Complexity, Phase Transitions, and Inference

Cristopher Moore, Santa Fe Institute

with Aurelien Decelle, Lenka Zdeborová, Florent Krzakala, Xiaoran Yan, Yaojia Zhu, Cosma Shalizi, Lise Getoor, Aaron Clauset, Mark Newman, Elchanan Mossel, Joe Neeman, Allan Sly, Pan Zhang, Jess Banks, Praneeth Netrapalli, Thibault Lesieur, Caterina de Bacco, Roman Vershynin, and Jiaming Xu
Statistical inference ⇔ statistical physics
Statistical inference ⇔ statistical physics

How can we find patterns in noisy data?
Statistical inference ⇔ statistical physics

How can we find patterns in noisy data?
Phase transitions and fundamental limits
Statistical inference ⇔ statistical physics

How can we find patterns in noisy data?
Phase transitions and fundamental limits
Optimal algorithms
Statistical inference \iff statistical physics

How can we find patterns in noisy data?
Phase transitions and fundamental limits
Optimal algorithms
Information vs. efficient computation
Statistical inference ⇔ statistical physics

How can we find patterns in noisy data?
Phase transitions and fundamental limits
Optimal algorithms
Information vs. efficient computation
Interdisciplinary exchange
Why least squares?
Why least squares?

the most common way to fit a line to noisy data
Why least squares?

the most common way to fit a line to noisy data
data points $Y = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
Why least squares?

the most common way to fit a line to noisy data

data points $Y = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

model: $y_i = ax_i + b$
Why least squares?

the most common way to fit a line to noisy data

data points $Y = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

model: $y_i = ax_i + b$

find $a, b$ that minimize

$$\sum_i (y_i - (ax_i + b))^2$$
Why least squares?

the most common way to fit a line to noisy data

data points $Y = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

model: $y_i = ax_i + b$

find $a, b$ that minimize

$$\sum_i (y_i - (ax_i + b))^2$$

but why?
A model of noise
A model of noise

A model with noise: \( y_i = a x_i + b + w \)
A model of noise

A model with noise: \( y_i = ax_i + b + w \)

where \( w \) is Gaussian, \( P(w) \propto \exp\left(-\frac{1}{2\sigma} w^2\right) \)
A model of noise

a model with noise: \( y_i = ax_i + b + w \)

where \( w \) is Gaussian, \( P(w) \propto \exp\left(-\frac{1}{2\sigma} w^2\right) \)

total probability of the data is

\[
P(Y | a, b) = \prod_i P(y_i | a, b) \\
\propto \exp\left(-\frac{1}{2\sigma} \sum_i (y_i - (ax_i + b))^2 \right)
\]
A model of noise

a model with noise:  \( y_i = ax_i + b + w \)

where \( w \) is Gaussian,  \( P(w) \propto \exp\left(-\frac{1}{2\sigma}w^2\right) \)

total probability of the data is

\[
P(Y \mid a, b) = \prod_i P(y_i \mid a, b) \propto \exp\left(-\frac{1}{2\sigma} \sum_i (y_i - (ax_i + b))^2\right)
\]

Bayes: posterior (with flat prior)  \( P(a, b \mid Y) \propto P(Y \mid a, b) \)
A model of noise

A model with noise: \( y_i = ax_i + b + w \)

where \( w \) is Gaussian, \( P(w) \propto \exp\left(-\frac{1}{2\sigma} w^2\right) \)

total probability of the data is

\[
P(Y \mid a, b) = \prod_i P(y_i \mid a, b)
\]

\[
\propto \exp\left(-\frac{1}{2\sigma} \sum_i (y_i - (ax_i + b))^2\right)
\]

