2019

Exploratory factor analysis of graphical features for link prediction in social networks

Lale Madahali
University of Nebraska at Omaha, lmadahali@unomaha.edu

Lotfi Najjar
University of Nebraska at Omaha, lnajjar@unomaha.edu

Margeret Hall
University of Nebraska at Omaha, mahall@unomaha.edu

Recommended Citation

Madahali L., Najjar L., Hall M. (2019) Exploratory Factor Analysis of Graphical Features for Link Prediction in Social Networks. In: Cornelius S., Granell Martorell C., Gómez-Gardeñes J., Gonçalves B. (eds) Complex Networks X. CompleNet 2019. Springer Proceedings in Complexity. Springer, Cham. https://doi.org/10.1007/978-3-030-14459-3_2

Exploratory factor analysis of graphical features for link prediction in social networks

Lale Madahali[1*], Lotfi Najjar[1], Margeret Hall[2]

[1]PhD student at University of
Nebraska at Omaha, Omaha, United
States
lmadahali@unomaha.edu
[2]Associate Professor at University of
Nebraska at Omaha, Omaha, United
States
lnajjar@unomaha.edu

[3]Assistant Professor at University of
Nebraska at Omaha, Omaha, United
States
mahall@unomaha.edu

Abstract. Social Networks attract much attention due to their ability to replicate social interactions at scale. Link prediction, or the assessment of which unconnected nodes are likely to connect in the future, is an interesting but non-trivial research area. Three approaches exist to deal with the link prediction problem: feature-based models, Bayesian probabilistic models, probabilistic relational models. In feature-based methods, graphical features are extracted and used for classification. Usually, these features are subdivided into three feature groups based on their formula. Some formulas are extracted based on neighborhood graph traverse. Accordingly, there exists three groups of features, neighborhood features, path-based features, node-based features. In this paper, we attempt to validate the underlying structure of topological features used in feature-based link prediction. The results of our analysis indicate differing results from the prevailing grouping of these features, which indicates that current literatures' classification of feature groups should be redefined. Thus, the contribution of this work is exploring the factor loading of graphical features in link prediction in social networks. To the best of our knowledge, there is no prior studies had addressed it.

Keywords: Social networks analysis, Exploratory factor analysis, Data mining

1 Introduction

A social network is a social structure that is composed of a set of actors (also known as players, agents or nodes) and a set of the relationships between these actors. It can be represented as a graph in which nodes (vertices) represent people (actors) and edges represent relationships between them. Link prediction, or the prediction of future connections of unconnected nodes in a graph [1], has many applications within and outside of the domain of social networks, e.g., finding interactions between proteins in bioinformatics [2]; in e-commerce and recommender systems [3]; and in the defense and security domains in identifying terrorist cells [4].

Traditionally there are three approaches for addressing link prediction problem: feature-based link prediction, Bayesian probabilistic models, and probabilistic relational models. In feature-based link prediction, the problem is seen as a supervised classification problem in which each record corresponds to a relationship. In Bayesian probabilistic models, the main idea is assessing a probability score denoting the existence of a future relationship between two nodes that are not connected. This score can also be used as a feature in classification. Probabilistic relational models (PRM) can incorporate attributes of edges and vertices to create the combined probability distribution of a set of nodes and edges. There are two approaches of PRM, Bayesian networks based which is used for directed links and relational Markov networks based, which is used for undirected links [5].

In most feature-based link prediction studies, researchers categorize graphical features as neighborhood features, node-based features, and path-based features [5]. The general aspects considered in each category are as follows:

- neighborhood: common neighbors, Jaccard coefficient, Adamic/Adar
- path-based: shortest path count, Katz, hitting time, rooted pagerank
- node-based: preferential attachment, clustering coefficient, simrank

The features' formulae are elaborated later. There are different ways researchers approach link prediction problems. In some studies, researchers investigate the structure of the networks to try to introduce new graphical features for improving prediction quality [6]. Other authors work on Machine learning techniques. In these studies, algorithms are created or manipulated to enhance performance, since supervised link prediction culminates in a classification problem [7].

