A Hypothesis Test for Network Comparison

THE OHIO STATE UNIVERSITY

Jinzhao Chen, Kartik Lovekar, Ha Khanh Nguyen

Faculty Advisor: Vishesh Karwa

Introduction

A principled method for comparing network can find applications in many areas. An often-used approach to achieve this goal is the use of network summary statistics fed into a standard statistical test (K-S test, two-sample t-test etc.) Such methods are generally not invariant to the size of the network and overlook local topological features. Also, it is not possible to find a "one-size-fits-all" metric for comparing two networks.

The goal of this project is to define a statistical framework for comparing two networks. Our key contribution is that given any metric that measures the distance between two networks, we propose a method to state the statistical significance of the difference between the two networks.

Background and Notation

Let G_1 and G_2 be graphs with n_1 and n_2 vertices respectively.

$$G_1 \overset{\text{iid}}{\sim} \mathbb{P}_1 \text{ and } G_2 \overset{\text{iid}}{\sim} \mathbb{P}_2.$$

We want to test:

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

Input: G_1, G_2, α (type I error) and a graph metric $\rho(u,v)$.

Given two graphs u and v, $\rho(u,v)$ has to satisfy the following 4 conditions:

1. $\rho(u,v) \geq 0$ and $\rho(u,u) = 0$.
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.

Hypothesis Test

We are interested in testing:

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

Step 1: Generate **M** samples each from G_1 and G_2.

$$X_1, \ldots, X_M \overset{\text{sample}}{\sim} G_1 \overset{\text{iid}}{\sim} \mathbb{P}_1$$

$$Y_1, \ldots, Y_M \overset{\text{sample}}{\sim} G_2 \overset{\text{iid}}{\sim} \mathbb{P}_2$$

The choice of the sampling method depends on the metric chosen such that the metric is not distorted in the pseudo samples.

$$X_1, \ldots, X_M \overset{\text{pseudo-sample}}{\sim} \mathbb{P}_1,$$

$$Y_1, \ldots, Y_M \overset{\text{pseudo-sample}}{\sim} \mathbb{P}_2.$$

Step 2: We will use the idea of permutation test to compare these two sets of sampled networks.

Method 1: **Sample Set Permutation**

1. Compute:

$$\delta_1 = \frac{1}{\binom{M}{2}} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \rho(X_i, X_j), \quad \delta_2 = \frac{1}{\binom{M}{2}} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \rho(Y_i, Y_j).$$

These two values, δ_1 and δ_2, present the average *within* distances of G_1 and G_2 respectively.

2. Our test statistic is the sum of the average *within* distances of the two networks we are trying to compare,

$$T_{\text{obs}} = \delta_1 + \delta_2.$$

3. Permutation: Permute the networks in these two sets B times. For the k-th permutation ($1 \leq k \leq B$), we compute:

$$T^{(k)} = \delta_1^{(k)} + \delta_2^{(k)} \text{ with}$$

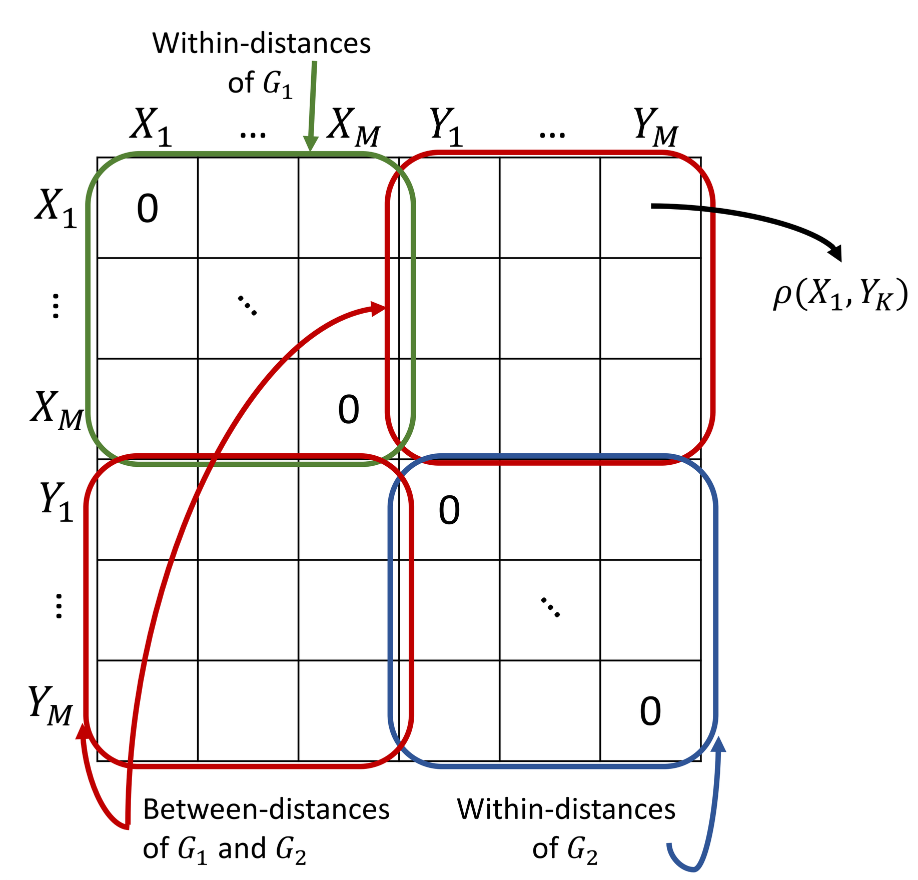
$$\delta_1^{(k)} = \frac{1}{\binom{M}{2}} \sum_{h=1}^{M-1} \sum_{l=h+1}^{M} \rho(X_h^{(k)}, X_l^{(k)}) \text{ and } \delta_2^{(k)} \text{ is defined similarly.}$$

After B permutations, we get $T^{(1)}, \ldots, T^{(B)}$, these form our sampling distribution for the test statistic T.

