technologies

MDPI

Article

User Similarity Determination in Social Networks

Sadia Tariq [1,*], Muhammad Saleem [2] and Muhammad Shahbaz [2]

[1] Department of Computer Science & Engineering, Narowal Campus, University of Engineering & Technology Lahore, Lahore 51600, Pakistan
[2] Department of Computer Science & Engineering, University of Engineering & Technology Lahore, Lahore 54890, Pakistan; saleemnawaz@gmail.com (M.S.); m.shahbaz@uet.edu.pk (M.S.)
* Correspondence: sadia.tariq@uog.edu.pk; Tel.: +92-333-8870846

Received: 8 March 2019; Accepted: 3 April 2019; Published: 15 April 2019

Abstract: Online social networks have provided a promising communication platform for an activity inherently dear to the human heart, to find friends. People are recommended to each other as potential future friends by comparing their profiles which require numerical quantifiers to determine the extent of user similarity. From similarity-based methods to artificial intelligent machine learning methods, several metrics enable us to characterize social networks from different perspectives. This research focuses on the collaborative employment of neighbor based and graphical distance-based similarity measurement methods with text classification tools such as the feature matrix and feature vector. Likeminded nodes are predicted accurately and effectively as compared to other methods.

Keywords: neighbour based similarity metrics; Euclidean distance; distance based similarity metrics; similarity score vector; data objects; adjacency matrix; link formation; node profile attributes

1. Introduction

Social network analysis is a measurement of relationships and the flow of information between the people of an organization, portraying them as actors of a network or as inter-connected, knowledge-sharing entities. The actors in the network and the relationships between them are modeled in the form of mathematical graphs having nodes and links. The links or edges between people or actors carry the information in the form of emails, telephonic conversations and text and picture messages, etc., counted as the weight of the link. Similarly, links can be directed to show the initiator of conversion.

Models determine whether localized network measures are sufficient to explain the global properties of the network or not. Conclusively, social network analysis provides a visual representation in the form of network graphs and a statistical representation in the form of network centrality measures and network similarity measures about human relationships.

With the passage of time, dynamic social networks tend to shift properties. There exists a finite set of unique features for social networks. The properties exhibited by a random communication network at a given point in time is merely a subset of the above-mentioned superset. The superset of salient characteristics identified to be present in social networks is given below:

- Directed ties
- Weighted ties
- Disconnected components
- Isolated nodes
- Bi-directional ties
- Non-normally distributed data

- Near complete node information
- Near complete link information
- Missing node information
- Missing link information

The temporal evolution undergone by dynamic social networks is classified into two kinds: expansion (when nodes are being added) and shrinking (when nodes are being removed). They might grow stronger (when the number of ties or tie weights increase) and weak (when the number of ties or tie weights decrease). The shift in size and shape corresponds to the properties they possess at a specific time.

A second concept utilized in the paper, and needing an introduction, is that of feature matrix and feature vectors. A feature matrix as shown in Table 1 is a set of features that characterize a given set of linguistic units for a finite set of properties. In lexical semantics, feature matrices can be used to determine the meaning of specific word fields.

Table 1. Feature Matrix.

	Feature A	**Feature B**	**Feature C**	**Feature D**
Feature A	✓	✓	✓	✓
Feature B	✓	✓	✓	✓
Feature C	✓	✓	✓	✓
Feature D	✓	✓	✓	✓

The approaches and metrics used for accurate similarity detection are large in number. Many combinations have been tried and tested. Therefore, we attempt to reap the benefits of a rarely used idea by jointly employing two well-known categories of similarity metrics. This paper presents the theoretical justification of loosely integrating neighborhood information and geographic distance information for precise identification of similar users within a social network. The fundamental idea is to formulate a feature matrix (adjacency matrix in case of social networks) from neighbor-based similarity scores and work out distance-based similarity metrics for two feature vectors (referred to as similarity score vectors in this study) at a time.

To accomplish our goal, we selected numerous more or less commonly known similarity measurement methods to be applied to social network graph. First of all, we provide a brief survey of the similarity metrics under consideration. Secondly, we perform social network analysis by populating an adjacency matrix for the weighted social network graph using neighbor-based similarity metrics. Every row in the matrix represented the similarity scores of a particular node with its neighbors, hence it was named the similarity score vector. Lastly, we bring into action vector-based similarity metrics for measuring the pairwise distance between vectors on two-dimensional plane.

The study aims to elaborate and reveal the productivity of various combinations of similarity-based methods in constructing a user similarity determination system without additional information about the suggested content.

In the following section a consolidated literature survey that scampers around the studies related to the proposed scheme is presented.

2. Related Work

Social networks have become an integral, part of our daily lives, enabling us to keep in touch with our friends and families [1]. The existing models incorporate either node profile attributes or link formation information or both [2]. Most of them represent the members and relationships among them in the form of a mathematical graph with nodes as vertices and ties as edges. In 1960's Stanley Milgram presented the term, "small-world phenomenon" [3]. It stated that we are all linked by short chains of acquaintances and was the pioneer to quantify the phenomenon allowing people to speak of the

"six degrees of separation" between any two people in the United States. Since then, many network models and frameworks were proposed to analytically study the phenomenon.

The concept of Similarity has been mathematically explained and formally defined in reference [4] and a graphical overview of applications of similarity algorithms and similarity metrics is presented in Figure 1.

