

Deep Learning based Network Similarity for Model Selection

Kushal veer singh ^{a,*}, Ajay kumar verma ^a and Lovekesh vig ^b

^a School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, Delhi, India

E-mail: kushalveer@gmail.com

^b TCS Innovation Labs, Tata Research Development and Design Centre, New Delhi, Delhi, India

E-mails: ajayverma81@gmail.com, lovekeshvigin@gmail.com

Abstract.

Capturing data in the form of network's is becoming an increasingly popular approach for modeling, analyzing and visualising complex phenomena, to understand the important properties of the underlying complex processes. Access to many large-scale network datasets is restricted due to the privacy and security concerns. Also for several applications (such as functional connectivity networks), generating large scale real data is expensive. For these reasons, there is a growing need for advanced mathematical and statistical models (also called generative models) that can account for the structure of these large-scale networks, without having to materialize them in the real world. The objective is to provide a comprehensible description of the network properties and to be able to infer previously unobserved properties. Various models have been developed by researchers, which generate synthetic networks that adhere to the structural properties of real networks. However, the selection of the appropriate generative model for a given real-world network remains an important challenge.

In this paper, we investigate this problem and provide a novel technique (named as TripletFit) for model selection (or network classification) and estimation of structural similarities of the complex networks. The goal of network model selection is to select a generative model that is able to generate a structurally similar synthetic network for a given real-world (target) network. We consider six outstanding generative models as the candidate models. The existing model selection methods mostly suffer from sensitivity to network perturbations, dependency on the size of the networks, and low accuracy. To overcome these limitations, we considered a broad array of network features, with the aim of representing different structural aspects of the network and employed deep learning techniques such as deep triplet network architecture and simple feed-forward network for model selection and estimation of structural similarities of the complex networks. Our proposed method, outperforms existing methods with respect to accuracy, noise-tolerance, and size independence on a number of gold standard data set used in previous studies.

Keywords: Complex Networks, Deep Learning, Generative Models, Model Selection

1. Introduction

Datasets emerging from different fields such as biology, neuroscience, engineering, social science, economics, etc. are often represented as networks to understand the complex systems in these fields.

To understand the formation and evolution of real-world networks various generative models have been proposed to generate synthetic networks that follow the non-trivial topological properties of real-world networks [1–6]. For example, Watts-Strogatz model [4] synthesizes small-world networks with small average path length and high clustering coefficient, and Barabási - Albert model [1] generate scale-free networks with long-tail (power law) degree distribution. In addition to degree distribution, clustering

*Corresponding author. E-mail: kushalveer@gmail.com.

and path lengths, other structural properties such as modularity, assortativity and special eigenvalues - are also supported in newer generative models [3, 7].

Despite the progress made in proposing many generative models, there is currently no universal generative model that is applicable for all applications. Therefore, prior to network generation, we have to perform the non-trivial task of choosing the appropriate generative model for a particular application (also called model selection). Since we want to choose the model that is most representative of the real network, model selection involves deep analysis of the properties of the given network (called target network), and accordingly the most appropriate model is chosen. Essentially, model selection attempts to evaluate a library of candidate generative models and predict which the most appropriate for generating complex network instances similar to the real network. There are many applications of model selection including network sampling [8–11], simulation of network dynamics [12–14] and summarization [15–17] etc.

In order to perform an effective model selection, we require a similarity measure to compare networks across their topological properties such as average path length, transitivity, clustering coefficient, modularity, etc.. Lots of literature discussed the importance of structural similarity metric for complex networks, an appropriate definition of distance similarity metric is the basis for many machine learning and data analysis tasks such as classification and clustering [15, 18–21]. An important property of the chosen distance metric is that it should be agnostic to network size so that it can compare networks of different scales. This is a departure from other similarity/dissimilarity notions including graph similarity with known node correspondence [21], classical graph similarity approaches (including graph alignment, graph matching, graph isomorphism) [22, 23]. One related approach to developing a network similarity metric is to create a feature vector for each network based on existing network topological properties and then computing the similarity of feature vectors based on Euclidean distance in feature space [15, 18, 24]. In this paper, we propose a novel method for automatic learning of the similarity metric via a specialized deep neural architecture. The model learns via supervised training wherein it learns from pairs of similar and dissimilar networks and maps the features onto space where distances between similar networks are smaller than distances between dissimilar networks.

2. Related work

In the literature, several network model selection (or network classification) methods are available most of them are based on graphlet counting feature [19, 25, 26], and combination of local and global features of network topology [17, 27, 28] for selecting the best generative model. Other methods are also developed for model selection problem [29, 30].

To measure the structural similarities between two networks various quantitative measures have been reported [15, 18, 20, 23, 24, 26, 28, 31, 32]. Graph isomorphism is one of the classical approaches to compare two graphs. Two graphs are said to be isomorphic if they have an identical topology. Some variants of this approach are also proposed, including subgraph isomorphism and maximum common subgraphs [33]. Several isomorphism-inspired methods based on counting the number of spanning trees [34] and computing the node-edge similarity scores are also available [23]. These different methods are computationally expensive and not suitable for the large complex networks. Various approaches utilized graphlet counts as a measure of network similarity [19, 26, 35] and distance measures for network comparison in which network are represented in the form of feature vectors that summarize the network topology [3, 15, 17, 33, 36].

1 3. Materials and Methods

2 In this section we describe our model, formerly known as 'TripleFit', for complex networks. The model
 3 consists of two important processes : (a) model selection and (b) estimate network structural similarity.
 4 The model selection is based on the structural similarity between two complex networks. For a given
 5 target (realistic) network instance, the proposed method chooses the appropriate generative model among
 6 six generative models that can generate a synthetic network similar to the target (realistic) network. The
 7 selection of the best model is based on the embedded feature space of the target network and various
 8 synthetic networks generated from different generative models. In the proposed method, we developed
 9 a network distance metric by utilizing topological features for separating the various types of network
 10 classes. The simplest choice would be to compute the Euclidean distance between the two topological
 11 feature vectors of two networks. In this way, we utilize the triplet neural network architecture to learn
 12 the best distance metric [43]; the goal will be to learn a transformation from the original feature space
 13 to an embedded feature space, in which the euclidean distance between similar networks is smaller than
 14 distances between dissimilar networks. We also quantify the structural similarity between two networks
 15 by computing the Euclidean distance between the corresponding network topological feature vectors in
 16 transformed space.

18 3.1. Dataset description:

19 In this study, we have taken a gold standard data set ¹ named as Reference Networks Datasets (RND)
 20 used in the previous study [28] that describes a Noise-tolerant model selection method for complex
 21 networks. For evaluating the network model selection they consider six network generative models:
 22 Kronecker Graphs (KG) [3], Forest Fire model (FF) [2], Barabási-Albert model (BA) [1], Watts-
 23 Strogatz model (WS) [4], Erdős-Rényi (ER) [53] and Random Power Law (RP) [54]. The detailed
 24 description of the dataset is given in the referred study [28].

