Link Prediction in Complex Networks Based on Cluster Information

Article *in* Lecture Notes in Computer Science · October 2012

DOI: 10.1007/978-3-642-34459-6_10

2 authors:

Jorge Valverde-Rebaza
Visibilia
34 PUBLICATIONS **381** CITATIONS
SEE PROFILE

Alneu de Andrade Lopes
University of São Paulo
99 PUBLICATIONS **1,177** CITATIONS
SEE PROFILE

Link Prediction in Complex Networks Based on Cluster Information

Jorge Carlos Valverde-Rebaza and Alneu de Andrade Lopes

Departamento de Ciências de Computação
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo - Campus de São Carlos
Caixa Postal 668
13560-970 São Carlos, SP, Brazil
{jvalverr,alneu}@icmc.usp.br

Abstract. Cluster in graphs is densely connected group of vertices sparsely connected to other groups. Hence, for prediction of a future link between a pair of vertices, these vertices common neighbors may play different roles depending on if they belong or not to the same cluster. Based on that, we propose a new measure (WIC) for link prediction between a pair of vertices considering the sets of their intra-cluster or within-cluster (W) and between-cluster or inter-cluster (IC) common neighbors. Also, we propose a set of measures, referred to as W forms, using only the set given by the within-cluster common neighbors instead of using the set of all common neighbors as usually considered in the basic local similarity measures. Consequently, a previous clustering scheme must be applied on the graph. Using three different clustering algorithms, we compared WIC measure with ten basic local similarity measures and their counterpart W forms on ten real networks. Our analyses suggest that clustering information, no matter the clustering algorithm used, improves link prediction accuracy.

Keywords: Link Prediction, Complex Networks, Clustering.

1 Introduction

Many social, biological, and information systems can be naturally described as networks, where vertices represent entities (individuals or organizations) and links denote relations or interactions between vertices [18], [30]. Networks or graphs are a powerful representation that has been employed in different tasks of machine learning (ML) and data mining (DM). This growing interest in the use of graph can be justified by the expressiveness of this representation and its applications include: supervised learning [16], [4], [19]; unsupervised learning [6], [25], [24], [20]; and semi-supervised learning [5], [3], [12], to cite just a few.

An important scientific issue regarding network analysis that has attracted increasing attention in recent years is the link prediction. The link prediction problem aims to estimate the likelihood of the future existence of a link between two disconnected vertices in a network, based on the observed links [13].

L.N. Barros, M. Finger, and A.T.R. Pozo (Eds.): SBIA 2012, LNAI 7589, pp. 92–101, 2012.

Many methods for link prediction based on similarity between vertices have been proposed since similar vertices likely share the same relations (links). When the similarity between vertices is based solely on network structure, it is called structural similarity. Structural similarity measures can be classified in different ways, such as the based on local or global information, refer to [18] for details.

Liben-Nowell and Kleinberg [13] and Zhou *et al.* [30] systematically compared a number of structural similarity measures on real networks. According to the authors, global measure can provide higher accuracy, but its computation is very time-consuming and usually infeasible for large-scale networks, while local measure is generally faster but with lower accuracy.

Common Neighbors (CN) [17] is one of the simplest similarity measures based on local information that leads to a good performance. In networks with large clustering coefficient, i.e., if there are connections between a vertex a and two vertices b and c, probably there is a link between b and c, CN provides accurate predictions compared to measures based on global information [30]. The basic assumption of CN is that two vertices are more likely to be connected if they have more common neighbors. Thus, each common neighbor gives equal contribution to the connection likelihood. However, sometimes different common neighbors may play different roles and by identifying them may lead to more accurate prediction than CN [15]. For instance, the common friends in a same social group of two people who do not know each other may contribute more to their possibly future friendship than their common friends from different social groups.

Furthermore, in recent experiments on synthetic and real-world networks, Feng and colleagues found that for a network with low clustering structure link prediction measures based on structural similarity perform poorly. Nonetheless, as the clustering structure of the network grows, the accuracy of these measures drastically improves [9]. Inspired by these results, here, we firstly apply a partitioning scheme to divide the network into communities and then we explicitly use the obtained clustering structure information in the link prediction.