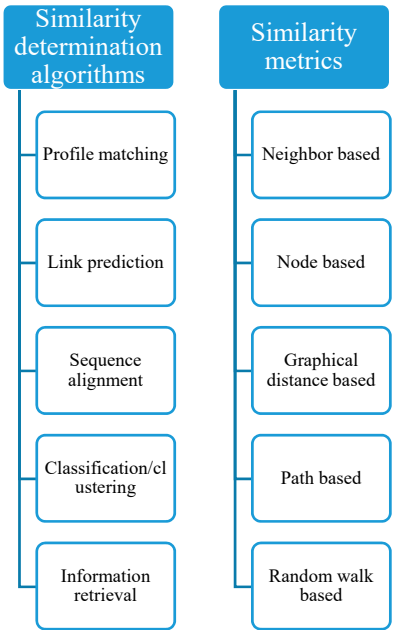

Figure 1. Algorithms and kinds of similarity metrics.

2.1. Profile Matching

A hybrid similarity measure that combines network similarity with node profile similarityhas been proposed in reference [5].A brief survey in reference [6] illustrates the variety of similarity measures developed for social networks and the difficulty of selecting a similarity measure for problems such as link prediction or community detection.

The friendship recommendation system proposed in reference [2] employed different similarity metrics for various features. When applied to real world social network data, it suggested that interest, age, gender, and location contribute to friendship formation.

The hierarchical arrangement of the network in the form of a tree with nodes and leaves relies on the belief that a pair of nodes is likely to be similar if they are equally close to a specific common predecessor. Kleinberg in reference [3] proposed that there is a decentralized algorithm capable of finding short paths with high probability. Decentralized search algorithms using combinations of homophily and node degree parameters in reference [7]. The work in reference [8] presents an extension of decentralized search algorithm which enables it to search in complete unconscious mode when the user is unaware of his or her position as well as that of others within the network. Also, community associations and common keywords play a vital role in identification of like-minded users [9].

Link prediction is mostly used to determine the most liked item by a user based on his/her buying patterns in past, and recommender systems and link prediction go side by side. In reference [10], a link prediction system has been constructed. Tested on an online judge, it recommended programming problems and tasks to users to be solved similar to test-session or practice-session. Based on performance in one problem, a next more difficult problem is presented to the user. Such automatic judgment systems are used for onsite training of programming contests.

Adamic and Adar in reference [11] developed functions to analyse similarities between users as a function of the frequency of a shared item. The reference [12] highlights construction of a measure of similarity of whole graphs based on the similarities of nodes.

User interest patterns have been studied chronologically to understand the underlying dynamics of the similarity determination process [13]. It has been proved that both path-based and neighbor-based approaches can significantly do better than the standard user-based algorithms [14]. Most of the studies on link prediction focus on un-weighted networks. Extensive work has been done to widen the information carrying capacity of similarity metrics by creating their weighted versions [15]. However, their performance has been questionable when tested on real world networks. The experimental study has been conducted to determine the role of weak links to improve the quality of results. Many other methods have been proposed to determine the real purpose of tie weights.

Other research works such as in reference [16] use projections mechanism to convert bi-partite graphs into non-bipartite and vice versa for application of suitable mathematical models. Many kinds of research jointly utilize similarity-based methods along with other approaches. The work in reference [17] chooses to enhance famous similarity-based metrics with network centrality measures such as betweenness, closeness and degree.

2.2. Distance between Feature Vectors

Though loosely based on references [18–20], our proposed methodology is distinguished by its attempt to incorporate local structural information by giving weight to the distances between similarity scores with neighbors. Two nodes are considered similar based on nearness between their similarity scores with their neighbors. Various combinations of metrics were tried and their results were compared. The ultimate goal was to precisely predict future friends for a particular node in its neighborhood. Similar approach has been attempted for graph matching in reference [21] by using the structural similarity of local neighborhoods to derive pair-wise similarity scores for the nodes of two different graphs, and present a related similarity measure that uses a linear update to generateboth node and edge similarity scores.

2.3. Similarity Measurement Methods

Similarity-based methods compute the pairwise similarity between nodes for the sake of prediction of possible future links among them [10]. Some similarity metrics are in common use for analyzing the affinity between two data objects. Be it a pair of vectors in two-dimensional plane or two nodes in a graph or more precisely two actors in a social network. These can also be classified broadly intosimilarity based methods and learning based methods [22]:

- Node-based metrics, which incorporate node features information for computing similarity.
- Neighbor-based metrics, which incorporate node's neighborhood information for computing similarity.
- Path-based metrics, which utilize the knowledge of various different alternative paths between two nodes.
- Random walk-based metrics, which use transition probabilities from a node to its neighbors and non-connected nodes to simulate social interactions.

In the proposed model the neighbor-based metrics and vector or geographical distance-based metrics will be jointly employed. Therefore, in the upcoming sections we give a brief overview of the chosen metrics from among these two categories.

3. Neighbor Based Similarity Metrics

Some of the most commonly used neighborhood-based similarity metrics [10] have been briefly introduced in the following paragraphs:

Edge Weight (EW), which measures the weight of the edge that links a pair of nodes in a social network graph. The presence of an edge between two problem nodes is mandatory and therefore cannot be used for an alikeness prediction.

$E\omega(\mathrm{x}, \mathrm{y}) = \mathrm{A}_{xy}$, where A_{xy} is the weight of the edge that connects nodes x and y

Common Neighbors (CN), which measure the number of neighbors that a pair of nodes have in common. Common neighbors are found using intersection operation on the two sets. The greater the intersection of the neighbor sets of any two nodes, the greater the chance of future association between them, and vice versa.

$\mathrm{CN}(\mathrm{x},\mathrm{y}) = \left|N(\mathrm{x}) \cap \mathrm{N}(\mathrm{y})\right|$, where |N(x)| represents the number of neighbors of node x or its degree.

