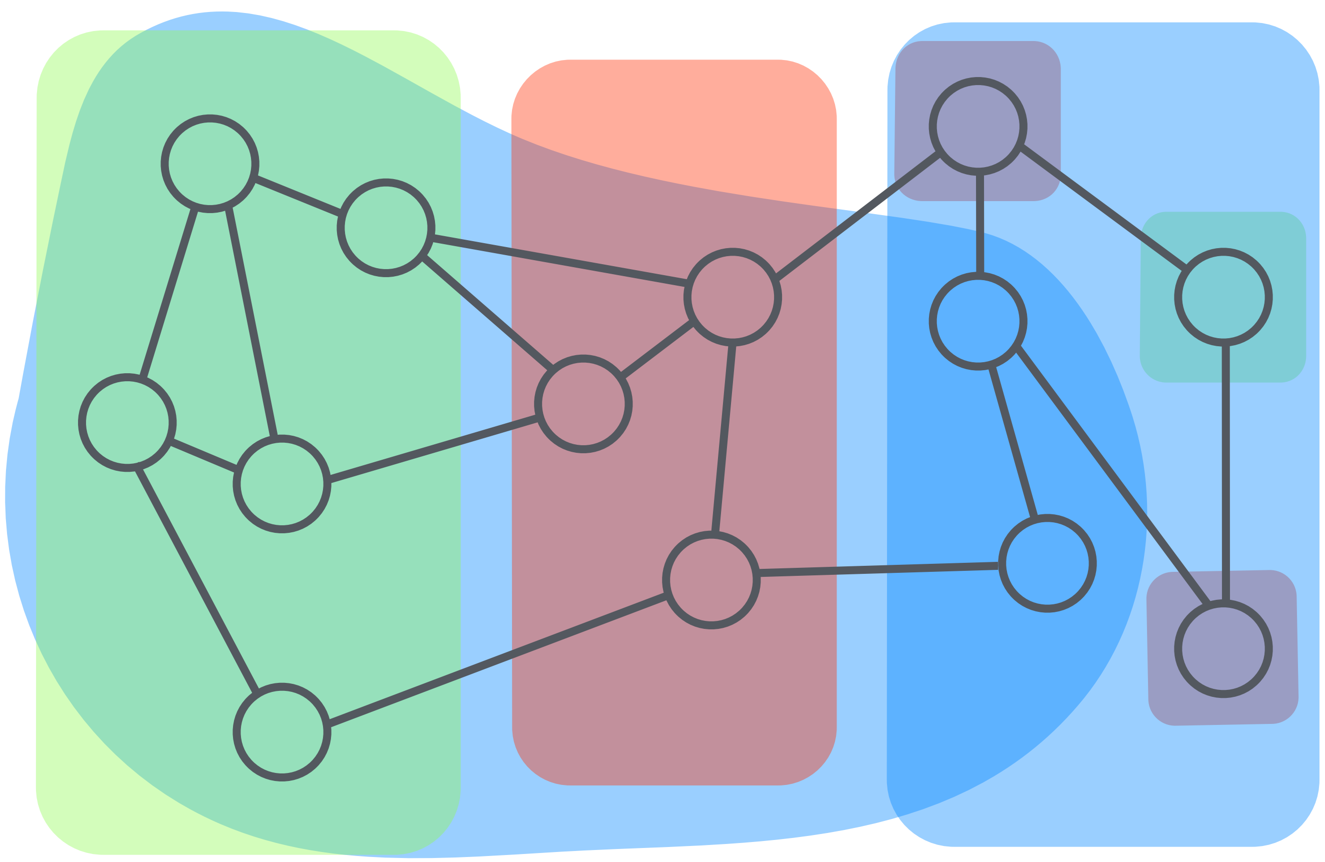
# Learning graphical models: hardness and tractability

Guy Bresler    David Gamarnik    Devavrat Shah
MIT

# graphical models

$G = (V, E) \quad |V| = p \quad |\partial i| \leq d$



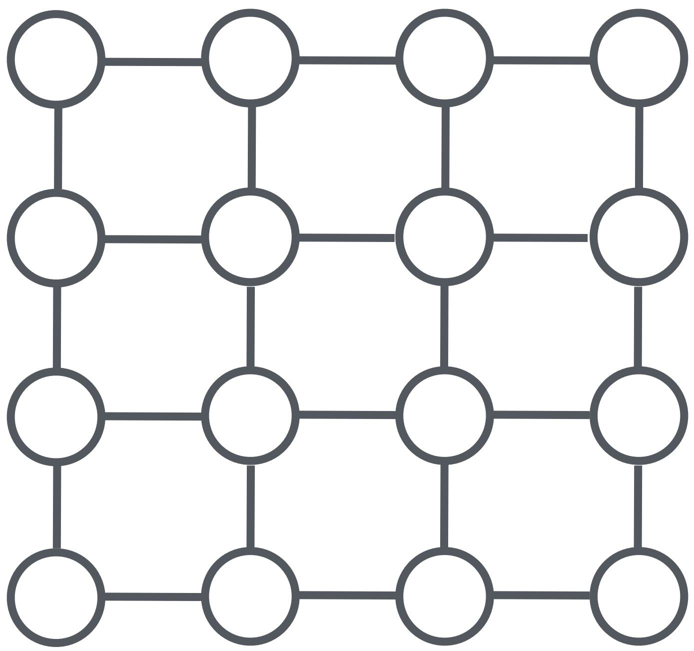
$X_A \perp\!\!\!\perp X_B \mid X_S$

$X_i \perp\!\!\!\perp X_{V \setminus \partial i \cup \{i\}} \mid X_{\partial i}$

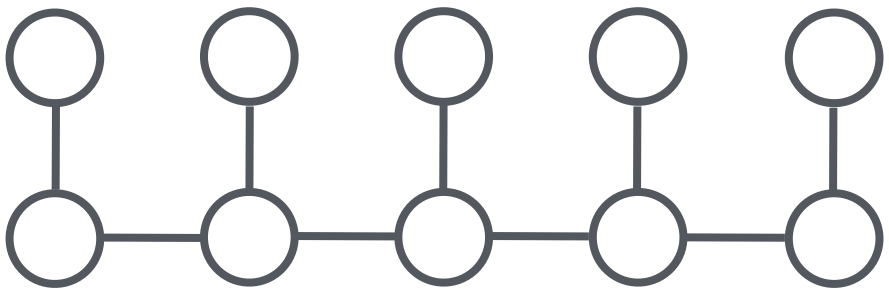
# efficient inference

belief propagation can be used to do inference

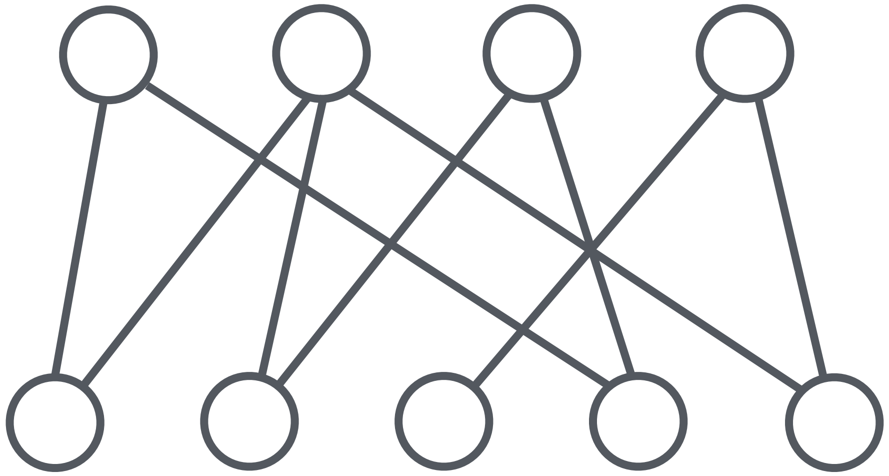
historically people knew what models to use