Considering that, to the connection likelihood between a pair of vertices, their common neighbors may contribute in different ways depending on if they belong or not to the same cluster, here we propose a new measure (WIC) for predicting link between a pair of vertices using information from intra-cluster or within-cluster (W) and inter-cluster (IC) common neighbors of these vertices. Also, considering solely the subset of within-cluster common neighbors instead of the set of all common neighbors, we also propose other measures, called *W forms*, derived from the basic local measures. Consequently, a clustering scheme must be applied on the graph analyzed before computing these measures. Thus, using three different clustering algorithms: FastQ (FQ) [6], an algorithm based on edge clustering coefficient (ECC) [25], and WalkTrap (WT) [24]. We compare the WIC measure with ten local similarity measures and their corresponding W forms on ten real networks. We show that cluster information about vertices, no matter the clustering algorithm used, improves the accuracy of link prediction over local similarity measures.

The remainder of the paper is organized as follows. In Section 2 we present the WIC measure. In Section 3 we present ten different local similarity measures and their corresponding W forms. In Section 4 we present experimental results on ten real networks drawn from disparate fields. Finally, in Section 5 we present the conclusions and discuss future work.

2 A Link Prediction Measure Based on Cluster Information

Consider an undirected network $G(V, E)$, where V is the set of vertices and E is the set of links. Multiple links and self-connections are not allowed. Consider the universal set, denote by U, containing all $\frac{|V|\cdot(|V|-1)}{2}$ possible links between vertices in V, where $|V|$ denotes the number of elements in V. The link prediction task is to find out the missing links (future links) in the set $U - E$ (set of nonexistent links) [30].

Given a pair of disconnected vertices (x, y), our task is to determine a similarity measure[1] that uses cluster information from the common neighbors of this pair of vertices. Consider that each vertex $v_i \in V$ is associated with a cluster label C that represents the cluster, community or any group of vertices that share some common properties and/or play similar roles within the network.

In network G exist $M > 1$ cluster labels $C_\alpha, C_\beta, \ldots, C_M$. When a vertex $x \in V$ belongs to a cluster with label C, this vertex is represented as x^C. Consider that a vertex belongs to a unique cluster.

Considering $\Gamma(x)$ denote the set of neighbors of vertex x, we denote by $\Lambda_{x,y} = \Gamma(x) \cap \Gamma(y)$ the set of common neighbors of the pair of disconnected vertices (x, y). According to Bayesian theory [11], the posterior probability that the same cluster label, C_α, be assigned to this pair of vertices, given their common neighbors $\Lambda_{x,y}$, is defined by Eq. 1.

$$P(x^{C_\alpha}, y^{C_\alpha} \mid \Lambda_{x,y}) = \frac{P(\Lambda_{x,y} \mid x^{C_\alpha}, y^{C_\alpha}) P(x^{C_\alpha}, y^{C_\alpha})}{P(\Lambda_{x,y})} \tag{1}$$

Similarly, the posterior probability that different cluster labels, C_α and C_β, be assigned to the vertices (x, y), given their common neighbors $\Lambda_{x,y}$, is defined by Eq. 2.

$$P(x^{C_\alpha}, y^{C_\beta} \mid \Lambda_{x,y}) = \frac{P(\Lambda_{x,y} \mid x^{C_\alpha}, y^{C_\beta}) P(x^{C_\alpha}, y^{C_\beta})}{P(\Lambda_{x,y})} \tag{2}$$

Eqs. 1 and 2 can not tell us which nonexistent links are more likely to exist than others. Nevertheless, we can derive an score measure for pairs of disconnected vertices (x, y) as the ratio of Eq. 1 to 2, as stated in Eq 3.

$$s_{x,y} = \frac{P(\Lambda_{x,y} \mid x^{C_\alpha}, y^{C_\alpha}) P(x^{C_\alpha}, y^{C_\alpha})}{P(\Lambda_{x,y} \mid x^{C_\alpha}, y^{C_\beta}) P(x^{C_\alpha}, y^{C_\beta})} \tag{3}$$

[1] We do not distinguish *similarity measure* and *score*.