Literature to date focusses on introducing new features and manipulating algorithms to increase performance. Our estimation and the present foundational work considers that current research has not considered the underlying real structure of these features in prediction, and whether all these features are needed in and for prediction. This is akin to the curse of dimensionality problem in machine learning [8], and leads us to position that this under-consideration by the research community has caused structural biases in existing analyses. This is what we exam in this paper. Factor analysis is a useful tool for probing relationships between variables. By collapsing a large number of variables into a few understandable factors, it allows investigating concepts that are not easily measured directly. It is also computationally efficient, well-benchmarked in literature, and follows the premise that simple solutions are preferable to complex solutions in the case of similar or the same results. Factor analysis is employed to analyze the relationship between factors and to see how they can be loaded under underlying factors.

We investigate which features load under each produced factor, and whether they deviate from expectation. To the extent of our knowledge, no prior studies have completed a factor analysis of topological features in link prediction problems. Until now it has been conceptually assumed that the attributes must go together. Until now, literature has not shown that these features must belong together mathematically. This paper tries to mathematically show if the features belong together. Therefore, the research question in this study is "Considering graphical features in a social network graph for the link prediction, what is the structure of retained factors for further analysis?"

Features are extracted from the Hep-ph dataset which is a coauthorship network from 1992 to 2000. It has 16,402 nodes and 156,742 edges where nodes are authors and edges are their relationship based on some type of collaboration, like publishing a paper or working on a project together. The related work, methodology, experimental setup, results, implications, and conclusion are described below.

2. Related work

2.1 Link Prediction

As stated earlier, one of the favored approaches to link prediction is feature-based link prediction. In their study Madahali et al., [7] used data mining techniques to improve the performance of machine learning algorithms. Applying preprocessing techniques to the data and combining algorithms, they came up with performance improvement in terms of F-measure and AUC (Area Under the Curve). Liben-Nowell et al., [1] extracted and worked with graphical features considering a core set of co-authors who have collaborated at least three times during the train and test interval. They considered each of these features as predictors and then compared their performance with a random predictor. Their result is a list of links with associated probabilities. Graph-based features are the most common features used in feature-based link prediction [5]. Cukierski et al., [9] extracted 94 distinct graph features and used them as their classification input along with using Random Forests. Furthermore, they mentioned big data problem in large graphs. They came up with this conclusion that the best classification performance achieved through a combination of different categories of features, as they show different aspects of the graph structure. They reported the area under the curve of 0.9695.

Introducing new futures for improvement is a common approach to contribute in this area. In [6] the authors define two new features, friends-measure, and same-community. Their definition is as follows: friends-measure is the number of connections between two nodes' neighbors; same-community determines whether two nodes are members of a common community. They ran their experiments on 10 datasets and showed their improved performance. In addition to topological features, other features such as node and edge features can contribute to improvement. [4] introduced new features like keyword match count or the sum of papers in coauthorship networks.

Scellato et al., [10] used Gowalla which is a location-based social network. They introduced a new feature called "place-friends" which determines which users visited the same place. They defined new features helpful for prediction relying on the properties of the places visited by users. Finally, through a supervised link prediction framework based on these features, they established their prediction.

The excessive change in social networks makes them extremely dynamic; millions of nodes and edges are added and deleted in a day. In order to deal with this challenge, Song et al., [11] introduced proximity measures with an algorithm to estimate proximity in very dynamic networks. Proximity measures show how far or close nodes are in a social network. Measuring proximity forms a range of applications in the social sciences, business, information technology, computer networks, and cybersecurity. Defining various proximity measures, they found first the effectiveness of using different proximity measures varies among networks. Second, considering several proximity measures contribute to better accuracy.

Link prediction problem is sometimes formulated as a random walk from a source node to the target. Supervised random walks can be applied to many problems that require learning to rank nodes in a graph, link recommendations, anomaly detection, missing link, and searching and ranking [12]. They provided a random walk solution on Facebook for a learning function which assigns a score to each edge, in turn, the algorithm could see nodes that are more probable to be connected. They found that their algorithm outperforms unsupervised approaches and feature-based prediction in terms of AUC. Using game theory, Zappella et al., [13] introduced a new approach based on Graph Transduction Game. Using dataset from Tuenti social network, they proved that their method excels standard local measures and also significant enhancement in terms of mean average precision and reciprocal rank.

