Intelligent Hybrid Cluster Based Classification Algorithm for Social Network Analysis

S. Muthurajkumar[#1], P. Indira Priya[#2], M. Vijayalakshmi[#3], S. Indira Gandhi[*#4], A. Kannan[#5]

[#] Department of Information Science and Technology, College of Engineering Guindy, Anna University, Chennai, Tamil Nadu, India 600 025.

[1] muthurajkumarss@gmail.com

[2] indrapriyap@gmail.com

[3] vijim@annauniv.edu

[5] kannan@annauniv.edu

[*#] Department of Electronics and Engineering, Madras Institute of Technology, Anna University, Chennai, Tamil Nadu, India 600 025.

[4] indira@mitindia.edu

Abstract – In this paper, we propose an hybrid clustering based classification algorithm based on mean approach to effectively classify to mine the ordered sequences (paths) from weblog data in order to perform social network analysis. In the system proposed in this work for social pattern analysis, the sequences of human activities are typically analyzed by switching behaviors, which are likely to produce overlapping clusters. In this proposed system, a robust Modified Boosting algorithm is proposed to hybrid clustering based classification for clustering the data. This work is useful to provide connection between the aggregated features from the network data and traditional indices used in social network analysis. Experimental results show that the proposed algorithm improves the decision results from data clustering when combined with the proposed classification algorithm and hence it is proved that of provides better classification accuracy when tested with Weblog dataset. In addition, this algorithm improves the predictive performance especially for multiclass datasets which can increases the accuracy.

Keywords: Clustering, Enhanced K-Means Clustering Algorithm, Social network, Fuzzy logic, Classification, Human Behavior.

I. INTRODUCTION

Cluster analysis is a set of different methodologies for clustering of data's into a number of groups using distance measure. Clustering is very important for many examining and finding tasks including machine learning, pattern recognition, and data mining. A number of research work has been done on building clustering algorithms and numerous clustering algorithms are proposed. Every approach has its own merits and demerits. Clustering are used to perform investigative data analysis technique, it attempts to partition a given data set in to dissimilar groups such that data patterns within a group are more similar to one another than those belonging to different groups[5].

Clustering techniques are classified into supervised and unsupervised methods. The unsupervised clustering method is used to detect the underlying structure in the data set for classification [2]. Supervised clustering method involves the human interaction. Moreover, the unsupervised clustering techniques are most popular due to the minimal knowledge about the dataset. Similarity is fundamental aspect to clustering, and various clustering techniques use different similarity definitions and techniques. The very famous distance measure is the Euclidean distance. But, it has takes more time to cluster the data's compared to other distance metrics like Minkowski.

Classification is one of the most frequently used decision making techniques among all the human activities. A problem that occurs in classification is when a person or an object needs to be assigned into a predefined group or class based on a number of reasonable attributes related to that person or object. Traditional statistical classification techniques such as discriminate analysis are based on the Bayesian decision theory **[6]**. An underlying probability model is assumed in those systems to calculate the posterior probability upon which the classification decision is made. The effectiveness of these methods depends to a large extent on the various assumptions made or conditions under which the models are developed. Therefore, users must possess adequate knowledge of both data properties and model capabilities before the models can be successfully applied.

In social networks, link mining focus on link-based classification which is classifying samples graphical using the relations or links that are present among them. A social network is a structure comprising nodes and relations in which the actors are humans who use the internet. The traditional methods **[7]**, assume that the complete dataset can be deal with at once. However, if a huge amount of this data can be modeled as labels

associated with individuals, and are represented as nodes within a graph or graph-like structure they provide effective navigation patterns. In social networks, these labels come in many forms including especially people in particular labels such as age, gender and location. Moreover, labels may represent political or religious beliefs and may labels that encode interests, hobbies, and affiliations. It also deals with characteristics that capture the aspects of an individual's preferences or behavior. These labels can be created from the user's profile within the network, or attached to other objects in the network. In this paper, we propose a new classifier called neuro fuzzy Link Based Classification Algorithm (NFLBCA), to classify the relationships of different social network links based on fuzzy rules in order to analyze the social relationships using multi graphs representing such relationships. The main advantage of this proposed algorithm is that it helps to capture the relationships based on discussion carried out.