Bayes: posterior (with flat prior) \( P(a, b \mid Y) \propto P(Y \mid a, b) \)

least squares = maximum likelihood estimate
From probability to energy
From probability to energy

define the energy of \((a,b)\) as \(E = -\log P\)
From probability to energy

define the energy of \((a,b)\) as \(E = -\log P\)

\[
E = \frac{1}{2\sigma} \sum_i (y_i - (ax_i + b))^2
\]
From probability to energy

define the energy of \((a,b)\) as
\[
E = -\log P
\]

\[
E = \frac{1}{2\sigma} \sum_i (y_i - (ax_i + b))^2
\]

springs between the model and data

\[
E = \frac{1}{2} k x^2
\]
From probability to energy

define the energy of \((a,b)\) as \(E = -\log P\)

\[
E = \frac{1}{2\sigma} \sum_i (y_i - (ax_i + b))^2
\]

springs between the model and data

\[
E = \frac{1}{2} kx^2
\]

maximizing \(P = \text{minimizing } E\)
From probability to energy

define the energy of \((a, b)\) as \(E = -\log P\)

\[
E = \frac{1}{2\sigma} \sum_i (y_i - (ax_i + b))^2
\]

springs between the model and data

\[
E = \frac{1}{2} kx^2
\]

maximizing \(P = \text{minimizing } E\)

maximum likelihood estimate = ground state
From probability to energy

define the energy of \((a,b)\) as \( E = -\log P \)

\[
E = \frac{1}{2\sigma} \sum_i (y_i - (ax_i + b))^2
\]
springs between the model and data

\[
E = \frac{1}{2} kx^2
\]

maximizing \( P \) = minimizing \( E \)

maximum likelihood estimate = ground state

but what if the energy were different?
Changing the model
Changing the model

outliers skew our estimates
Changing the model

outliers skew our estimates

use a noise model with heavier tails

\[ P(w) \]
outliers skew our estimates

use a noise model with heavier tails

“gooey springs” that exert less force at large distances

Changing the model

\[
P(w) \quad E(w)
\]
Uncertainty, equilibrium, and the energy landscape
Uncertainty, equilibrium, and the energy landscape

[Bayes] don’t just give an estimate! what’s the posterior distribution?
Uncertainty, equilibrium, and the energy landscape

Bayes] don’t just give an estimate!
what’s the posterior distribution?

[Boltzmann] at thermal equilibrium,

\[ P(s) \propto e^{-E(s)/T} \]
Uncertainty, equilibrium, and the energy landscape

[Bayes] don’t just give an estimate! what’s the posterior distribution?

[Boltzmann] at thermal equilibrium,

\[ P(s) \propto e^{-E(s)/T} \]

low \( T \): concentrated on ground states
Uncertainty, equilibrium, and the energy landscape

[Bayes] don’t just give an estimate! what’s the posterior distribution?

[Boltzmann] at thermal equilibrium,

\[ P(s) \propto e^{-E(s)/T} \]

low \( T \): concentrated on ground states

high \( T \): uniform
[Bayes] don’t just give an estimate! what’s the posterior distribution?

[Boltzmann] at thermal equilibrium,

\[ P(s) \propto e^{-E(s)/T} \]

low \( T \): concentrated on ground states

high \( T \): uniform

thermal noise: \( T=\sigma \) (or looser springs)
Uncertainty, equilibrium, and the energy landscape

[Bayes] don’t just give an estimate! what’s the posterior distribution?

[Boltzmann] at thermal equilibrium,

\[ P(s) \propto e^{-E(s)/T} \]

low \( T \): concentrated on ground states

high \( T \): uniform

thermal noise: \( T=\sigma \) (or looser springs)

\( E(a,b) \) defined by model and data
[Bayes] don’t just give an estimate! what’s the posterior distribution?

[Boltzmann] at thermal equilibrium,

\[ P(s) \propto e^{-E(s)/T} \]

low \( T \): concentrated on ground states

high \( T \): uniform

thermal noise: \( T=\sigma \) (or looser springs)

\( E(a,b) \) defined by model and data

posterior distribution = equilibrium
Uncertainty, equilibrium, and the energy landscape

[Bayes] don’t just give an estimate!
what’s the posterior distribution?

