EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

GLOS: a global and local features oriented link prediction technique in social network

Mamoona Qadir [1], Abdul Samad[2,*], Hafeez Ur Rehman Siddiqui[1]

[1]Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan Pakistan
[2]Capital University of Science and Technology, Islamabad Pakistan

Abstract

The link prediction has attracted majority of researchers from various domains since the beginning of behavioral science. For instance, online social networks such as Twitter, LinkedIn and Facebook change rapidly as new users appear in the graph. For all these networks, the more challenging task is to find and recommend friends to the users. In case of social graph, the foremost objective of link prediction is to predict which new links are likely to be appearing from the actual state of the graph. Varieties of methods have been developed such as probabilistic, maximum likelihood and similarity-based techniques where similarity-based techniques are considered as the best prediction methods. Similarity-based methods uses a strategy, where each pair of nodes assigned a similarity score such that more similar nodes have more chances to connect in a future. Similarity estimation works on the global and local features i.e. path, random walk and neighbors. Local features are those features of node that consider at node level i.e. adjacent neighbors nodes. On the other hand, global features are those type of features that considers at graph level i.e. path between two nodes. Our hypothesis is that the combination of both local and global features is more powerful predictor for link formation. Here in this study, we have evaluated global, local and hybrid similarity measures. Moreover, we also proposed a hybrid approach GLOS. We performed experiments on five different dataset (Astor, CondMat, GrQc, HepPh and HepTh). After the result evaluation, it is found that, hybrid approach GLOS obtained the highest accuracy by 1 on all the dataset, while, global approaches could not produced lowest accuracy on all dataset. On the other hand, HP from local similarity outperformed than rest of the local and global approaches.

Received on 25 June 2021; accepted on 07 August 2021; published on 13 August 2021

Keywords: Centrality, Social Network Analysis, Ranking, Influential Users

Copyright © 2021 Mamoona Qadir et al., licensed to EAI. This is an open access article distributed under the terms of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/), which permits unlimited use, distribution and reproduction in any medium so long as the original work is properly cited.

doi:10.4108/eai.13-8-2021.170672

1. Introduction

In 1954 J.A. Barnes was the first person who use word Social Network in the Class and Committees of Norwegian Island Parish to describe human links [34][36]. The current social network theory was intensify by Stanly Milgram [26]. Many individuals relate to each other and connected in a way of map called Social Network. It (Social Network) represents as a social structure made up of many different network nodes, and each node represents unique existence; it could be an individual person or a group. In general, Social Network is a collection or map of nodes and set of edges representing relation, link, connection among these nodes; these links could be family, friends colleagues and so on. In essence, a social network is a type of social structure having nodes connected to each other with particular type of one or more interconnection, like a map of all suitable links among the nodes (Samad et al., 2020a) as representing in Figure 1.

The creation of Social Networking is just to build more opportunities to develop friendships, share information and promote ideas for business in the network. For instance social network services, such as:

- Twitter,
- Facebook,
- Instagram,
- LinkedIn,

*Corresponding author. Email: writetosamadalvi@gmail.com

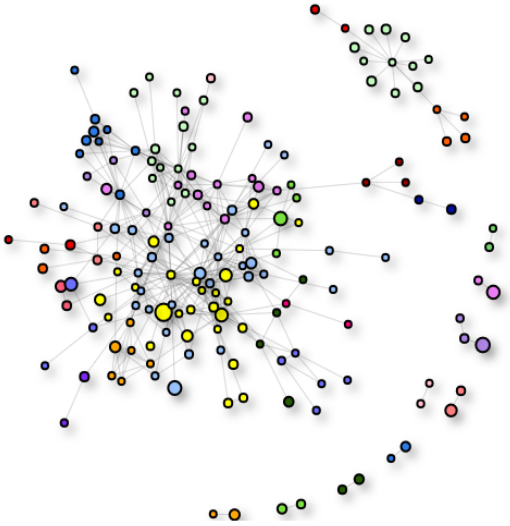

Figure 1. Social Network Graph

- Live Journal,
- Researchgate etc

One of the most fundamental characteristic of a social network is its huge broadening and dynamic nature that has led many researchers to study the network properties. In fact, huge scale of structural and behavioral of social network. Studying these graphs and their dynamic evolution processes, researchers have find some valuable facts which can solve many real world problem in practical [10].

In contrast to old-time when a Social Network were only made via face-to-face communications [33], nowadays people are good to use different virtual networking applications to make new relationships and keep connected with their old ones known as Online Social Networking System(OSNS). Deep and detailed analysis of social network social is known as Social Network Analysis (SNA). SNA is the measurement and mapping of links, relationships between nodes (people, groups, computers, organizations) and many other connected entities which provides some knowledge and information. Nodes/Vertices in a network are people and groups while relation among them is represented by links. It could be mathematical analysis and visual demonstration of human relationships through SNA that help us to make sense of complex social network to acknowledge its evolution and dynamic behavior of social network to discover complex communication patterns and characteristic features of networks [8]. Growth of social network can be understood accurately by predicting the edges that will be created between two nodes during some time intervals from time t to future time t+1 and the dynamics which it holds back from past [41]. For analyzing social network link prediction is an important task in this information and modern era, as link prediction research on social behaviors of an individual or community can be helpful in the quantitative and qualitative assessment of human interaction also when people are engaging online in virtual social world. Moreover the result of this prediction algorithm and its methods can be applied over a wide variety of applications such as molecular biology, behavioral sciencesp, helpfulness prediction [11][12][13][14], Contour Prediction [42],service QoS prediction [29], predicting web page [22] and terrorism and criminal investigations or in any commercial businesses as mentioned earlier. Furthermore this can be extended or changed to adapt different studies [25]. Link prediction is an important task in social network analysis. Most of the similarity approaches uses global and local features for the link prediction. However, hybrid approaches are very few. In this study, we have proposed an hybrid approach namely GLOS (which is extension of our previous work [34]) and evaluated local, global and hybrid based similarity approaches for the links prediction. In the previous work, we have proposed a similarity measure namely SAM, which uses a local (i.e., neighbors) features for the link prediction.

In the next section, problem statement is discussed. In section 3, state-of-the-art similarity based approaches are categorized and discussed. Section 4 represent the methodology and working detail of GLOS approach for link prediction. In section 5, in terms of accuracy, results about each similarity measures are discussed. In the last section, conclusion is presented.

2. Problem Statement

Consider a social network as graph $G(V, E_t)$ at time t, where V is the set of nodes in the graph i.e., $V = v1, v2, v3, ..., vn$ and E_t corresponds to the set of edges i.e., $E_t = e1, e2, e3, ..., en$ at time t. goal is to predict which new edges are likely to become visible at time $t + 1$ i.e., to predict the set of edges $e_i : e_i \epsilon E_{t+1} \bigcap e_i \not\epsilon E_t$. Clearly, the assumption is that, there is no change in the set of nodes V during time $t1$ and $t + 1$. In addition, the graph G is unweighted and undirected. Hypothesis in this study is that combination of both local and global complex topological features are more powerful predictors of link prediction than common neighbors i.e., a shallow structural information. This study will be helpful in the accurate prediction of future or missing links in the graph.

2.1. Link Prediction Approaches Categorize

In the analysis of link prediction problem, there are variety of ways to consider it; some of the researchers treat this problem as a probabilistic where prediction of the edge is deemed as estimating the probability between two nodes in an undirected graph [25].

While, some of the researchers treat this problem as a classification where the prediction of the edges considers the similarity score ranking between two nodes [16]. These approaches are often categorized into two groups, one is so-called network evolution modeling, and other is named as feature based link prediction. The previous one predicts the long run edges of a network taking some well-known attributes of social networks from Social Theory like the ability law distribution and tiny world phenomenon [25]. Whereas feature based link prediction is principlaly related to solve the supervised classification, while, using a training set the task of classification is to predict unknown nodes. Machine learning approaches are considered to be highly associated with this approach and may be extended into broader domains [2]. Figure 2 shows an outline of the listed the categorization of the common link prediction techniques [4]. Hence, the most instinctive solution for the link prediction problem is similarity based techniques. The basic idea behind this is the association of any node with the graph which holds most similarity between nodes pair if and only if it was not linked with the graph before. Another technique Topology-Based Metrics being used to solve problem of link prediction. In Addition, the technique of social theory are generally considered to be the primogenitor of link prediction problem. Generally, these methods are based on the classical social analysis method, and a few more smart ideas are presented in next subsection.

It is worth mentioning that the learning based method is totally different from other metrics. Till date, node-based metrics and topology based metrics have been constructed to take an account of similarities between pairs of nodes and to generate a similarity list. Eventually the prediction grades can be list down in descending order. Although, the learning based method deals with link evaluation problem somewhat differently, and the main application of recommendation system is very suitable for this kind of method. This method of estimating correlation is a binary classification problem [39]. Machine learning models solve this problem by using model like classifier and collaborative filtering[15]. The basic idea of node pair in a graph is classified and correspond to using features or can be regarded as labeled. Then if a pair of nodes has high possibility to link will be labeled as positive otherwise negative [39].

3. Similarity-Based Approaches

Similarity-based approaches believe that nodes attempt to make links with other similar nodes [35]. These approaches works on the hypothesis that the nodes are similar if they need a standard connected node or they need a shortest distance within the network. A similarity function $S(u, v)$ is employed by these approaches which allocates similarity score to every non-connected pair of nodes u and v. Finally, pair of nodes sorted in descending order according similarity score. A high score represents high probability that nodes are linked near in an exceedingly future, while low score shows that nodes would not be linked.

3.1. Node-Based

Similarity computation between pair of node is a stimulating solution certain the task of link prediction. It builds on the easy idea: Maximum the similarity between a pair, the more chances a link between them. The theory behind the computation of similarity is very simple: two nodes will be connected with a link if they have more similar parameters and attributes. Interestingly, there is an old Chinese saying goes (i.e., birds of a feather flock together) which is entirely in keeping with this idea. This reflects the very fact that folks try and make relationship with those that are similar in religions, language, educations, locations and interest. This relationship will be measure by computing similarity, where score (known as similarity between u and v) is assigned to every pair of nodes (u, v). A high similarity score means u and v will b linked, while low similarity score means u and v would not be linked.

In an exceedingly practical SN, a node (i.e., people) has profile in online SN containing set of attributes like gender, age, location, language, interest, bio, country and city [25]. These attributes values is accustomed compute similarity between pair of nodes. Most of your time, these attributes are in textual form, where textual-based similarities [40] are used. Discussing similarity based approaches is against the aim of this study, reader can read some comprehensive survey [35] within the area of citation network, evaluated both textual and topological similarity measures so as to predict the link between research papers. Where, they need to use profiles of research papers containing textual attributes including title and abstract. Their observation is that predicting link between nodes through topological similarity is best than textual similarity. They also observe that increasing text in attributes lowers the similarity between nodes. Bhattacharyya et al. define tree model with multiple categorize to review and analyze the keywords of profiles, then compute distance of keywords to search out similarity between users [17]. They need found that similarity between users is sort of equal apart from direct friends. Additionally, the maximum amount as keywords and friends increased, similarity between users decreased.

Akcora et al. observe that the majority of the user profiles don't seem to be publicly available in current social networks or missing [5]. Keeping this limitation

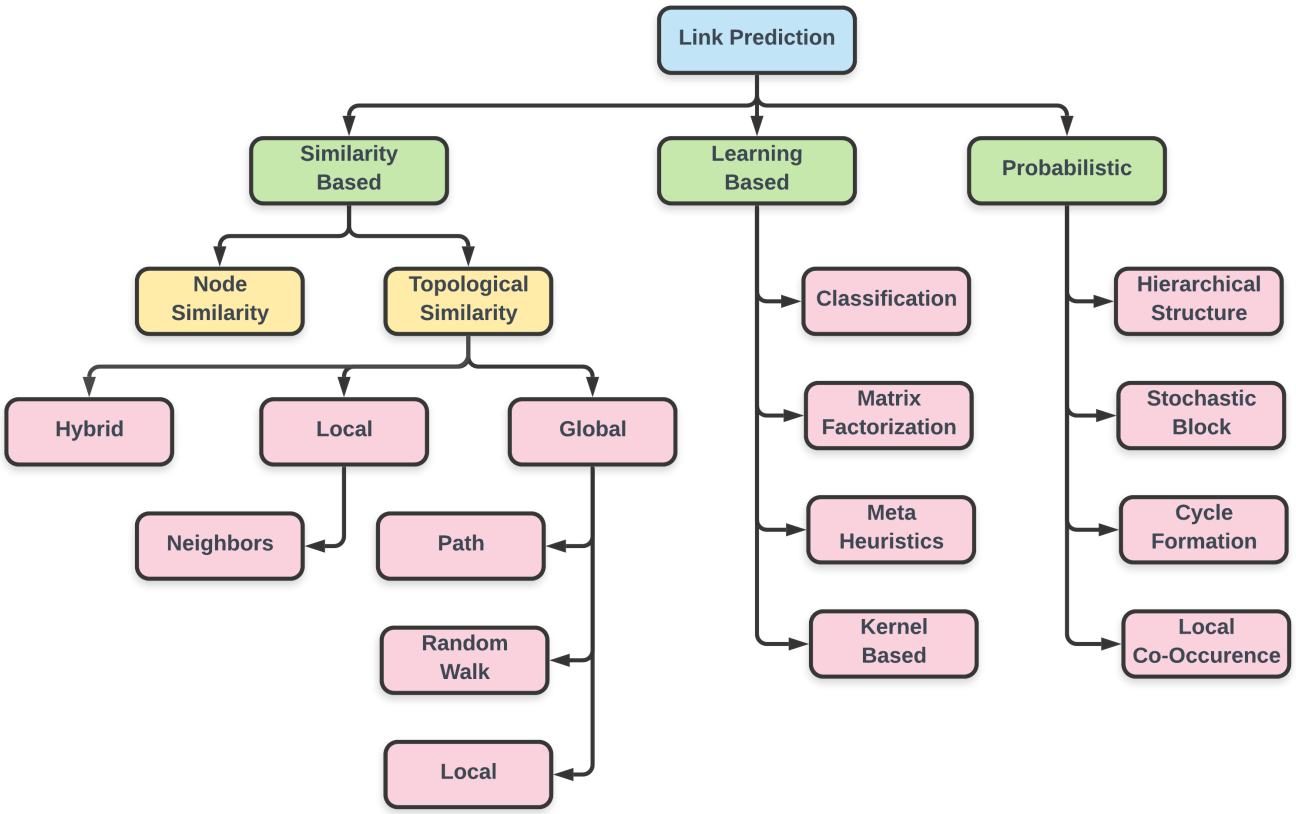

Figure 2. Link Prediction Categorize

in mind, they devise a technique that before computing users similarity, estimate the portion of missing values of profile [19]. Anderson et al. use the commonality of users interest to live similarity ([18]. Users interests are actions that user takes, like asking question, editing article, reading blog and bookmark items. of these actions taken by user is represented as vector, and users similarity is that the cosine between action vectors. Samad et. al, within the context of face-to-face contact networks, evaluate six different social attributes so as to predict the link [33]. They need found that, language and country are such attributes that plays a crucial role connected prediction. They need observed that individuals tend to contact those that are similar in language and country. Last, actions and attributes are mostly employed in node-based similarity approaches. These actions and attributes reflect the private behaviors and interests. Just in case of getting social attributes and behaviors, node based approaches are useful [1].