Consider that $\Lambda_{x,y} = \Lambda^W_{x,y} \cup \Lambda^{IC}_{x,y}$, where $\Lambda^W_{x,y} = \left\{ z \in \Lambda_{x,y} \mid x^C, y^C, z^C \right\}$ is the set of within-cluster (W) common neighbors and the complement $\Lambda^{IC}_{x,y} = \Lambda_{x,y} \setminus \Lambda^W_{x,y}$ is the set of inter-cluster (IC) common neighbors (common neighbors belonging to C_α, i.e., the same cluster of x, or C_β, the same cluster of y, or C_γ, any other cluster). Clearly, $\Lambda^W_{x,y} \cap \Lambda^{IC}_{x,y} = \varnothing$.

Hence, to estimate the probability of the common neighbors $\Lambda_{x,y}$ given x^{C_α} and y^{C_α}, we must consider the number of common neighbors with the same cluster label C_α by the total of common neighbors, i.e., the more the number of common neighbors in a same cluster the more the likelyhood of x and y belong to this cluster, Eq. 4.

$$P(\Lambda_{x,y} \mid x^{C_\alpha}, y^{C_\alpha}) = \frac{\left|\Lambda^W_{x,y}\right|}{|\Lambda_{x,y}|} \tag{4}$$

Similarly, to estimate the probability of the common neighbors $\Lambda_{x,y}$ given x^{C_α} and y^{C_β}, here we consider the number of common neighbors that may be associated with the cluster labels C_α or C_β or with another cluster label C_γ by the total of common neighbors, as stated in Eq. 5.

$$P(\Lambda_{x,y} \mid x^{C_\alpha}, y^{C_\beta}) = \frac{\left|\Lambda^{IC}_{x,y}\right|}{|\Lambda_{x,y}|} \tag{5}$$

Substituting Eqs. 4 and 5, the likelihood score of the pair of vertices x and y is

$$s_{x,y} = \frac{\left|\Lambda^W_{x,y}\right|}{\left|\Lambda^{IC}_{x,y}\right|} \times \frac{P(x^{C_\alpha}, y^{C_\alpha})}{P(x^{C_\alpha}, y^{C_\beta})} \tag{6}$$

The $\frac{P(x^{C_\alpha}, y^{C_\alpha})}{P(x^{C_\alpha}, y^{C_\beta})}$ ratio can be neglected since either this fraction value is 1 (when $\alpha = \beta$) leading $s_{x,y}$ to $\frac{\left|\Lambda^W_{x,y}\right|}{\left|\Lambda^{IC}_{x,y}\right|}$ or the score $s_{x,y}$ is 0 (when $\alpha \neq \beta$, in this case $\Lambda^W_{x,y} = \varnothing$).

Moreover, to prevent division by zero in the case when $\Lambda^W_{x,y} = \Lambda_{x,y}$ leading to $\Lambda^{IC}_{x,y} = \varnothing$ we add a small value constant $\delta \approx 0$ in denominator. The final score measure is computed by Eq. 7. We notice that $\delta \approx 0$ increases the score when $\Lambda^{IC}_{x,y} = \varnothing$, however it does not modify the evaluation based on the AUC measure, but it may modify the precision.

$$s^{WIC}_{x,y} = \frac{\left|\Lambda^W_{x,y}\right|}{\left|\Lambda^{IC}_{x,y}\right| + \delta} \tag{7}$$

3 Local Similarity Measures and Their W Forms

Different measures based on local information exist such as Common Neighbors (CN), Salton (Sal), Jaccard (Jac), Sørensen (Sor), Hub Promoted Index (HPI), Hub Depressed Index (HDI), Leicht-Holme-Newman index (LHN), Adamic-Adar

(AA), Resource Allocation (RA) and Preferential Attachment (PA), refer to [18] for details. All but PA measures use the set of common neighbors.

The simple counting of the number of common neighbors indicates that each common neighbor gives the same contribution to the connection likelihood. However, as already commented, different common neighbors may give different contributions to the connection probability [15].

Considering that within-cluster common neighbors may contribute more to the connection likelihood than inter-cluster common neighbors, we consider the subset of W common neighbors Λ_{xy}^{W} instead of using the set of all common neighbors Λ_{xy}, obtaining new measures referred to as W forms from the basic measures based on local information. Considering $k(x)$ is the degree of vertex x, Table 1 shows the local similarity measures and their corresponding W forms.

