Network Similarity using Distribution of Distance Matrices

Ralucca Gera  
Department of Applied Mathematics  
Naval Postgraduate School,  
Monterey, CA 93943 USA  
rgera@nps.edu

Ruriko Yoshida  
Department of Operations Research  
Naval Postgraduate School,  
Monterey, CA 93943 USA  
ryoshida@nps.edu

Abstract—Decision makers use partial information of networks to guide their decision, yet when they act, they act in the real network or the ground truth. Therefore, a way of comparing the partial information to ground truth is required. We introduce a statistical measure that analyzes the network obtained from the partially observed information and ground truth, which of course can be applied to the comparison of any networks. As a first step, in the current research, we restrict ourselves to networks of the same size to introduce such a method, which can be generalized to different size networks. We perform mathematical analysis on the random graph, and then apply our methodology to synthetic networks generated using five different generating models. We conclude with a statistical hypothesis test to decide whether two graphs are correlated or not correlated.

I. INTRODUCTION

Understanding and exploring data in today’s digital world is difficult due to the complex interconnections of systems we study and the big data they generate. Network Science is commonly used to describe and study such systems, by creating networks of interactions between people, objects, and abstract concepts. Consider an environment in which capturing the data is done progressively, such as assembling the social network of person based on information gathered from a cellphone. The question we would like to answer is: at what point have we captured enough information about the social network (partial information network), to know that it is representative of the real social network of the person (ground truth). That is, humans tend to base decisions on the partial information that is available, yet act in the ground truth one. An accurate depiction of a network that is similar to the ground truth is important for a meaningful analysis.

Comparing two graphs has always been a problem of interest for Graph Theorists [1], and lately for Network Scientists. Researchers have used various metrics to compare two given graphs like degree distributions, density, clustering coefficient, average path length, Maximum Common Subgraph, number of spanning trees, Hamming Distance, etc. These properties only compare the macro-level properties of the networks, but the micro-level binding of the network is not considered.

Kashima and Inokuchi use graph kernel method to classify the graphs where a graph kernel is defined using a random walk on the vertex product graph of two graphs [2]. Borgwardt and Kriegel proposed graph kernels based on shortest paths and verified it by classifying the proteins graphs model into functional classes [3]. Pržulj proposed a measure based on the graphlet degree distribution to compare two large networks, but the proposed measure is computationally costly [4]. Wilson and Zhu compared two graphs using spectra and show that the Euclidean distance between spectra of two graphs is highly correlated with their edit distance [5]. Zager and Verghese use the structural similarity of local neighborhoods of the nodes two compare different graphs [6]. Aliakbary et al. proposed a genetic algorithms based approach to compute the similarity measure between two networks of different sizes [7]. Davis et al. propose a random walk based method to compare two networks by comparing the explored percentages of nodes and edges at some iteration while exploring the networks [8]. Crawford et al. introduced two metrics to compare the topologies of two networks using the eigenvalues of the Adjacency matrix and the Laplacian Matrix [9]. Roginski et al. used the neighbor matrix to compare graphs and show that it captures eleven important characteristics of the networks like degree, distances, closeness, and betweenness, etc. [10].

A pragmatic approach to isomorphism is to identify if two graphs are almost the same, namely similar. In this research, we assume that two graphs are similar if they were created in a similar fashion by following the same rules. This would then extend to real networks by assuming that similar real-world networks became what they are through a similar process. This could include the case of comparing a depiction of a network to the ground truth itself due to insufficient information or the impossibility of capturing the entire network. We applied our methodology to synthetic networks whose number of nodes we can control, allowing us to compare their topology rather than sizes. As our experiment was very successful, we plan on extending our methodology to real-world networks having different node counts.

In this paper, we investigate the distribution of a distance metric between random graphs. We wish to compare graphs, or sets of graphs, to each other in a statistical framework and ask whether or not they are significantly different, more specifically: Suppose \(G_1\) and \(G_2\) have the same number of nodes. Then, we have the following hypotheses.
H₀ : two given graphs G₁ and G₂ do not have similar structure to each other.
H₁ : two given graphs have similar structure.