27 3.2. Overall model selection strategy:

29 Figure 1 shows a high-level view of the model selection process. The methodology is configurable by
 30 several decision points, such as the set of considered network features, the candidate generative models.
 31 The steps for constructing the final network classifier are described here:

- 33 (1) We took 1000 network instances of different size and densities (using different parameters) for
 34 each candidate network generative model from RND, for capturing their growth mechanism and
 35 generation process. These network instances will form the dataset for learning the triplet neural
 36 network.
- 37 (2) We extract the topological features (clustering coefficient, modularity, degree distribution, etc.) of
 38 each network instance. The result is a dataset of labeled structural features in which each record
 39 consists of topological features of a synthesized network along with the label of its generative
 40 model.
- 41 (3) Construct the triplets a from the labeled dataset which comprises of the positive, negative and
 42 anchor samples, where the positive and anchor sample is of the same class (or same generative
 43 model), and the negative sample is of a different class. These triplets are utilized for learning the

45 ¹Downloaded from <http://ce.sharif.edu/~aliakbary/datasets.html> on May 5, 2016.

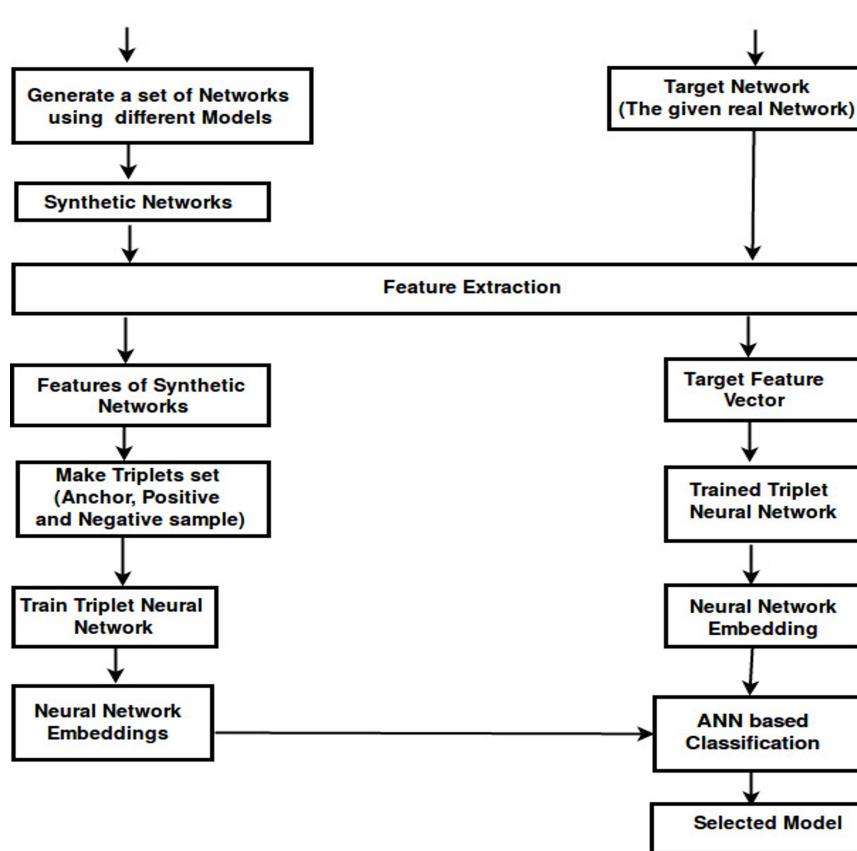


Fig. 1. The proposed methodology for model selection.

distance metric using a triplet network. So, we train the triplet neural network using the different triplets. This trained triplet neural network transforms the feature space into another feature space (called embedding feature space) in which the distance between the positive and anchor sample is smaller than the distance between negative and anchor sample. This new embedding feature space will form the dataset for learning a network classifier.

- (4) The labeled embedding feature dataset forms the training and test data for a supervised learning algorithm. The learning algorithm will return a network classifier which can predict the class (the best generative model) for a given network instance.
- (5) The same topological features used in Step 2 are also extracted from the real world (target) network. Then pass this feature vector into trained triplet neural network, which we trained in Step 3 and generate the embedding feature vector of the target network. This embedding feature vector of the target network is used as the input of the learned classifier, which is trained in Step 4.

The learned network classifier is a customized model selector for finding the model that fits the target network. It gets the topological features of the target network as the input and returns the most compatible generative model.

Besides the network model selection, proposed method also estimate the structural similarities of networks. For measuring the structural similarity between two networks, we extract the topological

1 features of two given networks as in Step 2. Then we pass both feature vectors into trained triplet neural
 2 network, trained in Step 3, and generate embedding feature vectors for both networks. Then we compute
 3 the Euclidean distance between the two embedding feature vectors, representing the similarity between
 4 two networks. The distances between similar networks are smaller compared to dissimilar networks.
 5 Section 4.3 describes the computation of structural similarities between the network instances generated
 6 in Step 1. Detailed description of the steps described in the above section are presented below:
 7

8 3.3. Network Features

9 The process of model selection, as described in Figure 1, utilizes network topological features in
 10 the second and fifth steps. There are various features are defined in network literature to quantify the
 11 topological properties of the network. We considered only some well-known and frequently studied
 12 measures, which are relevant to our study. A diverse set of local and global network features were
 13 utilized to construct feature vectors. A brief description of the calculated measures is as follows:
 14

- 15 • **Degree distribution:** Degree distribution defined as the probability distribution of the degrees of
 16 all nodes of the network. We quantify a degree distribution by computing its variance (V) [42], gini
 17 coefficient (G) [42]. The Variance and gini coefficient of degree distribution are defined as follows:
 18

$$20 V = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^N (k_i - \mu)^2}}{\mu} \quad (1)$$

$$25 G = \frac{1}{\mu} * \frac{1}{2N^2} \sum_{i=1}^N \sum_{j=1}^N |k_i - k_j| \quad (2)$$

28 Where k_i and k_j are the degrees of the nodes. N is the number of nodes in the network and μ is the
 29 mean of the degree.
 30

- 31 • **Entropy:** The entropy of the degree distribution provides an average measurement of the hetero-
 32 geneity which in turn determines the robustness of the network. Formally, the entropy is defined as
 33 [41]:

$$34 E = - \sum_k p(k) \log(p(k)) \quad (3)$$

37 Where $p(k)$ refer the probability of fraction of nodes having degree k .
 38

- 39 • **Clustering coefficient:** Clustering coefficient is a measure of segregation in network analysis. Clus-
 40 tering coefficient $C(v)$ of a node v is defined as the number of links that exist between the direct
 41 neighbours of that node divided by the maximum number of possible links. Formally, given a node
 42 v with N_v neighbors and l_v links within the neighbors, the clustering coefficient $C(v)$ is defined as
 43 follows:
 44

$$45 C(v) = \frac{2 \times l_v}{N_v \cdot (N_v - 1)} \quad (4)$$

The clustering coefficient C for the whole network is calculated by the average value of $C(v)$ over all v [4].