Table 1. Local similarity measures and their corresponding W forms

Local measure	W form
$s_{x,y}^{CN} = \lvert\Lambda_{xy}\rvert$	$s_{x,y}^{CN-W} = \lvert\Lambda_{xy}^{W}\rvert$
$s_{x,y}^{Sal} = \frac{\lvert\Lambda_{xy}\rvert}{\sqrt{k(x)\times k(y)}}$	$s_{x,y}^{Sal-W} = \frac{\lvert\Lambda_{xy}^{W}\rvert}{\sqrt{k(x)\times k(y)}}$
$s_{x,y}^{Jac} = \frac{\lvert\Lambda_{xy}\rvert}{\lvert\Gamma(x)\cup\Gamma(y)\rvert}$	$s_{x,y}^{Jac-W} = \frac{\lvert\Lambda_{xy}^{W}\rvert}{\lvert\Gamma(x)\cup\Gamma(y)\rvert}$
$s_{x,y}^{Sor} = \frac{2\lvert\Lambda_{xy}\rvert}{k(x)+k(y)}$	$s_{x,y}^{Sor-W} = \frac{2\lvert\Lambda_{xy}^{W}\rvert}{k(x)+k(y)}$
$s_{x,y}^{HPI} = \frac{\lvert\Lambda_{xy}\rvert}{min\{k(x),k(y)\}}$	$s_{x,y}^{HPI-W} = \frac{\lvert\Lambda_{xy}^{W}\rvert}{min\{k(x),k(y)\}}$
$s_{x,y}^{HDI} = \frac{\lvert\Lambda_{xy}\rvert}{max\{k(x),k(y)\}}$	$s_{x,y}^{HDI-W} = \frac{\lvert\Lambda_{xy}^{W}\rvert}{max\{k(x),k(y)\}}$
$s_{x,y}^{LHN} = \frac{\lvert\Lambda_{xy}\rvert}{k(x)\times k(y)}$	$s_{x,y}^{LHN-W} = \frac{\lvert\Lambda_{xy}^{W}\rvert}{k(x)\times k(y)}$
$s_{x,y}^{AA} = \sum_{z\in\Lambda_{xy}} \frac{1}{\log k(z)}$	$s_{x,y}^{AA-W} = \sum_{z\in\Lambda_{xy}^{W}} \frac{1}{\log k(z)}$
$s_{x,y}^{RA} = \sum_{z\in\Lambda_{xy}} \frac{1}{k(z)}$	$s_{x,y}^{RA-W} = \sum_{z\in\Lambda_{xy}^{W}} \frac{1}{k(z)}$
$s_{x,y}^{PA} = k(x) \times k(y)$	-

4 Experiments

We consider a scenario where new links of ten real networks from different fields must be predicted. In each one of these networks three different clustering algorithms are used, the FastQ (FQ) [6], an algorithm based on edge clustering coefficient, referred to as ECC [25] and WalkTrap (WT) [24], to assign a cluster label to each vertex. Finally, we compare the performance of our proposal and the W forms to the ones where similarity measures are based only on local information.

4.1 Datasets

The networks considered in our experiments are (i) Airline (AL) [2] (a network of US air transportation system), (ii) Football (FB) [10] (a network of American

football games between Division IA colleges), (iii) Industry (IDT) [8] (a network of companies linked via cooccurrence in 35318 PR Newswire press), (iv) Karate (KT) [29] (a social network of friendships between members of a karate club), (v) Imdb (MN) [21](this network contains movies linked if they share a producer), (vi) NetScience (NS) [23] (a network of coautorship between scientists), (vii) Political Blogs (PB) [1] (a network of the US political blogs. The original links are directed, here we treat them as undirected links), (viii) Yeast (PPI) [27] (a protein-protein interaction network), (ix) Power (PW) [28] (an electrical power grid of western US), (x) Router (RT) [26] (a router-level topology of the Internet). The basic topological features [22] of these ten networks are summarized in Table 2.

4.2 Experimental Results

In this section, we present and discuss the results of using the WIC measure (using $\delta = 0.001$), the local similarity measures and their W forms. Each of the link prediction measures was implemented in `C++` based on LPmade platform [14]. We use the experimental setup presented in [30], [15] and [18]. To test the prediction accuracy, the set of observed links, E, is randomly divided into two parts: the training set, containing 90% of the links, and the probe set, containing the remaining 10% of links. Table 3 summarizes the prediction accuracy results, measured by AUC on the ten networks whose vertices have a cluster label assigned by one of three different algorithms (FQ, ECC and WT).