Jaccard Neighbors (JN), which compares the number of common neighbors of x and y with respect to the total neighbors of x and y. Total number of neighbors is found using union operation on the two sets. This metric is based on set overlapping and uses a ratio of intersection operation to union operation. Empty set in denominator will cause numerical problems such as division by zero.

$$JN(\mathrm{x},\mathrm{y}) = \frac{\left|N(\mathrm{x}) \cap \mathrm{N}(\mathrm{y})\right|}{\left|N(\mathrm{x}) \cup \mathrm{N}(\mathrm{x})\right|}$$

Adar/Adamic (AA), which evaluate the likelihood that a node x is linked to a node y as the sum of the number of neighbors they have in common. Also, it can measure the intersection of neighbor-sets of two nodes in the graph but emphasizing in the smaller overlap. This metric is dependent upon commonality among the two problem nodes.

$$AA(\mathrm{x},\mathrm{y}) = \sum_{z \varepsilon N(\mathrm{x}) \cap (\mathrm{y})} \frac{1}{\log\left|\mathrm{N}(\mathrm{z})\right|}$$

Preferential Attachment (PA), which calculates the product of the degree of the nodes x and y, so he higher the degree of both nodes, the higher is the similarity between them. This metric has the drawback of leading to high similarity values for highly connected nodes to the detriment of the less connected ones in the network.

$$PA(\mathrm{x},\mathrm{y}) = \left|N(\mathrm{x})\right|.\left|N(\mathrm{y})\right|$$

In the upcoming section we discuss in detail the concept of feature matrix and feature vector, named here as similarity score vector. The similarity score vector (SSV) gets computed from the neighbor-based metrics discussed in above sections.

4. The Similarity Score Vector

In pattern recognition and machine learning, a feature vector is an n-dimensional vector of numerical features that represent some object. Many algorithms in machine learning require a numerical representation of objects since such representations facilitate processing and statistical analysis. Two data objects or more familiarly, feature vectors, are comparable on the basis of their numerical attributes or dimensions provided they are equal in number. When simulated on a two-dimensional plane, the distances between these vectors could be used to measure the farness or nearness between them. The lesser the difference between the values of their attributes, the more alike they are.

Utilization of adjacency matrix for mathematical representation of social networks is not an unfamiliar phenomenon. For example, to build a movie recommender system, a particular critic's review score (c), for a movie (m), will be represented by an element (c,m) in an adjacency matrix called feature matrix [23].

	*m*1	*m*2	*m*3	*m*4	*m*5	*m*6
*c*1	3	7	4	9	9	7
*c*2	7	5	5	3	8	8
*c*3	7	5	5	0	8	4
*c*4	5	6	8	5	9	8
*c*5	5	8	8	8	10	9
*c*6	7	7	8	4	7	8

Each row in the matrix represents the ratings of a particular critic c for all the movies under consideration in the review/rating space. The difference in opinion of two critics happens to be equal to the farness between their corresponding rating vectors on two-dimensional plane. If we want a feature vector per critic, then we can just take the rows of the matrix:

$$\mathbf{x1} = \mathbf{(3, 7, 4, 9, 9, 7)}$$
$$..$$
$$..$$
$$..$$
$$\mathbf{x6} = \mathbf{(7, 7, 8, 4, 7, 8)}$$

where x1 corresponds to feature vector of Critic 1, x2 corresponds to feature vector of Critic 2, and so on. Now, we propose that instead of measuring dissimilarity between the critic's opinion, we should measure the dissimilarity between movies. We compute distance between movies in the space of the critic's ratings. Transposing the matrix, we get feature vectors for movies m1 to m6:

$$\mathbf{w1} = \mathbf{(3, 7, 7, 5, 5, 7)^T}$$
$$.$$
$$.$$
$$\mathbf{w6} = \mathbf{(7, 8, 4, 8, 9, 8)^T}$$

In our proposed model, each row of adjacency matrix is composed of similarity score of particular node x with all of its neighboring nodes. Since a square matrix possess a uniform number of rows and columns, therefore each row represents the similarity score vector of the node x. A pictorial representation of a similarity score vector for various nodes from a particular network snapshot is shown in Figure 2. It shows PA-SSV for node B grows linearly from node A to D respectively whereas that for node c remains a straight line. Thus, node C holds equal similarity with respect to PA measure with all of its neighbors from A to D. And node B's similarity with respect to PA measure increases gradually from node A to D.

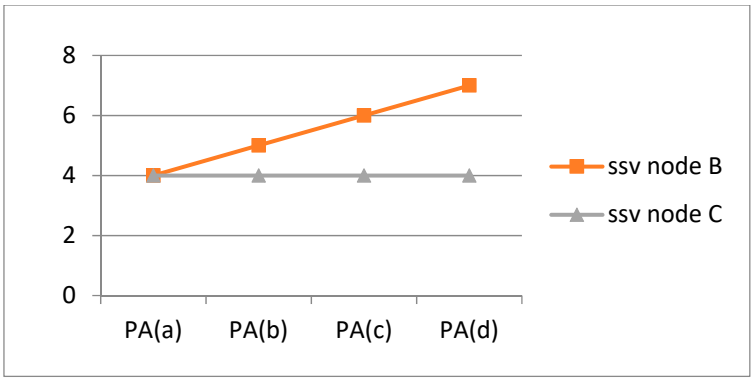

Figure 2. Preferential Attachment (PA) similarity score vectors (SSV) of nodes B and C with neighbors A, B, C, D, respectively.

This approach can be tested to perform link prediction using similarity scores for an experimental dataset which is split into two halves at particular timestamp t. The former set will contain the data collected before time t and the latter set will be used for evaluation purpose. Our work however purely generates theoretical results and is needed be tested on real world social network data.