In terms of statistics, these hypotheses as the following:

H₀ : two given graphs G₁ and G₂ are not correlated,
H₁ : two given graphs are correlated.

A key to conduct this hypothesis test is to understand the distribution of a distance measure d(G₁, G₂) between G₁ and G₂, which can be used as a test statistic for the hypotheses above.

If we understand the distribution of the test statistics, then we can conduct the hypothesis test or significance test. Thus here we investigate the distribution of d(G₁, G₂) where G₁ and G₂ are sampled from some distribution, where d(G₁, G₂) is the path distance defined in Section III.

This paper is organized as follows: Section II introduces the motivation and the generating models to create the synthetic networks. Section III explains the distance matrix followed by the sensitivity analysis on d(G₁, G₂) if we perturb G₁ by one edge. In Section III-C we compute the mean and variance of d²(G₁, G₂) if we sample G₁ and G₂ uniformly. The experimental results on synthetic graphs are discussed in Section IV. We then present our conclusions and further directions in Section V.

II. PROBLEM DEFINITION AND PRELIMINARIES

In the current research, we investigate a network similarity method to compare given networks based on limited information of a network, such as knowing just the network’s topology.

The standard graph comparison is done through isomorphisms of graphs. However, the isomorphism is hard, its decision problem being still open. In addition, we would like to move away from the limitation given by the binary answer: given two graphs, they are isomorphic or nonisomorphic. We would like to determine if the graphs are similar instead. We acknowledge that similarity is a very complex and vast problem. In this work, we propose a metric that recognizes similar graphs; we assume that two graphs are similar if they were created by the same algorithm. In our study, we use synthetic networks to create networks of the same type.

In network science, the understanding of large-scale networks can be done both by the analysis of real networks as well as by modeling them through synthetic networks. The analysis of these synthetic complex networks is allowing us to bring theory into place while mimicking reality. Next, we discuss five distinct synthetic models that we use for generating networks.

A. Erdős-Rényi (ER) Network

For p ∈ [0, 1], n ∈ N, and N = \binom{n}{2}, let Gₚ,n be the probability space (Ω, F, P) which is the product of probability spaces (Ωᵢⱼ, Fᵢⱼ, Pᵢⱼ) for i,j ∈ {1, 2, . . . , n} such that Ωᵢⱼ = {0, 1} and Pᵢⱼ is the probability function of the Bernoulli distribution with a parameter p; for each pair of nodes (i,j), we assign the probability p independently that decides to assign an edge between a node i and a node j in the graph (if a random variable X ∼ Bernoulli(p). If X = 1, then we assign an edge between a node i and a node j in the graph. If X = 0, then we do not assign the edge between a node i and a node j in the graph). Thus for any graph G in Gₚ,n, we have

\[ P(G) = p^{e(G)}(1-p)^{N-e(G)}, \]

where e(G) is the number of edges in G. This is called Erdős-Rényi model of a random graph [11].

B. Barabási-Albert (BA) Network

The Barabási-Albert Model (BA) is a basic scale-free graph that grows through preferential attachment of new nodes based on the degree of the existing nodes in the network. The user determines the parameters for the number of nodes and number of edges per incoming node, which will produce the power law degree distribution. Formally, we obtain a Barabási-Albert model by starting with a small number m₀ of vertices, and at every time step, new vertices yᵢ (i ≥ 1) are added with m < m₀ edges. These new edges link the incoming vertex yᵢ to m different vertices that already exist in the network at the step i when yᵢ is added. To incorporate preferential attachment, we assume that the probability, P, that a new vertex will be connected to a vertex i depends on the degree kᵢ of that vertex, namely \( \frac{kᵢ}{\sum kᵢ} \).

After t steps the model leads to a scale-free network with t + m₀ vertices and nt edges. The preferential attachment process gives rise to the “rich become richer” phenomenon that leads to power-law degree distribution [12].

