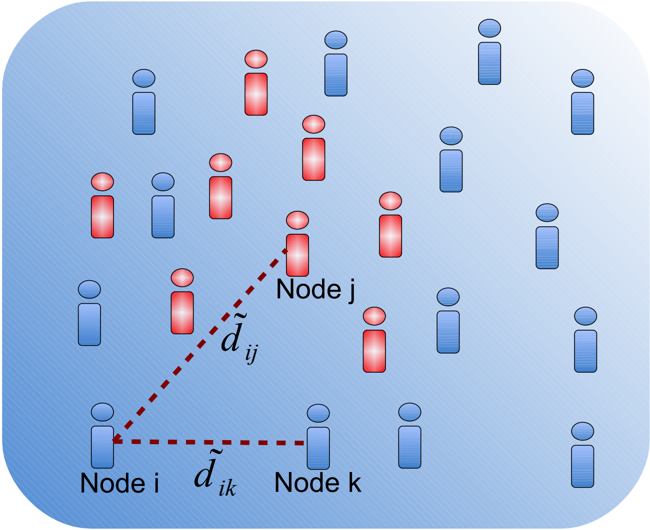
Identifying latent space geometry of network formation models via analysis of curvature

Shane Lubold, Arun Chandrasekhar, Tyler McCormick

May 17th, 2021

Latent Space Model



Talk Overview

- Different geometries lead to networks with different properties [Asta and Shalizi, 2015].

Talk Overview

- Different geometries lead to networks with different properties [Asta and Shalizi, 2015].
- Estimating the geometry of a network is critical!

Talk Overview

- Different geometries lead to networks with different properties [Asta and Shalizi, 2015].
- Estimating the geometry of a network is critical!
- **Our contribution: propose the first data-driven, HT framework for estimating geometry!**

Generative LS Model

Model Inputs

1. Latent space $\mathcal{M}^p(\kappa)$ with dimension p and curvature κ.

Generative LS Model

Model Inputs

1. Latent space $\mathcal{M}^p(\kappa)$ with dimension p and curvature κ.
2. Node-specific effects $\nu_1, \ldots, \nu_n \in (-\infty, 0]$.

Generative LS Model

Model Inputs

1. Latent space $\mathcal{M}^p(\kappa)$ with dimension p and curvature κ.
2. Node-specific effects $\nu_1, \ldots, \nu_n \in (-\infty, 0]$.
3. Node locations $z_1, \ldots, z_n$ on $\mathcal{M}^p(\kappa)$.

Generative LS Model

Model Inputs

1. Latent space $\mathcal{M}^p(\kappa)$ with dimension p and curvature κ.
2. Node-specific effects $\nu_1, \ldots, \nu_n \in (-\infty, 0]$.
3. Node locations $z_1, \ldots, z_n$ on $\mathcal{M}^p(\kappa)$.

We consider un-directed graph G drawn according to

$$P(G_{ij} = 1 | z, \nu) = \exp\left(\nu_i + \nu_j - d_{\mathcal{M}}(z_i, z_j)\right) \tag{1}$$

Generative LS Model

Model Inputs

1. Latent space $\mathcal{M}^p(\kappa)$ with dimension p and curvature κ.
2. Node-specific effects $\nu_1, \ldots, \nu_n \in (-\infty, 0]$.
3. Node locations $z_1, \ldots, z_n$ on $\mathcal{M}^p(\kappa)$.

We consider un-directed graph G drawn according to

$$\mathrm{P}(G_{ij} = 1 | z, \nu) = \exp\left(\nu_i + \nu_j - d_{\mathcal{M}}(z_i, z_j)\right) \tag{1}$$

Question: Given G drawn from (1), what geometry generated G?

Generative LS Model

Model Inputs

1. Latent space $\mathcal{M}^p(\kappa)$ with dimension p and curvature κ.
2. Node-specific effects $\nu_1, \ldots, \nu_n \in (-\infty, 0]$.
3. Node locations $z_1, \ldots, z_n$ on $\mathcal{M}^p(\kappa)$.

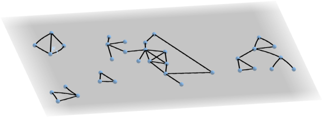
We consider un-directed graph G drawn according to

$$\mathrm{P}(G_{ij} = 1 | z, \nu) = \exp\left(\nu_i + \nu_j - d_{\mathcal{M}}(z_i, z_j)\right) \tag{1}$$

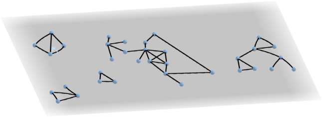
Question: Given G drawn from (1), what geometry generated G?

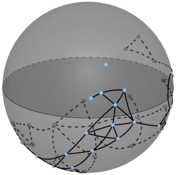
Contribution: Provide a HT framework to answer this!