Vector Based Similarity Metrics

The pair wise distance between two nodes is computed by applying a variety of distance-based similarity metrics on their respective similarity score vectors:

Euclidean distance, which is the difference between individual attributes of two data objects, provided the number of attributes and the scale used to normalize them is uniform for the pair [23], mathematically,

$$d(\mathrm{p},\mathrm{q}) = \sqrt{(\mathrm{p}_1-\mathrm{q}_1)^2+(\mathrm{p}_2-\mathrm{q}_2)^2+\ldots\ldots\ldots+(\mathrm{p}_n-\mathrm{q}_n)^2} = \sqrt{\sum_{i=1}^{n}(\mathrm{p}_i-\mathrm{q}_i)^2}$$

Pearson's coefficient, a linear similarity measure which uses mean centering and normalization of profiles. It uses a best fit line that runs through the attributes of the two data objects, plotted as data points on a two-dimensional plane. It is calculated by dividing the covariance of the two data objects by the product of their standard deviations. It performs better than Euclidean distance for non-normally distributed data.

$$\rho_{x,y} = \frac{\mathrm{Cov}(\mathrm{x},\mathrm{y})}{\sigma_x\sigma_y}$$

Cosine similarity, a linear similarity measure that uses normalization of profiles. It measures the similarity between two data objects treated as vectors, as the angle between them. This method is useful when finding the similarity between two text documents whose attributes are word frequencies A perfect similarity will have a score of 1 (or an angle of 0) and no similarity will have a score of zero (or an angle of 90 degrees).

$$\cos(A,B) = \frac{A\times B}{\| A \|\| B \|}$$

Tanimoto coefficient, a generalized version of Jaccard Coefficient, a set of overlap measures. Here A and B are data objects represented by vectors. The similarity score is the dot product of A and B divided by the squared magnitudes of A and B minus the dot product. A perfect similarity score would be one.

$$T(A,B) = \frac{A.B}{\| A \|^2 + \| B \|^2 - A.B}$$

Using the grocery store example, the Tanimoto coefficient ensures that a customer who buys five apples and one orange will be different from a customer who buys five oranges and an apple.

All of these measures perform well for non-bipartite graphs. The social network graph under consideration is however strictly non-bipartite. Its adjacency matrix comprises of nodes along rows as well as along columns, thus producing a people-people network or social network.

The Euclidean distance and its variations are easily computed using built-in functions provided by mathematical and statistical tools preferably Matlab and R. Hamming, Manhattan, Minkowski and Cityblock are few to name famous variations of this universal metric.

Similarity is further worked out as a decreasing function of geographic or Euclidean distances with formula given below [23]:

$$\mathrm{sim}(\mathrm{x},\mathrm{y}) = \frac{1}{1+d(\mathrm{x},\mathrm{y})}$$

The Euclidean distance is thus converted into a measure of similarity. It compares the pair wise affinity of nodes using their similarity score vectors. The higher the Euclidean distance score, lesser will be their similarity score and vice versa.

Hence a distance score of zero corresponds to a similarity of one (the largest value it can take) and a distance score of one corresponds to a similarity of zero (the smallest it can take).

Pearson's coefficient uses trend line to decide whether the data points are correlated positively or negatively. It is also possible that two sets of data points are simply uncorrelated. An evaluation chart which replicates a scattered plot of data points on two-dimensional plane serves the purpose.

5. The Proposed Model

The proposed model followed a simple set of steps in order to achieve its objectives. The first step was to compute various neighbor-based similarity metrics for sample social network graph shown in Figure 3 and craft an adjacency matrix.

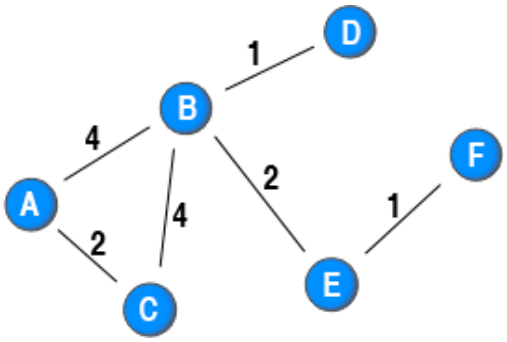

Figure 3. Weighted network graph.

The adjacency matrix for edge weight measure is tabulated below.

	A	*B*	*C*	*D*	*E*	*F*	*mean EW*
A	0	4	2	0	0	0	1
B	4	0	4	1	2	0	1.83
C	2	4	0	0	0	0	1
D	0	1	0	0	0	0	0.167
E	0	2	0	0	0	1	0.5
F	0	0	0	0	1	0	0.167

The social network graph simulates actors as nodes, communications as edges and number of messages transmitted or received as edge weights. The average similarity score per node was obtained from the adjacency matrix. Secondly, we selected two similarity score vectors from the adjacency matrix and computed vector-based similarity metrics for them. We repeated the procedure for node A's vector with all of its neighboring nodes' vectors in order to find most similar and least similar nodes for node A. Next, we employed line charts to illustrate the average neighbor-based metrics and bar charts to depict the similarity rank for vector-based metrics. Lastly, we compared the results of both charts. A high-level diagram of the whole course of action is shown in Figure 4 [2].

The network graph under consideration shows that for node A, the most similar node is node C with same number of ties, tie weights and same connected neighbors. The output variables used for evaluation of generated results are listed below:

- most similar node for node A
- least similar node for node A

In the next section we discuss the empirical setup, the experiment and the analysis of results of the proposed model.

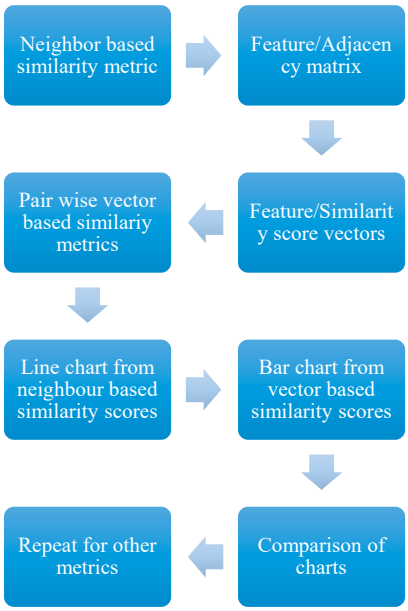

Figure 4. High level view of the proposed model.

