Physica A 501 (2018) 12–23

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

A new similarity measure for link prediction based on local structures in social networks

Farshad Aghabozorgi, Mohammad Reza Khayyambashi *

Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran

HIGHLIGHTS

- A new similarity measure between two vertices of the network is proposed.
- The proposed measure relies on structural units of online networks named motifs.
- A supervised learning experiment framework is applied to test this measure.
- The results indicate that this proposed measure outperforms others of its kind.
- The model trained with this measure outperforms other models in the link prediction.

ARTICLE INFO

Article history:
Received 14 September 2017
Received in revised form 4 February 2018
Available online 22 February 2018

Keywords:
Network motifs
Link prediction
Node similarity
Supervised learning
Social networks

ABSTRACT

Link prediction is a fundamental problem in social network analysis. There exist a variety of techniques for link prediction which applies the similarity measures to estimate proximity of vertices in the network. Complex networks like social networks contain structural units named network motifs. In this study, a newly developed similarity measure is proposed where these structural units are applied as the source of similarity estimation. This similarity measure is tested through a supervised learning experiment framework, where other similarity measures are compared with this similarity measure. The classification model trained with this similarity measure outperforms others of its kind.

1. Introduction

Social networks are dynamic structures that evolve over time either through addition of new vertices or through new links that form among vertices. Thus in many research articles the study and modeling of the dynamics in network structure are of concern [1,2]. Social networks like other complex networks have global statistical features such as the "small world" property of short paths between any two vertices [3,4]. These networks are "scale free", where the vertex degrees follow a power law distribution [5,6]. Many studies have found that complex networks contain small building blocks named "network motifs", that is, patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks [7,8].

Such motifs could be found in many scientific disciplines: social networks, electronic networks, biological interaction networks, etc. There exist many articles which concentrate on network motifs enumeration [9,10], and explanations of network motifs in the networks [7]. In this article a new similarity measure between two vertices of the network is applied which uses this structural units of networks as an input.

* Corresponding author.
E-mail addresses: aghabozorgi@eng.ui.ac.ir (F. Aghabozorgi), m.r.khayyambashi@comp.ui.ac.ir (M.R. Khayyambashi).

https://doi.org/10.1016/j.physa.2018.02.010
0378-4371/

Link prediction, introduced by Liben-Nowell and Kleinberg [11], refers to a fundamental problem related to the evolution of social networks in time. Given snapshots of a social network at time t and t', the problem is to predict the upcoming links that are likely to appear in the network within the time interval $[t, t']$. This issue can be considered as a problem of supervised learning models, the objective of which is to predict existence of edge between a pair of nodes [12,13].

The link prediction learning method can be divided into two broad categories: (a) link prediction based on unsupervised learning methods [11,14,15], and (b) link prediction based on supervised methods [12,16,17]. The earlier studies on link prediction are conducted on the category (a) and have become restricted due to their incapability to handle imbalance of social network data. Due to different studies the supervised learning methods provide capabilities to improve link prediction results [12,13].

The main contributions of this paper can be summarized as follows: (1) Introducing a new neighborhood-based similarity measure which outperforms others of its kind. (2) Applying local structures occurs between nodes, called network motifs, in proposed similarity measure. (3) Evaluate the proposed measure through supervised learning framework used to solve the link prediction problem.

The rest of the paper is organized as follows. Section 2, reviews the related work on link prediction and network motifs. Section 3, describes applied methodology in detail. Section 4 briefly describes the datasets and explains experiments settings. Section 5 discusses the experimental results. The conclusion is provided in Section 6.

2. Literature review

The proposed measure is related to two different fields of research: the network motifs and the link prediction. Therefore, summarized reviews of both fields are presented. This article applies the concept of network motifs and findings of related literature in similarity calculation. The proposed measure is evaluated through a link prediction experimental framework. The issue of link prediction in social network is a general concern consisting of several topics. Various research articles related to the link prediction are reviewed and the contribution of this article to the state-of-the-art is discussed.

2.1. The network motifs

There exist some articles which assess the discovery of complex networks through the motif analysis. Some of them refer to biologic network such as protein networks [18], some refer to ecologic networks such as food chain networks [19], and some refer to properties of social networks [20–23]. The theme of most of these studies are concerned with network motif detection and numeration problems, where the objective is to detect subgraphs of network and their numeration in a rapid manner in order to identify network properties statistically.

There exist $\binom{g}{3}$ distinct 3-subgraphs and their corresponding lines in a directed graph. These subgraphs are classified by their isomorphism type by [8]. There are 13 isomorphism classes of three node motifs for directed graphs which is organized in Fig. 1.

There exist numerous studies regarding algorithms of detecting network motifs in the networks. Most of them assume exhaust enumeration of all subgraphs with a given number of nodes in the network. Their computational cost increases dramatically with the network size [20]. However, it was recently found that it is possible to use random sampling to estimate concentrations of network motifs effectively [24]. But, random sampling cannot be used to applications of network motifs when the exact number and type of the network motifs are required.

Despite this numerous references, a few of them seek to apply these structural units of network as the source of knowledge. This article proposes an innovative similarity measure which is based on the quality and quantity of network motifs shared between two vertices of network. The vertices sharing more common motifs with greater quality are more similar to each other.

2.2. Link prediction

The link prediction is the task of inferring the missing links in the network based on the given snapshot of the network. It is a generic task for the analysis of networked data which appears in many applications, such as Viral Marketing [25], Recommender Systems [26,27], and Online Social Network Analysis [28,29].

The link prediction is one of the hot research fields in social network analysis [30]. Several studies, for example [31–33] and [34], have been conducted on applying link prediction methods in various applications. Authors in [31] Applied different similarity measures to find out the efficient measure for a complex military network. In [32], authors proposed a phase-dynamic algorithm of the directed network nodes to analyze the role of link directions and demonstrated that the bi-directional links and the one-directional links have different roles in link prediction and network structure formation. In [33] the authors studied a projection based algorithm for link prediction in bipartite networks. Authors in [34] defined significant influence and applied it in link prediction.