2.2 Applied Factor Analysis

Generally speaking, the goal of factor analysis is to reduce the dimensionality of the original space and ends in a new space. Its goal is to classify intercorrelated variables under more general variables [14]. Therefore, factor analysis brings about two advantages: the possibility of gaining a clear view of the data and the possibility of using the output in the subsequent analysis [15],[16] by reducing the space of the feature vector. Many studies give a thorough theoretical overview of factor analysis [15,16]. Factor analysis plays a major role in improving psychological researches [17]. It has been used as an analytic technique for extracting interrelationship patterns, reducing data, classifying and describing data, data transformation, hypothesis testing, and mapping construct space [16]. The journal Psychometrika (the primary journal of the Psychometric Society, a professional body devoted to psychometrics and quantitative psychology) has devoted more pages to factor analysis than to any other quantitative topic in the behavioral science [18], and the number of studies applying factor analysis has experienced a dramatic increase [19].

As stated, EFA is usually used in psychology and recognizing influential contributing factors. In [20] the authors measure personality and trait affect to see how various psychological factors contribute to the degree and nature of posting status updates on Facebook. They introduced an instrument to measure two types of status updates on Facebook, positive and negative content. Using this instrument along with instruments that measure personality and trait effect, they decided how much different psychological factors contribute to determining the degree and nature of posting status updates on Facebook. They used Pearson correlations and partial-least-squares structural equation modeling as their statistical techniques. In turn, they found support for the role of personality and trait affect in understanding the types of people that post status updates on Facebook.

In [21] the authors try to answer two questions at the intersection of psychometrics, data mining, machine learning, and physics education research. They analyzed a large set of students' responses to an assessment designed to assess strategic knowledge of the Mechanics Reasoning Inventory (MRI). They use EFA to identify the basic mental constructs of students in order to reduce the number of variables necessary to explain the data below the number of items. They mention their reason to use EFA it is the model that shows their theoretical assumption about the relationship between mental constructs and observed items responses since factors are formulated as the cause of the correlated item responses. Secondly, EFA models measurement errors and unique sources of variance in item responses.

2.3 Factor Analysis in Computing

Marsden et al., [22] present the results of an exploratory factor analysis (EFA) in professionals' attitudes toward online communication. They conduct a survey of 100 professionals based on the theoretical approaches to the scientific study of online communication identified by [23]. They found three constructs: 1) media choice, 2) the hyperpersonal options of online communication, and 3) social cues in online communication. These constructs reflect the scientific approaches that can explain the dynamics of online communication. These attitudinal approaches have several advantages. Firstly, it can influence the communication itself and secondly, an impact on the use and the quality of online communication. Thirdly, the results showed attitudes which will be useful for successful online communication.

In [24] the authors developed a scenario such that each scenario contained a single essay, related to identifying privacy, accuracy, property, and access (PAPA), within a context, and with the respondent as the individual encountering the dilemma. According to Mason [25], these are the four main ethical issues of the information age. Violation of each issue shows the injured participant's perspective to identify consequences coming from the use of information and information technology in an unethical way. They carried out two pilot tests to refine the wording of the 16 scenarios used in the final questionnaire [24]. The survey results showed that people have high egocentricity and concern for themselves; few are concerned with or aware of other stakeholders. [24] hypothesize that the lack of awareness of the stakeholder is the result of the mediation of technology which creates a gap between computer user and stakeholder. Consequently, this psychological distance may explain the growing rate of unethical computer usages, such as break-ins and viruses.

Researchers claim that there is value in studying social networks simultaneously and the advantage of social network use will be clear when multiple services are reviewed together[26,27]. [27] conducted a study of 198 Facebook users and predicted how it is probable that a Facebook user become a twitter user based on their Facebook usage. They collected twelve activity measures via Facebook API and concluded five discrete usage dimensions. The factor analysis result shows that Facebook usage for participants is multidimensional. This implies that people can be classified along these dimensions and into groups with very different usage styles. They claim that considering usage behavior a multidimensional concept provides a more accurate definition of individual behaviors than a general measure that describes Facebook usage. Therefore, their findings show that distinguished features are drivers for adopting one social network over another. Conversely, if multiple services share the same features, users tend to use those features in only one social network.