6. Empirical Setup and Evaluation

In this section we evaluate the performance of the previously discussed neighbor-based similarity metrics to recommend new potential friends in comparison to the joint incorporation of neighbor based and -vector based similarity metrics.

Figure 5 indicates the mean similarity score of all nodes. The most and least similar nodes for node A as determined by various neighbor-based metrics is tabularized in Table 2. According to the presented results, the most similar node for node A, agreed upon by all metrics except AA. For AA node B happens to be most similar.

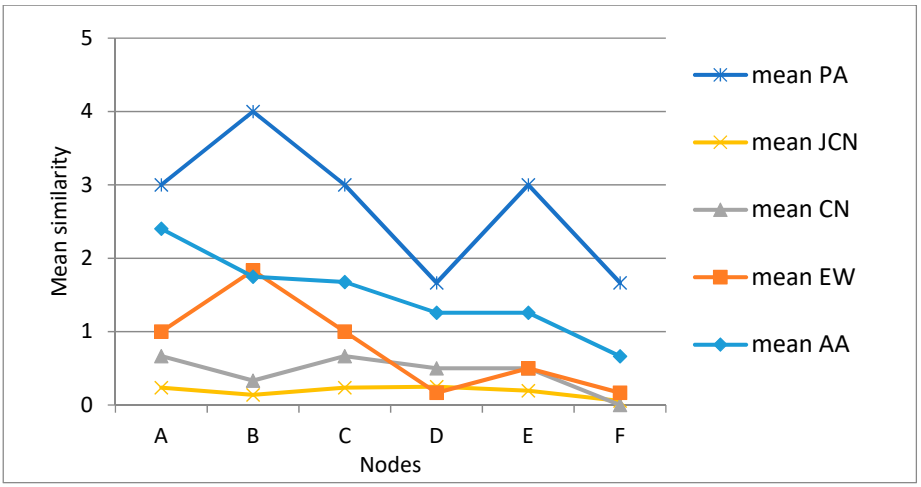

Figure 5. Neighbor based similarity scores for all nodes.

Table 2. Most and least similar nodes for node A determined by neighbor-based metrics.

	Most Similar to Node A	Least Similar to Node A
PA	C	D,F
JCN	C	F
CN	C	F
EW	C	D,F
AA	B	F

The least similar node is obviously node F, agreed upon by all.

It is evident that only JCN and CN metrics successfully determined similarity closer to reality. The most precise proximity prediction is accomplished by JCN and CN. The work in reference [10] compared various neighbor-based similarity measures and discovered that edge weight is best among its competitors.

The Figures 6–8 illustrate the most similar and least similar nodes for node A identified by vector-based similarity metrics. The obtained results for most similar node for node A are summarized in Table 3. The work in reference [12] compares two commonly used distance measures in vector models, namely the Euclidean distance and cosine angle distance for nearest neighbor determination in high dimensional data spaces. Through theoretical analysis as well as experimental evaluation it was revealed that the retrieval results based on Euclidean distance are similar for Cosine angle distance when dimension is high.

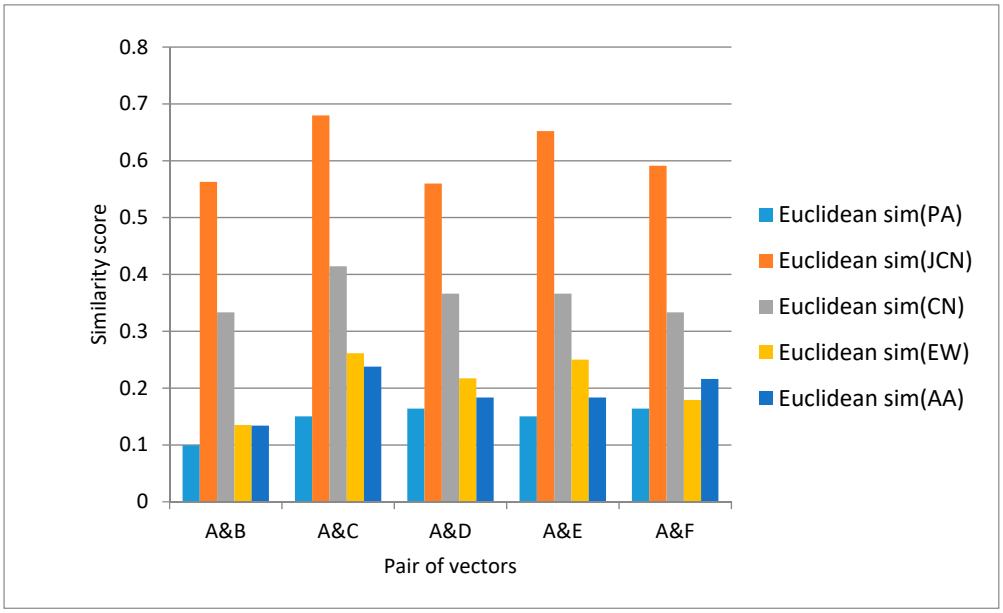

Figure 6. Similarity for neighbor-based metrics.