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

Method 2: **Matrix Permutation**

1. Compute the matrix D shown in the figure below.



Now, compute:

$$T_{\text{obs}} = \frac{\text{mean(within-distances)}}{\text{mean(between-distances)}} = \frac{\frac{1}{2M(M-1)}\left[\sum_{i=1}^{M}\sum_{j=1, i\neq j}^{M} \rho(X_i, X_j) + \sum_{i=1}^{K}\sum_{j=1, i\neq j}^{M} \rho(Y_i, Y_j)\right]}{\frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} \rho(X_i, Y_j)}$$

2. Permutation: Exchange the labels of the first M rows of D with the last M rows of D as well as exchanging the labels of the corresponding columns in D.

For the k-th permutation, we get permuted matrix, $D^{(k)}$, with rows and columns labeled as $D_1^{(k)}, \ldots, D_{2M}^{(k)}$.

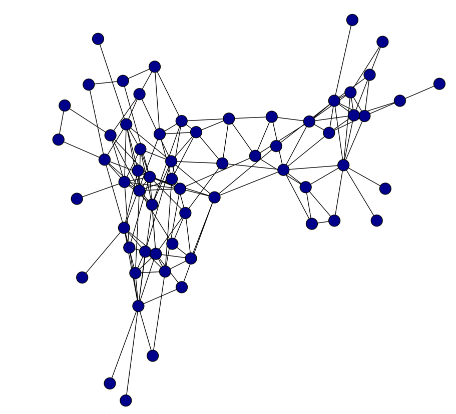
$$T^{(k)} = \frac{\text{mean(within-distances}^{(k)})}{\text{mean(between-distances}^{(k)})}$$

where mean(within-distances$^{(k)}$) and mean(between-distances$^{(k)}$) are calculated similarly to above.

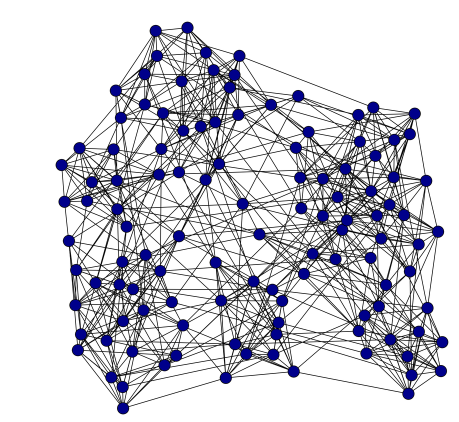
After permuting B times, $T^{(1)}, \ldots, T^{(B)}$ forms the sampling distribution of the test statistic T.

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

Real-world Network Example

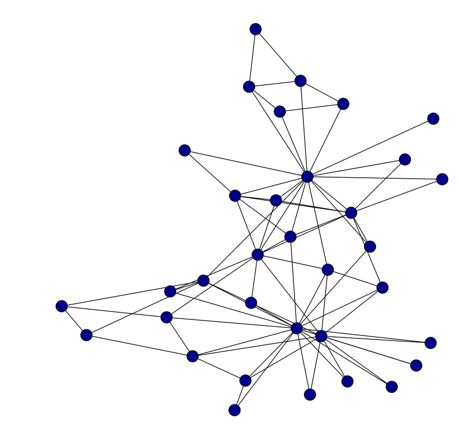


Dolphins Network



Football Network

We run the proposed hypothesis test (method 2) on the three famous real-world networks: Dolphin, Karate, and Football. The sampling method we use is TIES (edge-based node selection with graph induction).



Karate Network

We get the following results for Type I Error and Power of the test:

Network	Sampling Rate	# of Samples (M)	# of Perms (B)	Type I Error
Dolphin	0.7	30	5000	0.045
Football	0.7	30	5000	0.045
Karate	0.7	30	5000	0.035

Networks	Sampl. Rate	# of Samples (M)	# of Perms (B)	Power
Dolphin vs. Karate	0.7	30	5000	0.93
Football vs. Karate	0.7	30	5000	0.91
Football vs. Dolphin	0.7	30	5000	0.95

References

[1] N. Ahmed, J. Neville, R. Kompella, Network Sampling via Edge-based Node Selection with Graph Induction, *Purdue University e-Pubs* (2011).

[2] D. Asta, C. Shalizi, Geometric Network Comparisons, *Proceedings of the 31st Annual Conference on Uncertainty in AI* (2015).

[3] P. Good, Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses, 2nd Ed *Springer* (2000).

[4] 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).