lattice

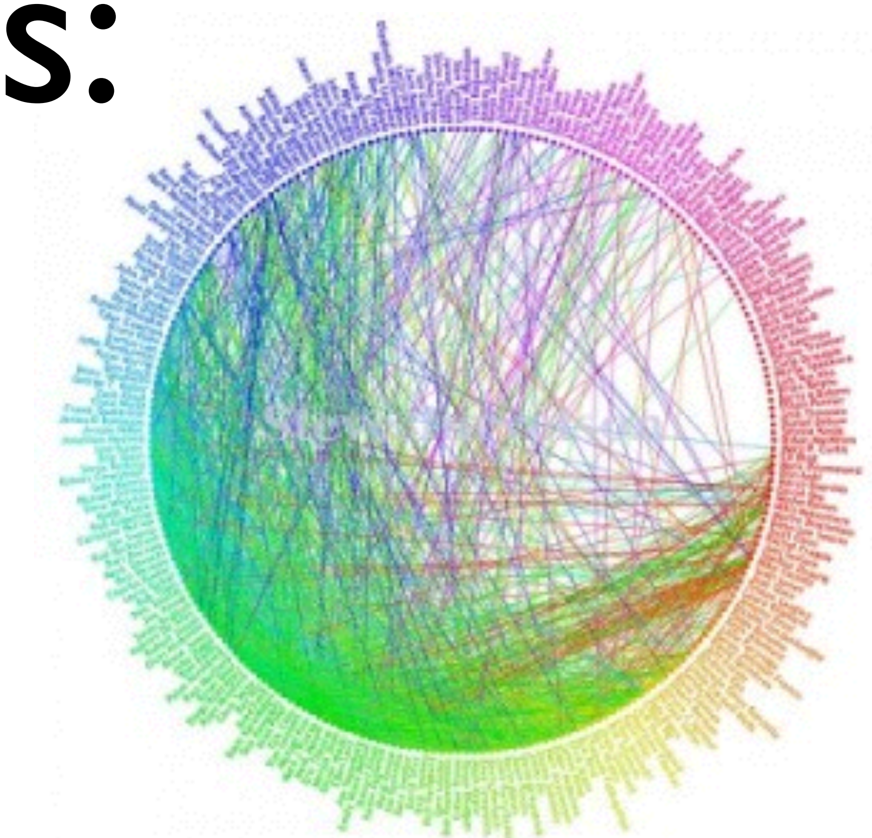


hidden Markov model



LDPC code

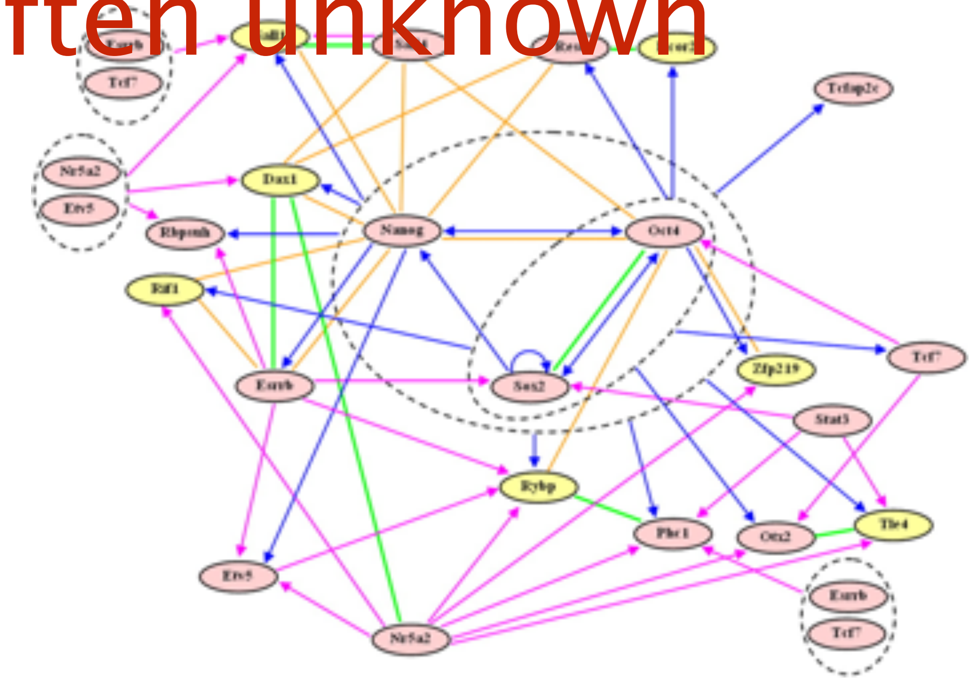
# modern applications: unknown structure



social network data


financial data

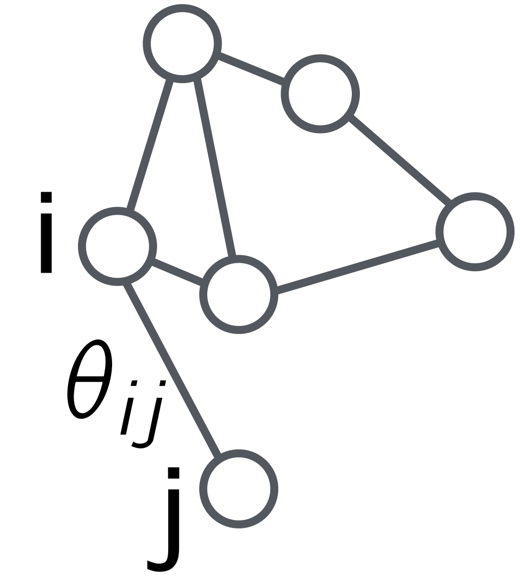
structure for modern data is often unknown



gene regulatory network

# learning graphical model

$G = (V, E)$
$|V| = p$
$|\partial i| \leq d$
$X \in \{0, 1\}^p$




**graphical model:** $$\mathrm{P}(X) = \frac{1}{Z} \exp\left( \sum_{\{i,j\} \in E} \theta_{ij} X_i X_j + \sum_{i \in V} \theta_i X_i \right)$$

$$\alpha \leq |\theta_{ij}| \leq \beta$$

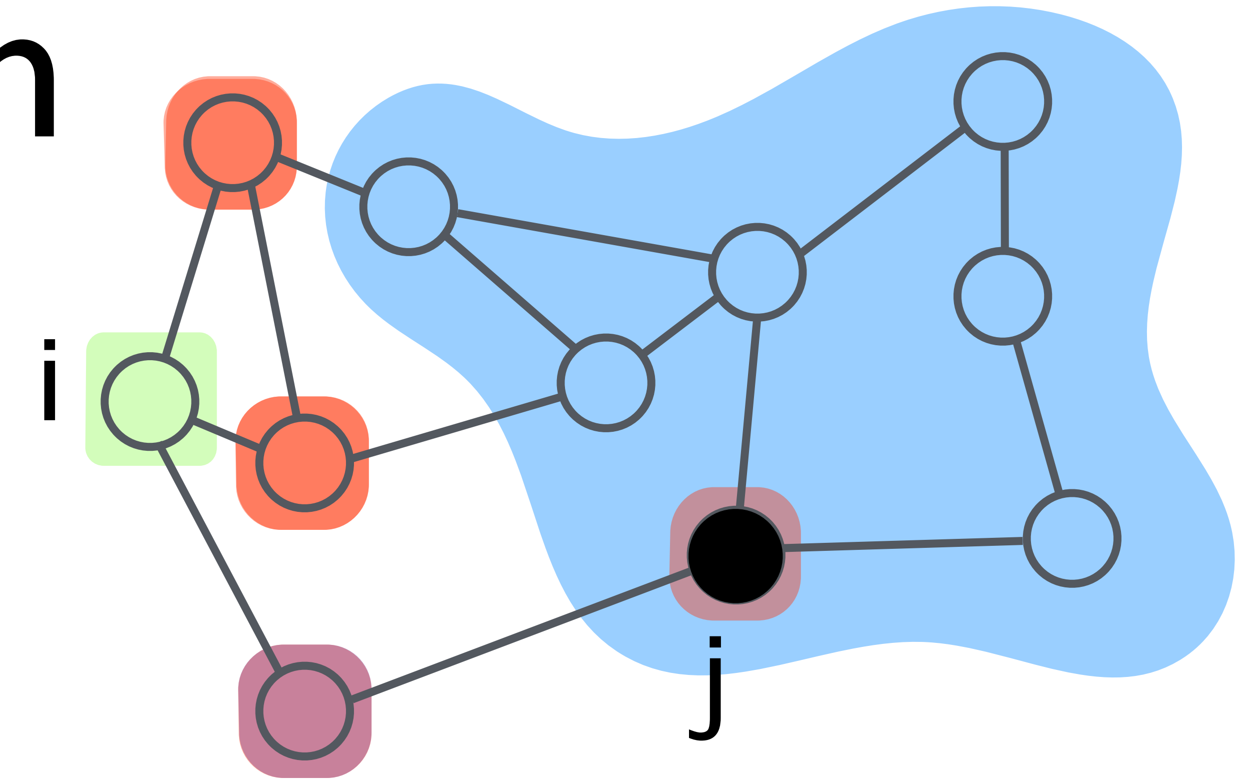
**data:** $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ $\quad X \sim \mathrm{P}$ (i.i.d. samples)