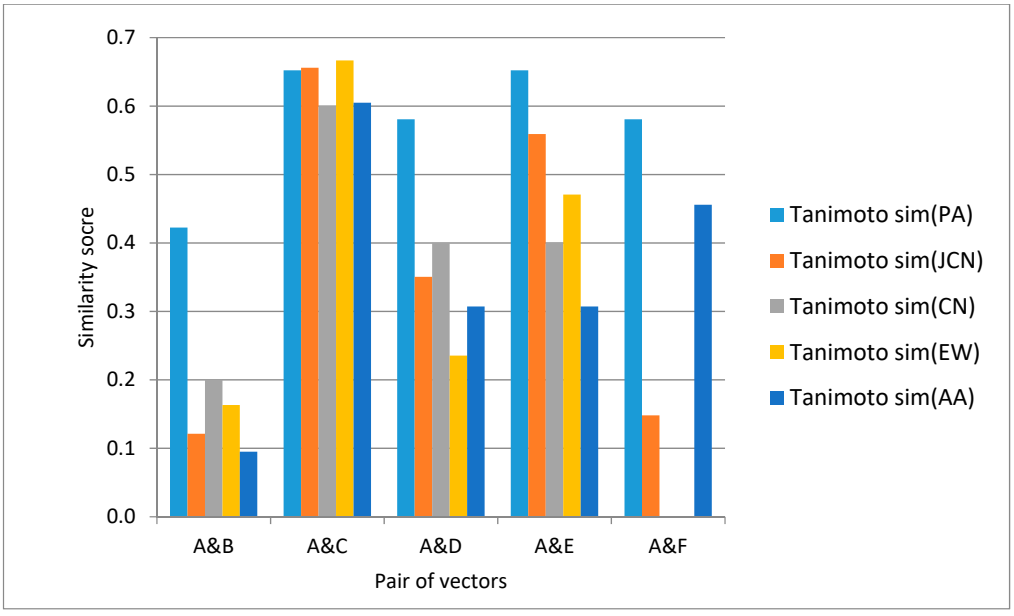

Figure 7. Coefficient similarity for neighbor based metrics.

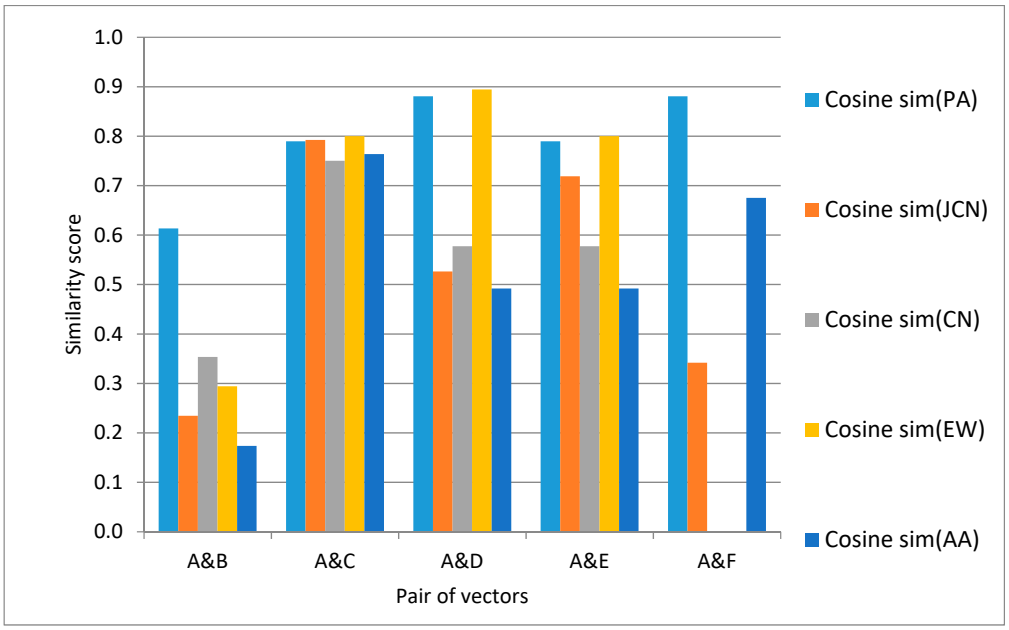

Figure 8. Similarity for neighbor based metrics.

Table 3. Most similar node for node A.

	Msn (PA)	Msn (JCN)	Msn (CN)	Msn (EW)	Msn (AA)
Sim(x,y) Euclidean	C,E	C	C,D	C	C
Sim(x,y) Tanimoto	C,E	C	C,D	C	C
Sim(x,y) Cosine	D,F	C	C,D	D	C

- Table 3 clearly proves that when teamed up with vector-based measures, JCN and AA determined node similarity most precisely.
- Furthermore, the table told us that among vector-based measures, Euclidean distance and Tanimoto coefficient presented analogous results.

The value of Pearson's coefficient computed for various neighbor-based metrics is presented in Table 4 and is evaluated by using the Figure 9. The key rules for evaluation are listed below:

1. A value above 0 represent positive correlation
2. A value below 0 represent negative correlation
3. A value of 0 represents that the two vectors in question are uncorrelated

Table 4. Pearson's coefficient for neighbor-based metrics.

	ρ (PA)	ρ (JCN)	ρ (CN)	ρ (EW)	ρ (AA)
A & B	−0.4	−0.6	−0.2	−0.2	0
A & C	0.2	0.4	0.2	0.6	0
A & D	0.5	−0.1	0.2	0.7	−0.8
A & E	0.2	0.3	0.0	0.6	−0.8
A & F	0.5	0.0	0.0	−0.2	0.8

Figure 9. Chart for Pearson's coefficient.

- According to the tabularized results, all metrics agree at the similarity or positive correlation of node C with node A except for Adar/Adamic which evaluates them to be uncorrelated.
- It is evident that Pearson's coefficient did not work well with AA.
- From metric perspective, JCN appears to be best predictor with CN standing at second position.

In the next section we justify our submitted results by comparing them with latest research on the subject. Our list of conclusions is also supported by many cited references.

7. Analysis of Results

In order to deeply envision the results produced, let us have a quick glance at the performance and robustness of neighbor based and vector-based metrics from existing researches.