Candidate Geometries

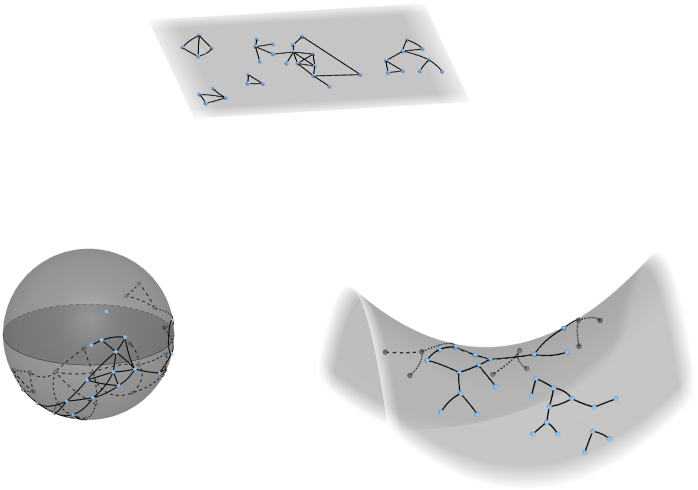


Candidate Geometries





Candidate Geometries



Literature Review

Literature Review

- Original paper:
  [Hoff et al., 2002, Handcock et al., 2007, Hoff, 2005]

Literature Review

- Original paper:
  [Hoff et al., 2002, Handcock et al., 2007, Hoff, 2005]
- Selecting dimension: [Oh and Raftery, 2001]

Literature Review

- Original paper:
  [Hoff et al., 2002, Handcock et al., 2007, Hoff, 2005]
- Selecting dimension: [Oh and Raftery, 2001]
- Network properties $\rightarrow$ geometry type. [Smith et al., 2019]

Literature Review

- Original paper:
  [Hoff et al., 2002, Handcock et al., 2007, Hoff, 2005]
- Selecting dimension: [Oh and Raftery, 2001]
- Network properties $\rightarrow$ geometry type. [Smith et al., 2019]
- Consistency of parameters: [Shalizi and Asta, 2017]

Literature Review

- Original paper:
  [Hoff et al., 2002, Handcock et al., 2007, Hoff, 2005]
- Selecting dimension: [Oh and Raftery, 2001]
- Network properties $\rightarrow$ geometry type. [Smith et al., 2019]
- Consistency of parameters: [Shalizi and Asta, 2017]

- Applications
  - CEO behavior [Friel et al., 2016]
  - Triadic data [Sosa and Rodriguez, 2018],
  - Anomaly detection [Lee et al., 2020]
  - Many many others!

- Extensions

Literature Review

- Original paper: [Hoff et al., 2002, Handcock et al., 2007, Hoff, 2005]
- Selecting dimension: [Oh and Raftery, 2001]
- Network properties $\rightarrow$ geometry type. [Smith et al., 2019]
- Consistency of parameters: [Shalizi and Asta, 2017]

- Applications
  - CEO behavior [Friel et al., 2016]
  - Triadic data [Sosa and Rodriguez, 2018],
  - Anomaly detection [Lee et al., 2020]
  - Many many others!

- Extensions
  - time-series [Salter-Townshend and McCormick, 2017]

Literature Review

- Original paper:
  [Hoff et al., 2002, Handcock et al., 2007, Hoff, 2005]
- Selecting dimension: [Oh and Raftery, 2001]
- Network properties $\rightarrow$ geometry type. [Smith et al., 2019]
- Consistency of parameters: [Shalizi and Asta, 2017]

- Applications
  - CEO behavior [Friel et al., 2016]
  - Triadic data [Sosa and Rodriguez, 2018],
  - Anomaly detection [Lee et al., 2020]
  - Many many others!

- Extensions
  - time-series [Salter-Townshend and McCormick, 2017]
  - various geometry types [Asta and Shalizi, 2015]

Literature Review

- Original paper: [Hoff et al., 2002, Handcock et al., 2007, Hoff, 2005]
- Selecting dimension: [Oh and Raftery, 2001]
- Network properties $\rightarrow$ geometry type. [Smith et al., 2019]
- Consistency of parameters: [Shalizi and Asta, 2017]

- Applications
  - CEO behavior [Friel et al., 2016]
  - Triadic data [Sosa and Rodriguez, 2018],
  - Anomaly detection [Lee et al., 2020]
  - Many many others!