The major contributions of this paper are summarized as follows: In this paper, a new Minkowski distance based K-means clustering algorithm for clustering the data is proposed. In addition, the performance analysis is compared with our proposed algorithm. In this work, Weblog data set is used to perform the analysis of these two algorithms and the merits and demerits their distance metrics. The main advantage of this proposed algorithm is that the execution time compare to existing K-means algorithm is reduced to a considerable extent. Finally we propose a neuro fuzzy classifier which applies a set of fuzzy rules to effectively generate and analyze features that are commonly used in social studies. This research work addresses the classification issues by using the neuro fuzzy link based classification from a novel approach. Because some social features are as novel and useful, finding such the features has been incorporated in this work for improving the classification accuracy of the neuro fuzzy link based classification.

The reminder of this paper is structured as follows: Section 2 gives the literature survey. Section 3 discusses the system architecture. Section 4 explains the proposed work and implementation. Section 5 provides the results and discussion. Section 6 gives conclusion on this work and some future enhancements.

II. LITERATURE SURVEY

Tingting Cui and Fangshi Li [1] presented a Weight Computing in Competitive K-Means Algorithm which is derived from Improved K-means method and subspace clustering. By adding weights to the objective function, the contributions from each feature of each clustering could simultaneously minimize the separations within clusters and maximize the separation between clusters. Shi Na et al [2], proposes an improved k-means algorithm in order to solve this question, requiring a simple data structure to store some information in every iteration, which is to be used in the next interation. The improved method avoids computing the distance of each data object to the cluster centers repeatly, saving the running time.

Ran Vijay Singh and M.P.S Bhatia[3], proposed an modified K-Means algorithm based on the improvement of the sensitivity of initial center (seed point) of clusters. This algorithm partitions the whole space into different segments and calculates the frequency of data point in each segment. In this paper we also define a threshold distance for each cluster's centroid to compare the distance between data point and cluster's centroid with this threshold distance through which we can minimize the computational effort during calculation of distance between data point and cluster's centroid. Yufang Liu, Shibin Xiao, XueqiangLv and Shuicai Shi [4], proposed an semantic based K-means text clustering model,to solve the problem on high-dimensional and sparse characteristics of text data set. The model reduces the semantic loss of the text data and improves the quality of text clustering.

GeXiufeng, Xing Changzheng, [5],presented a new clustering method: KMCP algorithm. By the use of chromosome retreading and focus operators, the algorithm has higher accuracy and convergence speed. Prepare a comparative test program, and repeatedly running test program in the analysis of large amounts of data. General Statistics proves that KMCP algorithm presented in this paper is a feasible and efficient clustering algorithm. ManishaPujari and Rushed Kanawati, [8] proposed a Link Prediction model for Complex Networks by applying Supervised Rank Aggregation. Their approach is based on supervised rank aggregation and is motivated by the belief that each attribute can provide some unique information which can be aggregated in the end to make a better prediction of association between two unconnected entities in a network.

Dunlavy DM, et.al.[9] proposed a Temporal link prediction using matrix and tensor factorizations. It concentrated mostly on predicting how a social network may grow by adding new links. In other words, most of the previous works on link prediction have either limited their research on the prediction of the links that will be added to the network during the interval from time t to a given futuretime t_0 or implicitly have devoted the link prediction in specific domains such as co-authorship. Lu and Zhou,[10] proposed a new Link Prediction method for analysis Complex Networks: According to that, one of the major problems in studying dynamic evolution of complex networks, is the problem of link prediction. This link prediction problem aims to finding new associations (edges) in a network at a given point of time *t* when provided with the information about the network's temporal history before time *t*. Allali et al. ,[11] has introduced a new approach for link prediction called internal links and weighted projection to predict links in bipartite graphs. This approach performs well in

some social networks to predict links that will appear in the future however, their algorithm does not predict when links are lost.

Comparing with all these works, the system proposed in this paper is different in many ways. First, it uses a next distance measure. Second, it performed classification on clustered data. Third, it analyses the behaviors. Finally, it provides a facility for predictive analysis.

III. SYSTEM ARCHITECTURE

The architecture of the system proposed in this work consists of six major components namely Training data set, User Interface module, Clustering module, classification module, knowledge Base, and Decision manager module as shown in Fig. 1.