[Boltzmann] at thermal equilibrium,

\[ P(s) \propto e^{-E(s)/T} \]

low \( T \): concentrated on ground states

high \( T \): uniform

thermal noise: \( T=\sigma \) (or looser springs)

\( E(a,b) \) defined by model and data

posterior distribution = equilibrium

in this case, landscape is simple and convex
The Ising model of magnetism
The Ising model of magnetism

the atoms of a block of iron interact with their neighbors
The Ising model of magnetism

the atoms of a block of iron interact with their neighbors when these interactions are strong enough, or the temperature is low enough, they line up and form a magnetic field
The Ising model of magnetism

the atoms of a block of iron interact with their neighbors when these interactions are strong enough, or the temperature is low enough, they line up and form a magnetic field

each site has a spin $s_i = \pm 1$ and (ferromagnet) $E = -J \sum_{(i,j)} s_i s_j$
The Ising model of magnetism

the atoms of a block of iron interact with their neighbors

when these interactions are strong enough, or the temperature is low enough, they line up and form a magnetic field

each site has a spin \( s_i = \pm 1 \) and (ferromagnet) \( E = -J \sum_{(i,j)} s_i s_j \)

ground state: all up or all down
The Ising model of magnetism

the atoms of a block of iron interact with their neighbors when these interactions are strong enough, or the temperature is low enough, they line up and form a magnetic field

each site has a spin $s_i = \pm 1$ and (ferromagnet) $E = -J \sum_{(i,j)} s_i s_j$

ground state: all up or all down

how does the magnetization $\left| \frac{1}{n} \sum_i s_i \right|$ vary with temperature?
The Ising model of magnetism
The Ising model of magnetism

At a critical temperature, the iron suddenly loses its magnetic field.
The Ising model of magnetism

at a critical temperature, the iron suddenly loses its magnetic field.
The Ising model of magnetism

at a critical temperature, the iron suddenly loses its magnetic field
atoms become uncorrelated: \( \langle s_is_j \rangle \sim e^{-r/\ell} \), no long-range information
The Ising model of magnetism

at a critical temperature, the iron suddenly loses its magnetic field

atoms become uncorrelated: $\langle s_i s_j \rangle \sim e^{-r/\ell}$, no long-range information
The Ising model of magnetism

at a critical temperature, the iron suddenly loses its magnetic field

atoms become uncorrelated: $\langle s_i s_j \rangle \sim e^{-r/\ell}$, no long-range information
Bumpy landscapes
Bumpy landscapes

least squares has a landscape with one optimum, and the Ising model has two
Bumpy landscapes

least squares has a landscape with one optimum, and the Ising model has two
Bumpy landscapes

least squares has a landscape with one optimum, and the Ising model has two
but a “spin glass” with energy $E = - \sum_{(i,j)} J_{ij} s_i s_j$ can have exponentially many
Bumpy landscapes

least squares has a landscape with one optimum, and the Ising model has two but a "spin glass" with energy $E = - \sum_{(i,j)} J_{ij} s_i s_j$ can have exponentially many

suppose the interactions $J_{ij}$ depend on the data and the model
Bumpy landscapes

least squares has a landscape with one optimum, and the Ising model has two but a “spin glass” with energy $E = - \sum_{(i,j)} J_{ij} s_i s_j$ can have exponentially many suppose the interactions $J_{ij}$ depend on the data and the model which local optimum is the true one?
Bumpy landscapes

least squares has a landscape with one optimum, and the Ising model has two but a “spin glass” with energy $E = - \sum_{(i,j)} J_{ij} s_i s_j$ can have exponentially many suppose the interactions $J_{ij}$ depend on the data and the model which local optimum is the true one?