- **Characteristic path length:** The mean of the shortest path length between all pairs of nodes also called the characteristic path length [4] and is a measure of network integration. For a graph G with n nodes, the network characteristic path length is given by:

$$L = \frac{1}{n.(n-1)} \sum_{u,v \in G: u \neq v} d(u, v) \quad (5)$$

where $d(u, v)$ is the shortest path length between vertices u and v in G .

- **Efficiency:** Global efficiency is a measure of integration, that is the calculated by average inverse shortest path length between all pairs of nodes in the network. For a graph G with n nodes and k edges, the global efficiency of a network is defined as [37, 38]:

$$E_g = \frac{1}{n.(n-1)} \sum_{u,v \in G: u \neq v} \frac{1}{d(u, v)} \quad (6)$$

where $d(u, v)$ is the shortest path length between vertices u and v in G .

- **Assortativity coefficient:** Assortativity is a measure of how well nodes of similar degree are linked to one another in the network. A network is said to show assortative mixing if high degree nodes have a high tendency to connect to other high degree nodes and similarly low degree nodes have a bias towards connecting to low degree nodes. To quantify the level of assortative mixing in a network we define an assortativity coefficient which is defined as [39]:

$$r = \frac{l^{-1} \sum_{e \in E} d_e d'_e - [\frac{l-1}{2} \sum_{e \in E} (d_e + d'_e)]^2}{\frac{l-1}{2} \sum_{e \in E} (d_e^2 + d'^2_e) - [\frac{l-1}{2} \sum_{e \in E} (d_e + d'_e)]^2} \quad (7)$$

Where d_e and d'_e are the degrees of the two nodes at either end of the e^{th} edge in the network. l is the total number of links, E is the set of all links in the network.

- **Modularity:** Modularity has been widely used as a quality measure for measure the strength of division of a network into modules (also called cluster or communities). Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules. The modularity of an unweighted graph partitioned into modules is evaluated by [40]:

$$Q = \frac{1}{2.m} \sum_{u,v} [A_{u,v} - \frac{k_u k_v}{2.m}] \delta(c_u, c_v) \quad (8)$$

Where k_u and k_v are the degrees of u and v , m is the number of links of the network, c_u the community of node u and $\delta(c_u, c_v) = 1$ if $c_u = c_v$ and $\delta(c_u, c_v) = 0$ otherwise. Thus the problem of discovering the modules of a network reduces to optimizing modularity.

Network features are not standardized i.e. there is no universal best set of features for networks and other features such as shrinking diameter, densification, vulnerability, network resilience, rich-club phenomenon, etc. have also been used. The proposed methodology is not limited to the specified network feature set. Hence, one can also utilize another set of features, according to their application domain. In this study, we utilized only 8 network topological features.

3.4. Triplet Neural Network

Triplet neural network is a deep learning models, which aims to learn useful representations of data by distance comparisons [43]. Recently, triplet network architecture have successfully applied in many computer vision tasks [44–47]. In past few years, deep learning models have been widely exploited to solve various machine learning tasks. Deep Learning algorithm is automating the extraction of high-level meaningful complex abstractions as data representations (features) through the use of a hierarchical learning approach. The notion of hierarchical features stems from neuroscientific discoveries of the visual cortex that indicate a hierarchy of cells with successively higher level cells firing for more complex visual patterns [48–52].

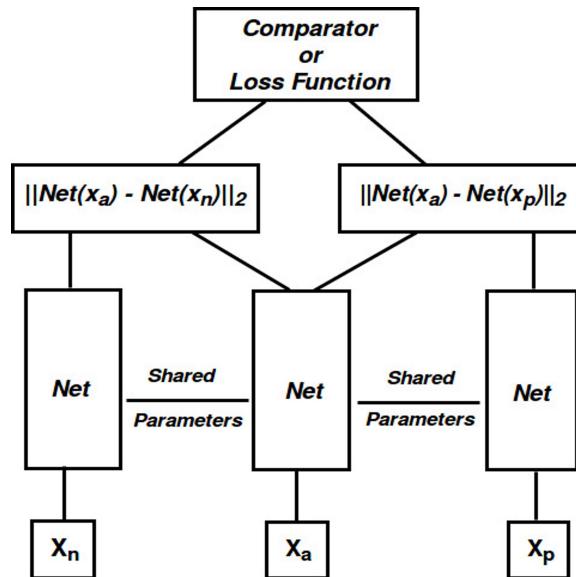


Fig. 2. A schematic representation of Triplet Neural Network architecture, which consists three feed forward neural networks of the same instance with shared parameters.

Triplet network, as shown in the Figure 2, is a model inspired by the Siamese Network which comprise three feed forward networks of the same instance with shared parameters. The network gives two intermediate values when fed three samples to it. These intermediate values come from L_2 distances between the embedded representation of two of its inputs from the representation of third. Symbolically, if we denote these inputs as X_a (anchor), X_p (positive), and X_n (negative) and the embedded representation of network as $Net(X)$ then the triplet score just before the last layer computed as:

$$TripletNet(X_p, X_a, X_n) = \left\{ \begin{array}{l} \left\| Net(X_a) - Net(X_p) \right\|_2 \\ \left\| Net(X_a) - Net(X_n) \right\|_2 \end{array} \right\} \in R^2_+ \quad (9)$$

We train this network for a 2-class classification problem using the triplet input (X_p, X_a, X_n) where (X_a, X_p) are chosen from same class and (X_a, X_n) from the different class. The training process encourages the network to find an embedding where the distance between X_a and X_n is larger than the distance between X_a and X_p (*i.e.*, minimizes the distance between a pair of examples with the same class label) plus some margin. We compared the SoftMax output of this TripletNet vector with vector $(1, 0)$. Softmax function for vector $X = [x_1, x_2, x_3, \dots, x_n]$ is defined as:

$$SoftMax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}} \quad (10)$$

The loss function is defined as:

$$Loss(d_+, d_-) = \|d_+ - d_- - 1\|_2^2 \quad (11)$$

Where

$$d_+ = \frac{e^{\|Net(X_a) - Net(X_p)\|_2}}{e^{\|Net(X_a) - Net(X_p)\|_2} + e^{\|Net(X_a) - Net(X_n)\|_2}} \quad (12)$$

and

$$d_- = \frac{e^{\|Net(X_a) - Net(X_n)\|_2}}{e^{\|Net(X_a) - Net(X_p)\|_2} + e^{\|Net(X_a) - Net(X_n)\|_2}} \quad (13)$$

We note that, $Loss(d_+, d_-) \rightarrow 0$, if and only if

$$\frac{\|Net(X_a) - Net(X_p)\|_2}{\|Net(X_a) - Net(X_n)\|_2} \rightarrow 0, \quad (14)$$

which is the required objective. We utilized the back-propagation algorithm to update the model on all three samples (X_a, X_p and X_n) simultaneously, using the same shared parameters.