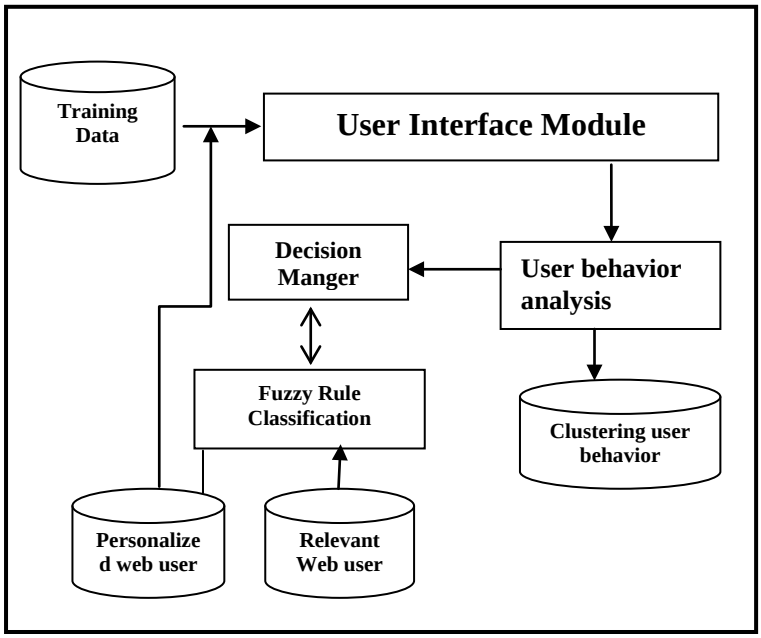

Fig. 1. System architecture

A. Training data and UI

Training data is a offline dataset collected from online database. The data's are retrieved from internet using standard program. Then this data's are sent into user behavior analysis for identifying the similarity between the different data Users can interact with the system through this user interface. User interface gives the request to the model and gets the result.

B. Clustering Module

The clustering module detects the dissimilar data from the given dataset using the five factor analysis. This clustering module distinguishes the dissimilar from similar data using the clustering algorithm called enhanced k means clustering algorithm.

C. Classification Module

The classification module classifies the dissimilar data from the given data using the proposed classification algorithm called NFLBCA. The classified results are used to form most relevant interest group.

D. Knowledge Base

This module find the relevant web user and to objective of the application. This knowledge is frequently presented in the form of linguistic.

IV. PROPOSED SYSTEM

In this paper to provide a new seven factor analysis finding friends with more similarities with respect to Frequency, Duration, Friends, Gender, Qualification, age and area were considered. The authors used some ranges for all factors for clustering the data.

A. Hybrid Clustering based Classification algorithm

Suppose a dataset X ={x1, x2 ,..., xn }include n data points and a feature set F ={f1 , f2 ,..., fd} comprise d features that describe the characteristics of each data point. A data point x= (xi1, xi2,...,xij ,..., xid)and *x* is the value of *j* th dimension of the *i* th point.

U = 1 if $l = arg\min d(xi, ct)$, $t = 1,...,k$

0 otherwise. (1)

If the distance measurements between one point and two cluster centers are equal, the point is assigned to the cluster with the smaller cluster index number. Minkowski distance, which is the most used dissimilarity measure, is evaluated by summing the differences between two data points in terms of all features. The distance between the point x and cluster center c is written as d(x,c) d(x,c) with the Minkowski distance is written as below:

$$d = (\Sigma \bmod ((M k - Nk) * r))^{\wedge} (1/ r)) \quad (2)$$

Where M & N are data objects r= Parameter

After an iteration of assignment is done, all the data points are assigned into k clusters. Then compute the geometric center of each cluster. The center of l th cluster C l is as follows:

$$C = \Sigma ux / \Sigma u \quad (3)$$

for and $1 \leq i \leq n$ and $1 \leq l \leq k$.

The objective of k-means algorithm is to minimize the sum of the dissimilarity between all data points and their corresponding cluster centers, which is shown as below:

$$\text{Min } E = \Sigma\Sigma\, uild(xi, cl) \quad (4)$$

In most cases, the k-means algorithm minimizes the following Mean Square Error (MSE) function based on Minkowski distance

$$\text{Min } E = \Sigma\Sigma \text{ umod } (x - c)^{\wedge} 2 \quad (5)$$

The hybrid clustering based classification algorithm is described below steps:

Step1: Randomly choose k number of points as the initial centers of k clusters.

