High-Dimensional Distribution Testing

Constantinos “Costis” Daskalakis

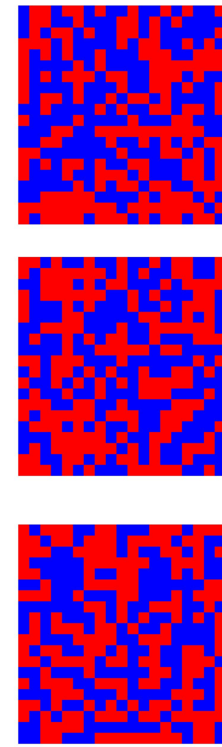
CSAIL and EECS, MIT

What properties do your BIG distributions have?

e.g. 1 Testing Uniformity

- Consider source p generating n-bit strings $\in \{0,1\}^n$
  - 0011010101 (sample 1)
  - 0101001110 (sample 2)
  - 0011110100 (sample 3)
  - …
- Is $p = U_{\{0,1\}^n}$ or is it far from uniform?

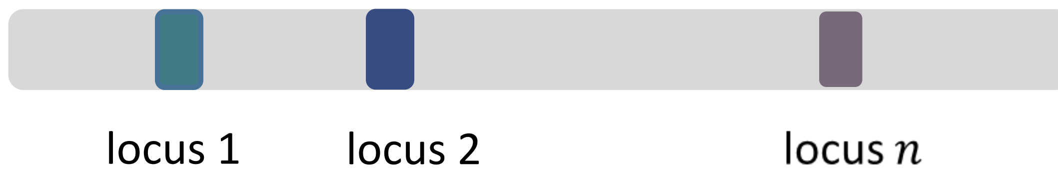
n bit images



⋮

e.g.2: Linkage Disequilibrium

Genome



Single Nucleotide Polymorphisms (SNPs), are they independent?

Suppose n loci, 2 possible states each, then:

- state of one's genome $\in \{0,1\}^n$
- **humans:** some distribution p over $\{0,1\}^n$

Question: Is p a product dist'n OR ***far*** from all product dist'ns?

1000 samples (you patients)

e.g.3: Behavior in a Social Network

Q: Are nodes behaving independently or far from independently?

Q': Do adopted technologies exhibit ***weak*** or ***strong*** network effects?

1 sample

Problem formulation

Distribution Property:

$\mathcal{P}$: subset of all distributions over $D = \Sigma^n$

– e.g. $\mathcal{P}$ = product measures, $\mathcal{P}$ = {uniform distribution over D}

Problem:

Given: samples from **unknown** p

TV (*c.f.* G's talk)

w/ prob ≥ 0.9, distinguish: $p \in \mathcal{P}$ **vs** $d(p, \mathcal{P}) > \varepsilon$

Objective

Minimize sample and time complexity $\ll |D|$?

[Acharya-Daskalakis-Kamath NIPS'15]: A broad set of properties $\mathcal{P}$ can be tested efficiently from an optimal $\Theta\left(\sqrt{|D|}/\varepsilon^2\right)$ number of samples.

- e.g. monotonicity and independence of high-dimensional dist'ns, unimodality, log-concavity, monotone-hazard rate of one-dimensional dist'ns
- c.f. **[Paninksi'04], [Valiant-Valiant'14], [Canonne et al'16]**

The sample complexity of $\Theta\left(|\Sigma|^{n/2}/\varepsilon^2\right)$ is optimal, but unsettling

What do we ***really*** know about our BIG distributions of interest?

Inspecting the LB Instance

- **Task:** Distinguish $p = U_{\{0,1\}^n}$ vs $d_{TV}(p, U_{\{0,1\}^n}) > \epsilon$?
  - **[Paninski'04]:** $\Theta\left(\frac{2^{n/2}}{\epsilon^2}\right)$ samples are necessary and sufficient
  - **"Proof:"**
    - Universe 1: p is uniform over $\{0,1\}^n$
    - Universe 2: p is randomly chosen as follows

      u.a.r.

      - if u, v differ only in last bit, set $(p_u, p_v) = \left(\frac{1}{2^n}(1 \pm \epsilon), \frac{1}{2^n}(1 \mp \epsilon)\right)$
    - average distribution in Universe 2 = uniform (formally use LeCam)

- To index a dist'n in Universe 2, need $2^n/2$ bits
- Nature doesn't have this many bits
  - often high dimensional systems have structure,
  - modeled as Markov Random Fields (MRFs), Bayesian Networks, etc

Testing high-dimensional distributions with structure?

