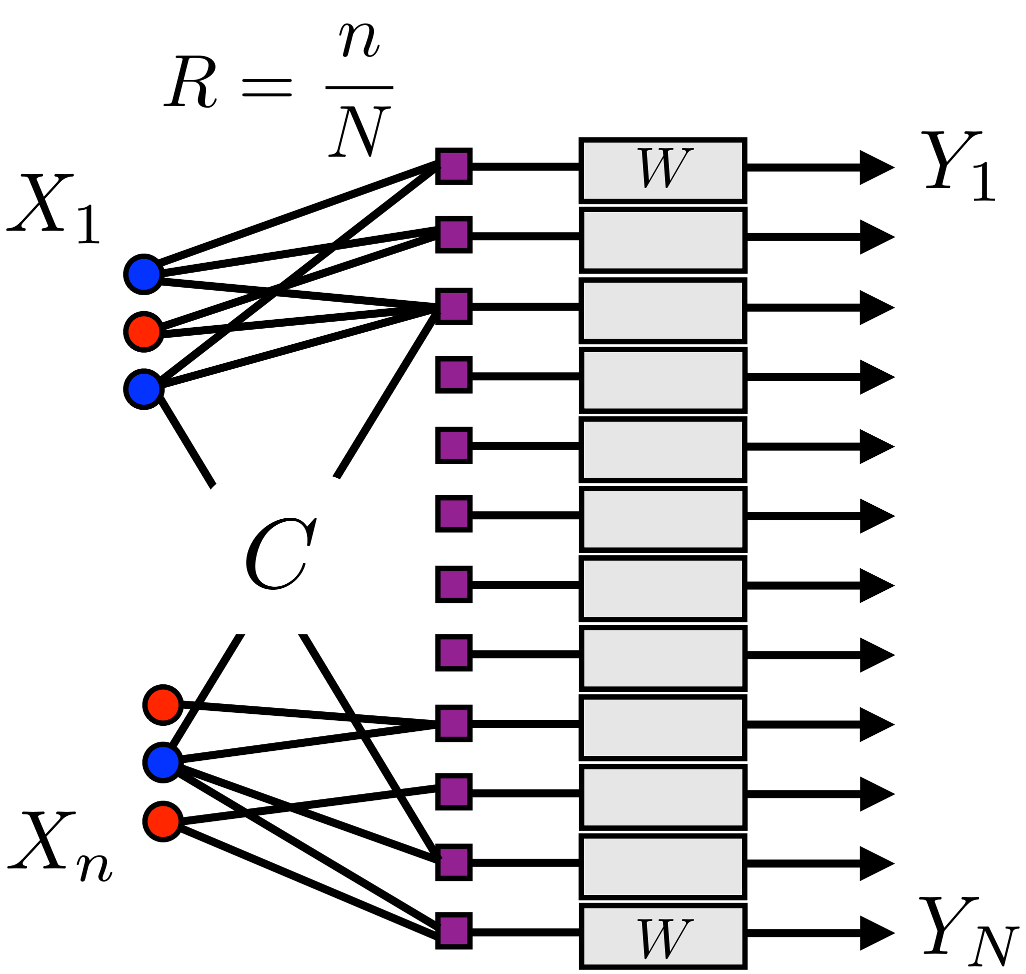
$$\mathbb{P}(\hat{X}^n = X^n) \to 1$$

Bui, Chaudhuri, Leighton, Sipser '84	maxflow-mincut	$p = \Omega(1/n), q = o(n^{-1-4/((p+q)n)})$
Boppana '87	spectral meth.	$(p-q)/\sqrt{p+q} = \Omega(\sqrt{\log(n)/n})$
Dyer, Frieze '89	min-cut via degrees	$p - q = \Omega(1)$
Snijders, Nowicki '97	EM algo.	$p - q = \Omega(1)$
Jerrum, Sorkin '98	Metropolis aglo.	$p - q = \Omega(n^{-1/6+\epsilon})$
Condon, Karp '99	augmentation algo.	$p - q = \Omega(n^{-1/2+\epsilon})$
Carson, Impagliazzo '01	hill-climbing algo.	$p - q = \Omega(n^{-1/2} \log^4(n))$
Mcsherry '01	spectral meth.	$(p-q)/\sqrt{p} \geq \Omega(\sqrt{\log(n)/n})$
Bickel, Chen '09	N-G modularity	$(p-q)/\sqrt{p+q} = \Omega(\log(n)/\sqrt{n})$
Rohe, Chatterjee, Yu '11	spectral meth.	$p - q = \Omega(1)$

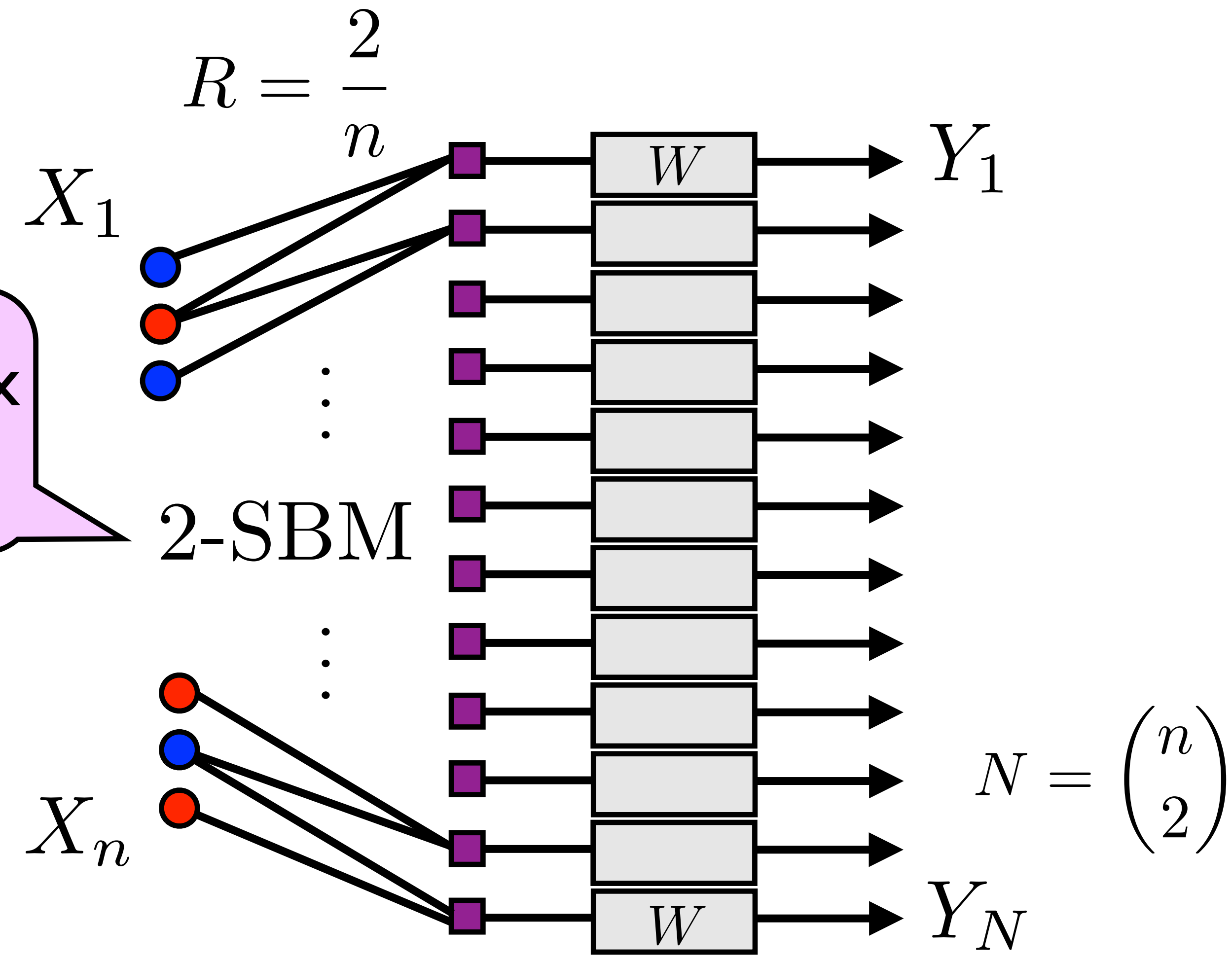
algorithms driven...

Instead of 'how', when can we recover the clusters (IT)?

Information-theoretic view of clustering



unorthodox code!



$$W = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}$$

reliable comm. iff R < 1-H(**ε**)

$$W = \begin{pmatrix} 1-p & p \\ 1-q & q \end{pmatrix}$$

reliable comm. iff ???

⇕

exact recovery

Some history for 2-SBM

Recovery problem

$$\mathbb{P}(\hat{X}^n = X^n) \to 1$$

1983 2010 2014

Holland
Laskey
Leinhardt

Boppana

Dyer
Frieze

Bui, Chaudhuri,
Leighton, Sipser

Snijders
Nowicki

Jerrum
Sorkin

Condon
Karp

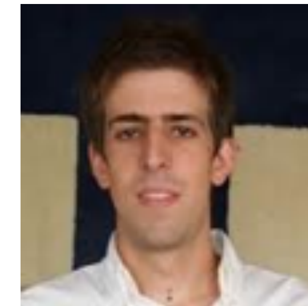
Carson
Impagliazzo

McSherry

Bickel
Chen

Rohe
Chatterjee
Yu

Abbe-Bandeira-Hall

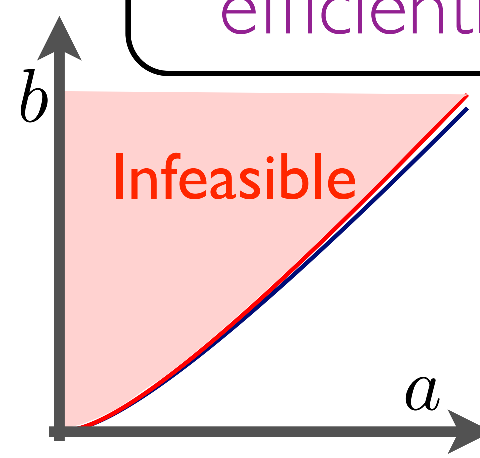



Bui, Chaudhuri, Leighton, Sipser '84	maxflow-mincut	$p=\Omega(1/n), q=o(n^{-1-4/((p+q)n)})$
Boppana '87	spectral meth.	$(p-q)/\sqrt{p+q}=\Omega(\sqrt{\log(n)/n})$
Dyer, Frieze '89	min-cut via degrees	$p-q=\Omega(1)$
Snijders, Nowicki '97	EM algo.	$p-q=\Omega(1)$
Jerrum, Sorkin '98	Metropolis aglo.	$p-q=\Omega(n^{-1/6+\epsilon})$
Condon, Karp '99	augmentation algo.	$p-q=\Omega(n^{-1/2+\epsilon})$
Carson, Impagliazzo '01	hill-climbing algo.	$p-q=\Omega(n^{-1/2}\log^4(n))$
Mcsherry '01	spectral meth.	$(p-q)/\sqrt{p}\geq\Omega(\sqrt{\log(n)/n})$
Bickel, Chen '09	N-G modularity	$(p-q)/\sqrt{p+q}=\Omega(\log(n)/\sqrt{n})$
Rohe, Chatterjee, Yu '11	spectral meth.	$p-q=\Omega(1)$

$$p = \frac{a\log(n)}{n}, q = \frac{b\log(n)}{n}$$

Recovery iff $\frac{a+b}{2} \geq 1+\sqrt{ab}$

efficiently achievable



Some history for 2-SBM

Recovery problem

$$\mathbb{P}(\hat{X}^n = X^n) \to 1$$

Detection problem

$$\exists \epsilon > 0 : \mathbb{P}\left(\frac{d(\hat{X}^n, X^n)}{n} < \frac{1}{2} - \epsilon\right) \to 1$$