**task:** reconstruct graph and parameters from the data

w.prob. $\to 1$ as $n, p \to \infty$

# baseline: exhaustive search algorithm

[Abbeel-Koller-Ng '06]
[Bresler-Mossel-Sly '08]




$$X_i \perp\!\!\!\perp X_{V\setminus \partial i \cup \{i\}} \mid X_{\partial i}$$

test whether $U \subseteq \partial i$

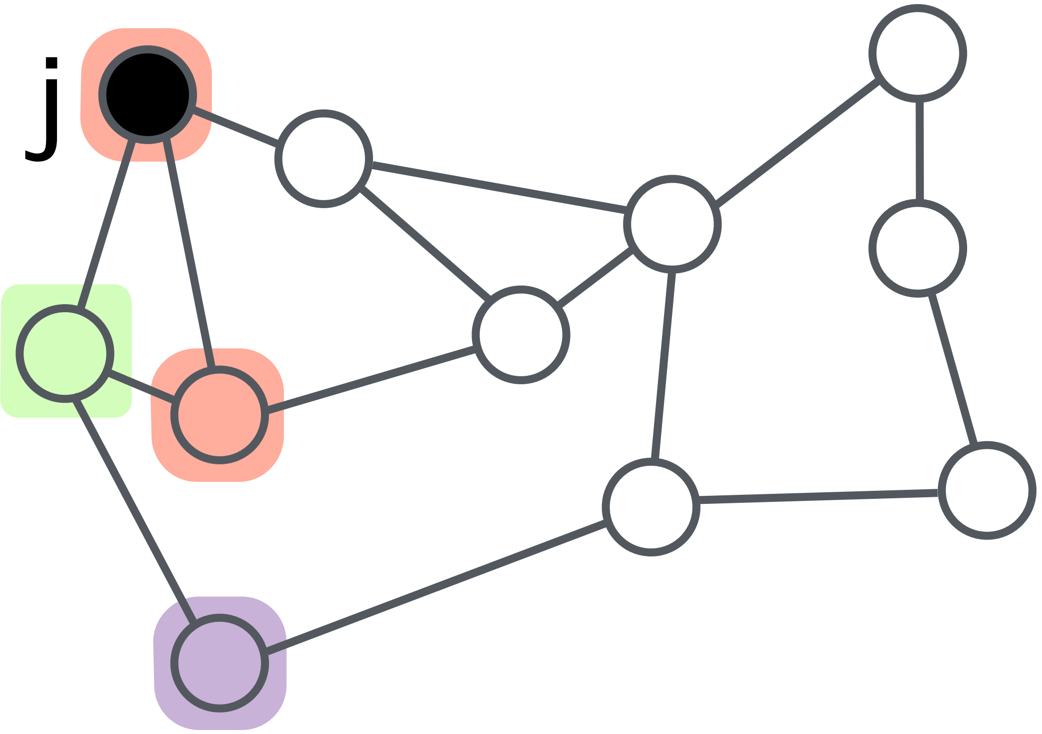
if $U \subsetneq \partial i$ then for some $j \in U$ and $W \supseteq \partial i \setminus U$

$$|P(X_i = +1 | X_U = x_U, X_W = x_W)$$
$$- P(X_i = +1 | X_U = \mathrm{flip}_j(x_U), X_W = x_W)| = 0$$

"*U* fails test"

# baseline: exhaustive search algorithm

[Bresler-Mossel-Sly '08]




$$X_i \perp\!\!\!\perp X_{V \setminus \partial i \cup \{i\}} \mid X_{\partial i}$$

**test whether** $U \subseteq \partial i$

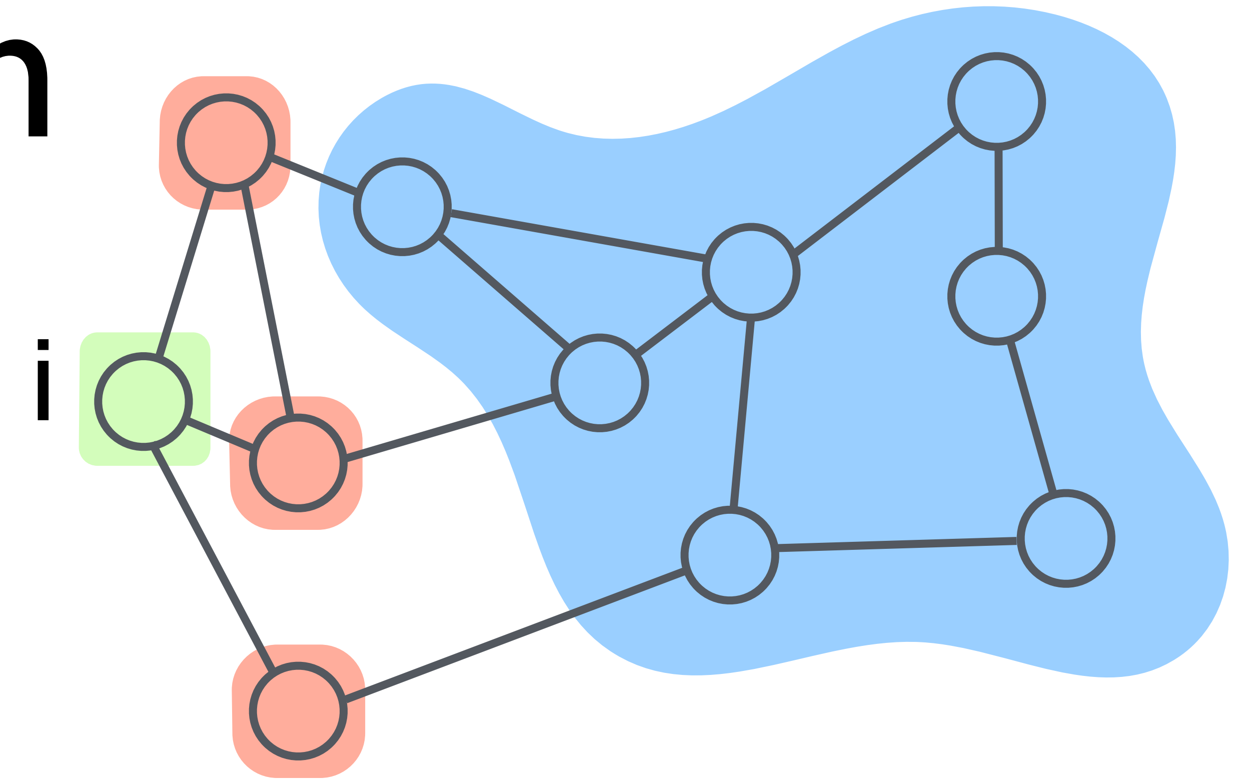
**if** $U \subseteq \partial i$ **then for all** $j \in U$ **and** $W \supseteq \partial i \setminus U$

$$|\mathrm{P}(X_i = +1 | X_U = x_U, X_W = x_W) - \mathrm{P}(X_i = +1 | X_U = \mathrm{flip}_j(x_U), X_W = x_W)| \geq \frac{\tanh 2\alpha}{2e^{2(d-1)\beta}}$$

"$U$ passes test"

# baseline: exhaustive search algorithm

$X_i \perp\!\!\!\perp X_{V \setminus \partial i \cup \{i\}} \mid X_{\partial i}$




Algorithm:

for each node $i$

test all possible neighborhoods $U$

choose largest $U$ passing test

**Theorem:** [Bresler-Mossel-Sly '08]

algorithm recovers with prob. $1 - o(1)$ using $n = O(2^{2d} e^{(4\beta+h)d} \log p)$ samples, w runtime $\widetilde{O}\left(p^{2d+1}\right)$

# our notion of computational efficiency

exhaustive search: $p^{\Theta(d)}$

indep. of *d*!

efficient: $f(d)\,p^{c}$

can be exponential

want to have no restrictions on graph structure

question: for what types of interactions can we learn *efficiently*?