C. Watts-Strogatz (WS) Network

The Watts-Strogatz model was one of the original models introduced to capture the small world effect, yet allowing the user to choose how much randomness and how much chaos to bring into the creation of the graph. The model starts with a one dimensional lattice of n nodes arranged into a circle, and each node is adjacent to the first k neighbors to its left and the first k neighbors to its right (just like Harary graphs [13] or the k⁰th power of an n-cycle, namely Cₖⁿ). This part creates the structure of the small world. The chaos is then brought in through rewiring some of the edges: with a fixed probability p some end vertices of edges are picked up and randomly rewired to different vertices, thus creating shorter connections between the people that were diametrically opposed in the existing lattice. Note that if p = 1 then one would obtain a random graph as each edge would be rewired randomly, and if p = 0 one obtains the originally structured lattice. The choice of other values of p allows for the combination of chaos and structure. This was introduced by Watts and Strogatz [14].

D. PowerLaw-Cluster (PC) Network

In real-world social networks, people prefer to make connections with the friends of their friends. It increases the number of closed triad connections and the average clustering [15] in the network. Holme and Kim proposed the PowerLaw-Cluster generating model based on both preferential attachment as well as triad connections phenomena [16]. The
network is generated as the Barabási-Albert (BA) growth model with an extra step that at each iteration a node is given a chance of making an edge to one of its neighbors’ neighbors with probability \( p \) and forms a triangle.

E. Community Structured (CS) Network

Community structure is a well known studied meso-scale structure in real-world networks. Here we use a very basic Community Structured generating model, where the nodes are divided into the groups of same sizes. The edges are placed uniformly at random between the nodes in the same group with probability \( p_{in} \) and between the nodes of different groups with probability \( p_{out} \) [17]. The probability of having an intra-community link is higher than the inter-community link.

III. THE PROPOSED METHODOLOGY

A. Distance Matrix

We propose to use the distance matrix (introduced in 1977 by Bonchev and Trinajstic) to find out the graph similarity.

The following definitions for adjacency matrix and distance matrix have been studied in the literature, and we review them here for completeness.

Let \( G_{\nu}(n) \) be the probability space \((G, F, \nu)\) which is the product of probability spaces where \( G \) is the set of all possible connect graphs with \( n \) nodes, \( F \) be the sigma algebra on \( G \) and \( \nu \) is the probability measure on \( F \). We consider a graph \( G \in G_{\nu}(n) \). Note that \( G_{\nu}(n) \) can be any synthetnic network generating model like Barabási-Albert or Erdős-Rényi model.

**Definition 1.** The Adjacency Matrix of \( G \) is a \( n \times n \) matrix \( A(G) \) such that:

\[
    a_{ij}(G) = \begin{cases} 
        1, & \text{node } i \text{ adjacent to node } j \\
        0, & \text{otherwise}
    \end{cases}
\]

\( \forall i, j (1 \leq i, j \leq n) \), where \( n \) is the number of vertices in the graph.

**Definition 2.** The distance \( d(i,j)(G) \), between two vertices \( i \) and \( j \) in \( G \), is the length of a shortest path between \( i \) and \( j \). The Distance Matrix of \( G \) is a \( n \times n \) matrix \( D(G) \) such that

\[
    d_{ij}(G) = d(i,j)(G)
\]

\( \forall i, j (1 \leq i, j \leq n) \), where \( n \) is the number of vertices in the graph.

We vectorize a graph \( G \) by a map \( v: G \rightarrow \mathbb{Z}^N \) as

\[
    v(G) = (d_{12}(G), d_{13}(G), \ldots, d_{n(n-1)}(G)).
\]

**Definition 3.** For connected graphs \( G_1, G_2 \in G \) with \( n \) nodes, we define the path difference distance \( d(G_1, G_2) \) between \( G_1, G_2 \) as

\[
    d(G_1, G_2) := ||v(G_1) - v(G_2)||_2,
\]

where \( || \cdot ||_2 \) is the \( l_2 \) norm on the Euclidean space \( \mathbb{R}^N \).

