Testing the Equality in Distribution of Two Random Graphs

Maurilio Gutzeit, OvGU Magdeburg
Joint project with A. Carpentier,
U. von Luxburg and D. Ghoshdastidar.

14 February 2018, Potsdam

BASIC MODEL

Undirected graph $G = (V_n, E)$

- Set of vertices $V_n := \{1, 2, \ldots, n\}$, $n \in \mathbb{N}$, $n \geq 2$.
- Set of edges $E \subseteq \{(i, j) \in V_n^2 \mid i < j\}$.
- In particular, no loops and

$$|E| \leq d := \binom{n}{2}.$$

- Graph represented by a symmetric adjacency matrix

$$A \in \{0, 1\}^{n \times n} \text{ with } A_{ij} = \mathbb{1}_{\{(i,j) \in E\}}, \; i < j.$$

Basic Model, cont.

Inhomogeneous Erdős-Rényi graphs

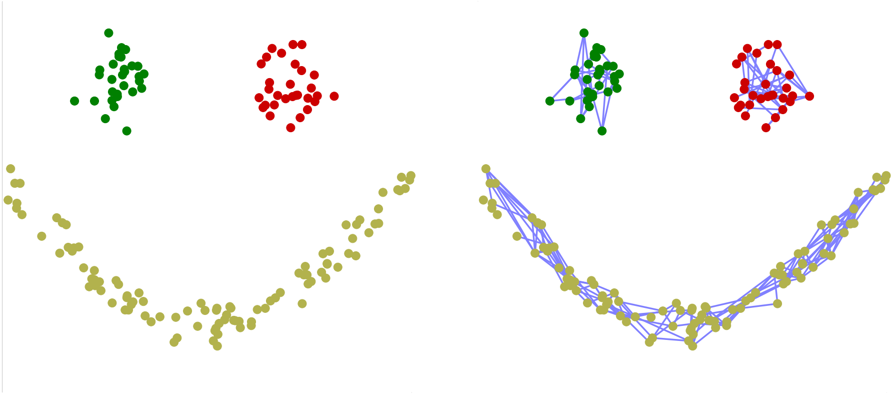
- E as a random variable: For each (i, j), independently
$$\mathbb{P}((i, j) \in E) = P_{ij} \in [0, 1].$$

- Whole model conveniently depicted in one matrix:
$$P \in [0, 1]^{n \times n}, \text{ symmetric, zero diagonal.}$$

- Let $\mathcal{G}_n$ be the set of all such matrices.

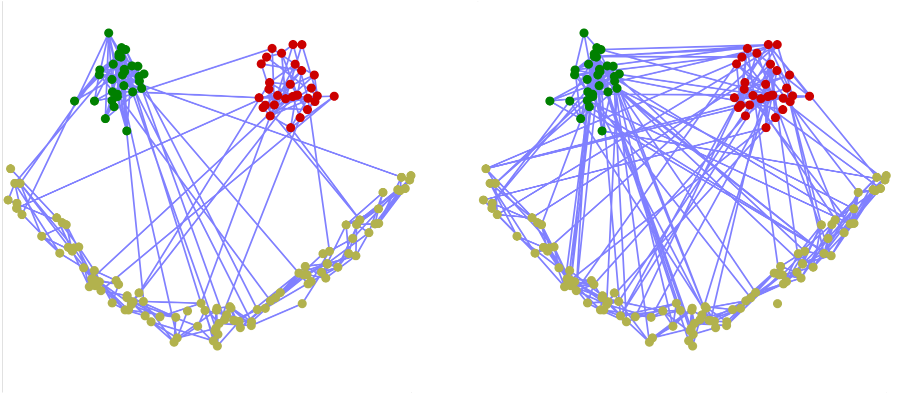
EXAMPLE

Stochastic Block Model (almost)



Example

Stochastic Block Model (almost)