3.2. Topological-Based

There are many type of metrics exist to compute similarity between two nodes even without node or edge attributes. These metrics used topological information and known as topological-based measures. node-based approaches considers many attributes to work, while, topology-based approaches that are to be discussing in this section do not need any parameters or attributes of the edges and nodes. It had been named as topological-based methods as these methods requires only topological information instead of the information of nodes and edges [9]. There are many topology-based metrics proposed in the last few decades. Here, a general description of some of the most popular topological-based metrics are provided in the realm of link prediction problem. Based on the characteristics of these metrics, normally they can be divided into three categories namely neighbor-based metrics, path-based metrics, and random-walk-based metrics [9]. These metrics are categorized into local and global metrics.

Local Similarity Measures. In a SN, to estimate the similarity of each node with other nodes, local similarity-based methods rely on structural information like neighborhood. These methods are faster, effective and highly parallelizable as compare to nonlocal methods. Moreover, local methods enable us to adequately deal with link prediction issue in changing and dynamic networks like online SN [10]. The primary defect of these methods is that local information (such neighborhood)

restricts nodes to find contacts within neighbors of neighbors. In real world networks, it is shown that many connections between nodes are formed at greater distance (i.e., more than two). Nevertheless, local methods have shown competitive prediction results as compare to complex methods. In addition, it is noticeable that, although these approaches are restricted to two-hop, their time complexity is $O(xk2f(m))$, where $O(xk2)$ is spatial complexity and $f(m)$ is similarity computing. Table 1 shows the standard notations for link prediction used in this study.

Table 1. Common Notations used in the Link Prediction approaches

Description	Notation
Set of Neighbors of node x	$\Gamma(x)$
Set of Neighbors of node y	$\Gamma(y)$
Number of neighborhood of node x	$\|\Gamma(x)\|$
Number of neighborhood of node y	$\|\Gamma(y)\|$

- **Common Neighbors (CN):** This [27] is the most simplest and famous method to predict links in the network. It is considered as basics of all similarity-based methods as many as other had been derived from it. For a pair of nodes (x, y), the number of neighbors that both x and y have to interact with CN. The larger the amount of common neighbors the higher likelihood that x and y will be connected in the future. The Equation is as follows 1.

$$CN(u,v) = |\Gamma u \bigcap \Gamma v| \tag{1}$$

  Since common neighbor is not normalized, sometimes it can only reflect the relative similarities between a node pair. Therefore many other neighbor-based methods consider how to normalize the common neighbors as an improvement of the method as shown below.

- **Jaccard Coefficient:** Jaccard Coefficient is known as Jaccard Index, and is basically the normalizes the similarity score of common neighbors by considering intersection over union [20]. For similarity of two nodes u and v, it take in account the common neighbors and total neighbors of both nodes. Besides, Liben et al. showed that Jaccard produced worst results as compare to common neighbors. It can be defined as in Equation 2:

$$JC(u,v) = \frac{|\Gamma(u) \bigcap \Gamma(v)|}{|\Gamma(u) \bigcup \Gamma(v)|} \tag{2}$$

- **SAM:** This method is recently published by [34]. This works on the simple idea that both nodes have their own similarity, i.e., it is possible that one node is 100% similar to another node, but at the same time other node is not similar as first node. SAM similarity can be defined as Equation 3.

$$SAM(u,v) = \frac{\frac{|\Gamma(u) \bigcap \Gamma(v)|}{|\Gamma(u)|} + \frac{|\Gamma(u) \bigcap \Gamma(v)|}{|\Gamma(v)|}}{2} \tag{3}$$

- **Adamic Adar(AA):** Initially, this method was proposed to find similarity among two pages [3]. Later, Liben et al. [25] used the customize version for link prediction problem as shown in Equation 4. In fact, this measure torches the common neighbors along with high degree. It can be defined as in Equation 4.

$$AA(u,v) = \sum_{z \in \Gamma(u) \bigcap \Gamma(v)} \frac{1}{log|\Gamma(z)|} \tag{4}$$

- **Resource Allocation(RA):** This measure is inspired by the process of resource allocation in operating systems. Resource allocation is same as adamic adar, but it gives more punishment to common neighbors along with high degree [43]. This is why, both resource allocation and adamic adar have close results. Its foremost feature is that it considers neighbors of neighbors along with direct neighbors. It is defined as in Equation 5.

$$RA(u,v) = \sum_{z \in \Gamma(u) \bigcap \Gamma(v)} \frac{1}{|\Gamma(z)|} \tag{5}$$

- **Preferential Attachment(PA):** This method is proposed by Barbasi et al. [7]. Its main feature is new node will be connected with node having high degree instead of node with low degree. Method can be defined as in Equation 6.

$$PA(u,v) = |\Gamma(u)| \cdot |\Gamma(v)| \tag{6}$$

- **Sorenson Index(SI):** This method was proposed by Thorvald Sorensen to find similarity between data samples of ecological community [38]. The foremost objective of this method is to motivate the lower degree nodes in order to find their links. Similarity can be computed as in Equation 7.

$$SI(u,v) = \frac{|\Gamma(u) \bigcap \Gamma(v)|}{|\Gamma(u) + \Gamma(v)|} \tag{7}$$

- **Salton Cosine(SC):** This method is also known as cosine similarity [31]. This method is similar as Sørensen Index and Jaccard Index. Through some studies, it is found that value produces by Salton Cosine is twice the Jaccard Index . Value can be

computed as in Equation 8.

$$SC(u,v) = \frac{|\Gamma(u) \bigcap \Gamma(v)|}{\sqrt{|\Gamma(u) \cdot \Gamma(v)|}} \tag{8}$$

- **Hub Promoted (HP):** Hub Promoted measure proposed by Ravasz et al. during the study of metabolic network [30]. It defines overlap between nodes u and v on the base of topology. Similarity computation defined as in Equation 9.

$$HP(u,v) = \frac{|\Gamma(u) \bigcap \Gamma(v)|}{Min(|\Gamma(u)|, |\Gamma(v)|)} \tag{9}$$

- **Hub Depressed (HD):** This measure is same as Hub Promoted, but the similarity value can be computed by nodes with higher degree [44]. Similarity can be defined as in Equation 10.

$$HD(u,v) = \frac{|\Gamma(u) \bigcap \Gamma(v)|}{Max(|\Gamma(u)|, |\Gamma(v)|)} \tag{10}$$

- **Leicht-Holme-Nerman(LHN):** This measure assigns high similarity score to pair of nodes with more common neighbors [23]. This method takes in account the number of actual paths and number of expected paths of length two between two nodes. The authors claimed that it is more sensitive than others in terms of structural equivalence. Similarity can be computed as in Equation 11.

$$LHN(u,v) = \frac{|\Gamma(u) \bigcap \Gamma(v)|}{|\Gamma(u)| \cdot |\Gamma(v)|} \tag{11}$$

Global Similarity Measures. In order to estimate the similarity between pair of nodes, global similarity-based methods rely on whole structure of network [25]. These methods are not restricted to two node distance as local methods, however, their complexity make them impractical for large networks. In addition, their parallelization is more complex as whole topology of network may not be known by computational agent. Regardless, they shows very diverse time complexities, $O(k^2)$ is their spatial complexity as they store similarity score of each pair. Global similarity-based methods are further categorized into path-based and random walk.

Besides neighbors and nodes information, path is another feature that can be used to estimate similarity between nodes, and this feature is used in path-based methods. Similarity approaches are as follows:

- **Katz:** this method [21] is based on ensemble of all paths between two nodes. The paths are damped exponentially by length that can give more importance to shorter paths. The expression is defined as in Equation 12.

$$Katz(u,v) = \sum_{k=1}^{\infty} \beta^k \cdot |path_{u,v}^k| = \beta A + \beta^2 A^2 + \beta^3 A^3 + ... \tag{12}$$

Where, $|path_{u,v}^k|$ represents all paths of length k that are connecting x and y, and β is the damping factor that is controlling the weights of all paths. In case of very small β, Katz method behaves like Common Neighbors, since short paths perform extra ordinary in final similarity.

- **Friend Link:** This method finds similarity by traversing all the paths [28]. It works on the hypothesis that social network users can use all the paths between them. Therefore, similarity between pair of nodes u and v can be estimated as in Equation 13.

$$FL(u,v) = \sum_{i=1}^{l} \frac{1}{i-1} \cdot \frac{|path_{u,v}^i|}{\prod_{j=2}^{i}(n-j)} \tag{13}$$

Where, n is the size of network, l is the path length between x and y, $|path_{u,v}^i|$ denotes all paths between x and y with l length. In addition, higher l will cause for the poor performance.

Hybrid Similarity Measures. Hybrid approaches are those kinds of link prediction approaches that uses the global as well as local features simultaneously. To the best of my knowledge, there are very few studies found in the literature that uses the global and local features in the similarity computation. The hybrid approaches are as follows:

- **Parameter-Free (PF):** The parameter-free link prediction hybrid approach is proposed by [6]. In this metric they have used the path as global and neighborhood as local feature. They consider the number of paths between two nodes x and y. According to them, two persons can contact near in the future if they have more number of paths between them. The similarity using parameter-free approach can be computed as follows 14.

$$FL(u,v) = \frac{|Path_{xy}^{length \leq l}|}{|\Gamma(x)| + |\Gamma(y)| + 1} \tag{14}$$

Where,

  - $|Path_{xy}^{length \leq l}|$ is the number of simple paths between two nodes x and y of length l,
  - $\Gamma(y)$ represents the number of direct neighbors of node y.

4. Research Methodology

Numerous existing world domains can be describe by using networks, where nodes represent individuals and links show connection or relation between nodes (Huang et al., 2014). Example of an Email System like Enron contains 250,000 emails that connect 28,000 people, while in a single day AT&T a telephone call network records 355 million people for making 275 million calls. Lastly, co-authorship events in the research publications networks interprets over 770,000 authors of 730,000 research papers as cite seer archives.

Social networks are very dynamics in nature as new edges and nodes are added to graph at any time [37]. Social network Analysis has attracted majority of the researchers from various fields as it become a hot topic. Nowadays SNA highly depends on link prediction, also it is serving many other different domains with its applications. From the past decade, majority of the researcher showing the great interest the social network in the way in it is represented in a graphical form and the approach to predict the graph topology and similarity of the nodes [32]. This chapter discussed the detailed methodology of proposed technique for link prediction. In our proposed approach, local as well as global features have been utilized to find the similarity between pair of nodes. Furthermore, the results have been verified using five different dataset (i.e., Astro, GrQc, CondMat, HepTh, HepPh) [24]. To check the accuracy of proposed technique, accuracy measure has been used. Figure 3 shows a graphical representation of the methodology. Detail about each part of methodology is given below.

4.1. Dataset Description

In order to evaluate the link prediction algorithms based on a similarity measure, we have conducted the experiments on five collaboration network datasets. The description of the used networks is as follows:

- **Astro:** Arvix the Astro-PH (Astro Physics) cooperation network is taken from e-print arxiv as it cover all scientific collaboration between the authorś papers presented in the Astro-physics category. If the author i is the co-author of an article with another author j, the edge in graph be an undirected link from I to j. if the paper is written in collaboration with k authors, it creates a fully integrated (sub) graph on the k nodes. Data refers to documents from January 1993 to April 2003 (124 months). It begins in a few months after the launch of the Arvix and thus basically represents the entire history of its Astro-PH division.

- **GrQc:** Arvix the GR-QC (General Relative and Quantum Cosmology) cooperation network is taken from e-print arxiv as it cover all scientific collaboration between the authorś papers presented in the General Relative and Quantum Cosmology category. If the author i is the co-author of an article with another author j, the edge in graph be an undirected link from I to j. if the paper is written in collaboration with k authors, it creates a fully integrated (sub) graph on the k nodes. Data refers to documents from January 1993 to April 2003 (124 months). It begins in a few months after the launch of the Arvix and thus basically represents the entire history of its GR-QC division.

- **CondMat:** Arvix the COND-MAT (Condense Problem Physics) cooperation network is taken from e-print arxiv as it cover all scientific collaboration between the authorś papers presented in the Condense Problem Physics category. If the author i is the co-author of an article with another author j, the edge in graph be an undirected link from I to j. if the paper is written in collaboration with k authors, it creates a fully integrated (sub) graph on the k nodes. Data refers to documents from January 1993 to April 2003 (124 months). It begins in a few months after the launch of the Arvix and thus basically represents the entire history of its COND-MAT division.

- **HepPh:** Arvix the Hep-PH (High Energy Physics - Phenomenology) cooperation network is taken from e-print arxiv as it cover all scientific collaboration between the authorś papers presented in High Energy Physics Phenomenology category. If the author i is the co-author of an article with another author j, the edge in graph be an undirected link from I to j. if the paper is written in collaboration with k authors, it creates a fully integrated (sub) graph on the k nodes. Data refers to documents from January 1993 to April 2003 (124 months). It begins in a few months after the launch of the Arvix and thus basically represents the entire history of its Hep-PH division.

- **HepTh:** Arvix the Hep-TH (High Energy Physics-Theory) cooperation network is taken from e-print arxiv as it cover all scientific collaboration between the authorś papers presented in High Energy Physics Theory category. If the author i is the co-author of an article with another author j, the edge in graph be an undirected link from I to j. if the paper is written in collaboration with k authors, it creates a fully integrated (sub) graph on the k nodes. Data refers to documents from January 1993 to April 2003 (124 months). It begins in a few months after the launch of the Arvix and thus basically represents the entire history of its Hep-TH division.

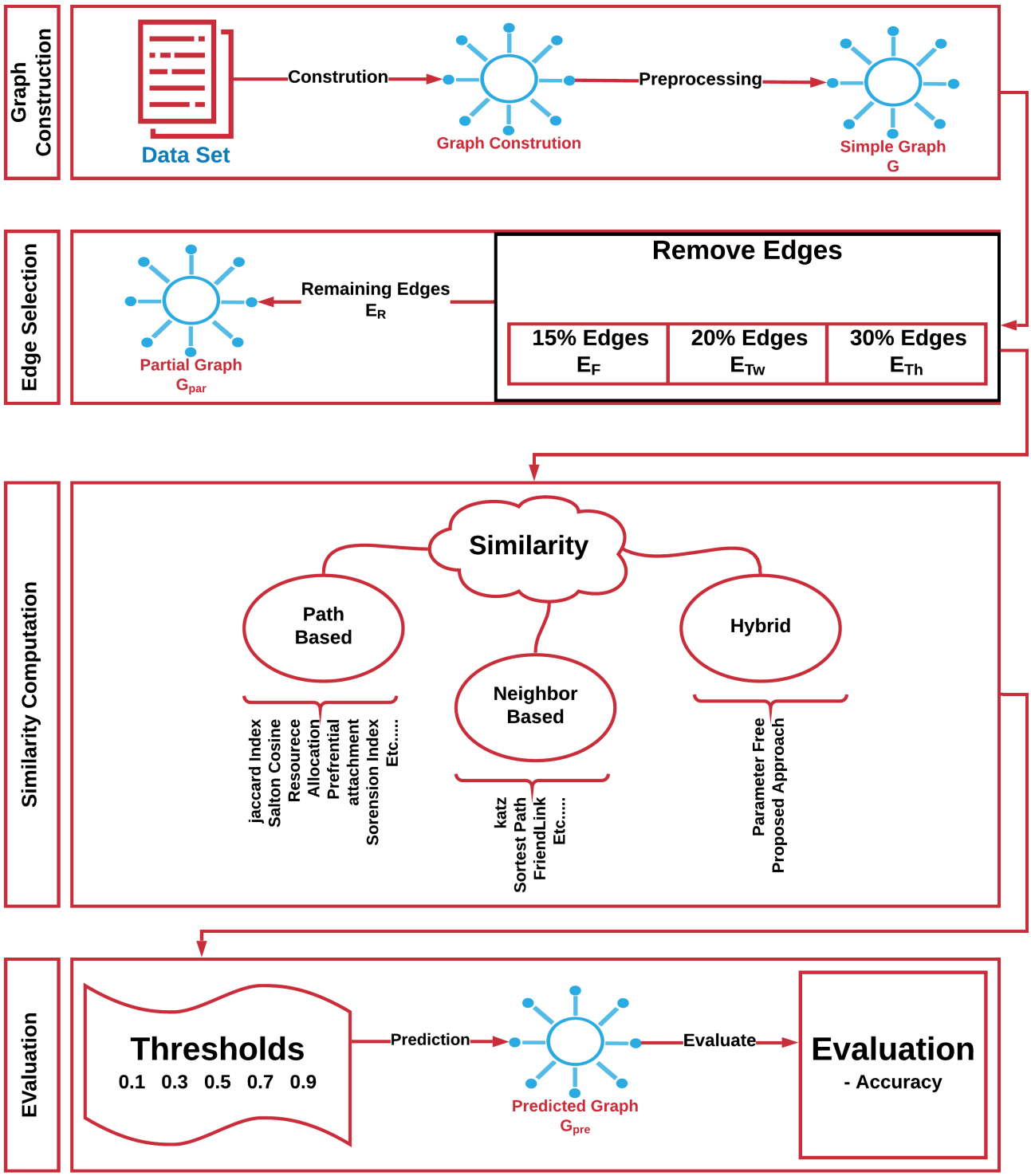

Figure 3. Research Methodology Diagram

In this way, all datasets are converted into excel format. Finally, the excel file is further used as a network in the experiments. A detailed statistics about each dataset are given in Table 2.