1983 — 2010 — 2014

Holland, Laskey, Leinhardt

Bui, Chaudhuri, Leighton, Sipser

Boppana

Dyer, Frieze

Snijders, Nowicki

Jerrum, Sorkin

Condon, Karp

Carson, Impagliazzo

McSherry

Bickel, Chen

Rohe, Chatterjee, Yu

Decelle, Krzakala, Moore, Zdeborova

Coja-Oghlan

Massoulié

Mossel, Neeman, Sly

Mossel-Neeman-Sly

Abbe-Bandeira-Hall

Bui, Chaudhuri, Leighton, Sipser '84	maxflow-mincut	$p = \Omega(1/n), q = o(n^{-1-4/((p+q)n)})$
Boppana '87	spectral meth.	$(p-q)/\sqrt{p+q} = \Omega(\sqrt{\log(n)/n})$
Dyer, Frieze '89	min-cut via degrees	$p - q = \Omega(1)$
Snijders, Nowicki '97	EM algo.	$p - q = \Omega(1)$
Jerrum, Sorkin '98	Metropolis aglo.	$p - q = \Omega(n^{-1/6+\epsilon})$
Condon, Karp '99	augmentation algo.	$p - q = \Omega(n^{-1/2+\epsilon})$
Carson, Impagliazzo '01	hill-climbing algo.	$p - q = \Omega(n^{-1/2}\log^4(n))$
Mcsherry '01	spectral meth.	$(p-q)/\sqrt{p} \geq \Omega(\sqrt{\log(n)/n})$
Bickel, Chen '09	N-G modularity	$(p-q)/\sqrt{p+q} = \Omega(\log(n)/\sqrt{n})$
Rohe, Chatterjee, Yu '11	spectral meth.	$p - q = \Omega(1)$

$$p = \frac{a\log(n)}{n}, q = \frac{b\log(n)}{n}$$

Recovery iff $\frac{a+b}{2} \geq 1 + \sqrt{ab}$

efficiently achievable

$$p = \frac{a}{n}, q = \frac{b}{n}$$

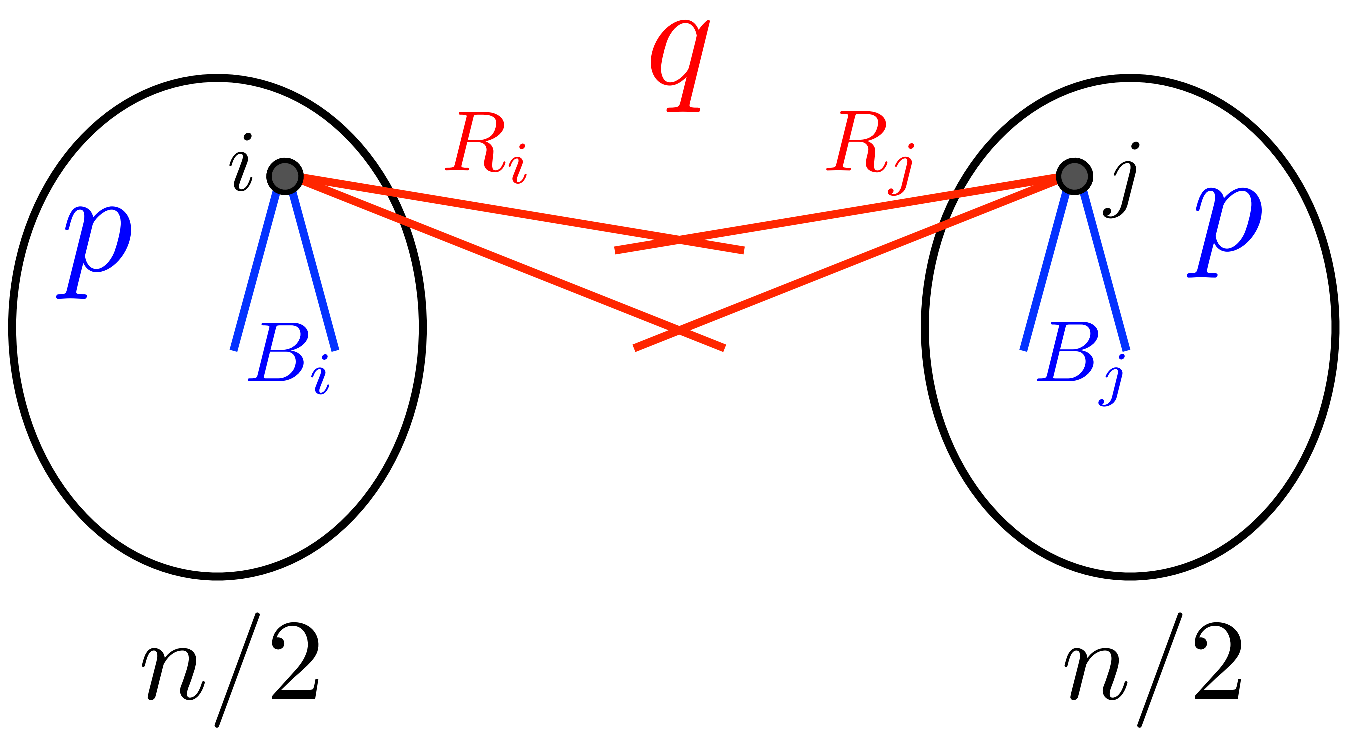
Detection iff $(a-b)^2 > 2(a+b)$

What about multiple/asymm. communities?

Conjecture: detection changes with 5 or more

Recovery in the 2-SBM: IT limit

Converse: If $\frac{a+b}{2} - \sqrt{ab} < 1$ then ML fails w.h.p.



$$p = \frac{a \log n}{n}, q = \frac{b \log n}{n}$$

what is ML? $\rightarrow$ min-bisection

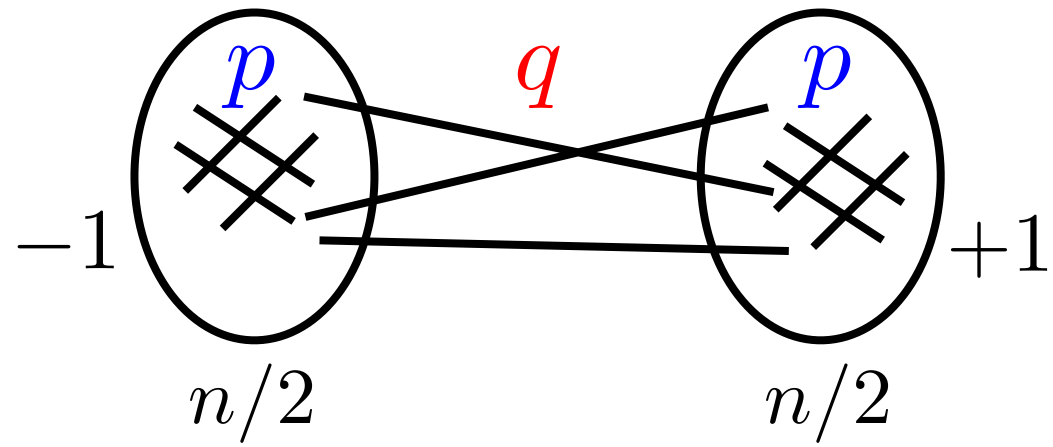
ML fails if two nodes can be swapped to reduce the cut

$$P(\exists i : B_i \leq R_i) = \; ? \; \asymp n\, P(B_1 \leq R_1) \text{ (weak correlations)}$$

$$P(B_1 \leq R_1) = n^{-((a+b)/2 - \sqrt{ab}) + o(1)}$$

Abbe-Bandeira-Hall '14

Recovery in the 2-SBM: efficient algorithms



lifting:

$X = xx^t$

spectral:

$$\max\ x^T A x$$
$$\text{s.t. } \|x\| = 1$$
$$1^t x = 0$$

ML-decoding:

$$\max\ x^T A x$$
$$\text{s.t. } x_i = \pm 1$$
$$1^t x = 0$$

NP-hard

SDP:

$$\max\ \mathrm{tr}(AX)$$
$$\text{s.t. } X_{ii} = 1$$
$$1^t X = 0$$
$$X \succeq 0$$
$$\cancel{\mathrm{rank}(X) = 1}$$

Abbe-Bandeira-Hall '14

Recovery in the 2-SBM: efficient algorithms

Theorem. The SDP solves recovery if $2L_{\mathrm{SBM}} + 11^t + I_n \succeq 0$ where $L_{\mathrm{SBM}} = D_{G_+} - D_{G_-} - A$.

-> Analyze the spectral norm of a random matrix
[Abbe-Bandeira-Hall '14] Bernstein: slightly loose

[Xu-Hanjek-Wu '15] Seginer bound
[Bandeira, Bandeira-Van Handel '15] tight bound

Note that SDP can be expensive...

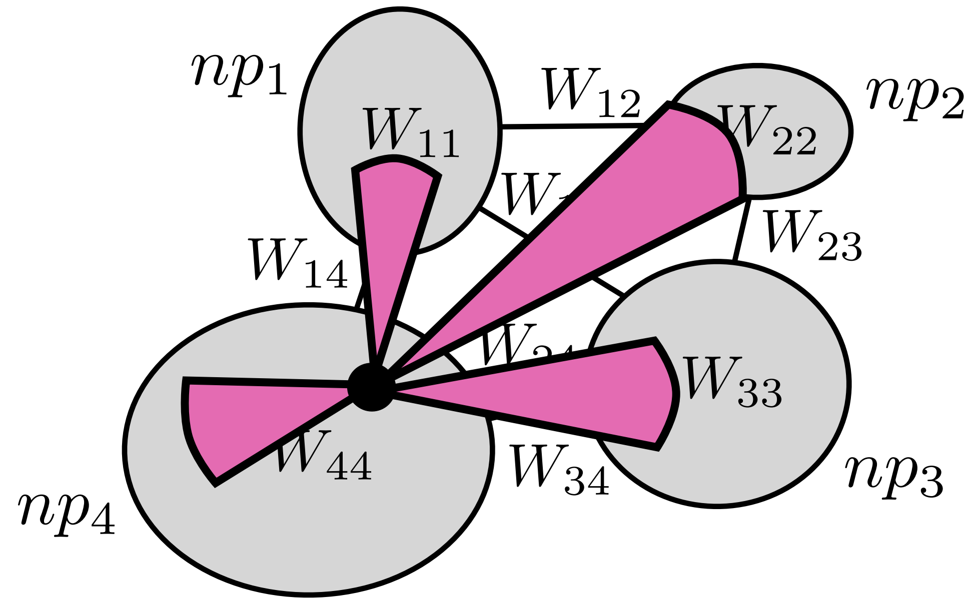
Abbe-Bandeira-Hall '14

The general SBM

$\mathrm{SBM}(n, p, W)$

Quiz: If a node is in community i, how many neighbors does it have in expectation in community j ?

