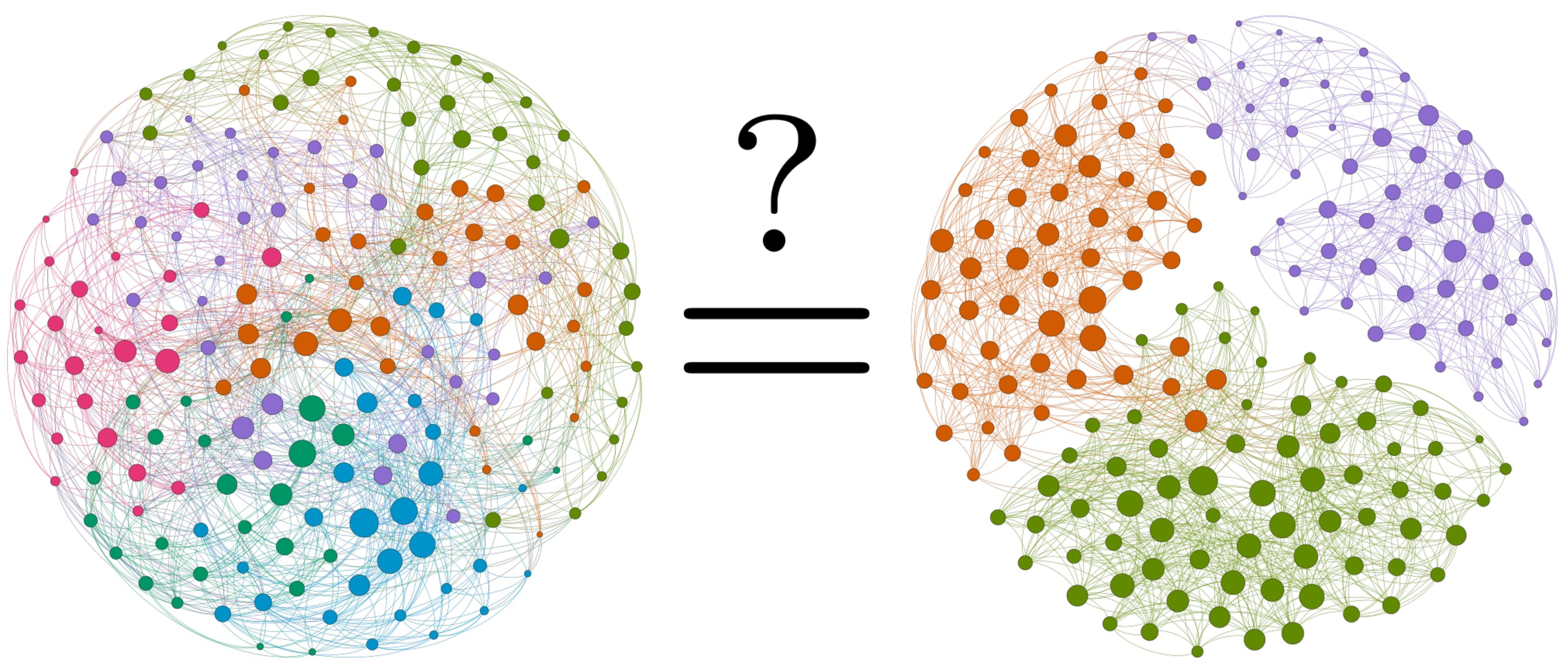
Practical Methods for Graph Two-Sample Testing

Debarghya Ghoshdastidar, Ulrike von Luxburg

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Max Planck Institute for Intelligent Systems Tübingen Campus

Graph two-sample testing



How to decide whether two (families of) graphs come from the same underlying population?