4.2. Pre-Processing

The datasets were received in the form of text file containing edge list of the network. Initially, transformed the dataset from text file into excel sheet

Table 2. Statistics of all dataset used in the research

Description	Astro	CondMat	GrQc	HepTh	HepPh
Nodes	18772	23133	5242	12008	9877
Edges	198110	93497	14496	118521	25998
Nodes in largest WCC	17903	21363	4158	11204	8638
Edges in largest WCC	197031	91342	13428	117649	24827
Nodes in largest SCC	17903	21363	4158	11204	8638
Edges in largest SCC	197031	91342	13428	117649	24827
Average Clustering Coefficient	0.6306	0.6334	0.5296	0.6115	0.4714
Number of Triangles	1351441	173361	48260	3358499	28339
Fraction of Closed Triangles	0.1345	0.107	0.3619	0.3923	0.1168
Diameter	14	14	17	13	17

to make it easy for graph construction. Initially, the version of datasets recieved in the form of multiple edges between a pair of nodes (as example shown in Figure 4(A)). In graph theory, multiple edges (also called parallel edges or a multi-edge), are, in an undirected graph, two or more edges that are incident to the same two vertices, or in a directed graph, two or more edges with both the same tail vertex and the same head vertex. These multiple edges considered as noise in similarity computation. On the other hand, a simple graph has no multiple edges. Therefore, in the next step, I have simplified the dataset into simple graph without having multiple edges (as an example shown in Figure 4(B)).

4.3. Graph Construction

In this study, five commonly used datasets (i.e., Astro, GrQc, CondMat, HepTh and HepPh) are being used. To construct the graphs of these datasets, the R tool is used which supports the igraph library. By importing excel files containing edge lists; the edge list placed in the graph construction function supported by Igraph library. After graph construction, five different network graphs obtained. As an example, Figure 3 represents the edge list, while, Figure 5 shows the graph constructed from the edge list.

4.4. Edge Lists Generation

In the previous section, a graph $G = (V, E)$ have been constructed from datasets (i.e., Astro, GrQc, CondMat, HepTh, HepPh), where V represents the set of nodes and E represents the set of edges. G is a simple graph i.e., no multiple edges or loop allowed. From the graph G of *Astro* dataset, here we defined three different subset of E such that E_R will be the remaining edges of G, E_{10} will be the 1000 random edges of G, E_{15} will be the 1500 random edges of G and $E20$ will be the 2000 random edges of G.

$$E = E_R \bigcup E_{10}$$

$$E = E_R \bigcup E_{15}$$

$$E = E_R \bigcup E_{20}$$

All these set of edges (I.e., $E10, E15$ and $E20$) will be used as ground truth in link prediction. Furthermore, we will remove these subsets (I.e., $E10, E15$ and $E20$) of edges one by one from the original graph G in order to make another partial graph G_{par} for each subset of edges. Graph G_{par} will be then used as a social network upon which similarity measures will be applied in order to compute similarity between each pair of nodes.

4.5. Proposed Approach (GLOS)

Introducing a new link prediction approach to demonstrate similarity among pair of nodes. Let a graph G have two nodes x and y; $S(x, y)$ calculates the similarity between x and y. Chances of the formation of link between x and y depends on their similarity value at the current time t. In this approach the main focuss is the number of paths between a pair of nodes and their neighborhood that contribute to provide the path between them. Let u and v be two users in the social network, u and v will interact if there are high number of paths between them and high, and then, there will be a chance that they will be connected in the future. In addition, let node x has k units to share with his neighbors, then each neighbor will get k(x), that is to say, the higher the degree (number of friends) on the network, the lower the number of units they will receive from the node x. This is why the proposed metric diminishes the role of nodes with high degree. The similarity $S(x, y)$ between two nodes x and y is

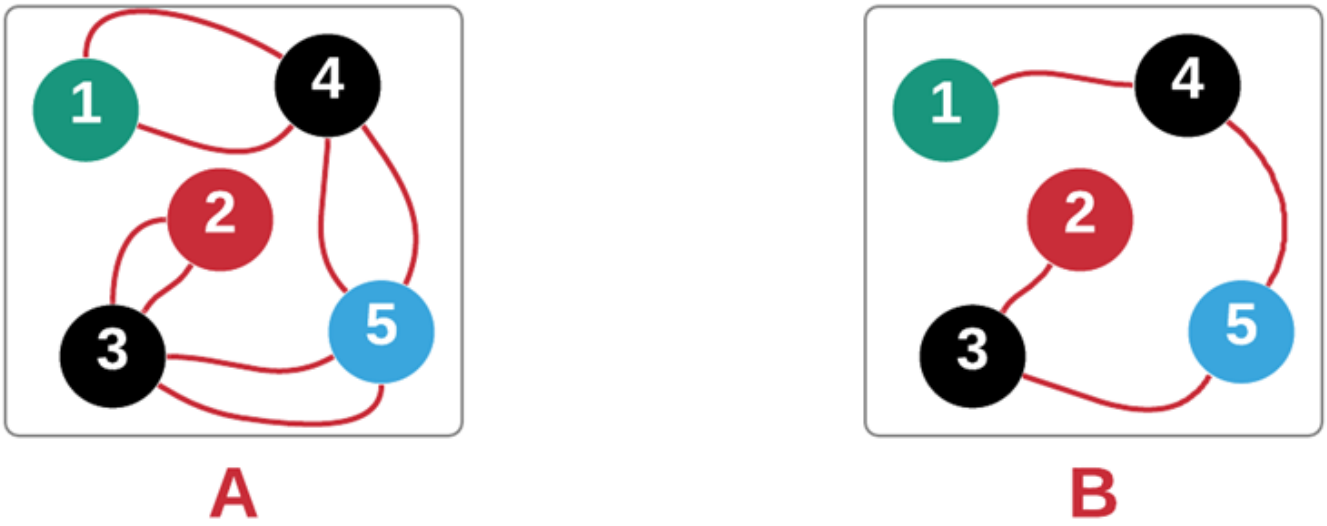

Figure 4. Example of graph with multiple edges and without multiple edges.

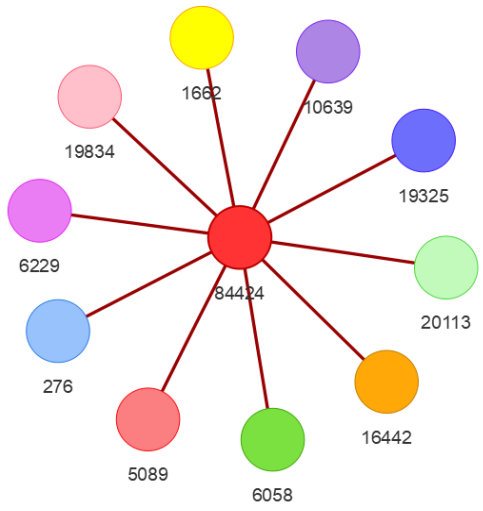

Figure 5. Example of Graph construction through edge list

defined as follows 15:

$$GLOS(u,v) = \frac{|Path^k_{x,y}|}{\Gamma(x) \bigcap \sigma(z)_{Path^k_{x,y}}} + \frac{|Path^k_{x,y}|}{\Gamma(y) \bigcap \sigma(z)_{Path^k_{x,y}}} \tag{15}$$

Where,

- $|Path^k_{x,y}|$ represents the count of simple paths of length k between two nodes x and y,
- $\sigma(z)_{Path^k_{x,y}}$ represents the number of neighbors that are participating in the paths,
- k represents the path length,
- $|\Gamma(x)|$ represents the length of neighbors of node x,
- $|\Gamma(y)|$ represents the length of neighbors of node y.

4.6. Similarity Computation

Similarity considers as great subject in human history since a long time ago. Even before machines invented, humans have been found in looking for similarity in everything. Similarity computation is the process of compute similarity of items and then to choose the most similar set of items. The simple and basic idea, that is consider by the majority of researchers, in similarity computation between two objects u and v is to first make a list of properties which belongs to these objects and then to apply a similarity computation technique to determine the similarity of u and v. Here, in this thesis, similarity between papers is computed on textual as well as topological parameters.

Path-Based. Path-based similarity is considers the path length between two nodes in graph. It uses the simple strategy: the pair of node will be similar if there are the more number of shortest paths between them, and vice versa. It can be computed by the similarity, where, each non-connected pair of nodes $(u, v$ is assigned a score signifying similarity between u and v. A high score indicates high probability that u will connect to v, while a low score also indicates high probability that u will not connect v. Therefore, using the rank of similarity scores, we can predict and recommend users for a friendship, talk, or any other relation. In a network graph, nodes can have many paths of different lengths. Here, path between nodes is known as a global feature within a graph.

Neighbor-Based. Neighbor-based similarity is a simplest approach to find similarity for the prediction

of link. It considers the neighborhood of the pair of nodes. The idea behind the neighbor-based similarity is that the more similar the pair is the more number of common neighbors between them, and vice versa. It also can be measured by the similarity, in which each non-connected pair of nodes (u, v) is assigned a score signifying similarity between u and v. Where, a high score represents high chances that there will be a link between u and v, while a low score also indicates high chances that there will be no link between u and v. Here, the neighbor of the node is known as a local feature of the node within a graph.

Hybrid. Majority of the studies, found in the literature, worked with global features (i.e., path) or local features (i.e., neighbors) separately. However, very few studies found which promoted the way of combining global and local features for similarity computation. The idea behind the hybrid approach is that the more similar the pair is the more number of common neighbors and paths between them. This thesis evaluated the both local and global features in order to predict the links. Moreover, a hybrid approach by combining local and global features is also proposed as shown in Equation 15.

4.7. Evaluation

In order to evaluate the proposed technique, accuracy measure is used. Model Accuracy is the ratio of number of correct predictions to the total number of input samples. Here in this thesis, the input is edges of the citation graph.

Accuracy. For the evaluation a model devised to compute the accuracy score between real graph and predicted graph. The accuracy score for the predicted graph Gp and real graph Gr is calculated using following the Equation 4.

$$Accuracy = 1 - \frac{E(Gr) + E(Gp - 2E(Gr \bigcap Gp)}{Max(E(Gr), E(Gp))} \quad (16)$$

Where,

- E represents the Edges of the citation graph,
- Gr is the original social network graph,
- Gp is the predicted social network graph,
- Max function will return the maximum number of edges from original and predicted social network graph.

Consider a graph Original in Figure 6. Let suppose the edge between A and C is removed in order to predict it again. After applying similarity, the graph A is predicted. As all the removed edges predicted, the accuracy of Predicted graph A will be 1 as follows.

$$Ac(A) = \frac{4 + 4 - 2(4)}{4} = 1$$

Similarly, suppose after applying similarity, the graph B predicted. Here, the removed edges could not predicted, the accuracy of graph B will be 0.75 as follows.

$$Ac(B) = \frac{4 + 3 - 2(3)}{4} = 0.75$$

5. Results and Discussion

5.1. Experimental Section

The experiments according to the methodology (as discussed in the previous chapter) are performed step-by-step. For the experiments, 5 different co-authors dataset (i.e. GrQc, Astro, CondMat, HepTh and HepPh) from the e-print arXiv are used. Initially, these dataset were contains the edge lists in the form of text file (.txt). First, the initial step was to copied these edge lists from text file into excel file manually. These edge lists contained edges i.e., two authors i and j, co-author a the edge lists shows a undirected edge from i to j. Secondly, the excel files of 5 dataset are imported in the R tool and created the graphs through Igraph library. After creation the graph, applied preprocessing steps including removal of duplicate edges and loops. Further these graphs have used in the link prediction analysis. Thirdly, a graph picked and extracted randomly 3 different set (i.e.,the first list consisted of 1000edges, the second 1500 and the third 2000) of edge list in order to predict these edges through similarity measures and made partial graph by removing these edges from original graph. Similarly, the edges picked for all dataset. Fourth, the similarity measures applied on partial graph by computing the similarity between edges that are extracted earlier. Finally, the thresholds applied and the accuracy computed for each similarity measure on each graph.

5.2. Similarity Computation

As discussed above, 15 different edge lists have extracted from 5 dataset and used for the similarity computation on partial graphs. For the similarity computation, local, global and hybrid similarity measures are used. Where, local similarity measures uses the local information such as neighbors, global similarity measures uses the global information such as path and random walk and hybrid similarity measures uses the global and local information such as combination of neighbor and path. Furthermore, the similarity through proposed technique GLOS is also computed. In the end, thresholds are applied on the similarity scores and predicted the edges.