STATISTICAL TESTING PROBLEM

Included quantities

- Known: n.
- Unknown: $P, Q \in \mathcal{G}_n$.
- Observations: $M \in \mathbb{N}$ sampled adjacency matrices from P and Q each, write

$$A_1, A_2, \ldots, A_M \overset{\text{iid}}{\sim} \text{IER}(P),$$
$$B_1, B_2, \ldots, B_M \overset{\text{iid}}{\sim} \text{IER}(Q).$$

Hypotheses

Given $\rho > 0$, consider

$$H_0: \; P = Q \; \text{ vs. } \; H_\rho: \; \|P - Q\| \geq \rho,$$

where $\|\cdot\| = \|\cdot\|_F$ (Frobenius norm) or $\|\cdot\| = \|\cdot\|_S$ (spectral norm).

QUANTITY OF INTEREST

Minimax separation rate

- The minimax type- I and type- II errors of a test φ:

$$\begin{aligned}\mathbb{P}_{H_0}(\varphi = 1) &= \sup_{P=Q} \mathbb{P}_{(P,Q)}(\varphi = 1), \\ \mathbb{P}_{H_\rho}(\varphi = 0) &= \sup_{\|P-Q\|\geq\rho} \mathbb{P}_{(P,Q)}(\varphi = 0).\end{aligned}$$

- Find the smallest ρ such that there is a test φ with

$$\mathbb{P}_{H_0}(\varphi = 1) + \mathbb{P}_{H_\rho}(\varphi = 0) \leq \eta \in (0,1)\,.$$

- Call this quantity ρ^*.
  Focus on n and M (thus "rate").

Applications and Literature

Testing equality in distribution could be useful for...
brain connectivity networks, molecular interaction networks (genomic data), social networks etc.

Previous results
Mostly asymptotic and with stronger model assumptions (RDPG, geometric graphs), see [GGCvL17a] for references.

Our work
Closest to this talk: [GGCvL17a]; broader perspective: [GGCvL17b].
Relation to classical signal detection, see e.g. [Bar02].

THE GENERAL RATE

THEOREM 1

In our testing problem with $\|\cdot\| = \|\cdot\|_F$, we have

$$\rho^* \sim \begin{cases} n, & \text{if } M = 1 \\ \sqrt{\dfrac{n}{M}}, & \text{if } M > 1. \end{cases}$$

Transition between $M = 1$ and $M > 1$

Note that for any (P, Q), we have

$$\|P - Q\|_F \leq \sqrt{n(n-1)} \sim n.$$

General Proof Tactics

Upper Bound

- Create a test φ with $\mathbb{P}_{H_0}(\varphi = 1) \leq \frac{\eta}{2}$.
- Tune ρ such that

$$\mathbb{P}_{H_\rho}(\varphi = 0) \leq \frac{\eta}{2}.$$

Lower Bound

- Want a difficult case, i.e. roughly

$$\mathbb{P}_{H_0} \approx \mathbb{P}_{H_\rho}, \quad \|P - Q\|_{\mathsf{F}} \gg 0.$$

- So, create appropriate priors ν_0 and ν_ρ for (P, Q) consistent with H_0 and H_ρ and control $\|\mathbb{P}_{(P,Q)\sim\nu_0} - \mathbb{P}_{(P,Q)\sim\nu_\rho}\|_{\mathsf{TV}}$.

Problem-Dependent Bounds on the Rate

Results of a different flavour through normalisation

We consider the alternative hypothesis

$$H'_\varrho : \; \frac{\|P - Q\|_F}{\sqrt{\|P + Q\|_F}} \geq \varrho.$$

Theorem 2

In our testing problem with $\|\cdot\| = \|\cdot\|_F$ and $M \geq 2$, we have

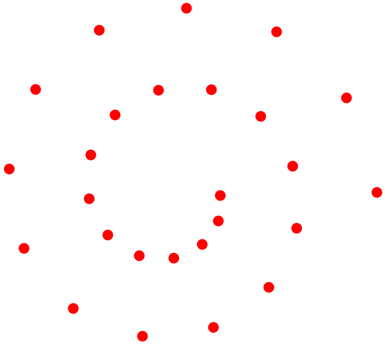
$$\sqrt{\frac{1}{M}} \lesssim \varrho^* \lesssim \sqrt{\frac{\ln(n)}{M}}$$

