Independence Testing for Bounded Degree Bayesian Network

Arnab Bhattacharyya[1] , Clément L. Canonne[2], Joy Qiping Yang[1]

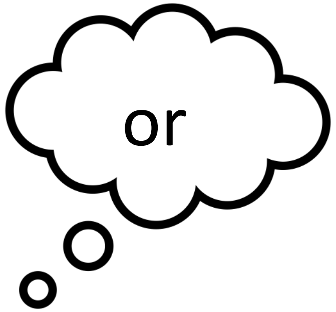
1. National University of Singapore
2. University of Sydney

Independence testing on hypercube

Product distribution: $Q = Q_{X_1} \otimes Q_{X_2} \otimes \cdots \otimes Q_{X_n}$

Sample access to P on $\{0,1\}^n$; is P a **product distribution** or ***ε*-far from it** -- ε-far from **every** product distribution?

- By confidence: correct probability at least 2/3 in both cases;
- Sample complexity: how many samples does it take?
- 2/3 prob can be boosted by a standard amplification trick (relatively cheaply).



Motivation – why independence testing?

(Statistical) independence is great!

- **Cheaper**: time/sample complexity. **Better** algorithms!
- **Applications**: drug response tests; genome analysis.

In the context of Bayesian Networks (bounded in-degree)

- **Stepping stone**: testing degree-k Bayes net. (more test problems!)
  - Independence test (degree-0 Bayes net; an empty graph).

Question of interest **(sample complexity)**:

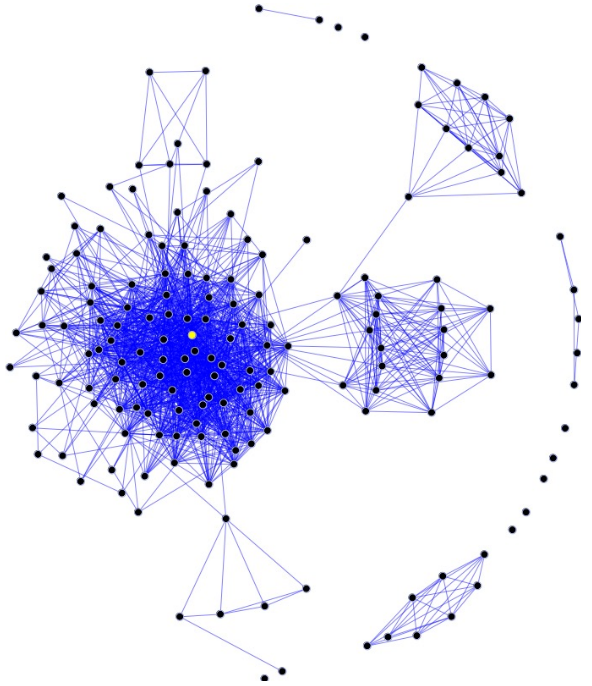
- Can we efficiently test this property (better than learning at least)?

Why bounded in-degree Bayesian Network?

- Joint density factorizes according to a directed acyclic graph (DAG):

$$P[X_1, \ldots, X_n] = \prod_{i=1}^{n} P[X_i | \mathrm{pa}(X_i)]$$

- (in)degree-d means that each node has **at most d parents**.
- Our regime of interest: small d and huge n.
- (Motivation again!) Common/natural setting:
  Genome; social media:
  (hundreds of local connections; millions of nodes).
- In particular: $d = O(\log n)$.



Equation credit: Wikipedia
Image credit: https://en.wikipedia.org/wiki/Social_graph

Prior results: high dim independence testing

Can we efficiently test independence?

- No -- not **in general**!
- **Known results** independence testing on $\{0,1\}^n$ in TV:
- Sample complexity on independence testing, in general. Known complexity results, tight, not great:

$$\Theta(2^{n/2}/\varepsilon^2).$$

- If P is known to be bounded degree Bayes net

$$\widetilde{O}(2^d n/\varepsilon^2)$$

Can we do better?

Sample complexity – what we show

- **Our key results**:
- Lower bound on testing **easy-to-learn** distributions* (e.g., **Bayes nets**).
- Assume bounded degree-d bayes net -- a near optimal bound at

$$\tilde{\Theta}(2^{d/2}n/\varepsilon^2)$$

- A tester that takes $\tilde{O}(2^{d/2}n/\varepsilon^2)$ to test.
- An information theoretic argument saying every tester needs at least $\Omega(2^{d/2}n/\varepsilon^2)$.

*: includes uniform distribution. ~: hides polylogarithmic, like $d, \log(\mathrm{n})$; $d = O(\log n)$;

Thanks!

(and please come to our poster if interested)!