EFA can be used for reordering different measurements in various areas. In the study carried out by Rashmi Jha et al., [28] researchers tried to explore and prioritized the factors that influence, control and empower the modern ERP implementation for small and medium enterprises. In order to accomplish this, they examined the latest trend of modern ERP functioning of Delhi-NCR companies through EFA and the reliability of all the constructs that emerged during their work. They concluded that optimizing software engineering, project management and lean six sigma techniques helped in having successful implementation of ERP in small and medium enterprises.

In the paper written by Schreiber et al [29], researchers try to figure out a valid categorization and to examine the performance and properties of a range of h-type indices. Using EFA, they studied the relationship between the h-index, its variants, and some standard bibliometric indicators of 26 physicists from the Science Citation Index in the Web of Science. They showed that for their dataset a distinction is possible to quality and quantity of scientific output.

3. Methodology

In this Section, we explicate the features extracted from the dataset graph then explain the EFA procedure. As mentioned, the Hep-ph dataset and the graphical features in this study are common in link prediction problems. LPmade [30] was used to extract these features.

3.1 Link prediction graphical features

The following are feature formulas used in most link prediction studies.

Adamic/Adar: in the context of web mining, Adamic/Adar measures the similarity between two webpages and determines how two pages are related. It computes the similarity between the two pages [31].

$$\sum_{z:\, common\ features\ between\ a,b} \frac{1}{\log(frequency(z))} \quad (1)$$

Thus, in link prediction this converts to:

$$\sum_{z: common\ neighbors\ between\ a,b} \frac{1}{\log|\Gamma(z)|} \quad (2)$$

Common neighbors: this feature calculates the number of common neighbors between two nodes [1]. Suppose Γ(a) is the set of node a neighbors. Common neighbors between two nodes are defined as follows:

$$|\Gamma(a)\cap \Gamma(b)| \quad (3)$$

Clustering coefficient: illustrate how your friends are friends with each other [32].

$$CC = \frac{3 \times number\ of\ triangles}{number\ of\ connected\ triplets\ of\ vertices} \quad (4)$$

Jaccard Coefficient: it is used for measuring the similarity and diversity. It is the quotient of the intersection of the two neighbor sets and the union of the two neighbor sets [33].

$$J(a,b) = \frac{|a \cap b|}{|a \cup b|} \quad (5)$$

Preferential attachment: is a network formation model. The more the degree of a node, the more probably it will connect to other nodes. The probability of collaboration of nodes a and b is proportional to the product of the number of neighbors they have [34].

$$|\Gamma(a)|.|\Gamma(b)| \quad (6)$$

Katz: it keeps a collection of paths. It assigns higher weights to shorter paths. This weight is calculated by graph traversing rather than matrix operation. Its first parameter is a maximum distance away from the source and the second parameter is the damping factor β.

$$\sum_{l=1}^{\infty} \beta^l . |paths_{a,b}^{<l>}| \quad (7)$$

Where $paths_{,b}^{<l>}$is the set of all l-length paths from a to b, and β>0 is a parameter of predictor. Katz have two variants (1) unweighted, in which $paths_{a,b}^{<l>} = 1$ if a and b have cooperated and 0 otherwise (2) weighted, in which $paths_{a,b}^{<l>}$is the number of times that a and b have cooperated [35].
Prop flow: according to [36] this algorithm runs breath first search algorithm that receives a maximum distance to explore the network before terminating.
Rooted pagerank: pagerank is a commonly used algorithm in Web Mining, and specifically, Web Structure Mining. It is an indicator of the importance of a page [37, 38]. Rooted pagerank is pagerank algorithm in which random walk has a restart parameter α [1]. A simple version of pagerank is as follows:

$$PR(u) = (1-d) + d \sum_{v \in B(u)} \frac{PR(v)}{N_v} \qquad (8)$$

Where d is the probability of the users' following the direct links. (1-d) is the page rank distribution from pages that are not linked directly. It is called damping factor and is frequently set to 0.85 [39].

Simrank: a recursive algorithm which implies the similarity of two nodes is proportional to their similar neighbors [1]. The starting point is similarity(a,a):=1

$$similarity(a,b) := \gamma . \frac{\sum_{x \epsilon L(a)} \sum_{y \epsilon L(b)} similarity(x,y)}{|\Gamma(a)|.|\Gamma(b)|} \quad \gamma \in [0,1] \qquad (9)$$

Shortest Path Count: this feature calculates the number of shortest paths between two nodes. This means that it executes a breadth first search terminating at the level at which the target is found and counts the number of times the target is encountered at that level [7].
Idegree: the degree of the target node.
Jdegree: the degree of the source node.