Today’s Menu

- **Motivation**
- Testing Bayesian Networks
- Testing Ising Models
- Closing Thoughts

Today’s Menu

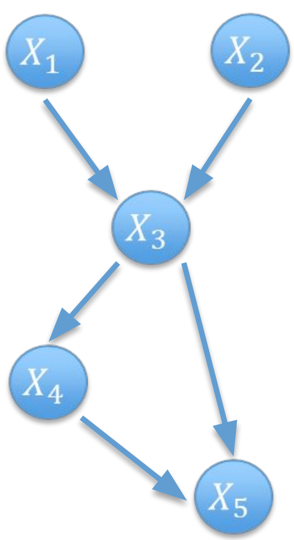
- **Motivation**
- **Testing Bayesian Networks**
- Testing Ising Models
- Closing Thoughts

Bayesian Networks

- Probability distribution defined in terms of a DAG $G = (V, E)$
- Nodes v associated w/ random variable $X_v \in \Sigma$
- Distribution factorizable in terms of parenthood relationships

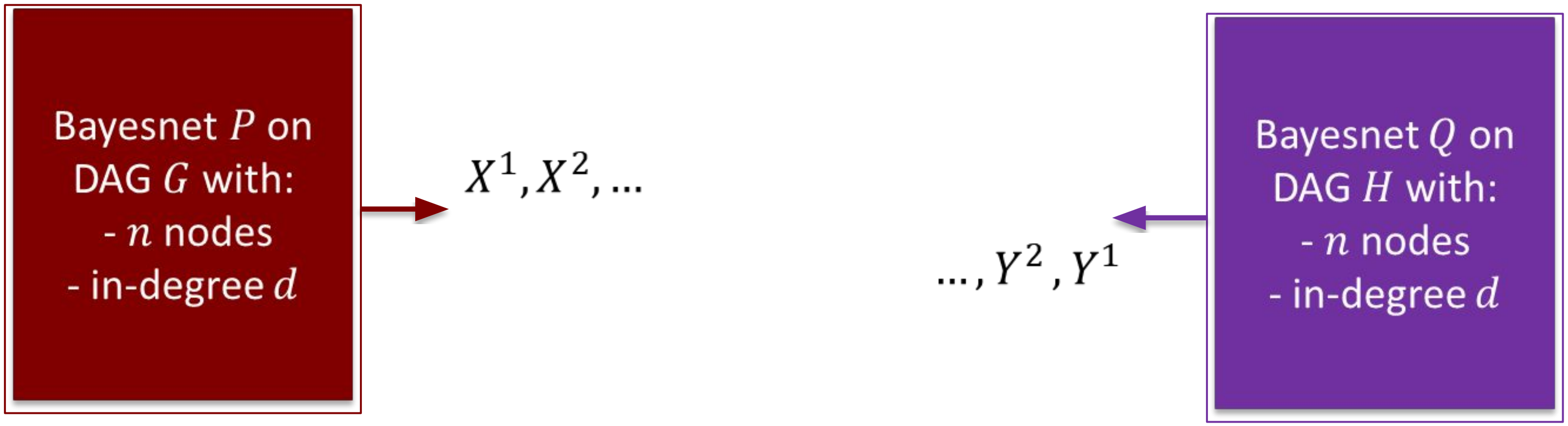
$$\Pr(x) = \prod_{v} \Pr_{X_v | X_{\Pi_v}}(x_v | x_{\Pi_v}), \forall x \in \Sigma^V$$

Parents of v in G



$$\Pr[\vec{x}] = \Pr[\mathrm{x}_1] \cdot \Pr[x_2] \cdot \Pr[x_3 | x_1, x_2] \cdot \Pr[x_4 | x_3] \cdot \Pr[x_5 | x_3, x_4]$$

Testing Bayesian Networks



Goal: distinguish $P = Q$ vs $d_{TV}(P, Q) > \varepsilon$

[Daskalakis-Pan COLT'17]: There exist efficient testers using:

- $\tilde{O}\left(\frac{|\Sigma|^{0.75\,(d+1)} n}{\varepsilon^2}\right)$ samples, if DAGs $G = H$ and unknown
- $\tilde{O}\left(\frac{|\Sigma|^{9/2} n}{\varepsilon^2}\right)$ samples, if G and H are unknown and potentially different trees

Moreover, the dependence on n, ε of both bounds is tight up to a $O(\log n)$ factor, and the exponential in d dependence is necessary and essentially tight.

[Canonne et al. COLT'17]: Identify conditions under which dependence on n can be made $\sqrt{n}$ when one of the two Bayesnets is known (goodness-of-fit problem)

Testing Bayesian Networks (cont'd)

Goal: distinguish $P = Q$ vs $d_{TV}(P,Q) > \varepsilon$

Idea: distance localization

- prove statements of the form: "If P and Q are far in TV, there exists a small size witness set S of variables such that P_S and Q_S, the marginals of P and Q on variables S, are also somewhat far away"
- reduces the original problem to identity testing on small size sets

Question: which distance to localize in?