3.5. Triplet Neural Network Architecture

Here, we briefly describe the implementation details of Triplet Neural Network. A Triplet Neural Network consists of three feedforward neural networks of shared weights, followed by two layers. An advanced activation function ELU (Exponential Linear Unit) are applied between two consecutive layers. Network configuration (ordered from input to output) consists of layers dimensions $\{8, 32, 16\}$ where a 16 the final embedded representation of the feature vector on which the distance has been measured.

3.6. Network Models

Among several existing network generative models, we have selected six important models: Kronecker Graphs (KG) [3], Forest Fire model (FF) [2], Barabási-Albert model (BA) [1], Watts-Strogatz model (WS) [4], Erdős-Rényi (ER) [53] and Random Power Law (RP) [54]. The selected models are the state of the art methods of network generation. Existing model selection methods have ignored some new and important generative models such as Kronecker graphs and Forest Fire [28]. All the models are briefly described in the study [28]. The characteristic of above six generative models are defined as follows:

- **Erdős-Rényi (ER):** This model generates completely random networks. The number of nodes and edges are configurable [53].
- **Barabási-Albert model (BA):** This model generates random scale-free networks using a preferential attachment mechanism [1].
- **Watts-Strogatz model (WS):** This model generates synthetic networks with small characteristic path lengths and high clustering coefficient. It starts with a regular lattice and then rewrites some edges of the network randomly [4].
- **Forest Fire model (FF):** In this model, edges are added in a process similar to a fire-spreading process. This model is inspired by Copying model and Community Guided Attachment but supports the shrinking diameter property [2].
- **Random Power Law Model (RP):** The RP model generates synthetic networks by following a variation of ER model that supports the power law degree distribution property [54].
- **Kronecker Graphs (KG):** This model generates realistic synthetic networks by applying a non standard matrix operation (the kronecker product) on a small initiator matrix. This model is mathematically tractable and supports many network features, such as heavy tail degree distribution, small diameters, heavy tails for eigenvalues and eigenvectors, and densification and shrinking diameters over time [3].

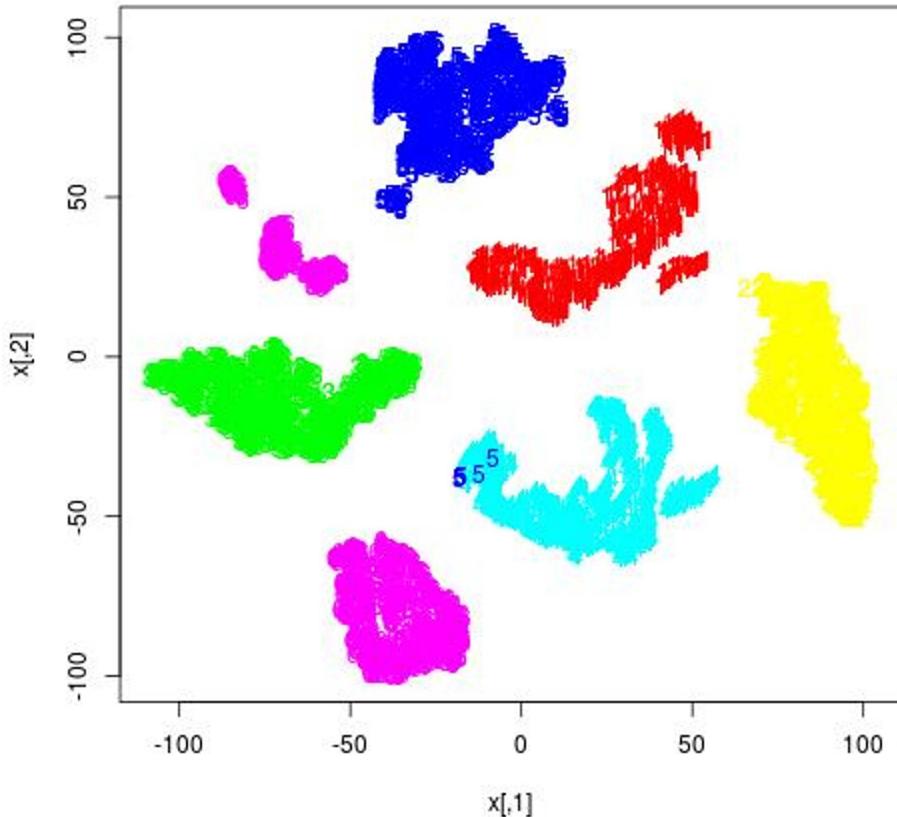
4. Results

In this section, we evaluate our proposed method for network model selection and network structural similarity. We utilize the RND to evaluate our model against other existing methods. The proposed model is evaluated by first transforming the networks dataset into the feature dataset using the topological features mentioned in Section 3.3. Thus we have several network instances generated using six generative models and many real world networks that correspond to one feature vector representation in feature dataset. In previous methods size (number of nodes) and/or density of a target network is considered in the generation of training data [19, 25, 27]. In our methodology, the size and density of the target network are not considered in the generation of the training data. In the proposed method, a Triplet Neural Network is utilized to find the best network distance metric, which is capable of separating networks of different classes.

4.1. Evaluation of network distance metric

TripletFit method is based on learning of network distance metric. The distance metric learning problem is concerned with learning a distance function, which can separate networks of different generative models. In this study, we choose euclidean distance (L_2 norm) as a distance function. We utilized Triplet

1 Neural Network [43] to learn the distance function. We train the triplet neural network on the feature
 2 dataset generated in Step 2 of Section 3.2. The trained triplet neural network transforms the feature
 3 dataset into embedded feature dataset, where each generative model is clearly separate (or clustered) in
 4 euclidean space (or embedded space).



34 Fig. 3. **Visualization of embedded feature set into two dimensional space using t-SNE method. The colors correspond to**
 35 **each individual generative model viz. yellow for ER, red for BA, blue for RP, cyan for KG, green for FF, pink for WS**
 36 **respectively.**

37 In this section, we show that learned distance function is capable of separating different class
 38 generative models. In this order, we visualize the high-dimensional embedded feature dataset using *T*-
 39 *Distributed Stochastic Neighbor Embedding* (t-SNE) technique, which projects the high-dimensional
 40 embedding data into low dimensional embedding. t-SNE is a non-linear dimensionality reduction tech-
 41 nique, allowing visualize of high-dimensional data. It learned the low dimensional embedding by min-
 42 imizing the Kullback-Leibler divergence between the two distributions: a distribution that measures
 43 pairwise similarities of the input data points and a distribution that measures pairwise similarities of
 44 the corresponding low-dimensional points in the embedding [55, 56]. t-SNE representation of high-
 45

dimensional embedded features dataset into two-dimensional space is given in Figure 3, where different colors are shown, different generative models.

