Structural Similarity Based Link Prediction in Social Networks Using Firefly Algorithm

Srilatha P*† and Manjula R*
*School of Computer Science and Engineering (SCOPE),
VIT University, Vellore, Tamil Nadu
Email: rmanjula@vit.ac.in
†Department of Computer Science and Engineering (CSE),
Anurag Group of Institutions, Hyderabad, Telangana
Email: srilatha@ieee.org

***Abstract*—Link prediction problem in social networks has received significant interest in the recent past from the researchers in diverse fields. Understanding and analyzing the present links in the social network or any complex networks either to understand their evolution or to predict the future possible links from the existing network (or links) forms the interesting link prediction problem. Link prediction based on Firefly optimization algorithm is proposed in this paper for social networks. The proposed algorithm is executed on a logical graph similar to social network and tested over real networks taking the benchmark data sets. Experimental values are compared with the other methods existing in the literature. From the comparison we can see that the proposed method performs better in terms of *precision* over the other methods.**

I. Introduction

Understanding and analyzing social networks has been of interest to researchers in the recent past. With million of users being part of one or the other social networking sites, provides rich information about the user attributes and the behaviour of users. Using the information rich networks for further study (either with identity of the users or only with anonymous attributes) and provide customized solutions to the users for purchasing (or recommending) products or visiting places or like the pages of similar interests or recommending nodes of similar kind have become research trends in social network analysis research [1], [2], [3], [4]. Apart from finding links, link prediction also helps in other domains such as finding spurious links in a network and biological networks with protein protein interactions etc. Such study of networks to predict the possibility of a node to connect with other nodes and products (or commodity) in general forms the link prediction problem.

To be precise, link prediction is calculating or predicting whether a given set of nodes will form a link or not in the future which is not present currently. In other words, if the social network is represented by a graph $G = (V, E)$ with V being the vertex set and E being the edge set of graph, then the link prediction problem is identifying the edge set E' that may happen over the graph G at a future instant of time which is not present currently [5], [6]. Edge set E' is the set of edges that may occur in the graph G in future. $E \cup E'$ constitute all the possible edges that may occur in graph G, constituting a complete graph with maximum number of possible edges U = $\frac{|V|\times(|V|-1)}{2}$. In such a setting, the set of existing edges E and the set of edges that may happen in future E' are non overlapping, i.e., $E \cap E' = \phi$. Link prediction is identifying such edge set E' by exploiting the properties of graph theoretic framework of social networks. Also, $E' + E = U$ and link prediction is finding the elements of set E', i.e., $E' = U - E$.

Various link prediction methods are proposed in literature considering multiple factors [7], [5], [8]. One such method to predict links is by using the structure of graph names as Similarity based Link Prediction. In similarity based link prediction methods, each node pair is assigned with a score called as similarity score or similarity index and the node pair with highest value will form a link in future. Similarity score will be calculated for all the node pairs in the network. If x and y are a node pair, then a score S_{xy} is assigned for the node pair. Node pair with higher value of S_{xy} is assumed to establish a link or connection in future. In graph theoretical terms, the nodes with higher values of S_{xy} are likely to form an edge in future.

Node pair connected by an edge or link is considered as connected by a path of length 1. Similarly, nodes connected with an intermediate node is considered as connected by a path of length 2. Structural link prediction methods are defined over path of length 2 and more. Various desired path lengths of a given social network or graph in general can be obtained by taking the powers of adjacency matrix A of graph G, i.e., A^2 gives the nodes which are connected by path 2 and A^3 gives the nodes that are connected by path 3 and so on [6], [5], [9]. Based on the path length considered the link prediction methods are classified as local similarity based method, global similarity based method and quasi local similarity based method. Local similarity based methods are defined over path length 2 whereas global similarity based methods are defined over path length strictly greater than 2. Quasi local link prediction method is defined based on combination of both local similarity and global similarity based link prediction methods.

Local similarity index based methods use neighborhood

978-1-5386-0569-1$31.00 ©2017 IEEE

information to compute similarity score. Some of the popular local similarity based methods are Common Neighbor [10], Salton Index [11], Jaccard index [12], Sørenson index [13], Hub promoted index [14], Hub depressed index [15], Leicht Holme Newman index-1 (LHN1) [16] , Preferential Attachment index [17], Adamic-Adar index [18] and Resource Allocation index [15]. Global similarity based methods are used in large networks to understand the nature of links that may happen in future. Some of the popular global similarity based methods are Katz Index [19], Average Commute Time [7], Random Walk with Restart [20], SimRank [21], Escape Probability [22], [23], [24] and Leicht Holme Newman index-2 [16].