3.2 Factor analysis

Factor analysis goal is to classify intercorrelated variables together under more general variables [40]. Factor analysis starts with a correlation matrix, in which the interrelations between variables are shown. The goal is to classify variables that highly correlate with a group of other variables under one variable. These variables should have low correlation with variables outside of that group [15]. All measured variables are related to a latent factor. The resultant factors show a new dimension that visualizes classification axes along which measurement variables can be plotted [15]. This projection of the scores of the original variables on the factor show two different information: factor scores and factor loadings. Factor scores are the scores of a variable (feature) on a factor [14]. Factor loadings are the correlation of the original variables with a factor. The factor scores can be used as new scores in multiple regression analysis. On the other side, factor loadings shows the importance of a particular variable to a factor [15]. This information is used for interpreting and naming the factors.

Measurements: As factor analysis starts with a correlation matrix, variables should at least be at an interval level. Normality is not required in EFA except in cases of statistical tests for significance of factors. The sample size is important as correlations are not resistant [42] and can affect reliability [15]. According to [15] there are a host of studies explaining the necessary sample size for factor analysis leading to many 'rules-of-thumb'. The general conclusion of these studies is that the most important factors in a reliable factor analysis are the absolute sample size and the absolute magnitude of factor loadings [15]. The Kaiser-Meyer-Olkin measure of sampling adequacy (KMO) can be employed as a measure of sample size adequacy. If the KMO value is greater than 0.5, the sample is adequate. Additionally, the Anti-image matrix of covariances and correlations can be employed. In this matrix, the sample size is adequate if all elements on the diagonal of this matrix are greater than 0.5 [15].

Correlation matrix: Regarding the correlation matrix, there are two important points. The variables have to be intercorrelated, but not too highly correlated. This will make recognizing the unique contribution of the variables to a factor [15]. Bartlett's test of sphericity is used to check the intercorrelation. When the correlation matrix is an identity matrix, there are no correlations between the variables.

The number of factors to be retained: The number of positive eigenvalues of the correlation matrix determines the number of factors to be retained. However, this is not always true, since sometimes some eigenvalues are positive but very close to zero. Thus, for determining the number of retained factors there are some rules of thumb [14,15]. Based on Guttman-Kaiser [40], only keep factors with an eigenvalue larger than 1. Then, there are two ways of selecting which factors to keep. The first one is to keep factors which contribute to 70-80% of the variance. The second way is to generate a scree-plot and keep all factors before the threshold. After factor extraction, we have to check the communalities. The extracted factors account for only a small part of the variance if the communalities are low. Thus, more factors might be retained in order to provide a better variance value.

Factor rotation. Two types of rotation are standard, orthogonal and oblique. The difference is that in the orthogonal rotation there is no correlation between the extracted factors, whereas, in the oblique rotation there is. A straightforward solution to choose the type of rotation is to perform the analysis with the two types of rotation. If the oblique rotation shows a trivial correlation between the extracted factors, then choosing orthogonal rotation makes sense [15].

4. Experimental setup, results, and discussion

SPSS 23.0 is used in this study. For all the experiments, principal component analysis with varimax rotation was carried out to construct the factor structure. The sample size is adequate; the KMO measure of sphericity exceeds the threshold of 0.5. Bartlett's test of sphericity is significant at $p < 0.005$. In the Anti-image matrix, all the correlations are greater than the threshold of 0.5. There were no cross loadings and the cutting point for factor loading was set at 0.5.

For the experiments, we set first set three then four for the retained number of factors. In the 3-component structure Clustering coefficient unexpectedly does not load on any of the components. In the 4-component analysis (*Table 1*), all the Anti-image correlation diagonal values are greater than 0.5. Cumulative variance is 82.360%. The only feature that is loaded under factor 4 is Clustering coefficient with 0.958 factor loading. Values for Cronbach's α for four factors are 0.002, 0.643, 1.000, and 0.000 respectively.