B. Sensitivity of the Path Difference

In this section we would like to see how sensitive the path difference \( d(G_1, G_2) \) between two graphs \( G_1, G_2 \in G \) by removing one edge \( e \in G_1 \).

Let \( p_G(i,j) \) be the set of all shortest paths in \( G \in G \) from a node \( i \) to a node \( j \), and let \( p_G(e) := \{(i,j) | e \in p_G(i,j), i < j \} \). Let \( P_e(G) := \{ p_G(e) ||p_G(i,j)| = 1 \ \text{with} \ (i,j) \in p_G(e) \} \) for an edge \( e \in G \) and \( G \in G \). Intuitively, for \( e \) a fixed edge of \( G \in G \), \( P_e(G) \) is the all shortest and unique paths from all possible node in \( G \) to another node in \( G \) such that each shortest path goes through an edge \( e \).

**Theorem 1.** Suppose \( G \in G \) are connected graphs with \( n \) nodes. If we remove an edge \( e \) in \( G \), then

\[
    d(G, (G - \{e\})) \geq \sqrt{|P_e(G)|}.
\]

**Proof.** We now compute \( d = d(G, (G - \{e\}) \)

\[
    d = \sqrt{\sum_{i<j} (d_{ij}(G) - d_{ij}(G - \{e\}))^2}
    = \sqrt{\sum_{i<j | e \in p_G(i,j)} (d_{ij}(G) - d_{ij}(G - \{e\}))^2}
    = \left( \sum_{i<j | e \in p_G(i,j), |p_G(i,j)| = 1} (d_{ij}(G) - d_{ij}(G - \{e\}))^2 \right)^{1/2}
    \geq \left( \sum_{i<j | e \in p_G(i,j), |p_G(i,j)| = 1} (d_{ij}(G) - d_{ij}(G - \{e\}))^2 \right)^{1/2}
    = \sqrt{|P_e(G)|}.
\]

**Theorem 2.** \( d(\cdot, \cdot) \) is metric over \( \mathbb{R}^N \).

**Proof.** \( v: G \rightarrow \mathbb{Z}^N \) maps a graph \( G \in G \) to a vector in \( \mathbb{Z}^N \) and \( d(G_1, G_2) := ||v(G_1) - v(G_2)||_2 \) is an \( l_2 \) norm of a vector \( v(G_1) - v(G_2) \in \mathbb{R}^N \) by the definition. Since \( l_2 \) norm in \( \mathbb{R}^N \) is metric, \( d(\cdot, \cdot) \) is metric.

C. Mean and Variance on the Squared Path Difference

In this section, we investigate theoretical results on the distance between networks on random graphs using the path difference distance. We find a closed form for the mean and the variance of the path difference distance in terms of the shortest path between two vertices.

For the following theorems, we assume that \( G_{\nu}(n) \) is the probability space with the uniform probability distribution \( \nu \) over \( G \).

**Theorem 3.** Assume that \( G_{\nu}(n) \) the probability space with the uniform distribution \( \nu \) on \( G \). Let \( G, G' \in G_{\nu}(n) \) be a random connected graphs, and \( \mu(n) \) be the mean value of \( d^2 \). Then

\[
    \mu(n) := \mathbb{E}\{d^2(G_1, G_2)\} = 2 \cdot \binom{n}{2} \mathbb{V}ar[d_{ij}(G)].
\]

