Similarity Measures for Link Prediction Using Power Law Degree Distribution

Srinivas Virinchi* and Pabitra Mitra

Dept of Computer Science and Engineering, Indian Institute of Technology Kharagpur-721302, India
{virinchimnm,pabitra}@cse.iitkgp.ernet.in

Abstract. Link Prediction deals with predicting important non-existent edges in a social network that are likely to occur in the near future. Typically a link between two nodes is predicted if the two nodes have high similarity. Well-known local similarity functions make use of information from all the local neighbors. These functions use the same form irrespective of the degree of the common neighbors. Based on the power law degree distributions that social networks generally follow, we propose non-linear schemes based on monotonically non-increasing functions that give more emphasis to low-degree common nodes than high-degree common nodes. We conducted experiments on several benchmark datasets and observed that the proposed schemes outperform the popular similarity function based methods in terms of accuracy.

Keywords: Degree of neighbors, Common Neighbors, Adamic Adar, Resource Allocation Index, Markov Inequality.

1 Introduction

Link prediction problem employs similarity that could be based on either local or global neighborhood. Here, we concentrate on some of the state-of-the art local neighborhood measures. One of the local measures which is popularly employed is the Resouore Allocation Index. The similarity function employed by this index makes the contribution of each of the common neighbors inversely proportional to its degree.

Social networks follow the power law degree distribution. According to this law, the probability of encountering a high degree node is very small. We use an appropriate threshold to split the set of nodes based on their degree into low and high degree node sets. We have given different weights to common nodes in different clusters while computing the similarity between a pair of nodes. Specifically, we have compared the performance of the modified algorithms with common neighbors, Adamic Adar and resource allocation index. We emphasize the role of low degree common nodes in terms of their contribution to the similarity and either deemphasize or ignore the contributions of high degree common

* Corresponding author: virinchimnm@gmail.com

M. Lee et al. (Eds.): ICONIP 2013, Part II, LNCS 8227, pp. 257–264, 2013.

nodes. We justify the proposed scheme formally. The modified algorithms have resulted in an improved performance in terms of classification accuracy on several benchmark datasets. Our specific contributions in this paper are:

1. We establish formally that the number of high degree common neighbors is insignificant compared to the low degree common neighbors.
2. To deemphasize the role of high degree neighbors and increase the contribution of low degree neighbors in computing similarity.

2 Background and Related Work

We can view any social network as a graph G. Each link can be viewed as an edge in the graph. We can represent the graph as G = (V, E) where V is the set of vertices and E is the set of edges in the graph. Now, let us consider that at some future instance t' the graph after addition of some edges has become $G' = (V', E')$ where V' is the set of vertices of graph G' and E' is the set of edges of G'. Link Prediction problem deals with the prediction of edges from the set of edges $E' - E$ accurately.

According to the power law, the probability of finding a k degree node in the social network which is denoted by p_k is directly proportional to $k^{-\alpha}$ where α is some positive constant usually between 2 and 3. Hence, the probability of finding a high degree node in the graph is very less as the corresponding value of k is very high.In general, given G, it is not clear as to which are high degree nodes and which are low degree nodes. For differentiating between high degree and low degree nodes we make use of the Markov Inequality. It can be defined as: if X is a non-negative random variable and there exists some positive constant $b > 0$ then,

$$P(X \geq b) \leq \frac{E[X]}{b}. \tag{1}$$

We make use of the above inequality to find a threshold value (T) which divides the set of nodes based on degrees into low degree nodes and high degree nodes. In this case, as degree is always non-negative we can make degree as the non-negative random variable and T will be positive, we can represent the above inequality as:

$$P(degree \geq T) \leq \frac{E[degree]}{T} \Rightarrow T \leq \frac{E[degree]}{P(degree \geq T)}. \tag{2}$$

Thus, from inequality (2) we can calculate the threshold that we require based on the number of high degree nodes that we need. The similarity functions could be either local or global (takes the whole graph into account along with some distance based measures). [1] and [2] present the survey of various similarity measures and higher level approaches used in the area of link prediction. Further, they show that Resource Allocation Index outperforms the other local similarity measures. [3] refines the common neighbors approach by giving the less connected neighbors higher weight and is popularly known as the Adamic-Adar index (AA)

known after the authors. The authors in [4] designed the Resource Allocation (RA) Index which is motivated by the resource allocation dynamics where node x can transmit a resource through a common neighbor to another node y. In [5], the authors make use of the latent information from communities and showed that embedding community information into the state-of-the-art similarity functions can improve the classification accuracy.