Liben-Nowell and Kleinberg in reference [24] used a co-authorship network of scientists to prove that an evolutionary dynamic spatio-temporal network is affected more by endogenous changes rather than exogenous changes. They concluded that a similarity determination algorithm for link prediction based on network topology only outperforms direct or neighbor-based measures. The correctness of prediction rises by a factor of 40–50 for network structure based measures as compared to random prediction as elaborated by Table 5. Graph distance predictors lie between the two extremes with mediocre performance.

Table 5. Improvement factor of various techniques over correct random prediction [24].

Technique	Probability of Correct Random Prediction	Improvement Factor
Graphical distance metrics	0.153%	29.2
CN	0.153%	47.2
PA	0.207%	15.2
AA	0.147%	54.8
Jaccard/Tanimoto	0.147%	42.3

Haihang and Zhang in reference [25] attempted to lift the performance of four chosen graph distance descriptors using time as active factor. The idea was to convert existing structural similarity measures into temporal similarity measures. They used two datasets: one was a citation network which derived the citation data in high energy physical phenomenon between 1992 and 2001 on Arxiv.org website, while the second was data of cooperation network extracted from thesis information concerning high energy physical theory part between 1991 and 2012. They concluded that PA was enhanced more by active factor than other measures and produced more accurate results.

Sharma and Khatri in reference [26] compared twelve graphical distance metrics using the Epinion dataset, which is a general consumer review website for large range of products. It belongs to the who-trusts-whom category of online social networks. They concluded that node neighbor behaved best in terms of precision of results produced, whereas PA behaved worst as indicated by Table 6.

Table 6. Precision of various metrics [25].

Technique	Precision
Node Neighbor	16.4
Jaccard/Tanimoto	15.8
Adar/Adamic	15.1
Preferential attachment	7.1

The results produced by our theoretical analysis and evaluation are summarized below in Table 7:

Table 7. The best proposed metrics as proposed by experiment in comparison to recent researches.

Best Metric	Experimental	Cited
JCN/CN	Table 2	Tables 5 and 6
JCN + Euclidean/Tanimoto/Cosine	Table 3	Table 5
AA+Euclidean/Tanimoto/Cosine	Table 3	Table 6
Pearson+JCN/CN	Table 4	Table 8

Table 8. Properties and limitations of distance-based measures.

Metric	Properties	Limitations
Euclidean	-geometric distance measure -alignment of vectors	-less optimal in higher dimensional space -sparse vectors -sensitive to data distortion and outliers
Cosine	-normalized inner product -symmetric -incorporate semantics -use relative frequencies -best for sparse vector space -same for superset and its subset -multiplicative and additive identity invariant	-less optimal in lower dimensional space -scale and origin shift variant -effected by liner transformation -
Tanimoto	-intersection of objects/union of objects -extension of Jaccard -best for asymmetric binary variables -range between 0 and 1	-less optimal in lower molecular complexity space -not deliberately used as distance measure
Pearson	-origin and scale shift invariant -symmetric -range between −1 and +1 -normalized covariance -different for superset and its subset - multiplicative and additive identity invariant	-neglects role of third variable -assumption of linearity -time consuming -sensitive to extreme values

The case of distance or vector-based measures is different from the neighbor-based measures. Euclidean distance is affected by every term which is non-zero in either vector. It tends to suffer from the curse of dimensionality. As dimension grow higher and vector space grows sparse, its performance slowly downgrades. It means that they are not reliable with larger social network datasets. A possible solution to this problem is normalization of the outlier/noisy feature provided its minimum, and maximum possible value is known. Scaling can also be applicable for certain datasets.

Cosine, on the other hand, depicts the angular orientation of the vectors rather than magnitudes. It is affected by the terms which two vectors have in common and understands the meaningful semantics of similarity. Properties and limitations of distance metrics have been tabularized below:

8. Conclusions and Future Work

Incorporation of a useful and accurate user similarity determination system is an integral part of social networks. In this paper, we have considered a simple network graph and used social network analysis and mathematical techniques for determination of similarity between nodes.

The data analysis showed that the JCN and AA when vigilantly combined with Euclidean distance, Tanimoto coefficient or cosine similarity, provided most accurate results. However, one of the biggest anomalies in neighbor-based metrics is their absolute dependence upon the existence of a tie between a pair of nodes. It is possible only if they are friends already and therefore is of little use for similarity prediction. Among the distance-based measures, Pearson's coefficient performed admirably well when paired with JCN and CN but produced surprising results with AA.

A theoretical foundation for proposed model has been established. Practical evaluation on real world social network dataset is a lavishing consideration for future work. Though the results produced are fairly justified by similar studies, there is no clear winner among the metrics discussed. It is quite likely that similarity metrics when paired differently, work well together under certain circumstances and behave weirdly under others. Trying various combinations of metrics might be promising. The use of weighted versions of metrics could be tried to produce more precise results for weighted networks. The limitations of certain metrics such as edge weight (the existence of a tie is mandatory between two nodes and for that they have to be friends already) could be taken into account and studied. Also, the reasons behind the poor performance of AA with Pearson needs to be explored.

The roots of variation in results lie in the underlying dynamics and randomly shifting set of characteristics of a real-world social network. However, it can be safely concluded that the more node and network structure properties a model reveals, the better will be the results produced. To achieve this goal, better and newer methods and metrics should be incorporated and evaluated.

Author Contributions: Conceptualization, S.T. and M.S. (Muhammad Shahbaz); methodology, S.T.; software, M.S. (Muhammad Saleem).; validation, M.S. (Muhammad Shahbaz) and M.S. (Muhammad Saleem); formal analysis, M.S. (Muhammad Shahbaz) and M.S. (Muhammad Saleem); investigation, S.T. and M.S. (Muhammad Saleem).; resources, S.T. and M.S. (Muhammad Shahbaz); data curation, M.S. (Muhammad Saleem); writing—original draft preparation, S.T.; writing—review and editing, M.S. (Muhammad Shahbaz); visualization, M.S. (Muhammad Shahbaz) and M.S. (Muhammad Saleem); supervision, S.T. and M.S. (Muhammad Shahbaz); funding; S.T. and M.S. (Muhammad Saleem).