**Proof.** Let \( g(n) = |G| \). By definition of \( d^2 \) we have:

\[
    d^2(G, G') = ||v(G) - v(G')||_2^2 = \sum_{i<j} [d_{ij}(G) - d_{ij}(G')]^2,
\]
where $G$ and $G'$ are two random connected graphs in $\mathcal{G}_n(n)$. So the mean is:

$$
\mu(n) = \mathbb{E}[d^2(G, G')] = \sum_{G, G'} \Pr(G) \Pr(G') d^2(G, G')
$$

$$
= \sum_{G, G'} \frac{1}{g(n)^2} \sum_{i<j} [d_{ij}(G) - d_{ij}(G')]^2
$$

$$
= \frac{1}{g(n)^2} \sum_{G, G'} [\sum_{G} d_{ij}(G)^2 + \sum_{G} d_{ij}(G')^2 - 2 \sum_{i<j} d_{ij}(G)d_{ij}(G')] = \frac{1}{g(n)^2} \sum_{G, G'} \frac{\Pr(G) \Pr(G')}{{g(n)}^2} [\sum_{i<j} (d_{ij}(G) - d_{ij}(G'))^2]
$$

$$
= \frac{1}{g(n)^2} \sum_{i<j} \left[ 2 g(n) \sum_{G} d_{ij}(G)^2 - 2 \left( \sum_{G} d_{ij}(G) \right)^2 \right].
$$

Notice that $\sum_{G} f(d_{ij}(G))$ does not depend on the selection of $i$ and $j$ because of the symmetry of labeling (it is easy to prove by contradiction and switching the labels). Therefore, $\sum_{G} f(d_{ij}(G)) = \sum_{G} f(d_{kl}(G))$ with $i < j, k < l$, and thus we have:

$$
\mu(n) = \frac{2}{g(n)^2} \left( \frac{n}{2} \right) \left[ g(n) \sum_{G} d_{ij}(G)^2 - \left( \sum_{G} d_{ij}(G) \right)^2 \right] = 2 \left( \frac{n}{2} \right) \left[ \sum_{G} d_{ij}(G)^2 g(n) - \left( \sum_{G} d_{ij}(G) g(n) \right)^2 \right] = 2 \left( \frac{n}{2} \right) \left[ \mathbb{E}[d_{ij}(G)^2] - \mathbb{E}[d_{ij}(G)]^2 \right] = 2 \left( \frac{n}{2} \right) \mathbb{V}ar(d_{ij}(G))
$$

with any selection of $i$ and $j$. 

**Theorem 4.** Assume that $\mathcal{G}_n(n)$ the probability space with the uniform distribution $\nu$ on $\mathcal{G}$. Let $\sigma^2(n)$ be the variance of $d^2$. Then

$$
\sigma^2(n) = \frac{1}{g(n)^2} \left\{ \sum_{G, G'} \left[ \sum_{i<j} d_{ij}(G)^2 \right]^2 + \sum_{G, G'} \left[ \sum_{i<j} d_{ij}(G')^2 \right]^2 \right\}
$$

$$
= \frac{1}{g(n)^2} \left\{ \sum_{G, G'} \left[ \sum_{i<j} d_{ij}(G)d_{ij}(G') \right]^2 + 4 \left( \frac{n}{2} \right) g(n) \mathbb{E}[d_{ij}(G)]^2 \right\}
$$

$$
= \frac{1}{g(n)^2} \left\{ \sum_{G, G'} \left[ \sum_{i<j} d_{ij}(G) \right]^2 + \sum_{G, G'} \left[ \sum_{i<j} d_{ij}(G') \right]^2 \right\}
$$

$$
= \frac{1}{g(n)^2} \left\{ \sum_{G} \left[ \sum_{i<j} d_{ij}(G) \right]^2 + \sum_{G} \left[ \sum_{i<j} d_{ij}(G') \right]^2 \right\}
$$

$$
= \left\{ \left( \frac{n}{2} \right) g(n) \mathbb{E}[d_{ij}(G)]^2 \right\}.
$$