- Standard tests need large sample size
- Existing graph tests need bootstrap samples — difficult for small population

Problem: Fix m, and let V be a common set of n vertices

Given graphs $G_1, \ldots, G_m \sim_{\text{iid}} \mathcal{P}_n$ and $H_1, \ldots, H_m \sim_{\text{iid}} \mathcal{Q}_n$ defined on V

Test: $\mathcal{H}_0 : \mathcal{P}_n = \mathcal{Q}_n$ or $\mathcal{H}_1 : \|\mathcal{P}_n - \mathcal{Q}_n\| > \delta_n$

New tests for IER graphs based on asymptotic distributions

IER (Inhomogeneous Erdős-Rényi graph) **:** Edges independent, but have arbitrary probabilities

$\mathcal{P}_n = \text{IER}(P_n)$ and $\mathcal{Q}_n = \text{IER}(Q_n)$ parameterized by $n \times n$ matrices

Asymptotic normal test

- Applicable for sample size $m > 1$
- Test based on entry-wise difference in adjacency matrices

$$T_n = \frac{\sum\limits_{i<j} \left(\sum\limits_{k \le m/2} (A_{G_k})_{ij} - (A_{H_k})_{ij} \right) \left(\sum\limits_{k > m/2} (A_{G_k})_{ij} - (A_{H_k})_{ij} \right)}{\sqrt{\sum\limits_{i<j} \left(\sum\limits_{k \le m/2} (A_{G_k})_{ij} + (A_{H_k})_{ij} \right) \left(\sum\limits_{k > m/2} (A_{G_k})_{ij} + (A_{H_k})_{ij} \right)}}$$

Result:

$\mathcal{H}_0 : \lim\limits_{n \to \infty} T_n$ dominated by $\mathcal{N}(0,1)$

$\mathcal{H}_1 : T_n \to \infty$ if $\delta_n \gg \sqrt{\frac{1}{m}} (\|P_n\|_F \vee \|Q_n\|_F)$

Asymptotic Tracy-Widom test

- Applicable for unit sample size ($m = 1$)
- Test captures difference in graph spectrum

$$T_n = n^{2/3} \left(\left\| (A_{G_1} - A_{H_1}) \circ S \right\|_2 - 2 \right)$$

where S does entry-wise re-scaling

Result:

$\mathcal{H}_0 : \lim\limits_{n \to \infty} T_n$ nearly follows TW_1 law

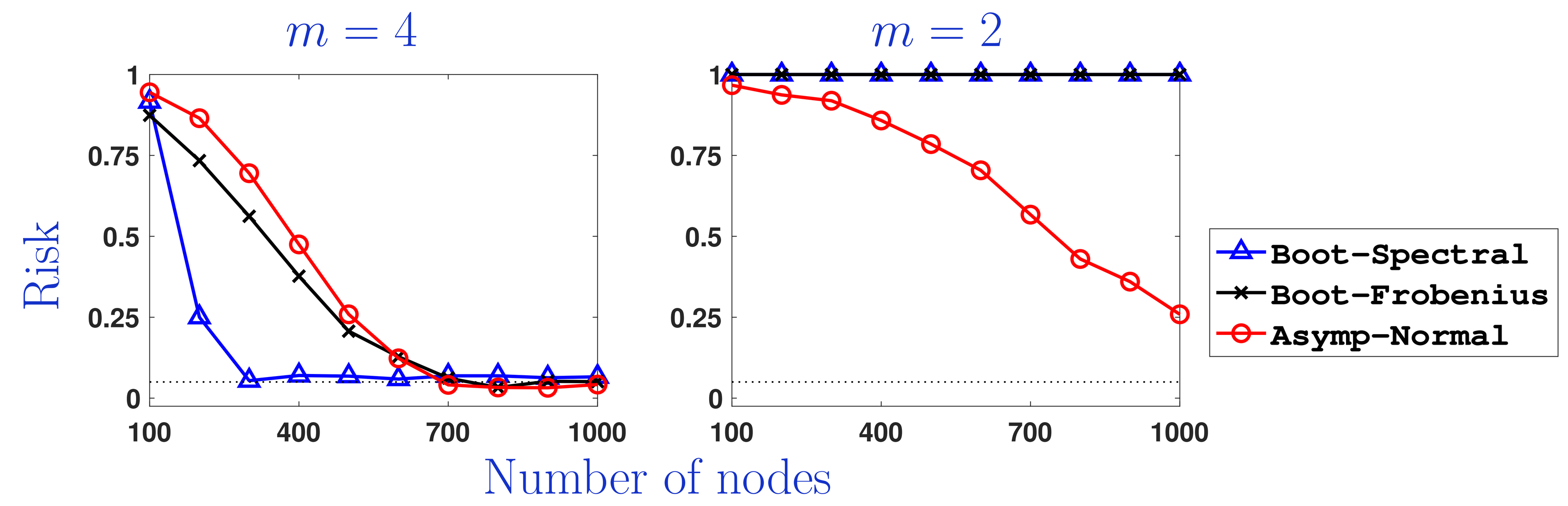
$\mathcal{H}_1 : T_n \to \infty$ if $\delta_n \gg k\sqrt{n\rho}$ for k-SBM

