Black Cats in Dark Rooms: Inference, Phase Transitions, and Networks

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

How can we find patterns in noisy data?
Phase transitions
Information vs. computation
The dynamics of algorithms
at a critical temperature, iron suddenly loses its magnetic field:
atoms become uncorrelated, no long-range information

Magnetick bodies made red hot lose their vertue
— Newton (and Curie)

at a critical temperature, iron suddenly loses its magnetic field:
atoms become uncorrelated, no long-range information
Phase transitions in inference: finding patterns in noisy data

when data is too noisy or too sparse, it suddenly becomes impossible to find the underlying pattern
The analogy between physics and inference

states of the atoms = hidden variables we want to infer
fields and interactions between them = data (and our model)
Gibbs distribution = posterior distribution given the data
temperature = noise
magnetization = accuracy
critical temperature = critical amount of noise where it becomes impossible to find patterns in data, or even tell if they exist
where are these phase transitions?
are there algorithms that succeed all the way up to these transitions?
let’s look at a classic problem in social networks…
Divided we blog

As pointed out also by other authors [11, 30], the non-degree-corrected block-model 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.

[Adamic & Glance]
Blogs again: hierarchical clustering

divide a network into subnetworks, until the remaining pieces have no statistically significant communities

[Zhang and Moore, *PNAS* 2014]
image by Tiago de Paula Peixoto
Ground states vs. the landscape

the most likely labeling (MLE, MAP) is the “ground state”

even random 3-regular graphs have labelings with only 11% edges crossing [Zdeborová & Boettcher] — exponentially many, which don’t agree!

we need to understand the entire landscape, not just the optimum
we, and our algorithms, are prone to false positives

fitting data with fancy models is tempting…

but often we’re really fitting the noise, not the underlying process
explore the landscape of models, not just the best one

if there is real structure in the data, there is a robust optimum

but the landscape can be “glassy”: many local optima with nothing in common

even if there is enough information in the data, the “true” solution might take exponential time to find: a dynamical transition
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:

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

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

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

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

using $P \sim e^{-\beta E}$ where $\beta = 1/T$, 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

what can we learn from the “physics” of the block model?
Belief propagation (a.k.a. the cavity method)

we want the marginals: the probability that each node belongs to each group.

each node $i$ sends a “message” to each of its neighbors $j$, giving an estimate of $i$’s marginal distribution...

based on the estimates it receives from its other neighbors $k$

avoids an “echo chamber” between pairs of nodes

update until we reach a fixed point (how many iterations? does it converge?)
Updating the beliefs

\[ \mu_{s}^{\rightarrow j} = \frac{1}{Z_{i \rightarrow j}} q_s \prod_{k \neq j} \sum_{(i,k) \in E} \mu_{r}^{k \rightarrow i} p_{rs} \times \prod_{k \neq j} \sum_{(i,k) \notin E} \mu_{r}^{k \rightarrow i} (1 - p_{rs}) \]

sparse case: can simplify by assuming that \( \mu_{r}^{k \rightarrow i} = \mu_{r}^{k} \) for all non-neighbors \( i \)

each update takes \( O(n+m) \) time, for constant \( k \)

WARNING: EXACT ONLY ON TREES

conditional independence
A phase transition: detectable to undetectable communities

when \( c_{\text{out}} / c_{\text{in}} \) is small enough, BP can find the communities

there is a regime where it can’t, and no algorithm can!

for 2 groups, the threshold is at

\[
|c_{\text{in}} - c_{\text{out}}| = 2\sqrt{c}
\]

there is a fixed point where all nodes have uniform marginals...

at the transition, it becomes stable

conjectured by [Decelle, Krzakala, Moore, Zdeborová, ‘11]
proved by [Mossel, Neeman, Sly, ‘13; Massoulié ‘13]
Information in the block model: 
the effect of a link

$k$ equal groups, $p = \frac{1}{n} \left( \begin{array}{ccc} c_{\text{in}} & \cdots & c_{\text{out}} \\ \vdots & \ddots & \vdots \\ c_{\text{out}} & \cdots & c_{\text{in}} \end{array} \right)$: 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

$$\frac{1}{k} \left( \begin{array}{ccc} c_{\text{in}} & \cdots & c_{\text{out}} \\ \vdots & \ddots & \vdots \\ c_{\text{out}} & \cdots & c_{\text{in}} \end{array} \right) = \lambda \mathbb{1} + (1 - \lambda) \left( \begin{array}{ccc} 1/k & \cdots & 1/k \\ \vdots & \ddots & \vdots \\ 1/k & \cdots & 1/k \end{array} \right)$$

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

if $\lambda$ is fixed, community detection gets easier as $c$ increases…
Detectability threshold for two groups: when the trivial fixed point becomes unstable

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

- easy: efficient algorithms (belief propagation, spectral)
- information-theoretically impossible; can’t do better than a coin flip, or even distinguish from a purely random graph $G(n, p = c/n)$
Another regime: detectable but hard

trivial fixed point is locally stable; accurate one has a small basin of attraction

a free energy barrier between “paramagnetic” and “ferromagnetic” phases

detection is information-theoretically possible, but we believe it’s computationally (and dynamically!) hard
Solutions hidden behind energy barriers: spin glasses

A glassy material can stay stuck in an amorphous state for exponential time, even though the minimum energy configuration is crystalline.
Solutions hidden behind energy barriers: spin glasses

Belief propagation falls into the trivial fixed point for almost all initial conditions: Monte Carlo would take exponential time to cross barrier.

Information-theoretically possible (using exhaustive search) but computationally / dynamically hard.
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?
- \( O\left(\frac{\log k}{k\lambda^2}\right) \)
- Easy: efficient algorithms (belief propagation, spectral)

Accuracy

\[ \frac{1}{\lambda^2} \]
Phase transitions with metadata: what if we know some labels?

suppose we are given the correct labels for $\alpha n$ nodes for free

can we extend this information to the rest of the graph?

when $\alpha$ is large enough, knowledge percolates from the known nodes to the rest of the network

a line of discontinuities in the $(c, \alpha)$ plane, ending at a critical point

[Zhang, Moore, Zdeborová ’14]
What if we don’t know how strong the structure is?

lower temperature = greedier algorithm = assume stronger structure

if we get too greedy, we enter a “spin glass” where BP fails to converge

[Diagram showing assumed structure vs. true structure with regions labeled SG, M, R, and P, along with Nishimori line and detectability transition]
Extensions to richer data

can add metadata to nodes and edges: signed or weighted edges, nodes with social status, location, content...

for networks of documents, a model that combines overlapping communities with standard models of word frequencies

a network of 1,000 microprocessor patents (joint work with Sergi Valverde):

using both text and links does better than either one alone

[Zhu, Yan, Getoor, Moore, KDD 2013]
many problems involving sparse, noisy data have phase transitions beyond which no algorithm can find the underlying pattern

networks, high-dimensional clustering, compressed sensing, sparse regression…

ideas from physics can help us find optimal algorithms that succeed all the way up to these transitions

the dynamics of these algorithms tells us about the computational complexity of these problems, and has analogies with glassy systems

“as simple as possible, but no simpler”
Everything is an ecology
Shameless plug

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