A Hypothesis Test for Network Comparison

Ha Khanh Nguyen
The Ohio State University

Dec 10, 2018

Introduction

Big Network Analysis Challenge: *Researchers want to compare two observed networks and decide whether they come from the same network model or not.*

- Identify different groups of brain networks
- Compare protein-protein interaction networks
- Compare communication interactions in different social groups

This is an ongoing research project with Jinzhao (Daniel) Chen, Kartik Lovekar, and Dr. Vishesh Karwa.

Goal: *Define a statistical framework for comparing two networks*

Result: *Given any metric that measures the distance between two networks, we propose a hypothesis test to calibrate that test to the right type I error.*

Project Goals

1. Examine the proposed test under the light of the permutation test theory
2. Implement the test in R and simulate networks from different network models to estimate the test type I error and power
3. Explore the effect of different sampling methods and the sampling rate on the test performance

Background and Notation

Assume $G_1 \sim \mathbb{P}_1$ and $G_2 \sim \mathbb{P}_2$.

Hypotheses:

$$H_0 : \mathbb{P}_1 = \mathbb{P}_2 \textit{ vs. } H_1 : \mathbb{P}_1 \neq \mathbb{P}_2$$

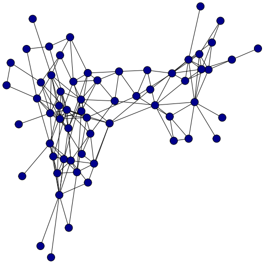
Input: G_1, G_2, α (type I error) and a graph metric $\rho(u, v)$.
$\rho(u, v)$ has to satisfy the following 4 conditions:

1. $\rho(u, u) = \rho(v, v) = 0$
2. $\rho(u, v) = \rho(v, u)$
3. $\rho(u, v)$ is graph invariant.
4. $\rho(u, v)$ does not depend on the sizes of u and v.

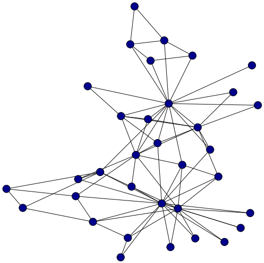
Output: p-value, reject H_0/fail to reject H_0.

Example

Consider the graph G_1 and G_2 in the figure below.



Dolphins Network (n = 62)



Karate Network (n = 34)

Proposed Test

Part 1: Generate M samples each from G_1 and G_2.

$$X_1, \ldots, X_M \overset{\text{sample}}{\sim} G_1$$

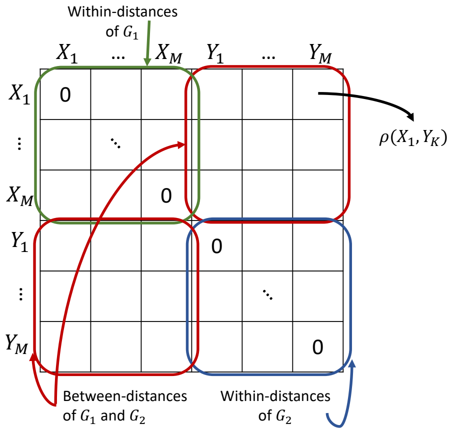
$$Y_1, \ldots, Y_M \overset{\text{sample}}{\sim} G_2$$

Assume the sampling method preserves the properties of the original graph and the samples are independent from one another.

$$X_1, \ldots, X_M \overset{iid}{\sim} \mathbb{P}_1,$$
$$Y_1, \ldots, Y_M \overset{iid}{\sim} \mathbb{P}_2.$$

Part 2: Matrix Permutation

1. Compute the distance matrix, D:



Part 2 (cont.): Matrix Permutation

2. Compute the test statistic

$$T_{\text{obs}} = \frac{\text{mean(within-distances)}}{\text{mean(between-distances)}}$$

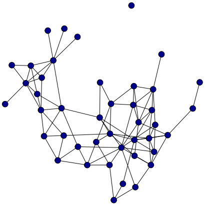
3. Permutation Step
   For the k-th permutation, compute

$$T^{(k)} = \frac{\text{mean}(\text{within-distances}^{(k)})}{\text{mean}(\text{between-distances}^{(k)})}$$

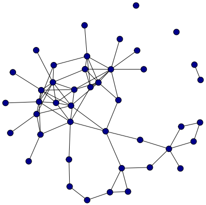
Repeat the permutation step B times. We have

$$p\text{-value} = \frac{\#\text{ of } T^{(i)} \leq T_{\text{obs}}}{B}, 1 \leq i \leq B$$

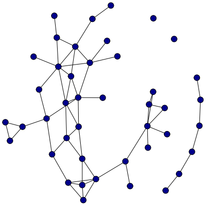
Example (Dolphins Network)



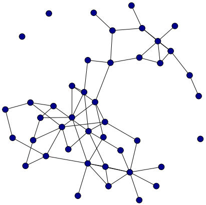
Subgraph 1



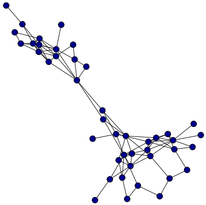
Subgraph 2



Subgraph 3

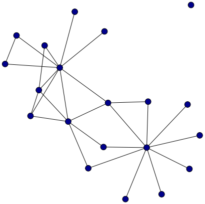


Subgraph 4

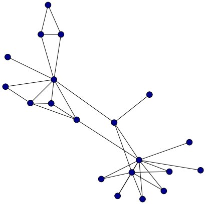


Subgraph 5

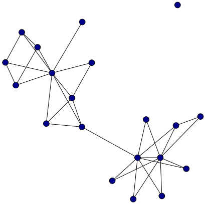
Example (Karate Network)