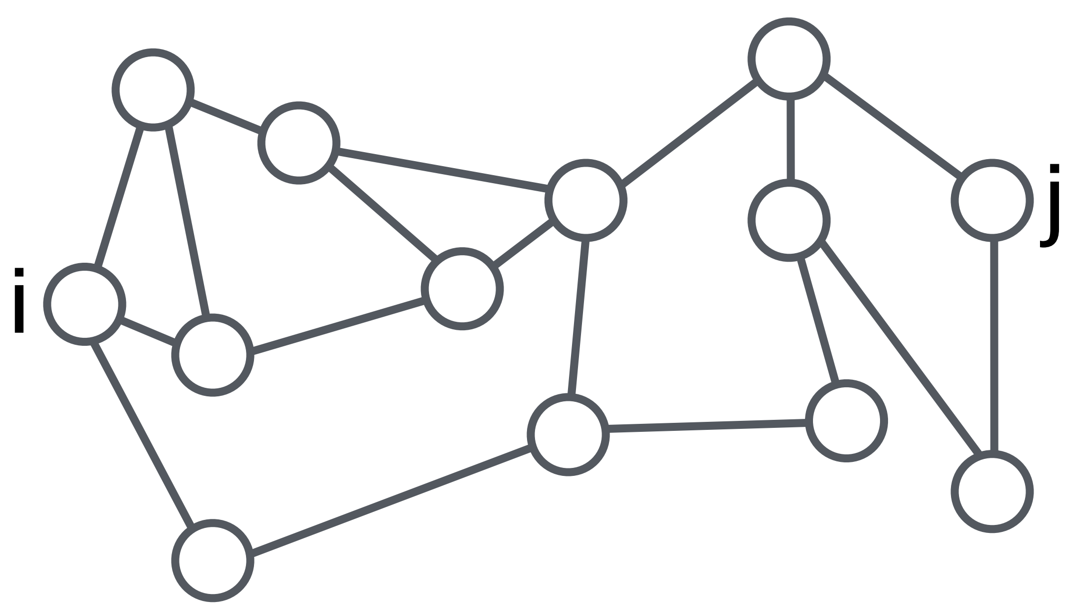
# correlation decay

**Theorem:** [Bresler-Mossel-Sly '08]

if have *correlation decay* and $\mathrm{E}X_iX_j \geq \kappa$ for $\{i,j\} \in E$
can learn using $n = O(2^{8d} e^{16\beta d} \log p)$ samples in time $O(np^2)$

$$f(d)p^c : \quad c = 2, \; f(d) = 2^{8d} e^{16\beta d}$$

exponential decay of correlations: $\mathrm{E}X_iX_j \leq (1-\gamma)^{\mathrm{dist}(i,j)}$

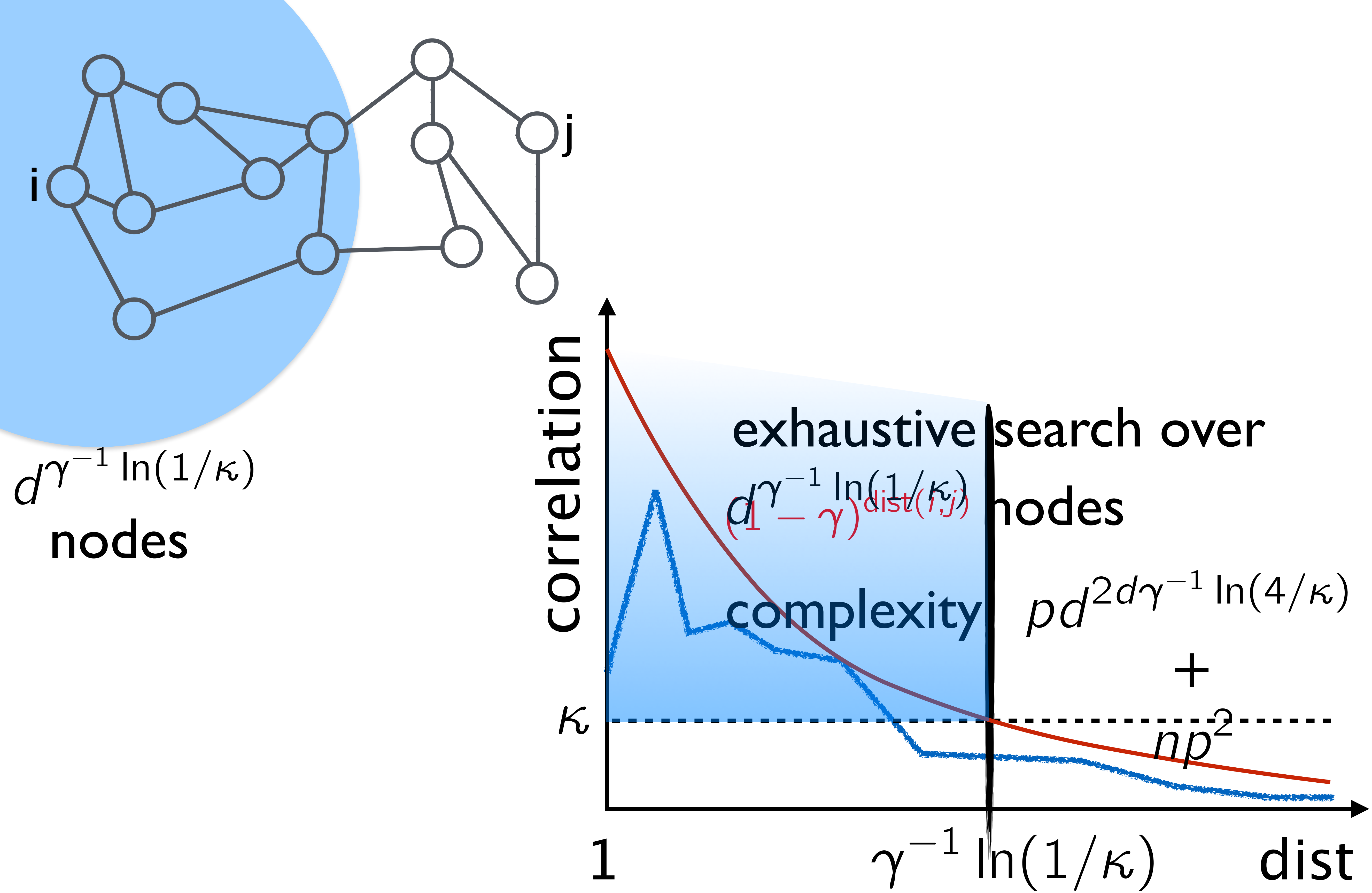



various models satisfy CDP:

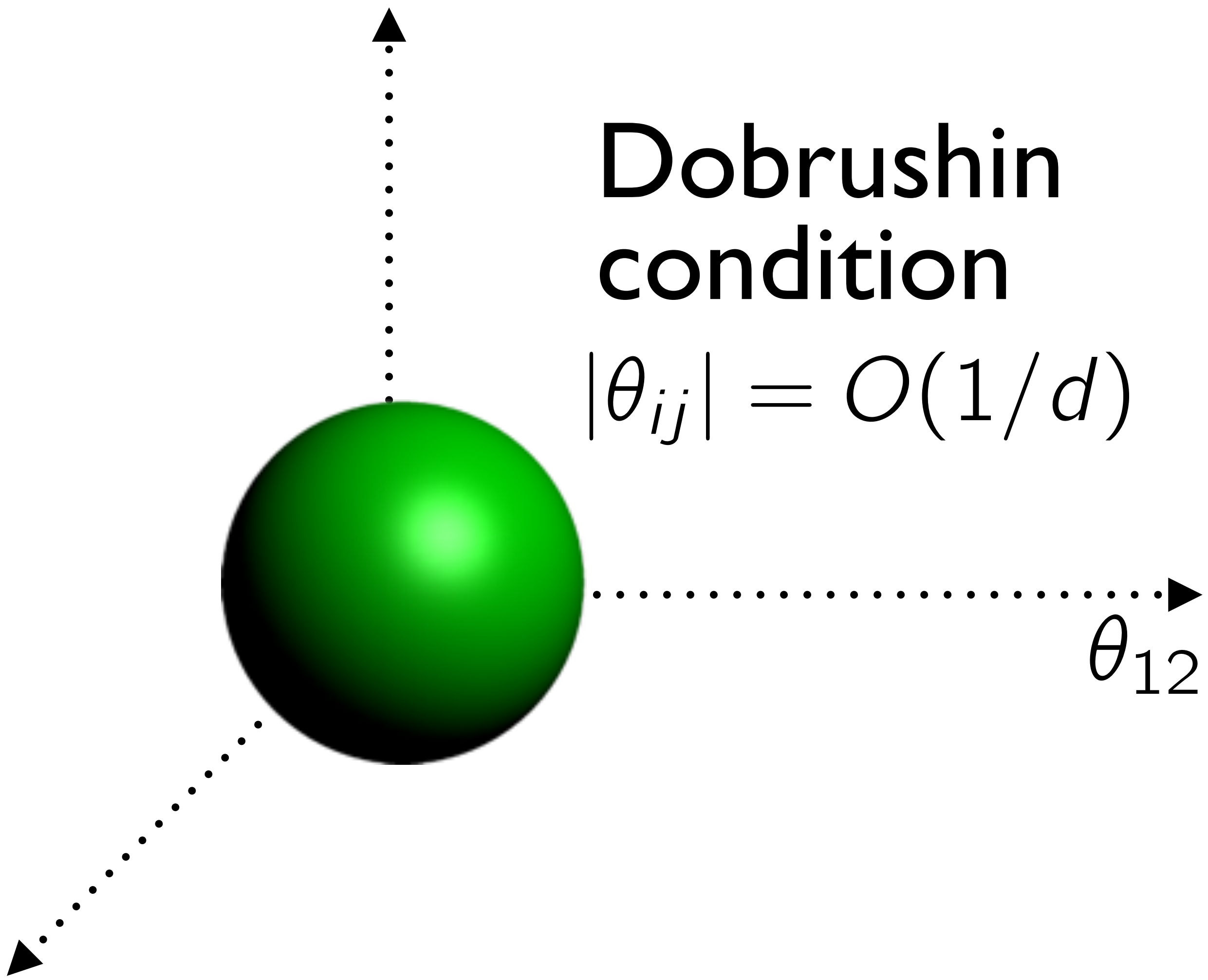
[Dobrushin '70, Dobrushin-Shlosman '85, Martinelli '95, Weitz '06, Salas-Sokal '97, Bandyopadhyay-Gmarnik '08, Gamarnik-Goldberg-Weber '13, and many others…]