As described in Figure 3, each class of generative model is clearly separate in embedding feature space. Hence, we utilized the embedded feature dataset for both model selection (network classification) and network similarities (network comparison).

4.2. Evaluation of the model selection approach

First, we evaluate the TripletFit for model selection, on the synthetic labeled dataset, which is constructed by the network instances of known generative models mentioned in Section 3.6. To the best of our knowledge, these are the most widely used generative models in network generation applications, and they also cover a wide range of network structures.

Each generative model offers a set of parameters for tuning the synthesized networks so that they follow the properties of real (target) networks. The Number of nodes (or network size) is an important parameter of considered models. Unlike other network model selection methods [17, 19, 25–27], we select the all the parameters randomly. The size of the network is randomly chosen from 1000 to 5000 nodes, and other network parameters are also chosen randomly from the parameter ranges described in study [28]. As compared with the size of real (target) network, we generate smaller sizes of the network instances for training the proposed method. This feature increases the efficiency and performance of our proposed method. For each generative model, 1000 of network instances are generated using different parameters.

The generative model selection treated as a network classification problem. In this study, we developed a supervised learning algorithm for network model selection. In this way, we utilized an Artificial Neural Network (ANN) model [57, 58] as a classifier to select an appropriate generative model for a given real network. For training the ANN model, we utilized embedded features dataset generated in Step 3 of Section 3.2. We performed 10-fold cross-validation process for evaluation of our proposed method, where the whole dataset is randomly divided into 10 equal subsets. From the 10 subsets, a single subset is retained as a test set, and the remaining subsets are used as a training data. This process repeated 10 times. We construct the test set in such a way that, in each iteration it contained an equal number of networks instances (*i.e.*, 100) of each generative model. The final accuracy of proposed method is calculated by mean of accuracies of each iteration. The detailed average results (average outcome of 10-folds) of the classifier is given in Table 1.

Table 1
Results of TripletFit method on synthetic network dataset.

| True Pred \ Pred | BA | ER | FF | KG | RP | SW | Class Precision |
|---------------------|------|------|------|------|-----|------|------------------------------------|
| BA | 100 | 0 | 0 | 0 | 0 | 0 | 100% |
| ER | 0 | 100 | 0 | 0 | 0 | 0 | 100% |
| FF | 0 | 0 | 100 | 0 | 0 | 0 | 100% |
| KG | 0 | 0 | 0 | 100 | 2 | 0 | 98.04% |
| RP | 0 | 0 | 0 | 0 | 98 | 0 | 100% |
| SW | 0 | 0 | 0 | 0 | 0 | 100 | 100% |
| Class Recall | 100% | 100% | 100% | 100% | 98% | 100% | Accuracy: 99.70 % |

Aliakbary et. al. [28] compared their proposed method (ModelFit) with various existing model selection methods: GMSCN [27], SVMFit [57, 59], RGF [24], AvgBased, Naïve-Manhattan. GMSCN utilized LADTree decision tree classifier on RND. SVMFit is an SVM-based classification method. AvgBased is a distance based classifier which considers an average distance of the given network with neighbor networks. RGF-method is another distance based method, utilizes the concept of the graphlet count features. Finally, the Naïve-Manhattan distance can be defined as pure Manhattan distance of the network features, where all the network features shares the equal weights during distance calculation. More details about these methods can be found through the research paper [28]. We compared our proposed method with other methods reported in the work of Aliakbary et. al. [28], a comparative summary of results is shown in Figure 4.

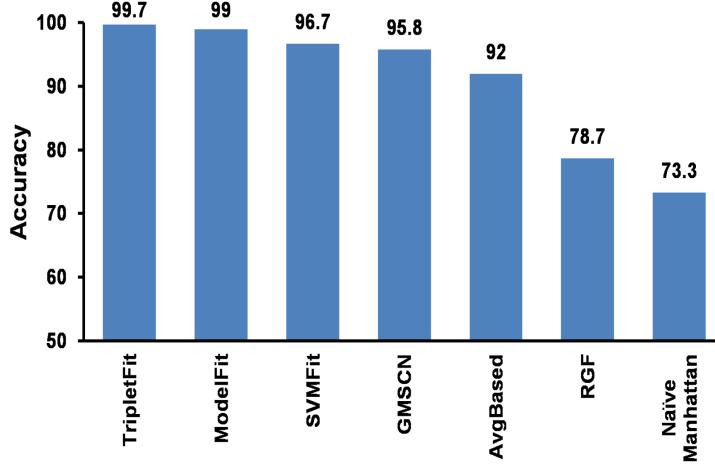


Fig. 4. The accuracy of different network selection methods.

We also evaluate the robustness of our proposed method by introduce varying noise levels (noise = 5%, 10%, 15%, 20%, and 25%) into the test-case network. The noise was introduced in the network by rewiring a particular fraction of network edges between the randomly selected pair of nodes. For example, to introduce 5% of the noise level in the network, five percent of the network edges were rewired between the randomly chosen pair of nodes. Figure 5 shows the the average accuracy of the proposed method for different noise levels. We observed that upto 10% noise level our model outperforms the rest. Additional noise beyond 10% results in some degradation in performance compared to ModelFit but our model still remains robust compared to other models published in [28].

4.3. Evaluation of the network similarity approach

In this section, we measure the structural similarities between different synthetic and real world complex networks. In order to measure the network structural similarity between two networks, we compute

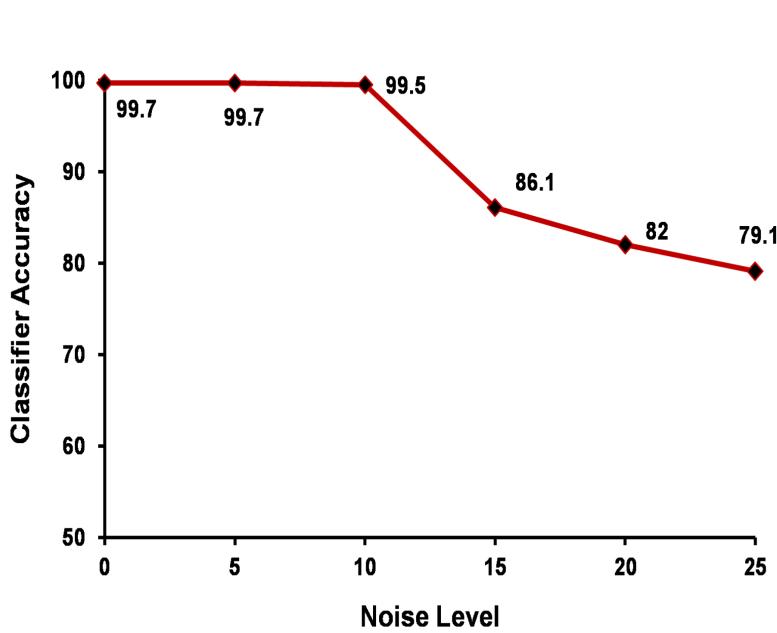


Fig. 5. The robustness of TripleFit with respect to different levels of noise.

