Feature Extraction from Degree Distribution for Comparison and Analysis of Complex Networks

SADEGH ALIAKBARY, JAFAR HABIBI, AND ALI MOVAGHAR

Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
Email: aliakbary@ce.sharif.edu, jhabibi@sharif.edu, movaghar@sharif.edu

The degree distribution is an important characteristic of complex networks. In many data analysis applications, the networks should be represented as fixed-length feature vectors and therefore the feature extraction from the degree distribution is a necessary step. Moreover, many applications need a similarity function for comparison of complex networks based on their degree distributions. Such a similarity measure has many applications including classification and clustering of network instances, evaluation of network sampling methods, anomaly detection, and study of epidemic dynamics. The existing methods are unable to effectively capture the similarity of degree distributions, particularly when the corresponding networks have different sizes. In this paper, we propose a feature extraction method and a similarity function for the degree distributions in complex networks. We propose to calculate the feature values based on the mean and standard deviation of the node degrees in order to decrease the effect of the network size on the extracted features. Experiments on a wide range of real and artificial networks confirms the accuracy, stability, and effectiveness of the proposed method.

Keywords: Complex Network; Degree Distribution; Feature Extraction; Distance Function; Classification; Social Networks

Received 09 June 2014; revised 23 October 2014

1. INTRODUCTION

Degree distribution is an important and informative characteristic of a complex network [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17]. Degree distribution is also a sign of link formation process in the network. The degree distribution of complex networks often follows a heavy-tailed distribution [6, 7, 8, 9, 18], but networks still show different characteristics in their degree distributions. Hence, we frequently need an appropriate similarity function in order to compare the degree distribution of the complex networks. Such a similarity function plays an important role in many network analysis applications, such as evaluation of network models [1, 2, 19, 5, 7, 16], evaluation of sampling methods [9, 20, 11, 21, 22], model selection [23, 24, 25], classification or clustering of network instances [3, 12, 24, 25, 23], anomaly detection [26, 27], and study of epidemic dynamics [28, 29, 30]. For example, in evaluation of sampling algorithms, the given network instance is compared with its sampled counterpart in order to ensure that the structure of the degree distribution is preserved [9, 20, 11, 21, 22, 12].

In addition to the need for network comparison, representing the network as a fixed-size feature vector is also an important step in every data analysis process [23, 24, 25]. In order to employ the degree distribution in such applications, a procedure is needed for extracting a feature vector from the degree distribution. The extracted feature vector is also useful in developing a distance function for comparing two degree distributions. Although there exist accepted quantification methods for many network features (e.g., average clustering coefficient, modularity, and average shortest path length), the quantification and comparison of the degree distribution is not a trivial task. We will show that the existing methods have major weaknesses in feature extraction and comparison of networks, particularly when the considered networks have different sizes. Even if no network comparison is required, feature extraction from the degree distributions has independent applications. For example, data-analysis algorithms require the network features, including the degree distribution, to be represented as some real numbers in the form of a fixed-length feature vector [23, 12, 24, 25].

According to the mentioned applications for the demanded similarity function, two degree distributions are regarded similar if the corresponding networks follow similar connection patterns and similar link formation processes, even if the networks are of different scales. The state of the art approaches for comparing degree distributions are eye-balling the distribution diagrams (usually, to satisfy a heavy-tailed distribution) [2, 7, 13], Kolmogorov-Smirnov (KS) test [6, 8, 9, 31, 32, 33], comparison based on fitted power-law exponent [13, 34], and distribution percentiles [24]. Eyeballing is obviously an inaccurate, error-prone and manual task. Comparison based on power-law exponent is based on the assumption that the degree distributions obeys a power-law, which is invalid for many complex networks [10, 35, 36, 37]. KS-test is based on a point-to-point comparison of the cumulative distribution function, which is not a good approach for comparing networks with different ranges of node degrees. Percentile method is too sensitive to the outlier values of node degrees. As a result, the existing methods are actually inappropriate for comparing the degree distribution of networks.

In order to reveal the limitations of the existing methods and the main ideas of our proposed method, Figure 1 and Figure 2 provide intuitive examples of networks with different sizes over time. Figure 1 illustrates the degree distribution of two citation networks and two collaboration (co-authorship) networks which are extracted from CiteSeerX digital library [38]. The networks represent two snapshots in the years 1992 and 2010. For example, *Citation_1992* represents the graph of citations in the papers of the CiteSeerX repository which are published before 1992. The figure shows two similar distributions for the two citation networks (resembling a power-law distribution) and two similar distributions for the two collaboration networks (similar to log-normal model). Obviously, the degree distributions of the citation networks are dissimilar to those of the collaboration networks. The existing similarity functions are unable to capture such similarity and dissimilarity in the shape of the degree distributions appropriately. For example, the Kolmogorov-Smirnov test will return the least similarity between the two citation networks, the power-law exponents are uninformatively different for the four networks, and the percentile quantification method [24] returns an equal vector for all the four networks. Although the shape of the degree distributions in this example clearly reveals the similarity of the networks, in many situations the similarity of the degree distributions are not that obvious. Therefore, we need an automatic method for feature extraction and/or quantified network comparison using the degree distributions. It is worth noting that many real networks do not follow a specific distribution model such as the power-law or log-normal models. As a result, fitting the degree distribution to some predefined distribution models is not a good approach for feature extraction or comparison of the networks.

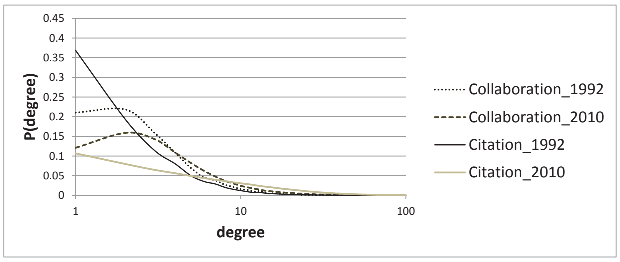


FIGURE 1: Degree distribution of four network snapshots. Although the degree distribution in all of these networks have a long tail, the degree distributions of the two citation networks are similar to each other and different from those of the collaboration networks.

As another example, Figure 2 shows the degree distribution of the citation networks, extracted from the same CiteSeerX repository for different snapshots from 1990 to 2010. These networks have become bigger over years, with 191,443 nodes (papers) in 1990 and 1,039,952 nodes in 2010. Although all the networks show similar degree distributions, the bigger networks contain a wider range of node degrees. As the network size increases, maximum node degree, average degree, and standard deviation of the node degrees also increase. As a result, it seems that a degree distribution should be normalized according to its mean and deviation so that the comparison of the networks with different scales become valid. We follow this idea in our proposed methods of feature extraction and comparison of the network degree distributions. In our proposed "feature extraction" method, a fixed-length feature vector is extracted from the degree distribution, which can be used in data analysis applications, data-mining algorithms and comparison of degree distributions. In the "comparison method", we propose a distance function that computes the distance (amount of dissimilarity) between two given network degree distributions. With such a distance function, we can figure out how similar the given networks are, according to their degree distributions.

Although our proposed approach is of general nature and equally applicable to other network types, in this paper we focus on simple undirected networks. In the rest of this paper, our proposed method is called "Degree Distribution Quantification and Comparison (DDQC)". The rest of this paper is organized as follows: In section 2, we briefly overview the related works. In section 3, we propose a method for degree distribution quantification and comparison. In section 4, we evaluate the proposed method and we compare it with baseline methods. Finally, we conclude the paper in section 5.