can we find it efficiently? can we find it at all, given the posterior distribution?
Bumpy landscapes

least squares has a landscape with one optimum, and the Ising model has two
but a “spin glass” with energy $E = - \sum_{(i,j)} J_{ij} s_i s_j$ can have exponentially many
suppose the interactions $J_{ij}$ depend on the data and the model
which local optimum is the true one?
can we find it efficiently? can we find it at all, given the posterior distribution?
let’s look at a classic problem in social networks…
Divided we blog

ing corrected and uncorrected blockmodels with \( K = 2 \), we find the results shown in Fig. 1. As pointed out also by other authors [11, 30], the non-degree-corrected blockmodel fails to split the network into the known factions (indicated by the dashed line in the figure), instead splitting it into a group composed of high-degree vertices and another of low. The degree-corrected model, on the other hand, splits the vertices according to the known communities, except for the misidentification of one vertex on the boundary of the two groups. (The same vertex is also misplaced by a number of other commonly used community detection algorithms.

The failure of the uncorrected model in this context is precisely because it does not take the degree sequence into account. The \textit{apriori} probability of an edge between two vertices varies as the product of their degrees, a variation that can be fit by the uncorrected blockmodel if we divide the network into high- and low-degree groups. Given that we have only one set of groups to assign, however, we are obliged to choose between this fit and the true community structure. In the present case it turns out that the division into high and low degrees gives the higher likelihood and so it is this division that the algorithm returns. In the degree-corrected blockmodel, by contrast, the variation of edge probability with degree is already included in the functional form of the likelihood, which frees up the block structure for fitting to the true communities.

Moreover it is apparent that this behavior is not limited to the case \( K = 2 \). For \( K = 3 \), the ordinary stochastic blockmodel will, for sufficiently heterogeneous degrees, be biased towards splitting into three groups by degree—high, medium, and low—and similarly for higher values of \( K \). It is of course possible that the true community structure itself corresponds entirely or mainly to groups of high and low degree, but we only want our model to find this structure if it is still statistically surprising once we know about the degree sequence, and this is precisely what the corrected model does.

As a second real-world example we show in Fig. 2 an application to a network of political blogs assembled by Adamic and Glance [31]. This network is composed of blogs (i.e., personal or group web diaries) about US politics and the web links between them, as captured on a single day in 2005. The blogs have known political leanings and were labeled by Adamic and Glance as either liberal or conservative in the data set. We consider the network in undirected form and examine only the largest connected component, which has 1222 vertices. Figure 2 shows that, as with the karate club, the uncorrected stochastic blockmodel splits the vertices into high- and low-degree groups, while the degree-corrected model finds a split more aligned with the political division of the network. While not matching the known labeling exactly, the split generated by the degree-corrected model has a normalized mutual information of 0.72 with the labeling of Adamic and Glance, compared with 0.0001 for the uncorrected model.

(a) Without degree-correction
(b) With degree-correction
FIG. 2: Divisions of the political blog network found using the (a) uncorrected and (b) corrected blockmodels. The size of a vertex is proportional to its degree and vertex color reflects inferred group membership. The division in (b) corresponds roughly to the division between liberal and conservative blogs given in [31].

(To make sure that these results were not due to a failure of the heuristic optimization scheme, we also checked that the group assignments found by the heuristic have a higher objective score than the known group assignments, and that using the known assignments as the initial condition for the optimization recovers the same group assignments as found with random initial conditions.)

B. Generation of synthetic networks

We turn now to synthetic networks. The networks we use are themselves generated from the degree-corrected [Adamic & Glance]
Who eats whom
I record that I was born on a Friday
The stochastic block model
The stochastic block model

nodes have discrete labels: $k$ “groups” or types of nodes
The stochastic block model

nodes have discrete labels: $k$ “groups” or types of nodes

$k \times k$ matrix $p$ of connection probabilities
The stochastic block model

nodes have discrete labels: $k$ “groups” or types of nodes

$k \times k$ matrix $p$ of connection probabilities

if $t_i = r$ and $t_j = s$, there is a link $i \rightarrow j$ with probability $p_{rs}$
The stochastic block model