**Theorem:** if have *correlation decay* and $\mathrm{E}X_iX_j \geq \kappa$ for $\{i,j\} \in E$:

can learn using $n = O(2^{8d} e^{16\beta d} \log p)$ samples in time $O(np^2)$



# correlation decay




[Bento-Montanari '09]

other low-complexity approaches to learning:

[Ravikumar-Lafferty-Wainwright '06]
[Lee-Ganapathi-Koller '06]
[Anandkumar-Tan-Huang-Willsky '12]
[Ray-Sanghavi-Shakkottai '12]
[Wu-Srikant-Ni '13]
[many many others]

several low-complexity algorithms
fail without correlation decay

all known low-complexity algorithms
explicitly or implicitly require correlation decay

can we learn efficiently without correlation decay?

# repelling models

$$\mathrm{P}(X) = \frac{1}{Z} \exp\left( \sum_{\{i,j\} \in E} \theta_{ij} X_i X_j + \sum_{i \in V} \theta_i X_i \right) \qquad X \in \{0,1\}^p$$

$$\alpha \le |\theta_{ij}| \le \beta$$

repelling

$$\theta_{ij} \le -\alpha$$

$$\theta_i \le h$$

$$\alpha \ge d(h + \ln 2)$$

**Theorem:** [Bresler-Gamarnik-Shah '14a]

can learn these models with prob. $1 - o(1)$ using $n = O(2^{2d} e^{4\beta d} \log p)$ samples, with runtime $O(np^2)$

# repelling models

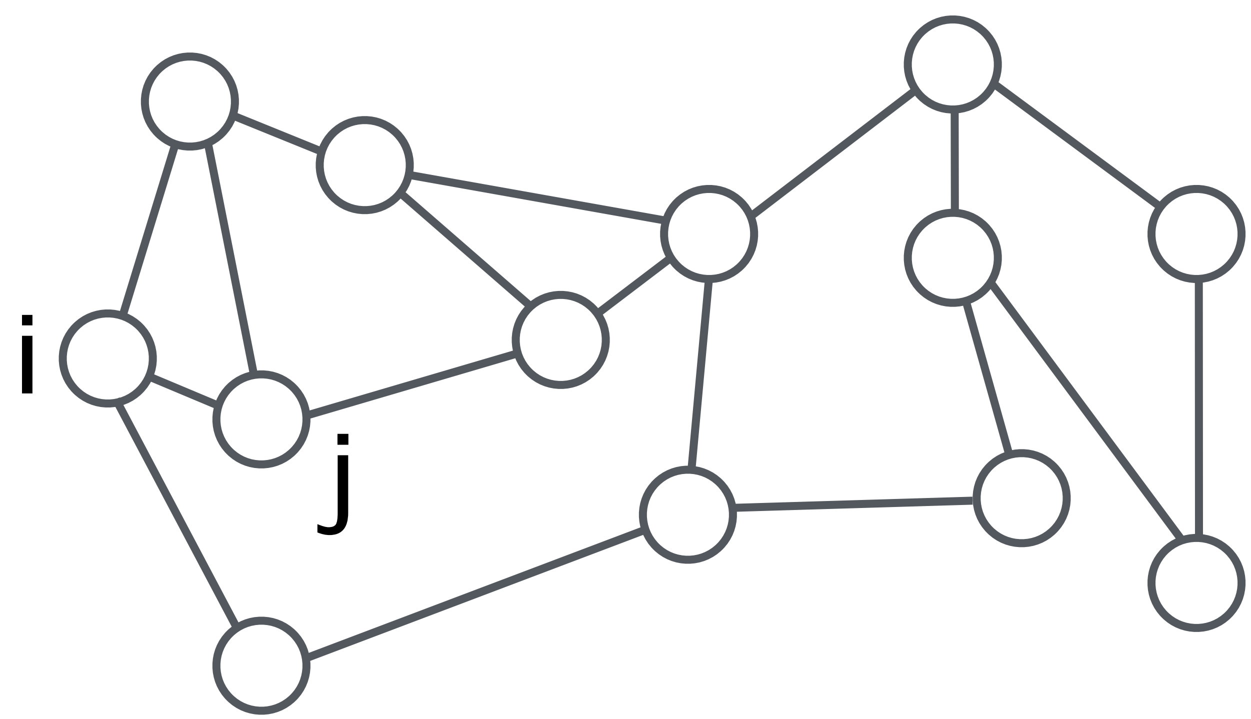
## Ex. Independent set model

$$\alpha \to \infty, \text{ i.e. } \theta_{ij} \to -\infty$$

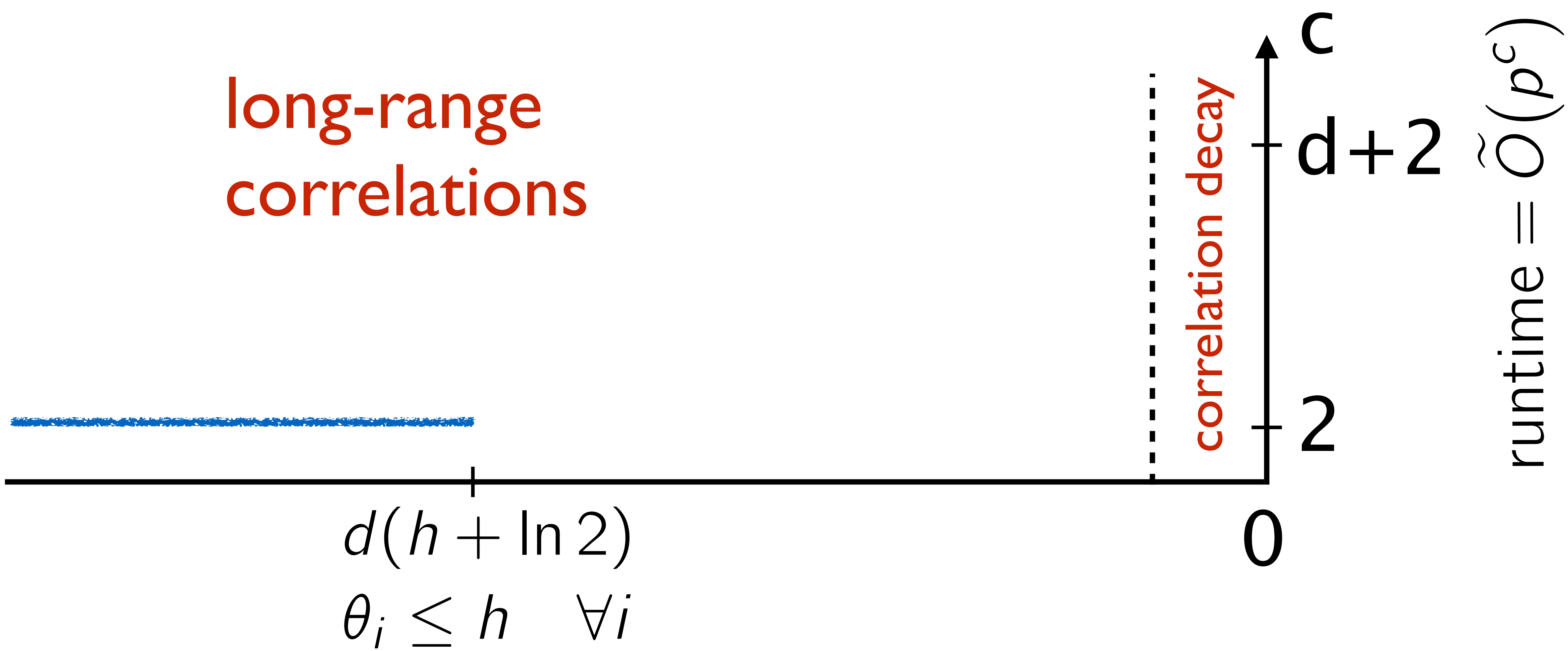
Key observation:

i and j are neighbors: $(X_i, X_j) \neq (1,1)\ w.p.\ 1$

i and j are not neighbors: $(X_i, X_j) = (1,1)\ w.p. \geq \gamma(d)$



# repelling models



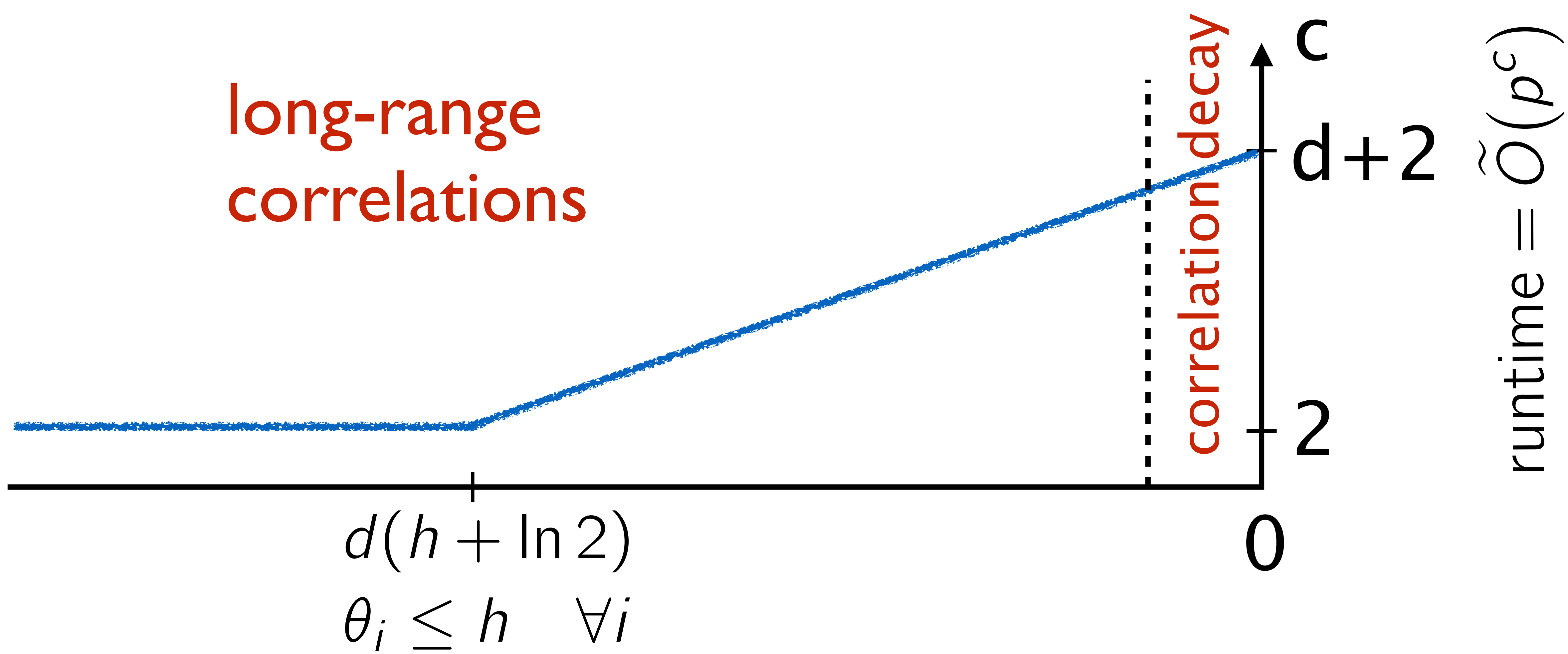

**Theorem:** [Bresler-Gamarnik-Shah '14a]

can learn these models with prob. $1-o(1)$ using $n = O(2^{2d} e^{4\beta d} \log p)$ samples, with runtime $O(np^2)$

# repelling models

$\cancel{\alpha \geq d(h + \ln 2)}$

$\alpha \geq (d - \tau)(h + \ln 2)$




**Theorem:** [Bresler-Gamarnik-Shah '14a]

can learn these models with prob. $1 - o(1)$ using $n = O(2^{2d} e^{4\beta d} \log p)$ samples, with runtime $\cancel{O(np^2)}$ $O(np^{2+\tau})$

# learning parameters

**once you know the graph, learning parameters is easy**

Algorithm (node-wise parameters)

1. consider samples with $X_{\partial i} = \mathbf{0}$
2. this allows to estimate

$$\mathsf{P}(X_i = 1 | X_{\partial i} = \mathbf{0}) = \frac{e^{h_i}}{1 + e^{h_i}}$$

3. solve for $h_i$

**Theorem:**

algorithm recovers with prob. $1 - o(1)$ using $n = O(2^{2d} e^{(4\beta+h)d} \log p)$ samples, with runtime $O(np)$

# can we do better in general

exhaustive search: $p^{\Theta(d)}$

**Theorem:** [Bresler-Gamarnik-Shah '14a]

No algorithm can do better than $p^{\Theta(d)}$ under the computation model of "statistical algorithms" in general.

a general approach to simplifying:

reduce to sufficient statistics

# reducing to sufficient statistics

$$\mathrm{P}(X) = \frac{1}{Z} \exp\left( \sum_{\{i,j\} \in E} \theta_{ij} X_i X_j + \sum_{i \in V} \theta_i X_i \right) \qquad X \in \{0,1\}^p$$

$$\left.\begin{aligned} (\mu_i)_i &= (\mathrm{E}X_i)_i = (\mathrm{P}(X_i = i))_i \\ (\mu_{ij})_{ij} &= (\mathrm{E}X_i X_j)_{ij} = (\mathrm{P}(X_i = X_j i))_{ij} \end{aligned}\right] \text{ sufficient statistics}$$

try to estimate $\mu \mapsto \theta$ (feasible in principle!)

physicists try to estimate this map using various "expansions"

[Ricci-Tersenghi '12]
[Sessak-Monasson '08]
[Cocco-Monasson '12]
...many others

# reducing to sufficient statistics

$$\mathrm{P}(X) = \frac{1}{Z}\exp\left(\sum_{\{i,j\}\in E}\theta_{ij}X_iX_j + \sum_{i\in V}\theta_i X_i\right) \qquad X\in\{0,1\}^p$$

$\theta_{ij} \to -\infty$

(special case of repelling model)

$$\mathrm{P}(X) = \frac{1}{Z}\exp\left(\sum_{i\in V}\theta_i X_i\right), \quad X \text{ is an independent set}$$

$\mu = (\mathrm{E}(X_i))_i = (\mathrm{P}(X_i = 1))_i$ sufficient statistics

**Theorem:** [Bresler-Gamarnik-Shah '14b] [Montanari '14]

learning parameters of graphical models from sufficient statistics is NP-hard

# some remarks on proof

Reduction:

Suppose there exists efficient algorithm for $\mu \mapsto \theta$

Use it as a black-box to solve a known difficult problem

The difficult problem: given $\theta$ find corresponding $\mu \equiv \mu(\theta)$

For independent set with $\theta = \mathbf{0}$ this corresponds to

counting # of independent sets in G

a known hard (to approximate) problem

[Dyer-Frieze-Jerrum ’02]
[Sly ’10]
[Sly-Sun ’12]

# some remarks on proof

## Reduction:

The difficult problem: given $\theta$ find corresponding $\mu \equiv \mu(\theta)$

Solve using black-box $\mu \mapsto \theta$

$$\mu(\theta) \in \arg\max_{\nu \in \mathcal{M}} \ \langle \nu, \theta \rangle + H_{\mathrm{ER}}(\nu)$$

Gradient ascent:

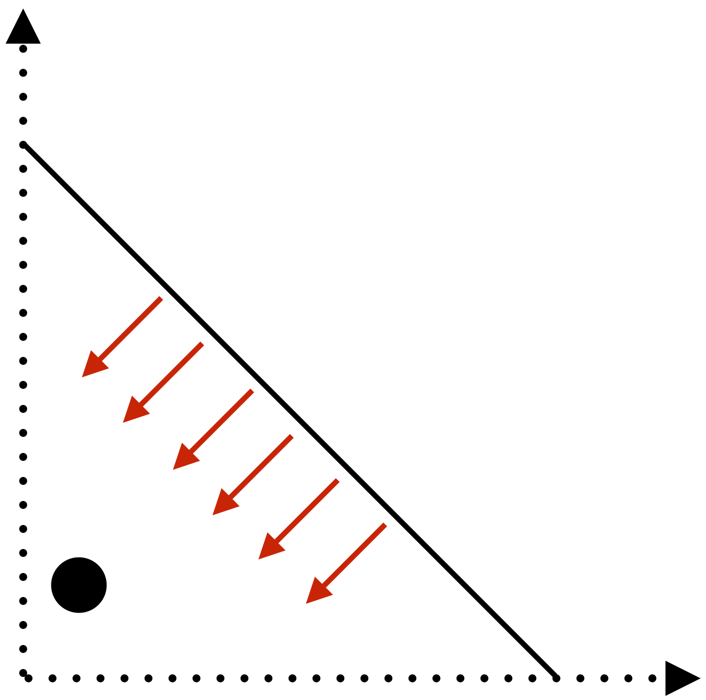
$\mu^t \to \theta^t$

$$\mu^{t+1} = \mu^t + \frac{1}{t}\left(\theta - \theta^t\right)$$

Key challenge:

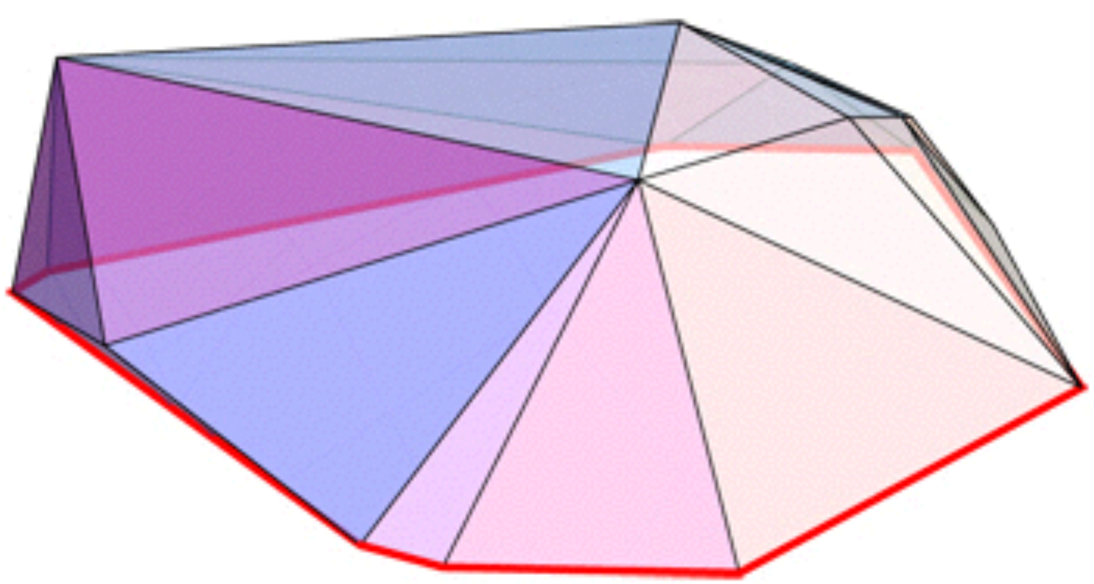
$\mu^{t+1}$ needs to be projected on marginal polytope $\mathcal{M}$

# some remarks on proof



what if algorithm naturally avoids boundary

**Lemma:** [Bresler-Gamarnik-Shah '14b]
For the objective of interest, the polytope boundary has an inherent repulsion property



marginal polytope is *very complicated*

once you know the graph,
learning parameters is easy

graph tells you on which
higher order statistics to focus

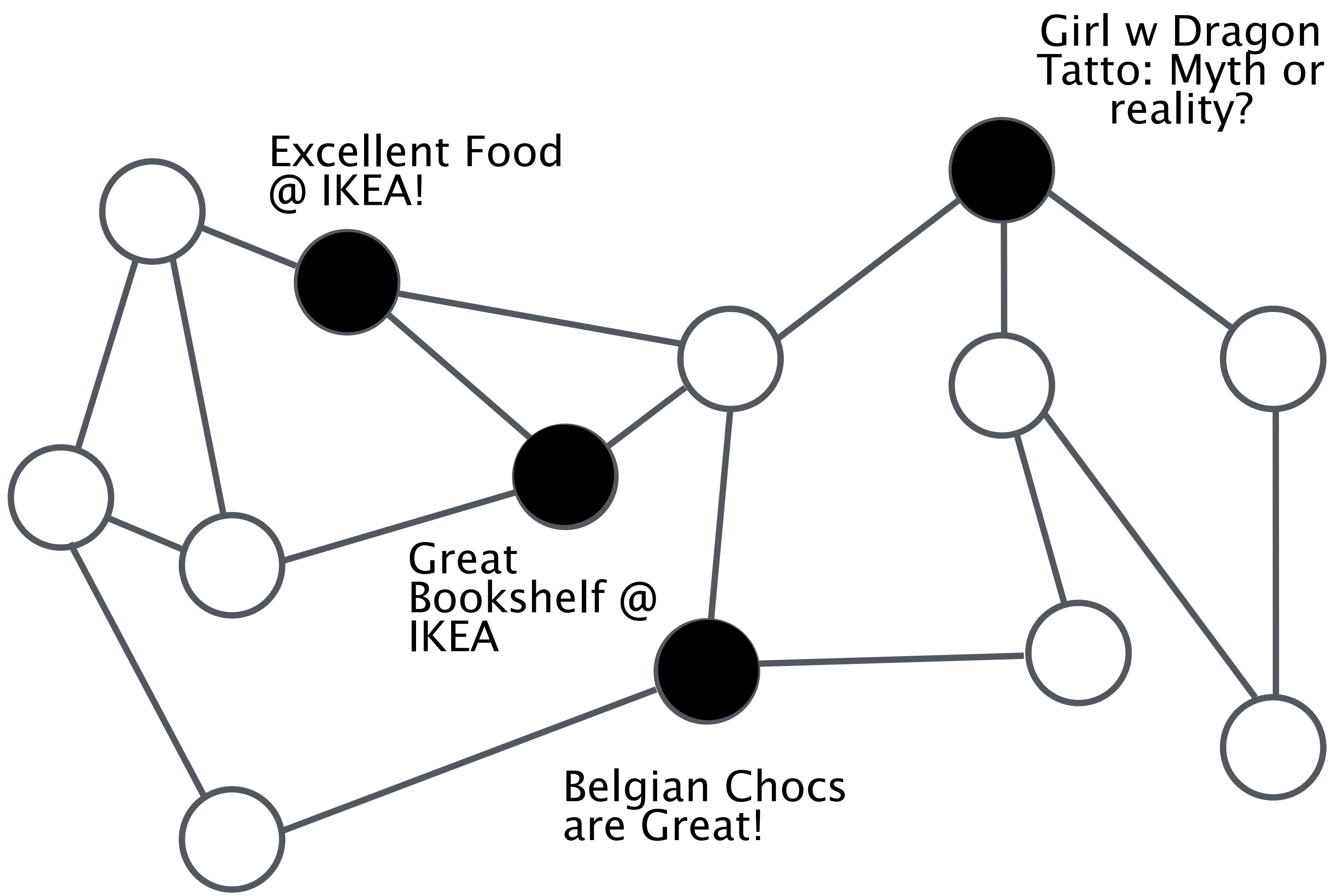
learning from sufficient statistics
is probably not a good idea

so far: i.i.d. data

revisit original goal:
learning from data

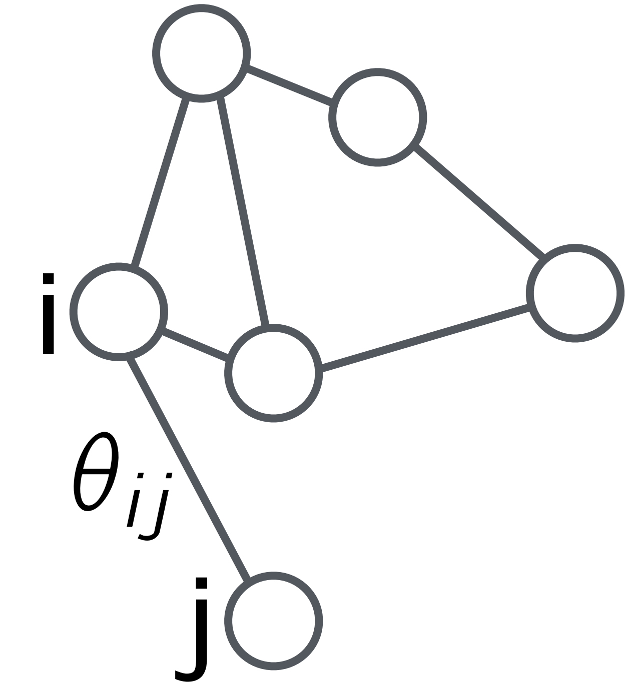
# what sort of data?