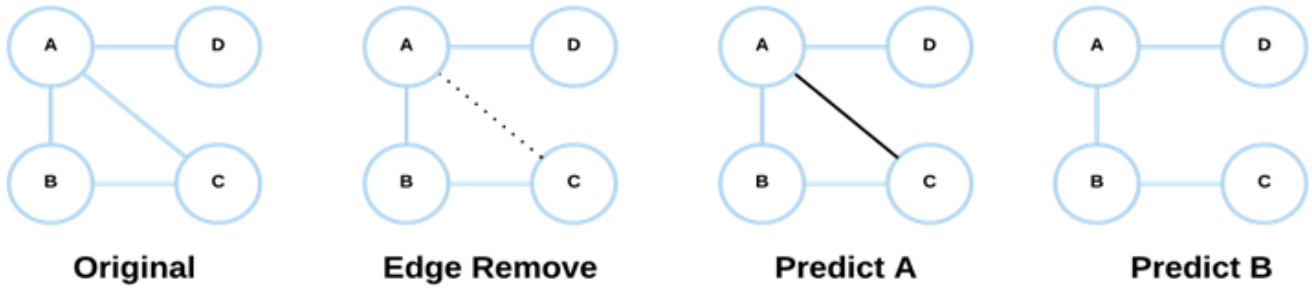

Figure 6. Accuracy Measure Example

Astro Dataset. This section presented the results of Astro dataset, from which, 3 different set of edge lists (i.e., Astro-1000, Astro-1500 & Astro-2000) extracted. After the extraction of these edge lists, these edges removed from the original graph and made partial graph upon which similarity measures (i.e., Hybrid, Global & Local) applied and computed the similarity between extracted edges. After applying the thresholds on similarity scores, the prediction accuracy of similarity measures are shown in Figures 7, 8 and 9. Where, Figure 7 shows the result of first edge list (i.e., Astro-1000) from Astro dataset. In Figure 7, pattern of the bars shows the thresholds (i.e., 0.2, 0.4, 0.6 and 0.8); similarity measures are shown on X-axis and on the Y-axis, accuracy of the similarity measures is shown. Furthermore, the X-axis is divided into three sections. Where, Hybrid section represents the hybrid similarity measures, Global section represents the global similarity measures and Local section represents the local similarity measures. And the same pattern is followed in all the remaining figures. The resultant thresholds shown that the highest accuracy is achieved at threshold 0.2. Where, both hybrid approaches GLOS and PF succeed in getting accuracy 1. On the other hand, at threshold 0.2, from the Local approaches AA, SAM, SI, SC and PD achieved highest accuracy by 0.848. Similarly, at threshold 0.2, from the Global approaches, both Katz and FL obtained highest accuracy by 0.832. At threshold 0.4, the hybrid approach GLOS achieved highest accuracy by 0.989 and global approaches Katz and FL achieved the lowest accuracy by 0.571. However, at threshold 0.4, local approaches performed better than global approaches, where, HP obtained the highest accuracy by 0.825. At threshold 0.6, again hybrid approach GLOS obtained highest accuracy and global approaches could not performed well. The main and interesting thing which can be seen in Figure 1 is the GLOS similarity. In case of all the thresholds, except 0.8, GLOS outperformed than rest of the similarity measures. While, at the threshold 0.8, local approach HP obtained highest results by 0.703 and hybrid approach GLOS achieved the second highest results by 0.669. Result of second edge list Astro-1500 is shown in Figure 8. Compared to the previous edge list Astro-1000, almost all the similarity measures improved their results here. GLOS is the only measures who performed well at all thresholds except 0.8, where, HP outperformed than other. At threshold 0.2, hybrid approaches GLOS and PF obtained accuracy 1, both global approaches Katz and FL achieved 0.767 and local approach AA succeeded in getting 0.861. Similarly, at threshold 0.2, hybrid approaches GLOS and PF achieved the highest accuracy and global approaches Katz and FL obtained the lowest accuracy. Moreover, at threshold 0.4, hybrid approach GLOS achieved the highest result by 0.986 and global approaches Katz and FL obtained the lowest accuracy by 0.541. Likewise, at threshold 0.6, again highest accuracy achieved by GLOS and both global approaches (i.e., Katz and FL) obtained the lowest accuracy. However, at threshold 0.8, hybrid approach GLOS could not perform well and obtained only 0.657, while, local approach HP succeeded in getting the high accuracy by 0.727. At threshold 0.8, global approaches Katz and FL were the only who obtained the lowest accuracy by 0.358. Figure 9 showing the result of third edge list Astro-2000 from Astro dataset. At threshold 0.2, the hybrid approaches GLOS and PF obtained the highest accuracy by 1 and global approaches Katz and FL succeeded in getting the lowest accuracy by 0.763. Similarly, at threshold 0.4, again hybrid approach GLOS obtained the highest accuracy by 0.985, while, global approaches Katz and FL achieved the lowest results by 0.539. On the other hand, at threshold 0.6, local approach LHN performed better than rest of the local and global approaches, where, LHN obtained accuracy 0.764. Similarly, at threshold 0.8, Local approach obtained the highest score by 0.725, while, RA obtained the second highest accuracy by 0.761. Overall, the performance of GLOS was better than rest of the approaches. Moreover, a detailed stats about accuracy on Astro dataset are presented in Table 3.

CondMat Dataset. Here, the results of ContMat dataset are presented, from which, three different set of

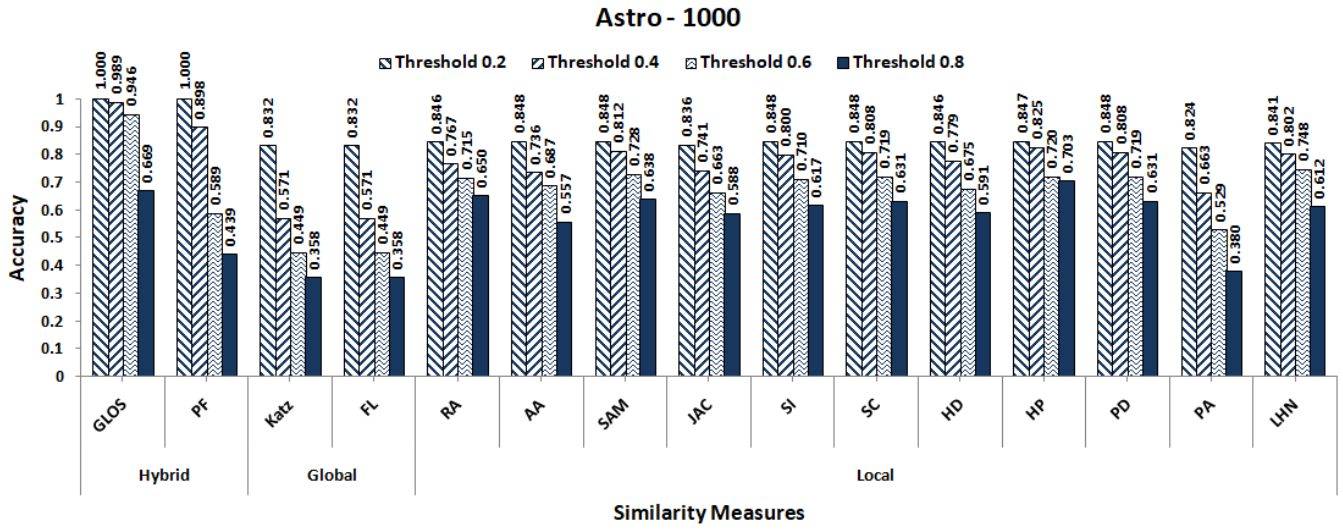

Figure 7. Comparisons of similarity measures on Astro-1000 edge list from Astro dataset

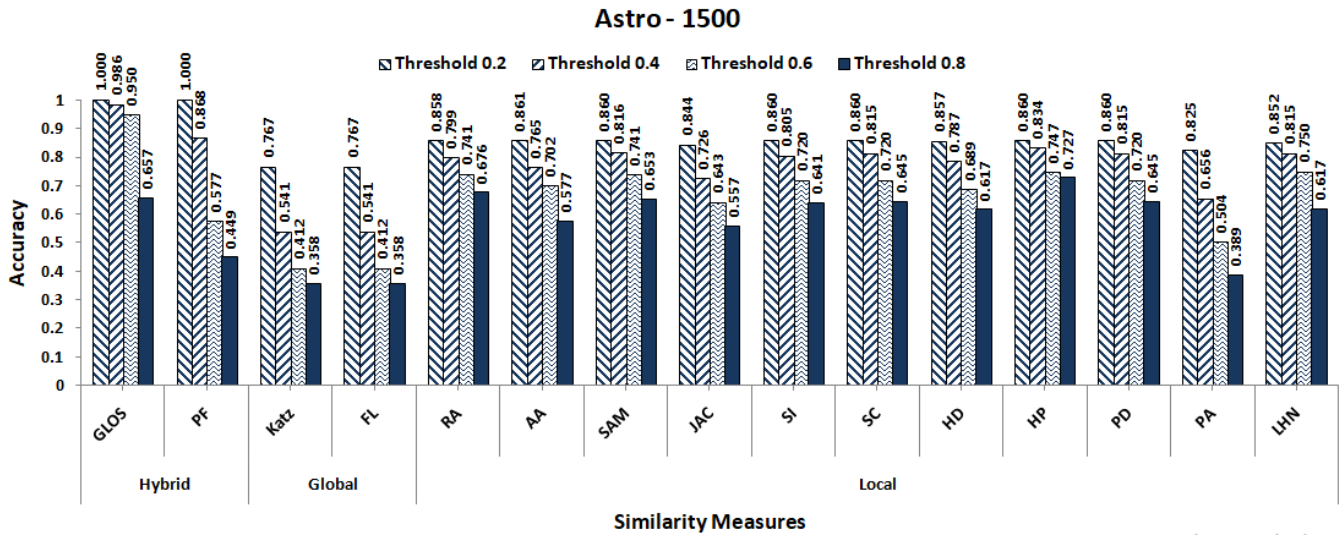

Figure 8. Comparisons of similarity measures on Astro-1500 edge list from Astro dataset

edge lists (i.e., ContMat -1000, ContMat -1500 & ContMat -2000) have extracted. In order to predict these edges, the edges from the original graph are removed and made partial graph upon which similarity measures (i.e., Hybrid, Global & Local) are applied and computed the similarity between these edges. Further the thresholds on similarity scores are applied, and presented the prediction accuracy of similarity measures in Figures 10, 11 and 12. Where, Figure 13 shows the result of first edge list (i.e., ContMat -1000) from ContMat dataset. In Figure 10, pattern of the bars shows the thresholds, similarity measures are shown on X-axis and accuracy of similarity measures is shown on Y-axis. And the same pattern is followed in all the remaining figures of ContMat dataset. At threshold 0.2, the highest accuracy achieved, while, the lowest accuracy is achieved at threshold 0.8. Where, at threshold 0.2, both hybrid approaches GLOS and PF succeed in getting accuracy 1. On the other hand, at threshold 0.2, from the Local approaches SAM, SC and PD achieved highest accuracy by 0.929. Similarly, at threshold 0.2, from the Global approaches, both Katz and FL obtained highest accuracy by 0.807. At threshold 0.4, the hybrid approach GLOS achieved highest accuracy by 0.998 and local approach PA achieved the lowest accuracy by 0.402. However, at threshold 0.4, rest of the local approaches performed better than global approaches, where, HP obtained the highest accuracy by 0.829. At threshold 0.6, again hybrid approach GLOS obtained highest accuracy by 0.971 and local approach PA could not perform well and obtained 0.319. The main and interesting thing

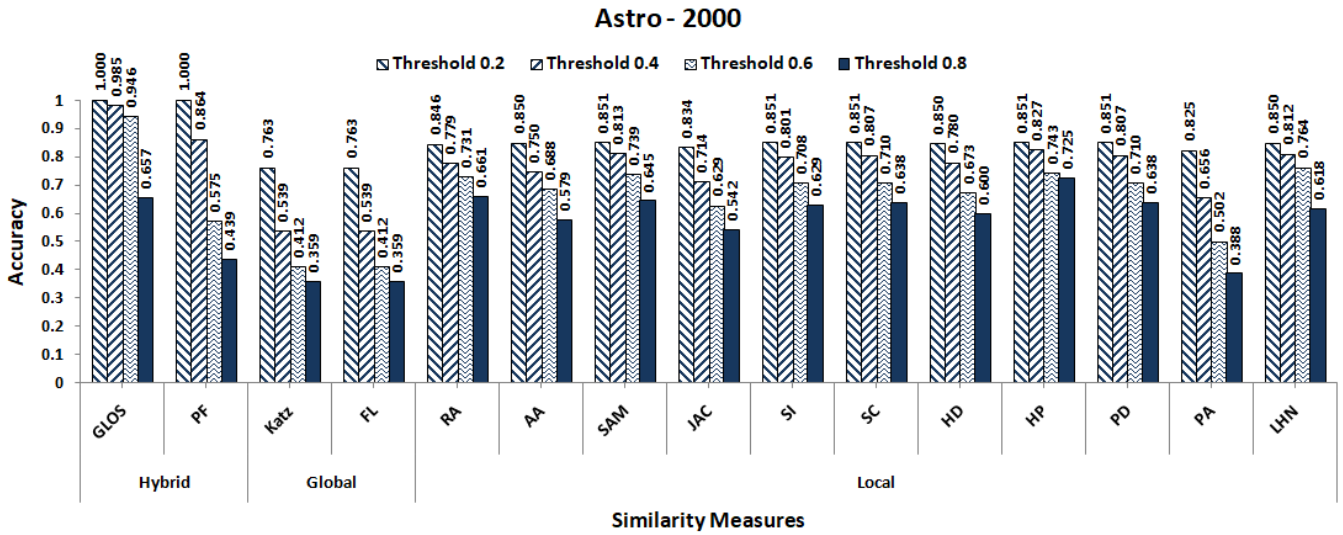

Figure 9. Comparisons of similarity measures on Astro–2000 edge list from Astro dataset

Table 3. The table presents the results of accuracy on Astro Dataset. Where, (T) represents the threshold, High represents the similarity measures obtaining highest accuracy, 2nd High corresponds to the similarity measures obtained second highest accuracy and low represents the similarity measures obtained lowest accuracy. The red color shows the Global approaches, the blue color shows the Local approaches and the green color shows the Hybrid approaches.

Edge List	(T)	High	2nd High	Low
Astro-1000	0.2	GLOS/PF	AA/SAM/SI/SC/PD	FL/KATZ
	0.4	GLOS	PF	FL/KATZ
	0.6	GLOS	LHN	FL/KATZ
	0.8	HP	GLOS	FL/KATZ
Astro-1500	0.2	GLOS/PF	AA	FL/KATZ
	0.4	GLOS	PF	FL/KATZ
	0.6	GLOS	HP	FL/KATZ
	0.8	HP	RA	FL/KATZ
Astro-2000	0.2	GLOS/PF	PD/HP/SC/SI/SAM	FL/KATZ
	0.4	GLOS	PF	FL/KATZ
	0.6	GLOS	LHN	FL/KATZ
	0.8	HP	RA	FL/KATZ

which can be seen in Figure 10 is the GLOS similarity. In case of all the thresholds, GLOS outperformed than rest of the similarity measures. As compared to the previous dataset, GLOS improved the accuracy, while, local approach PA degraded its results. While, at the threshold 0.8, hybrid approach GLOS obtained the highest accuracy by 0.780 and local approach HP achieved the second highest results by 0.646. Figure 11 showing the result of second edge list ContMat -1500 from ContMat dataset. Compared to the previous Astro dataset, almost all the similarity measures improved their results here; however, local approach PA reduced its results. GLOS is the only measures who performed well at all thresholds, while, PA obtained low accuracy than other similarity measures. At threshold 0.2, hybrid approaches GLOS and PF obtained accuracy 1, both global approaches Katz and FL achieved 0.804 and local approach SAM succeeded in getting the high accuracy by 0.861 than rest of the local approaches. Similarly, at threshold 0.2, hybrid approaches GLOS and PF achieved the highest accuracy and local approach PA obtained the lowest accuracy. Moreover, at threshold 0.4, hybrid approach GLOS achieved the highest result by 0.999 and local approach PA obtained the lowest accuracy by 0.427. On the other hand, at threshold 0.4, global approaches Katz and FL obtained the second lowest accuracy by 0.499. Likewise, at threshold 0.6, again highest accuracy achieved by GLOS and local approach PA obtained the lowest accuracy. Overall, at all the thresholds, hybrid approach GLOS