Step2: Generate k number of new clusters by assigning each point to its closest cluster center by (1).

Step3: Calculate new cluster centers by (4).

Step 4: Keep repeating Step 2 and Step 3 until the cluster centers are stable or the Mean Square Error (MSE) function converges to a threshold value.

Step 5: Find out the set of users belonging to the same cluster as V_i, and all those users more likely will have the same relationship $\ell_i$with v.

Step6: This model is used to update the classifications of the objects. Create a sharedinformation table (SIT) to keep the latest labeling information for each link. Eachadjacent user is able to access and update the information of the corresponding link stored in SIT.

Step 7: Apply fuzzy rules from training patterns, perform classification, and generate linguistic rules. It refines the class boundaries by iteratively minimizing the error rate using the training data provide to the program.

Step 8: The algorithm terminates when it converges (there are no longer any updates to the categories) or a maximum number of steps has been reached. Otherwise repeat step 5 through7.

Step 9 : Take a test sample and test the classification performed using neuro fuzzy rules.

Step 10: Modified the links with suitable labels.

V. RESULT AND DISSCUSSION

In order to evaluate *HCBCA*, the algorithm is tested on a benchmark dataset, the network traffic data from the facebook dataset [16]. Facebook data set is usually used as a standard dataset to evaluate the performance of clustering algorithm. The proposed algorithm evaluates by using the recall and precision. Precision and recall defined as follows,

$$\text{Precision} = [TP / (TP+FN)] * 100 \quad (6)$$

$$\text{Recall} = [TP / (TP+FP)] * 100 \quad (7)$$

Table 1 shows the comparison of recall and precision values of the proposed algorithm with the existing algorithms. It is inferred that the proposed algorithm improves the performance of existing algorithms.

TABLE 1 Performance evaluation of the clustering algorithms

Dataset	EKMA		HCBC	
	Precision	Recall	Precision	Recall
BlogCatalog3	85.77	85.14	94.58	93.27
Weblog	84.32	85.63	94.26	94.01
YouTube7	81.46	82.51	87.35	88.29

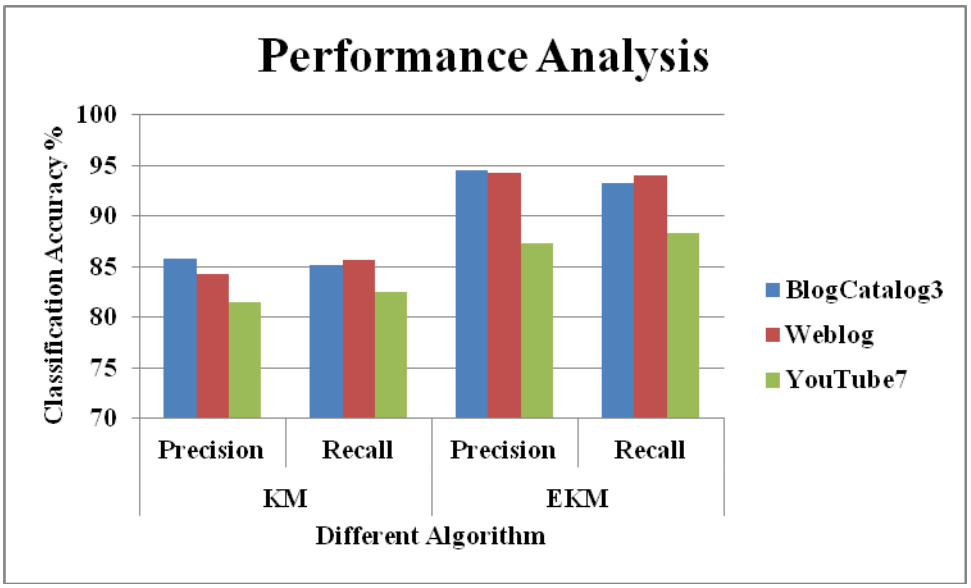

Fig. 2. Performance Analysis for precision and recall in different clustering technique

In Fig. 2 shows the performance analysis of the online purchase for different set of age groups.