In all these applications, resemblance among the network vertices is considered through the similarity measures. Among all of the similarity measures, neighborhood-based measures are commonly used, due to their simplicity, accessibility in complex networks and their acceptable computational complexity [35]. The proposed similarity measure intend to improve neighborhood-based measures through applying existing local structures among nodes. Therefore, the proposed measure is

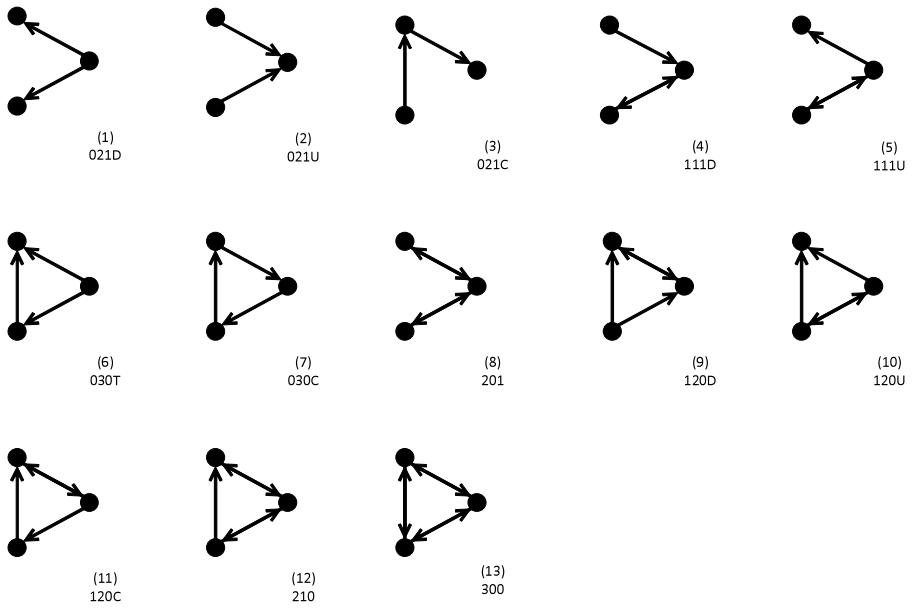

Fig. 1. The 13 types of three node motifs in a directed graph.

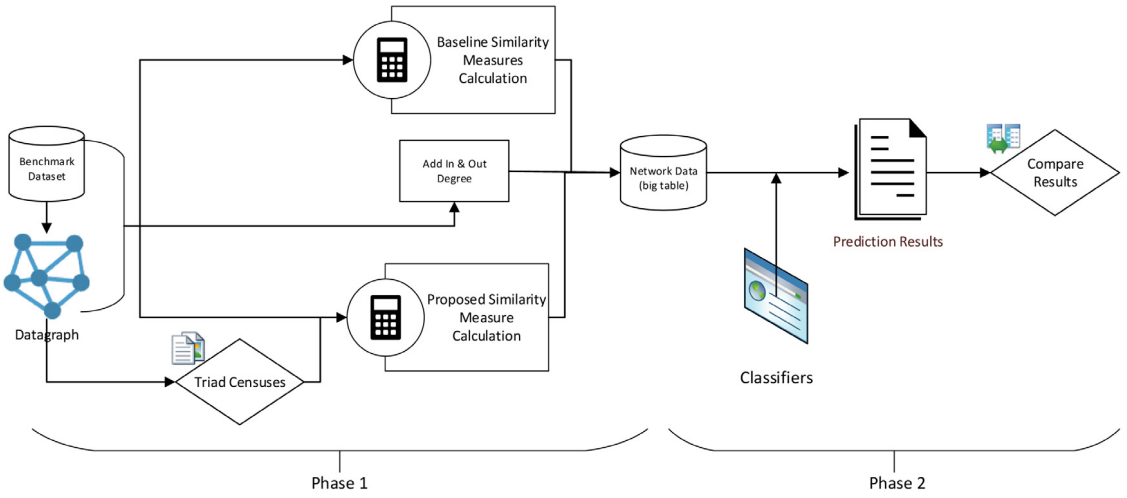

Fig. 2. The process pattern of research methodology.

compared with some baseline similarity measures, including: Common Neighbors [11]; Jaccard's Coefficient [11]; Adamic Adar [15]; Preferential Attachment [36].

The link prediction problem is usually formalized as a binary classification problem (supervised learning) or a ranking problem on the node pairs (unsupervised learning). Different studies indicate that the latter are fundamentally unable to cope with dynamics, interdependencies, and other properties in networks [13]. Therefore, a supervised learning framework is applied.

3. Methodology

The process pattern of this methodology is drawn in Fig. 2. This proposed methodology consists of two major phases: (1) construction of a dataset *Big Table* containing the social network edge information with similarity measures of the source and the destination vertices. Moreover, contains other structural information on the social network vertices. (2) Adopting the supervised learning of prediction model.

The two phases with their corresponding steps are explained as follows. The aim of phase 1 is to provide information needed for supervised learning. These information should be gathered in a table (*Big Table*). There is a need to access the social network graph to calculate common neighbors and triadic censuses. The social network graph (*Datagraph*) is constructed based on the information of vertices presented in the dataset. The proposed similarity measure values are calculated for every pair of nodes in the dataset and these values are added to the big table. The new similarity measure should be compared with other baseline neighborhood-based similarity measures. Therefore, these similarity measures are calculated and added to the big table. Moreover, the intrinsic attributes of vertices which constitute of the Out-degree of the edge origin and In-degree of the edge destination are added.

The next phase of this article's methodology is related to the link prediction base on a supervised learning platform. Training a classifier based on current snapshot of the network is applied to predict upcoming edges. For this purpose the dataset should be partitioned into two sections. The major section consists of the data used as training set and the minor section is used to investigate the testing set. This phase of experiments can be done differently due to various purposes. A linear classifier is applied in this article to compare different similarity measures individually. A boosting algorithm for classification is applied to investigate different aspects of link prediction as a binary classification problem. Also combinations of different similarity measures are used as classification feature set to study impact of using proposed similarity measure.