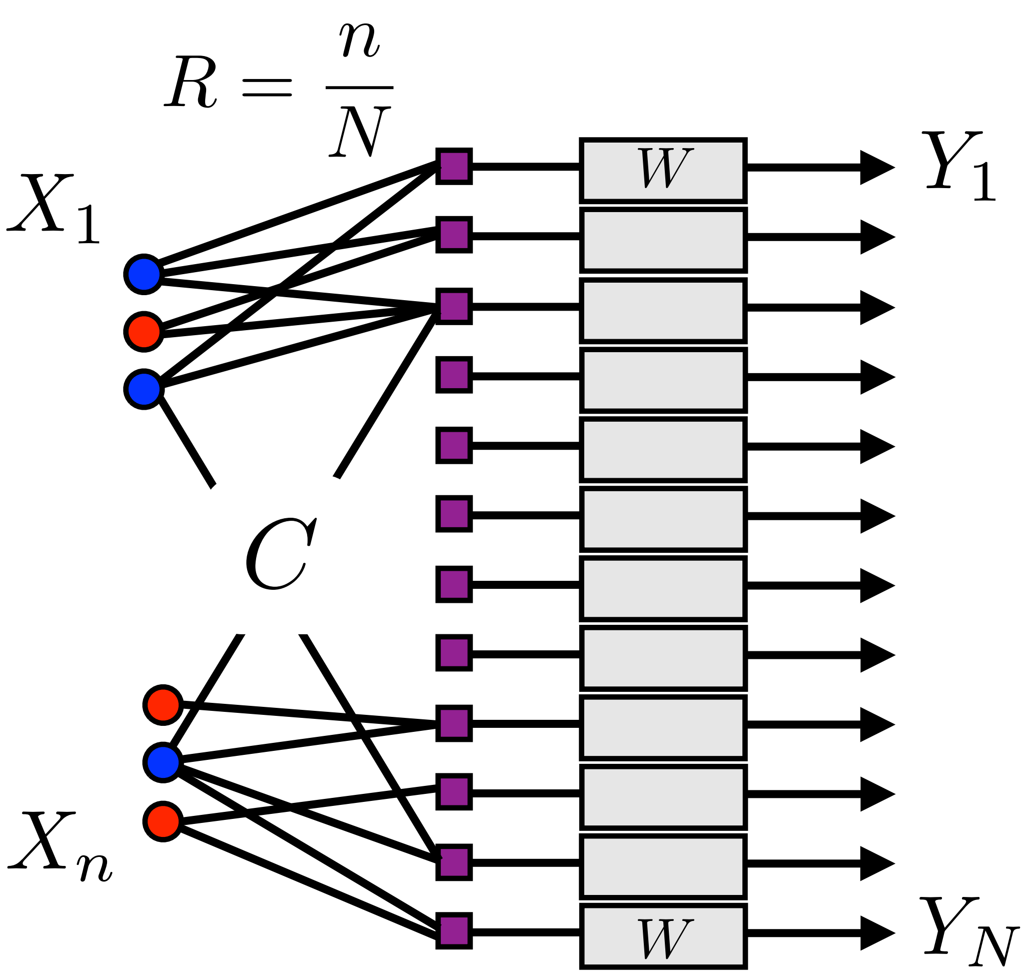
1. np_j
2. $np_j W_{ij}$
3. $np_i W_{ij}$
4. 7

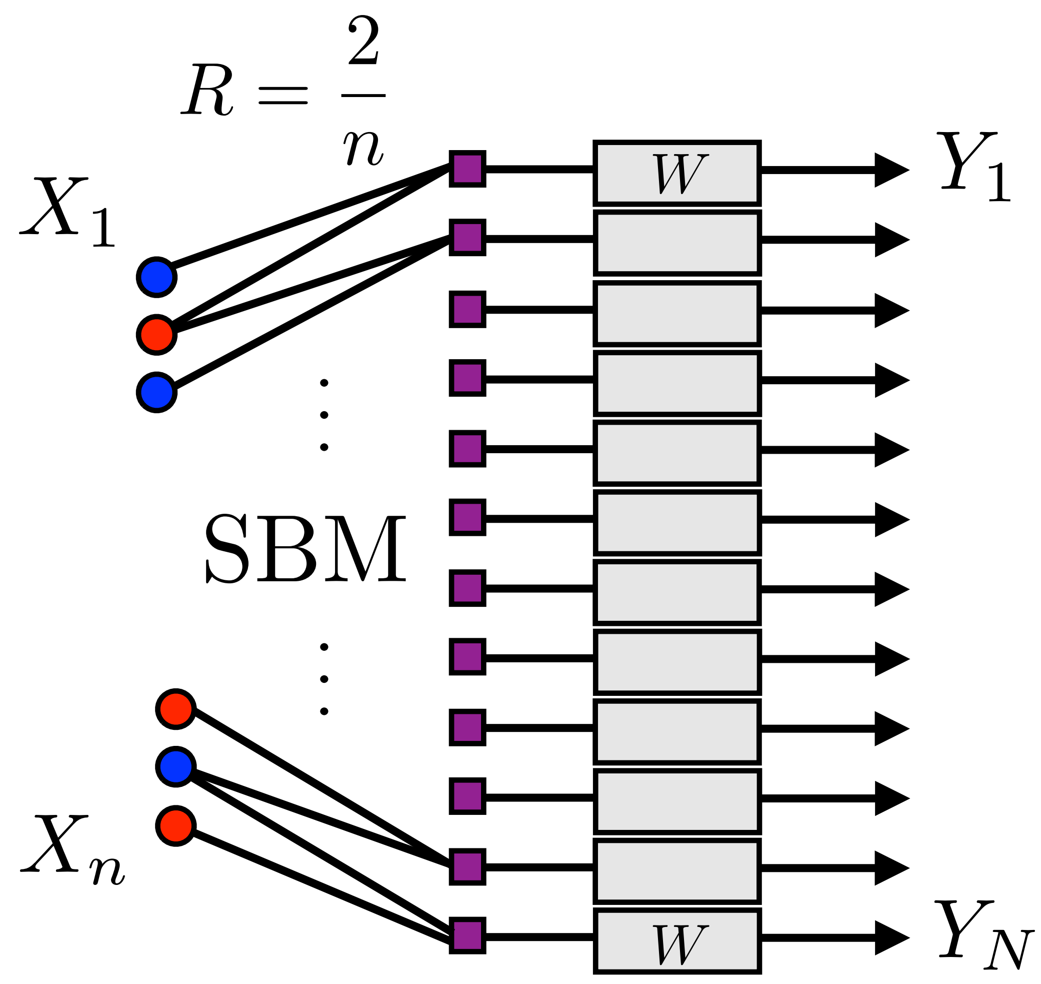


$$\begin{array}{cc} & i \\ \Bigg(nPW & \boxed{} \Bigg) \end{array}$$

"degree profile matrix"

Back to the Information-theoretic view of clustering





$$\left(W \right) \qquad\qquad \left(W \right)$$

reliable comm. iff R < 1-H(**ε**)

reliable comm. iff 1 < (a+b)/2-$\sqrt{ab}$

reliable comm. iff R < $\max_p \underbrace{I(p,W)}_{\text{KL-divergence}}$

reliable comm. iff 1 < J(p,W) ???

Main results

Theorem 1. Recovery is solvable in $\mathrm{SBM}(n, p, Q\log(n)/n)$ if and only if

$$J(p,Q) := \min_{i<j} D_+((PQ)_i, (PQ)_j) \geq 1$$

where

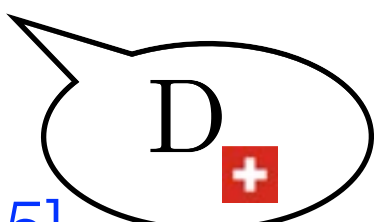
$$D_+(\mu,\nu) := \max_{t\in[0,1]} \underbrace{\sum_{\ell\in[k]} \left(t\mu_\ell + (1-t)\nu_\ell - \mu_\ell^t \nu_\ell^{1-t}\right)}_{D_t(\mu,\nu)}$$

$\frac{1}{2}(\sqrt{a}-\sqrt{b})^2 \geq 1 \longleftarrow$

Abbe-Bandeira-Hall '14

- $D_{1/2}(\mu,\nu) = \frac{1}{2}\|\sqrt{\mu}-\sqrt{\nu}\|_2^2$ is the Hellinger divergence (distance)
- D_t is an f-divergence: $\sum_i \nu_i f(\mu_i/\nu_i)$ $\qquad f(x) = 1 - t + tx - x^t$
- $-\log \max_t \sum_i \mu_i^t \nu_i^t$ is the Chernoff divergence

We call D_+ the CH-divergence.



[Abbe-Sandon '15]

Main results

Theorem 1. Recovery is solvable in $\mathrm{SBM}(n, p, Q\log(n)/n)$ if and only if

$$J(p,Q) := \min_{i<j} D_+((PQ)_i, (PQ)_j) \geq 1$$

where

$$D_+(\mu,\nu) := \max_{t\in[0,1]} \sum_{\ell\in[k]} \left(t\mu_\ell + (1-t)\nu_\ell - \mu_\ell^t \nu_\ell^{1-t}\right)$$

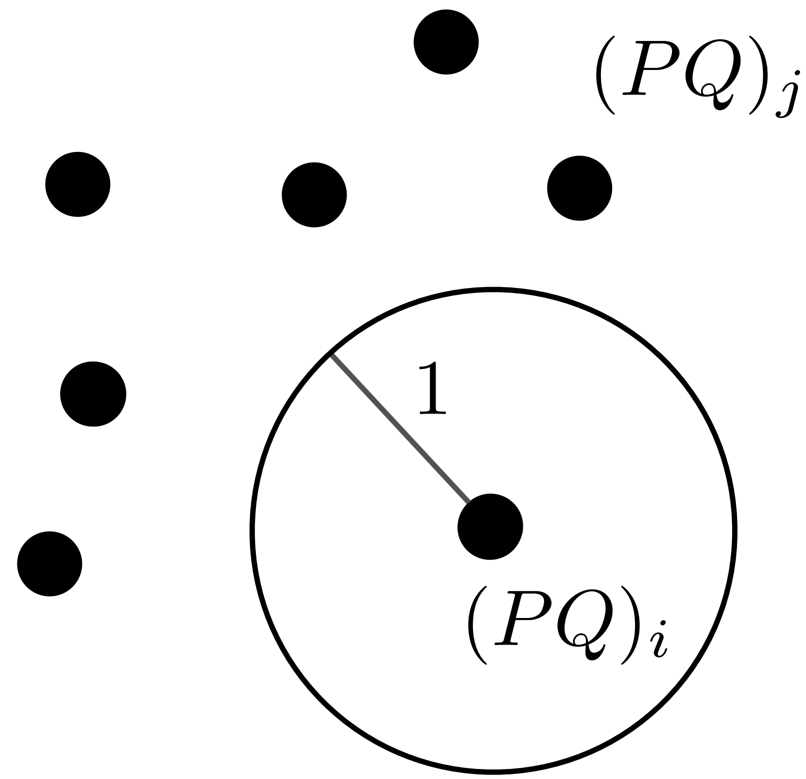
Is recovery in the general SBM solvable efficiently down the information theoretic threshold? **YES!**

Theorem 2. The `degree-profiling` algorithm achieves the threshold and runs in quasi-linear time.