In general, WIC measure has better performance on six of ten networks. We notice this performance is obtained on networks with large clustering coefficient and large degree of heterogeneity – defined as $\mathbf{H} = \langle k^2 \rangle / \langle k \rangle^2$, where $\langle k \rangle$ denotes the network average degree – such as AL, PPI, IDT and RT, where measures using the set of all common neighbors have few remarkable differences. In addition, in different networks, several W forms outperform, with significant difference, their corresponding basic forms. Note that, the performance of our propo- sals are observed independently of the clustering algorithm used. PA has the worst overall performance.

To show how the cluster information, no matter the clustering algorithm used, improves the link prediction accuracy, we present two different analyses from the results of Table 3. In the first, we analyze the difference between WIC measure and five local similarity measures. In the second, we analyze the statistical distributions of performance of all measures to emphasize the differences between the W forms and their corresponding basic forms.

The measures for link prediction mostly used in the literature are CN, Jac and AA. PA is interesting for requiring minimal information and RA due its similarity to AA. To analyze the difference between WIC measure and these five measures, we use a post-hoc test [7] from the results of the Table 3 and whose analysis is shown in three diagrams in Fig. 1. We show the critical difference (CD) on the top of each diagram. The axis in each diagram is the axis on which we plot the average ranks of measures. In the axis the lowest (best) ranks are in the left side. The measures analyzed have no significant difference, so they are connected by

Table 2. The basic topological features of ten experimental networks. Where $|V|$ and $|E|$ are the number of vertices and links. N_{C-FQ}, N_{C-ECC} and N_{C-WT} are the number of clusters found by the algorithms FQ, ECC and WT, respectively. Each entry for N_{C-FQ}, N_{C-ECC} and N_{C-WT} represents the total number of clusters detected and the number of clusters with a single node, for example, the entry 141/136 for AL with N_{C-FQ} means that the network has 141 clusters, detected by FQ algorithm, with 136 of these clusters are formed by a single node. **C** and **r** are clustering and assortative coeficient, respectively. **H** denotes the heterogeneity degree.

Nets	$\|V\|$	$\|E\|$	N_{C-FQ}	N_{C-ECC}	N_{C-WT}	**C**	**r**	**H**
AL	332	2126	141/136	5/0	32/17	0.6252	-0.2079	3.4639
FB	115	613	51/45	5/0	10/0	0.4032	0.1624	1.0069
IDT	2189	11666	1008/807	396/94	502/308	0.3297	0.1842	3.4122
KT	34	78	16/13	6/2	3/0	0.5706	-0.4756	1.6933
MN	1441	20317	793/523	321/282	359/315	0.5843	0.3492	2.0982
NS	1461	2742	740/524	269/0	1098/1082	0.6937	0.4616	1.8486
PB	1224	16716	598/584	122/92	13/2	0.3203	-0.2211	2.9749
PPI	2617	11855	1091/979	344/79	358/242	0.2844	0.461	3.7284
PW	4941	6594	2023/1985	738/175	234/0	0.0801	0.0034	1.4504
RT	5022	6258	1856/1786	165/2	208/0	0.0116	-0.1384	5.5031

Table 3. The prediction accuracy measured by AUC on ten networks. At each vertex of each network is assigned a cluster label using three different clustering algorithms (FQ, ECC and WT). Each AUC value is obtained by averaging over 100 implementations with independently random divisions of training set and probe set. The entries corresponding to the highest accuracies among the evaluated measures are emphasized by black.