- Extensions
  - time-series [Salter-Townshend and McCormick, 2017]
  - various geometry types [Asta and Shalizi, 2015]
  - Partial network data [McCormick and Zheng, 2015]

Hypothesis Testing

- Given $G \sim LS(\mathcal{M})$, we want to test

Hypothesis Testing

- Given $G \sim LS(\mathcal{M})$, we want to test

$$(\textit{Euclidean}) \quad H_0 : \mathcal{M} = \mathbb{R}^p, \quad H_a : \mathcal{M} \neq \mathbb{R}^p$$

Hypothesis Testing

- Given $G \sim LS(\mathcal{M})$, we want to test

$$(Euclidean) \quad H_0 : \mathcal{M} = \mathbb{R}^p, \quad H_a : \mathcal{M} \neq \mathbb{R}^p$$

$$(Spherical) \quad H_0 : \mathcal{M} = \mathbb{S}^p, \quad H_a : \mathcal{M} \neq \mathbb{S}^p$$

Hypothesis Testing

- Given $G \sim LS(\mathcal{M})$, we want to test

$$(Euclidean) \quad H_0 : \mathcal{M} = \mathbb{R}^p, \quad H_a : \mathcal{M} \neq \mathbb{R}^p$$

$$(Spherical) \quad H_0 : \mathcal{M} = \mathbb{S}^p, \quad H_a : \mathcal{M} \neq \mathbb{S}^p$$

$$(Hyperbolic) \quad H_0 : \mathcal{M} = \mathbb{H}^p, \quad H_a : \mathcal{M} \neq \mathbb{H}^p$$

Hypothesis Testing

- Given $G \sim LS(\mathcal{M})$, we want to test

$$(\textit{Euclidean}) \quad H_0 : \mathcal{M} = \mathbb{R}^p, \quad H_a : \mathcal{M} \neq \mathbb{R}^p$$

$$(\textit{Spherical}) \quad H_0 : \mathcal{M} = \mathbb{S}^p, \quad H_a : \mathcal{M} \neq \mathbb{S}^p$$

$$(\textit{Hyperbolic}) \quad H_0 : \mathcal{M} = \mathbb{H}^p, \quad H_a : \mathcal{M} \neq \mathbb{H}^p$$

Assume that $\kappa \in [-b, -a] \cup \{0\} \cup [a, b]$**, with** $b > a > 0$.

Convert hypotheses into easily tested conditions!

HT Framework

To create hypothesis testing framework, we have three steps:

HT Framework

To create hypothesis testing framework, we have three steps:

1. Show that the eigenvalues of distance matrices for points on $\mathcal{M}$ tell us the geometry. So the eigenvalues become our test statistics.

HT Framework

To create hypothesis testing framework, we have three steps:

1. Show that the eigenvalues of distance matrices for points on $\mathcal{M}$ tell us the geometry. So the eigenvalues become our test statistics.
2. Estimate distances between LS positions using graph.

HT Framework

To create hypothesis testing framework, we have three steps:

1. Show that the eigenvalues of distance matrices for points on $\mathcal{M}$ tell us the geometry. So the eigenvalues become our test statistics.
2. Estimate distances between LS positions using graph.
3. Use sampling uncertainty to construct distribution of test statistic.

HT Framework

To create hypothesis testing framework, we have three steps:

1. Show that the eigenvalues of distance matrices for points on $\mathcal{M}$ tell us the geometry. So the eigenvalues become our test statistics.
2. Estimate distances between LS positions using graph.
3. Use sampling uncertainty to construct distribution of test statistic.

Embedding Conditions

Theorem ([Schoenberg, 1935])

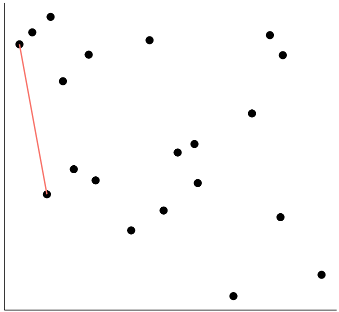
Let D contains distances between $z_1, \ldots, z_K$. Then, $\{z_1, \ldots, z_K\} \overset{isom}{\to} \mathbb{R}^p$ for some p if and only if $F(D)$ is psd, where

$$F(D) = -1/2 JD \circ DJ \ .$$

and J is the $K \times K$ centering matrix.