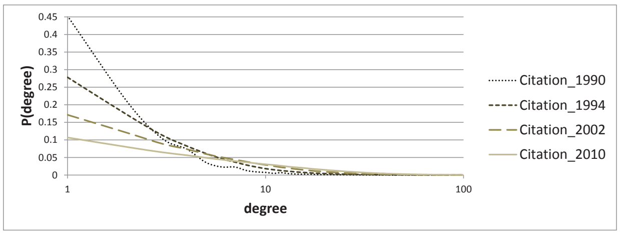


FIGURE 2: Degree distribution of a citation network over different times. As the network grows, the maximum, average and deviation of the node degrees also increase.

2. RELATED WORKS

The degree distribution of many real-world networks are heavy tailed [6, 7, 8, 9, 18], and the power-law distribution is the most suggested model for complex networks [4, 6, 8]. In power-law degree distribution the number of nodes with degree d is proportional to $d^{-\gamma}$ ($N_d \propto d^{-\gamma}$) where γ is a positive number called "the power-law exponent". The value of γ is typically in the range $2 < \gamma < 3$ [2, 39, 36]. The fitted power-law exponent can be used to characterize graphs [34]. A common approach for feature extraction from the degree distribution is to fit it on a power-law model and estimate the power-law exponent (γ). As a result, it will be possible to compare networks according to their fitted power-law exponents. One of the drawbacks of this approach is that power-law exponent is too limited to represent a whole degree distribution. This approach also follows the assumption that the degree distribution is power-law, which is not always valid, because many networks follow other degree distribution models such as log-normal distribution [10, 35, 36, 37]. In addition, the power-law exponent does not reflect the deviation of the degree distribution from the fitted power-law distribution. As a result, two completely different distributions may have similar quantified feature (fitted power-law exponent).

An alternative approach for comparing the degree distributions is to utilize statistical methods of probability distribution comparison. Degree distribution is a kind of probability distribution and there are a variety of measures for calculating the distance between two probability distributions. In this context, the most common method is the Kolmogorov-Smirnov (KS) test, which is defined as the maximum distance between the cumulative distribution functions (CDF) of the two probability distributions [6]. KS-test is used for comparing two degree distributions (two-sample KS test) [9, 31] and also for comparing a degree distribution with a baseline (usually the power-law) distribution [6, 10, 40, 41]. The KS distance of two distributions is calculated according to Equation 1, in which $S_1(d)$ and $S_2(d)$ are the CDFs of the two degree distributions, and d indicates the node degree. KS-test is largely utilized in the literature for comparing degree distribution of complex networks [6, 8, 9, 31, 32, 33]. KS-test is a method for comparing the degree distributions and calculating their distance, and it does not provide feature extraction mechanism. As a result, we should maintain the CDF of the degree distributions so that we can compare them according to KS-test. This is a drawback of KS-test since in other investigated approaches the degree distribution is summarized in a small fixed-length feature vector. Additionally, the KS-test is not applicable in data analysis applications that rely on feature vector representation of networks. KS-test is also sensitive to the scale and size of the networks, since it performs a point-to-point comparison of CDFs. Therefore, for two networks with different ranges of node degrees, the KS-test may return a large value as their distance even if the overall views of the degree distributions are similar (refer to Figure 2).

$$distance_{KS}(S_1, S_2) = \max_{d} |S_1(d) - S_2(d)| \qquad (1)$$

Janssen et. al., [24] propose an alternative method for feature extraction from the degree distribution. In this method, the degree distribution is divided into eight equal-size regions and the sum of degree probabilities in each region is extracted as distribution percentiles. This method is sensitive to the range of node degrees and also to outlier values of degrees. We recall this technique as "Percentiles" and we include it in the set of the baseline methods, along with "KS-test" and "Power-law" (the power-law exponent) in order to evaluate our proposed distance metric which is called "DDQC".

3. PROPOSED METHOD

The degree distribution of a network is described in Equation 2 as a probability distribution function. The equation shows the probability of node degrees in the given graph G, in which $D(v)$ is the degree of node v, and $P(d)$ is the probability that the degree of a node is equal to d. Based on our observations in networks of different sizes, we propose a method for feature extraction and comparison of the degree distributions. In this method, a vector of eight real numbers is extracted from the degree distribution. A distance function is also suggested for comparing the feature vectors. In order to reduce the impact of the network size on the extracted features, we considered the mean and standard deviation of the degree distribution in the feature extraction procedure. Equation 3 and Equation 4 show the mean and standard deviation of the degree distribution respectively.

According to Equation 5, for any network G, we divide the range of node degrees into four regions ($R(r), r = 1..4$). The regions are defined based on the mean, standard deviation, minimum, and maximum of node degrees. Equation 6 shows the length of each region ($|R(r)|$), in which $left(R)$ indicates the lower-bound (minimum degree) in region R and $right(R)$ is its upper-bound. As Equation 7 shows, each region is further divided into two equal-size intervals ($I(i), i = 1..8$). Although it is possible to consider more than two intervals in each region, our experiments showed that considering more than two intervals per region results in no considerable improvements in the accuracy of the distance metric. Equation 8 defines the “interval degree probability” (IDP) which shows the probability that the specified interval I includes the degree of a randomly chosen node. We have defined eight intervals in the degree distribution (four regions and two intervals per region), and the degree distribution can be quantified based on the IDP of the eight intervals. Equation 9 shows the final quantification (feature-vector) of the degree distribution. The proposed feature extraction method can be utilized in network analysis applications which rely on feature extraction and/or comparison of network degree distributions.

$$P(d) = P(D(v) = d); v \in V(G) \tag{2}$$

$$\mu = \sum_{d=D_{min}}^{D_{max}} d \times P(d) \tag{3}$$

$$\sigma = \sqrt{\sum_{d=D_{min}}^{D_{max}} P(d) \times (d-\mu)^2} \tag{4}$$

$$R(r) = \begin{cases} [D_{min}, \mu - \sigma] & r = 1 \\ [\mu - \sigma, \mu] & r = 2 \\ [\mu, \mu + \sigma] & r = 3 \\ [\mu + \sigma, D_{max}] & r = 4. \end{cases} \tag{5}$$

$$|R(r)| = max\Big(right\big(R(r)\big) - left\big(R(r)\big), 0\Big) \tag{6}$$

$$I(i) = \begin{cases} [left(R(\lceil \frac{i}{2} \rceil)), left(R(\lceil \frac{i}{2} \rceil)) + \frac{|R(\lceil \frac{i}{2} \rceil)|}{2}] & i \text{ is odd} \\ [left(R(\lceil \frac{i}{2} \rceil)) + \frac{|R(\lceil \frac{i}{2} \rceil)|}{2}, right(R(\lceil \frac{i}{2} \rceil))] & i \text{ is even} \end{cases} \tag{7}$$

$$IDP(I) = P(left(I) \leq D < right(I)) \tag{8}$$

$$Q(G) = \langle IDP(I(i)) \rangle_{i=1..8} \tag{9}$$

After the feature extraction phase, we can compare the degree distribution of two networks G_1 and G_2 according to their quantified feature vectors. We propose the Equation 10 for comparing two degree distributions. This equation compares two networks based on their corresponding feature vectors (Q_i values). Intuitively, $d(G_1, G_2)$ compares the corresponding interval degree probabilities of the two networks and sums their differences. Equation 10 is a distance function for degree distribution of networks, and it is the result of a comprehensive study of different real, artificial and temporal networks.

$$d(G_1, G_2) = distance(G_1, G_2) = \sum_{i=1}^{8} |Q_i(G_1) - Q_i(G_2)| \tag{10}$$

4. EVALUATION

In this section, we evaluate our proposed method. In subsection 4.1, we describe different network datasets which are used in our evaluations. In subsection 4.2, we describe the evaluation criteria and in subsection 4.3, we compare our proposed method with baseline methods.