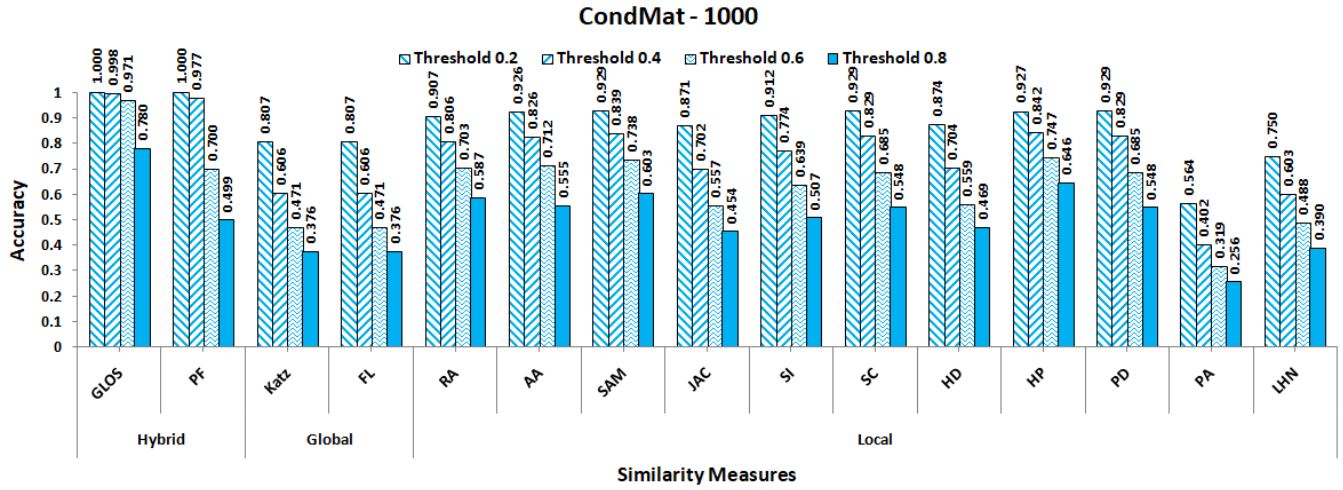

Figure 10. Comparisons of similarity measures on CondMat-1000 edge list from CondMat dataset

outperformed than others and PA degraded its results by obtaining lowest accuracy. At threshold 0.8, hybrid approach GLOS performed well and obtained only 0.759, while, local approach SAM obtained the second highest accuracy by 0.618. On the other hand, at threshold 0.8, local approach PA was the only similarity measure who obtained the lowest accuracy by 0.283. Figure 12 showing the result of third edge list ContMat -2000 from ContMat dataset. At threshold 0.2, the hybrid approaches GLOS and PF obtained the highest accuracy by 1 and local approach PA succeeded in getting the lowest accuracy by 0.579. Similarly, at threshold 0.4, the hybrid approach GLOS obtained the highest accuracy by 0.998, while, global local approach PA achieved the lowest results by 0.415. On the other hand, at threshold 0.6, local approach SAM performed better than rest of the local and global approaches, where, LHN obtained accuracy 0.738. Similarly, at threshold 0.8, hybrid approach GLOS obtained the highest score by 0.775, while, HP obtained the second highest accuracy by 0.648. Overall, the performance of GLOS was better than rest of the approaches. Detailed about accuracy on CondMat dataset is presented in Table 4.

GrQc Dataset. Here, the results of GrQc dataset are presented, from which, three different set of edge lists (i.e., GrQc-1000, GrQc -1500 & GrQc -2000) have used for the prediction. In order to predict these edges, the edges from the original graph are removed and made partial graph upon which different types of similarity measures (i.e., Hybrid, Global & Local) are applied and computed the similarity between these edges. After that the thresholds on similarity scores are applied, and presented the prediction accuracy of similarity measures in Figures 13, 14 and 15. Where, Figure 13 shows the result of first edge list (i.e., GrQc -1000) from GrQc dataset. In Figure 13, pattern of the bars shows the thresholds, similarity measures are shown on X-axis and accuracy of similarity measures is shown on Y-axis. And the same pattern is followed in all the remaining figures of GrQc dataset. At threshold 0.2, the highest accuracy achieved, while, the lowest accuracy is achieved at threshold 0.8. Where, at threshold 0.2, both hybrid approaches GLOS obtained the maximum accuracy by 1. On the other hand, at threshold 0.2, from the Local approach HP achieved the highest accuracy by 0.868. Similarly, at threshold 0.2, from the Global approaches, both Katz and FL obtained accuracy by 0.350. At threshold 0.4, the hybrid approach GLOS achieved highest accuracy by 0.991 and global approaches Katz and FL achieved the lowest accuracy by 0.307. However, at threshold 0.4, the local approaches performed better than global approaches, where, SAM obtained the highest accuracy by 0.797. At threshold 0.6, again hybrid approach GLOS obtained highest accuracy by 0.890 and global approaches Katz and FL could not perform well and obtained 0.275. The main and interesting thing which can be seen in Figure 7 is the GLOS similarity. In case of all the thresholds, except threshold 0.8, GLOS outperformed than rest of the similarity measures. At threshold 0.8, our proposed approach GLOS could not perform well and obtained accuracy by 0.563. On the other hand, at threshold 0.8, HP obtained the highest accuracy by 0.694. As compared to the previous dataset, GLOS maintained its accuracy, while, global approaches degraded its results. Surprisingly, the hybrid approach PF degraded its results for the thresholds 0.4, 0.6 and 0.8. Overall, at the threshold 0.8, local approach HP obtained the highest accuracy by 0.694 and global approaches Katz and FL obtained lowest results by 0.260.

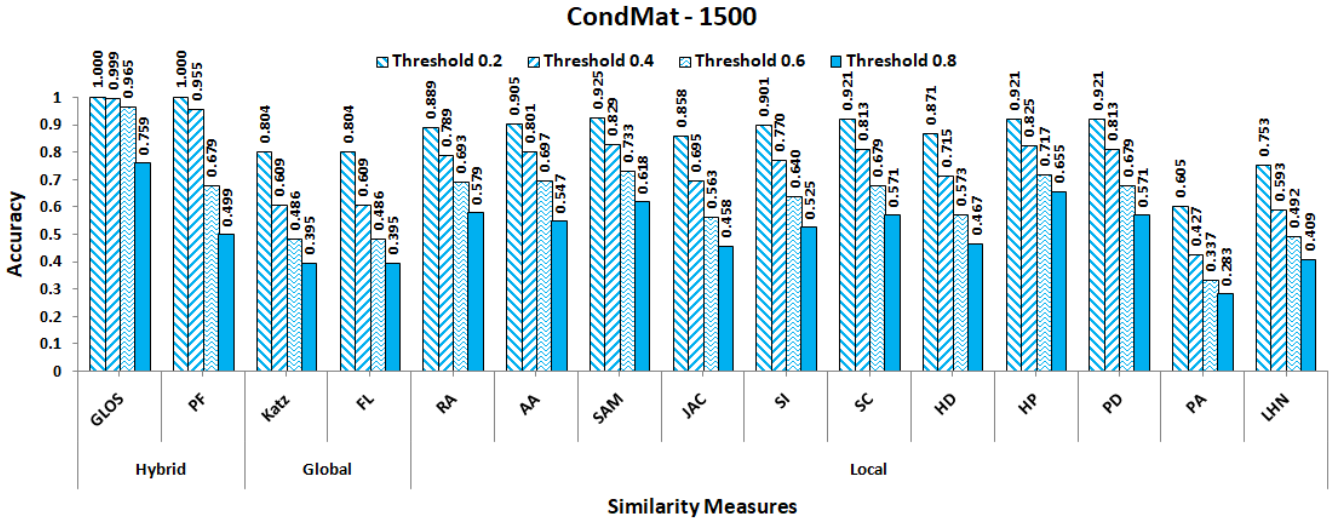

Figure 11. Comparisons of similarity measures on CondMat-1500 edge list from CondMat dataset

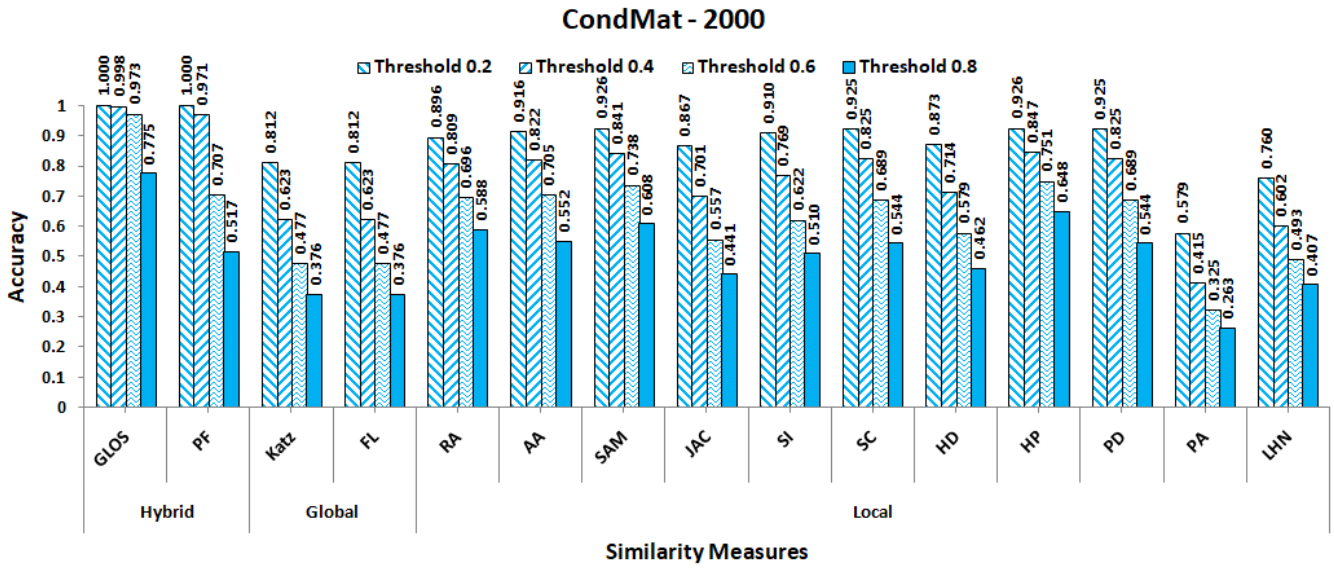

Figure 12. Comparisons of similarity measures on CondMat-2000 edge list from CondMat dataset

Figure 14 showing the result of second edge list GrQc-1500 from GrQc dataset. Compared to the previous CondMat dataset, almost all the similarity measures degraded their results here. GLOS is the only measures who performed well at all thresholds except 0.8, while, global approaches Katz and FL obtained lowest accuracy than other similarity measures. At threshold 0.2, hybrid approach GLOS obtained accuracy 0.999, both global approaches Katz and FL achieved 0.767 and local approach SAM succeeded in getting the high accuracy by 0.861 than rest of the local approaches. Similarly, at threshold 0.2, hybrid approach GLOS achieved the highest accuracy and global approaches Katz and FL obtained the lowest accuracy. Moreover, at threshold 0.4, hybrid approach GLOS achieved the highest result by 0.987 and global approaches Katz and FL obtained the lowest accuracy by 0.308. On the other hand, at threshold 0.4, local approach PA obtained the second lowest accuracy by 0.375. Likewise, at threshold 0.6, again highest accuracy achieved by GLOS and global approaches Katz and FL obtained the lowest accuracy. Overall, at all the thresholds, except threshold 0.8, hybrid approach GLOS outperformed than others and global approaches Katz and FL degraded its results by obtaining lowest accuracy. At threshold 0.8, local approach HP performed well and obtained only 0.689, while, local approach SAM obtained the second highest accuracy by 0.584.

Figure 15 showing the result of third edge list GrQc-2000 from GrQc dataset. At threshold 0.2, the hybrid

Table 4. The table presents the results of accuracy on CondMat Dataset. Where, (T) represents the threshold, High represents the similarity measures obtaining highest accuracy, 2nd High corresponds to the similarity measures obtained second highest accuracy and low represents the similarity measures obtained lowest accuracy. The red color shows the Global approaches, the blue color shows the Local approaches and the green color shows the Hybrid approaches.

Edge List	(T)	High	2nd High	Low
CondMat-1000	0.2	GLOS/PF	SAM/SC/PD	PA
	0.4	GLOS	PF	PA
	0.6	GLOS	HP	PA
	0.8	GLOS	HP	PA
CondMat-1500	0.2	GLOS/PF	AA	PA
	0.4	GLOS	PF	PA
	0.6	GLOS	SAM	PA
	0.8	GLOS	HP	PA
CondMat-2000	0.2	GLOS/PF	PD/HP/SC/SI/SAM	PA
	0.4	GLOS	PF	PA
	0.6	GLOS	HP	PA
	0.8	GLOS	HP	PA

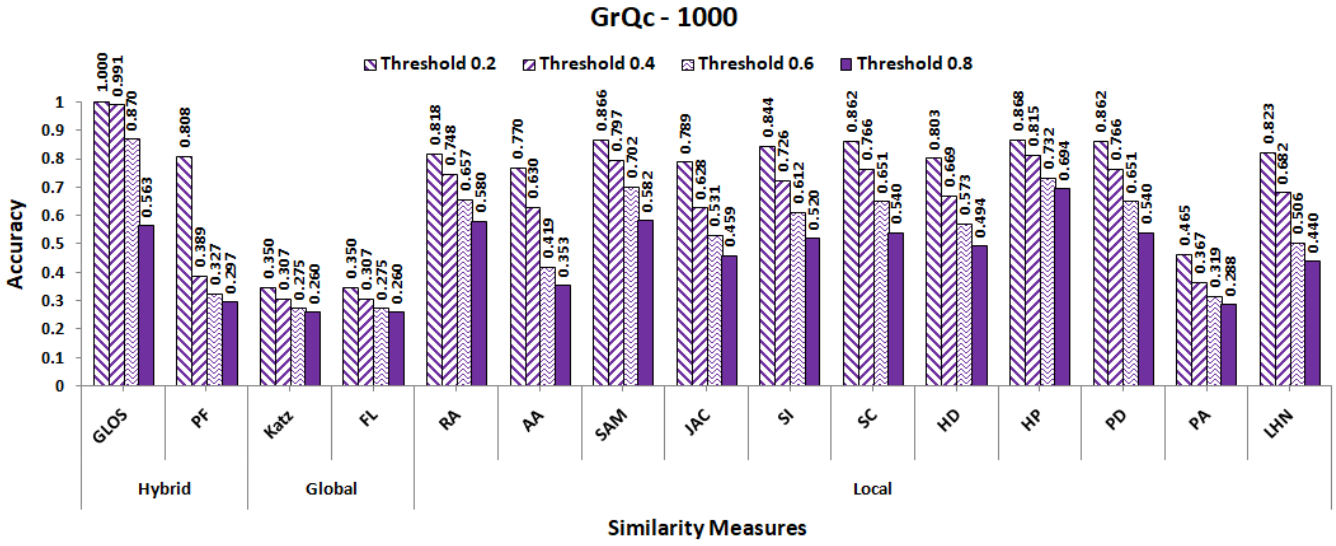

Figure 13. Comparisons of similarity measures on GrQc-1000 edge list from GrQc dataset

approaches GLOS obtained the highest accuracy by 1 and global approaches Katz FL succeeded in getting the lowest accuracy by 0.354. Similarly, at threshold 0.4, the hybrid approach GLOS obtained the highest accuracy by 0.990, while, global approaches Katz and FL achieved the lowest results by 0.304. On the other hand, at threshold 0.6, local approach HP performed better than rest of the local and global approaches, where, HP obtained accuracy 0.725. Similarly, at threshold 0.8, local approach HP obtained the highest score by 0.688, while, SAM obtained the second highest accuracy by 0.575. Overall, the performance of GLOS was better than rest of the approaches, however, HP outperformed than rest of the approaches. Detailed information is presented in Table 5.