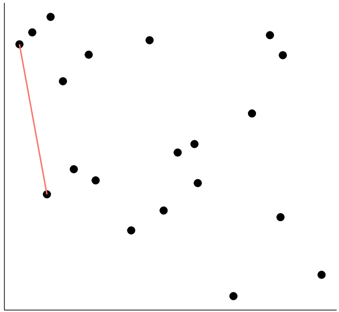
Embedding Conditions - Example

$z_i \in \mathbb{R}^2$



Embedding Conditions - Example

$z_i \in \mathbb{R}^2$

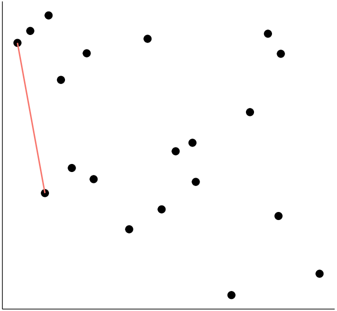


$\lambda_1(F(D)) = 0$

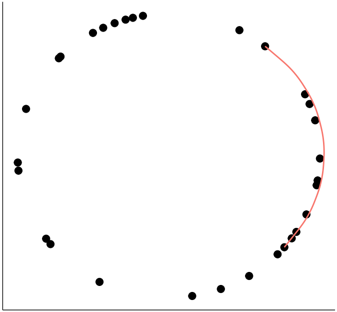
$F(D)$ is psd ✓

Embedding Conditions - Example

$z_i \in \mathbb{R}^2$



$z_i \in \mathbb{S}^1(1)$

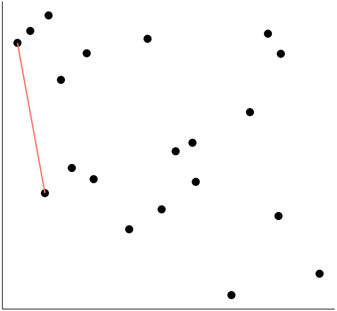


$\lambda_1(F(D)) = 0$

$F(D)$ is psd ✓

Embedding Conditions - Example

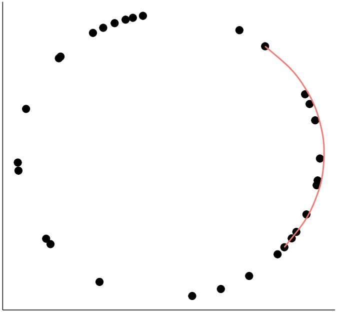
$z_i \in \mathbb{R}^2$



$$\lambda_1(F(D)) = 0$$

$F(D)$ is psd

$z_i \in \mathbb{S}^1(1)$



$$\lambda_1(F(D)) < 0$$

$F(D)$ is not psd

HT Framework

If D is **ANY** distance matrix for points on $\mathcal{M}$:

HT Framework

If D is **ANY** distance matrix for points on $\mathcal{M}$:

$$H_0 : \mathcal{M} = \mathbb{R}^p, \quad H_a : \mathcal{M} \neq \mathbb{R}^p$$

$$\Updownarrow$$

$$H_0 : \lambda_1(F(D)) \geq 0, \quad H_a : \lambda_1(F(D)) < 0 \ .$$

HT Framework

If D is **ANY** distance matrix for points on $\mathcal{M}$:

$$H_0 : \mathcal{M} = \mathbb{R}^p, \quad H_a : \mathcal{M} \neq \mathbb{R}^p$$

$$\Updownarrow$$

$$H_0 : \lambda_1(F(D)) \geq 0, \quad H_a : \lambda_1(F(D)) < 0 \; .$$

- Similar results hold for spherical and hyperbolic spaces.

HT Framework

If D is **ANY** distance matrix for points on $\mathcal{M}$:

$$H_0 : \mathcal{M} = \mathbb{R}^p, \quad H_a : \mathcal{M} \neq \mathbb{R}^p$$

$$\Updownarrow$$

$$H_0 : \lambda_1(F(D)) \geq 0, \quad H_a : \lambda_1(F(D)) < 0 \;.$$

- Similar results hold for spherical and hyperbolic spaces.
- **Main point: Eigenvalues of distance matrix tell us the geometry! So how do we find points on $\mathcal{M}$?**

HT Framework

To create hypothesis testing framework, we have three steps:

HT Framework

To create hypothesis testing framework, we have three steps:

1. Show that the eigenvalues of distance matrices for points on $\mathcal{M}$ tell us the geometry. So the eigenvalues become our test statistics.
2. Estimate distances between LS positions using graph.
3. Use sampling uncertainty to construct distribution of test statistic.

Convert probabilities to distances

Suppose $\nu_i = 0$, so

$$P(G_{ij} = 1|z) = \exp(-d(z_i, z_j)) \ .$$

Convert probabilities to distances

Suppose $\nu_i = 0$, so

$$P(G_{ij} = 1|z) = \exp(-d(z_i, z_j)) \ .$$

Now solve for distance

$$d(z_i, z_j) = -\log(P(G_{ij} = 1|z)) \ .$$

Convert probabilities to distances

Suppose $\nu_i = 0$, so

$$P(G_{ij} = 1|z) = \exp(-d(z_i, z_j)) \ .$$

Now solve for distance

$$d(z_i, z_j) = -\log(P(G_{ij} = 1|z)) \ .$$

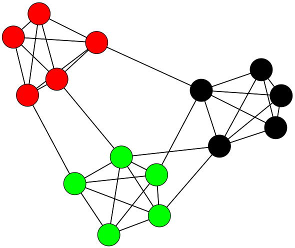
But how do we estimate $P(G_{ij} = 1|z)$?

Embedding Problem

Let's use cliques!