4.1. Datasets

In our problem setting, we aim a distance function that given the degree distribution of two networks, calculates how similar they are. But what does this “similarity” mean for degree distributions? What benchmark is available for evaluating such a distance function? For evaluating different distance metrics, an approved dataset of networks with known distances of its instances is sufficient. Although there is no such an accepted benchmark of networks with known “distance values”, there exist some similarity witnesses among the networks. For evaluating different distance metrics, we have prepared two network datasets with admissible similarity witnesses among the networks of these datasets:

- ***Artificial Networks***. We generated a dataset of 6,000 labeled artificial networks using six different models. The utilized models are network generation methods for synthesizing graphs that resemble the topological properties of real-world networks. We considered six models in the evaluations: Barabási-Albert [5], Erdős-Rényi [42], Forest Fire [7], Kronecker model [16], random power-law [17], and Small-world (Watts-Strogatz) model [19]. Each evaluation scenario consists of 100 iterations of network generation and the average of the evaluation results of the 100 iterations are reported. In each iteration 60 networks are generated using the six models (10 networks per model). As a result, the evaluations on the artificial networks includes the generation of a total of 6,000 network instances using different parameters. The number of nodes in generated networks ranges from 1,000 to 5,000 nodes with the average of 2,936.34 nodes in each network instance. The average number of edges is 13,714.75. In this dataset, the models are the witnesses of the similarity: The networks generated by the same model follow identical link formation rules, and their degree distributions are considered similar. The networks of this data-set are further described in the Appendix A, along with an overview of the selected models.
- ***Real-world Networks***. We have collected a dataset of 33 real-world networks of different types. The networks are selected from six different network classes: Friendship networks, communication networks, collaboration networks, citation networks, peer to peer networks and graph

of linked web pages. The category of networks is a sign of similarity: networks of the same type usually follow similar link formation procedures and produce similar degree distributions. So, when comparing two network instances, we expect the distance metric to return small distances (in average) for networks of the same type and relatively larger distances for networks with different types. The “real-world networks” dataset is described in the Appendix A, along with the basic properties and the source of its members.

We assume that the category of the networks in the artificial and real datasets is a sign of their similarity. Networks of the same type follow similar link formation procedures and produce networks with similar structures. Although it is possible for two different-class networks to be more similar than two same-class networks, we assume that the overall “expected similarity” among networks of the same class is more than the expected similarity of different-class networks. This definition of the network similarity which is based on the network types is frequently utilized in the literature [23, 24, 25, 43, 44, 45, 46].

4.2. Evaluation Criteria

In the section 4.1, we described our two network datasets and we introduced different signs and witnesses of similarities among networks of these datasets. We can consider these witnesses in the evaluation of the proposed method. In this subsection, we describe the criteria that we utilized in order to evaluate the proposed method.

The accuracy of the K-Nearest-Neighbor (KNN) classifier [47] is the first investigated evaluation criterion. The KNN rule is a common method for classification. It categorizes an unlabeled example by the majority label of its k-nearest neighbors in the training set. The performance of KNN is essentially dependent on the way that similarities are computed between different examples. Hence, better distance metrics result in better classification accuracy of KNN. In order to evaluate the accuracy of different distance functions, we employ them in KNN classification and we test the accuracy of this classifier. This evaluation is performed for both labeled datasets of real-world and artificial networks. As Equation 11 shows, in a *dataset* of labeled instances, the KNN-accuracy of a distance metric d is the probability that the predicted class of an instance is equal to its actual class, when the distance metric d is used in the KNN classifier. In classification of artificial networks, the models (e.g., Barabási-Albert and Watts-Strogatz) are the network classes. In the case of real networks, the network types (e.g., collaboration and citation networks) play the role of the classes.

$$KNN\text{-}Accuracy(d) = P(KNN\text{-}Classify_d(x) = class(x)),\quad x \in dataset \tag{11}$$

Precision-at-K (P@K) criterion [48, 49] is another metric which is described in Equation 12. P@k is defined to be the average number of classmates in the k nearest neighbors of an instance, divided by k. In other words, P@K for a given network is the mean fraction of instances of the same model in its k most similar networks. P@K is dependent on the distance metric d that is utilized for computing the distances among the instances, therefore, an accurate distance function results in a high P@K.

$$P@K(d) = \frac{E(c)}{k}; c = count(m), m \in KNN_d(x)\ and\ class(m) = class(x), x \in dataset \tag{12}$$

An appropriate distance metric should return smaller distances for networks of the same class. In this context, Dunn Index [50] is a suitable measure for comparing the inter/intra class distances. For any partition setting U, in which the set of instances are clustered or classified into c groups ($U \leftrightarrow X = X_1 \cup X_2 \cup ... \cup X_c$), Dunn defined the *separation index* of U as described in Equation 13 [50]. Dunn index investigates the ratio between the average distance of the two nearest classes (Equation 14) and the average distance between the members of the most extended class (Equation 15).

$$DI(U) = \underbrace{min}_{1\le i \le c}\{\underbrace{min}_{\substack{1\le j\le c\\ j\ne i}}\{\frac{\delta(X_i, X_j)}{\underbrace{max}_{1\le k\le c}\{\Delta(X_k)\}}\}\} \tag{13}$$

$$\delta(S,T) = \delta_{avg}(S,T) = \frac{1}{|S||T|}\sum_{x\in S, y\in T} d(x,y) \tag{14}$$

$$\Delta(S) = \frac{1}{|S|\cdot(|S|-1)}\sum_{x,y\in S, x\ne y} d(x,y) \tag{15}$$

4.3. Evaluation Results

In this subsection, we evaluate our proposed method based on the described criteria, and we compare it with baseline methods. It is also worth noting that the evaluation results in the artificial networks are more stable and reliable than those of the real networks dataset. This is because the artificial networks provides a large set of networks (6000 instances) while the real networks dataset provides only 33 networks.

Figure 3 and Figure 4 illustrate the KNN evaluation in the artificial and real network datasets respectively. In these evaluations, KNN rule is performed with different values of k from 1 to 10, and the average KNN accuracy is computed. Figure 5 and Figure 6 present the results of KNN evaluation for different values of k. The average KNN accuracy is also displayed separately for different classes in Figure 7 and Figure 8. As the figures show, the proposed method results in the best KNN accuracy among the baseline methods. Although in some few cases DDQC is not the top method (e.g., in classification of the citation networks in the real networks dataset), DDQC is undoubtedly the most accurate method in general, in all of the presented evaluation scenarios.