3 Standard Similarity Measures and Our Approach

We formally explain different similarity functions which are used for experimental comparison against our approach. Here, $N(x)$ is the set of nodes adjacent to node x.

Common Neighbors (CN): Score between nodes x and y is calculated as the number of common neighbors between x and y.

$$CN(x, y) =\mid N(x) \cap N(y) \mid . \tag{3}$$

Adamic Adar (AA): Score between nodes x and y is calculated as the sum of inverse of the log of degree of each of the common neighbors z between x and y.

$$AA(x, y) = \sum_{z \in N(x) \cap N(y)} \frac{1}{log(degree(z))}. \tag{4}$$

Resource Allocation Index (RA): Score between nodes x and y is calculated as the sum of inverse of the degree of each of the common neighbors z between x and y.

$$RA(x, y) = \sum_{z \in N(x) \cap N(y)} \frac{1}{degree(z)}. \tag{5}$$

We use (2) discussed in section 2 to bound the threshold. For conducting the experiment, we use $P(degree \geq T)$ between 0.01 and 0.1 and on experimenting we find the best threshold for each dataset. The thresholds are listed in table 2. Once we have the threshold value, we divide the node set V based on the threshold into low degree and high degree node sets. We justify our deemphasis of the high degree common neighbors using the theorem below. Let,

- n = Number of nodes in the graph, n_L = Number of Low degree nodes, n_H = Number of High Degree nodes, p_k = Probability existence of a k degree node in graph
- L_{avg} = Average degree of a node in the low degree region , H_{avg} = Average degree of a node in the high degree node
- K_L = Expected number of low degree common neighbors for any pair of nodes, K_H = Expected number of low degree common neighbors for any pair of nodes
- T = Threshold on degree, max = Maximum degree, $L = \{x|degree(x) < T\}$, $H = \{x|degree(x) \geq T\}$

Theorem 1. *For any pair of nodes x and y, the expected number of common neighbors of high degree is very small when compared to expected number of common neighbors of low degree.*

Proof: Consider the possibility that both $x, y \in L$. The probability that a common neighbour $z \in L$ is given by

$$P(z \in L) = \frac{n_L}{n} * p_{L_{avg}} * \frac{n_L}{n} * p_{L_{avg}} * \frac{n_L}{n} * p_{L_{avg}} * \frac{n_L}{n} * p_{L_{avg}}. \tag{6}$$

where $\frac{n_L}{n}$ accounts for the selection of a node from L and $p_{L_{avg}}$ accounts for the average probability of existence of a low degree node. Note that the first four terms correspond to the probability of existence of an edge between x and z and the last four terms correspond to the probability of existence of an edge between y and z. The above equation can be simplified to the following form :

$$P[z \in L] = \frac{n_L}{n} * p_{L_{avg}} * \frac{n_L}{n} * p_{L_{avg}} * (\frac{n_L}{n} * p_{L_{avg}})^2. \tag{7}$$

In a similar way the probability that a common neighbor $z \in H$ is given by

$$P[z \in H] = \frac{n_L}{n} * p_{L_{avg}} * \frac{n_L}{n} * p_{L_{avg}} * (\frac{n_H}{n} * p_{H_{avg}})^2. \tag{8}$$

So,

$$K_L = n * P[z \in L] = n * \frac{n_L}{n} * p_{L_{avg}} * \frac{n_L}{n} * p_{L_{avg}} * (\frac{n_L}{n} * p_{L_{avg}})^2. \tag{9}$$

and

$$K_H = n * P[z \in H] = n * \frac{n_L}{n} * p_{L_{avg}} * \frac{n_L}{n} * p_{L_{avg}} * (\frac{n_H}{n} * p_{H_{avg}})^2. \tag{10}$$

Thus, the ratio of K_H to K_L, from (4) and (5) is given by

$$\frac{K_H}{K_L} = (\frac{n_H}{n_L})^2 * (\frac{p_{H_{avg}}}{p_{L_{avg}}})^2. \tag{11}$$

Note that there are 3 other possbilities for assigning x and y to L and H; these are: 1. $x \in L$ and $y \in H$, 2. $x \in H$ and $y \in L$ and 3. $x \in H$ and $y \in H$ Note that in all these 3 cases also, the value of $\frac{K_H}{K_L}$ is the same as the one given in (6). Using power law

$$p_{L_{avg}} = C * (L_{avg})^{-\alpha}. \tag{12}$$

From (6) and (7), we have

$$\frac{K_H}{K_L} = (\frac{n_H}{n_L})^2 * (\frac{L_{avg}}{H_{avg}})^{2\alpha}. \tag{13}$$