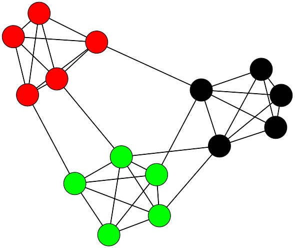
Embedding Problem

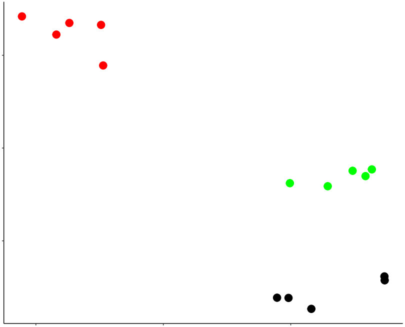
Let's use cliques!



Embedding Problem

Let's use cliques!





How to estimate D from Cliques

Nodes in cliques C_k and $C_{k'}$ are at roughly the same positions,

$$P(G_{kk'} = 1|\nu, z)$$

How to estimate D from Cliques

Nodes in cliques C_k and $C_{k'}$ are at roughly the same positions,

$$P(G_{kk'} = 1 | \nu, z) \approx \frac{1}{\ell^2} \sum_{i \in C_k} \sum_{j \in C_{k'}} G_{ij}$$

How to estimate D from Cliques

Nodes in cliques C_k and $C_{k'}$ are at roughly the same positions,

$$P(G_{kk'} = 1|\nu, z) \approx \frac{1}{\ell^2} \sum_{i \in C_k} \sum_{j \in C_{k'}} G_{ij} := \hat{p}_{k,k'} \ .$$

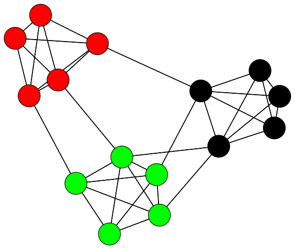
and $\hat{d}_{k,k'} = -\log(\hat{p}_{k,k'})$ by graph model!

How to estimate D from Cliques

Nodes in cliques C_k and $C_{k'}$ are at roughly the same positions,

$$P(G_{kk'} = 1|\nu, z) \approx \frac{1}{\ell^2} \sum_{i \in C_k} \sum_{j \in C_{k'}} G_{ij} := \hat{p}_{k,k'} \ .$$

and $\hat{d}_{k,k'} = -\log(\hat{p}_{k,k'})$ by graph model!

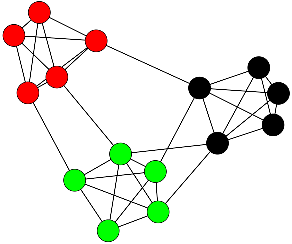


How to estimate D from Cliques

Nodes in cliques C_k and $C_{k'}$ are at roughly the same positions,

$$P(G_{kk'} = 1 | \nu, z) \approx \frac{1}{\ell^2} \sum_{i \in C_k} \sum_{j \in C_{k'}} G_{ij} := \hat{p}_{k,k'} \ .$$

and $\hat{d}_{k,k'} = -\log(\hat{p}_{k,k'})$ by graph model!



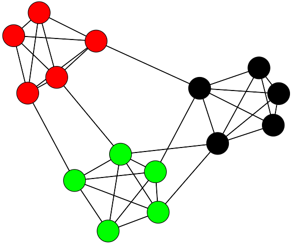
$$\hat{D} = \begin{pmatrix} 0 & -\log\left(\frac{2}{25}\right) & -\log\left(\frac{3}{25}\right) \\ -\log\left(\frac{2}{25}\right) & 0 & -\log\left(\frac{1}{25}\right) \\ -\log(\frac{3}{25}) & -\log(\frac{1}{25}) & 0 \end{pmatrix}$$

How to estimate D from Cliques

Nodes in cliques C_k and $C_{k'}$ are at roughly the same positions,

$$P(G_{kk'} = 1|\nu, z) \approx \frac{1}{\ell^2} \sum_{i \in C_k} \sum_{j \in C_{k'}} G_{ij} := \hat{p}_{k,k'} \ .$$

and $\hat{d}_{k,k'} = -\log(\hat{p}_{k,k'})$ by graph model!



$$\hat{D} = \begin{pmatrix} 0 & -\log\left(\frac{2}{25}\right) & -\log\left(\frac{3}{25}\right) \\ -\log\left(\frac{2}{25}\right) & 0 & -\log\left(\frac{1}{25}\right) \\ -\log\left(\frac{3}{25}\right) & -\log\left(\frac{1}{25}\right) & 0 \end{pmatrix}$$

Find K cliques of size ℓ and compute $\hat{D}$ using these cliques.