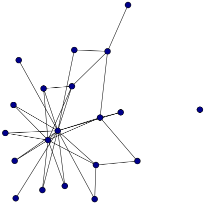
Subgraph 1



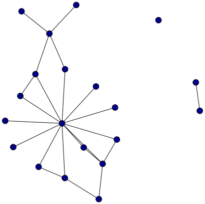
Subgraph 2



Subgraph 3



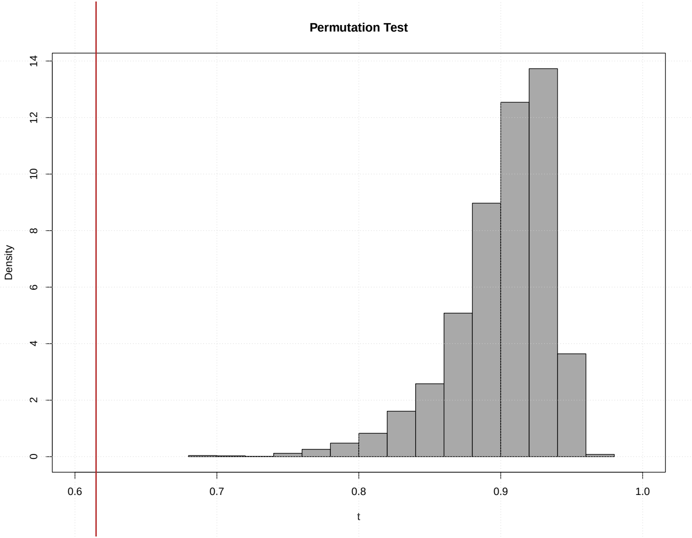
Subgraph 4



Subgraph 5

Example

We have: $T_{obs} = 0.6149$ and p-value $= 0$.



Simulation Results (with Ideal Sampling Assumption)

Network Model	Test Statistic	Graph Metric	# of Samples (K)	# of Perms (B)	Type I Error
Barabasi-Albert	Ratio	KS-dist	10	5000	0.052
	Ratio	Var-Cov Matrix	10	5000	0.062
	Sum	KS-dist	10	5000	0.048
	Sum	Var-Cov Matrix	10	5000	0.042
Erdos-Renyi	Ratio	KS-dist	10	5000	0.064
	Ratio	Var-Cov Matrix	10	5000	0.054
	Sum	KS-dist	10	5000	0.044
	Sum	Var-Cov Matrix	10	5000	0.044
Geometric Random Graph	Ratio	KS-dist	10	5000	0.046
	Ratio	Var-Cov Matrix	10	5000	0.054
	Sum	KS-dist	10	5000	0.048
	Sum	Var-Cov Matrix	10	5000	0.034

Simulation Results (with Ideal Sampling Assumption)

Network Model	Test Statistic	Graph Metric	# of Samples (K)	# of Perms (B)	Power
Barabasi-Albert vs. Erdos-Renyi	Ratio	KS-dist	10	5000	1
	Ratio	Var-Cov Matrix	10	5000	1
	Sum	KS-dist	10	5000	0.98
	Sum	Var-Cov Matrix	10	5000	0.98
Erdos-Renyi vs. Geometric	Ratio	KS-dist	10	5000	1
	Ratio	Var-Cov Matrix	10	5000	1
	Sum	KS-dist	10	5000	0.99
	Sum	Var-Cov Matrix	10	5000	0.98
Barabasi-Albert vs. Geometric	Ratio	KS-dist	10	5000	1
	Ratio	Var-Cov Matrix	10	5000	1
	Sum	KS-dist	10	5000	0.99
	Sum	Var-Cov Matrix	10	5000	0.99

Simulation Results

Network Model	Test Statistic	Graph Metric	# of Samples (K)	Sampling Rate	# of Perms (B)	Type I Error
Barabasi-Albert	Ratio	KS-dist	10	0.4	10000	0.09
	Ratio	Var-Cov Matrix	10	0.4	10000	0.07
	Sum	KS-dist	10	0.4	10000	0.05
	Sum	Var-Cov Matrix	10	0.4	10000	0.07
Erdos-Renyi	Ratio	KS-dist	10	0.4	10000	**0.228**
	Ratio	Var-Cov Matrix	10	0.4	10000	**0.128**
	Sum	KS-dist	10	0.4	10000	**0.214**
	Sum	Var-Cov Matrix	10	0.4	10000	0.106
Geometric Random Graph	Ratio	KS-dist	10	0.4	5000	**0.18**
	Ratio	Var-Cov Matrix	10	0.4	5000	**0.18**
	Sum	KS-dist	10	0.4	5000	**0.22**
	Sum	Var-Cov Matrix	10	0.4	5000	**0.17**

Discussion

1. More investigation on the effects of sampling method and sampling rate on the type 1 error and power of the test
2. Try the test on other popular network models such as ERGM, Stochastic Block Models, etc.
3. Apply the test to solve a real-world problem

References

- N. Ahmed, J. Neville, R. Kompella, *Network Sampling via Edge-based Node Selection with Graph Induction*, Purdue University e-Pubs (2011).
- D. Asta, C. Shalizi, *Geometric Network Comparisons*, Proceedings of the 31st Annual Conference on Uncertainty in AI (2015).
- P. Good, *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, 2nd Ed, Springer (2000).
- S. Simpsons, R. Lyday, S. Hayasaka, A. Marsh, and P. Laurienti, *A Permutation Test Framework to Compare Groups of Brain Networks*, Frontiers in Computational Neuroscience (2013).