Nodes have discrete labels: $k$ “groups” or types of nodes

$k \times k$ matrix $p$ of connection probabilities

If $t_i = r$ and $t_j = s$, there is a link $i \rightarrow j$ with probability $p_{rs}$

Sparse: $p = O(1/n)$
The stochastic block model

nodes have discrete labels: $k$ “groups” or types of nodes

$k \times k$ matrix $p$ of connection probabilities

if $t_i=r$ and $t_j=s$, there is a link $i \to j$ with probability $p_{rs}$

sparse: $p=O(1/n)$

popular special case:

$$p = \frac{1}{n} \begin{pmatrix} c_{in} & \cdots & c_{out} \\ \vdots & \ddots & \vdots \\ c_{out} & & c_{in} \end{pmatrix}$$
The stochastic block model

nodes have discrete labels: $k$ “groups” or types of nodes

$k \times k$ matrix $p$ of connection probabilities

if $t_i=r$ and $t_j=s$, there is a link $i \rightarrow j$ with probability $p_{rs}$

sparse: $p=O(1/n)$

popular special case:

$$p = \frac{1}{n} \begin{pmatrix} c_{in} & \cdots & c_{out} \\ \vdots & \ddots & \vdots \\ c_{out} & \cdots & c_{in} \end{pmatrix}$$

ferromagnetic ( assortative, homophilic) if $c_{in} > c_{out}$
Likelihood and energy
Likelihood and energy

the probability of $G$ given the types $t$ is a product over edges and non-edges:

$$P(G \mid t) = \prod_{(i,j) \in E} p_{t_i,t_j} \prod_{(i,j) \notin E} (1 - p_{t_i,t_j})$$
Likelihood and energy

the probability of $G$ given the types $t$ is a product over edges and non-edges:

$$P(G \mid t) = \prod_{(i,j) \in E} p_{t_i,t_j} \prod_{(i,j) \notin E} (1 - p_{t_i,t_j})$$

the corresponding energy is

$$E(t) = -\log P(G \mid t) = - \sum_{(i,j) \in E} \log p_{t_i,t_j} - \sum_{(i,j) \notin E} \log(1 - p_{t_i,t_j})$$
Likelihood and energy

The probability of $G$ given the types $t$ is a product over edges and non-edges:

$$P(G | t) = \prod_{(i,j) \in E} p_{t_i,t_j} \prod_{(i,j) \notin E} (1 - p_{t_i,t_j})$$

The corresponding energy is

$$E(t) = -\log P(G | t) = - \sum_{(i,j) \in E} \log p_{t_i,t_j} - \sum_{(i,j) \notin E} \log(1 - p_{t_i,t_j})$$

Like Ising model, but with weak antiferromagnetic interactions on non-edges.
Likelihood and energy

the probability of $G$ given the types $t$ is a product over edges and non-edges:

$$P(G \mid t) = \prod_{(i,j) \in E} p_{t_i,t_j} \prod_{(i,j) \notin E} (1 - p_{t_i,t_j})$$

the corresponding energy is

$$E(t) = -\log P(G \mid t) = - \sum_{(i,j) \in E} \log p_{t_i,t_j} - \sum_{(i,j) \notin E} \log(1 - p_{t_i,t_j})$$

like Ising model, but with weak antiferromagnetic interactions on non-edges

what can we learn from the “physics” of the block model?
Ground states vs. the landscape
Ground states vs. the landscape

even random graphs have good-looking communities: only 11% of edges cross!
Ground states vs. the landscape

even random graphs have good-looking communities: only 11% of edges cross!
Ground states vs. the landscape

even random graphs have good-looking communities: only 11% of edges cross!
many local optima, with nothing in common
Ground states vs. the landscape

even random graphs have good-looking communities: only 11% of edges cross!
many local optima, with nothing in common
we need to understand the entire landscape, not just the optimum
Ground states vs. the landscape