## Social Behavior: Purchases, Likes, …




dynamics over time

# learning models from data




$$\mathrm{P}(X) = \frac{1}{Z} \exp\left( \sum_{\{i,j\}\in E} \theta_{ij} X_i X_j + \sum_{i \in V} \theta_i X_i \right)$$

$$X \in \{0,1\}^p \quad \alpha \le |\theta_{ij}| \le \beta$$

data: $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ ~~i.i.d. samples~~
*n* steps of some process

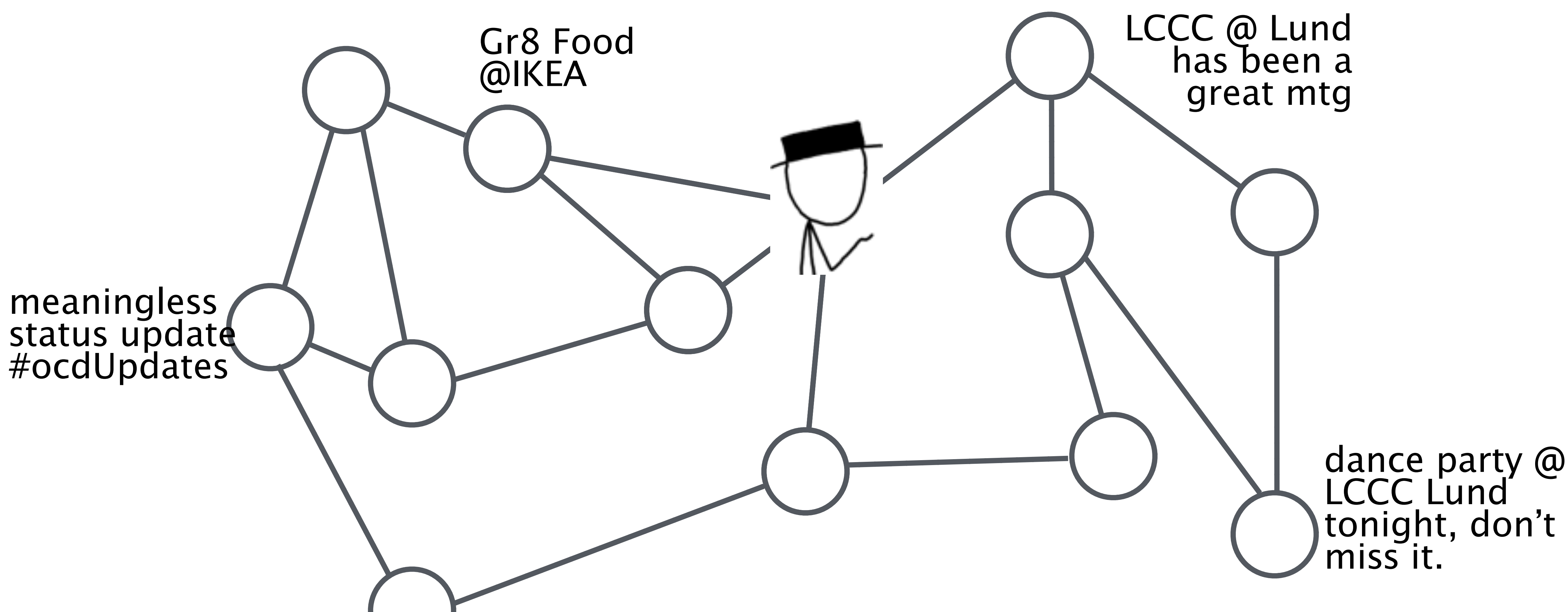
task: reconstruct graph and parameters from the data
w.prob. $\to 1$ as $n, p \to \infty$

# Glauber dynamics

1. each node has a Poisson(1) clock

2. when clock rings, update variable according to

$$P(X_i = 1 | X^t_{\partial i}) = \frac{\exp\left(2 \sum_{j \in \partial i} \theta_{ij} X^t_j\right)}{1 + \exp\left(2 \sum_{j \in \partial i} \theta_{ij} X^t_j\right)}$$



# slow mixing

i.i.d. sampling is NP-hard for some models
but Glauber dynamics defined for any graphical model

for models without correlation decay, the Glauber dynamics is known to mix exponentially slowly in $p$

samples will be *far* from i.i.d.

# efficient learning from the Glauber dynamics

**Theorem:** [Bresler-Gamarnik-Shah '14c] with $n = O(e^{4d\beta} \log p)$ samples per node, and runtime $O(np^2)$ can learn *any pairwise model* *even* *without correlation decay*

learning theory:

[Aldous-Vazirani '90]

[Bartlett-Fischer-Hoffgen '94]

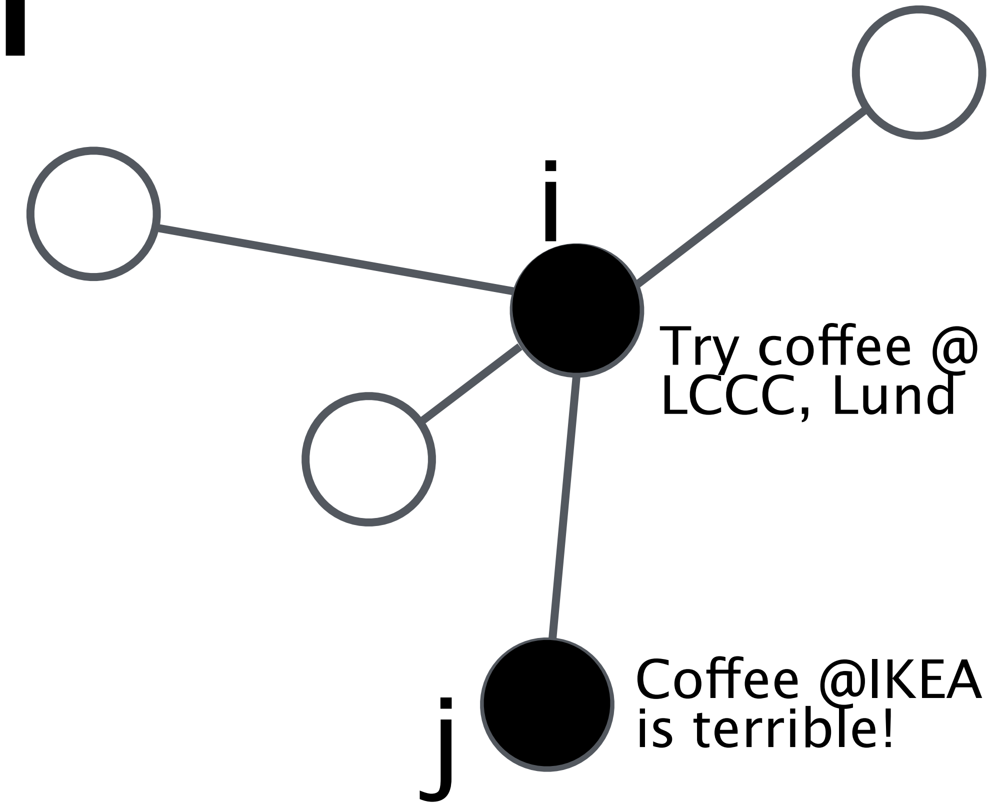
[Bshouty-Mossel-O'Donnell-Servedio '03]

epidemic models:

[Netrapalli-Sanghavi '12]

[Dahleh-Tsitsiklis-Zoumpoulis '13]

# estimating effect of a neighbor




imaginary scenario: node i updates, then node j flips, then node i again

test for existence of an edge:

$$\exp\left(\theta_{ij}\right) = \frac{p^{+}(1-p^{-})}{p^{-}(1-p^{+})}$$

$$p^{+} = \mathbb{P}\big(X_i = +1 | X_{\partial i \setminus j} = +\mathbf{1}, X_j = +1\big)$$

$$p^{-} = \mathbb{P}\big(X_i = +1 | X_{\partial i \setminus j} = +\mathbf{1}, X_j = -1\big)$$

this would require $\Omega(e^{d\beta} p^2)$ samples per node

a more delicate argument is required to get to $O_d(\log p)$

# summary

correlation decay is not necessary to learn efficiently

however exhaustive algorithm seems the best in general

reducing to sufficient statistics is computationally suboptimal

observing dynamics over time can make things easy

insight: often makes sense to learn structure first
and only then estimate parameters