**Proof.** Since $\sigma^2(n) = \mathbb{V}ar(d^2) = \mathbb{E}[d^4] - \mu(n)^2$, where the explicit formula of $\mu(n)$ is known, we have to consider only $\mathbb{E}[d^4]$:

$$
\mathbb{E}[d^4(G, G')]
$$

$$
= \Pr(G) \Pr(G') [d^2(G, G')]^2
$$

$$
= \frac{1}{g(n)^2} \sum_{G, G'} \frac{\Pr(G) \Pr(G')}{{g(n)}^2} [\sum_{i<j} (d_{ij}(G) - d_{ij}(G'))^2]^2
$$

$$
= \frac{1}{g(n)^2} \sum_{G, G'} \frac{\Pr(G) \Pr(G')}{{g(n)}^2} [\sum_{i<j} d_{ij}(G)^2 + \sum_{i<j} d_{ij}(G')^2 - 2 \sum_{i<j} d_{ij}(G)d_{ij}(G')]^2
$$

$$
= \frac{1}{g(n)^2} \sum_{G, G'} \frac{\Pr(G) \Pr(G')}{{g(n)}^2} [\sum_{i<j} d_{ij}(G)^2 + \sum_{i<j} d_{ij}(G')^2 - 2 \sum_{i<j} d_{ij}(G)d_{ij}(G')]^2
$$

In this equation, two terms can be simplified as:
Theorem 3 and our simulation results, the distribution distance measure $G$. 

### A. Hypothesis test for same generating models

Suppose we have two arbitrary graphs $G_1$ and $G_2$ constructed using the same generating model, each having the same number of $n$ nodes. Then we are interested in testing whether $G_1$ and $G_2$ are similar. Here we are using our path distance measure $d^2(G_1, G_2)$ as a test statistic to conduct a hypothesis test. Thus, our hypotheses become as follows:

$$H_0 : d^2(G_1, G_2) = \mu(n)$$
$$H_1 : d^2(G_1, G_2) \neq \mu(n),$$

where $\mu(n)$ is the mean of $d^2$ for the graphs $G_1$ and $G_2$. With Theorems 3 and 4 and our simulation results, the distribution of $d^2(G_1, G_2)$ can be approximated by $N(\mu(n), \sigma^2(n))$, the normal distribution with the mean $\mu(n)$ and the variance $\sigma^2(n)$. Therefore, we can conduct the hypothesis test with the null and alternative hypotheses on (4) as follows:

1. Compute $d_0 := d^2(G_1, G_2)$ and compute the z-score $z_0 := \frac{d_0 - \mu(n)}{\sigma(n)}$.
2. Compute the rejection region, i.e.,
   $$(-\infty, -z_{1-\alpha/2}), (z_{1-\alpha/2}, \infty)$$
   where $z_{1-\alpha/2}$ is the z-score for $100 \cdot (1 - \alpha/2)$th percentile.
3. If $z_0$ is in the rejection region then return “Reject the null hypothesis.” If $z_0$ is not in the rejection region then return “Not reject the null hypothesis.”

To test our hypothesis, we test 1000 generations of the same model on the same number of nodes and edges (for the experiment). The generation models we use here are introduced in Section II. Each of the graph created was of 1000 nodes and around 6000 edges. We chose them all to have the same number of nodes and edges so that the emphasis is on the structure (namely the emphasis is on the path distance distributions they come from) rather than the size.

The histogram of $z$-scores under each model are shown in Figure 1. Under all models except Watts-Strogatz, the histograms look like the standard normal distribution. The histogram under Watts-Strogatz model seems to be skewed. However, the rejection rate for each model is controlled by the significance level, which is set $\alpha = 0.05$ in this simulation study. Our experiment shows that the hypothesis test which we proposed controls the type I error rate under the significance level (Table I).

### B. Significance test for different generating models

In Section III-C we have worked on a distribution of the path difference distances between two graphs which are generated by the same random process. In Section IV-A we have focused on hypothesis testing on samples of graphs generated under a fixed graph generating process. In the current section we generalize it further to a hypothesis testing on two samples of graphs of different generating models. This shows that we can differentiate between graphs that were created through different models, as they have different rules that created them.

Suppose we have the cumulative distribution function (CDF) $F_a$ for a distribution of path differences between two random graphs generated under the model $a$, and the cumulative distribution function $F_b$ for a distribution of path differences between two random graphs generated under the model $b$. That is, $F_a(x) = P(d(G_1, G_2) \leq x | G_1, G_2$ are generated under the model $a)$, and $F_b(x) = P(d(G_1, G_2) \leq x | G_1, G_2$ are generated under the model $b)$.