3.1. Triad similarity measure

The proposed similarity measure is constructed from distributions of vertex involvement in network motifs. These patterns are small building blocks of the information networks. Each pair of network vertices and any of their common neighbors establish a network motif. Therefore every pairs of the vertices in the network could be a member of many triadic network motifs.

Influence of these triadic network motifs is the main difference between proposed similarity measures and other neighborhood based similarity measures. There are many facts in social network analysis and graph theory which are applied in proposition of this measure.

Fact 1. Each pair of nodes and one of their common neighbors construct a three node motif (triad) in the graph.

Fact 2. There are 13 different motifs that consist of three node; these are called 13 isomorphism classes of triads. These isomorphism are ranked based on the number of edges and information presented through them.

Fact 3. Isomorph number 13 is a complete graph; which means that if a pair of nodes and one of their common neighbors construct this kind of motifs it can be interpreted as full interactions (send and receive) between them. On the other hand the low ranked isomorphs can be interpreted as low interactive common neighbors because of their narrow number of edges.

According to above facts the new triadic similarity measure is proposed here. This measure is one of the neighborhood based similarity measures which each of the common neighbors are weighted by the isomorphism class they construct with the pair of nodes. The neighbors which construct complete triadic motif are considered with maximum impact. The common neighbors which construct other types of triadic motifs are considered in accordance with the rank of the isomorphism class they make.

Triadic Similarity measure between two nodes is proposed as follows:

$$score\,(x, y) = \frac{\sum_{z \in\ \Gamma(x) \cap \Gamma(y)} \Phi(x, y, z) \times 1/13}{|\Gamma(x) \cap \Gamma(y)|}. \tag{1}$$

where, $\Phi(x, y, z)$ is the number of the triadic network motif with x, y and z as the participant nodes. The numbering of the triadic network motifs is based on Leinhardt's ordering illustrated in Fig. 1. Network motifs with larger numbers are more effective on the similarity. $\Gamma(x)$ and $\Gamma\ (y)$ denote the neighbor sets of these nodes.

The implementation manner of developing this proposed similarity measure is arranged in Fig. 3.

The complexity of computational time is a crucial challenge in online social networks. The neighborhood-based similarity measures calculation follow up the same procedure [35,37]. If N denotes number of nodes in the network, and k is the average degree which indicate the time complexity to traverse the neighborhood of a node [35], the time complexity of triadic similarity measure is $O(nk^2)$. This time complexity is acceptable for social network analysis applications.

4. Experiments settings

In this section, the experiments settings, the benchmark datasets and classification models applied in the experiments are introduced. The definitions of baseline similarity measures applied to investigate proposed measure are illustrated. All the experiments in this study are developed through R language and environment [38]. The igraph package [39] is applied in order to handle graph related functions. And the Caret package [40] is applied for supervised learning functions of the experiments. The experiments are run on a windows 32-bit platform.

```
(1)  Create Social Network graph (DataGraph) based on the Dataset
(2)  For each vertices pairs in Dataset
     a.   Extract start node's neighbors Γ(x)
     b.   Extract end node's neighbors Γ(y)
     c. Create Common Neighbors set A = Γ(x) ∩ Γ(y)
     d. for each z ∈ A
          I. Identify triadic Census type
          II.  Update Triadic_Sim(x, y)
```

Fig. 3. Developing of the triadic similarity measure through pseudo code.

Table 1
The number of nodes and links for benchmark datasets.

Dataset	Nodes	Edges
Highschool	70	366
Residence hall	217	2672
Advogato	6541	51 127

4.1. The applied data

For the purpose of experiments, three social network datasets are chosen. All of these datasets are used as benchmark in many studies. The results of experiments on these datasets can be generalized to many similar social networks.

The Highschool network dataset [41] which contains friendship among boys in a small high school in Illinois, USA is used for this study. This is a directed, positive weighted social network with 70 vertices and 366 edges, where every boy was asked to choose his friends among others in two consecutive academic year. A node in dataset represents a boy and an edge between two boys shows that the left boy chose the right boy as a friend. The edge weights show how often that happened, hence edge values could be from 1 to 2.

The Residence hall directed network [42] contains friendship ratings among 217 residence living at a residence hall located on the Australian National University campus. Each node represents a person and each edge represents a friendship tie.

The first two datasets are relatively small. Real world social networks are known as huge networks. Therefore, the network of Advogato [43] is chosen to be the third dataset. Advogato is an online community platform for developers of free software. This network is a professional network, defined in [44]. Nodes are users of Advogato and the directed edges represent trust relationships. The dataset contains 6541 vertices and 51127 edges. All of the described datasets are summarized in Table 1.

4.2. Classification methods

There exist many classification algorithms in the field of machine learning. The performance of each classification algorithm is different based on its dataset and attributes. In this article two classification learning algorithm of (1) Gradient Boosting Machine (GBM) and (2) Linear Discriminant Analysis (LDA) are applied.

Linear discriminant analysis [45] is a method used in machine learning to find a linear combination of features which characterizes two classes of objects. The resulting combination may be used as a linear classifier. It is shown that the LDA is a good classification method for binary classification, which link prediction is one of them. On the other hand a classification method with linear characteristic is needed in order to comparison of different features. Two LDA classifier learnt through a same set of features apart a single one can be applied to compare different features. LDA classifiers learnt through different similarity measures in their feature sets are applied for this purpose. The classifier with better prediction capability shows the excellence of its similarity measure.

The accuracy and productivity of a predictive model can be boosted in two ways: either by including new features or by applying boosting algorithms. This article benefitted by both ways; a new similarity measure is proposed for applying in feature sets, and GBM classification method is applied to construct boosted predictive model.

Gradient boosting machine [46] is a machine learning technique for classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in an iterative fashion like other boosting methods do.

