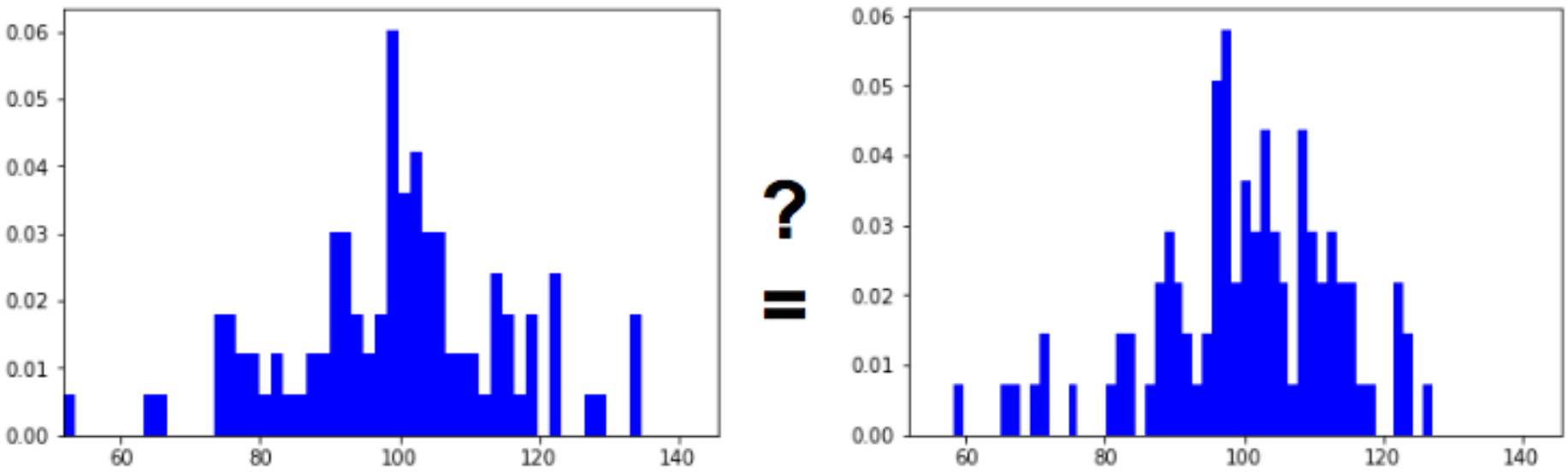
A central limit theorem for an omnibus embedding of random dot product graphs

Keith Levin[1]
with Avanti Athreya[2], Minh Tang[2], Vince Lyzinski[3] and Carey E. Priebe[2]

[1]University of Michigan, [2]Johns Hopkins University, [3]University of Massachusetts Amherst