Funding: This research received no external funding and the APC was funded by [Sadia Tariq and Muhammad Saleem].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liang, X.; Li, X.; Zhang, K.; Lu, R.; Lin, X.; Shen, X.S. Fully anonymous profile matching in mobile social networks. *IEEE J. Sel. Areas Commun.* **2013**, *31*, 641–655. [CrossRef]
2. Mazhari, S.; Fakhrahmad, S.M.; Sadeghbeygi, H. A user-profile-based friendship recommendation solution in social networks. *J. Inf. Sci.* **2015**, *41*, 284–295. [CrossRef]
3. Kleinberg, J. The small-world phenomenon: An algorithmic perspective. In Proceedings of the thirty-second annual ACM symposium on Theory of computing, Portland, OR, USA, 21–23 May 2000.
4. Lin, D. An Information-Theoretic Definition of Similarity. Available online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.55.1832&rep=rep1&type=pdf (accessed on 15 April 2019).
5. Gurcan Akcora, C.; Carminati, B.; Ferrari, E. User similarities on social networks. *Soc. Netw. Anal. Min.* **2013**, *3*, 475–495. [CrossRef]
6. Rawashdeh, A.; Ralescu, A.L. Similarity Measure for Social Networks—A Brief Survey. In Proceedings of the Modern AI and Cognitive Science Conference (MAICS), Greensboro, NC, USA, 25–26 April 2015; Curran Associates, Inc.: Greensboro, NC, USA, 2015.
7. Şimşek, Ö.; Jensen, D. Navigating networks by using homophily and degree. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 12758–12762. [CrossRef] [PubMed]
8. Sandberg, O. *The Structure and Dynamics of Navigable Networks*; Division of Mathematical Statistics, Department of Mathematical Sciences, Chalmers University of Technology and Göteborg University: Göteborg, Sweden, 2007.
9. Banks, L.; Bhattacharyya, P.; Spear, M.; Wu, S.F. Davis social links: Leveraging social networks for future internet communication. In Proceedings of the Ninth Annual International Symposium on Applications and the Internet (SAINT'09), Bellevue, WA, USA, 20–24 July 2009.
10. Jimenez-Diaz, G.; Gómez-Martín, P.P.; Gómez-Martín, M.A.; Sánchez-Ruiz, A.A. Similarity metrics from social network analysis for content recommender systems. *AI Commun.* **2017**, *30*, 223–234. [CrossRef]
11. Adamic, L.A.; Adar, E. Friends and neighbors on the web. *Soc. Netw.* **2003**, *25*, 211–230. [CrossRef]
12. Nikolić, M. Measuring similarity of graph nodes by neighbor matching. *Intell. Data Anal.* **2012**, *16*, 865–878. [CrossRef]

13. Ricci, F.; Rokach, L.; Shapira, B. Recommender systems: Introduction and challenges. In *Recommender Systems Handbook*; Springer: Boston, MA, USA, 2015; pp. 1–34.
14. Chen, H.; Li, X.; Huang, Z. Link prediction approach to collaborative filtering. In Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'05), Denver, CO, USA, 7–11 June 2005.
15. Lü, L.; Zhou, T. Link prediction in weighted networks: The role of weak ties. *EPL (Europhys. Lett.)* **2010**, *89*, 18001.
16. Zhou, T.; Ren, J.; Medo, M.; Zhang, Y.-C. Bipartite network projection and personal recommendation. *Phys. Rev. E* **2007**, *76*, 046115. [CrossRef] [PubMed]
17. Liu, H.; Hu, Z.; Haddadi, H.; Tian, H. Hidden link prediction based on node centrality and weak ties. *EPL (Europhys. Lett.)* **2013**, *101*, 18004. [CrossRef]
18. Mafteiu-Scai, L.O. A new dissimilarity measure between feature-vectors. *Int. J. Comput. Appl.* **2013**, *64*, 39–44.
19. Balmachnova, E.; Florack, L.; ter Haar Romeny, B. Feature vector similarity based on local structure. Presented at the International Conference on Scale Space and Variational Methods in Computer Vision, Ischia, Italy, 30 May–2 June 2007.
20. Qian, G.; Sural, S.; Gu, Y.; Pramanik, S. Similarity between Euclidean and cosine angle distance for nearest neighbor queries. In Proceedings of the 2004 ACM Symposium on Applied Computing, Nicosia, Cyprus, 14–17 March 2004.
21. Zager, L.A.; Verghese, G.C. Graph similarity scoring and matching. *Appl. Math. Lett.* **2008**, *21*, 86–94. [CrossRef]
22. Wang, P.; Xu, B.; Wu, Y.; Zhou, X. Link prediction in social networks: The state-of-the-art. *Sci. China Inf. Sci.* **2015**, *58*, 1–38. [CrossRef]
23. Shimodaira, H. *Similarity and Recommender Systems*; School of Informatic, The University of Eidenburgh: Eidenburgh, UK, 2014; p. 21.
24. Liben-Nowell, D.; Kleinberg, J. The link prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.* **2007**, *58*, 1019–1031. [CrossRef]
25. Xu, H.H.; Zhang, L.J. Application of link prediction in temporal networks. *Adv. Mater. Res.* **2013**, *756*, 2231–2236. [CrossRef]
26. Sharma, D.; Sharma, U.; Khatr, S.K. An experimental comparison of the link prediction techniques in social networks. *Int. J. Model. Optim.* **2014**, *4*, 21. [CrossRef]