Bio inspired algorithms are also employed to improve the accuracy of similarity based link prediction algorithms. Chen and Chen [6] used Ant colony optimization algorithm along with Common Neighbor similarity index to predict missing links in the social network. The fitness function was constructed with the node's centrality. The node pair was assigned with similarity score based on the amount of pheromone left at the end of iteration and all node pairs were ranked based on the remaining pheromone, the links with highest amount of pheromone will be likely to form the link. Also, they considered node attributes apart from structural similarity by constructing an augmented graph to improve the link prediction accuracy. However, they have not mentioned explicitly the number of ants employed in obtaining the reported results. Sherkat et al. [25] proposed Ant Colony Optimization Link Prediction (ACOLP) wherein the triangular structures were predicted using ant colony optimization and further with finding sub structures of the graphs that may form links in future. Bliss et al. [26] proposed Covariance Matrix Adaptation Evolution Strategy (CMA-ES) to predict links in Twitter social network. Score matrix (or similarity matrix) S_i was constructed by considering 16 different similarity measures and evolutionary algorithm approach was used to evolve the coefficients w_i.

In this paper, swarm firefly algorithm based link prediction algorithm is proposed to predict links in the social network. Fireflies are attracted to each other based on intensity of brightness and move towards other fireflies having higher brightness. The foraging behaviour of fireflies is employed to predict the links in social network. Similarity score is computed between the node pair at the end of iterations and the node pair with higher score are likely to form the links. Area under the receiver operating characteristic curve (AUC) and Precision are used as metrics to evaluate the performance of the proposed algorithms. Standard datasets are considered for testing the proposed algorithm with 10-fold cross-validation. Obtained results are compared with existing works from the literature.

The paper is organized as follows: Section II provides the preliminaries required for the proposed algorithm. In section III, firefly based structural link prediction model is defined with fitness function. In section IV, proposed model is evaluated using standard datasets available in literature. In section V, proposed model is compared with the existing methods and section VI deals with the conclusion.

II. Preliminaries

In this section preliminaries required for proposing the link prediction algorithm based on firefly optimization technique is discussed.

A. Firefly algorithm

Xin-She Yang introduced Firefly algorithm in 2008 [27]. Firefly algorithm is a meta heuristic algorithm inspired by foraging behaviour of fireflies. Each firefly will produce light by the process of bioluminescence either to attract other fireflies or to find mating partners. Based on produced light intensity, fireflies will be attracted to the fireflies with bright light intensity. The light intensity varies with distance and obeys inverse square law.

The light intensity of a firefly at a distance r is observed as $I(r) \propto \frac{1}{r^2}$, as the distance increases then the intensity or brightness decreases.

If the intensity of source is given as I_s, then the intensity $I(r)$ at a distance r is given as

$$I(r) = \frac{I_s}{r^2} \tag{1}$$

In the presence of medium absorbing light with absorption coefficient γ, the light intensity varies with r, i.e.,

$$I = I_0 \exp^{-\gamma r} \tag{2}$$

where, I_0 is the original light intensity.

By combining both inverse square law and absorption coefficient light intensity $I(r)$ is given as in Eq. 3 with Gaussian approximation

$$I(r) = I_0 \exp^{-\gamma r^2} \tag{3}$$

Similarly, for a firefly β attractiveness is given by

$$\beta = \beta_0 \exp^{-\gamma r^2} \approx \frac{\beta_0}{1+\gamma r^2} \tag{4}$$

where, β is intensity at $r = 0$. Exponential part in Eq. 4 is approximated for faster computation.

The distance r between the fireflies can be chosen based on the quantities of interest either as Cartesian distance or Euclidean distance.

The movement of a firefly i attracted to another more attractive firefly j is given by

$$x_i = x_i + \beta_0 \exp^{-\gamma r_{ij}^2}(x_j - x_i) + \alpha\epsilon_i \tag{5}$$

where, β_0 is the intensity at source, α is randomization parameter and ϵ_i is a random vector containing values from a distribution function.

The movement of firefly depend on second term in Eq. 5. If β = 0 then the movement of firefly will reduce to simple random walk.

III. Proposed Model

A. Basic idea of the algorithm

Link prediction algorithm in the social network can be modeled as a graph G with G = (V, E), where, V is the vertex set and E is the edge set. A logical complete graph G' is assumed with addition of missing links to the graph G. Fireflies are employed on the graph to find the nodes in the graph that are likely to form links in future. Each firefly will have its own brightness or light intensity value, based on that fireflies will mode towards the ones having higher intensity. It is known that nodes with existing links or common links are tend to form links. Similarly, even in firefly algorithm the fireflies tend to move towards nodes having higher degree. At every iteration, the fireflies will update their brightness value and score matrix s will be updated accordingly. Similar, procedure of moving of fireflies towards the brighter one will takes place and process continues. At the end of all iterations all nodes will be ranked based on score matrix s and the ones with higher values will tend to form link. Outline of the above briefed procedure is given in algorithm 1.