HT Framework

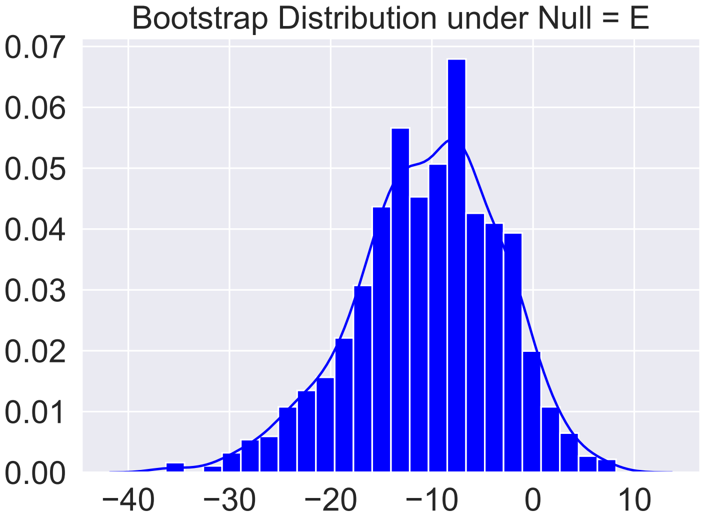
To create hypothesis testing framework, we have three steps:

HT Framework

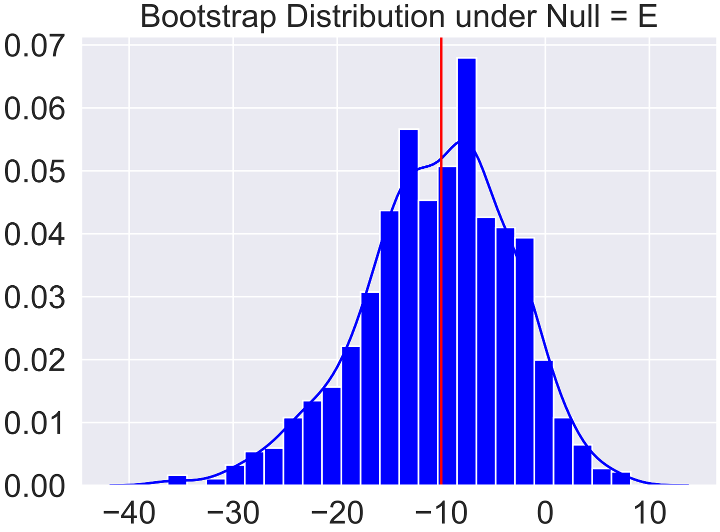
To create hypothesis testing framework, we have three steps:

1. Show that the eigenvalues of distance matrices for points on $\mathcal{M}$ tell us the geometry. So the eigenvalues become our test statistics.
2. Estimate distances between LS positions using graph.
3. Use sampling uncertainty to construct distribution of test statistic.

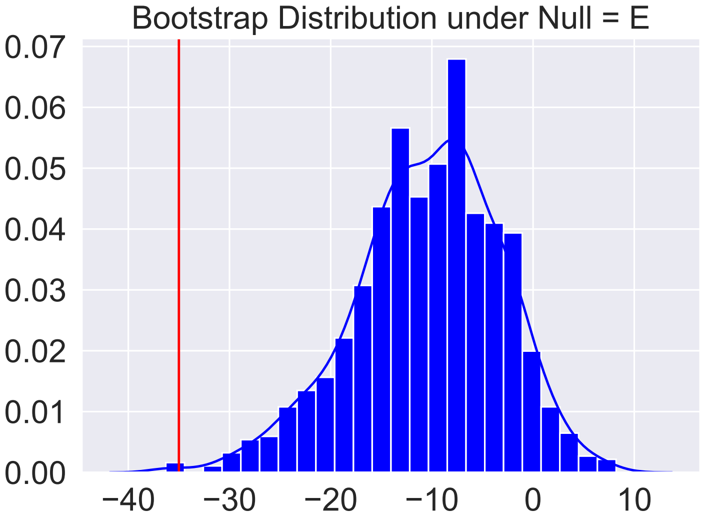
Bootstrapping $\lambda_{\min}(F(D))$



Bootstrapping - Euclidean!



Bootstrapping - Not Euclidean!



Review of HT Framework

To create hypothesis testing framework, we have three steps:

Review of HT Framework

To create hypothesis testing framework, we have three steps:

1. Show that the eigenvalues of distance matrices for points on $\mathcal{M}$ tell us the geometry. So the eigenvalues become our test statistics.

Review of HT Framework

To create hypothesis testing framework, we have three steps:

1. Show that the eigenvalues of distance matrices for points on $\mathcal{M}$ tell us the geometry. So the eigenvalues become our test statistics.
2. Estimate distances between LS positions using graph.

Review of HT Framework

To create hypothesis testing framework, we have three steps:

1. Show that the eigenvalues of distance matrices for points on $\mathcal{M}$ tell us the geometry. So the eigenvalues become our test statistics.
2. Estimate distances between LS positions using graph.
3. Use sampling uncertainty to construct distribution of test statistic.

Review of HT Framework

To create hypothesis testing framework, we have three steps:

1. Show that the eigenvalues of distance matrices for points on $\mathcal{M}$ tell us the geometry. So the eigenvalues become our test statistics.
2. Estimate distances between LS positions using graph.
3. Use sampling uncertainty to construct distribution of test statistic.