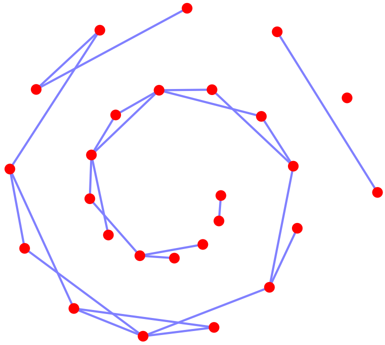
Comparison to previous theorem

E.g. $\|P - Q\|_F \gtrsim \sqrt{\frac{n}{M}}$ as opposed to $\|P - Q\|_F \gtrsim \sqrt{\frac{\|P + Q\|_F}{M}}$.

EXAMPLE

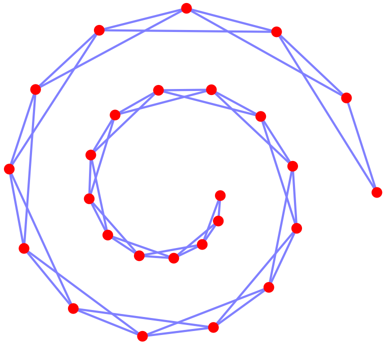
Only "4 nearest neighbors" allowed, otherwise $P_{ij} \equiv \frac{1}{2}$

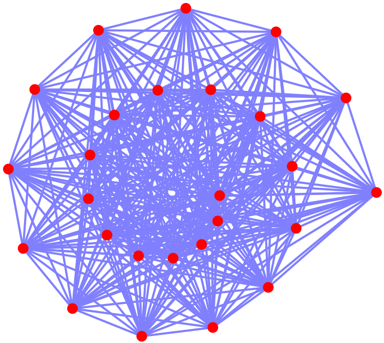




Example

Full graphs for "4 nearest neighbors" and "no restriction"





THE GENERAL RATE

THEOREM 3

In our testing problem with $\|\cdot\| = \|\cdot\|_S$, we have

$$\rho^* \sim \sqrt{\frac{n}{M}}.$$

No transition between $M = 1$ and $M > 1$

Note that

$$\|P - Q\|_S \leq \|P - Q\|_F \leq \sqrt{\mathrm{rk}(P - Q)} \cdot \|P - Q\|_S$$

and

$$\|P - Q\|_F \gtrsim n \Rightarrow \|P - Q\|_S \gtrsim \sqrt{n}.$$

Problem-Dependent Bounds on the Rate

With the row sum norm $\|\cdot\|_{\mathrm{r}}$, we consider the alternative hypothesis

$$H'_\varrho : \ \frac{\|P-Q\|_{\mathrm{S}}}{\sqrt{\|P+Q\|_{\mathrm{r}}}} \geq \varrho.$$

Theorem 4

In our testing problem with H'_ϱ, we have

$$\sqrt{\frac{1}{M}} \lesssim \varrho^* \lesssim \sqrt{\frac{\ln(n)}{M}}.$$

Plan

- Revise the paper.
- Try to improve the problem-dependent bounds, i.e. understand/get rid of $\ln -$factors.

Literature

[Bar02] Yannick Baraud. Non-asymptotic minimax rates of testing in signal detection. Bernoulli, 8(5):577--606, 2002.

[GGCvL17a] Debarghya Ghoshdastidar, Maurilio Gutzeit, Alexandra Carpentier, and Ulrike von Luxburg. Two-sample hypothesis testing for inhomogeneous random graphs. arXiv preprint, 2017.

[GGCvL17b] Debarghya Ghoshdastidar, Maurilio Gutzeit, Alexandra Carpentier, and Ulrike von Luxburg. Two-sample tests for large random graphs using network statistics. COLT 2017, 2017.

Thank you for your attention!

&

Happy Valentine's Day!