B. Structural link prediction with Firefly algorithm

Algorithm 1 Structural Link prediction with Firefly algorithm

```
Input: A: Adjacency matrix of network, Intensity $I$ of fireflies determined by $f(x)$
N: Number of iterations
Objective function: $f(x) : x = (x_1, x_2, \dots, x_d)^T$
Output: Score matrix $s$
while $t < N$ do
    for $i$ = 1 to $n$ do
        for $j$ = 1 to $n$ do
            if $(I_i < I_j)$ then
                Move firefly $i$ towards $j$;
            end if
            Update $r$ and $I$
        end for
    end for
    Rank the fireflies and store values in $s$
end while
```

C. Fitness function

The links in the network are tend to form in the neighborhood of nodes having higher degree [25]. Higher the degree, the greater the influence of such nodes in the network. Fitness of path is given in terms of edges as the clustering co-efficient. Clustering co-efficient is defined as the number of triangles associated with it. For an edge e_{ij} connecting nodes v_i and v_j the clustering coefficient is given as

$$C(e_{ij}) = \frac{z_{ij} + 1}{\min[(d_i - 1), (d_j - 1)]} \tag{6}$$

where, z_{ij} is the number of triangles formed with edge e_{ij}, d_i and d_j are the degrees of nodes v_i and v_j respectively. Larger clustering coefficient indicates higher chances of clustering.

Fitness of path in terms of clustering coefficient is given as

$$\sum_{i=1}^{n-1} C(e_{i,i+1}) \tag{7}$$

where, $i = \{0, 1, \dots, n-1\}$.

At each iteration, fireflies will move towards the brighter ones based on light intensity

$$x_i = x_i + \beta_0 \exp^{-\gamma r_{ij}^2}(x_j - x_i) + \alpha\epsilon_i \tag{8}$$

where, r_{ij} is given as $(\Gamma(i,j) \times \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2})$ with $\Gamma(i,j)$ as common neighbors of nodes at location i and j and the Euclidean distance between fireflies. α is the randomness parameter in the interval [0,1] and ϵ_i is the standard normal distribution with μ = 0 and σ = 1.

D. Parameter Selection

All the parameters suggested by Yang [27] are considered with following values: β_0 = 1, γ = 1 and α = 0.2 for conducting experiments.

IV. Evaluation of proposed model

Proposed model is evaluated using AUC and $precision$ parameters.

1) AUC [5] is calculated as

$$AUC = \frac{n' + 0.5n''}{n} \tag{9}$$

where, n' is the number by which missing link is having higher score than randomly chosen non existing link. n'' is the number of times both number of making link is equal to non existing link. Higher the values of AUC than 0.5 higher the accuracy of algorithm. n is the number of comparisons.

2) *Precision* is defined as the ratio of relevant items selected to the numbers of items selected [5], [6], i.e., among the links predicted (L) which are relevant right (R) links are selected

$$precision = \frac{R}{L} \tag{10}$$

Following bench mark datasets are considered for evaluating the performance of proposed method: USAir, Political-Blogs(PB) and Power. Data sets are taken from repository [28]. 10-fold cross-validation is performed by dividing each data set into 10-subsets and 1 subset is used as probe set and remaining 9 are used as testing set. Each subset is used once as probe set and remaining as test set to perform 10-fold cross-validation.

From the experiments, AUC and $precision$ values are obtained for different datasets as given in Table I. Number of fireflies employed in the experiments are also given in the Table I.

TABLE I
AUC AND *Precision* VALUES

No. of Fireflies	Dataset	AUC	Precision
100	USAir	0.6566	0.7566
	PB	0.5412	0.0612
	Power	0.5038	1
200	USAir	0.7142	0.7599
	PB	0.6080	0.0494
	Power	0.5076	1

TABLE II
AUC VALUE COMPARISON WITH OTHER METHODS

Method	USAir	Power	PB
LHN1	0.7194	0.5896	0.7541
Jaccard	0.8854	0.4926	0.8714
Proposed Method	0.7142	0.5076	0.6080

V. COMPARISON WITH THE EXISTING METHODS

Obtained values are compared with other methods of similarity link prediction namely LHN1 [16] and Jaccard Index [12]. AUC and precision values of LHN1 and Jaccard index are taken from [6] to compare with the proposed method.

AUC value is compared with other methods in Table II. We can infer from Table II that, over the dataset Power, the proposed method perform slightly better compared to Jaccard index whereas almost similar performance is exhibited by the proposed algorithm compared with LHN1 index.