The GBM R package is an implementation of extensions to Freund and Schapire's AdaBoost algorithm and Friedman's gradient boosting machine.

There is a need for another attribute which would indicate the class of each record of the dataset. For this purpose, in the preprocessing stage the extra attribute which represents the class of each record is added to the dataset. The accessibility of only positive class instances of the dataset is an issue of concern in the link prediction of social networks which should be

Table 2
Prediction performance of *lda* classifier for Highschool dataset, learnt through similarity measures individually.

Similarity measure	Accuracy	AUC
Triadic similarity	0.7285	0.7923
Common neighbors	0.6785	0.7561
Jaccard's coefficient	0.7	0.7651
Adamic adar	0.7	0.7606
Preferential attachment	0.6428	0.6784

dealt with. To deal with this issue a weighted sample of negative instances of the network is added to the dataset according to the ratio of the positive instances.

The weights of edges and the identifiers of their origins and destinations are not proper as classification features. Therefore, these attributes are omitted from the dataset. It should be noted that any network dynamics could be applied as training attributes. In this article, the out-degree of edge origins and in-degree of destinations are applied. Sectioning the dataset consist of 80% as the training set and 20% as the testing set. Resampling method of cross validation of classifiers is run in 3 iterations.

Since the link prediction problem is a classic prediction task, two very well-known evaluation metrics are applied to measure prediction performance. The first one is *Accuracy* of the classifier which is the percentage of correct predictions, and the second one is *AUC* measure which is the area under the ROC curve [47]. A ROC curve represents the tradeoffs between true positive and false positive of the classic confusion matrix of the classifier.

4.2.1. Class imbalance

Link prediction problem in social networks suffer from extreme imbalance [13]. That is, the dataset distribution reflects a significant majority of the negative class and a minority positive class [48]. The edges of social network graphs are positive instances which are very rare in comparison with not existent edges.

While unsupervised methods cannot combat this imbalance because they are agnostic to class distributions by definition, supervised learning schemes are able to balance data and focus on class boundaries. It is proofed that by proper sampling of data and applying ensemble methods this problem is conquerable [13,48]. Here, the weighted sampling of data and GBM classification method is applied.

4.3. Baseline similarity measures

The proposed similarity measure can be categorized as one of the neighborhood based similarity measures. Therefore, it has to be compared with other well-known measures of this category. There exist many similarity measures applied in the link prediction field. Some of the well-known similarity measures are selected to be compared with the Triadic Similarity measure. These similarity measures are commonly used as baseline to test capabilities of new similarity measures [11,49]. The selected similarity measures in this study are as follows:

Common Neighbors. Let $\Gamma(x)$ represent the set of neighbors of node x in social network. For two nodes, x and y, the number of common neighbors is defined as [11]:

$$CN(x, y) = |\Gamma(x) \cap \Gamma(y)| . \tag{2}$$

Jaccard's Coefficient. This measure is computed by the probability of both x and y having common neighbors [11], Eq. (3).

$$JS(x, y) = |\Gamma(x) \cap \Gamma(y)| / |\Gamma(x) \cup \Gamma(y)| . \tag{3}$$

Adamic Adar. This similarity measure weighs the rare common neighbors more heavily [15]. It is defined as below.

$$AA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log\, |\Gamma(z)|}. \tag{4}$$

Preferential Attachment. This similarity measure is defined as the product of number of neighbors of vertices [36], Eq. (5).

$$PA(x, y) = |\Gamma(x)| . |\Gamma(y)| . \tag{5}$$

5. Results and discussions

The first section of experiments is dedicated to comparison of proposed measure with other baseline similarity measures, individually. An experiment is conducted which applies many *lda* classifiers, each trained with one of the experiments similarity measures. A simple comparison among classification models prediction results trained through these measures is

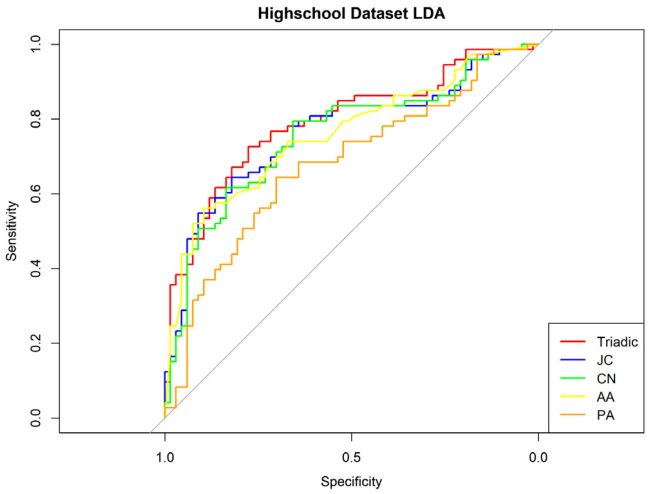

Fig. 4. AUC comparison of *lda* trained through each one of the similarity measures in highschool dataset.

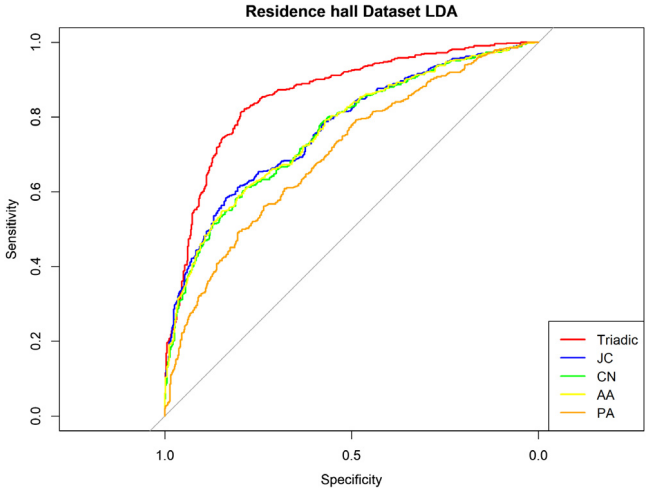

Fig. 5. AUC comparison of *lda* trained through each one of the similarity measures in Residence hall dataset.