November 18, 2017

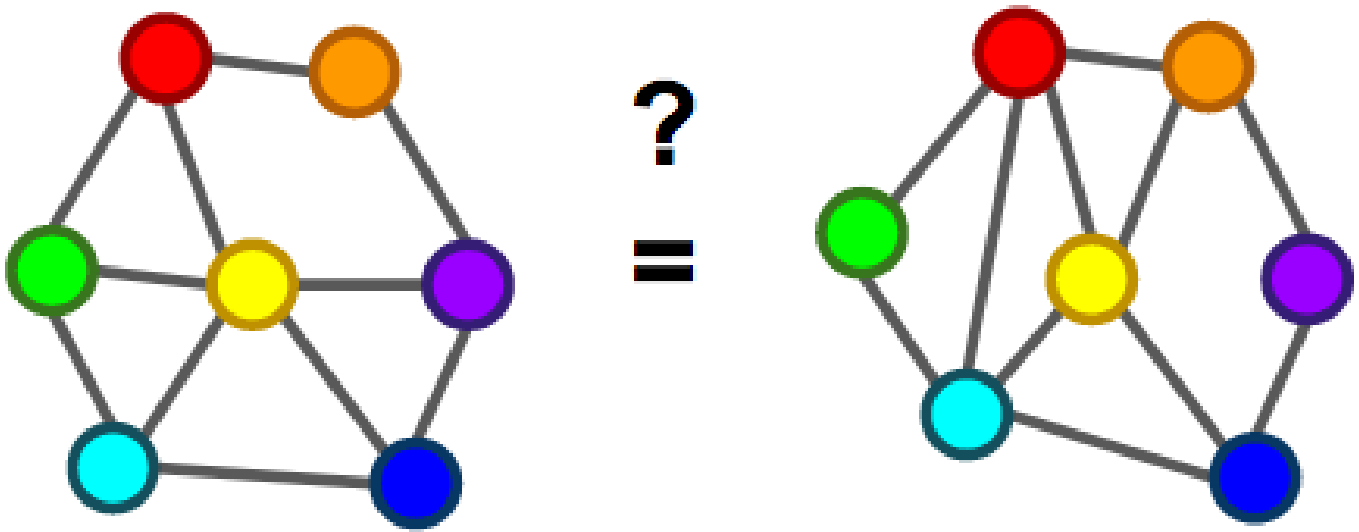
Classical two-sample hypothesis testing



Well-studied in statistics (indeed, the only thing we teach undergrads?)

Graph Hypothesis Testing

Q: how to tell if two (or more) graphs are from the same distribution?



Random Dot Product Graph (RDPG; Young and Scheinerman, 2007)

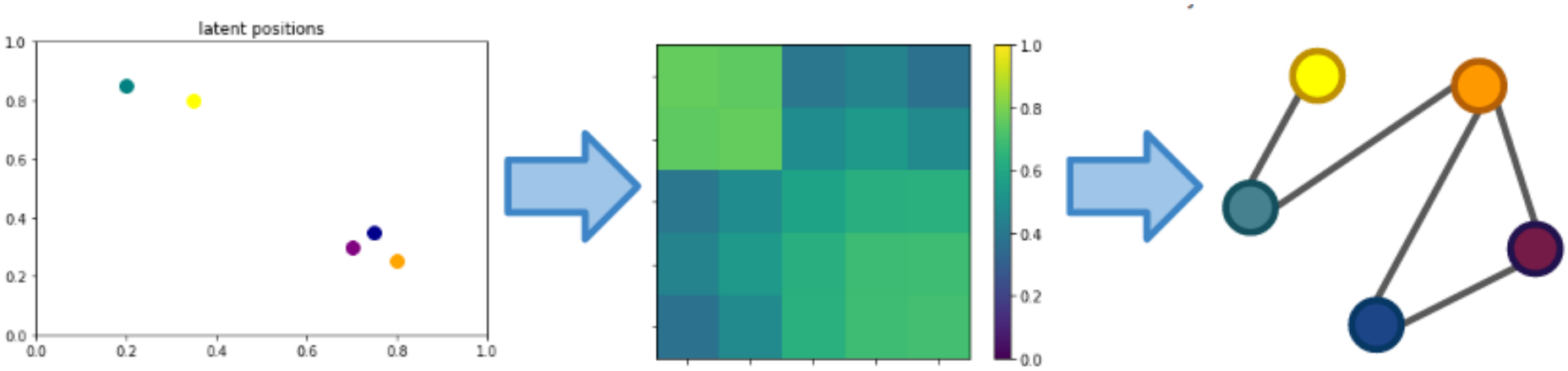
- Extends stochastic block model (SBM)
- Vertices assigned *latent positions*
  - drawn i.i.d. from d-dimensional distribution F
  - F constrained so that $0 \leq x^T y \leq 1$ whenever $x, y \in \text{supp}\, F$
  - Denote i-th latent position by $X_i \in R^d$
- Edges $\{i, j\}$ present or absent independently with probability $X_i^T X_j$.
- Collect latent positions in rows of $X \in \mathbb{R}^{n \times d}$.

Warning: Non-identifiability

Model specified only up to orthogonal rotation of latent positions.

Random Dot Product Graph (RDPG; Young and Scheinerman, 2007)

- Extends stochastic block model (SBM)
- Vertices assigned *latent positions*
  - drawn i.i.d. from d-dimensional distribution F
  - F constrained so that $0 \leq x^T y \leq 1$ whenever $x, y \in \text{supp}\, F$.
  - Denote i-th latent position by X_i
- Edges $\{i, j\}$ present or absent independently with probability $X_i^T X_j$.