[Abbe-Sandon '15]

When can we extract a specific community?

Theorem. If community i has a profile $(PQ)_i$ at D_+-distance at least 1 from all other profiles $(PQ)_j$, $j \neq i$, then it can be extracted w.h.p.



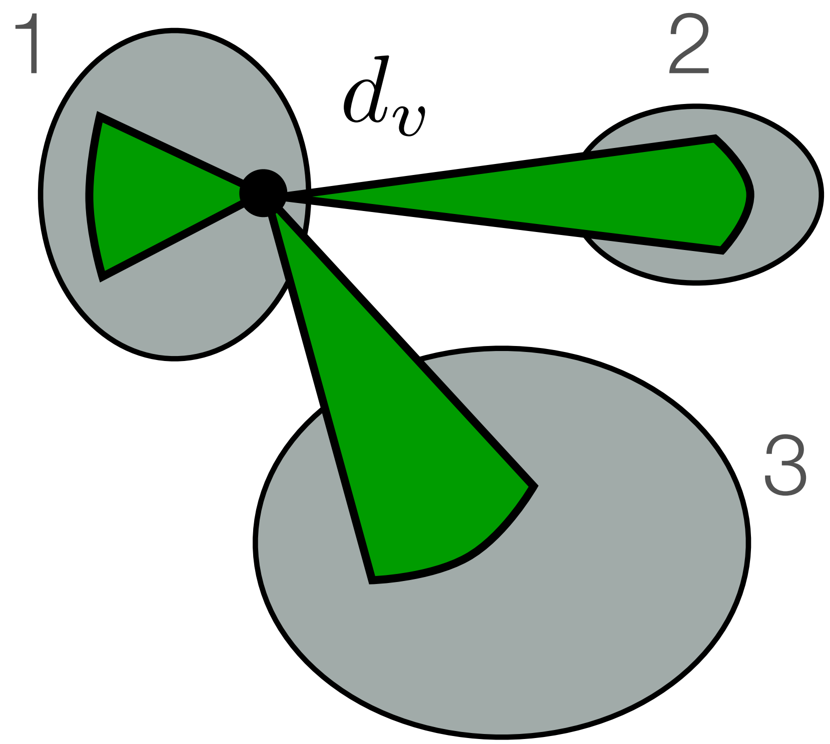
What if we do not know the parameters?

We can learn them on the fly: [Abbe-Sandon '15] (second paper)

[Abbe-Sandon '15]

Proof techniques and algorithms

A key step

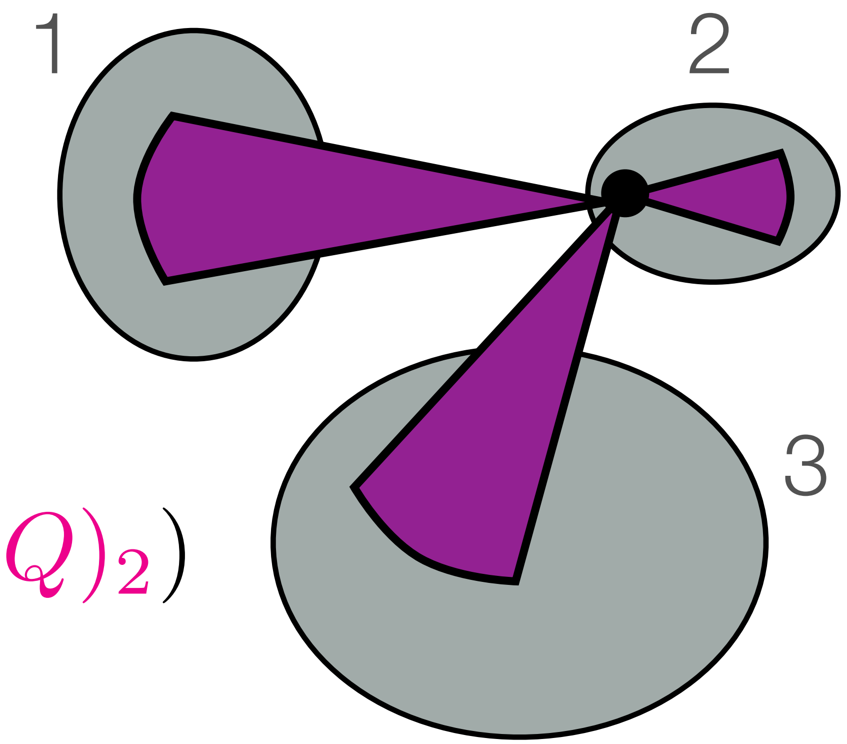


Hypothesis 1

$$d_v \dot{\sim} \mathcal{P}(\log(n)(PQ)_1)$$

Hypothesis 2

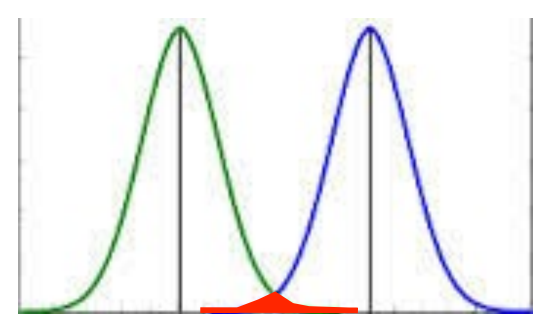
$$d_v \dot{\sim} \mathcal{P}(\log(n)(PQ)_2)$$



Theorem. For any $\theta_1, \theta_2 \in (\mathbb{R}_+ \setminus \{0\})^k$ with $\theta_1 \neq \theta_2$ and $p_1, p_2 \in \mathbb{R}_+ \setminus \{0\}$,

$$\sum_{x \in \mathbb{Z}_+^k} \min(\mathcal{P}_{\ln(n)\theta_1}(x)p_1, \mathcal{P}_{\ln(n)\theta_2}(x)p_2) = \Theta\left(n^{-D_+(\theta_1,\theta_2)-o(1)}\right),$$

where D_+ is the CH-divergence.



[Abbe-Sandon '15]

How to use this?

Plan:

put effort in recovering most of the nodes and
then finish greedily with local improvements

[Abbe-Sandon '15]

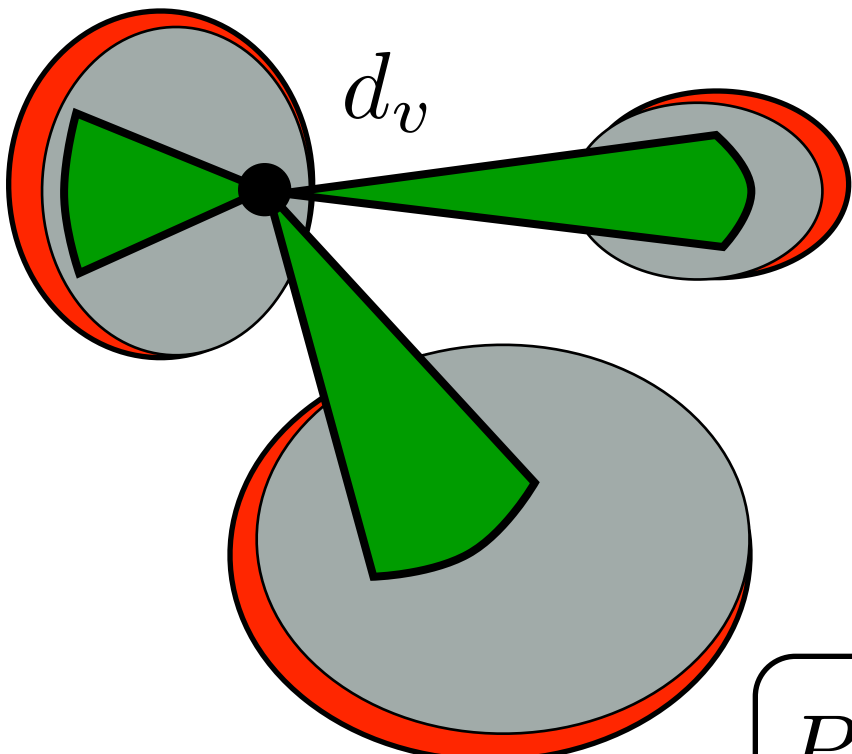
The degree-profiling algorithm (capacity-achieving)

(1) Split G into two graphs → G' loglog-degree, G'' log-degree

(2) Run `Sphere-comparison` on G'

-> gets a fraction 1-o(1) (see next)

(3) Take now G'' with the clustering of G'

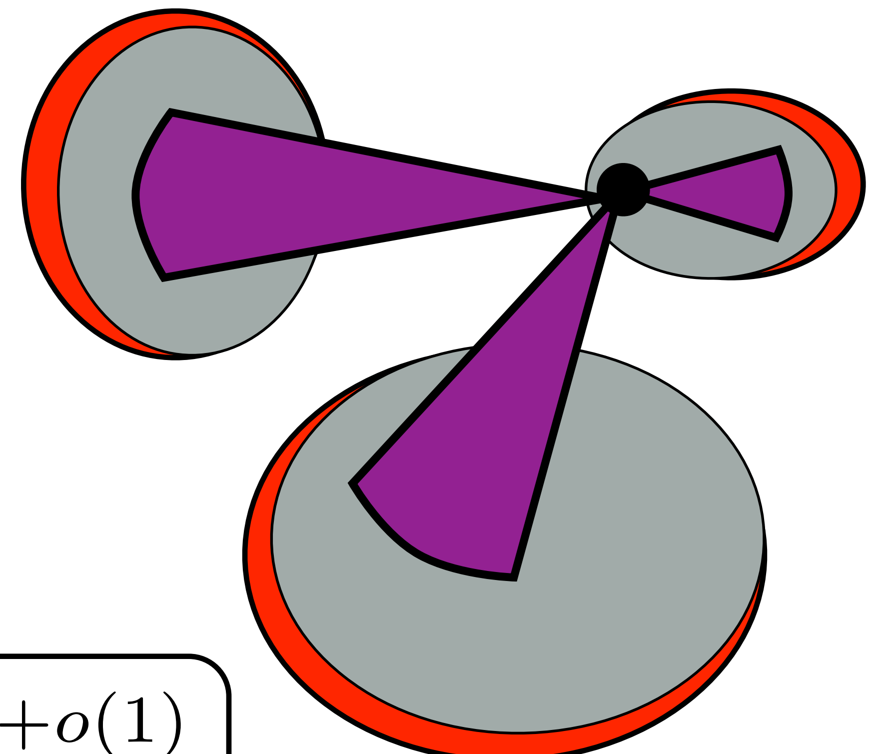


Hypothesis 1

$$d_v \overset{\cdot\cdot\cdot}{\sim} \mathcal{P}(\log(n)(PQ)_1)$$

Hypothesis 2

$$d_v \overset{\cdot\cdot\cdot}{\sim} \mathcal{P}(\log(n)(PQ)_2)$$



$$\boxed{P_e = n^{-D_+((PQ)_1,(PQ)_2)+o(1)}}$$

[Abbe-Sandon '15]

How do we get most nodes correctly?