the euclidean distance between the embedded feature vectors of corresponding networks. Let p and q be the two embedded feature vectors then euclidean distance between these two features are given by:

$$d_E(p, q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \quad (15)$$

We employ the Euclidean distance as a network structural similarity measure for comparison of real networks. To measures the structural similarity between two networks, we computes the euclidean distance between the embedded feature vectors of these two networks. The question is, Is the euclidean distance capable of describing the structural similarities in real network data? In order to evaluate this, we compute the euclidean distance between every pair of embedded feature vectors of 6000 network instances, generated by six different generative models as described in in Step 4 of Section 3.2. Then we plot a Heatmap of this 6000×6000 distance matrix. Figure 6 shows the Heatmap of pairwise distances between the embedded fute vectors of all network instances of six diffrent generative models. Figure 6 shows that euclidean distance measurement are consistent with the expectation: euclidean distance calculates the structural similarity of two networks in a way that networks instances of the same class (or same generative model) type are considered more similar (small distance) to each other than to networks of different class.

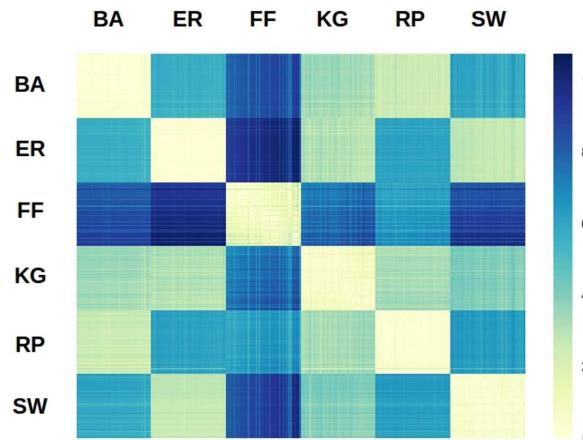


Fig. 6. Structural similarities between synthetic networks.

4.4. Case Study

In order to find the effectiveness of proposed method for real networks, we have applied TripleFit on some real networks. The real network instances and the results of TripleFit on these networks are described below:

- (1) “p2p-Gnutella08”, Gnutella is a small peer to peer file sharing network with about 6,301 nodes from August 8 2002. In Gnutella, network nodes represent the host and edges describe the connection between the host.
- (2) “Email-URV”, Email-URV is the network of emails interchanges between members of the University Rovira i Virgili of Tarragona, Spain. This data set is collected by Guimera *et al.* [60]. The data covers total 1,133 members including faculty, student and technicians *etc.*
- (3) “email-Enron”, email-Enron is the communication network covers all the email communication of Federal Energy Regulatory Commission during its investigation. In this network nodes represent the email addresses and an edge represents the email interchange between two address *i.e.*, an address sent at least one email to other address.
- (4) “cit-HepPh”, cit-HepPh is a arxiv High Energy Physics Phenomenology paper citation network. Which covers all the citations within a dataset of 34,546 papers with 421,578 edges. The data covers papers in the period from January 1993 to April 2003.
- (5) “cit-HepTh” cit-HepTh is a arxiv High Energy Physics Theory paper citation network. Which covers all the citations within a dataset of 27,770 papers with 352,807 edges. The data covers papers in the period from January 1993 to April 2003.

- (6) “ca-HepPh” ca-HepPh is a arxiv High Energy Physics Phenomenology collaboration network (or co-authorship network). Which covers scientific collaborations between authors papers submitted to High Energy Physics - Phenomenology category within a dataset of 12,008 authors with 118,521 edges. The data covers papers in the period from January 1993 to April 2003.
 - (7) “ca-HepTh” ca-HepPh is a arxiv High Energy Physics Theory collaboration network (or co-authorship network). Which covers scientific collaborations between authors papers submitted to High Energy Physics - Theory category within a dataset of 9,877 authors with 25,998 edges. The data covers papers in the period from January 1993 to April 2003.

Table 2 shows the results after applying Tripletfit on the real networks as described above. As Table 2 shows, various real-networks, which are selected from a wide range of sizes, densities and domains, are categorized in different network models by the classifier. This fact indicates that no generative model is dominated in proposed method for real networks and it suggests different models for different network structures. The case study also verifies that no generative model is sufficient for synthesizing networks similar to real networks and we should find the best model fitting the target network in each application. As a result, the task of generative model selection is an important stage before generating network instances.

Table 2
Real network samples and the selected generative models.

| Network | No. of Nodes | No. of Edges | Selected Model |
|---------------------|---------------------|---------------------|-----------------------|
| p2p-Gnutella08 [61] | 6301 | 20777 | SW |
| Email-URV [62] | 1133 | 10903 | FF |
| email-Enron [61] | 36692 | 367662 | SW |
| cit-HepPh [61] | 34546 | 421578 | BA |
| cit-HepTh [61] | 27770 | 352807 | ER |
| ca-HepPh [61] | 12008 | 237010 | SW |
| ca-HepTh [61] | 9877 | 51971 | BA |

5. Discussion

In this paper, we proposed a novel method for network model selection and network similarity. In this method, we investigated the network distance metric for comparing the topological properties of the complex networks. Despite most of the existing methods [19, 25, 27], the proposed distance based method is independent of the size and density of the input network. It is very difficult (sometimes not possible) to generate synthetic networks with exactly equal densities of input (real) network because some generative models are not configurable for finely tuning the exact density of synthesized networks. Equal size and/or density of the network in training data is undesirable because a good model is the one which tolerates variations in size and density of the networks. In our methodology, the size and density of the target network are not considered in the generation of the training data. Size and density independence is an important feature of our method. It enables the classifier to learn from a dataset of generated networks with different sizes and different densities, perhaps smaller from the size of the target network. For example, given a very large network instance as the target network, we can prepare the dataset of generated networks with smaller networks than the target network. This facility decreases the time of network generation and feature extraction considerably. In case size of our real network grows,

we can utilize a graph sampling techniques to reduce the size of the network for efficient computation of different network features [63–68]. The proposed method outperforms the various existing methods, which highlight the effectiveness of deep learning architectures in the learning of a distance metric for topological comparison of complex networks. Our proposed method (TripletFit), outperforms the state-of-the-art methods with respect to accuracy and noise tolerance.