Estimating latent positions: adjacency spectral embedding (Sussman et al, 2012)

Definition (Adjacency Spectral Embedding (ASE))

Given adjacency matrix A, embed vertices of $A = USU^T$ into $\mathbb{R}^d$ as rows of $\hat{X} = U_d S_d^{1/2} \in \mathbb{R}^{n \times d}$, where U_d denotes first d columns of U, S_d denotes truncation of S to top d eigenvalues.

- Under RDPG, $\exists W : \max_{1 \leq i \leq n} \|\hat{X}_i - WX_i\| = O_{\mathbb{P}}(n^{-1/2} \log n)$.
- Lyzinski, et al (2014): ASE yields a.a.s. perfect recovery of block memberships in SBM

RDPG: what do we mean by *same distribution*?

RDPG: what do we mean by *same distribution*?

Option 1: Test if latent positions are drawn from same distribution.

- G_1 positions drawn i.i.d. F_1, G_2 positions drawn i.i.d. F_2
- Test if $F_1 = F_2$
- "Nonparametric" testing

Tang, Athreya, Sussman, Lyzinski and Priebe (2017)

Estimate latent positions of G_1 and G_2 via ASE, apply maximum mean discrepancy (Gretton et al, 2012) to ASE estimates.

RDPG: what do we mean by *same distribution*?

Option 2: Test if latent positions are the same

- G_1 latent positions $X \in \mathbb{R}^{n \times d}$, G_2 latent positions $Y \in \mathbb{R}^{n \times d}$
- Test if $X = YW$ for some unitary W.
- "Semiparametric" testing

Tang, Athreya, Sussman, Lyzinski and Priebe (2015)

Embed both graphs via ASE, align estimated positions via Procrustes analysis (Gower, 1975). Reject H_0 if alignment is poor, i.e., if $T_{\text{Proc}} = \min_{W \in \mathcal{U}_d} \|\hat{X} - \hat{Y}W\|_F$ is large.

Challenges in semiparametric graph testing

Problem 1: Procrustes alignment introduces variance

- More variance $\Rightarrow$ less power.

Problem 2: How to generalize to multiple-graph hypothesis testing?

- Ultimately, we want something like ANOVA for graphs.

Goal: develop a technique that...

1. Avoids Procrustes alignment
2. Generalizes naturally to 3 or more graphs

Omnibus matrix: motivation

Definition (Omnibus matrix)

Let graphs G_1 and G_2 be d-dimensional RDPGs with adjacency matrices $A^{(1)}$ and $A^{(2)}$. We construct an *omnibus matrix* for the graphs as

$$M = \begin{bmatrix} A^{(1)} & \frac{A^{(1)}+A^{(2)}}{2} \\ \frac{A^{(1)}+A^{(2)}}{2} & A^{(2)} \end{bmatrix} \in \mathbb{R}^{2n \times 2n}$$

Note: generalizes naturally to m graphs, with (i,j)-block $(A^{(i)} + A^{(j)})/2$.

Omnibus embedding

Reminder

$$M = \begin{bmatrix} A^{(1)} & \frac{A^{(1)}+A^{(2)}}{2} \\ \frac{A^{(1)}+A^{(2)}}{2} & A^{(2)} \end{bmatrix} \in \mathbb{R}^{2n \times 2n}$$

Under H_0, we have $\mathbb{E}A^{(1)} = \mathbb{E}A^{(2)} = XX^T = P = U_P S_P U_P^T$

- $S_P \in \mathbb{R}^{d \times d}$ diagonal, $U_P \in \mathbb{R}^{n \times d}$ orthonormal columns
- $\mathbb{E}M = \tilde{P} = \begin{bmatrix} P & P \\ P & P \end{bmatrix} = \begin{bmatrix} U \\ U \end{bmatrix} S_P \begin{bmatrix} U^T U^T \end{bmatrix} = \begin{bmatrix} X \\ X \end{bmatrix} \begin{bmatrix} X^T X^T \end{bmatrix} = U_{\tilde{P}} S_{\tilde{P}} U_{\tilde{P}}^T.$