Other recovery requirements

Weak recovery or detection : $c = 1/k + \epsilon$ for some $\epsilon > 0$ (for the symmetric k-SBM).

Partial recovery: An algorithm solves partial recovery in the SBM with accuracy c if it produces a clustering which is correct on a fraction c of the nodes with high probability.

Almost exact recovery: $c = 1 - o(1)$

Exact recovery: $c = 1$

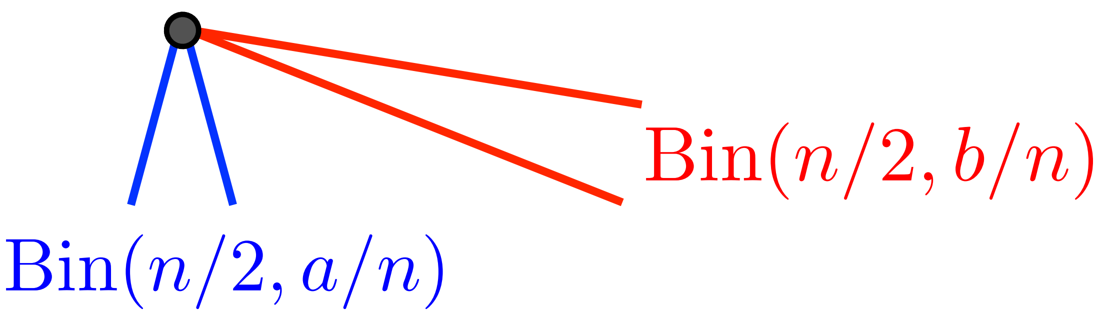
For all the above: what are the "efficient" VS. "information-theoretic" fundamental limits?

Partial recovery in $\mathrm{SBM}(n, p, Q/n)$

What is a good notion of SNR?

e.v. of PQ

Proposed notion of SNR: $\dfrac{|\lambda_{\min}|^2}{\lambda_{\max}}$



$$= \frac{(a-b)^2}{2(a+b)} \text{ for 2-symm. comm.}$$

$$= \frac{(a-b)^2}{k(a+(k-1)b)} \text{ for } k\text{-symm. comm.}$$

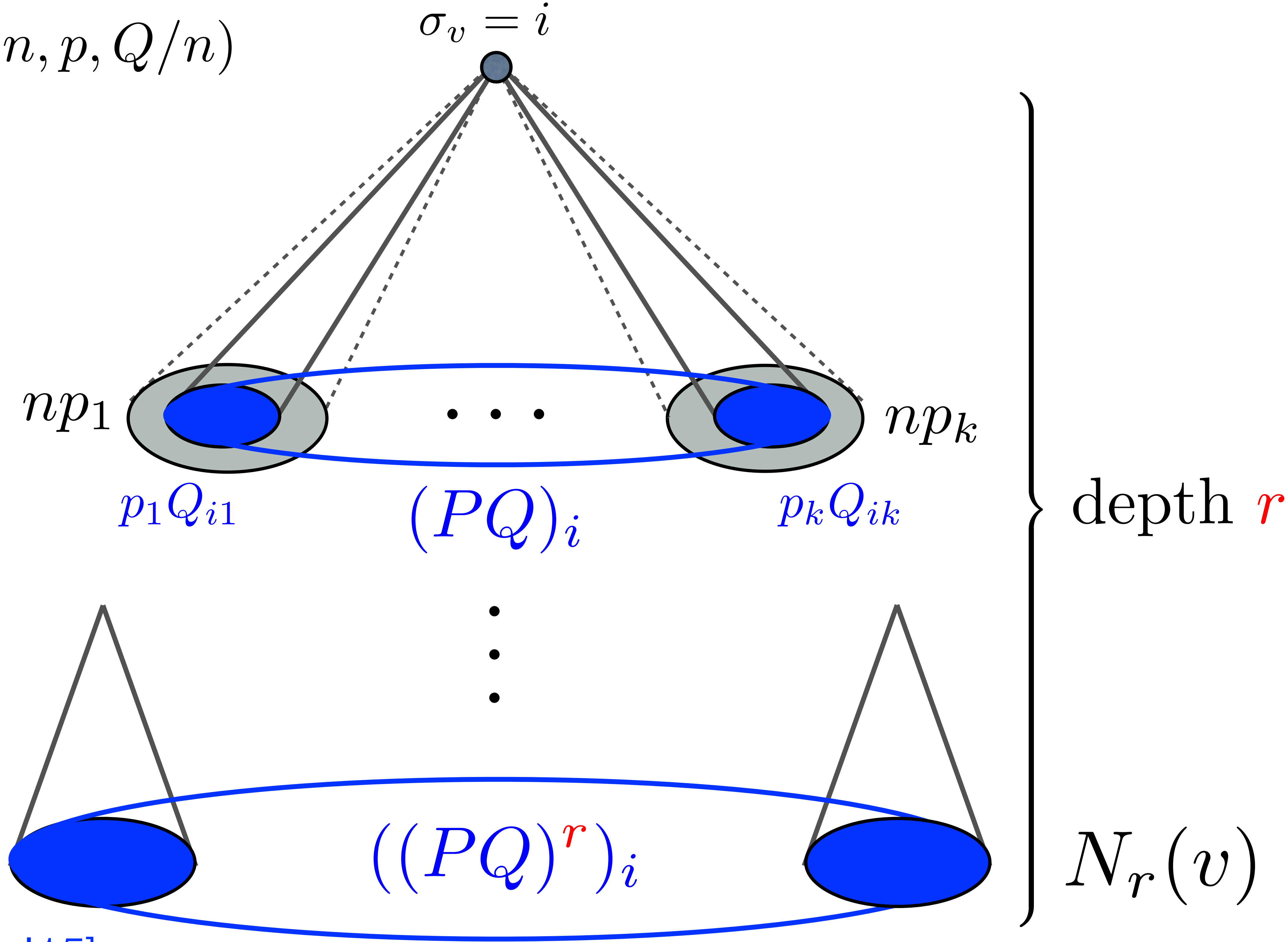
Theorem (informal). In the sparse $\mathrm{SBM}(n, p, Q/n)$, the `Sphere-comparison` algorithm recovers a fraction of nodes which approaches 1 when the SNR diverges.

Note that the SNR scales if Q scales!

[Abbe-Sandon '15]

Sphere-comparison

A node neighborhood in $\mathrm{SBM}(n, p, Q/n)$

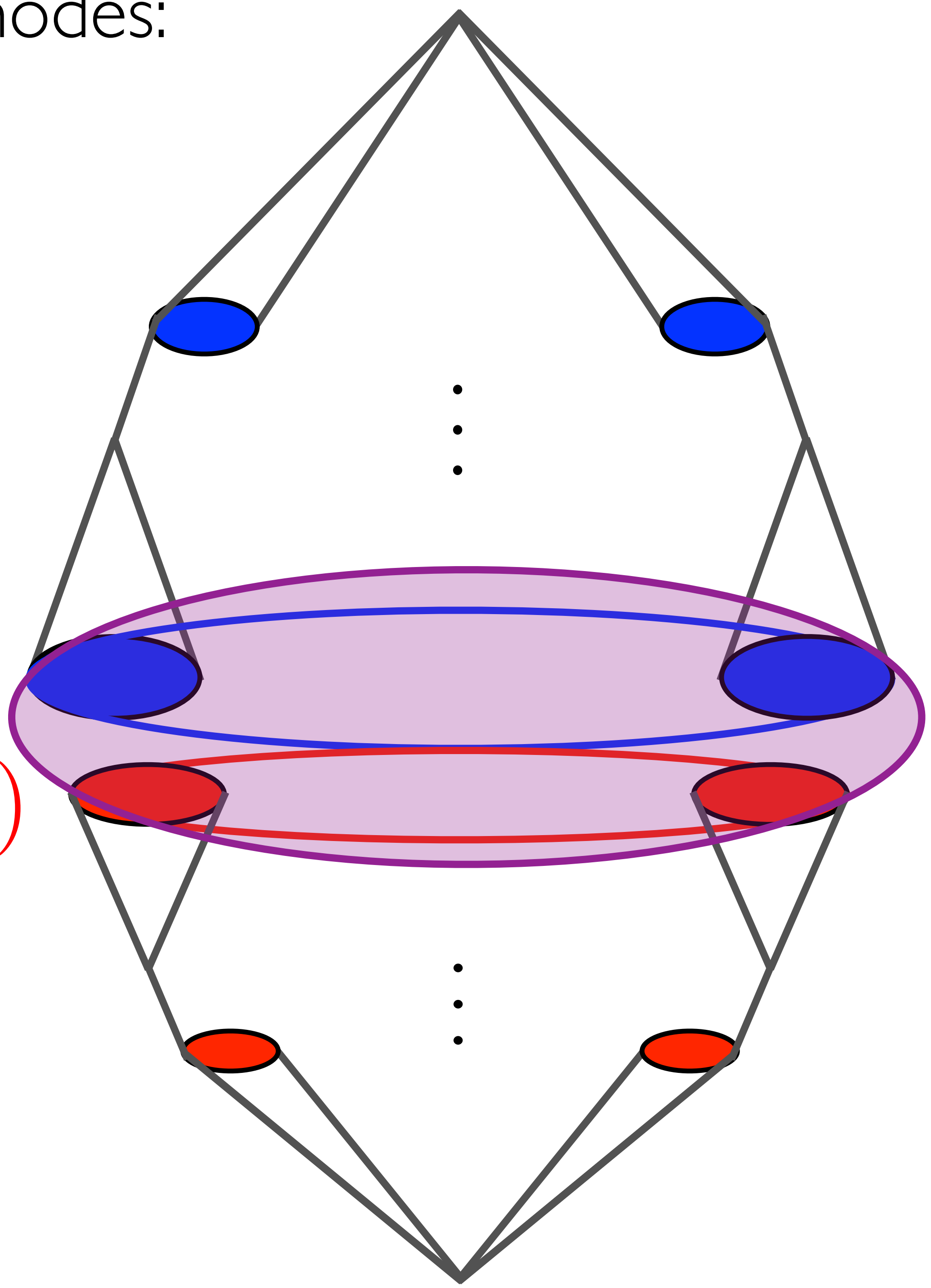


[Abbe-Sandon '15]