Author Contributions: K.V.S. and L.V. conceived and designed the study. K.V.S. and A.K.V. implemented the algorithm and prepared the figures of the numerical results. K.V.S., A.K.V. and L.V. analyzed and interpreted the results, and wrote the manuscript. All the authors have read and approved the final manuscript.

Competing Interests: We declare that there is no competing of interests for this work.

Acknowledgments: The authors would like to thank J.N.U. and U.G.C., India for providing the research fellowship to K.V.S. and A.K.V. .

References

- [1] Barabási, Albert-László, and Réka Albert. "Emergence of scaling in random networks." *science* 286, no. 5439 (1999): 509-512.
- [2] Leskovec, Jure, Jon Kleinberg, and Christos Faloutsos. "Graphs over time: densification laws, shrinking diameters and possible explanations." In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 177-187. ACM, 2005.
- [3] Leskovec, Jure, Deepayan Chakrabarti, Jon Kleinberg, Christos Faloutsos, and Zoubin Ghahramani. "Kronecker graphs: An approach to modeling networks." *Journal of Machine Learning Research* 11, no. Feb (2010): 985-1042.
- [4] Watts, Duncan J., and Steven H. Strogatz. "Collective dynamics of small-world networks." *nature* 393, no. 6684 (1998): 440-442.
- [5] Chakrabarti, Deepayan, and Christos Faloutsos. "Graph mining: Laws, generators, and algorithms." *ACM computing surveys (CSUR)* 38, no. 1 (2006): 2.
- [6] Volchenkov, Dmitri, and Ph Blanchard. "An algorithm generating random graphs with power law degree distributions." *Physica A: Statistical Mechanics and its Applications* 315, no. 3 (2002): 677-690.
- [7] Akoglu, Leman, and Christos Faloutsos. "RTG: A recursive realistic graph generator using random typing." In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 13-28. Springer Berlin Heidelberg, 2009.
- [8] Stumpf, Michael PH, and Carsten Wiuf. "Sampling properties of random graphs: the degree distribution." *Physical Review E* 72, no. 3 (2005): 036118.
- [9] Lee, Sang Hoon, Pan-Jun Kim, and Hawoong Jeong. "Statistical properties of sampled networks." *Physical Review E* 73, no. 1 (2006): 016102.
- [10] Han, Jing-Dong J., Denis Dupuy, Nicolas Bertin, Michael E. Cusick, and Marc Vidal. "Effect of sampling on topology predictions of protein-protein interaction networks." *Nature biotechnology* 23, no. 7 (2005): 839-844.
- [11] Leskovec, Jure, and Christos Faloutsos. "Sampling from large graphs." In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 631-636. ACM, 2006.
- [12] Pastor-Satorras, Romualdo, and Alessandro Vespignani. "Epidemic dynamics in finite size scale-free networks." *Physical Review E* 65, no. 3 (2002): 035108.
- [13] Montanari, Andrea, and Amin Saberi. "The spread of innovations in social networks." *Proceedings of the National Academy of Sciences* 107, no. 47 (2010): 20196-20201.
- [14] Briesemeister, Linda, Patrick Lincoln, and Phillip Porras. "Epidemic profiles and defense of scale-free networks." In *Proceedings of the 2003 ACM workshop on Rapid malcode*, pp. 67-75. ACM, 2003.
- [15] Sala, Alessandra, Lili Cao, Christo Wilson, Robert Zablit, Haitao Zheng, and Ben Y. Zhao. "Measurement-calibrated graph models for social network experiments." In *Proceedings of the 19th international conference on World wide web*, pp. 861-870. ACM, 2010.
- [16] Mahadevan, Priya, Dmitri Krioukov, Kevin Fall, and Amin Vahdat. "Systematic topology analysis and generation using degree correlations." In *ACM SIGCOMM Computer Communication Review*, vol. 36, no. 4, pp. 135-146. ACM, 2006.

- [17] Airoldi, Edoardo M., Xue Bai, and Kathleen M. Carley. "Network sampling and classification: An investigation of network model representations." *Decision support systems* 51, no. 3 (2011): 506-518.
- [18] Berlingero, Michele, Danai Koutra, Tina Eliassi-Rad, and Christos Faloutsos. "Network similarity via multiple social theories." In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pp. 1439-1440. IEEE, 2013.
- [19] Janssen, Jeannette, Matt Hurshman, and Nauzer Kalyaniwalla. "Model selection for social networks using graphlets." *Internet Mathematics* 8, no. 4 (2012): 338-363.
- [20] Mehler, Alexander. "Structural similarities of complex networks: A computational model by example of wiki graphs." *Applied Artificial Intelligence* 22, no. 7-8 (2008): 619-683.
- [21] Koutra, Danai, Joshua T. Vogelstein, and Christos Faloutsos. "Deltacon: A principled massive-graph similarity function." In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pp. 162-170. Society for Industrial and Applied Mathematics, 2013.
- [22] Wilson, Richard C., and Ping Zhu. "A study of graph spectra for comparing graphs and trees." *Pattern Recognition* 41, no. 9 (2008): 2833-2841.
- [23] Zager, Laura A., and George C. Verghese. "Graph similarity scoring and matching." *Applied mathematics letters* 21, no. 1 (2008): 86-94.
- [24] Pržulj, Nataša, Derek G. Corneil, and Igor Jurisica. "Modeling interactome: scale-free or geometric?" *Bioinformatics* 20, no. 18 (2004): 3508-3515.
- [25] Middendorf, Manuel, Etay Ziv, and Chris H. Wiggins. "Inferring network mechanisms: the *Drosophila melanogaster* protein interaction network." *Proceedings of the National Academy of Sciences of the United States of America* 102, no. 9 (2005): 3192-3197.
- [26] Pržulj, Nataša. "Biological network comparison using graphlet degree distribution." *Bioinformatics* 23, no. 2 (2007): e177-e183.
- [27] Motallebi, Sadegh, Sadegh Aliakbary, and Jafar Habibi. "Generative model selection using a scalable and size-independent complex network classifier." *Chaos: An Interdisciplinary Journal of Nonlinear Science* 23, no. 4 (2013): 043127.
- [28] Aliakbary, Sadegh, Sadegh Motallebi, Sina Rashidian, Jafar Habibi, and Ali Movaghar. "Noise-tolerant model selection and parameter estimation for complex networks." *Physica A: Statistical Mechanics and its Applications* 427 (2015): 100-112.
- [29] Goncalves, Wesley Nunes, Alexandre Souto Martinez, and Odemir Martinez Bruno. "Complex network classification using partially self-avoiding deterministic walks." *Chaos: An Interdisciplinary Journal of Nonlinear Science* 22, no. 3 (2012): 033139.
- [30] Jurman, Giuseppe, Roberto Visintainer, Michele Filosi, Samantha Riccadonna, and Cesare Furlanello. "The HIM glocal metric and kernel for network comparison and classification." In *Data Science and Advanced Analytics (DSAA), 2015-36678 2015. IEEE International Conference on*, pp. 1-10. IEEE, 2015.
- [31] Crawford, Brian, Raluca Gera, Jeffrey House, Thomas Knuth, and Ryan Miller. "Graph Structure Similarity using Spectral Graph Theory." In *International Workshop on Complex Networks and their Applications*, pp. 209-221. Springer International Publishing, 2016.
- [32] Schieber, Tiago A., Laura Carpi, Albert Díaz-Guilera, Panos M. Pardalos, Cristina Masoller, and Martín G. Ravetti. "Quantification of network structural dissimilarities." *Nature communications* 8 (2017): 13928.
- [33] Borgwardt, Karsten Michael. "Graph kernels." PhD diss., Ludwig-Maximilian University of Munich, 2007.
- [34] Kelmans, Alexander K. "Comparison of graphs by their number of spanning trees." *Discrete Mathematics* 16, no. 3 (1976): 241-261.
- [35] Kondor, Risi, Nino Shervashidze, and Karsten M. Borgwardt. "The graphlet spectrum." In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 529-536. ACM, 2009.
- [36] Mahadevan, Priya, Dmitri Krioukov, Kevin Fall, and Amin Vahdat. "Systematic topology analysis and generation using degree correlations." In *ACM SIGCOMM Computer Communication Review*, vol. 36, no. 4, pp. 135-146. ACM, 2006.
- [37] Achard, Sophie, and Ed Bullmore. "Efficiency and cost of economical brain functional networks." *PLoS Comput Biol* 3, no. 2 (2007): e17.
- [38] Latora, Vito, and Massimo Marchiori. "Efficient behavior of small-world networks." *Physical review letters* 87, no. 19 (2001): 198701.
- [39] Newman, Mark EJ. "Assortative mixing in networks." *Physical review letters* 89, no. 20 (2002): 208701.
- [40] Clauset, Aaron, Mark EJ Newman, and Cristopher Moore. "Finding community structure in very large networks." *Physical review E* 70, no. 6 (2004): 066111.
- [41] Costa, L. da F., Francisco A. Rodrigues, Gonzalo Travieso, and Paulino Ribeiro Villas Boas. "Characterization of complex networks: A survey of measurements." *Advances in physics* 56, no. 1 (2007): 167-242.
- [42] Badham JM. Commentary: measuring the shape of degree distributions. *Network Sci.* 1, (2013), 213-225. (doi:10.1017/nws.2013.10)