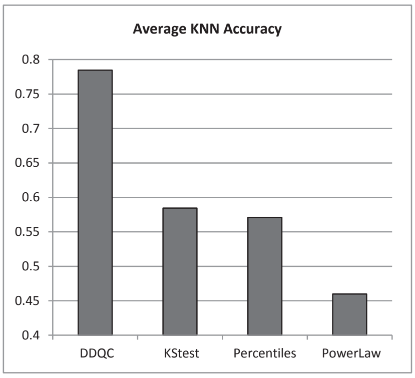


FIGURE 3: Average KNN accuracy in artificial networks dataset (for K=1..10). DDQC outperforms baseline methods by more than 20 percent with respect to average KNN accuracy in this dataset.

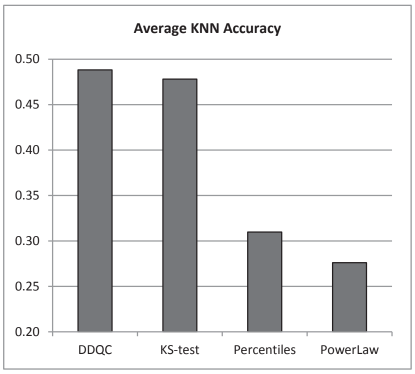


FIGURE 4: Average KNN accuracy in real networks dataset (for K=1..10). DDQC shows the best KNN accuracy in this dataset.

Figure 9 and Figure 10 show the average P@K for K=1..10 in the networks of artificial and real network datasets respectively, according to different distance metrics. As both the figures show, the proposed method outperforms all the baseline methods with respect to the average P@K measure.

Figure 11 and Figure 12 illustrate the Dunn index for different network distance functions in artificial and real network datasets respectively. The proposed method shows the best Dunn index among the baseline methods.

Another interesting characteristic of the proposed method is its stability for large networks. When a network grows, its degree distribution should stabilize, tending to the real distribution of the network. In this sense, degree distribution measurements may capture invariant features to the size of the network. Figure 13 and Figure 14 show the stability analysis for different methods in artificial and real networks respectively. In these figures, the bold line shows average distance among the networks of the same class, and the dotted line shows the average distance among all of the networks. In this experiment, we have also included smaller networks (with 100 to 1000 nodes) in the artificial networks dataset. The evaluations show that as the size of the network grows, our proposed method (DDQC) and the Percentiles method tend to become more stable, particularly for networks with more than 1000 nodes.

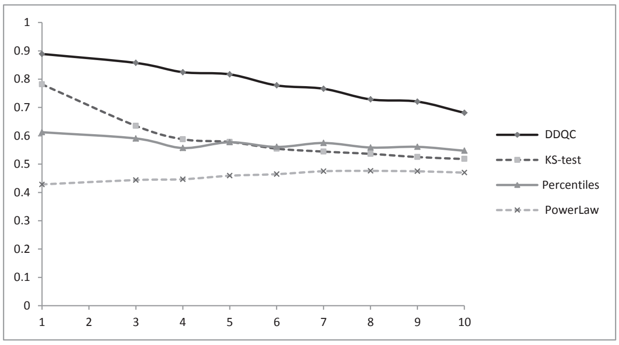


FIGURE 5: KNN accuracy in artificial networks dataset for different values of k. DDQC outperforms baseline methods in all values of k in this dataset.

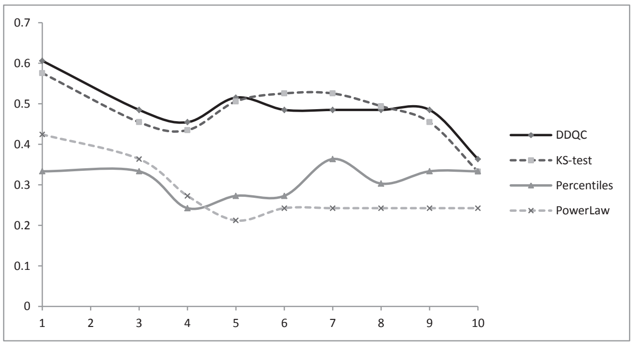


FIGURE 6: KNN accuracy in real networks dataset for different values of k. DDQC performs better than baseline methods in most values of k.

In the last experiment, we investigate the integration of the degree distribution features with other network features, such as clustering coefficient and assortativity. In this experiment, we represent the networks with feature vectors that not only include the degree distribution features, but also some other important features that reflect the structural characteristics of complex networks. Then, we utilize supervised machine learning algorithms in order to classify the networks of the artificial and real network datasets using the integrated feature vectors. The aim of this experiment is to find the degree distribution features that best improve the accuracy of the classifier. Along with the degree distribution features, we consider four well-known network features: 1- “Average clustering coefficient” [19] reflects the transitivity of connections. 2- “Average path length” [3] shows the average shortest path between any pair of nodes. 3- The “assortativity” measure [51] shows the degree correlation between pairs of linked nodes . 4- “Modularity” [52] is a measure for

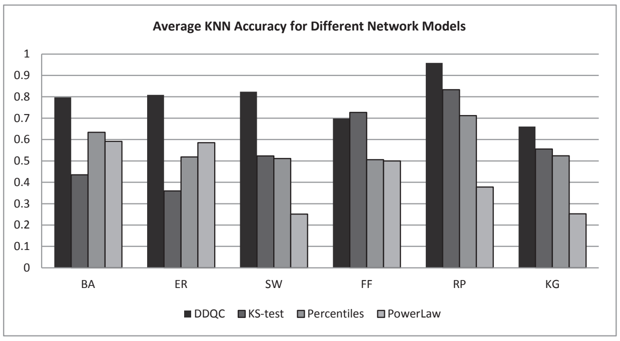


FIGURE 7: KNN accuracy in artificial networks dataset for different network models. DDQC outperforms the baseline methods in most of the models.

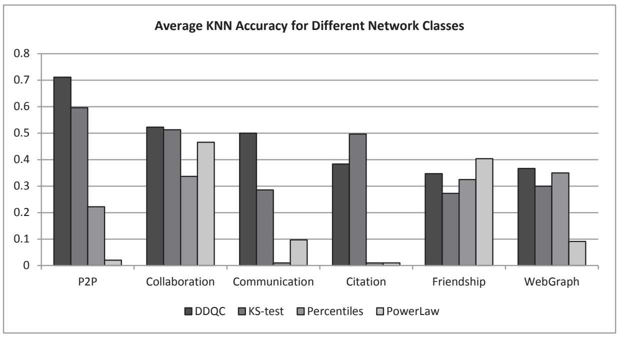


FIGURE 8: KNN accuracy in real networks dataset for different network types. DDQC performs better than baseline methods for most of the network types.

quantifying community structure of a network. When a network is represented with a feature vector of these four properties, we call the feature vector “Features”. We consider three other feature vector representations in which the degree distribution features are also considered: “Features+Powerlaw” adds the fitted power-law exponent, “Features+Percentiles” includes the eight percentile features and “Features+DDQC” adds our proposed features of degree distribution. Since Kolmogorov-Smirnov (KS) does not provide a feature extraction method, we do not consider it in this experiment. We used Support Vector Machines (SVM) [53] as the classification method with 10-fold cross-validation. SVM performs a classification by mapping the inputs into a high-dimensional feature space and constructing hyperplanes to categorize the data instances. We utilized Sequential Minimal Optimization (SMO) [53] which is a common method for solving the optimization problem. Figure 15 shows the accuracy of the classifier based on the described four versions of the feature vectors in the artificial networks dataset. Figure 16 shows the result of the same experiment for the real networks dataset. As both the figures show, among the feature extraction methods for degree distribution, our proposed method results in the best classifier. In other words, our proposed method extracts the most informative features from the degree distribution that can improve the accuracy of network classification.