Table 1. Rotated factors, factor loadings individual and cumulative variances

	Component		
	1	2	3
Commonneighbor		.954	
AdamicAdar		.852	
Jaccardcoefficient	.692		
Clusteringcoefficient			
Rootedpagerank	.847		
Katz		.907	
Idegree	-.685		
Jdegree			.969
Preferentialattachment		.578	
Propoflow	.812		
Shortestpathcount			.969
Simrank	.754		
% of Variance	36.077	23.721	14.878
Cumulative Variance	36.077	59.798	74.676
Cronbach's α	-0.002	.018	1.000

As it is shown in *Table 2* Jaccard coefficient, rooted page rank, Idegree, propflow, and simrank are loaded on one factor. While it may be expected that Idegree and Jdegree load on the same factor, they did not. Features common neighbor, AdamicAdar, katz, and preferential attachment loaded under factor 2. Jdegree and shortest path count were loaded under factor 3.

Given the Cronbach's α result, we conducted the experiments again on standardized data for four factors. The Anti-image diagonal values are greater than 0.5. The cumulative variance is 74.676%. Again, Clustering coefficient did not load under any factors. Thus, we removed Clustering coefficient and carried out the factor analysis again. The cumulative variance reached 80.488%.

In the next experiment, we set the number of retained factors at four. The results are shown in Error! Reference source not found.. The only feature loaded under factor 4 was Clustering coefficient. Furthermore, the cumulative variance in this condition reached 82.360%. As stated earlier the value of Cronbach's α for the first factor is 0.596, for the second factor is 0.877, and for the third factor is 1.000. Therefore, it is on average 0.824, which is an acceptable value (more than 0.7). Therefore, data standardization culminated in higher reliability. Figures 1 visualize the common graphical conceptual structure. Whereas *Figure 2* shows the statistical structure of loading each feature.

Table 2. Rotated factors, loadings, individual and cumulative variances for 4 components

	Component			
	1	2	3	4
Commonneighbor		.961		
AdamicAdar		.858		
Jaccardcoefficient	.739			
Clusteringcoefficient				.958
Rootedpagerank	.880			
Katz		.918		
Idegree	-.597			
Jdegree			.970	
Preferentialattachment		.571		
Propoflow	.817			
Shortestpathcount			.970	
simrank	.795			
% of Variance	36.077	23.721	14.878	7.684
Cumulative Variance	36.077	59.798	74.676	82.360
Cronbach's α	-0.002	.643	1.000	.000

neighborhood	Jaccard coefficient
	common neighbors
	Adamic/Adar
path-based	shortest path count
	Katz
	hitting time
	rooted pagerank
node-based	preferential attachment
	clustering coefficient
	simrank

Figure 1. the conceptual structure of feature grouping

factor 1	Jaccard coefficient
	rooted pagerank
	Idegree
	propflow
	simrank
factor 2	common neighbor
	Adamic/Adar
	Katz
	preferntial attachment
factor 3	Jdegree
	shortest path count
factor 4	clustering coefficient

Figure 2. statistical grouping of features

5. Implications

In most link prediction studies, researchers focus on defining new features, creating and manipulating learning algorithms, and manipulating data to deal with the problem. In order to analyze the real relationship and correlation between these features, and real groupings of these features, we ran EFA on many common features in this problem. Based on our analysis, the groupings differ from those employed in the standard literature. These new groupings are advantageous especially when there are a host of features to deal with. It may also help to reduce intercorrelated errors and biases in the analyses. Interestingly, Clusteringcoefficient was the only feature which did not load under any other factors. Future researchers may either remove Clustering coefficient or consider it as a new feature category. In the latter case, it will be interesting to introduce new features similar and related to it. Furthermore, Idegree and Jdegree do not load on the same factor, indicating that they are structurally dissimilar. This has an unknown impact on extant literature which must be validated in future work.

Future researchers may reduce the complexity of their datasets with learning classifiers via EFA to simplify and reduce factors in a way that reflects the structure of the data. This is helpful to reduce model and computational complexity (in order to avoid being overloaded) in terms of time and memory complexity. This simplified approach assists in increasing generalizability in measurement [41].