- [43] Hoffer, Elad, and Nir Ailon. "Deep metric learning using triplet network." In International Workshop on Similarity-Based Pattern Recognition, pp. 84-92. Springer International Publishing, 2015.
- [44] Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815-823. 2015.
- [45] Kumar, B. G., Gustavo Carneiro, and Ian Reid. "Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5385-5394. 2016.
- [46] Chechik, Gal, Varun Sharma, Uri Shalit, and Samy Bengio. "Large scale online learning of image similarity through ranking." Journal of Machine Learning Research 11, no. Mar (2010): 1109-1135.
- [47] Wang, Jiang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. "Learning fine-grained image similarity with deep ranking." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1386-1393. 2014.
- [48] Wurtz, Robert H. "Recounting the impact of Hubel and Wiesel." The Journal of physiology 587, no. 12 (2009): 2817-2823.
- [49] Bengio, Yoshua, and Yann LeCun. "Scaling learning algorithms towards AI." Large-scale kernel machines 34, no. 5 (2007): 1-41.
- [50] Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." IEEE transactions on pattern analysis and machine intelligence 35, no. 8 (2013): 1798-1828.
- [51] Arel, Itamar, Derek C. Rose, and Thomas P. Karnowski. "Deep machine learning-a new frontier in artificial intelligence research [research frontier]." IEEE Computational Intelligence Magazine 5, no. 4 (2010): 13-18.
- [52] Hinton, Geoffrey E. "Learning multiple layers of representation." Trends in cognitive sciences 11, no. 10 (2007): 428-434.
- [53] Erdős, Paul, and Alfréd Rényi. "On the central limit theorem for samples from a finite population." Publ. Math. Inst. Hungar. Acad. Sci 4 (1959): 49-61.
- [54] Volchenkov, Dmitri, and Ph Blanchard. "An algorithm generating random graphs with power law degree distributions." Physica A: Statistical Mechanics and its Applications 315, no. 3 (2002): 677-690.
- [55] Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of Machine Learning Research 9, no. Nov (2008): 2579-2605.
- [56] L.J.P. van der Maaten. Learning a Parametric Embedding by Preserving Local Structure. In Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AIS-TATS), JMLR W&CP, 5 (2009):384-391.
- [57] Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." (2007): 3-24.
- [58] Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators." Neural networks 2, no. 5 (1989): 359-366.
- [59] J.C. Platt, Fast training of support vector machines using sequential minimal optimization, in: Advances in Kernel Methods, MIT Press, 1999, pp. 185-208.
- [60] Guimera, Roger, Leon Danon, Albert Diaz-Guilera, Francesc Giralt, and Alex Arenas. "Self-similar community structure in a network of human interactions." Physical review E 68, no. 6 (2003): 065103.
- [61] SNAP: Stanford Network Analysis Project. Retrieved from <http://snap.stanford.edu/> on January 16, 2017.
- [62] E-mail network URV. Retrieved from <http://deim.urv.cat/~alexandre.arenas/data/welcome.htm> on January 16, 2017.
- [63] Leskovec, Jure, and Christos Faloutsos. "Sampling from large graphs." In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 631-636. ACM, 2006.
- [64] Lee, Sang Hoon, Pan-Jun Kim, and Hawoong Jeong. "Statistical properties of sampled networks."Physical review E 73, no. 1 (2006): 016102.
- [65] Ahn, Yong-Yeol, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. "Analysis of topological characteristics of huge online social networking services." In Proceedings of the 16th international conference on World Wide Web, pp. 835-844. ACM, 2007.
- [66] Gjoka, Minas, Maciej Kurant, Carter T. Butts, and Athina Markopoulou. "Walking in facebook: A case study of unbiased sampling of osns." In 2010 Proceedings IEEE Infocom, pp. 1-9. Ieee, 2010.
- [67] Ribeiro, Bruno, and Don Towsley. "Estimating and sampling graphs with multidimensional random walks." In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, pp. 390-403. ACM, 2010.
- [68] Hu, Pili, and Wing Cheong Lau. "A survey and taxonomy of graph sampling."arXiv preprint arXiv:1308.5865(2013).

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46