Omnibus embedding

Under H_0, we have $\mathbb{E}A^{(1)} = \mathbb{E}A^{(2)} = XX^T = P = U_P S_P U_P^T$

- $S_P \in \mathbb{R}^{d\times d}$ diagonal, $U_P \in \mathbb{R}^{n\times d}$ orthonormal columns

$$\mathbb{E}M = \tilde{P} = \begin{bmatrix} P & P \\ P & P \end{bmatrix} = \begin{bmatrix} U \\ U \end{bmatrix} S_P \begin{bmatrix} U^T U^T \end{bmatrix} = \begin{bmatrix} X \\ X \end{bmatrix} \begin{bmatrix} X^T X^T \end{bmatrix} = U_{\tilde{P}} S_{\tilde{P}} U_{\tilde{P}}^T.$$

Key point

Applying ASE to M, we get a $2n$-by-d matrix,

$$\hat{Z} = \begin{bmatrix} \hat{X} \\ \hat{Y} \end{bmatrix},$$

$\hat{X}, \hat{Y} \in \mathbb{R}^{n\times d}$ provide estimates of latent positions of G_1, G_2, in the *same* d-dimensional space *without* additional alignment step. Natural test statistic given by $T_{\text{Omni}} = \|\hat{X} - \hat{Y}\|_F$.

Main results: Notational preliminaries

In what follows, we assume the null hypothesis

So G_1 and G_2 have shared latent positions $X \in \mathbb{R}^{n \times d}$.

- $\mathbb{E}A^{(1)} = \mathbb{E}A^{(2)} = P = U_P S_P U_P^T = XX^T \in \mathbb{R}^{n \times n}$
- We denote the "true latent positions" of M by

$$Z = \begin{bmatrix} X \\ X \end{bmatrix} = \begin{bmatrix} U_P \\ U_P \end{bmatrix} S_P^{1/2} = U_{\tilde{P}} S_{\tilde{P}}^{1/2} \in \mathbb{R}^{2n \times d}$$

and their estimates by

$$\hat{Z} = U_M S_M^{1/2} = \begin{bmatrix} \hat{X} \\ \hat{Y} \end{bmatrix} \in \mathbb{R}^{2n \times d}$$

where $S_M \in \mathbb{R}^{d \times d}$ is the diagonal matrix of the top d eigenvalues of M and corresponding eigenvectors in columns of $U_M \in \mathbb{R}^{2n \times d}$.

Main results: Concentration inequality

Lemma (Uniform concentration of estimates)

Let $\{A^{(i)}\}_{i=1}^{m}$ be adjacency matrices of m independent RDPGs with shared latent positions $X = U_P S_P^{1/2} \in \mathbb{R}^{n \times d}$ and let $M \in \mathbb{R}^{mn \times mn}$ be their omnibus matrix with top eigenvalues collected in diagonal matrix $S_M \in \mathbb{R}^{d \times d}$ and corresponding eigenvalues in the columns of $U_M \in \mathbb{R}^{mn \times d}$. There exists a constant $C > 0$ such that with high probability, there exists an orthogonal matrix $W \in \mathbb{R}^{d \times d}$ such that

$$\max_{1 \leq h \leq mn} \|(U_M S_M^{1/2} - U_{\tilde{P}} S_{\tilde{P}}^{1/2} W)_{h,\cdot}\| \leq \frac{C m^{1/2} \log mn}{\sqrt{n}}.$$

Main results: CLT

Theorem (CLT: informally)

Let $\{A^{(i)}\}_{i=1}^{m}$ be adjacency matrices of m independent RDPGs with shared latent positions $X = U_P S_P^{1/2} \in \mathbb{R}^{n \times d}$ drawn i.i.d. from d-dimensional distribution F. Let $M \in \mathbb{R}^{mn \times mn}$ be their omnibus matrix with top eigenvalues collected in diagonal matrix $S_M \in \mathbb{R}^{d \times d}$ and corresponding eigenvalues in the columns of $U_M \in \mathbb{R}^{mn \times d}$. Fix $h = m(s-1) + i$ for $i \in [n]$ and $s \in [m]$. Then the error between the h-th position estimate and the (properly rotated) true h-th position is asymptotically a continuous mixture of normals, with mixing determined by F.