6. Conclusion and Limitations

In this study, we tried to examine the link prediction problem features by analyzing their relationship and correlation. Until now it was unknown if these groupings are an accurate representation of the underlying structure of the features. This novel approach looks at the problem from a very different perspective. Usually, these graphical features are grouped into three categories: neighborhood, path, and node features. The employed groupings are traditionally based only on their formula. Our results are intriguing in that features grouped in the same category did not load under a common factor in EFA. Further efforts are required to fully empirically validate the underlying structure of the features and its impact on the results of other literature. It is our position that given the structural differences in commonly-used features, current link-prediction research may be subject to unknown structural biases. Correcting these biases may improve the results and prediction quality of current and future works. A limitation of this work can be considering common features but not all possible features from literature. Another limitation is that this paper analyzes a coauthorship network. The results we found may not hold in the case of e.g. a social media network. Further research is required. Another direction could be considering as many features as possible to validate the relationship and factor loadings.

References

1. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. J. Am. Soc. Inf. Sci. Technol. 58, 1019–1031 (2007)

2. E.M. Airoldi, “Mixed membership stochastic block models for relational data with application to protein-protein interactions”, In Proceedings of the international biometrics society annual meeting. 2006, pp. 1–34. - Google Search, https://www.google.com/search?q=E.M.+Airoldi%2C+“Mixed+membership+stochastic+block+models+for+relational+data+with+application+to+protein-protein+interactions”%2C+In+Proceedings+of+the+international+biometrics+society+annual+meeting.+2006%2C+pp.+1–34.&oq=E.M.+Airoldi%2C+“Mixed+membership+stochastic+block+models+for+relational+data+with+application+to+protein-protein+interactions”%2C+In+Proceedings+of+the+international+biometrics+society+annual+meeting.+200

3. Huang, Z.H.Z., Li, X.L.X., Chen, H.C.H.: Link prediction approach to collaborative filtering. Proc. 5th ACM/IEEE-CS Jt. Conf. Digit. Libr. (JCDL ’05). (2005)

4. Al Hasan, M., Chaoji, V., Salem, S., Zaki, M.: Link prediction using supervised learning. In: SDM’06: Workshop on Link Analysis, Counter-terrorism and Security (2006)

5. Al Hasan, M., Zaki, M.J.: A survey of link prediction in social networks. In: Social network data analytics. pp. 243–275. Springer (2011)

6. Fire, M., Tenenboim-Chekina, L., Puzis, R., Lesser, O., Rokach, L., Elovici, Y.: Computationally efficient link prediction in a variety of social networks. ACM Trans. Intell. Syst. Technol. 5, 10 (2013)

7. Madahali, L., Sherkat, E., Hall, M.: A comprehensive study on improving supervised feature based link prediction in social networks. In: 1st North American Social Networks (NASN) conference. , Washington DC (2017)

8. Hastie, T., Tibshirani, R., Friedman, J.: Unsupervised learning. In: The elements of statistical learning. pp. 485–585. Springer (2009)

9. Cukierski, W., Hamner, B., Yang, B.: Graph-based features for supervised link prediction. In: Neural Networks (IJCNN), The 2011 International Joint Conference on. pp. 1237–1244. IEEE (2011)

10. Scellato, S., Noulas, A., Mascolo, C.: Exploiting place features in link prediction on location-based social networks. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’11. p. 1046. ACM Press, New York, New York, USA (2011)

11. Song, H.H., Cho, T.W., Dave, V., Zhang, Y., Qiu, L.: Scalable proximity estimation and link prediction in online social networks. In: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. pp. 322–335. ACM (2009)

12. Facebook, L.B., Leskovec, J.: Supervised Random Walks: Predicting and Recommending Links in Social Networks. (2010)

13. Zappella, G., Karatzoglou, A., Baltrunas, L.: Games of Friends: a game-theoretical approach for link prediction in online social networks. In: Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence (2013)

14. Rietveld, T., Van Hout, R.: Statistical techniques for the study of language and language behaviour. Walter de Gruyter (1993)

15. Field, A.: Discovering statistics using SPSS for Windows: Advanced techniques for beginners (Introducing Statistical Methods series), (2000)

16. Rummel, R.J.: Applied factor analysis. Northwestern University Press (1988)

17. Spearman, C.: &quot;General Intelligence,&quot; Objectively Determined and Measured. Am. J. Psychol. 15, 201 (1904). doi:10.2307/1412107