where $\rho = \|P_n\|_{\max} \vee \|Q_n\|_{\max}$

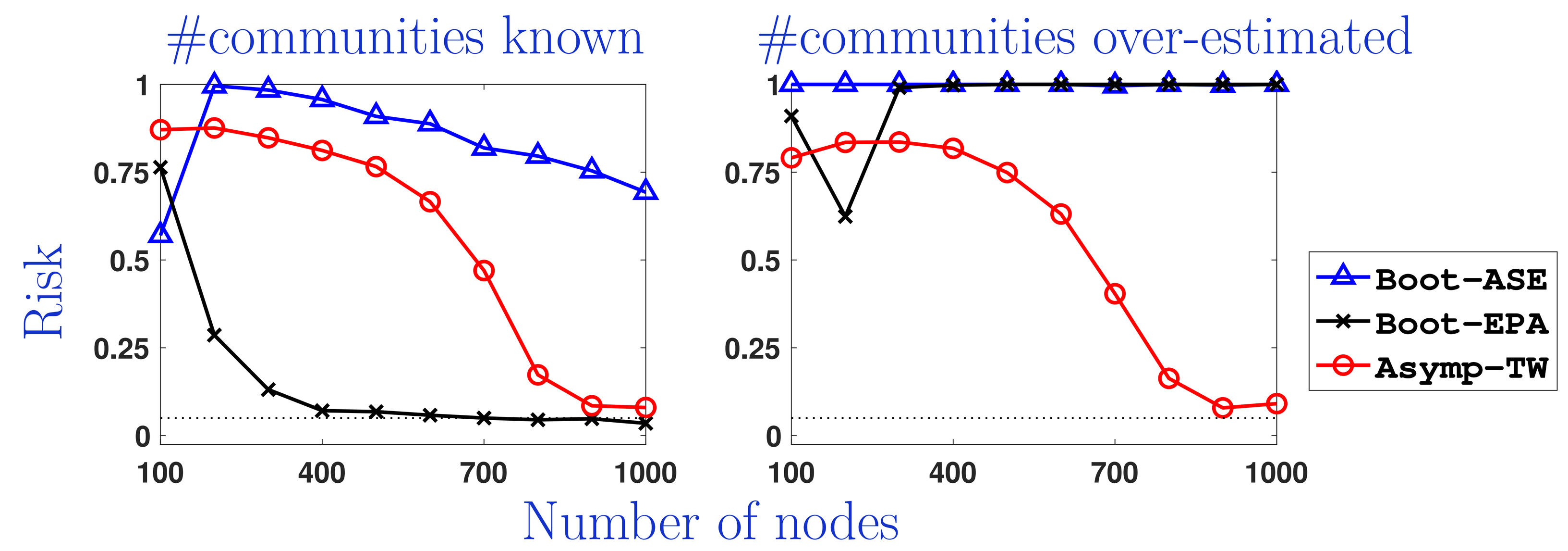
For graphs on a common vertex set, our asymptotic tests are fast and easy to use