$$n^{1/2}(U_M S_M^{1/2} - U_{\tilde{P}} S_{\tilde{P}}^{1/2} W_n)_{h,\cdot} \to \int \mathcal{N}(0, \Sigma(y)) dF(y).$$

Main results: CLT

Theorem (CLT: More formally)

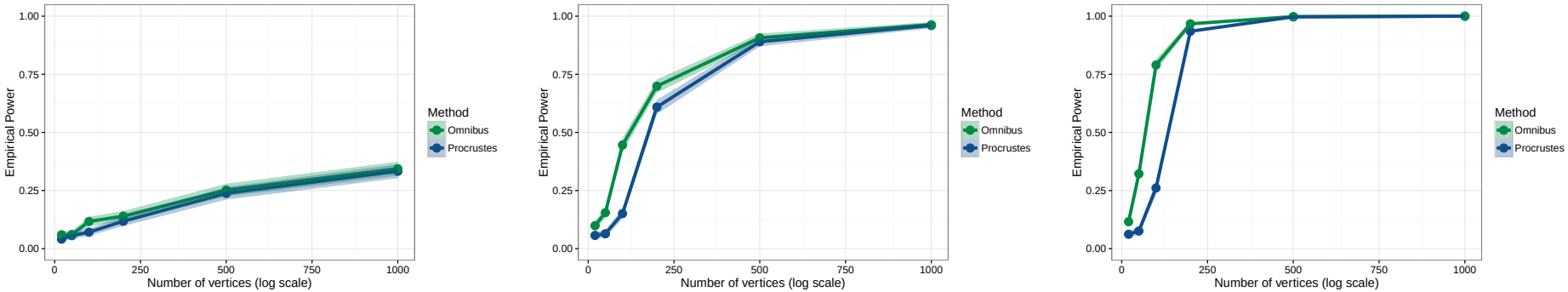
Let $\{A^{(i)}\}_{i=1}^{m}$ *be adjacency matrices of* m *independent RDPGs with shared latent positions* $X = U_P S_P^{1/2} \in \mathbb{R}^{n \times d}$ *drawn i.i.d. from* d*-dimensional distribution* F*. Let* $M \in \mathbb{R}^{mn \times mn}$ *be their omnibus matrix with top eigenvalues collected in diagonal matrix* $S_M \in \mathbb{R}^{d \times d}$ *and corresponding eigenvalues in the columns of* $U_M \in \mathbb{R}^{mn \times d}$*. Let* $\Phi(x, \Sigma)$ *denote the cdf of a multivariate Gaussian with mean* 0 *and covariance matrix* Σ*. Fix* $h = m(s-1) + i$ *for* $i \in [n]$ *and* $s \in [m]$*. There exists a sequence of* d*-by-*d *orthogonal matrices* $(W_n)_{n=1}^{\infty}$ *such that for all* $x \in \mathbb{R}^d$*,*

$$\lim_{n \to \infty} \Pr\left[n^{1/2} (U_M S_M^{1/2} - U_{\tilde{P}} S_{\tilde{P}}^{1/2} W_n)_{h,\cdot} \leq x \right] = \int \Phi\left(x, \Sigma(y)\right) dF(y),$$

where $\Sigma(y) = (m+3)\Delta^{-1} \tilde{\Sigma}(y) \Delta^{-1} / (4m)$ *and*

$$\Delta = \mathbb{E}_F X_1 X_1^T, \quad \tilde{\Sigma}(y) = \mathbb{E}_F (y^T X_1 - (y^T X_1)^2) X_1 X_1^T.$$

Experiments: hypothesis testing



(a) (b) (c)

Figure: Power of the Procrustes-based (blue) and omnibus-based (green) tests to detect when the two graphs being testing differ in (a) one, (b) five, and (c) ten of their latent positions. Each point is the proportion of 1000 trials for which the given technique correctly rejected the null hypothesis. Error bars denote two standard errors of this empirical mean.