even random graphs have good-looking communities: only 11% of edges cross!
many local optima, with nothing in common
we need to understand the entire landscape, not just the optimum
otherwise, we could be overfitting…
Overfitting
Overfitting

we, and our algorithms, are prone to false positives
Overfitting

we, and our algorithms, are prone to false positives
fitting the data with fancy models is tempting…
Overfitting

we, and our algorithms, are prone to false positives
fitting the data with fancy models is tempting...
we, and our algorithms, are prone to false positives
fitting the data with fancy models is tempting...
we, and our algorithms, are prone to false positives
fitting the data with fancy models is tempting...
Overfitting

we, and our algorithms, are prone to false positives

fitting the data with fancy models is tempting...
we, and our algorithms, are prone to false positives

fitting the data with fancy models is tempting…

but often we’re really fitting the noise, not the underlying process
we, and our algorithms, are prone to false positives
fitting the data with fancy models is tempting…
but often we’re really fitting the noise, not the underlying process
we want to understand the coin, not the coin flips
Information in the block model: the effect of a link
Information in the block model: the effect of a link

$k$ equal groups, $p = \frac{1}{n} \begin{pmatrix} c_{\text{in}} & \cdots & c_{\text{out}} \\ \vdots & \ddots & \vdots \\ c_{\text{out}} & & c_{\text{in}} \end{pmatrix}$: average degree $c = \frac{c_{\text{in}} + (k - 1)c_{\text{out}}}{k}$
Information in the block model: the effect of a link

\[ k \text{ equal groups, } p = \frac{1}{n} \begin{pmatrix} c_{\text{in}} & \cdots & c_{\text{out}} \\ \vdots & \ddots & \vdots \\ c_{\text{out}} & & c_{\text{in}} \end{pmatrix} \text{: average degree } c = \frac{c_{\text{in}} + (k - 1)c_{\text{out}}}{k} \]

if there is a link \( i \rightarrow j \), the probability distribution of \( t_j \) is related to that of \( t_i \) by a transition matrix
Information in the block model: the effect of a link

$k$ equal groups, $p = \frac{1}{n} \begin{pmatrix} c_{in} & \cdots & c_{out} \\ \vdots & \ddots & \vdots \\ c_{out} & & c_{in} \end{pmatrix}$: average degree $c = \frac{c_{in} + (k-1)c_{out}}{k}$

if there is a link $i \rightarrow j$, the probability distribution of $t_j$ is related to that of $t_i$ by a transition matrix

\[
\frac{1}{kc} \begin{pmatrix} c_{in} & \cdots & c_{out} \\ \vdots & \ddots & \vdots \\ c_{out} & & c_{in} \end{pmatrix} = \lambda \mathbf{1} + (1 - \lambda) \begin{pmatrix} 1/k & \cdots & 1/k \\ \vdots & \ddots & \vdots \\ 1/k & & 1/k \end{pmatrix}
\]

where $\lambda = \frac{c_{in} - c_{out}}{kc}$
Information in the block model: the effect of a link

\[ k \text{ equal groups, } p = \frac{1}{n} \begin{pmatrix} c_{\text{in}} & \cdots & c_{\text{out}} \\ \vdots & \ddots & \vdots \\ c_{\text{out}} & \cdots & c_{\text{in}} \end{pmatrix} : \text{average degree } c = \frac{c_{\text{in}} + (k - 1)c_{\text{out}}}{k} \]

if there is a link \( i \rightarrow j \), the probability distribution of \( t_j \) is related to that of \( t_i \) by a transition matrix

\[
\frac{1}{kc} \begin{pmatrix} c_{\text{in}} & \cdots & c_{\text{out}} \\ \vdots & \ddots & \vdots \\ c_{\text{out}} & \cdots & c_{\text{in}} \end{pmatrix} = \lambda \mathbb{1} + (1 - \lambda) \begin{pmatrix} 1/k & \cdots & 1/k \\ \vdots & \ddots & \vdots \\ 1/k & \cdots & 1/k \end{pmatrix}
\]

where
\[
\lambda = \frac{c_{\text{in}} - c_{\text{out}}}{kc}
\]

with probability \( \lambda \), copy from \( i \) to \( j \); with probability \( 1 - \lambda \), set \( j \)'s type randomly
Information in the block model: the effect of a link