Table 3
Prediction performance of *lda* classifier for Residence hall dataset, learnt through similarity measures individually.

Similarity measure	Accuracy	AUC
Triadic similarity	0.7926	0.8589
Common neighbors	0.6779	0.7605
Jaccard's coefficient	0.6981	0.7684
Adamic adar	0.6856	0.7628
Preferential attachment	0.6335	0.7006

applied to test which measure contribute higher to prediction task. It is common practice to normalize classifiers results by applying some intrinsic attributes. Therefore, each *lda* classifier's feature set contains intrinsic attributes too.

The prediction results obtained through *lda* classification model trained through different similarity measures for Highschool dataset are tabulated in Table 2 and illustrated in Fig. 4, respectively. This proposed similarity measure indicate better prediction performance compared to corresponding baseline similarity measures. This proposed triadic similarity measure has better *accuracy* and *AUC* than its counterparts. As *lda* learning phase is done through a linear fashion it can be concluded that triadic similarity measure is a better neighborhood based similarity measure.

For more accurate results, the first experiment is repeated with other datasets. The prediction results of learning *lda* classifier with different similarity measures are listed in Table 3 and *ROC* curves are shown in Fig. 5. Similar to other

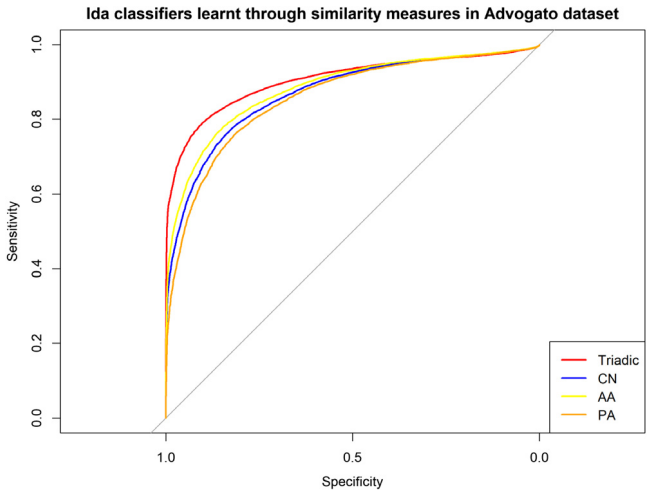

Fig. 6. AUC comparison of *lda* trained through each one of the similarity measures in Advogato dataset.

Table 4
Prediction performance of *lda* classifier for Advogato dataset, learnt through similarity measures individually.

Similarity measure	Accuracy	AUC
Triadic similarity	0.8436	0.905
Common neighbors	0.7937	0.8703
Jaccard's coefficient	0.8236	08958
Adamic adar	0.8093	0.8828
Preferential attachment	0.7812	0.8572

Table 5
Prediction performance of *gbm* classifier for Highschool dataset, trained through each one of the similarity measures.

Similarity measure	Accuracy	AUC
Triadic similarity	0.72	0.7876
Common neighbors	0.6714	0.7748
Jaccard's coefficient	0.7071	0.7838
Adamic adar	0.66	0.7554
Preferential attachment	0.6214	0.6779

experiments, model trained with this proposed similarity measure shows better prediction performance. Both of the area under the *ROC* curve and the accuracy related to the triadic similarity measure are higher, that is, it has a better prediction capability.

The final part of the first experiment consist of training the *lda* classification model by the similarity measures for the Advogato dataset. The Advogato dataset contain much bigger population of vertices and can estimate capability of proposed method in real world social networks. The prediction results for different similarity measures are tabulated in Table 4. The prediction result of the model learnt through proposed measure outperform others. The ROC curves of classifiers are graphically shown in Fig. 6. The results of this part indicate that the proposed measure is applicable in real world social networks.

In order to boost prediction results another classification method is applied. As discussed earlier the gradient boosting is applied in second part of experiments to confront class imbalance and boost results. The prediction results of using the *gbm* classifier trained with each one of the similarity measures in Highschool dataset are tabulated in Table 5. The results are somehow similar to that of the prior experiment. The classifier learnt through proposed similarity measure indicate better prediction performance compared to corresponding baseline similarity measures. Comparison of *ROC* curves of these prediction models is shown in Fig. 7. The related *ROC* curve to this proposed has higher *AUC*, that is, it has a better prediction capability. In comparison with prediction results of applying *lda* classifier tabulated in Table 2; it can be seen that the *gbm* classifier does not make much change. It can be concluded that applying *gbm* classifier with just one similarity measure in feature set does not improve prediction capability.

To extend this experiment the same classifier is trained with the similarity measure on Residence hall dataset. Prediction results and related *ROC* curves are shown in Table 6 and Fig. 8. Clearly the prediction model trained with proposed similarity

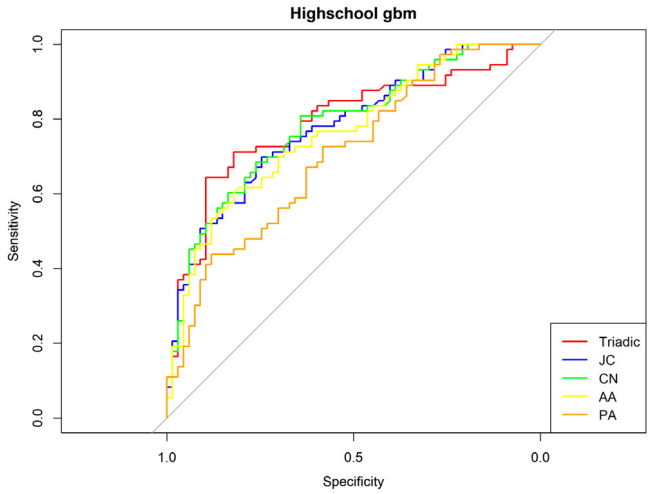

Fig. 7. AUC comparison of *gbm* trained through each one of the similarity measures in Highschool dataset.