Experiments: estimating latent positions

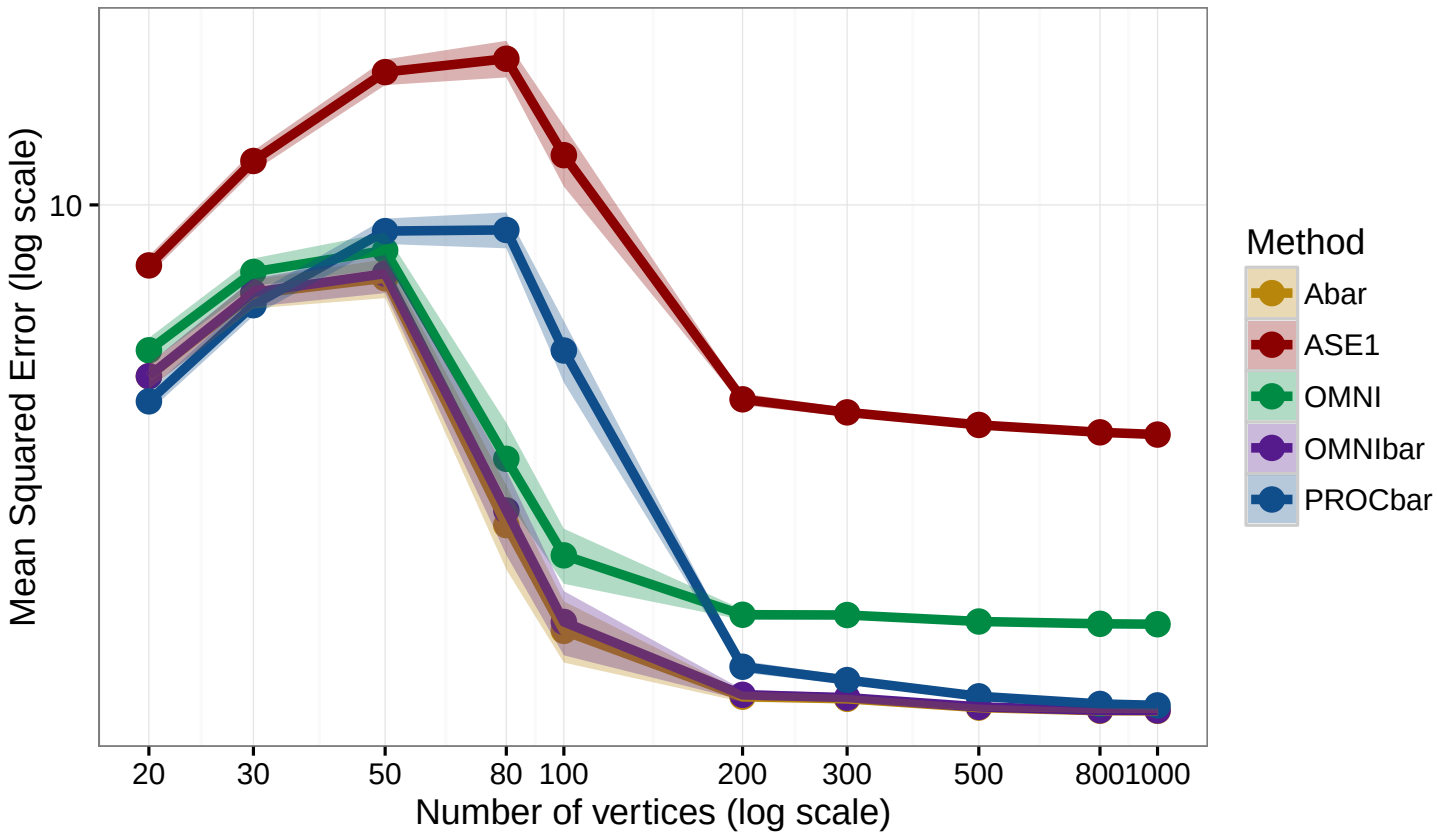


Figure: Mean squared error (MSE) in recovery of latent positions (up to rotation) in a 2-graph RDPG model as a function of the number of vertices for different estimation procedures.

Future Work

- Develop graph analogues of ANOVA and other multiple hypothesis testing procedures
- Improve techniques for choosing critical value in omnibus test
- Improve understanding of power under H_A

Thanks!

Full paper: `https://arxiv.org/abs/1705.09355`