Sphere-comparison

Take now two nodes:

$\sigma_v = i$



$N_r(v)$

$N_{r'}(v')$

$\sigma_{v'} = j$

Compare v and v' from:

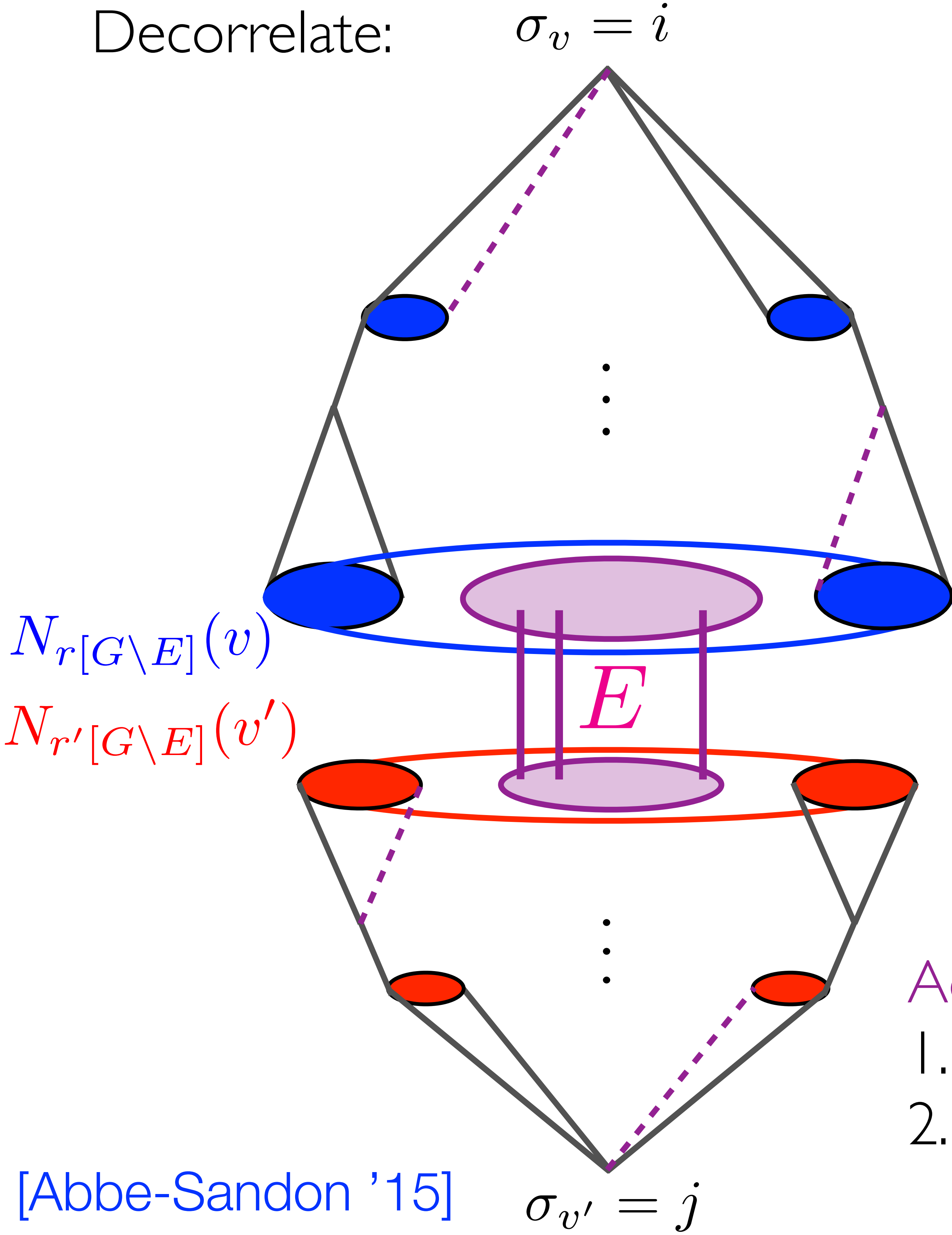
$$|N_r(v) \cap N_{r'}(v')|$$

hard to analyze...

[Abbe-Sandon '15]

Sphere-comparison

Decorrelate:



Subsample G with prob. c to get E

Compare v and v' from:

$$N_{r,r'[E]}(v \cdot v')$$

= number of crossing edges

$$\approx N_{r[G\setminus E]}(v) \cdot \frac{cQ}{n} N_{r'[G\setminus E]}(v')$$

$$\approx ((1-c)PQ)^r e_{\sigma_v} \cdot \frac{cQ}{n}((1-c)PQ)^{r'} e_{\sigma_{v'}}$$

$$= c(1-c)^{r+r'} e_{\sigma_v} \cdot Q(PQ)^{r+r'} e_{\sigma_{v'}}/n$$

Additional steps:

1. look at several depths -> Vandermonde syst.
2. use "anchor nodes"

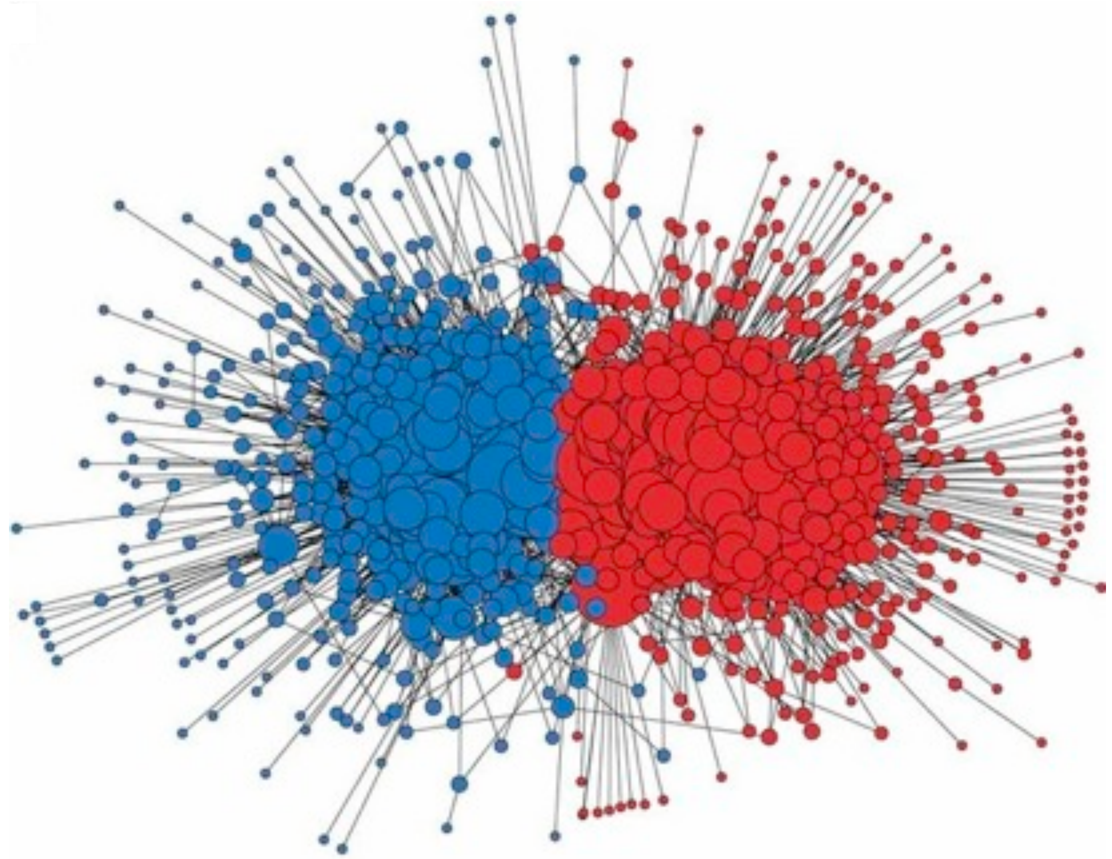
[Abbe-Sandon '15]

A real data example

The political blogs network

1222 blogs
(left- and right-leaning)
[Adamic and Glance '05]

edge = hyperlink between blogs



The CH-divergence is close to 1

We can recover 95% of the nodes correctly

Some open problems in community detection

Some open problems in community detection

I. The SBM:

a. Recovery

Growing nb. of communities? Sub-linear communities?

[Abbe-Sandon '15] should extend to $k = o(\log(n))$

[Chen-Xu '14] k,p,q scale with n polynomially

Some open problems in community detection

I. The SBM:

a. Recovery

b. Detection and broadcasting on trees

[Mossel-Neeman-Sly '13] Converse for detection in 2-SBM:

Galton-Watson tree

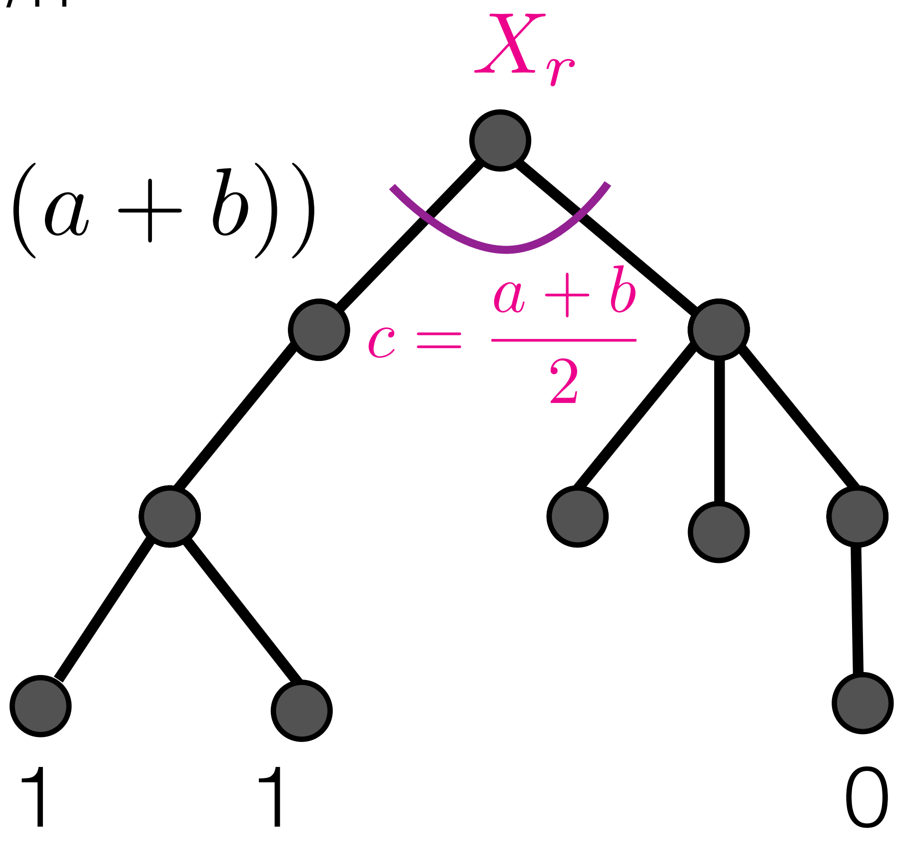
Poisson((a+b)/2)

If (a+b)/2<=1 the tree dies w.p. 1

If (a+b)/2>1 the tree survives w.p. >0

p=**a**/n, q=**b**/n

$\mathrm{BSC}(b/(a+b))$



Unorthodox broadcasting problem:

when can we detect the root-bit?