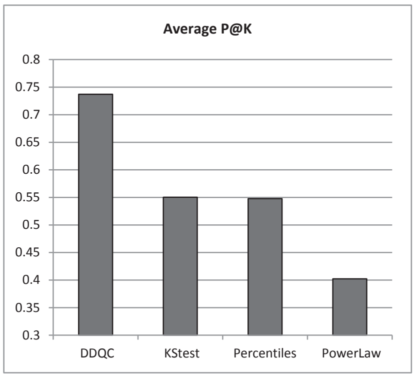


FIGURE 9: Average P@K in artificial networks dataset (for K=1..10). DDQC outperforms the baselines by more than 18 percent with respect to P@K in this dataset.

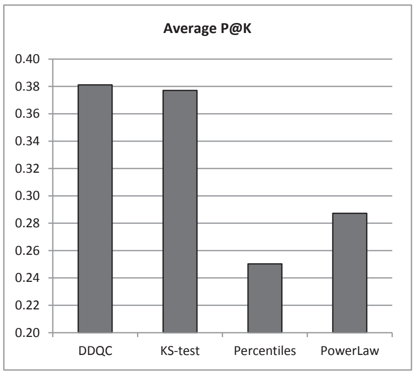


FIGURE 10: Average P@K in real networks dataset (for K=1..10). DDQC shows a higher Dunn index compared with other alternatives in this dataset.

5. CONCLUSION

In this paper, we proposed a method for quantification (feature extraction) and comparison of network degree distributions. The aim of “quantification” is extracting a fixed-length feature vector from the degree distribution. Such a feature vector is used in network analysis applications such as network comparison. The “network comparison” is performed by returning a real number as the distance between two degree distributions. The distance is the counterpart of “similarity” and larger distances indicate less similarity. The degree distribution is an indicator of the link formation process in the network. Similarly evolving networks have analogous degree distributions and we derive the similarity of degree distributions according to the similarity of link formation process in the networks. For deriving the amount of similarity of networks, we introduced admissible witnesses for network similarity:

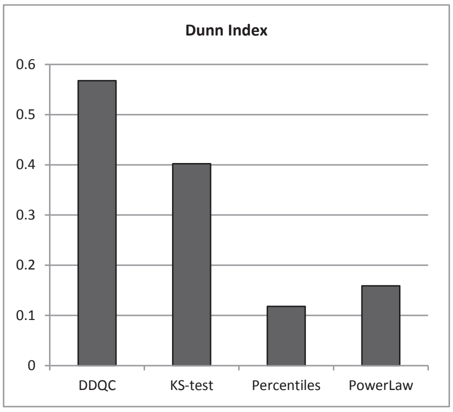


FIGURE 11: Dunn index for different distance metrics in artificial networks dataset. DDQC outperforms the baseline by more than 16 percent with respect to Dunn index in this dataset.

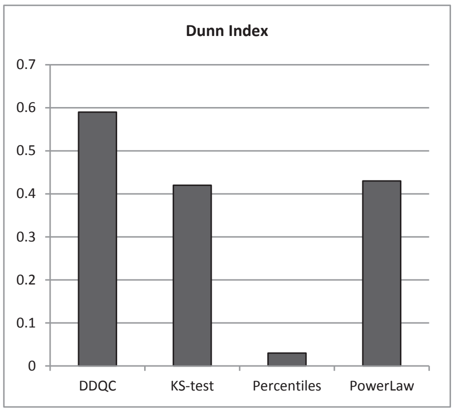
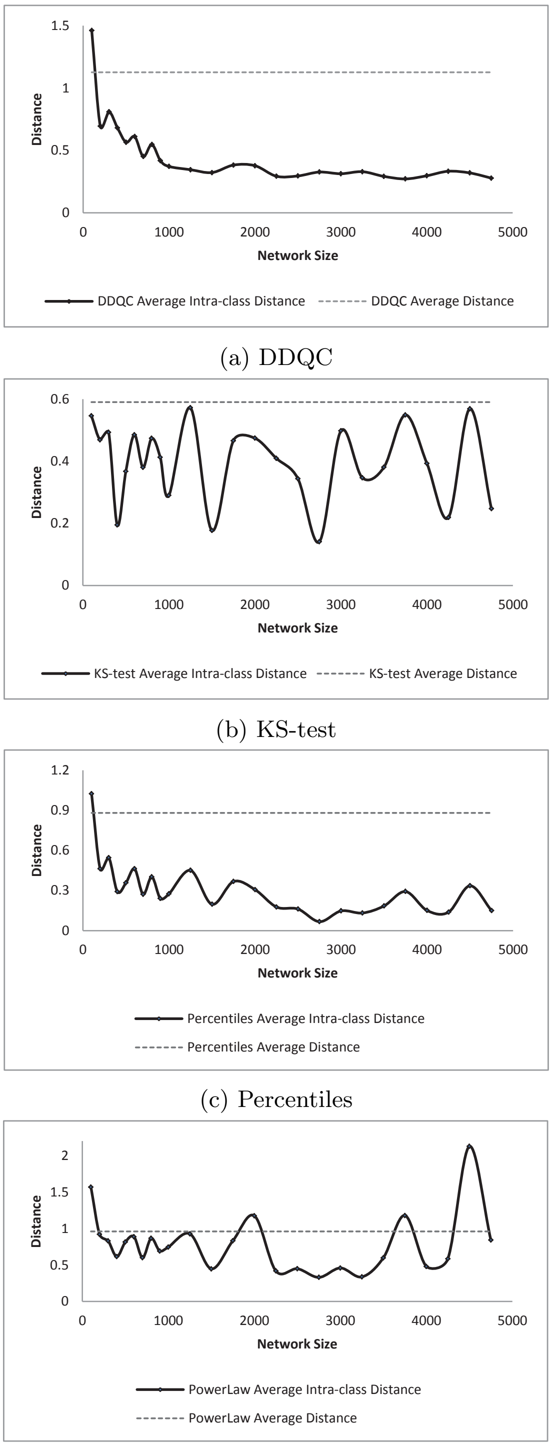


FIGURE 12: Dunn index for different distance metrics in real networks dataset. DDQC performs at least 16 percent better than the alternatives with respect to Dunn index in the real networks dataset.

Similarity among the networks in two categories (same-type real networks and same-model artificial networks). We assume the networks in each of these categories have similar degree distributions. This assumption is the base of our evaluations for different degree-distribution distance metrics. Our proposed method, named DDQC, outperforms the baseline methods with regard to its accuracy in various evaluation criteria. The evaluations are performed based on different criteria such as the ability of the similarity metric to classify networks and comparison of inter/intra class distances. Although the integration of other network features (such as assortativity and modularity) improves the accuracy of the network similarity function, a similarity metric which is only based on the degree distribution has independent and important applications.