	WIC	**CN**	**CN -W**	**Jac**	**Jac -W**	**Sal**	**Sal -W**	**Sor**	**Sor -W**	**HPI**	**HPI -W**	**HDI**	**HDI -W**	**LHN**	**LHN -W**	**AA**	**AA -W**	**RA**	**RA -W**	**PA**
AL (FQ)	**0.761**	0.669	0.617	0.737	0.717	0.725	0.756	0.668	0.759	0.672	0.755	0.675	0.756	0.677	0.739	0.663	0.756	0.661	0.757	0.535
AL (ECC)	**0.739**	0.669	0.668	0.737	0.697	0.725	0.648	0.668	0.669	0.672	0.674	0.675	0.615	0.677	0.68	0.663	0.589	0.661	0.585	0.535
AL (WT)	**0.738**	0.669	0.612	0.737	0.634	0.725	0.705	0.668	0.691	0.672	0.599	0.675	0.689	0.677	0.67	0.663	0.676	0.661	0.674	0.535
FB (FQ)	**0.929**	0.57	0.658	0.583	0.674	0.59	0.696	0.578	0.742	0.602	0.737	0.584	0.707	0.592	0.755	0.624	0.927	0.602	0.928	0.567
FB (ECC)	**0.839**	0.57	0.821	0.583	0.67	0.59	0.817	0.578	0.827	0.602	0.819	0.584	0.836	0.592	0.838	0.624	0.622	0.602	0.658	0.567
FB (WT)	**0.756**	0.57	0.731	0.583	0.585	0.59	0.706	0.578	0.741	0.602	0.682	0.584	0.728	0.592	0.714	0.624	0.754	0.602	0.749	0.567
IDT (FQ)	**0.729**	0.55	0.586	0.562	0.69	0.562	0.682	0.591	0.692	0.543	0.689	0.563	0.694	0.581	0.708	0.544	0.66	0.545	0.664	0.543
IDT (ECC)	**0.711**	0.55	0.659	0.562	0.656	0.562	0.674	0.591	0.653	0.543	0.637	0.563	0.649	0.581	0.614	0.544	0.655	0.545	0.662	0.543
IDT (WT)	**0.901**	0.55	0.58	0.562	0.883	0.562	0.882	0.591	0.725	0.543	0.711	0.563	0.705	0.581	0.667	0.544	0.743	0.545	0.813	0.543
KT (FQ)	0.657	0.739	0.605	0.697	0.776	0.796	0.705	0.742	0.515	0.756	0.605	0.683	0.515	0.708	0.515	0.832	**0.929**	0.836	**0.929**	0.866
KT (ECC)	0.736	0.739	0.742	0.697	0.639	0.796	**0.913**	0.742	0.838	0.756	0.75	0.683	0.75	0.708	0.805	0.832	0.661	0.836	0.643	0.866
KT (WT)	0.789	0.739	0.867	0.697	0.617	0.796	0.76	0.742	0.608	0.756	**0.929**	0.683	0.586	0.708	0.622	0.832	0.869	0.836	0.863	0.866
MN (FQ)	**0.722**	0.557	0.719	0.573	0.577	0.56	0.712	0.56	0.721	0.549	0.721	0.557	0.701	0.552	0.708	0.565	0.633	0.575	0.607	0.568
MN (ECC)	**0.63**	0.557	0.572	0.573	0.522	0.56	0.56	0.56	0.562	0.549	0.538	0.557	0.573	0.552	0.555	0.565	0.616	0.575	0.623	0.568
MN (WT)	**0.675**	0.557	0.673	0.573	0.567	0.56	0.657	0.56	0.667	0.549	0.633	0.557	0.664	0.552	0.587	0.565	0.622	0.575	0.606	0.568
NS (FQ)	0.743	**0.994**	0.864	0.979	0.848	0.99	0.864	0.993	0.894	0.991	0.823	0.992	0.793	0.985	0.796	0.983	0.803	0.909	0.81	0.784
NS (ECC)	0.774	**0.994**	0.869	0.979	0.793	0.99	0.827	0.993	0.827	0.991	0.877	0.992	0.823	0.985	0.75	0.983	0.781	0.909	0.812	0.784
NS (WT)	0.766	**0.994**	0.765	0.979	0.765	0.99	0.738	0.993	0.745	0.991	0.704	0.992	0.727	0.985	0.723	0.983	0.731	0.909	0.74	0.784
PB (FQ)	0.521	0.552	0.587	0.599	0.599	0.556	0.616	0.562	0.595	0.546	0.574	0.577	0.615	0.567	0.603	0.556	**0.628**	0.549	0.607	0.54
PB (ECC)	0.594	0.552	0.594	0.599	0.613	0.556	**0.642**	0.562	**0.642**	0.546	0.618	0.577	0.62	0.567	0.635	0.556	0.557	0.549	0.551	0.54
PB (WT)	0.511	0.552	0.512	0.599	0.513	0.556	**0.605**	0.562	0.561	0.546	0.547	0.577	0.523	0.567	0.524	0.556	0.556	0.549	0.536	0.54
PPI (FQ)	**0.904**	0.779	0.651	0.785	0.779	0.771	0.683	0.77	0.759	0.763	0.656	0.783	0.781	0.733	0.708	0.783	0.901	0.782	0.902	0.82
PPI (ECC)	**0.905**	0.779	0.899	0.785	0.784	0.771	0.833	0.77	0.804	0.763	0.78	0.783	0.852	0.733	0.693	0.783	0.8	0.782	0.784	0.82
PPI (WT)	**0.911**	0.779	0.892	0.785	0.761	0.771	0.795	0.77	0.815	0.763	0.782	0.783	0.808	0.733	0.576	0.783	0.785	0.782	0.782	0.82
PW (FQ)	0.575	0.575	0.575	0.575	0.505	0.555	0.555	0.57	0.505	0.565	0.505	0.575	0.505	0.555	0.505	0.575	**0.64**	0.575	0.637	0.45
PW (ECC)	0.581	0.575	0.571	0.575	0.635	0.555	0.674	0.57	0.643	0.565	**0.682**	0.575	0.635	0.555	0.67	0.575	0.581	0.575	0.578	0.45
PW (WT)	0.578	0.575	0.577	0.575	0.649	0.555	**0.656**	0.57	0.642	0.565	0.599	0.575	0.644	0.555	0.647	0.575	0.61	0.575	0.611	0.45
RT (FQ)	**0.909**	0.716	0.697	0.533	0.521	0.502	0.908	0.543	0.902	0.502	0.889	0.589	0.817	0.505	0.758	0.743	0.876	0.716	0.876	0.702
RT (ECC)	**0.76**	0.716	0.759	0.533	0.512	0.502	0.505	0.543	0.553	0.502	0.502	0.589	0.557	0.505	0.505	0.743	0.759	0.716	0.752	0.702
RT (WT)	**0.767**	0.716	0.765	0.533	0.526	0.502	0.505	0.543	0.523	0.502	0.511	0.589	0.537	0.505	0.505	0.743	0.765	0.716	0.759	0.702