Attempt 1: $d_{\mathrm{TV}}(P,Q) \le \sum_v d_{\mathrm{TV}}\left(P_{v\cup\Pi_v}, Q_{v\cup\Pi_v}\right) + \sum_v d_{\mathrm{TV}}\left(P_{\Pi_v}, Q_{\Pi_v}\right)$ (hybrid argument)

- Hence: $d_{\mathrm{TV}}(P,Q) > \varepsilon \;\Rightarrow\; \exists v \; s.t.\, d_{\mathrm{TV}}\left(P_{v\cup\Pi_v}, Q_{v\cup\Pi_v}\right) > \frac{\varepsilon}{2n}$ or $d_{\mathrm{TV}}\left(P_{\Pi_v}, Q_{\Pi_v}\right) > \frac{\varepsilon}{2n}$
- But leads to suboptimal sample complexity $\Omega_{d,|\Sigma|}\left(\frac{n^2}{\varepsilon^2}\right)$

Attempt 2: $KL(P||Q) \le \sum_v KL\left(P_{v\cup\Pi_v}||Q_{v\cup\Pi_v}\right)$ (chain rule of KL)

- Hence: $d_{\mathrm{TV}}(P,Q) > \varepsilon \;\Rightarrow KL(P||Q) > 2\varepsilon^2 \;\Rightarrow\; \exists v \; s.t.\, KL\left(P_{v\cup\Pi_v}||Q_{v\cup\Pi_v}\right) > \frac{2\varepsilon^2}{n}$
- But KL testing requires infinitely many samples, b.c. of low probability events ☹

Testing Bayesian Networks (cont'd)

Goal: distinguish $P = Q$ vs $d_{TV}(P,Q) > \varepsilon$
Idea: distance localization

- prove statements of the form: "If P and Q are far in TV, there exists a small size witness set S of variables such that P_S and Q_S, the marginals of P and Q on variables S, are also somewhat far away"
- reduces the original problem to identity testing on small size sets

Attempt 3: Use Hellinger distance!

- Defined as: $H(P,Q) = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_x \left(\sqrt{P(x)} - \sqrt{Q(x)}\right)^2}$
- Satisfies: $d_{\mathrm{TV}}(\mathrm{P},\mathrm{Q}) \leq \sqrt{2} \cdot H(P,Q) \leq \sqrt{KL(P||Q)}$

We show that H^2 satisfies subadditivity over neighborhoods:

$$H^2(P,Q) \leq \sum_v H^2\left(P_{v\cup\Pi_v}, Q_{v\cup\Pi_v}\right)$$

Hence: $d_{\mathrm{TV}}(P,Q) > \varepsilon \;\Rightarrow\; \exists v \; s.t. \, H^2\left(P_{v\cup\Pi_v}, Q_{v\cup\Pi_v}\right) > \frac{\varepsilon^2}{2n}$

c.f. G's talk: distinguishing $H^2\left(P_{v\cup\Pi_v}, Q_{v\cup\Pi_v}\right) = 0$ versus $> \frac{\varepsilon^2}{2n}$, requires $O_{d,|\Sigma|}\left(\frac{n}{\varepsilon^2}\right)$ samples

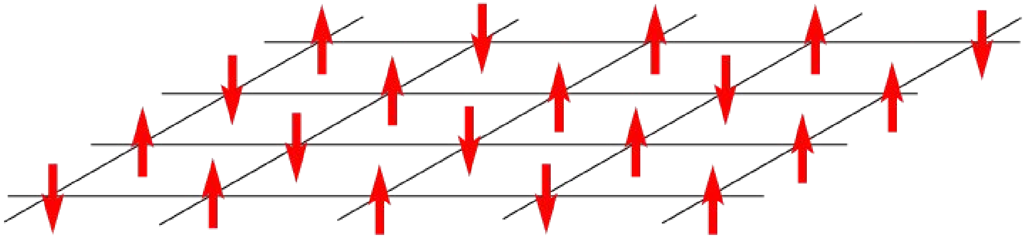
Today’s Menu

- **Motivation**
- **Testing Bayesian Networks**
- **Testing Ising Models**
- Closing Thoughts

Ising Model

- Probability distribution defined in terms of a graph $G = (V, E)$
- State space $\{\pm 1\}^V$
- Given edge potentials θ_e, node potentials θ_v:

$$p_\theta(x) \propto \exp\left(\sum_{e=(u,v)\in E} \theta_e x_u x_v + \sum_{v\in V} \theta_v x_v\right)$$