Our proposed method enables the data analysis applications and data mining algorithms to employ the degree distribution as a fixed-length feature vector. Hence, it is now possible to represent a network



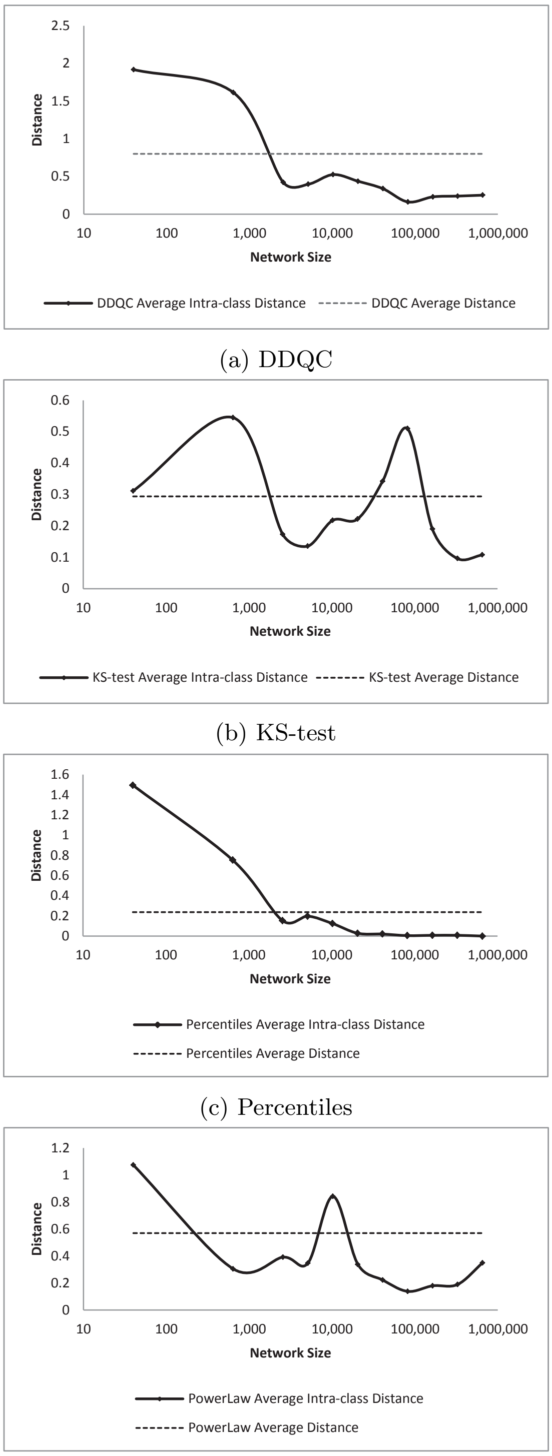
(a) DDQC

(b) KS-test

(c) Percentiles

(d) PowerLaw

FIGURE 13: Stability of different methods with respect to the network size in artificial networks dataset. The horizontal axis shows the considered network size. The vertical axis shows the average distance of the networks. The bold lines show the average distance for networks of the same class, and the dotted lines show the average distance for all of the networks. The figures show that DDQC and Percentiles methods become more stable when we consider larger networks. Particularly, DDQC shows stable distances for networks with more than 1000 networks.



(a) DDQC

(b) KS-test

(c) Percentiles

(d) PowerLaw

FIGURE 14: Stability of different methods with respect to the network size in real networks dataset. The horizontal axis shows the considered network size (logarithmically scaled with base 10). The vertical axis shows the average distance of the networks. The bold lines show the average distance for networks of the same class, and the dotted lines show the average distance for all of the networks. The figures show that Percentiles and DDQC methods show more stability when we consider larger networks

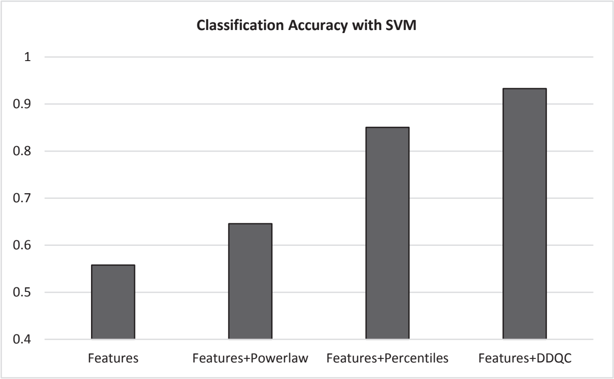


FIGURE 15: SVM classification accuracy in artificial networks dataset. Our proposed method results in features that produces the best classifier.

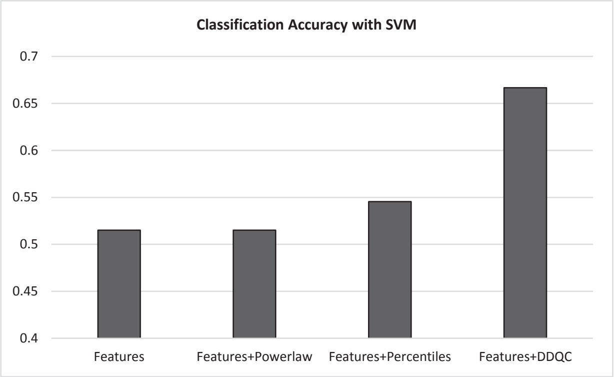


FIGURE 16: SVM classification accuracy in real networks dataset. Our proposed method results in features that produces the best classifier.

instance with a record of features (including clustering coefficient, average path length and the quantified degree distribution) and use such records in data analysis applications. As the future works, we will combine different network features along with the quantified degree distribution in an integrated distance metric for complex networks. Such an integrated distance metric will be the main building block of our future research in evaluation and selection of network models and sampling methods.

ACKNOWLEDGEMENTS

We appreciate Sadegh Motallebi, Sina Rashidian and Javad Gharechamani for their help in implementation and evaluation of the methods and in preparation of network datasets. We also thank Masoud Asadpour, Mahdieh Soleymani Baghshah and Hossein Rahmani for their valuable comments. This research is supported by Iran Telecommunication Research Center (ITRC).

APPENDIX A. OVERVIEW OF ARTIFICIAL AND REAL-WORLD NETWORK DATASETS

In this appendix, we briefly describe the utilized network datasets of this research. The "artificial networks" dataset consists of a total of 6,000 networks, which are synthesized using six generative models. The selected models are some of the important and widely used network generation methods which cover a wide range of degree distribution structures. The models are described in the following, along with their configuration parameters in generation of "artificial networks" dataset. The values of the model parameters are selected according to the hints and recommendations of the cited original papers, along with a concern of keeping the number of network edges balanced for different models. The datasets are available upon request.