HepPh Dataset. The result of HepPh dataset is presented here, from which, three different set of edge lists (i.e., HepPh-1000, HepPh-1500 & HepPh-2000) have extracted. In order to predict these edges, the edges from the original graph are removed and made partial graph upon which similarity measures (i.e., Hybrid, Global & Local) are applied and computed the similarity between these edges. Further, the thresholds on similarity scores are applied, and presented the prediction accuracy of similarity measures in Figures 16, 17 and 18. Where, Figure 16 shows the result of first edge list (i.e., HepPh-1000) from HepPh dataset.

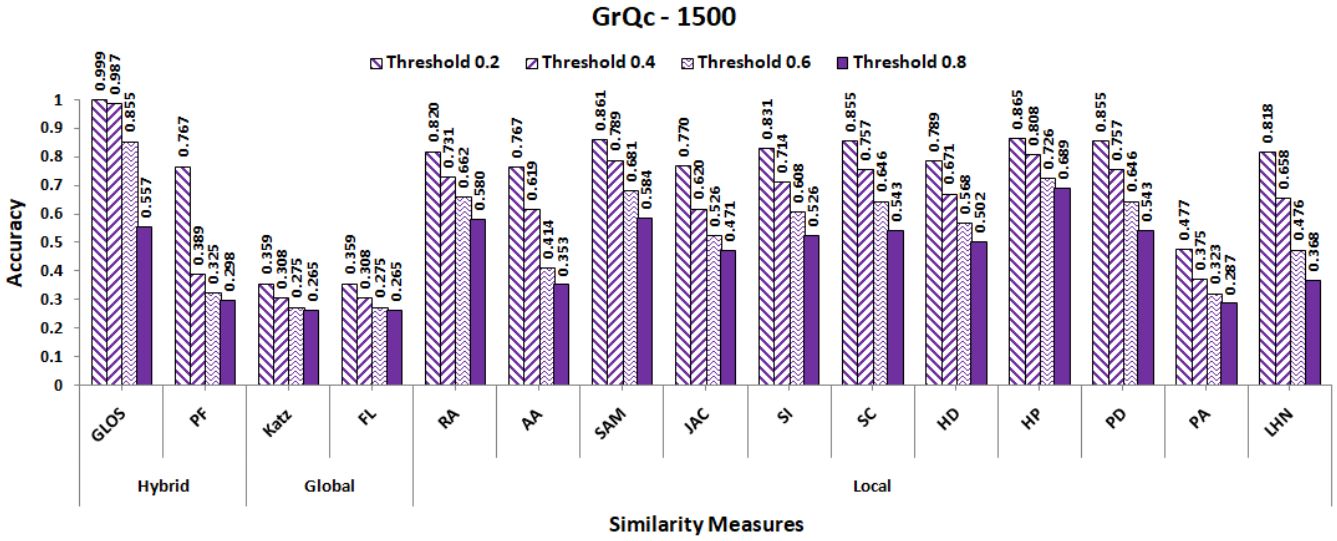

Figure 14. Comparisons of similarity measures on GrQc-1500 edge list from GrQc dataset

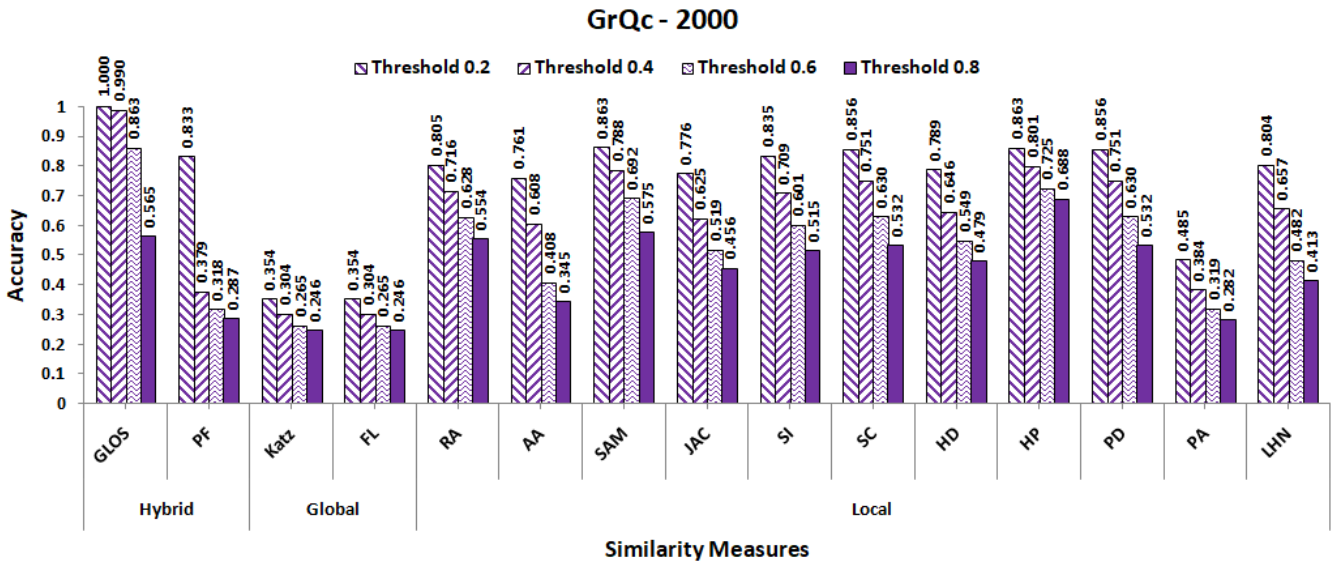

Figure 15. Comparisons of similarity measures on GrQc-1500 edge list from GrQc dataset

In Figure 16, pattern of the bars shows the thresholds, similarity measures are shown on X-axis and accuracy of similarity measures is shown on Y-axis. And the same pattern is followed in all the remaining figures of HepPh dataset. At threshold 0.2, the highest accuracy achieved, while, the lowest accuracy is achieved at threshold 0.8. Where, at threshold 0.2, both hybrid approaches GLOS and PF succeed in getting accuracy 1. On the other hand, at threshold 0.2, from the Local approaches PA achieved highest accuracy by 0.876. Similarly, at threshold 0.2, from the Global approaches, both Katz and FL obtained accuracy by 0.857. At threshold 0.4, the hybrid approach GLOS achieved highest accuracy by 0.997 and global approaches Katz and FL achieved the lowest accuracy by 0.602. However, at threshold 0.4, HP outperformed than rest of the local and global approaches, where, HP obtained the highest accuracy by 0.759. At threshold 0.6, again hybrid approach GLOS obtained highest accuracy by 0.971 and global approaches Katz and FL achieved the lowest accuracy by 0.432. The main and interesting thing which can be seen in Figure 10 is the GLOS similarity. In case of all the thresholds, GLOS outperformed than rest of the similarity measures. As compared to the GrQc dataset, GLOS as well as rest of the similarity measures improved their accuracy. While, at the threshold 0.8, hybrid approach GLOS obtained the highest accuracy

Table 5. The table presents the results of accuracy on GrQc Dataset. Where, (T) represents the threshold, High represents the similarity measures obtaining highest accuracy, 2nd High corresponds to the similarity measures obtained second highest accuracy and low represents the similarity measures obtained lowest accuracy. The red color shows the Global approaches, the blue color shows the Local approaches and the green color shows the Hybrid approaches.

Edge List	(T)	High	2nd High	Low
GrQc-1000	0.2	GLOS	HP	FL/KATZ
	0.4	GLOS	HP	FL/KATZ
	0.6	GLOS	HP	FL/KATZ
	0.8	HP	SAM	FL/KATZ
GrQc-1500	0.2	GLOS	HP	FL/KATZ
	0.4	GLOS	HP	FL/KATZ
	0.6	GLOS	HP	FL/KATZ
	0.8	HP	SAM	FL/KATZ
GrQc-2000	0.2	GLOS	HP	FL/KATZ
	0.4	GLOS	HP	FL/KATZ
	0.6	GLOS	HP	FL/KATZ
	0.8	HP	SAM	FL/KATZ

by 0.681 and local approach HP achieved the second highest results by 0.606.

Figure 17showing the result of second edge list HepPh-1500 from HepPh dataset. Compared to the previous GrQc dataset, almost all the similarity measures improved their results here. GLOS is the only measures who performed well at all thresholds, while, global approaches Katz and FL obtained low accuracy than other similarity measures. At threshold 0.2, hybrid approaches GLOS and PF obtained accuracy 1, both global approaches Katz and FL achieved 0.875 and local approach PA succeeded in getting the high accuracy by 0.884 than rest of the local approaches. Similarly, at threshold 0.2, hybrid approach GLOS achieved the highest accuracy and global approaches Katz and FL obtained the lowest accuracy. Moreover, at threshold 0.4, hybrid approach GLOS achieved the highest result by 0.997 and global approaches Katz and FL obtained the lowest accuracy by 0.557. On the other hand, at threshold 0.4, local approach JAC obtained the second lowest accuracy by 0.659. Likewise, at threshold 0.6, again highest accuracy achieved by GLOS and global approaches Katz and FL obtained the lowest accuracy. Overall, at all the thresholds, hybrid approach GLOS outperformed than others and global approaches Katz and FL produced lowest accuracy. At threshold 0.8, hybrid approach GLOS performed well and obtained only 0.676, while, local approach HP obtained the second highest accuracy by 0.614. On the other hand, at threshold 0.8, global approaches Katz and FL were the only similarity measures which obtained the lowest accuracy by 0.315.

Figure 17 showing the result of third edge list HepPh-2000 from HepPh dataset. At threshold 0.2, the hybrid approaches GLOS and PF obtained the highest accuracy by 1 and local approach JAC succeeded in getting the lowest accuracy by 0.803. Similarly, at threshold 0.4, the hybrid approach GLOS obtained the highest accuracy by 0.997, while, global approaches Katz and FL achieved the lowest results by 0.529. On the other hand, at threshold 0.6, local approach HP performed better than rest of the local and global approaches, where, HP obtained accuracy 0.711. Similarly, at threshold 0.8, hybrid approach GLOS obtained the highest score by 0650, while, HP obtained the second highest accuracy by 0.625. Overall, the performance of GLOS was better than rest of the hybrid, global and local approaches. Detailed information is presented in Table 6.

HepTh Dataset. Here, the results of HepTh dataset are presented, from which, three different set of edge lists (i.e., HepTh-1000, HepTh-1500 & HepTh-2000) have extracted. In order to predict these edges, the edges from the original graph are removed and made partial graph upon which similarity measures (i.e., Hybrid, Global & Local) are applied and computed the similarity between these edges. Furthermore, the thresholds on similarity scores are applied, and presented the prediction accuracy of similarity measures in Figures 19, 20 and 21. Where, Figure 22 shows the result of first edge list (i.e., HepTh-1000) from HepTh dataset. In Figure 19, pattern of the bars shows the thresholds, similarity measures are shown on X-axis and accuracy of similarity measures is shown on Y-axis. And the same pattern is followed in all the remaining figures of HepTh dataset. At threshold 0.2, the highest accuracy achieved, while, the lowest accuracy is achieved at threshold 0.8. Where, at threshold 0.2, both hybrid approaches GLOS and PF

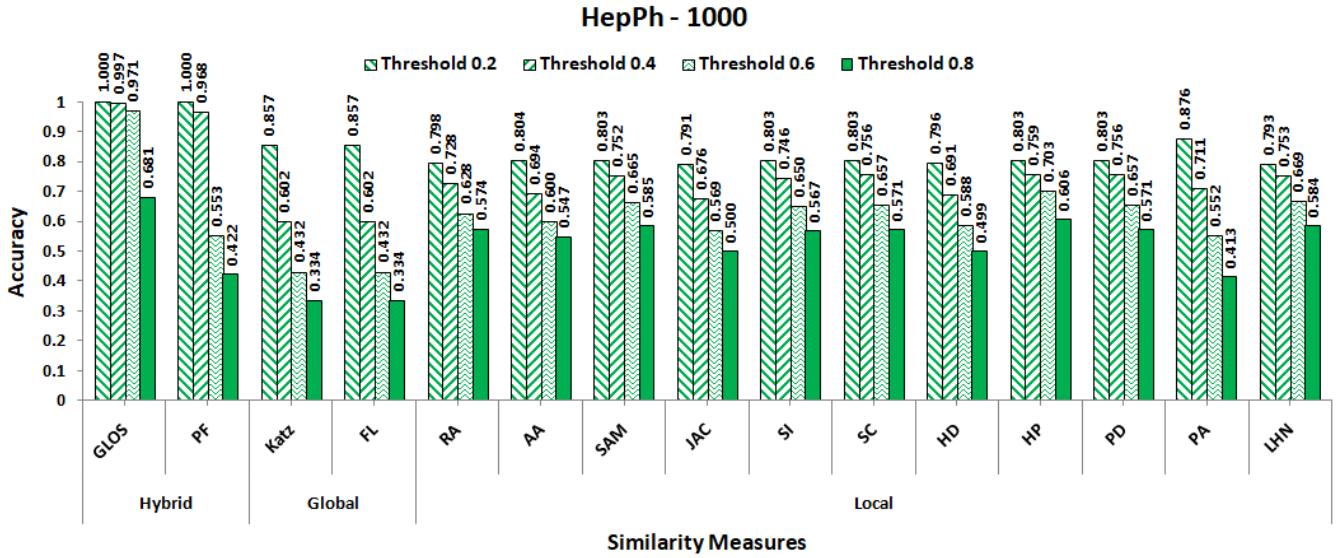

Figure 16. Comparisons of similarity measures on HepPh-1000 edge list from HepPh dataset

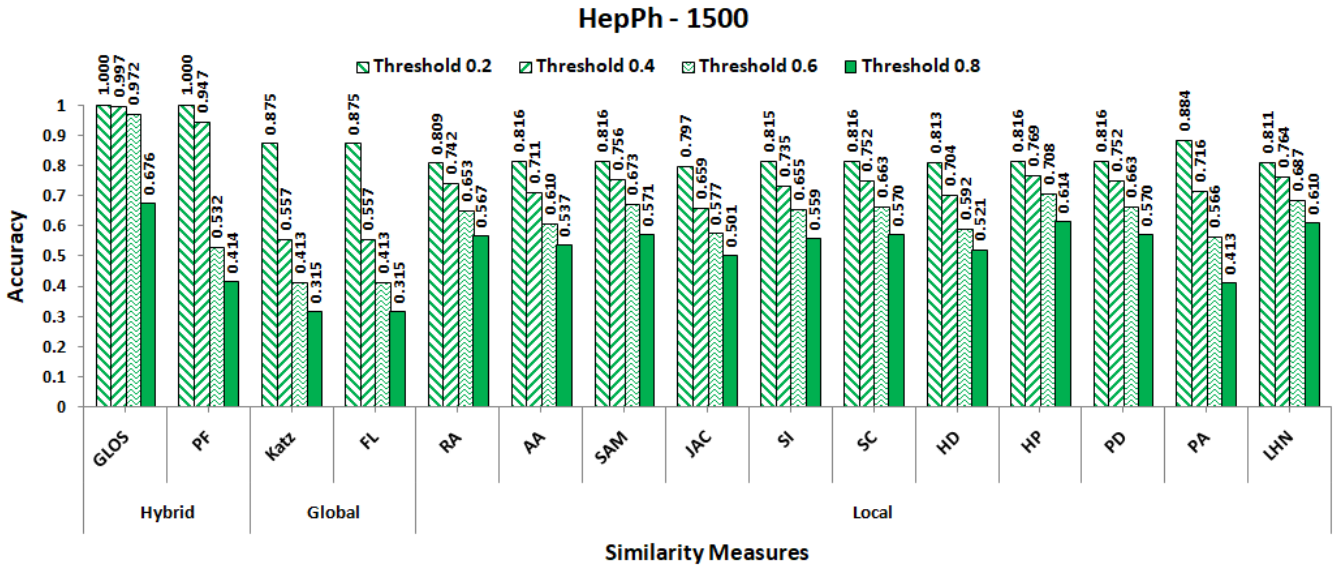

Figure 17. Comparisons of similarity measures on HepPh-1500 edge list from HepPh dataset

succeed in getting accuracy 1. On the other hand, at threshold 0.2, from the Local approaches AA achieved highest accuracy by 0.860. Similarly, at threshold 0.2, from the Global approaches, both Katz and FL obtained highest accuracy by 0.803. At threshold 0.4, the hybrid approach GLOS achieved highest accuracy by 0.994 and local approach PA achieved the lowest accuracy by 0.486. However, at threshold 0.4, rest of the local approaches performed better than global approaches, where, SAM obtained the highest accuracy by 0.769. At threshold 0.6, again hybrid approach GLOS obtained highest accuracy by 0.966 and global approaches Katz and FL could not perform well and obtained 0.373. Similarly, at the threshold 0.8, hybrid approach GLOS obtained the highest accuracy by 0.772 and local approach HP achieved the second highest results by 0.581. In case of all the thresholds, GLOS outperformed than rest of the similarity measures. As compared to the previous datasets, GLOS improved the accuracy.