$k$ equal groups,  $p = \frac{1}{n} \begin{pmatrix} c_{in} & \cdots & c_{out} \\ \vdots & \ddots & \vdots \\ c_{out} & \cdots & c_{in} \end{pmatrix}$: average degree $c = \frac{c_{in} + (k-1)c_{out}}{k}$

if there is a link $i \to j$, the probability distribution of $t_j$ is related to that of $t_i$ by a transition matrix

$$\frac{1}{kc} \begin{pmatrix} c_{in} & \cdots & c_{out} \\ \vdots & \ddots & \vdots \\ c_{out} & \cdots & c_{in} \end{pmatrix} = \lambda 1 + (1-\lambda) \begin{pmatrix} 1/k & \cdots & 1/k \\ \vdots & \ddots & \vdots \\ 1/k & \cdots & 1/k \end{pmatrix}$$

where $\lambda = \frac{c_{in} - c_{out}}{kc}$

with probability $\lambda$, copy from $i$ to $j$; with probability $1 - \lambda$, set $j$’s type randomly

if $\lambda$ is fixed, community detection gets easier as $c$ increases…
Detectability thresholds

For two groups of equal size [DKMZ, MNS, M, KMMNSSZ, BLM]: 
Detectability thresholds

For two groups of equal size [DKMZ, MNS, M, KMMNSSZ, BLM]:

\[ \frac{1}{\lambda^2} \]

Accuracy vs. \( c \)

(Chance)
Detectability thresholds

For two groups of equal size [DKMZ, MNS, M, KMMNSSZ, BLM]:

- easy: efficient algorithms (belief propagation, spectral)

Accuracy vs. $c$: 1 (chance) vs. $\frac{1}{\lambda^2}$
Detectability thresholds

For two groups of equal size [DKMZ, MNS, M, KMMNSSZ, BLM]:

- **Easy**: efficient algorithms (belief propagation, spectral)
- **Impossible**; can’t do better than a coin flip, or even distinguish from a purely random graph \( G(n, p = c/n) \)
Detectability thresholds

For $k \geq 4$ groups [DKMZ, KMMNSSZ, BLM, BMNN, AS]:
Detectability thresholds

For $k \geq 4$ groups [DKMZ, KMMNSSZ, BLM, BMNN, AS]:

$$O\left(\frac{\log k}{k\lambda^2}\right) \quad \text{to} \quad \frac{1}{\lambda^2}$$
Detectability thresholds

For $k \geq 4$ groups [DKMZ, KMMNSSZ, BLM, BMNN, AS]:

\[ O\left(\frac{\log k}{k \lambda^2}\right) \]  

\[ \frac{1}{\lambda^2} \]

easy: efficient algorithms (belief propagation, spectral)
Detectability thresholds

For $k \geq 4$ groups [DKMZ, KMMNSSZ, BLM, BMNN, AS]:

- Information-theoretically impossible accuracy $c_1 = \log k$.
- Easy: efficient algorithms (belief propagation, spectral).