Testing random graphs

- Graphs from stochastic block model — 2 communities; different parameters under $\mathcal{H}_0$ and $\mathcal{H}_1$
- **Sample size $m > 1$:** Our test works even for sample size 2, but existing (bootstrap) tests need more samples

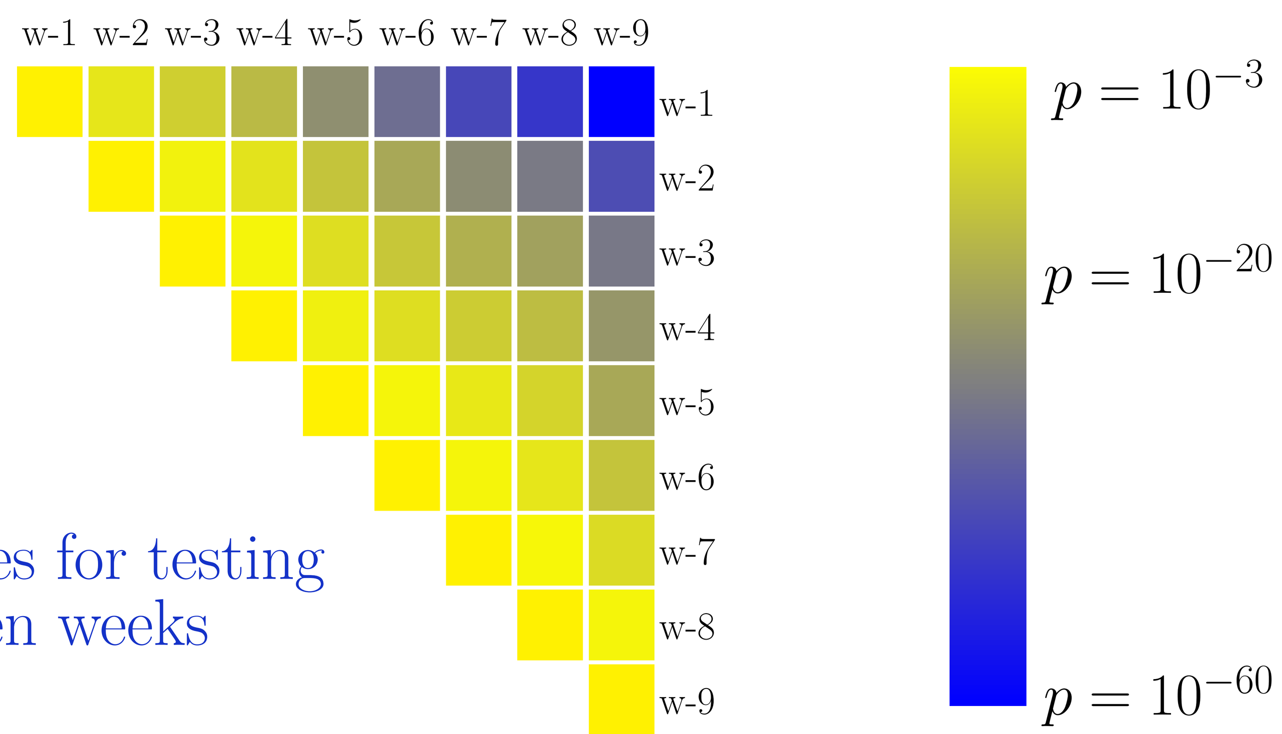


- **Sample size $m = 1$:** Bootstrap tests fail if number of communities not known correctly, but our test works



Testing networks in Oregon data set

- Peering networks of 11806 routers over 9 weeks
- Networks change considerably over the weeks



p-values for testing between weeks

References: • Gretton et al. *JMLR*, 2012 • Ginestet et al. *AOAP*, 2017 • Tang et al. *JCGS*, 2016; *Bernoulli*, 2017 • Ghoshdastidar et al. *COLT*, 2017; *arXiv:1707.00833*

Codes on Github: `gdebarghya/Network-TwoSampleTesting`