- ***Barabási-Albert model (BA)***. This is the classical preferential attachment model which generates scale free networks with power-law degree distributions [5]. In this model, new nodes are incrementally added to the graph, one at a time. Each new node is randomly connected to k existing nodes with a probability that is proportional to the degree of the available nodes. In the artificial networks dataset, k is randomly selected as an integer number from the range $1 \leqslant k \leqslant 10$.
- ***Erdős-Rényi (ER)***. This model generates completely random graphs with a specified density [42]. Network density is defined as the ratio of the existing edges to potential edges. The density of the ER networks in the artificial networks dataset is randomly selected from the range $0.002 \leqslant density \leqslant 0.005$.
- ***Forest Fire (FF)***. This model, in which edge creation is similar to fire-spreading process, supports shrinking diameter and densification properties along with heavy-tailed in-degrees and community structure [7]. This model is configured by two main parameters: Forward burning probability (p) and backward burning probability (p_b). For generating artificial networks dataset, we fixed $p_b = 0.32$ and selected p randomly from the range $0 \leqslant p \leqslant 0.3$.
- ***Kronecker graphs (KG)***. This model generates realistic synthetic networks by applying a matrix operation (the kronecker product) on a small initiator matrix [16]. The model is mathematically tractable and supports many network features including small path lengths, heavy tail degree distribution, heavy tails for eigenvalues and eigenvectors, densification, and shrinking diameters over time. The KG networks of the artificial networks dataset are generated using a 2×2 initiator matrix. The four elements of the initiator matrix are randomly selected from the ranges: $0.7 \leqslant P_{1,1} \leqslant 0.9, 0.5 \leqslant P_{1,2} \leqslant 0.7, 0.4 \leqslant P_{2,1} \leqslant 0.6, 0.2 \leqslant P_{2,2} \leqslant 0.4$.
- ***Random power-law (RP)***. This model follows a variation of ER model and generates networks with power law degree distribution [17]. This model is configured by the power-law degree exponent (γ). In our parameter setting, γ is randomly selected from the range $2.5 < \gamma < 3$.
- ***Watts-Strogatz model (WS)***. The classical Watts-Strogatz small-world model synthesizes networks with small path lengths and high clustering [19]. It starts with a regular lattice, in which each node is connected to k neighbors, and then randomly rewires some edges of the network with rewiring probability β. In WS networks of the artificial networks dataset, β is fixed as $\beta = 0.5$, and k is randomly selected from the integer numbers between 2 and 10 ($2 \leqslant k \leqslant 10$).

Table A.1 describes the graphs of the "real-world networks" dataset, along with the category, number of nodes and edges, and the source of these graphs. Most of these networks are publicly available datasets. Two temporal networks (Cit_CiteSeerX and Collab_CiteSeerX) are extracted from CiteSeerx digital library [38], using a web crawler software tool.

REFERENCES

[1] Albert, R. and Barabási, A.-L. (2002) Statistical mechanics of complex networks. *Reviews of modern physics*, **74**, 47–97.

[2] Newman, M. E. (2003) The structure and function of complex networks. *SIAM review*, **45**, 167–256.

[3] Costa, L. d. F., Rodrigues, F. A., Travieso, G., and Villas Boas, P. (2007) Characterization of complex networks: A survey of measurements. *Advances in Physics*, **56**, 167–242.

[4] Easley, D. and Kleinberg, J. (2010) *Networks, Crowds, and Markets: Reasoning about a highly connected world.* Cambridge university press, New York, NY, USA.

[5] Barabási, A.-L. and Albert, R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

[6] Clauset, A., Shalizi, C. R., and Newman, M. E. (2009) Power-law distributions in empirical data. *SIAM review*, **51**, 661–703.

[7] Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005) Graphs over time: Densification laws, shrinking diameters and possible explanations. *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, New York, NY, USA, pp. 177–187. ACM.

[8] Muchnik, L., Pei, S., Parra, L. C., Reis, S. D., Andrade Jr, J. S., Havlin, S., and Makse, H. A. (2013) Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Scientific reports*, **3**, 1783.

[9] Leskovec, J. and Faloutsos, C. (2006) Sampling from large graphs. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, pp. 631–636. ACM.

TABLE A.1: Dataset of real-world networks

Category	ID	Vertices	Edges	Source
Citation Network	Cit-HepPh	34,546	420,899	SNAP [54]
	Cit-HepTh	27,770	352,304	SNAP [54]
	dblp_cite	475,886	2,284,694	DBLP [55]
	Cit_CiteSeerX	1,106,431	11,791,228	CiteSeerX [38]
Collaboration Network	CA-AstroPh	18,772	198,080	SNAP [54]
	CA-CondMat	23,133	93,465	SNAP [54]
	CA-HepTh	9,877	25,985	SNAP [54]
	Collab_CiteSeerX	1,260,292	5,313,101	CiteSeerX [38]
	com-dblp	317,080	1,049,866	SNAP [54]
	dblp_collab	975,044	3,489,572	DBLP [55]
	dblp20080824	511,163	1,871,070	Sommer [56]
	IMDB-09	4,155	16,679	Rossetti [57]
	CA-GrQc	5,242	14,490	SNAP [54]
	CA-HepPh	12,008	118,505	SNAP [54]
Communication Network	EmailURV	1,133	5,451	Aarenas [58]
	Email-Enron	36,692	183,831	SNAP [54, 59]
	Email-EuAll	265,214	365,025	Konect [60]
	WikiTalk	2,394,385	4,659,565	SNAP [54]
Friendship Network	Dolphins	62	159	NetData [61]
	facebook-links	63,731	817,090	MaxPlanck [62]
	Slashdot0811	77,360	507,833	SNAP [54]
	Slashdot0902	82,168	543,381	SNAP [54]
	soc-Epinions1	75,879	405,740	SNAP [54]
	Twitter-Richmond	2,566	8,593	Rossetti [57]
	youtube-d-growth	1,138,499	2,990,443	MaxPlanck [62]
Graph of Web Pages	web-BerkStan	685,230	6,649,470	SNAP [54]
	web-Google	875,713	4,322,051	SNAP [54]
	web-NotreDame	325,729	1,103,835	SNAP [54]
	web-Stanford	281,903	1,992,636	SNAP [54]
P2P Network	p2p-Gnutella04	10,876	39,994	SNAP [54]
	p2p-Gnutella05	8,846	31,839	SNAP [54]
	p2p-Gnutella06	8,717	31,525	SNAP [54]
	p2p-Gnutella08	6,301	20,777	SNAP [54]

[10] Gómez, V., Kaltenbrunner, A., and López, V. (2008) Statistical analysis of the social network and discussion threads in slashdot. *Proceedings of the 17th International Conference on World Wide Web*, New York, NY, USA, pp. 645–654. ACM.

[11] Stumpf, M. P., Wiuf, C., and May, R. M. (2005) Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 4221–4224.

[12] Airoldi, E. M., Bai, X., and Carley, K. M. (2011) Network sampling and classification: An investigation of network model representations. *Decision support systems*, **51**, 506–518.

[13] Sala, A., Cao, L., Wilson, C., Zablit, R., Zheng, H., and Zhao, B. Y. (2010) Measurement-calibrated graph models for social network experiments. *Proceedings of the 19th International Conference on World Wide Web*, New York, NY, USA, pp. 861–870. ACM.

[14] Boginski, V., Butenko, S., and Pardalos, P. M. (2006) Mining market data: a network approach. *Computers & Operations Research*, **33**, 3171–3184.

[15] Stam, C. J. and Reijneveld, J. C. (2007) Graph theoretical analysis of complex networks in the brain. *Nonlinear biomedical physics*, **1**, 3.

[16] Leskovec, J., Chakrabarti, D., Kleinberg, J., Faloutsos, C., and Ghahramani, Z. (2010) Kronecker graphs: An approach to modeling networks. *The Journal of Machine Learning Research*, **11**, 985–1042.

[17] Volchenkov, D. and Blanchard, P. (2002) An algorithm generating random graphs with power law degree distributions. *Physica A: Statistical Mechanics and its Applications*, **315**, 677–690.