Figure 20 showing the result of second edge list HepTh-1500 from HepTh dataset. Compared to the previous GrQc dataset, almost all the similarity measures improved their results here; however, local approach PA reduced its results. GLOS is the only measures who performed well at all thresholds, while, PA, Katz and FL obtained low accuracy than

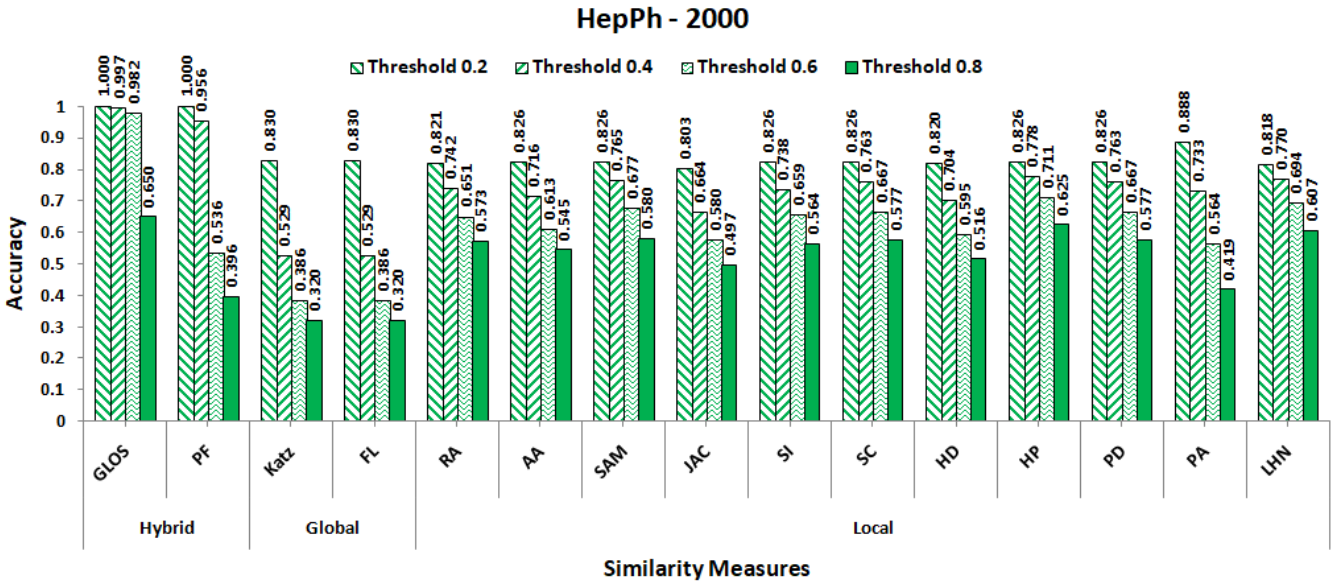

Figure 18. Comparisons of similarity measures on HepPh-2000 edge list from HepPh dataset

Table 6. The table presents the results of accuracy on HepPh Dataset. Where, (T) represents the threshold, High represents the similarity measures obtaining highest accuracy, 2nd High corresponds to the similarity measures obtained second highest accuracy and low represents the similarity measures obtained lowest accuracy. The red color shows the Global approaches, the blue color shows the Local approaches and the green color shows the Hybrid approaches.

Edge List	(T)	High	2nd High	Low
HepPh-1000	0.2	GLOS/PF	PA	JAC
	0.4	GLOS	PF	FL/KATZ
	0.6	GLOS	HP	FL/KATZ
	0.8	GLOS	HP	FL/KATZ
HepPh-1500	0.2	GLOS/PF	PA	JAC
	0.4	GLOS	PF	FL/KATZ
	0.6	GLOS	HP	FL/KATZ
	0.8	GLOS	HP	FL/KATZ
HepPh-2000	0.2	GLOS	PA	JAC
	0.4	GLOS	PF	FL/KATZ
	0.6	GLOS	HP	FL/KATZ
	0.8	GLOS	HP	FL/KATZ

other similarity measures. At threshold 0.2, hybrid approaches GLOS and PF obtained accuracy 1, both global approaches Katz and FL achieved 0.803 and local approach AA succeeded in getting the high accuracy by 0.860 than rest of the local approaches. Similarly, at threshold 0.2, hybrid approaches GLOS and PF achieved the highest accuracy and local approach PA obtained the lowest accuracy. Moreover, at threshold 0.4, hybrid approach GLOS achieved the highest result by 0.994 and local approach PA obtained the lowest accuracy by 0.486. On the other hand, at threshold 0.4, global approaches Katz and FL obtained the second lowest accuracy by 0.535. Likewise, at threshold 0.6, again highest accuracy achieved by GLOS and global approaches Katz and FL obtained the lowest accuracy. Overall, at all the thresholds, hybrid approach GLOS outperformed than others and PA, Katz and FL degraded its results by obtaining lowest accuracy. At threshold 0.8, hybrid approach GLOS performed well and obtained accuracy 0.772, while, local approach HP obtained the second highest accuracy by 0.581. On the other hand, at threshold 0.8, global approaches Katz and FL were the only similarity measures who obtained the lowest accuracy by 0.259.

Figure 21 showing the result of third edge list HepTh-2000 from HepTh dataset. At threshold 0.2, the hybrid

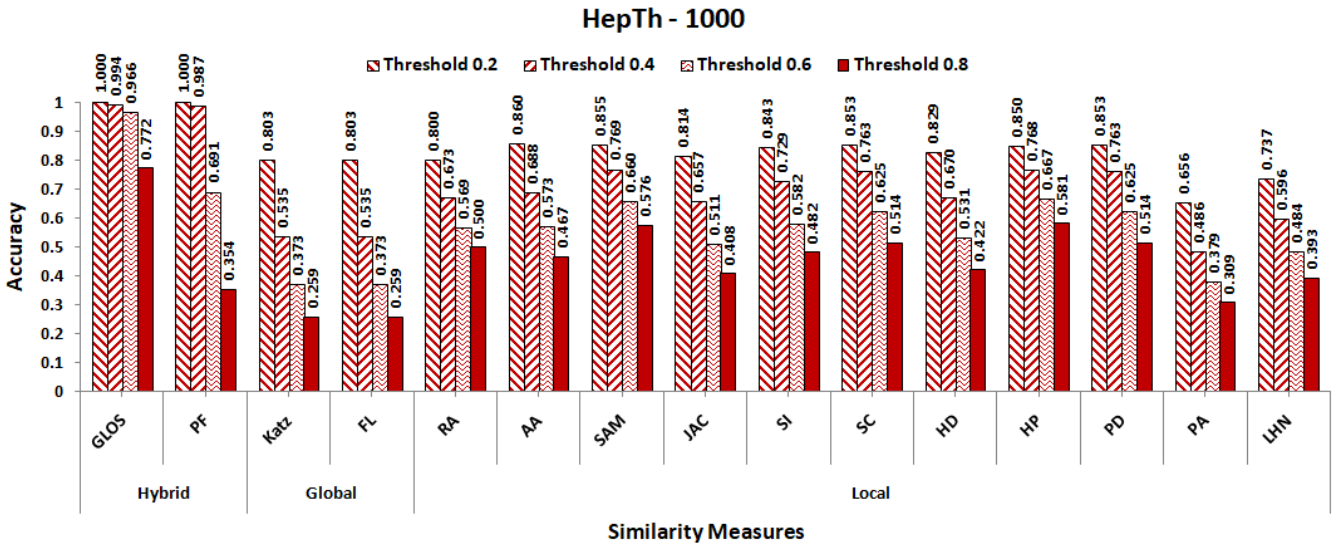

Figure 19. Comparisons of similarity measures on HepTh-1000 edge list from HepTh dataset

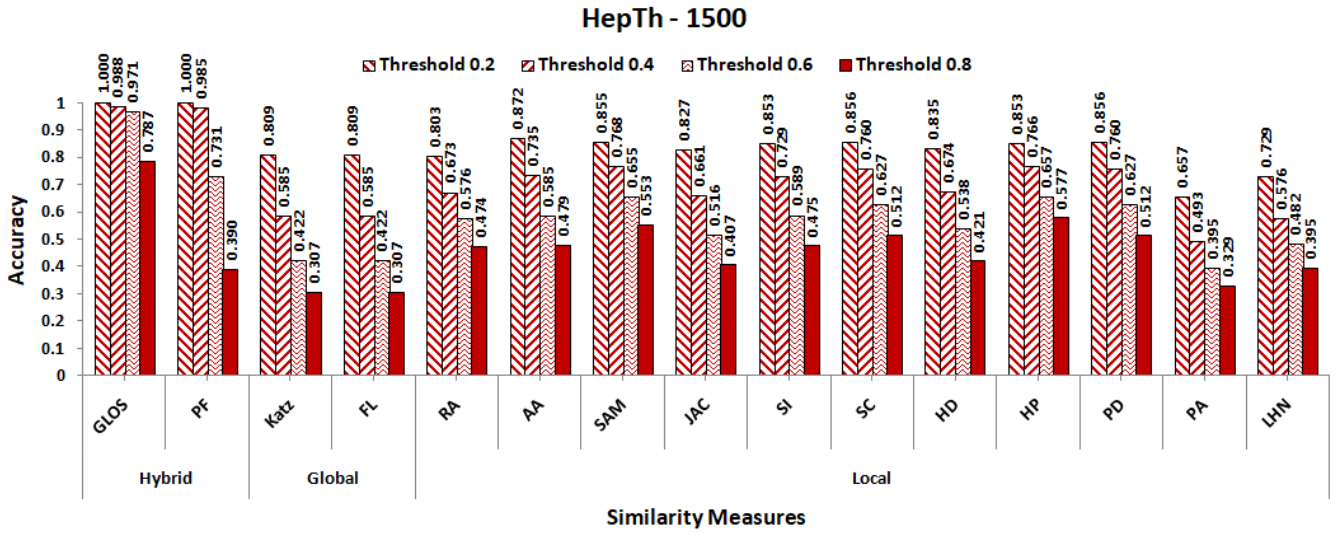

Figure 20. Comparisons of similarity measures on HepTh-1500 edge list from HepTh dataset

approaches GLOS and PF obtained the highest accuracy by 1 and local approach PA succeeded in getting the lowest accuracy by 0.648. Similarly, at threshold 0.4, the hybrid approach GLOS obtained the highest accuracy by 0.989, while, global local approach PA achieved the lowest results by 0.480. On the other hand, at threshold 0.6, local approach HP performed better than rest of the local and global approaches, where, HP obtained accuracy 0.674. Similarly, at threshold 0.8, hybrid approach GLOS obtained the highest score by 0.771, while, HP obtained the second highest accuracy by 0.586. Overall, the performance of GLOS was better than rest of the approaches. Similarly, at thresholds 0.2 and 0.4, local approach PA produced the lowest accuracy, while, at thresholds 0.6 and 0.8, global approaches Katz and Fl obtained the lowest accuracy. Detailed stats about accuracy are presented in Table 7.

5.3. Evaluation

In this thesis, experiments performed on 15 different edge lists from 5 dataset, where 3 edge lists belongs to each dataset. In this section, the performance of global, local and hybrid similarity measures is evaluated by giving the answer of the following questions.

- **Q1: Could global features hold the potential in accurate link prediction?**

  In order to evaluate the local similarity measures, both normalized and un-normalized local similarity measures (i.e., Resource Allocation,

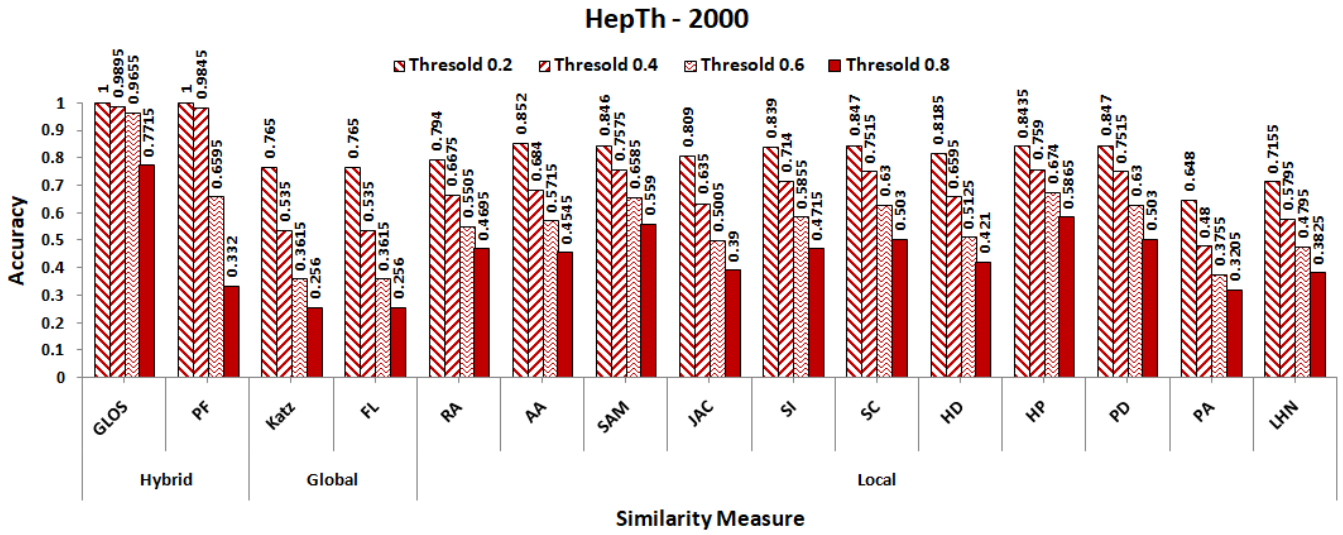

Figure 21. Comparisons of similarity measures on HepTh–2000 edge list from HepTh dataset

Table 7. The table presents the results of accuracy on HepTh Dataset. Where, (T) represents the threshold, High represents the similarity measures obtaining highest accuracy, 2nd High corresponds to the similarity measures obtained second highest accuracy and low represents the similarity measures obtained lowest accuracy. The red color shows the Global approaches, the blue color shows the Local approaches and the green color shows the Hybrid approaches.