In this study we have the following hypotheses:

$$H_0 : F_a(x) = F_b(x) \forall x \geq 0$$
$$H_1 : F_a(x) \neq F_b(x) \text{ for some } x \geq 0.$$  

A hypothesis test with the hypotheses in (5) can be used for “goodness-of-fit” test model selection.

In order to conduct the hypothesis test on two distributions of path differences between two graphs, we have applied a non-parametric method Kolmogorov-Smirnov (KS) test [18], [19]. Let $F_{a,N}$ be the empirical CDF for model $a$ for a sample with size $N$, and $F_{b,N'}$ be the empirical CDF for model $b$ for a sample with size $N'$. Then the test statistics for our test becomes

$$D_{N,N'} = \sup_x |F_{a,N}(x) - F_{b,N'}(x)|.$$
Input: The observed two set of arbitrary graphs $S_1 = \{G_1^1, G_2^1, \ldots, G_N^1\}$ and $S_2 = \{G_1^2, G_2^2, \ldots, G_N^2\}$. The Type I error rate (significance level) $\alpha$.

Output: Decision whether we reject or not reject the null hypothesis in (5).

1) Compute the test statistics $D_{N,N'}$ from the observation.
2) If $D_{N,N'} > c(\alpha) \sqrt{(N + N')/(N \cdot N')}$, where $c(\alpha)$ can be found for each level of $\alpha$ in [19], then we reject the null hypothesis. If not then we do not reject the null hypothesis.

We conducted a simulation study for this significance test on two samples. In this experiment we would like to see if the number of nodes in the graphs affect the type I error rate (i.e., the false positive rate) and the type II error rate (i.e., the false negative rate) of the statistical test under each model to generate random graphs. Under each model, we varied the number of nodes $n = (5 \cdot i) + 10$ for $i = 1, \ldots, 100$. Then, we generated 100 graphs with fixed number of nodes. We generated two such samples under each model with a fixed number of nodes.

First, we conducted tests for Type I errors. We compared for the number of nodes $n = (5 \cdot i) + 10$ for $i = 1, \ldots, 100$:

1) Two samples of size 100 generated under Erdős-Rényi model (see Figure 2(a)).
2) Two samples of size 100 generated Barabási-Albert model (see Figure 2(b)).
3) Two samples of size 100 generated under Watts-Strogatz model (see Figure 2(c)).

4) Two samples of size 100 generated under PowerLaw-Cluster model (see Figure 2(d)).
5) Two samples of size 100 generated under Community Structured model (see Figure 2(e)).

Type I error with the significance level $\alpha = 0.05$ in this study can be found at Table II.

<table>
<thead>
<tr>
<th>model</th>
<th>Type I error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erdős-Rényi model</td>
<td>0.09</td>
</tr>
<tr>
<td>Barabási-Albert model</td>
<td>0.04</td>
</tr>
<tr>
<td>Watts-Strogatz model</td>
<td>0.05</td>
</tr>
<tr>
<td>PowerLaw-Cluster model</td>
<td>0.06</td>
</tr>
<tr>
<td>Community Structured model</td>
<td>0.06</td>
</tr>
</tbody>
</table>

TABLE II: Type I error with the significance level $\alpha = 0.05$.

Next, we conduct tests for Type II errors. We compare One sample of size 100 generated under one model and one sample of size 100 generated under a different model for the number of nodes $n = (5 \cdot i) + 10$ for $i = 1, \ldots, 100$. The results are shown in Figure 3.