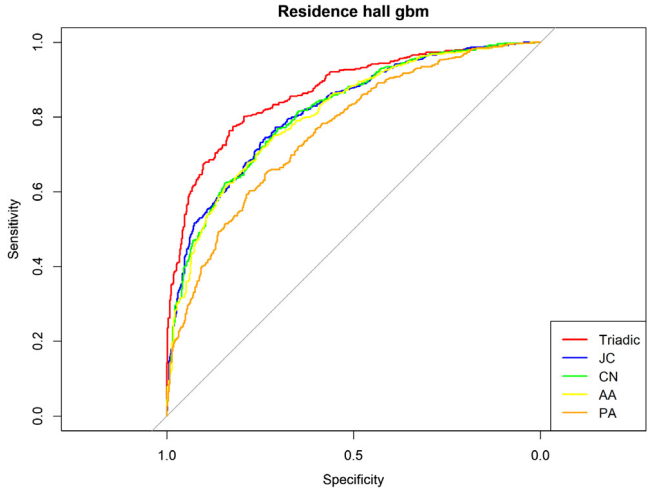

Fig. 8. AUC comparison of *gbm* trained through each one of the similarity measures in Residence hall dataset.

Table 6
Prediction performance of *gbm* classifier for Residence hall dataset, trained through each one of the similarity measures.

Similarity measure	Accuracy	AUC
Triadic similarity	0.7859	0.86
Common neighbors	0.727	0.8101
Jaccard's coefficient	0.7348	0.8167
Adamic adar	0.7357	0.8082
Preferential attachment	0.6788	0.7607

measure outperforms others of its kind. Based on the experiments run in this article, it can be deduced that there exists a direct relation the prediction performances of classification models and the greater volume of the dataset. But, in comparison with prediction results of *lda* classifier on Residence hall dataset tabulated in Table 3; the prediction results are not improved.

The *gbm* classifier learnt through different similarity measures for Advogato dataset. The prediction results shows interesting points. The prediction results and related *ROC* curves are presented in Table 7 and Fig. 9. The model trained with proposed similarity measure outperforms others. In comparison of applying *lda* classifier results tabulated in Table 4, the prediction results improved. It can be concluded that the boosting algorithms are more effective with bigger datasets.

Table 7
Prediction performance of *gbm* classifier for Advogato dataset trained through each one of the similarity measures.

Similarity measure	Accuracy	AUC
Triadic similarity	0.9262	0.9832
Common neighbors	0.9033	0.9661
Jaccard's coefficient	0.8509	0.8861
Adamic adar	0.8964	0.9631
Preferential attachment	0.8706	0.948

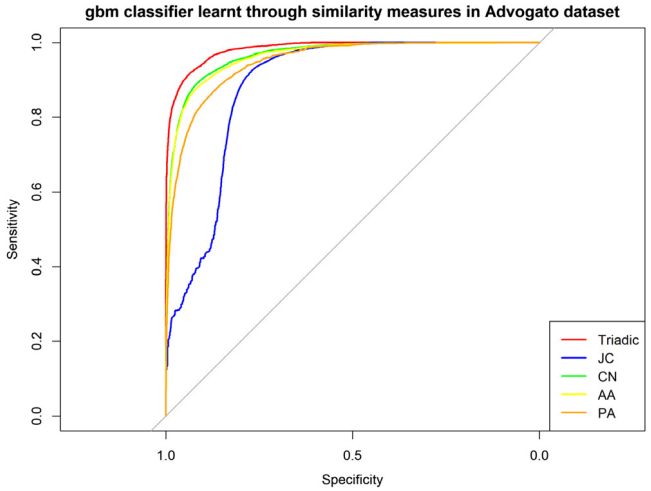

Fig. 9. AUC comparison of *gbm* trained through each one of the similarity measures in Advogato dataset.

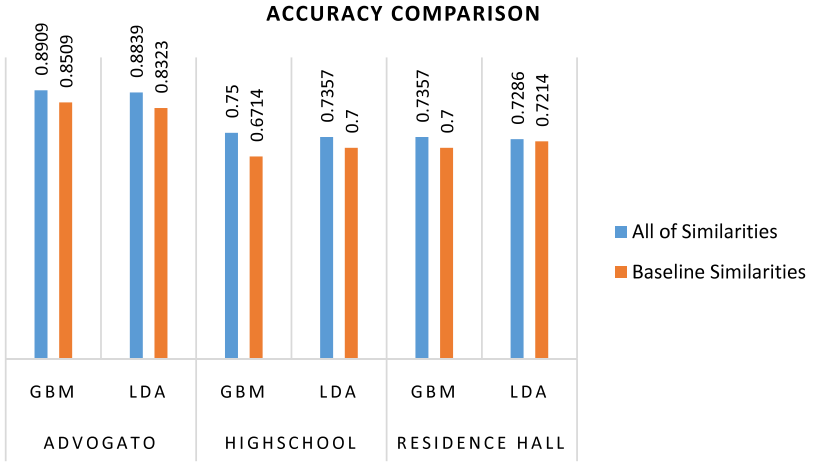

Fig. 10. Accuracy comparison of classification models when learnt though features sets with triadic similarity measure and without.

The third part of experiments is related to studying the impact of proposed measure on prediction results. The prediction model is once learnt through all similarity measures and once more by adding the triadic similarity measures. The classification models which contains proposed measure in their feature sets have better prediction capabilities. The comparison of accuracy and AUCs of these classification models are presented in Figs. 10 and 11. These results show that applying triadic similarity measure in the link prediction would enhance prediction capability.