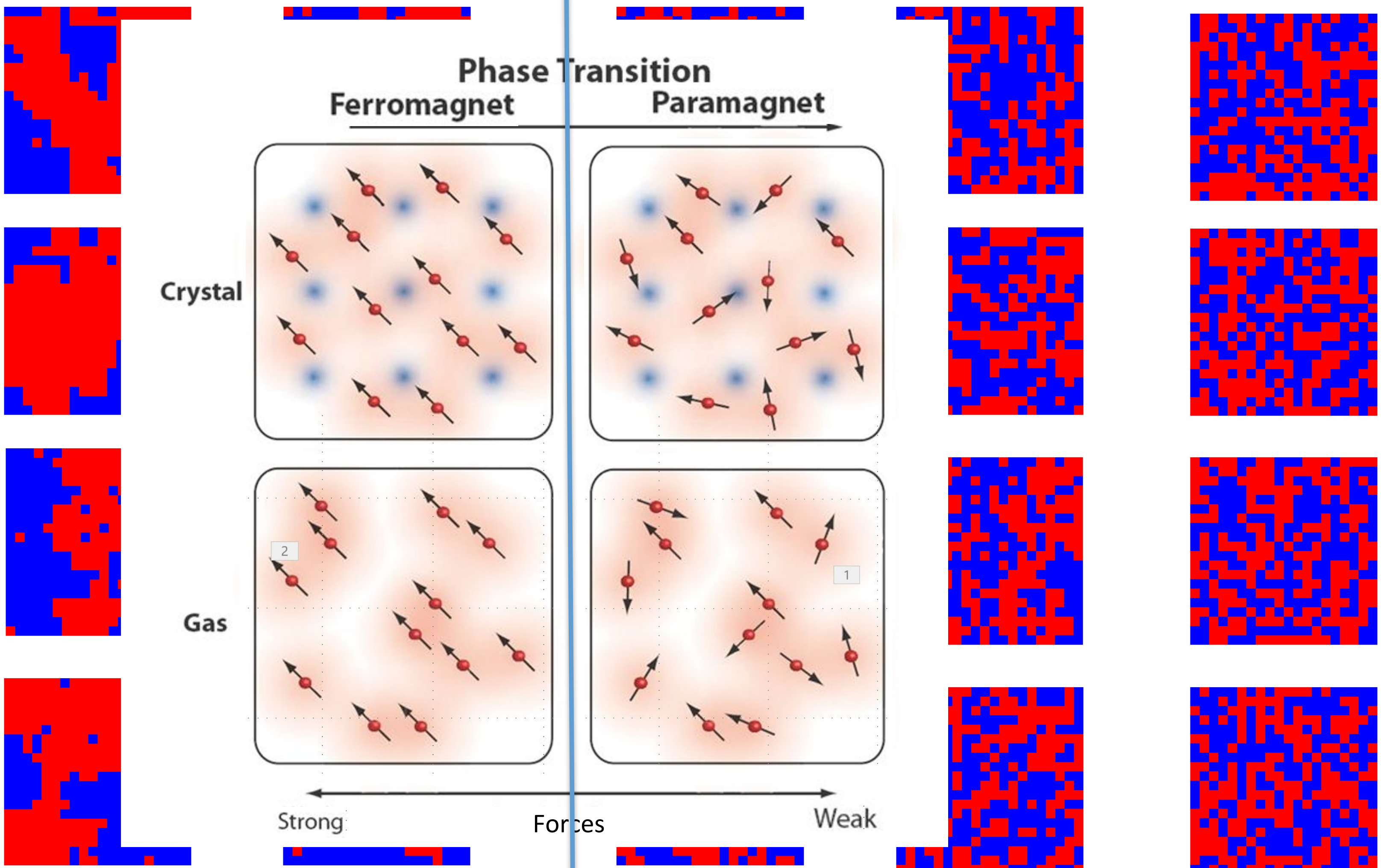
2–D Ising Model

  - High $|\theta_e|$'s $\Longrightarrow$ strongly (anti-)correlated spins
- Statistical physics, computer vision, neuroscience, social science

$\theta_v = 0$

Ising Model: Strong vs weak ties

"low temperature regime" "high temperature regime"



$\theta_e = 1$ $\theta_e = 0.5$ $\theta_e = 0.25$ $\theta_e = 0.125$ $\theta_e = 0$

Testing Ising Models

$$p_\theta(x) \propto \exp\left(\sum_{(u,v)} \theta_{uv} x_u x_v + \sum_{v \in V} \theta_v x_v\right)$$

- **Identity Testing:** Given sample access to two Ising models p_θ and $p_{\theta'}$,

  distinguish $p_\theta = p_{\theta'}$ vs $d_{\mathrm{TV}}(p_\theta, p_{\theta'}) > \varepsilon$

- **Independence Testing:** Given sample access to an Ising model p_θ,

  distinguish $p_\theta \in \mathcal{I}_{\{\pm 1\}^V}$ vs $\ell_1\left(p_\theta, \mathcal{I}_{\{\pm 1\}^V}\right) > \epsilon$

  product measures

- **[w/ Dikkala, Kamath SODA'18]:** small-poly $\left(n, \frac{1}{\epsilon}\right)$ samples suffice to do this efficiently
  - Poly depends on the regime: high vs low temperature, ferromagnetic ($\theta_{uv} \geq 0, \forall u, v$) vs non-ferromagnetic, non-external fields ($\theta_v = 0, \forall v$) vs external fields, tree vs general graph, independence vs identity, etc.
  - Technical vignettes: localization, concentration of measure