If and only if $c > \frac{1}{(1-2\varepsilon)^2}$

[Evans-Kenyon-Peres-Schulman '00]

$$\iff \quad \frac{(a-b)^2}{2(a+b)} > 1$$

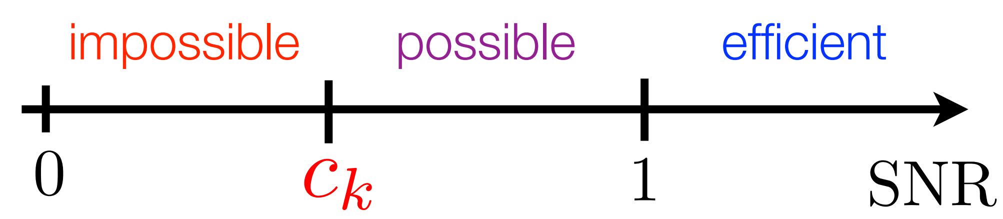
Some open problems in community detection

I. The SBM:

a. Recovery

b. Detection and broadcasting on trees

For k clusters? $\mathrm{SNR} = \dfrac{(a-b)^2}{k(a-(k-1)b)}$



Conjecture. For the symmetric k-SBM(n,a,b), there exists c_k s.t.
(1) If $\mathrm{SNR} < c_k$, then detection cannot be solved,
(2) If $c_k < \mathrm{SNR} < 1$, then detection can be solved information-theoretically but not efficiently,
(3) If $\mathrm{SNR} > 1$, then detection can be solved efficiently.

Moreover $c_k = 1$ for $k \in \{2,3,4\}$ and $c_k < 1$ for $k \geq 5$.

[Decelle-Krzakala-Zdeborova-Moore '11]

Some open problems in community detection

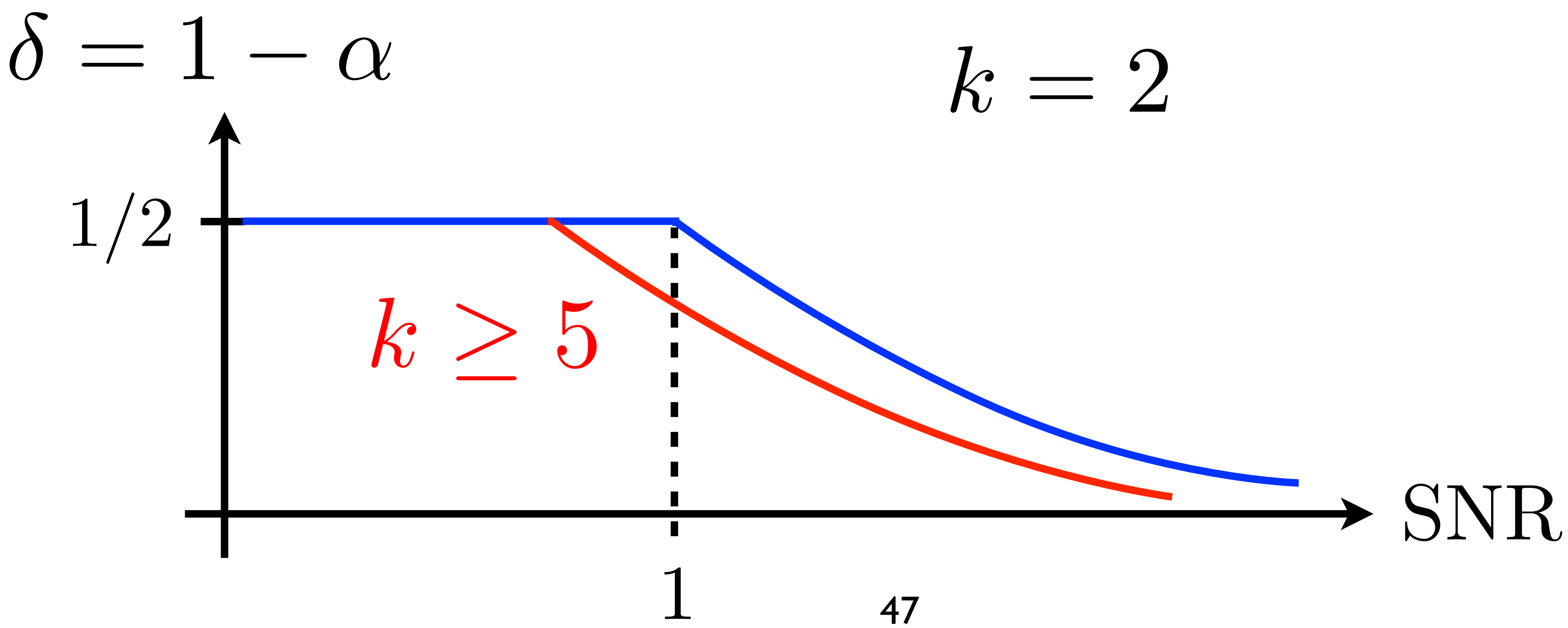
I. The SBM:

a. Recovery

b. Detection and broadcasting on trees

c. Partial recovery and the SNR-distortion curve

Conjecture. For the symmetric k-SBM(n, a, b) and $\alpha \in (1/k, 1)$, there exists β_k, γ_k s.t. partial-recovery of accuracy α is solvable if and only if $\text{SNR} > \beta_k$, and efficiently solvable iff $\text{SNR} > \gamma_k$.

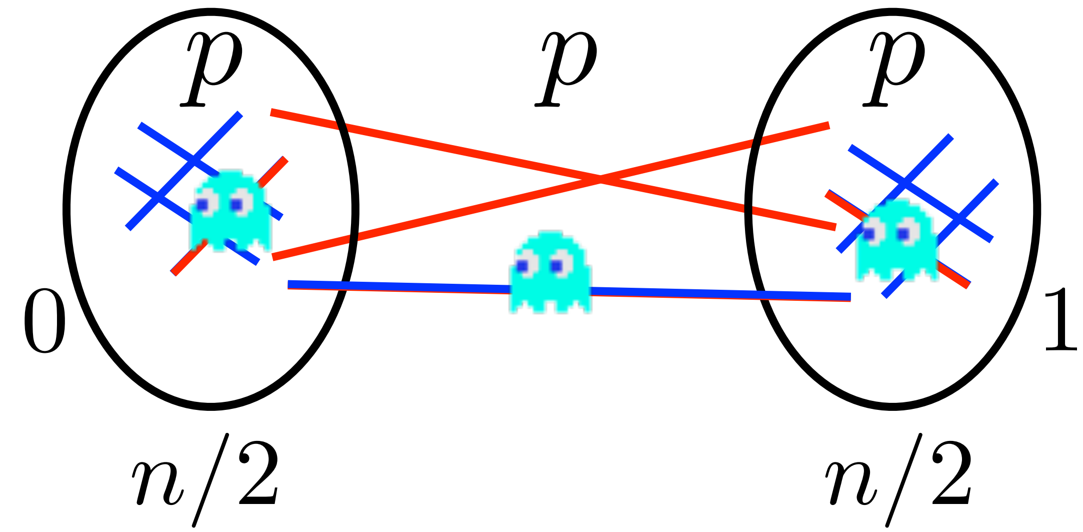


Some open problems in community detection

II. Other block models

- Censored block models
- Labelled block models

$$2\text{-CBM}(n, p, \epsilon)$$



[Abbe-Bandeira-Bracher-Singer '14]

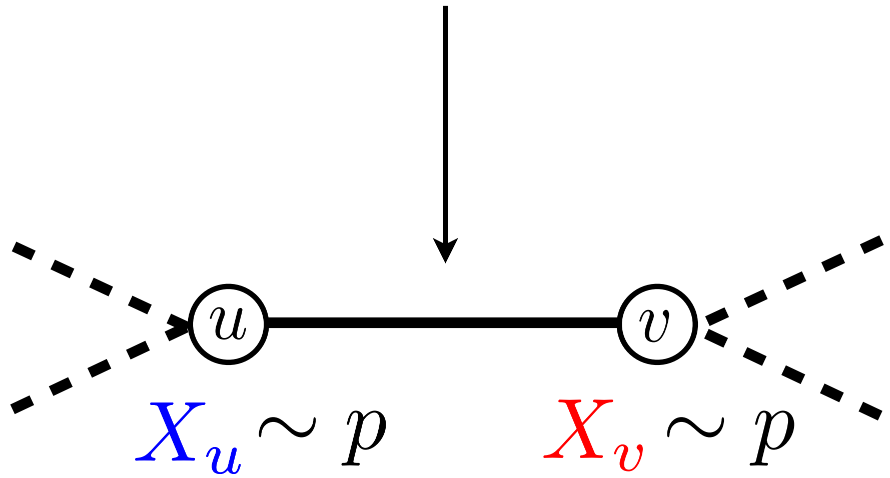
2-CBM
[Abbe-Bandeira-Bracher-Singer,
Xu-Lelarge-Massoulie,
Saad-Krzakala-Lelarge-Zdeborova,
Chin-Rao-Vu,
Hajek-Wu-Xu]

"correlation clustering" [Bansal, Blum, Chawla '04]
"LDGM codes" [Kumar, Pakzad, Salavati, Shokrollahi '12]
"labelled block model" [Heimlicher, Lelarge, Massoulié '12]
"soft CSPs" [Abbe, Montanari '13]
"pairwise measurements" [Chen, Suh, Goldsmith '14 and '15]
"bounded-size correlation clustering" [Puleo, Milenkovic '14]

Some open problems in community detection

II. Other block models

- Censored block models
- Labelled block models
- Degree-corrected block models   [Karrer-Newman '11]
- Mixed-membership block models   [Airoldi-Blei-Fienber-Xing '08]
- Overlapping block models   [Abbe-Sandon '15]