In link prediction problem very large training sets of millions of instances are common. Most often, the training data will not fit in memory. Therefore, classification model construction may become inefficient. More scalable approaches, capable of handling training data that are too large to fit in memory, are required. Earlier approaches included discretizing continuous-valued attributes and sampling data. More operational and modern approaches included using Random Forest framework

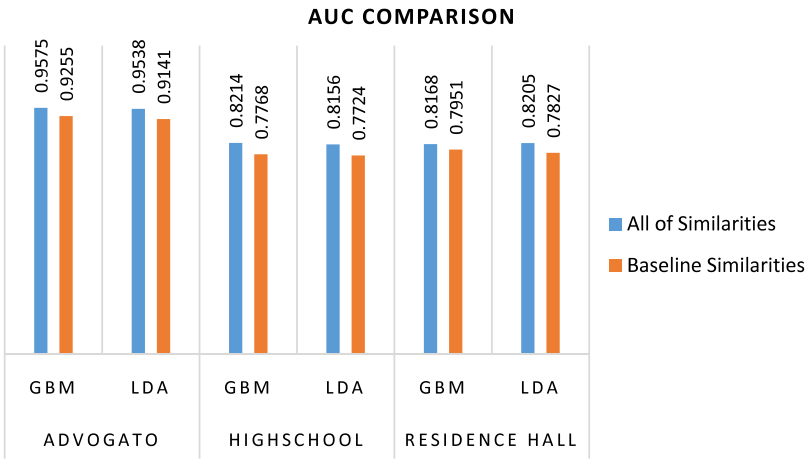

Fig. 11. AUC comparison of classification models when learnt though features sets with triadic similarity measure and without.

and BOAT algorithm [48]. The scalability of supervised learning in social networks is not an issue anymore, due to mentioned approaches.

6. Conclusion

One of the main goals of this experiment was to attempt to propose a new similarity measure that uses network motifs as an input. This article seek to answer some questions. Are network motifs applicable for estimating network users' similarities? Is such a new similarity measure usable in the link prediction problem? How this similarity measure works in the link prediction task in comparison to other baseline similarity measures?

Using similarity measures to predict occurrence probability of future interactions is one of the most accepted methods in link prediction problem. This proposed similarity measure applies network motifs in a new neighborhood based similarity measure calculation. Therefore, its prediction results should be compared with other well-known neighborhood based similarity measures.

A supervised learning experiment framework is applied in this study to answer the research questions. The experiment framework is applied variously to answer particular questions. Once the LDA classifier is applied to compare proposed measure with other baseline similarity measures. The GBM classifier is applied to investigate the impact of boosting methods on prediction results and its capability to confront imbalance problem. These classifiers are learnt through combinations of similarity measures to study the impact of proposed measure when it is considered in classification feature sets. In each section of this experiment, the classifiers are repeatedly trained according to the attributes of either of benchmark datasets. These classifiers prediction performance are evaluated through two reputable metrics; accuracy and AUC.

The results obtained through these experiments indicate that the triadic similarity measure has better performance than other similarity measures, individually. These experiments apply different similarity metrics of pair of nodes as attributes in training classifiers, which then become learned classification models in order to predict occurrence of edges in test set of dataset. These learned classification models are compared with one another and it is deduced that this model trained based on Triadic Similarity measure outperforms all tested models in this study.

It was also observed that applying the proposed similarity measure with other similarities in classification feature set, would significantly improve classifiers prediction capability. This observation strongly proves that local structures in social networks are a good source knowledge for similarity estimation.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

References

[1] J. Leskovec, L. Backstrom, R. Kumar, A. Tomkins, Microscopic evolution of social networks, in: Proceeding 14th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD 08, ACM Press, New York, New York, USA, 2008, p. 462. http://dx.doi.org/10.1145/1401890.1401948.
[2] Y. Dong, J. Tang, S. Wu, J. Tian, N.V. Chawla, J. Rao, H. Cao, Link prediction and recommendation across heterogeneous social networks, in: Data Min. (ICDM), 2012 IEEE 12th Int. Conf., 2012, pp. 181–190.
[3] S. Milgram, The small-world problem, Psychol. Today 1 (1967) 61–67.
[4] J.M. Kleinberg, Navigation in a small world, Nature 406 (2000) 845.