Consider $\frac{n_H}{n_L}$ which can be simplified as, (degree 1 nodes are ignored)

$$\frac{n_H}{n_L} = \frac{n * \sum_{i=T}^{max} p_i}{n * \sum_{i=2}^{T-1} p_i}. \tag{14}$$

Note that

$$2^{-\alpha} \leq \sum_{i=2}^{T} i^{-\alpha} \leq (T-1) * 2^{-\alpha}. \tag{15}$$

Using power law we can substitute $i^{-\alpha}$ for p_i and cancelling out n, we get

$$\frac{n_H}{n_L} = \frac{\sum_{i=T}^{max} i^{-\alpha}}{\sum_{i=2}^{T-1} i^{-\alpha}} \leq \frac{(max-T) * T^{-\alpha}}{2^{-\alpha}} = (max-T) * (\frac{2}{T})^{\alpha}. \tag{16}$$

Also by noting that $L_{avg} < T$ and $H_{avg} > T$ we can bound $\frac{L_{avg}}{H_{avg}}$ as follows,

$$\frac{L_{avg}}{H_{avg}} = \frac{T-\delta}{T+\delta} \; for \; some \; 1 < \delta < max \tag{17}$$

$$\frac{L_{avg}}{H_{avg}} = (1 - \frac{2\delta}{T+\delta}) < e^{-}\frac{2\delta}{T+\delta} \tag{18}$$

So from (13), (16) and (18), we get

$$\frac{K_H}{K_L} \leq (max-T)^2 * (\frac{2}{T})^2 * (e^{-}\frac{2\delta}{T+\delta})^{2\alpha} \tag{19}$$

which is a very small quantity and it tends to zero as T tends to a large value which happens when the graph is large. It is intuitively clear that $L_{avg} < H_{avg}$. Further, because of the power law and selection of an appropriate threshold value we can make $\frac{n_H}{n_L}$ as small as possible. For example, by selecting the value of T to be less than or equal to $\frac{max}{2}$ where $max = max_x(degree(x))$ we get $\frac{n_H}{n_l}$ to range between 0.003 to 0.04 for the datasets considered and for the same threshold the value of $\frac{L_{avg}}{H_{avg}}$ ranges from 0.07 to 0.14. So, the value of $\frac{K_H}{K_L}$ ranges from $0.000009 * (0.0049)^{\alpha}$ to $0.019 * (0.0016)^{\alpha}$. Further, the value of α lies between 2 and 3. So, $\frac{K_H}{K_L}$ can be very small. We show the corresponding values in Table 2

Now we present our modifications to the above metrics, let us call them $CN1$, $AA1$ and $RA1$ where these stand for the modified similarity functions for CN, AA and RA respectively. Let $R = N(x) \cap N(y)$.

$$CN1(x,y) \; = \mid S(z) \mid \; where \; S(z) = \{z | z \in R \wedge degree(z) < T\} \tag{20}$$

$$AA1(x,y) \; = \sum_{z \in R \wedge degree(z) < T} \frac{1}{log(degree(z))} \tag{21}$$

$$RA1(x,y) = \sum_{z \in R \wedge degree(z) < T} \frac{1}{\sqrt{degree(z)}} + \sum_{z \in R \wedge degree(z) \geq T} \frac{1}{degree(z)^2} \tag{22}$$

Thus, in general we can write the score function as a combination of two monotonically non-increasing functions *low* and *high* , where low is applied on common neighbors having degree less than threshold and high is applied on common neighbors having degree greater than the threshold.

$$score(x,y) = \sum_{z \in R} (low(z) + high(z)) \tag{23}$$

Table 1. $low(z)$ and $high(z)$ for various schemes

Metric	low(z)	high(z)
CN1	1	0
AA1	$\frac{1}{log(degree(z))}$	0
RA1	$\frac{1}{\sqrt{degree(z)}}$	$\frac{1}{degree(z)^2}$

where z is the common neighbor of x and y. In our approach we use values for functions low and high as shown in table 1. It is clear from our approach that we have given less importance or zero weight to high degree nodes. Let z_1 and z_2 be two common neighbors of nodes x and y with degrees p and q respectively such that $z_1 \in L$ and $z_2 \in H$ then their contributions are:

- contribution of z_1 to RA is $\frac{1}{p}$ and for RA1 it is $\frac{1}{\sqrt{p}}$; further, $\frac{1}{\sqrt{p}} > \frac{1}{p}$
- contribution of z_2 to RA is $\frac{1}{q}$ and for RA1 it is $\frac{1}{q^2}$; further, $\frac{1}{q^2} < \frac{1}{q}$
- contribution of z_1 to CN is same as its contribution in CN1 and contribution of z_2 does not contribute to CN1 as it contributes in CN. Similar interpretation holds for AA and AA1.