$OSBM(n, p, f)$

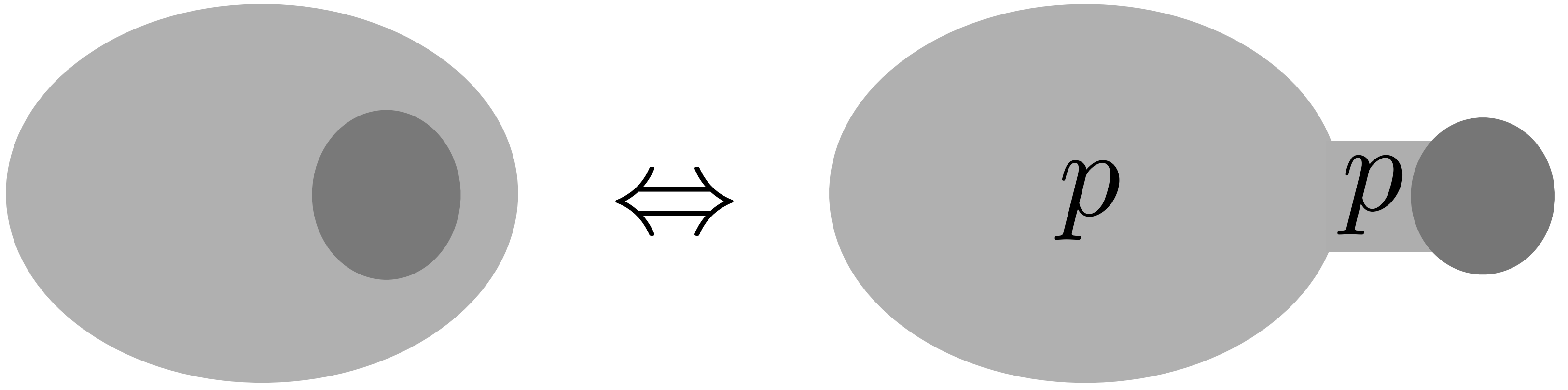
p on $\{0,1\}^s$

connect (u, v) with prob. $f(X_u, X_v)$ $f : \{0,1\}^s \times \{0,1\}^s \to [0,1]$

Some open problems in community detection

II. Other block models

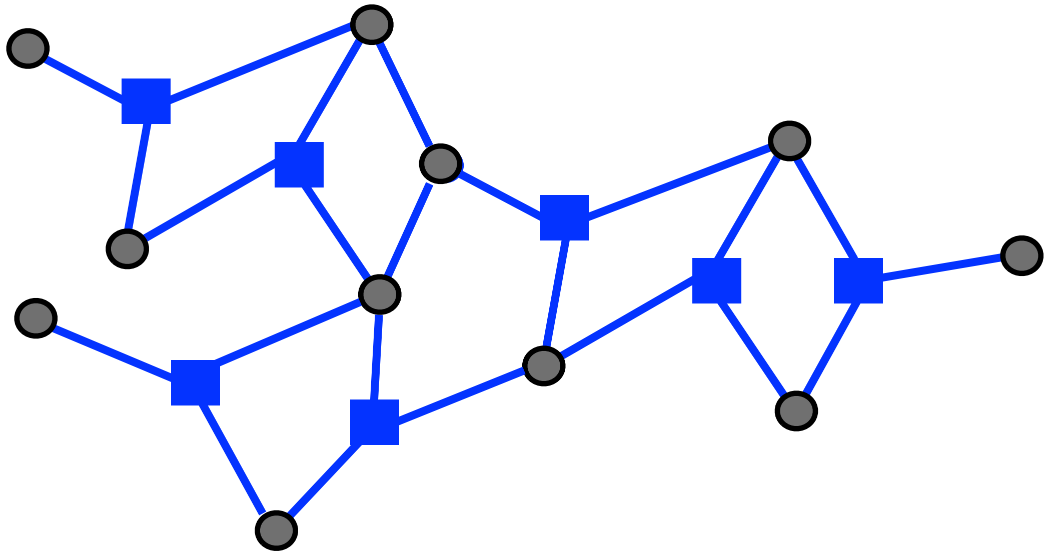
- Censored block models
- Labelled block models
- Degree-corrected block models
- Mixed-membership block models
- Overlapping block models
- Planted community [Deshpande-Montanari '14 , Montanari '15]



Some open problems in community detection

II. Other block models

- Censored block models
- Labelled block models
- Degree-corrected block models
- Mixed-membership block models
- Overlapping block models
- Planted community
- Hypergraph models



Some open problems in community detection

II. Other block models

- Censored block models
- Labelled block models
- Degree-corrected block models
- Mixed-membership block models
- Overlapping block models
- Planted community
- Hypergraph models

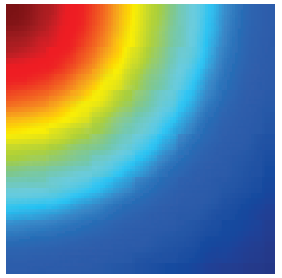
For all the above: Is there a CH-divergence behind?
A generalized notion of SNR? Detection gaps?
Efficient algorithms?

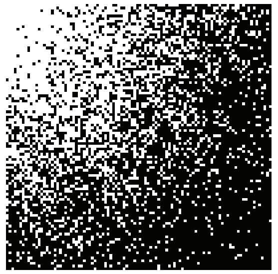
Some open problems in community detection

III. Beyond block models:

a. Exchangeable random arrays and graphons $w : [0,1]^2 \to [0,1]$

$$P(E_{ij} = 1 | X_i = x_i, X_j = x_j) = w(x_i, x_j)$$





[Lovasz]
SBMs can approximate graphons [Choi-Wolfe, Airoldi-Costa-Chan]

b. Graphical channels

A general information-theoretic model?

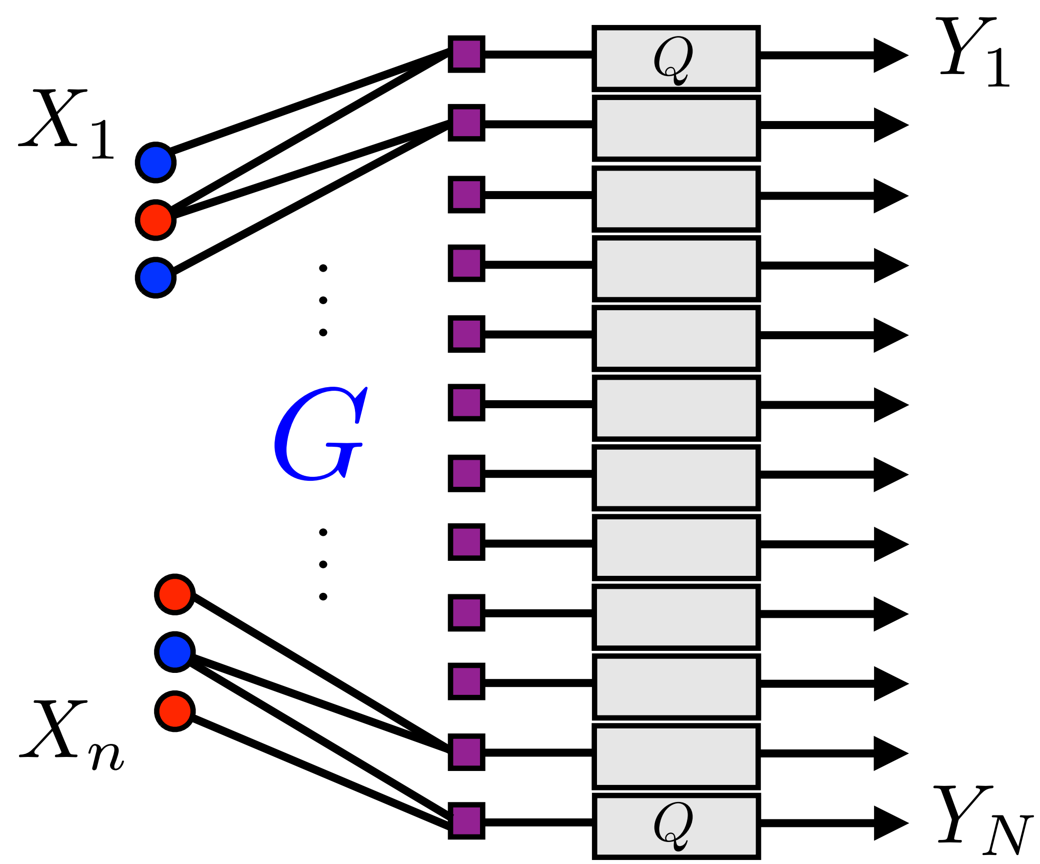
Graphical channels [Abbe-Montanari '13]

A family of channels motivated by inference on graph problems

- $G = (V, E)$ a k-hypergraph
- $Q : \mathcal{X}^k \to \mathcal{Y}$ a channel (kernel)

For $x \in \mathcal{X}^V, y \in \mathcal{Y}^E$,

$$\mathbb{P}(y|x) = \prod_{e \in E(G)} Q(y_e | x[e])$$



How much information do graphical channels carry?

Theorem. $\lim_{n\to\infty} \frac{1}{n} I(X^n; G)$ exists for ER graphs and some kernels

-> why not always? what is the limit?

Connection: sparse PCA and clustering

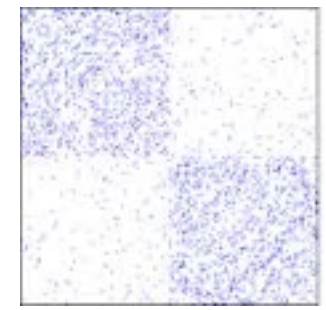
SBM and low-rank Gaussian model

Spiked Wigner model: $Y_\lambda = \sqrt{\frac{\lambda}{n}} X X^t + Z$ ← Gaussian symmetric $Y_{ij} = c X_i X_j + Z_{ij}$

- $X = (X_1, \ldots, X_n)$ i.i.d. Bernoulli(ϵ) $\to$ sparse-PCA

[Amini-Wainwright '09, Desphande-Montanari '14]

- $X = (X_1, \ldots, X_n)$ i.i.d. Radamacher$(1/2)$ $\to$ SBM???



[Deshpande-Abbe-Montanari '15]

Theorem. $\lim_{n\to\infty} \frac{1}{n} I(X; G(n, p_n, q_n)) \stackrel{?}{=} \lim_{n\to\infty} \frac{1}{n} I(X; Y_\lambda)$

$$= \frac{\lambda}{4} + \frac{\gamma_*^2}{4\lambda} - \frac{\gamma_*}{2} + I(\gamma_*)$$

If $\lambda_n = \frac{n(p_n - q_n)^2}{2(p_n + q_n)} \to \lambda$ (finite SNR), with $np_n, nq_n \to \infty$ (large degrees)

where γ_* solves $\gamma = \lambda(1 - \mathrm{MMSE}(\gamma))$ $\longleftarrow$ $Y_1(\gamma) = \sqrt{\gamma}\, X_1 + Z_1$

I-MMSE [Guo-Shamai-Verdú] (single-letter)

Conclusion

Community detection couples naturally with the channel view of information theory and more specifically with:

- graph-based codes
- f-divergences
- broadcasting problems
- I-MMSE
- ...

unorthodox versions...

More generally, the problem of inferring global similarity classes in data sets from noisy local interactions is at the center of many problems in ML, and an information-theoretic view of these problems seems needed and powerful.

Questions?

Documents related to the tutorial:
www.princeton.edu/~eabbe