Testing Ising Models

$$p_\theta(x) \propto \exp\left(\sum_{(u,v)} \theta_{uv} x_u x_v + \sum_{v \in V} \theta_v x_v\right)$$

- **Identity Testing:** Given sample access to two Ising models p_θ and $p_{\theta'}$, distinguish $p_\theta = p_{\theta'}$ vs $d_{\mathrm{TV}}(p_\theta, p_{\theta'}) > \varepsilon$
- **Independence Testing:** Given sample access to an Ising model p_θ, distinguish $p_\theta \in \mathcal{I}_{\{\pm 1\}^V}$ vs $\ell_1\left(p_\theta, \mathcal{I}_{\{\pm 1\}^V}\right) > \epsilon$

  product measures
- **[w/ Dikkala, Kamath SODA'18]:** small-poly $\left(n, \frac{1}{\epsilon}\right)$ samples suffice to do this efficiently
  - Poly depends on the regime: high vs low temperature, ferromagnetic ($\theta_{uv} \geq 0, \forall u, v$) vs non-ferromagnetic, non-external fields ($\theta_v = 0, \forall v$) vs external fields, tree vs general graph, independence vs identity, etc.
  - Technical vignettes: localization, **concentration of measure**

Testing Ising Models

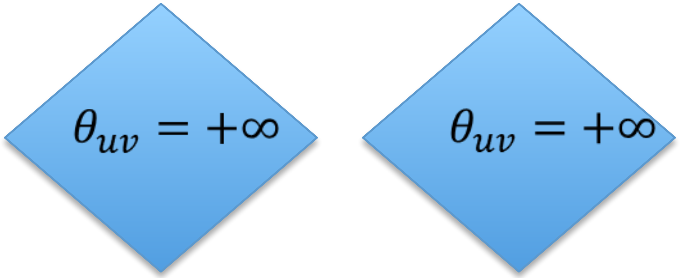
$$p_\theta(x) \propto \exp\left(\sum_{(u,v)} \theta_{uv} x_u x_v + \sum_{v \in V} \theta_v x_v\right)$$

- **Identity Testing:** Given sample access to two Ising models p_θ and $p_{\theta'}$,
  distinguish $p_\theta = p_{\theta'}$ vs $d_{\mathrm{TV}}(p_\theta, p_{\theta'}) > \varepsilon$
- **Independence Testing:** Given sample access to an Ising model p_θ,
  distinguish $p_\theta \in \mathcal{I}_{\{\pm1\}^V}$ vs $\ell_1\left(p_\theta, \mathcal{I}_{\{\pm1\}^V}\right) > \epsilon$
- Bi-linear functions of the Ising model serve as useful distinguishing statistics
- For $X \sim p_\theta$ consider:
  $f(X) = \sum_{u,v} c_{uv}(X_u - E[X_u])(X_v - E[X_v])$, where say $c_{uv} \in [\pm 1]$

- **Technical Challenge:** can't bound $\mathrm{Var}[f(X)]$ intelligently

- If $\theta_{uv} = 0, \forall uv$, then $\mathrm{Var}[f(X)] = n^2$
- O.w. best can say is (trivial) $Var[f(X)] = O(n^4)$
  - and, in fact, this is tight
    - consider two disjoint cliques with super-strong θ_{uv}'s inside, 0 across, and all θ_v's zero everywhere
    - suppose also $c_{u,v} = 1$, for all u, v
    - Then $f(X)$ dances around its mean by $\Omega(\mathrm{n}^2)$

Low temperature.
How about high temperature?



High Temperature Ising

- Several conditions
- Dobrushin's uniqueness criterion:

$$\max_{v} \sum_{u \neq v} \tanh(|\theta_{uv}|) < 1$$

- Think:

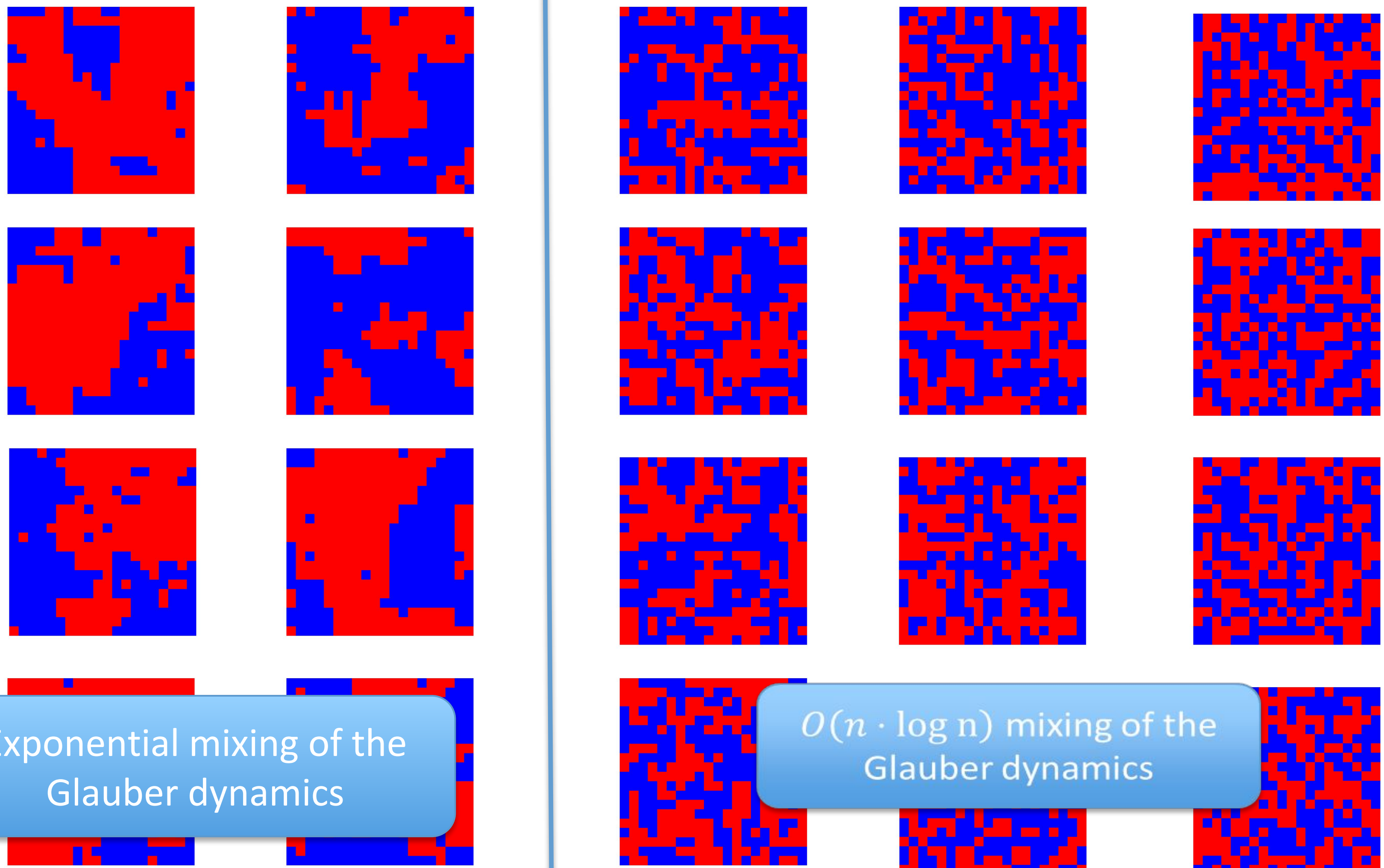
$$\max_{v} \sum_{u \neq v} |\theta_{uv}| < 1$$

- Implies:
  - $O(n \log n)$ mixing of natural MC
  - Correlation decay properties

$\theta_v = 0$

Ising Model: Strong vs weak ties

"low temperature regime" "high temperature regime"



Exponential mixing of the Glauber dynamics

$O(n \cdot \log n)$ mixing of the Glauber dynamics

$\theta_e = 1$ $\theta_e = 0.5$ θ $\theta_e = 0.25$ $\theta_e = 0.125$ $\theta_e = 0$

Testing Ising Models

$$p_\theta(x) \propto \exp\left(\sum_{(u,v)} \theta_{uv} x_u x_v + \sum_{v \in V} \theta_v x_v\right)$$

- **Identity Testing:** Given sample access to two Ising models p_θ and $p_{\theta'}$,
  distinguish $p_\theta = p_{\theta'}$ vs $d_{\mathrm{TV}}(p_\theta, p_{\theta'}) > \varepsilon$
- **Independence Testing:** Given sample access to an Ising model p_θ,
  distinguish $p_\theta \in \mathcal{I}_{\{\pm 1\}^V}$ vs $\ell_1\left(p_\theta, \mathcal{I}_{\{\pm 1\}^V}\right) > \epsilon$
- Bi-linear functions of the Ising model serve as useful distinguishing statistics
- For $X \sim p_\theta$ consider:

$$f(X) = \sum_{u,v} c_{uv} (X_u - E[X_u])(X_v - E[X_v])$$

- Low temperature: $Var[f(X)] = O(n^4)$
- **[w/ Dikkala, Kamath]:** High temperature: $Var[f(X)] = O(n^2)$
  - proof by tightening exchangeable pair technology **[Stein,…,Chatterjee 2006]**