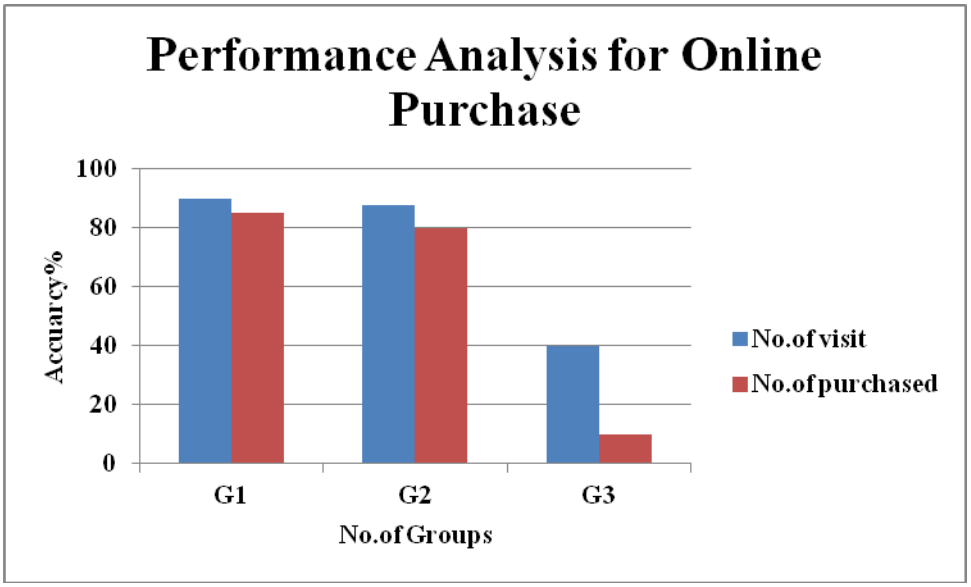

Fig. 3. Performance Analysis for Online purchase

From Fig. 3 it can be observed that the classification accuracy is more for No. of visits when it is compared with purchased due to the behaviour of the dataset.

Fig. 4 show the accuracy of number of visit analysis for Facebook Dataset.

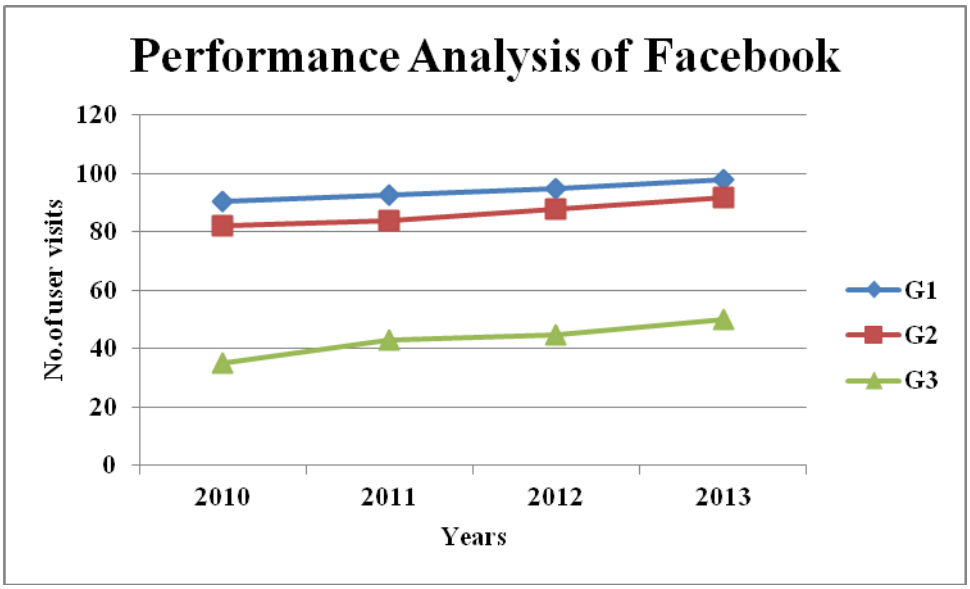

Fig. 4. Performance Analysis of facebook data set for different group of users

Proposed Result:

Selected attributes: 3,9,11,14,58,69,70,96,102,145,193,251 :12
aratio
url*media
url*blipverts
url*advertising+blipverts
url*468x60
url*wired.com
url*go2net
url*lycos.de
url*ng+spacedesc
url*img
url*www.uk.lycos.de
url*netscape.com
All the base classifiers:
C45 TREE
---
url*blipverts = 0
| url*www.uk.lycos.de = 0
| | aratio = 2.14: 0 (5.0)
| | aratio != 2.14
| | | url*468x60 = 0
| | | | aratio = 8.9523: 0 (2.0)
| | | | aratio != 8.9523: 1 (165.0/17.0)
| | | url*468x60 != 0: 0 (5.0)
| url*www.uk.lycos.de != 0: 0 (6.0)
url*blipverts != 0: 0 (16.0)
Number of Leaves : 6
Size of the tree : 11

From this it can be concluded that neuro tree detection paradigm performs better when specific user behavior are provided.The detailed accuracy for each class is shown below.