[18] Leskovec, J., Lang, K. J., and Mahoney, M. (2010) Empirical comparison of algorithms for network community detection. *Proceedings of the 19th International Conference on World Wide Web*, New York, NY, USA, pp. 631–640. ACM.

[19] Watts, D. J. and Strogatz, S. H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–442.

[20] Stumpf, M. P. and Wiuf, C. (2005) Sampling properties of random graphs: the degree distribution. *Physical Review E*, **72**, 036118.

[21] Lee, S. H., Kim, P.-J., and Jeong, H. (2006) Statistical properties of sampled networks. *Physical Review E*, **73**, 016102.

[22] Han, J.-D. J., Dupuy, D., Bertin, N., Cusick, M. E., and Vidal, M. (2005) Effect of sampling on topology

predictions of protein-protein interaction networks. *Nature biotechnology*, **23**, 839–844.

[23] Motallebi, S., Aliakbary, S., and Habibi, J. (2013) Generative model selection using a scalable and size-independent complex network classifier. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, **23**, 043127.

[24] Janssen, J., Hurshman, M., and Kalyaniwalla, N. (2012) Model selection for social networks using graphlets. *Internet Mathematics*, **8**, 338–363.

[25] Middendorf, M., Ziv, E., and Wiggins, C. H. (2005) Inferring network mechanisms: the drosophila melanogaster protein interaction network. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 3192–3197.

[26] Juszczyszyn, K., Nguyen, N. T., Kolaczek, G., Grzech, A., Pieczynska, A., and Katarzyniak, R. (2006) Agent-based approach for distributed intrusion detection system design. *Proceedings of the 6th International Conference on Computational Science - Volume Part III*, Berlin, Heidelberg, pp. 224–231. Springer-Verlag.

[27] Papadimitriou, P., Dasdan, A., and Garcia-Molina, H. (2010) Web graph similarity for anomaly detection. *Journal of Internet Services and Applications*, **1**, 19–30.

[28] Pastor-Satorras, R. and Vespignani, A. (2002) Epidemic dynamics in finite size scale-free networks. *Physical Review E*, **65**, 035108.

[29] Montanari, A. and Saberi, A. (2010) The spread of innovations in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 20196–20201.

[30] Briesemeister, L., Lincoln, P., and Porras, P. (2003) Epidemic profiles and defense of scale-free networks. *Proceedings of the 2003 ACM Workshop on Rapid Malcode*, New York, NY, USA, pp. 67–75. ACM.

[31] Kossinets, G. and Watts, D. J. (2006) Empirical analysis of an evolving social network. *Science*, **311**, 88–90.

[32] Goldstein, M. L., Morris, S. A., and Yen, G. G. (2004) Problems with fitting to the power-law distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, **41**, 255–258.

[33] Deng, W., Li, W., Cai, X., and Wang, Q. A. (2011) The exponential degree distribution in complex networks: Non-equilibrium network theory, numerical simulation and empirical data. *Physica A: Statistical Mechanics and its Applications*, **390**, 1481–1485.

[34] Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999) On power-law relationships of the internet topology. *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, New York, NY, USA, pp. 251–262. ACM.

[35] Gong, N. Z., Xu, W., Huang, L., Mittal, P., Stefanov, E., Sekar, V., and Song, D. (2012) Evolution of social-attribute networks: Measurements, modeling, and implications using Google+. *Proceedings of the 2012 ACM Conference on Internet Measurement Conference*, New York, NY, USA, pp. 131–144. ACM.

[36] Kwak, H., Lee, C., Park, H., and Moon, S. (2010) What is twitter, a social network or a news media? *Proceedings of the 19th International Conference on World Wide Web*, New York, NY, USA, pp. 591–600. ACM.

[37] Kim, M. and Leskovec, J. (2012) Multiplicative attribute graph model of real-world networks. *Internet Mathematics*, **8**, 113–160.

[38] Citeseerx digital library. http://citeseerx.ist.psu.edu.

[39] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.

[40] Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2007) Measurement and analysis of online social networks. *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, New York, NY, USA, pp. 29–42. ACM.

[41] Corlette, D. and Shipman III, F. (2009) Capturing on-line social network link dynamics using event-driven sampling. *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04*, Washington, DC, USA, pp. 284–291. IEEE Computer Society.

[42] Erdös, P. and Rényi, A. (1959) On the central limit theorem for samples from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy*, **4**, 49–61.

[43] Bagrow, J. P., Bollt, E. M., Skufca, J. D., and Ben-Avraham, D. (2008) Portraits of complex networks. *EPL (Europhysics Letters)*, **81**, 68004.

[44] Berlingerio, M., Koutra, D., Eliassi-Rad, T., and Faloutsos, C. (2013) Network similarity via multiple social theories. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, New York, NY, USA, pp. 1439–1440. ACM.

[45] Mehler, A. (2008) Structural similarities of complex networks: A computational model by example of wiki graphs. *Applied Artificial Intelligence*, **22**, 619–683.

[46] Onnela, J.-P., Fenn, D. J., Reid, S., Porter, M. A., Mucha, P. J., Fricker, M. D., and Jones, N. S. (2012) Taxonomies of networks from community structure. *Physical Review E*, **86**, 036104.

[47] Cover, T. and Hart, P. (1967) Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, **13**, 21–27.

[48] Yu, C.-N. J. and Joachims, T. (2009) Learning structural svms with latent variables. *Proceedings of the 26th Annual International Conference on Machine Learning*, New York, NY, USA, pp. 1169–1176. ACM.

[49] Bian, J., Liu, Y., Agichtein, E., and Zha, H. (2008) Finding the right facts in the crowd: Factoid question answering over social media. *Proceedings of the 17th International Conference on World Wide Web*, New York, NY, USA, pp. 467–476. ACM.

[50] Bezdek, J. C. and Pal, N. R. (1998) Some new indexes of cluster validity. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, **28**, 301–315.

[51] Newman, M. E. (2002) Assortative mixing in networks. *Physical review letters*, **89**, 208701.

[52] Newman, M. E. (2006) Modularity and community structure in networks. *Proceedings of the National*

Academy of Sciences of the United States of America, **103**, 8577–8582.

[53] Platt, J. C. (1999) Advances in kernel methods, chapter. Fast Training of Support Vector Machines Using Sequential Minimal Optimization, pp. 185–208. MIT Press, Cambridge, MA, USA.

[54] Stanford large network dataset collection. http://snap.stanford.edu/data/.

[55] Xml repository of DBLP library. http://dblp.uni-trier.de/xml/.

[56] Christian Sommer's graph datasets. http://www.sommer.jp/graphs/.

[57] Giulio Rossetti networks dataset. http://giuliorossetti.net/about/ongoing-works/datasets.

[58] Alex Arenas's network datasets. http://deim.urv.cat/ aarenas/data/welcome.htm.

[59] Klimt, B. and Yang, Y. (2004) The Enron corpus: A new dataset for email classification research. *Proceedings of the 15th European Conference on Machine Learning*, Berlin, Heidelberg, pp. 217–226. Springer.

[60] The Koblenz network collection. http://konect.uni-koblenz.de/.

[61] Newman's NetData collection. http://www-personal.umich.edu/ mejn/netdata/.

[62] Network datasets at Max Planck. http://socialnetworks.mpi-sws.org.