18. Nunnally, J.C., Bernstein, I.H.: Psychometric Theory McGraw-Hill New York Google Scholar. (1978)

19. Comrey, A.L., L., A.: Common methodological problems in factor analytic studies. J. Consult. Clin. Psychol. 46, 648–659 (1978). doi:10.1037/0022-006X.46.4.648

20. Dupuis, M., Khadeer, S., Huang, J.: “I Got the Job!”: An exploratory study examining the psychological factors related to status updates on facebook. Comput. Human Behav. 73, 132–140 (2017).

doi:10.1016/j.chb.2017.03.020

21. Lee, S., Kimn, A., Chen, Z., Paul, A., Pritchard, D.: Factor analysis reveals student thinking using the mechanics reasoning inventory. L@S 2017 - Proc. 4th ACM Conf. Learn. Scale. 197–200 (2017). doi:10.1145/3051457.3053984

22. Marsden, N.: Attitudes towards online communication: An exploratory factor analysis. 2013 ACM Conf. Comput. People Res. SIGMIS-CPR 2013. 147–152 (2013). doi:10.1145/2487294.2487326

23. Walther, J.B.: Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. Communic. Res. 23, 3–43 (1996)

24. Conger, S., Loch, K.D., Helft, B.L.: Information technology and ethics. In: Proceedings of the conference on Ethics in the computer age -. pp. 22–27. ACM Press, New York, New York, USA (1994)

25. Mason, R.O.: Four Ethical Issues of the Information Age. MIS Q. 10, 5 (1986). doi:10.2307/248873

26. Hall, M., Mazarakis, A., Peters, I., Chorley, M., Caton, S., Mai, J.-E., Strohmaier, M.: Following User Pathways: Cross Platform and Mixed Methods Analysis. In: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems. pp. 3400–3407. ACM Press, San Jose, USA (2016)

27. Spiliotopoulos, T., Oakley, I.: An exploratory study on the use of Twitter and Facebook in tandem. Proc. 2015 Br. HCI Conf. - Br. HCI '15. 299–300 (2015). doi:10.1145/2783446.2783620

28. Jha, R., Saini, A.K.: An exploratory factor analysis on pragmatic Lean ERP implementation for SMEs. Proc. 2012 2nd IEEE Int. Conf. Parallel, Distrib. Grid Comput. PDGC 2012. 474–479 (2012). doi:10.1109/PDGC.2012.6449867

29. Schreiber, M., Malesios, C.C., Psarakis, S.: Exploratory factor analysis for the Hirsch index, 17 h-type variants, and some traditional bibliometric indicators. J. Informetr. 6, 347–358 (2012). doi:10.1016/j.joi.2012.02.001

30. Lichtenwalter, R.N., Chawla, N. V: Lpmade: Link prediction made easy. J. Mach. Learn. Res. 12, 2489–2492 (2011)

31. Adamic, L.A., Adar, E.: Friends and neighbors on the web. Soc. Networks. 25, 211–230 (2003)

32. Zhou, T., Yan, G., Wang, B.-H.: Maximal planar networks with large clustering coefficient and power-law degree distribution. 8718, (2004)

33. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)

34. Newman, M.E.: Clustering and preferential attachment in growing networks. Phys. Rev. E. Stat. Nonlin. Soft Matter Phys. 64, 025102 (2001)

35. Katz, L.: A new status index derived from sociometric analysis. Psychometrika. 18, 39–43 (1953)

36. Lichtenwalter, R.N., Lussier, J.T., Chawla, N. V: New perspectives and methods in link prediction. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 243–252. ACM (2010)

37. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web. (1999)

38. Brin, S., Page, L.: The anatomy of a large scale hypertextual Web search engine. Comput. Networks ISDN Syst. 30, 107–17 (1998). doi:10.1.1.109.4049

39. Tyagi, N., Sharma, S.: Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page. Int. J. Soft Comput. Eng. 2231–2307 (2012)

40. Moore, D.S., Mccabe, G.P.: STATISTIEK IN DE PRAKTIJK Theorieboek.

41. Lee, A.S., Baskerville, R.L.: Generalizing generalizability in information systems research. Inf. Syst. Res. 14, 221–243 (2003)