Concentration of Measure

$$p_\theta(x) \propto \exp\left(\sum_{(u,v)} \theta_{uv} x_u x_v + \sum_{v \in V} \theta_v x_v\right)$$

- **[w/ Dikkala, Kamath NIPS'17]:** Under **high temperature**, any **centered polynomial function** of the Ising model concentrates **essentially as well as if the variables where independent**.
- **High temperature** = Dobrushin's condition holds, think $\|[\theta_{uv}]\|_\infty < 1$
- **Centered multi-linear function of degree d:**

$$f(X) = \sum_{S,|S| \le d} c_S \prod_{v \in S} (X_v - E[X_v])$$

- **Essentially as well as if the variables where independent:**

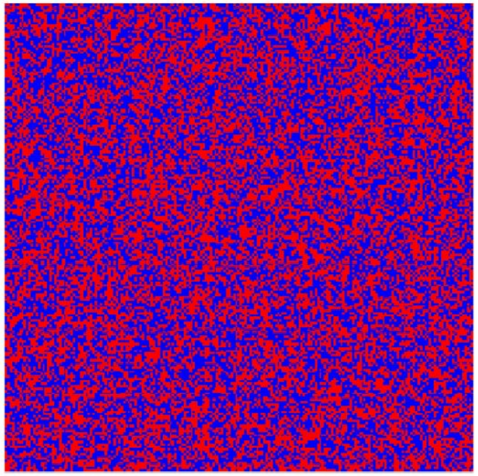
$$\Pr[|f(X) - E[f(X)]| > r] \le \exp\left(-\Omega_d\left(\frac{r^{\frac{2}{d}}}{n \log n}\right)\right)$$

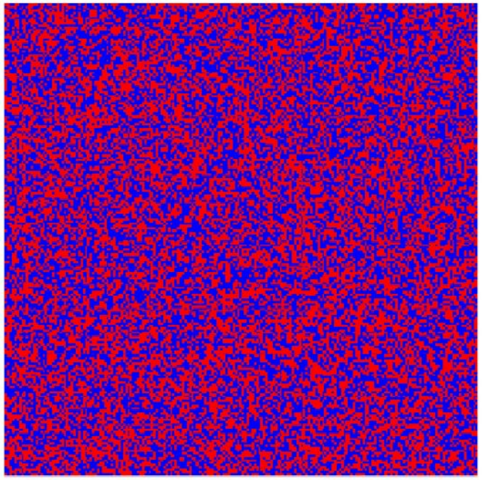
- Improves from known concentration results on Lipschitz fn's of Ising model
  - $n^{d-0.5} \to n^{d/2}$ radius of concentration

Using Concentration to Test

$$p_\theta(x) \propto \exp\left(\sum_{(u,v)} \theta\, x_u x_v + \sum_{v\in V} \theta_v x_v\right)$$

-
- Is it high-temperature Ising $\left(\theta < \theta_c = \frac{\ln(1+\sqrt{2})}{2}\right)$?



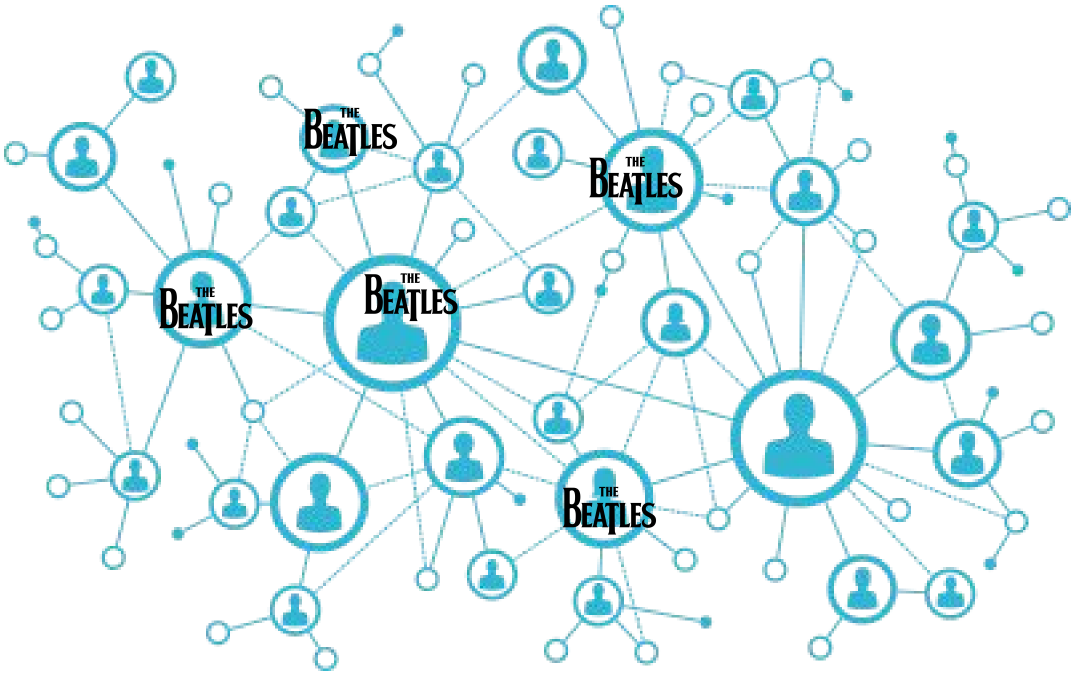


One is a sample from a product measure, the other is product measure but every node selects a friend or friend of friend and copies him with probability τ