Results show that the proposed hypothesis can clearly differentiate two networks generated using two different generating processes as shown in Figure 3(a-c,e,g-j). In Figure 3(d,f), p-values are higher, and the proposed hypothesis is not able to differentiate the models due to the following reasons. In Community Structured Model, the edges are placed within and between the communities uniformly at random with the given probability $p_{in}$ and $p_{out}$ respectively, as discussed in Section II-E. Thus the generating fashion of Community Structured model is very similar to the Erdős-Rényi model, and therefore the similarity value is very high. The highest p-valued for Community structured and Erdős-Rényi model is 0.443 and shows that the p-value captures the similarities between the models. Similarly, PowerLaw-Cluster model generates the network using preferential attachment model, and a triad edge is placed between a node and a neighbor
of its neighbors with probability $p$. In the experiments, we have taken a small value for $p$, $p = 0.1$. Thus the generated PowerLaw-Cluster networks are generated in a very similar fashion as Barabási-Albert networks for the small size networks, and the similarity is captured by p-values as shown in Figure 3(f). We further observe that as the network size increases, the p-values decreases and shows the differences between the generating models.

We also observe the biggest p-value for each pair of models and the results are shown in Table III.

<table>
<thead>
<tr>
<th>Models</th>
<th>Biggest p-value</th>
<th>Models</th>
<th>Biggest p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ER vs. BA</td>
<td>0.069</td>
<td>ER vs. WS</td>
<td>5.956e-06</td>
</tr>
<tr>
<td>ER vs. PC</td>
<td>0.443</td>
<td>ER vs. CS</td>
<td>0.443</td>
</tr>
<tr>
<td>BA vs. WS</td>
<td>9.122e-09</td>
<td>BA vs. PC</td>
<td>0.894</td>
</tr>
<tr>
<td>BA vs. CS</td>
<td>1.220e-05</td>
<td>WS vs. PC</td>
<td>3.695e-09</td>
</tr>
<tr>
<td>WS vs. CS</td>
<td>2.849e-06</td>
<td>PC vs. CS</td>
<td>0.001</td>
</tr>
</tbody>
</table>

TABLE III: Biggest p-values for different models for type II errors.

Our experiment shows that the hypothesis test which we proposed controls the type I error rate under the significance level (Table II and Figure 2), and it seems that the type I error rates is not affected by the number of nodes. In addition, the analysis shows that our hypothesis test has fairly high power and it does not depend on the number of nodes (Figure 3).

V. CONCLUSIONS AND FURTHER STUDIES

In the current research, we introduced the methodology of identifying similar graphs using the path difference distance. We analyzed it theoretically by looking at the mean and standard deviation for uniformly generated Erdősf-Rényi random graphs with $n$ nodes. We test the proposed hypotheses on Erdősf-Rényi and four more generating models. We created multiple copies of the same size graph using all different methods of obtaining synthetic models. First, we tested our hypothesis to capture the similarity between the networks generated using the same generating model. Our type I error analysis showed that the hypothesis test in Section IV-A and the hypothesis test in Section IV-B control the type I error rates under the significance level. Next, we tested our hypothesis of differentiating graphs that were created by different synthetic models, and our power analysis shows that the hypothesis test in Section IV-B has high power.

There has been much work to classify an observed network using supervised learning methods, such as classification trees and random forests, using graph statistics, such as mean distances and average degree (for example, [20] and [21]). It is not well understood how such statistics contribute classification accuracy. In the classification process, we learn behaviors of such graph statistics by generating random networks and then apply what we learned from random networks to a given data set. Chia suggests that the average distance contributes significantly in classifying a network, yet it is important to understand how a distance $d(G_1, G_2)$ between two networks, $G_1$, $G_2$, varies.

This project opens up various future directions. In this work, we compare graphs that we control in size, but having different generating processes. In a temporal view, the sizes
of real-world networks increase very fast, and thus researchers analyze samples of the ground true networks. It will be beneficial to identify what collected samples capture the network generating information.

ACKNOWLEDGMENTS

The authors thank the DoD for partially sponsoring the current research. R.Y. is supported by NSF Division of Mathematical Sciences: CDS&E-MSS program 1622369. Authors would also like to thank Akrati Saxena, IIT Ropar, India for the discussions and code implementations for the project.

REFERENCES