Graph showing accuracy vs. $c$, with $O\left(\frac{\log k}{k\lambda^2}\right)$ and $\frac{1}{\lambda^2}$ thresholds.
Detectability thresholds

For $k \geq 4$ groups [DKMZ, KMMNSSZ, BLM, BMNN, AS]:

- Information-theoretically impossible
- $O\left(\frac{\log k}{k \lambda^2}\right)$
- Information-theoretically possible; but computationally hard?
- Easy: efficient algorithms (belief propagation, spectral)
Detectability thresholds

For \( k \geq 4 \) groups [DKMZ, KMMNSSZ, BLM, BMNN, AS]:

- Information-theoretically impossible
- Clusters, but can’t tell which is the true one: overfitting
- Information-theoretically possible; but computationally hard?
- Easy: efficient algorithms (belief propagation, spectral)
Clustering high-dimensional data
Clustering high-dimensional data

$m$ points in $n$-dimensional space, where $m=O(n)$
Clustering high-dimensional data

$m$ points in $n$-dimensional space, where $m=O(n)$

$k$ clusters with Gaussian noise
Clustering high-dimensional data

$m$ points in $n$-dimensional space, where $m=O(n)$

$k$ clusters with Gaussian noise

when can we...
Clustering high-dimensional data

$m$ points in $n$-dimensional space, where $m=O(n)$

$k$ clusters with Gaussian noise

when can we...

find the cluster centers?
$m$ points in $n$-dimensional space, where $m=O(n)$

$k$ clusters with Gaussian noise

when can we...

find the cluster centers?

label the points better than chance?
Clustering high-dimensional data

$m$ points in $n$-dimensional space, where $m = O(n)$

$k$ clusters with Gaussian noise

when can we...

find the cluster centers?

label the points better than chance?

tell that there are clusters, i.e., distinguish from a null model with one big cluster?
Clustering high-dimensional data

$m$ points in $n$-dimensional space, where $m=O(n)$

$k$ clusters with Gaussian noise

when can we...

find the cluster centers?

label the points better than chance?

tell that there are clusters, i.e., distinguish from a null model with one big cluster?

phase transitions as a function of noise vs. cluster distances, and $m/n$
Clustering high-dimensional data

$m$ points in $n$-dimensional space, where $m=O(n)$

$k$ clusters with Gaussian noise

when can we...

find the cluster centers?

label the points better than chance?

tell that there are clusters, i.e., distinguish from a null model with one big cluster?

phase transitions as a function of noise vs. cluster distances, and $m/n$

when $k$ is large enough, we can do better (information-theoretically) than PCA
Techniques
Techniques

If we iteratively estimate the probabilities with which nodes belong to groups, can we avoid a fixed point where each node is equally likely to be in each group? What can we learn about the ancestor of a family tree from its descendants?
If we iteratively estimate the probabilities with which nodes belong to groups, can we avoid a fixed point where each node is equally likely to be in each group? What can we learn about the ancestor of a family tree from its descendants?

How does community structure affect random walks (or epidemics) on networks? When does it show up in the spectrum of the adjacency matrix? When is it dominated by the randomness in the graph?
Techniques

If we iteratively estimate the probabilities with which nodes belong to groups, can we avoid a fixed point where each node is equally likely to be in each group? What can we learn about the ancestor of a family tree from its descendants?

How does community structure affect random walks (or epidemics) on networks? When does it show up in the spectrum of the adjacency matrix? When is it dominated by the randomness in the graph?

How can we tell the difference between the block model and a null model with no community structure? Can we bound the likelihood ratio between them? How can we tell when an apparent community is real, instead of overfitting?
If we iteratively estimate the probabilities with which nodes belong to groups, can we avoid a fixed point where each node is equally likely to be in each group? What can we learn about the ancestor of a family tree from its descendants?

How does community structure affect random walks (or epidemics) on networks? When does it show up in the spectrum of the adjacency matrix? When is it dominated by the randomness in the graph?

How can we tell the difference between the block model and a null model with no community structure? Can we bound the likelihood ratio between them? How can we tell when an apparent community is real, instead of overfitting?

Next two lectures!
A little light reading

To put it bluntly: this book rocks! It somehow manages to combine the fun of a popular book with the intellectual heft of a textbook.

Scott Aaronson, MIT

This is, simply put, the best-written book on the theory of computation I have ever read; one of the best-written mathematical books I have ever read, period.

Cosma Shalizi, Carnegie Mellon

www.nature-of-computation.org