Bilinear statistics catch the deviation at 10x smaller τ value compared to MLE on θ and comparison to θ_c

Testing Weak vs Strong Network Ties

e.g. Who listens to the Beatles?



Q: Given one sample (from last.fm dataset) of who does/doesn't listen to a particular band, can we reject the hypothesis that this decision comes from high-temperature Ising model (lack of long range correlation)?

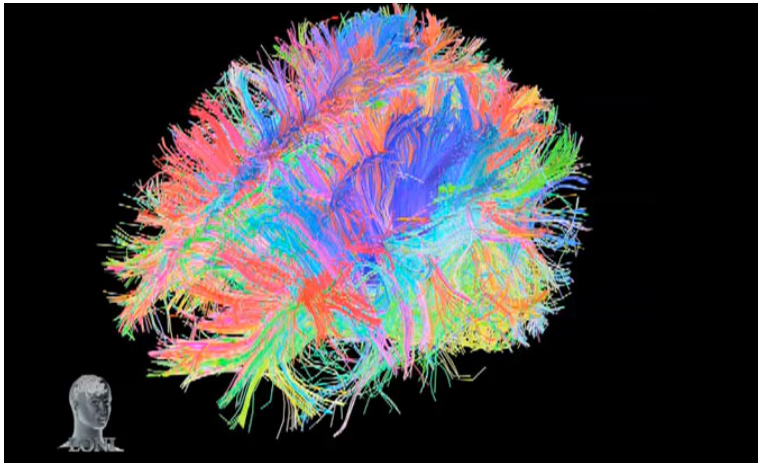
A: we can for Taylor Swift, Britney Spears, Katy Perry, Rihanna, Lady Gaga; we cannot for Beatles and Muse

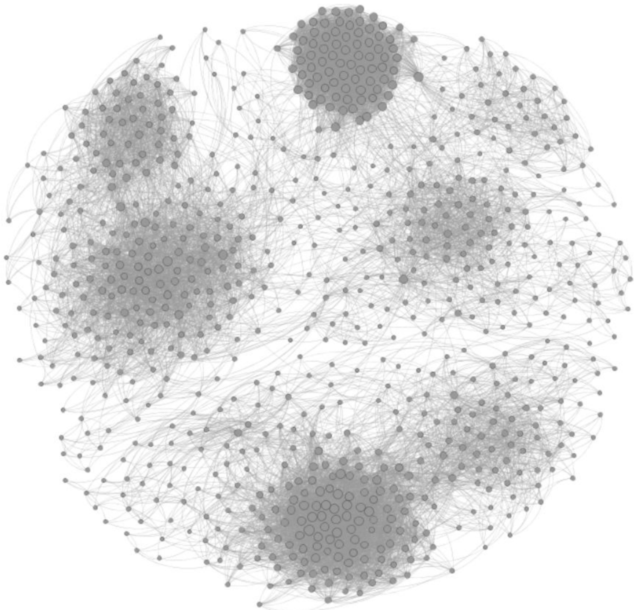
Conclusions

- Testing properties of high-dimensional distributions requires exponentially many samples
- Making assumptions about the distribution being sampled gives leverage
- [w/ Pan COLT'17]: Testing Bayes nets with linearly many samples
- [w/ Dikkala, Kamath SODA'18]: Testing Ising models with polynomially many samples
- [w/ Dikkala, Kamath NIPS'17]: Testing weak vs strong ties from **one sample**

Testing from a Single Sample

- Given **one** social network, **one** brain, etc., how can we test the validity of a certain generative model?
- Ongoing with Aliakbarpour-Rubinfeld-Zampetakis, testing preferential attachment models





Testing Markov Chains

- Given one trajectory of an unknown Markov Chain M, whose starting state we cannot control, can we test whether it came from a given Markov Chain M^* over n states?
- Question: test $M = M^* \quad \text{vs} \quad dist(M, M^*) > \varepsilon$

How to quantify distance between Markov chains?

- **[Ongoing w/ Dikkala, Gravin]:** We propose a distance measure capturing the limiting behavior of the TV distance between trajectories of the two chains

$$dist(M, M^*) = 1 - \rho\left(\sqrt{M_{ij} \cdot M_{ij}^*}\right)$$

- Show that **one trajectory** of n/ϵ^2 length suffices

Thanks!