Edge List	(T)	High	2nd High	Low
HepTh-1000	0.2	GLOS/PF	AA	PA
	0.4	GLOS	PF	PA
	0.6	GLOS	PF	FL/KATZ
	0.8	GLOS	HP	FL/KATZ
HepTh-1500	0.2	GLOS/PF	AA	PA
	0.4	GLOS	PF	PA
	0.6	GLOS	PF	PA
	0.8	GLOS	HP	FL/KATZ
HepTh-2000	0.2	GLOS/PF	AA	PA
	0.4	GLOS	PF	PA
	0.6	GLOS	HP	FL/KATZ
	0.8	GLOS	HP	FL/KATZ

Adamic Adar, SAM, Jaccard, Sorenson Index, Salton Cosine, Hub Depressed, Hub Promoted, Parameter-Dependent, Prefrential Attachment and Leicht-Holme-Nerman) are picked. After the comparisons it is found that local similarity measures hold the potential in accurate link prediction than global similarity measures. However, local similarity measures could not perform well than hybrid similarity measures. In case of Astro dataset, HP out performed than rest of the similarity local and global measures, while, PA could not produce better accuracy. Likewise, on all the dataset, PA obtained the lowest accuracy. Similarly, SAM, SC, HP and PD performed better, while, LHN obtained the second lowest accuracy. As compare to the global similarity measures, local similarity measures outperformed on all the thresholds. In case of comparison with the hybrid approaches, local approaches could not achieved good results. Furthermore, statistics about lowest and highest similarity measures are given in the following table.

- **Q2: Whether local features hold the potential in accurate link prediction?**

  In this study, we have used two global similarity measures (i.e., Katz and FL) for the comparisons with local and hybrid similarity measures. In the results of experiments, we have found that

global features did not hold the potential in order to find the accurate links. Both global similarity measures could not produce the better results as compare to local and global similarity measures. As shown in Figure 7 from Astro dataset, in case of all the thresholds, both global approaches obtained the lowest accuracy. Likewise, In case of remaining datasets, again the global similarity measures could not improve their results. In Figure 15 from GrQc dataset, both global approaches obtained the lowest similarity by 0.246. However, In case of CondMat dataset, global approaches outperformed than PA. Overall, the performance of global approaches was not good.

- **Q3: Combination of both local and global features could achieve the best accuracy?**

  Hybrid similarity measures are those types of measures which uses the local as well as global information. In the experiments, this study compared our hybrid approach GLOS with rest of the hybrid, global and local approaches. After the comparisons, it is shown that GLOS outperformed than rest of the approaches. In case Astro dataset, GLOS obtained accuracy results on all thresholds expect 0.8, where, GLOS obtained the second highest accuracy. Similarly, the same behavior of GLOS found in GrQc dataset, where, at threshold 0.8, GLOS obtained the second highest accuracy. On the other hand, in case of CondMat, HepPh and HepTh dataset, GLOS outperformed than rest of the approaches for all thresholds. Overall, hybrid approaches obtained better accuracy than local and global approaches. So, it is clear that combination of both local and global features can be more powerful predictor for link prediction. A detailed summary about highest, second highest, third highest and lowest accuracy for each dataset is presented in Table 1.

6. Conclusion

Social network (SN) is place where individuals or group of people connect to each other in order to share views, information and ideas. SN may be online via interaction of people on social sites (i.e., Facebook, LinkedIn, Twitter etc.) or offline via face-to-face interaction of people on public places (i.e., Colleges, Universities, Parks etc.). SN can be seen as a graph, where nodes correspond to the people and edges/ties represent relationship between them. During past two decades, attraction of people towards networking sites and applications have opened doors for new as well as expert researchers to study and analyzed SN properties and aspects of human behaviors in social world. Researchers, in the social network analysis, are facing various challenges. One of them and major challenge is link prediction.

Previous studies classified the link prediction techniques into three categorize: (1) probabilistic techniques, (2) maximum likelihood techniques and (3) similarity-based techniques. First two categorize, due to the algorithms complexity, could not deal with big networks. Similarity-based methods works on the similarity among users, as the similarity increases, chances of link formation also increase. In addition, similarity-based techniques consider two foremost features in estimation of the similarity: (1) local features (i.e., neighborhood) and (2) global features (i.e., path, random walk). The approaches that used the local features are called local similarity measures, while, the approaches that used the global features are called global similarity measures. Likewise, the approaches that use the global and local features are called hybrid approaches.

This study was an extension of our previous research where we have proposed a similarity measure namely SAM, based on local features, for link prediction in social network. In this study, we have addressed the challenging problem link prediction using local (such as neighbors) and global features (such as path). We have argued that combination of local and global features can have power to predict the link accurately. Moreover, we have proposed a hybrid approach GLOS, where, we have used path as global feature and neighborhood as local feature. Furthermore, we have compared the hybrid, global and local approaches on five dataset. In the results, we have found that hybrid approaches outperformed than rest of the local and global approaches. In addition, our proposed approach GLOS obtained highest accuracy on all the dataset. Moreover, global approaches produced the lowest accuracy on all the dataset and proved that there is a need to enhance the global approaches in order to predict the accurate results. In case of Astro dataset, GLOS obtained the highest accuracy by 1 at threshold 0.2, while, global approaches achieved the lowest accuracy by 0.358 at threshold 0.8. Similarly, in case of CondMat, HepPh and HepTh dataset, GLOS outperformed than rest of the similarity approaches by obtaining highest accuracy on all the thresholds. In case of threshold 0.6, GLOS obtained accuracy by 0.950 from Astro dataset, achieved accuracy by 0.973 from CondMat dataset, obtained accuracy 0.870 from GrQc dataset, achieved accuracy by 0.982 from HepPh dataset and succeeded in getting accuracy 0.971 from HepTh dataset. Similarly, PA from local approaches could not perform well on all dataset.

6.1. Future Work

In this study, similarity-based (i.e., local, global and hybrid) approaches for the link prediction are used. Moreover, we have compared the local, global and hybrid approaches in order to check their worth in accurate link prediction. In addition, we have proposed a hybrid approach GLOS using global features (i.e., path) and local features (i.e., neighborhood). Our future direction could be the use of other global features (i.e., random walk) with combination of local features for link prediction. In addition, nodes tend to connect with the nodes of their level or status. So, the structural importance of node for similarity computation could be another future direction.

7. Bibliography

References

[1] Abo Khedra, M.M., Abd EL-Aziz, A.A. and Hefny, H.A. (2019) Social network analysis through big data platform review. In *2019 International Conference on Computer and Information Sciences (ICCIS)*: 1–5. doi:10.1109/ICCISci.2019.8716484.

[2] Adafre, S. and de Rijke, M. (2005) Discovering missing links in wikipedia. In *Proceedings of the 3rd international workshop on Link discovery*: 90–97.

[3] Adamic, L.A. and Adar, E. (2003) Friends and neighbors on the web. *Social networks* **25**(3): 211–230.

[4] Aggarwal, C.C., Xie, Y. and Yu, P.S. (2014) A framework for dynamic link prediction in heterogeneous networks. *Statistical Analysis and Data Mining: The ASA Data Science Journal* **7**(1): 14–33.

[5] Akcora, C.G. (2010) *Using microblogs for crowdsourcing and public opinion mining* (State University of New York at Buffalo).

[6] Ayoub, J., Lotfi, D., El Marraki, M. and Hammouch, A. (2020) Accurate link prediction method based on path length between a pair of unlinked nodes and their degree. *Social Network Analysis and Mining* **10**(1): 9.

[7] Barabâsi, A.L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A. and Vicsek, T. (2002) Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications* **311**(3-4): 590–614.

[8] Buhai, I.S. and van der Leij, M.J. (2020) A social network analysis of occupational segregation. *arXiv preprint arXiv:2004.09293* .

[9] Cao, M., Zhang, H., Park, J., Daniels, N.M., Crovella, M.E., Cowen, L.J. and Hescott, B. (2013) Going the distance for protein function prediction: a new distance metric for protein interaction networks. *PloS one* **8**(10): e76339.

[10] Chen, B., Hua, Y., Yuan, Y. and Jin, Y. (2018) Link prediction on directed networks based on auc optimization. *IEEE Access* **6**: 28122–28136. doi:10.1109/ACCESS.2018.2838259.

[11] Du, J., Rong, J., Michalska, S., Wang, H. and Zhang, Y. (2019) Feature selection for helpfulness prediction of online product reviews: An empirical study. *PloS one* **14**(12): e0226902.

[12] Du, J., Rong, J., Wang, H. and Zhang, Y. (2020) Helpfulness prediction for online reviews with explicit content-rating interaction. In *International Conference on Web Information Systems Engineering* (Springer): 795–809.

[13] Du, J., Rong, J., Wang, H. and Zhang, Y. (2021) Neighbor-aware review helpfulness prediction. *Decision Support Systems* : 113581.

[14] Du, J., Rong, J., Wang, H. and Zhang, Y. (2021) Neighbor-aware review helpfulness prediction. *Decision Support Systems* : 113581.

[15] Duivesteijn, W., Pechenizkiy, M., Fletcher, G., Menkovski, V., Postma, E., Vanschoren, J. and van der Putten, P. [eds.] (2017) *Benelearn 2017: Proceedings of the Twenty-Sixth Benelux Conference on Machine Learning, TechnischeUniversiteit Eindhoven, 9-10 June 2017* (s.n.).

[16] Franz, T., Schultz, A., Sizov, S. and Staab, S. (2009) Triplerank: Ranking semantic web data by tensor decomposition. In *International semantic web conference* (Springer): 213–228.

[17] Guorong, X., Peiqi, C. and Minhui, W. (1996) Bhattacharyya distance feature selection. In *Proceedings of 13th International Conference on Pattern Recognition* (IEEE), **2**: 195–199.

[18] Huang, W., Zhaohui Wu, Mitra, P. and Giles, C.L. (2014) Refseer: A citation recommendation system. In *IEEE/ACM Joint Conference on Digital Libraries*: 371–374. doi:10.1109/JCDL.2014.6970192.

[19] Ibrahim, N.M.A. and Chen, L. (2015) Link prediction in dynamic social networks by integrating different types of information. *Applied Intelligence* **42**(4): 738–750.

[20] Jaccard, P. (1901) Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat* **37**: 547–579.

[21] Katz, L. (1953) A new status index derived from sociometric analysis. *Psychometrika* **18**(1): 39–43.

[22] Khalil, F., Wang, H. and Li, J. (2007) Integrating markov model with clustering for predicting web page accesses. In *Proceeding of the 13th Australasian world wide web conference (AusWeb07)* (AusWeb): 63–74.

[23] Leicht, E.A., Holme, P. and Newman, M.E. (2006) Vertex similarity in networks. *Physical Review E* **73**(2): 026120.

[24] Leskovec, J., Kleinberg, J. and Faloutsos, C. (2007) Graph evolution: Densification and shrinking diameters. *ACM transactions on Knowledge Discovery from Data (TKDD)* **1**(1): 2–es.

[25] Liben-Nowell, D. and Kleinberg, J. (2007) The link-prediction problem for social networks. *Journal of the American society for information science and technology* **58**(7): 1019–1031.

[26] Milgram, S. (1974) Obedience to authority. new york: Harper collins .

[27] Newman, M.E. (2001) Clustering and preferential attachment in growing networks. *Physical review E* **64**(2): 025102.

[28] Papadimitriou, A., Symeonidis, P. and Manolopoulos, Y. (2012) Fast and accurate link prediction in social networking systems. *Journal of Systems and Software* **85**(9): 2119–2132.

[29] Rangarajan, S., Liu, H. and Wang, H. (2020) Web service qos prediction using improved software source

code metrics. *Plos one* **15**(1): e0226867.

[30] Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. and Barabási, A.L. (2002) Hierarchical organization of modularity in metabolic networks. *science* **297**(5586): 1551–1555.

[31] Salton, G. and McGill, M.J. (1983) *Introduction to modern information retrieval* (mcgraw-hill).

[32] Samad, A., Azam, M. and Qadir, M. () Structural importance-based link prediction techniques in social network .

[33] Samad, A., Islam, M.A., Iqbal, M.A., Aleem, M. and Arshed, J.U. (2017) Evaluation of features for social contact prediction. In *2017 13th International Conference on Emerging Technologies (ICET)* (IEEE): 1–6.

[34] Samad, A., Qadir, M. and Nawaz, I. (2019) Sam: a similarity measure for link prediction in social network. In *2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)* (IEEE): 1–9.

[35] Samad, A., Qadir, M., Nawaz, I., Islam, M.A. and Aleem, M. (2020) A comprehensive survey of link prediction techniques for social network. *EAI Endorsed Trans. Indust. Netw. &Intellig. Syst.* **7**(23): e3.

[36] Samad, A., Qadir, M., Nawaz, I., Islam, M.A. and Aleem, M. (2020) Sam centrality: a hop-based centrality measure for ranking users in social network. *EAI Endorsed Trans. Indust. Netw. & Intellig. Syst.* **7**(23): e2.

[37] Scholz, C., Atzmueller, M., Barrat, A., Cattuto, C. and Stumme, G. (2013) New insights and methods for predicting face-to-face contacts. *Proceedings of the International AAAI Conference on Web and Social Media* **7**(1). URL https://ojs.aaai.org/index.php/ICWSM/article/view/14415.

[38] Sørensen, T., Sørensen, T., Sørensen, T., SORENSEN, T., Sorensen, T., Sorensen, T. and Biering-Sørensen, T. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons .

[39] Sutskever, I., Tenenbaum, J. and Salakhutdinov, R.R. (2009) Modelling relational data using bayesian clustered tensor factorization. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. and Culotta, A. [eds.] *Advances in Neural Information Processing Systems* (Curran Associates, Inc.), **22**: 1821–1828. URL https://proceedings.neurips.cc/paper/2009/file/5705e1164a8394aace6018e27d20d237-Paper.pdf.

[40] Tylenda, T., Angelova, R. and Bedathur, S. (2009) Towards time-aware link prediction in evolving social networks. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis*, SNA-KDD '09 (New York, NY, USA: Association for Computing Machinery). doi:10.1145/1731011.1731020, URL https://doi.org/10.1145/1731011.1731020.

[41] Wang, P., Xu, B., Wu, Y. and Zhou, X. (2015) Link prediction in social networks: the state-of-the-art. *Science China Information Sciences* **58**(1): 1–38.

[42] Wang, Z.J., Zhan, Z.H., Lin, Y., Yu, W.J., Wang, H., Kwong, S. and Zhang, J. (2019) Automatic niching differential evolution with contour prediction approach for multimodal optimization problems. *IEEE Transactions on Evolutionary Computation* **24**(1): 114–128.

[43] Zhou, T., Lü, L. and Zhang, Y.C. (2009) Predicting missing links via local information. *The European Physical Journal B* **71**(4): 623–630.

[44] Zhou, T., Lü, L. and Zhang, Y.C. (2009) Predicting missing links via local information. *The European Physical Journal B* **71**(4): 623–630.