TABLE III
PRECISION VALUE COMPARISON WITH OTHER METHODS

Method	USAir	Power	PB
LHN1	0.0122	0.003	0.0005
Jaccard	0.1037	0	0.0407
Proposed Method	0.7599	1	0.0494

AUC value is compared with other methods in Table III. We can infer from Table III that the proposed method performs significantly better than Jaccard and LHN1 index over USAir, Power and PB datasets. The proposed algorithm achieves higher precision values in comparison with other methods over all the datasets.

The proposed method performs better in terms of precision as it combines both brightness values of the fireflies as well as heuristics information to move towards the brighter fireflies.

VI. CONCLUSION

In this paper, firefly based link prediction algorithm is proposed. In the proposed approach fireflies were attracted towards the fireflies having bright intensity values and at each iteration the similarity matrix was updated. At the end of all iterations, the similarity matrix will output the nodes having higher similarity score. From the experimental results, it is evident that the proposed method outperforms the Jaccard and LHN1 similarity indices in terms of precision. As a future scope, it will be interesting to extend the proposed method to node attribute based link prediction.

REFERENCES

[1] S. Hill, F. Provost, and C. Volinsky, "Network-based marketing: Identifying likely adopters via consumer networks," *Statistical Science*, pp. 256–276, 2006.

[2] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[3] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, and J. Han, "Co-author relationship prediction in heterogeneous bibliographic networks," in *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pp. 121–128, IEEE, 2011.

[4] Z. Yin, M. Gupta, T. Weninger, and J. Han, "Linkrec: a unified framework for link recommendation with user attributes and graph structure," in *Proceedings of the 19th international conference on World wide web*, pp. 1211–1212, ACM, 2010.

[5] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 6, pp. 1150–1170, 2011.

[6] B. Chen and L. Chen, "A link prediction algorithm based on ant colony optimization," *Applied Intelligence*, vol. 41, no. 3, pp. 694–708, 2014.

[7] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American society for information science and technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[8] V. Martínez, F. Berzal, and J.-C. Cubero, "A survey of link prediction in complex networks," *ACM Computing Surveys (CSUR)*, vol. 49, no. 4, p. 69, 2016.

[9] P. Srilatha and R. Manjula, "Similarity index based link prediction algorithms in social networks: A survey," *Journal of Telecommunications and Information Technology*, no. 2, p. 87, 2016.

[10] F. Lorrain and H. C. White, "Structural equivalence of individuals in social networks," *The Journal of mathematical sociology*, vol. 1, no. 1, pp. 49–80, 1971.

[11] G. Salton, "Introduction to modern information retrieval," *McGraw-Hill*, 1983.

[12] P. Jaccard, *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz, 1901.

[13] T. Sørensen, *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*. Biologiske Skrifter // Det Kongelige Danske Videnskabernes Selskab, I kommission hos E. Munksgaard, 1948.

[14] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," *science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[15] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," *The European Physical Journal B*, vol. 71, no. 4, pp. 623–630, 2009.

[16] E. Leicht, P. Holme, and M. E. Newman, "Vertex similarity in networks," *Physical Review E*, vol. 73, no. 2, p. 026120, 2006.

[17] Z. Huang, X. Li, and H. Chen, "Link prediction approach to collaborative filtering," in *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pp. 141–142, ACM, 2005.

[18] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social networks*, vol. 25, no. 3, pp. 211–230, 2003.

[19] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

[20] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu, "Automatic multimedia cross-modal correlation discovery," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 653–658, ACM, 2004.

[21] G. Jeh and J. Widom, "Simrank: a measure of structural-context similarity," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 538–543, ACM, 2002.

[22] P. G. Doyle and J. L. Snell, "Random walks and electric networks," *AMC*, vol. 10, p. 12, 1984.

[23] H. H. Song, T. W. Cho, V. Dave, Y. Zhang, and L. Qiu, "Scalable proximity estimation and link prediction in online social networks," in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pp. 322–335, ACM, 2009.

[24] H. Tong, C. Faloutsos, and Y. Koren, “Fast direction-aware proximity for graph mining,” in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 747–756, ACM, 2007.

[25] E. Sherkat, M. Rahgozar, and M. Asadpour, “Structural link prediction based on ant colony approach in social networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 419, pp. 80–94, 2015.

[26] C. A. Bliss, M. R. Frank, C. M. Danforth, and P. S. Dodds, “An evolutionary algorithm approach to link prediction in dynamic social networks,” *Journal of Computational Science*, vol. 5, no. 5, pp. 750–764, 2014.

[27] X.-S. Yang, *Nature-inspired metaheuristic algorithms*. Luniver press, 2010.

[28] “Datasets.” http://www.linkprediction.org/index.php/link/resource/data/.