G $\rightarrow$ Cliques

Review of HT Framework

To create hypothesis testing framework, we have three steps:

1. Show that the eigenvalues of distance matrices for points on $\mathcal{M}$ tell us the geometry. So the eigenvalues become our test statistics.
2. Estimate distances between LS positions using graph.
3. Use sampling uncertainty to construct distribution of test statistic.

$$\text{G} \rightarrow \text{Cliques} \rightarrow \hat{D} \rightarrow \text{Dist. of Test Statistic} \,.$$

Simulations

- We generate graphs on $n = 1200$ nodes.

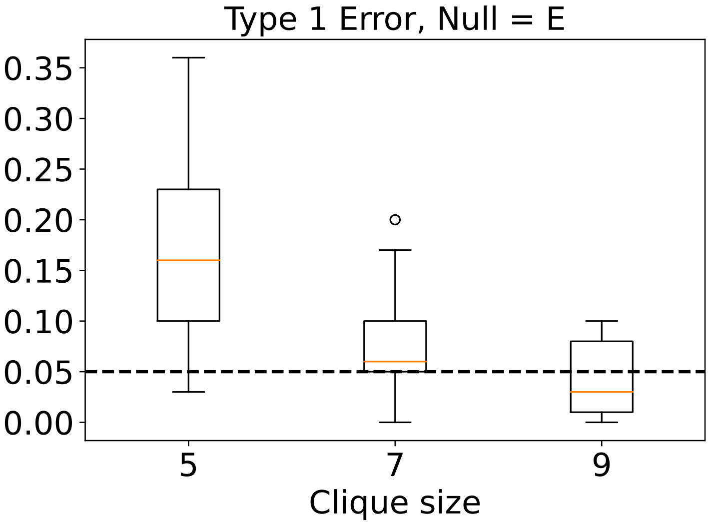
Simulations

- We generate graphs on $n = 1200$ nodes.
- We generate 25 sets of LS positions, then 100 networks from each set.
  - First draw K group centers. Create equal sized groups.
  - Distribute LS locations around these group centers.

Simulations

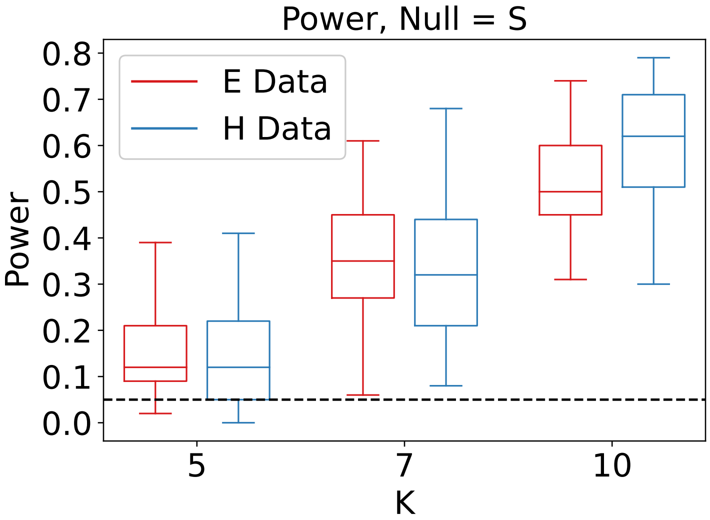
- We generate graphs on $n = 1200$ nodes.
- We generate 25 sets of LS positions, then 100 networks from each set.
  - First draw K group centers. Create equal sized groups.
  - Distribute LS locations around these group centers.
- We draw $\nu_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(-0.2, 0)$.

Type 1 error approaches α as clique size increases.



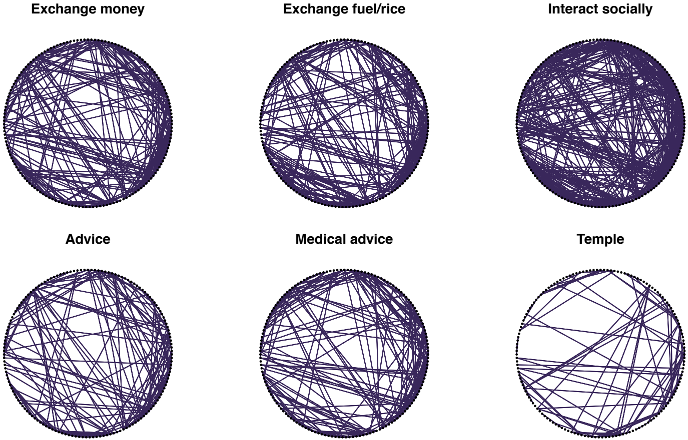
Power: As number of cliques increases, power increases.

Power: As number of cliques increases, power increases.



Banerjee, Duflo, Chandrasekhar and Jackson Data

- 75 villages from Karnakata, India.
- We explored, for example, how predicted geometry affects likelihood of loans between villagers.



Summary of LS work

- We proposed first tests of latent space geometry. (curvature and dimension not discussed).
- We showed our methods identify signal in data consistent with economic intuition.