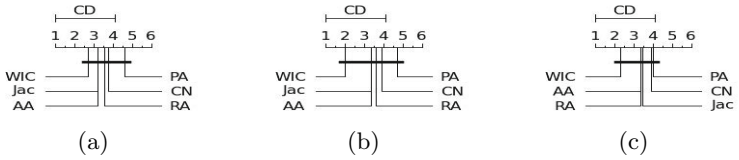

Fig. 1. Comparison of WIC measure with five measures (CN, Jac, AA, PA, RA) with a post-hoc test for results from Table 3. In (a), our proposal uses the information of the cluster labels assigned by the FQ algorithm in the ten networks analyzed. In (b), is used the ECC algorithm, and, in (c), is used the WT algorithm.

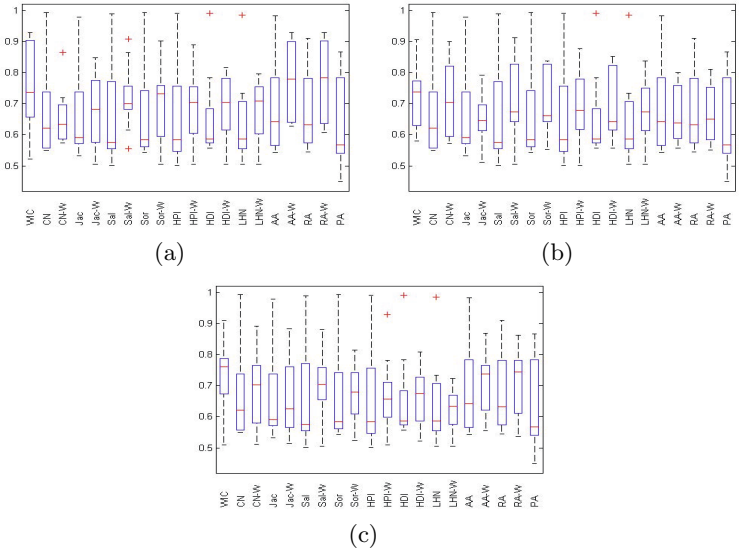

Fig. 2. Statistical distributions of performance of all measures analyzed. The line inside each box indicates the median of the prediction accuracies measured by AUC on ten networks. In (a), our proposals uses the information of the cluster labels assigned by FQ algorithm. In (b), is used the ECC algorithm, and, in (c) is used the WT algorithm.