So, common neighbors in L contribute more to RA1 than RA and those in H contribute less to RA1 than to RA.

Table 2. Details of various datasets

Dataset	$\mid V \mid$	$\mid E \mid$	T	L_{avg}	H_{avg}	n_L	n_H	$\frac{L_{avg}}{H_{avg}}$	$\frac{n_H}{n_L}$
GrQc	5241	28968	65	9	99	5168	73	0.09	0.014
HepTh	9875	51946	53	9	80	9792	83	0.1125	0.008
CondMat	23133	186878	45	13	92	22221	912	0.14	0.04
AstroPh	18771	396100	322	40	551	18714	57	0.07	0.003

4 Datasets and Experimental Methodology

For conducting our experiments we used collaboration graphs [6] which are undirected. In the case of collaboration graphs, the nodes represent the authors and the edge between two nodes x and y represents a collaboration between authors x and y. We conduct the experiments on the four datasets listed in table 2. We remove the self loops and nodes having degree one. The details of the datasets after preprocessing are given in Table 2. The datasets for papers from January 1993 to April 2003 are : AstroPh (Astro Physics), CondMat (Condense Matter Physics), GrQc (General Relativity and Quantum Cosmology), and HepTh (High Energy Physics - Theory).

We perform the link prediction on a static snapshot of the graph. We follow the approach to set up the data similar to the one explained in [2] and [5]. We take each of the graph datasets and randomly partition the graph into five parts by removing the edges randomly where each part has 20% of the edges. Now, we use one part as test data and the remaining 4 parts for training. That is we use a 5 fold cross validation. We repeat this five times each time taking a different part to be the test data; we report the average values.

5 Results and Discussions

On performing the experiments using the modified metrics described in section 3 we report the results in table 3 which indicate the accuracy on various datasets. From table 3 we observe that RA outperforms CN and AA. This result is consistent with the results shown in [2]. We can observe that our modified approach for each of the similarity functions CN, AA and RA performs better than the corresponding base metric as shown in boldface. We can observe that on the above datasets, CN1 performs almost as well as AA and AA1 performs as well as RA. Further, RA1 is the best. Also note that in some cases the accuracy has increased up to 12%.

Table 3. Percentage Classification Accuracy

Dataset	CN	CN1	AA	AA1	RA	RA1
GrQc	98.4	99.3	98.7	99.5	98.4	99.5
HepTh	78.1	85.4	90.4	93.8	92.2	94.5
CondMat	81.61	93.42	92.41	96.76	96.76	97.13
AstroPh	98.81	99.02	99.36	99.45	99.45	99.55

6 Conclusion

Based on experimentation, we observed that our approach performs better than the corresponding base similarity functions. RA1 performs the best among all the metrics. Thus, from the results we can conclude that our approximation of the state-of-the-art local neighborhood similarity functions performs better than the original similarity functions in terms of classification accuracy; further, it can decrease time when the contribution of high degree neighbors is ignored. We can conclude that high degree neighbors are not so useful in predicting new links. Thus, we can completely ignore or minimize the contributions of high degree nodes by making use of a suitable non-linear similarity function to weigh their contributions accordingly.

References

1. Nowell, L., Kleinberg, J.: The Link Prediction Problem for Social Networks. In: CIKM, pp. 556–559. ACM Press, New York (2003)
2. Zhou, T., Lu, L.: Link Prediction in Complex networks. J. Phys. A 390, 1150–1170 (2011)
3. Adamic, L., Adar, E.: Friends and neighbours on the Web. J. Social Networks 25, 211–230 (2003)
4. Zhou, T., Lu, L., Zhang, Y.-C.: Predicting Missing links via local information. J. Eur. Phys. B 71, 623–630 (2009)
5. Soundarajan, S., Hopcroft, J.: Using Community information to Improve the Precision of Link Prediction Methods. In: WWW, pp. 607–608. ACM Press, New York (2012)
6. Leskovec, J., Kleinberg, J., Faloutsos: Graph Evolution: Densification and shrinking Diameters. Technical report, arXiv.org (2007)