TP Rate	FP Rate	Precision	Recall	FMeasure	Class
0.987	0.042	0.964	0.987	0.975	normal
0.318	0.001	0.778	0.318	0.452	warezclient
0.999	0.004	0.991	0.999	0.995	neptune
0.989	0.007	0.829	0.989	0.902	ipsweep
0.8	0.001	0.857	0.8	0.828	teardrop
0.911	0.001	0.953	0.911	0.932	satan
0.855	0.001	0.93	0.855	0.891	portsweep
0.983	0.001	0.95	0.983	0.966	smurf
0.8	0.001	0.933	0.8	0.862	nmap
0	0	0	0	0	warezmaster
0.074	0	0.667	0.074	0.133	back
0	0	0	0	0	land
1	0	1	1	1	pod
0.833	0	0.625	0.833	0.714	buffer_overflow

VI. CONCLUSION

This system builds the patterns of the user behavior model over datasets labelled by the services. With the built patterns, the framework detects usage in the datasets using the classification algorithms. These results were useful in focusing research and highlighting the current capabilities and recent advances of existing algorithms. The main advantage of this algorithm is that it helps to reduce the execution time and provides better classification accuracy than the existing systems. Future work in this direction can be the use of intelligent agents for enhancing the execution time and classification accuracy since they can perform detective inference for effective decision making.

REFERENCES

[1] Tingting Cui, Fangshi Li, “Weight Computing in Competitive K-Means Algorithm”, IEEE, pp. 430-435, 2012.

[2] Shi Na, Liu Xumin, Guan yong, “Research on k-means Clustering Algorithm”, Third International Symposium on Intelligent Information Technology and Security Informatics, pp.63-67, 2010.

[3] Ran Vijay Singh, M.P.S Bhatia, “Data Clustering with Modified K-means Algorithm”, IEEE-International Conference on Recent Trends in Information Technology, pp.717-721, 2011.

[4] Yufang Liu, Shibin Xiao, XueqiangLv, Shuicai Shi, “Research on K-Means Text Clustering Algorithm Based on Semantic ”, International Conference on Computing, Control and Industrial Engineering, pp.124-127, 2010.

[5] GeXiufeng, Xing Changzheng, “K-means Multiple Clustering Research Based on Pseudo Parallel Genetic Algorithm”, International Forum on Information Technology and Applications, pp.30-33, 2010.

[6] L.Bolc, M.Makowski, and A.Wierzbicki :Label-Dependent Feature Extraction In Social Networks For Node Classification : SocInfo 2010 , LNCS 6430,pp.89-102,2010,Springer.

[7] J.N. Kok et al. (Eds): Generating Social Network Features forLink-based Classification, PKDD Springer2007,LNAI 4702,pp.127-139,2007.

[8] ManishaPujari , Rushed Kanawati, Link Prediction in Complex Networks by Supervised Rank Aggregation.2012 IEEE 24th International Conference on Tools with Artificial Intelligence (IEEE computer society)1082-3409/12 $26.00 © 2012 IEEE DOI 10.1109/ICTAI.2012.111.

[9] Dunlavy DM, Kolda TG, Acar E (2011) Temporal link prediction using matrix and tensor factorizations. ACM Trans KnowlDiscov Data (TKDD) 5(2), Art no 10.

[10] Lu T and Zhou T, “Link prediction in complex networks: A survey,” Physica A: Statistical Mechanics and its applications vol. 390, no. 6, pp. 1150–1170, Mar. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.physa.2010.11.02.

[11] J Allali O, Magnien C, Latapy M (2011) Link prediction in bipartite graphs using internal links and weighted projection. In: 2011 IEEE conference on computer communications workshops (INFOCOM WKSHPS), pp 936–941.