Summary of LS work

- We proposed first tests of latent space geometry. (curvature and dimension not discussed).
- We showed our methods identify signal in data consistent with economic intuition.

G

Summary of LS work

- We proposed first tests of latent space geometry. (curvature and dimension not discussed).
- We showed our methods identify signal in data consistent with economic intuition.

G $\rightarrow$ Cliques

Summary of LS work

- We proposed first tests of latent space geometry. (curvature and dimension not discussed).
- We showed our methods identify signal in data consistent with economic intuition.

$$\text{G} \rightarrow \text{Cliques} \rightarrow \hat{D}$$

Summary of LS work

- We proposed first tests of latent space geometry. (curvature and dimension not discussed).
- We showed our methods identify signal in data consistent with economic intuition.

$$\text{G} \to \text{Cliques} \to \hat{D} \to \text{Dist. of Test Statistic} \,.$$

Future Work

- How to combine three HTs into one estimate of geometry?
- Can we test if the constant curvature assumption is satisfied?
- What if there are few cliques? Different geometry characterizations might lead to better tests [Gu et al., 2019]?
- The sub-sampling method requires several parameters. Can we use a simpler resampling method, like [Levin and Levina, 2019]?

More Information

At https://slubold.github.io/

- Arxiv pre-print
- Code
- Blog post on major points of paper
- These slides!

References I

Asta, D. M. and Shalizi, C. R. (2015).
Geometric network comparisons.
In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 102–110. AUAI Press.

Friel, N., Rastelli, R., Wyse, J., and Raftery, A. E. (2016).
Interlocking directorates in irish companies using a latent space model for bipartite networks.
Proceedings of the National Academy of Sciences, 113(24):6629–6634.

Gu, A., Sala, F., Gunel, B., and Ré, C. (2019).
Learning mixed-curvature representations in products of model spaces.

References II

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007).
Model-based clustering for social networks.
J. R. Statist. Soc. A.

Hoff, P. (2005).
Bilinear mixed-effects models for dyadic data.
Journal of the American Statistical Association.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002).
Latent space approaches to social network analysis.
Journal of the American Statistical Association.

Lee, J., Li, G., and Wilson, J. D. (2020).
Varying-coefficient models for dynamic networks.
Computational Statistics Data Analysis.

Levin, K. and Levina, E. (2019).
Bootstrapping networks with latent space structure.

References III

McCormick, T. H. and Zheng, T. (2015).
Latent surface models for networks using aggregated relational data.
Journal of the American Statistical Association, 110(512):1684–1695.

Oh, M.-S. and Raftery, A. (2001).
Bayesian multidimensional scaling and choice of dimension.
Journal of the American Statistical Association.

Salter-Townshend, M. and McCormick, T. H. (2017).
Latent space models for multiview network data.
Annals of Applied Statistics.

Schoenberg, I. (1935).
A remark to maurece frechets article.
Ann. of Math, 36(3).

References IV

Shalizi, C. R. and Asta, D. (2017).
Consistency of maximum likelihood for continuous-space network models.
arXiv preprint arXiv:1711.02123.

Smith, A. L., Asta, D. M., Calder, C. A., et al. (2019).
The geometry of continuous latent space models for network data.
Statistical Science, 34(3):428–453.

Sosa, J. and Rodriguez, A. (2018).
A latent space model for cognitive social structures data.

Clique sizes

- We do not need all cliques!
- We

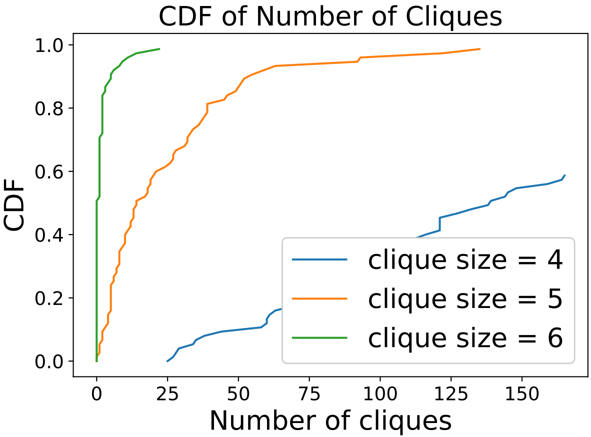


Figure 1: Number of cliques of size $\ell \in \{4, 5, 6\}$ in the Indian villages data

Picking Cliques

$$\begin{aligned}\hat{C}_1, \ldots, \hat{C}_K \in \underset{C_1,\ldots,C_k}{\text{argmin}} \sum_{i,j}^{K} |C_i \cap C_j| \\ \text{such that} \quad |C_i| = \ell \text{ for each } i \text{ and} \\ \hat{P}(C_1, \ldots, C_K) \text{ does not contains a 0.}\end{aligned} \tag{2}$$