[5] A.-L. Barabâsi, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, T. Vicsek, Evolution of the social network of scientific collaborations, Physica A 311 (2002) 590–614.
[6] R. Jiang, Z. Tu, T. Chen, F. Sun, Network motif identification in stochastic networks, Proc. Natl. Acad. Sci. 103 (2006) 9404–9409.
[7] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Network motifs: simple building blocks of complex networks, Science 298 (2002) 824–827 (80).
[8] S. Leinhardt, Local structure in social networks, Sociol. Methodol. 7 (1976) 1–45.
[9] J.A. Grochow, M. Kellis, Network motif discovery using subgraph enumeration and symmetry-breaking, in: Res. Comput. Mol. Biol., 2007, pp. 92–106.
[10] V. Lacroix, C.G. Fernandes, M.-F. Sagot, Motif search in graphs: application to metabolic networks, Comput. Biol. Bioinformatics, IEEE/ACM Trans. 3 (2006) 360–368.
[11] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, J. Am. Soc. Inf. Sci. Technol. 58 (2007) 1019–1031. http://onlinelibrary.wiley.com/doi/10.1002/asi.20591/full (accessed 03.05.13).
[12] M. Al Hasan, V. Chaoji, S. Salem, M. Zaki, Link prediction using supervised learning, in: SDM'06 Work. Link Anal. Counter-Terrorism Secur., 2006.
[13] R.N. Lichtenwalter, J.T. Lussier, N.V. Chawla, New perspectives and methods in link prediction, in: Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2010, pp. 243–252.
[14] A. Papadimitriou, P. Symeonidis, Y. Manolopoulos, Fast and accurate link prediction in social networking systems, J. Syst. Softw. 85 (2012) 2119–2132. http://dx.doi.org/10.1016/j.jss.2012.04.019.
[15] L.A. Adamic, E. Adar, Friends and neighbors on the web, Soc. Networks 25 (2003) 211–230.
[16] H.R. De Sá, R.B.C. Prudêncio, Supervised link prediction in weighted networks, in: Neural Networks (IJCNN), 2011 Int. Jt. Conf., 2011, pp. 2281–2288.
[17] Z. Lu, B. Savas, W. Tang, I.S. Dhillon, Supervised link prediction using multiple sources, in: 2010 IEEE Int. Conf. Data Min., 2010, pp. 923–928.
[18] S. Mangan, U. Alon, Structure and function of the feed-forward loop network motif, Proc. Natl. Acad. Sci. 100 (2003) 11980–11985.
[19] J. Camacho, D.B. Stouffer, L.A.N. Amaral, Quantitative analysis of the local structure of food webs, J. Theoret. Biol. 246 (2007) 260–268.
[20] K. Juszczyszyn, P. Kazienko, K. Musiał, Local topology of social network based on motif analysis, in: Knowledge-Based Intell. Inf. Eng. Syst., 2008, pp. 97–105.
[21] K. Juszczyszyn, K. Musial, M. Budka, Link prediction based on subgraph evolution in dynamic social networks, in: 2011 IEEE Third Int'l Conf. Privacy, Secur. Risk Trust 2011 IEEE Third Int'l Conf. Soc. Comput., IEEE, 2011, pp. 27–34. http://dx.doi.org/10.1109/PASSAT/SocialCom.2011.15.
[22] J. Ugander, L. Backstrom, M. Park, J. Kleinberg, Subgraph Frequencies : Mapping the Empirical and Extremal Geography of Large Graph Collections, 2013.
[23] T. Opsahl, Triadic closure in two-mode networks: Redefining the global and local clustering coefficients, Soc. Networks 35 (2013) 159–167.
[24] P. Wang, J. Lui, B. Ribeiro, D. Towsley, J. Zhao, X. Guan, Efficiently estimating motif statistics of large networks, ACM Trans. Knowl. Discov. Data 9 (2014) 8.
[25] H.H. Song, T.W. Cho, V. Dave, Y. Zhang, L. Qiu, Scalable proximity estimation and link prediction in online social networks, in: Proc. 9th ACM SIGCOMM Conf. Internet Meas. Conf. - IMC '09, ACM Press, New York, New York, USA, 2009, p. 322. http://dx.doi.org/10.1145/1644893.1644932.
[26] I. Esslimani, A. Brun, A. Boyer, Densifying a behavioral recommender system by social networks link prediction methods, Soc. Netw. Anal. Min. 1 (2011) 159–172. http://dx.doi.org/10.1007/s13278-010-0004-6.
[27] M. Slokom, R. Ayachi, A new social recommender system based on link prediction across heterogeneous networks, in: Int. Conf. Intell. Decis. Technol., 2017, pp. 330–340.
[28] T. Tylenda, R. Angelova, S. Bedathur, Towards time-aware link prediction in evolving social networks, in: Proc. 3rd Work. Soc. Netw. Min. Anal. - SNA-KDD '09, ACM Press, New York, New York, USA, 2009, pp. 1–10. http://dx.doi.org/10.1145/1731011.1731020.
[29] Y. Yasami, F. Safaei, A novel multilayer model for missing link prediction and future link forecasting in dynamic complex networks, Physica A 492 (2018) 2166–2197.
[30] P. Wang, B. Xu, Y. Wu, X. Zhou, Link prediction in social networks: the state-of-the-art, Sci. China Inf. Sci. 58 (2015) 1–38.
[31] C. Fan, Z. Liu, X. Lu, B. Xiu, Q. Chen, An efficient link prediction index for complex military organization, Physica A 469 (2017) 572–587.
[32] K. Shang, M. Small, W. Yan, Link direction for link prediction, Physica A 469 (2017) 767–776.
[33] M. Gao, L. Chen, B. Li, Y. Li, W. Liu, Y. Xu, Projection-based link prediction in a bipartite network, Inf. Sci. (Ny). 376 (2017) 158–171.
[34] Y. Yang, J. Zhang, X. Zhu, L. Tian, Link prediction via significant influence, Physica A 492 (2018) 1523–1530.
[35] L. Lü, C.-H. Jin, T. Zhou, Similarity index based on local paths for link prediction of complex networks, Phys. Rev. E 80 (2009) 46122.
[36] M.E.J. Newman, Clustering and preferential attachment in growing networks, Phys. Rev. E 64 (2001) 25102.
[37] F. Gao, K. Musial, C. Cooper, S. Tsoka, Link prediction methods and their accuracy for different social networks and network metrics, Sci. Program. 2015 (2015) 1.
[38] R Core Team, R: A Language and Environment for Statistical Computing, 2015. https://www.r-project.org/.
[39] G. Csardi, T. Nepusz, The igraph software package for complex network research, Int. J. Complex Syst. (2010) 1695 http://igraph.org.
[40] M.K.C. from Jed Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, the R Core Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, caret: Classification and Regression Training, 2015. http://cran.r-project.org/package=caret.
[41] J.S. Coleman, et al., Introduction to Mathematical Sociology, London Free Press Glencoe, 1964.
[42] L.C. Freeman, C.M. Webster, D.M. Kirke, Exploring social structure using dynamic three-dimensional color images, Soc. Networks 20 (1998) 109–118.
[43] P. Massa, M. Salvetti, D. Tomasoni, Bowling alone and trust decline in social network sites, in: Dependable, Auton. Secur. Comput. 2009. DASC'09. Eighth IEEE Int. Conf., 2009, pp. 658–663.
[44] M.A. Brandão, M.M. Moro, Social professional networks: A survey and taxonomy, Comput. Commun. 100 (2017) 20–31.
[45] A.J. Izenman, Linear discriminant analysis, in: Mod. Multivar. Stat. Tech., Springer, 2013, pp. 237–280.
[46] J.H. Friedman, Stochastic gradient boosting, Comput. Statist. Data Anal. 38 (2002) 367–378.
[47] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (2006) 861–874.
[48] J. Han, J. Pei, M. Kamber, Data Mining: Concepts and Techniques, Elsevier, 2011.
[49] T. Zhou, L. Lü, Y.-C. Zhang, Predicting missing links via local information, Eur. Phys. J. B 71 (2009) 623–630.