a black line in each diagram. Although there is no significant difference between these measures, WIC measure has a better average ranking, which in turn is considerably larger than the average ranking of the next best measure (especially when using ECC and WT clustering algorithms). The average ranking of CN, Jac, AA and RA are closer to each other in the three diagrams. The PA has the worst overall average ranking.

Figure 2 shows the statistical distributions of performance of all measures analyzed. There is no single clear winner among the measures, however we can observe the following. First, all medians of accuracies obtained using W forms outperform their corresponding basic forms, no matter the clustering algorithm

used, except for AA-W that is outperformed by a minimal margin by AA when using ECC algorithm. Second, while the median of WIC measure is always between 0.7 and 0.8, the medians of the W forms are always between 0.6 and 0.8, and the medians of the basic forms are always between 0.5 and 0.7. Furthermore, in most cases the difference between median of an W form and its basic form is considerable.

5 Conclusion

We proposed a new measure for link prediction in complex networks, called WIC measure. The WIC measure scores the likelihood of a link between a pair of vertices taking into account the clusters from which common neighbors of these vertices originated. This is, the measure uses the information denoted by high concentration of links within particular groups of vertices as well as by low concentration of links between these groups. Additionally, considering the subset of within-cluster common neighbors, we propose modifications to CN, Salton, Jaccard, Sørensen, HPI, HDI, LHN, AA and RA measures, obtaining their corresponding W forms.

Empirical analysis of our proposals compared with ten local similarity measures on ten real networks from different fields shows that there is no single clear winner but our proposals achieve better accuracies. Thus, the experiments carried out suggest that clustering information, independently of the clustering algorithm used, improves the link prediction accuracy. However the cost in the partitioning process must be considered. Finally, as social network analysis has become a hot topic in the last years, we intend to investigate how this proposal may be adapted to use semantics information of social networks.

Acknowledgements. This work is partially supported by CNPq agency.

References

1. Ackland, R.: Mapping the US political blogosphere: Are conservative bloggers more prominent? Presentation to BlogTalk, Downunder, Sydney (2005)
2. Batageli, V., Mrvar, A.: Pajek datasets (2006), `http://vlado.fmf.uni-lj.si/pub/networks/data/mix/usair97.net`
3. Bertini, J., Lopes, A., Zhao, L.: Partially labeled data stream classification with the semi-supervised K-associated graph. Journal of the Brazilian Computer Society, 1–12 (2012)
4. Bertini, J., Zhao, L., Motta, R., Lopes, A.: A nonparametric classification method based on k-associated graphs. Information Sciences 181(24), 5435–5456 (2011)
5. Blum, A., Chawla, S.: Learning from labeled and unlabeled data using graph mincuts. In: ICML, pp. 19–26 (2001)
6. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Phys. Rev. E 70, 066111 (2004)
7. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. JMLR 7, 1–30 (2006)

8. Fawcett, T., Provost, F.: Activity monitoring: Noticing interesting changes in behavior. In: Proc. of the Fifth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 53–62 (1999)
9. Feng, X., Zhao, J.C., Xu, K.: Link prediction in complex networks: a clustering perspective. Eur. Phys. J. B 85(1-3) (2012)
10. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. PNAS 99(12), 7821–7826 (2002)
11. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: data mining, inference and prediction, 2nd edn. Springer (2009)
12. Laguna, V., Lopes, A.: Combining local and global knn with cotraining. In: ECAI 2010 - 19th European Conference on Artificial Intelligence, vol. 215, pp. 815–820. IOS Press, Netherlands (2010)
13. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. JASIST 58(7), 1019–1031 (2007)
14. Lichtenwalter, R.N., Chawla, N.V.: Lpmade: Link prediction made easy. JMLR 12, 2489–2492 (2011)
15. Liu, Z., Zhang, Q.-M., Lü, L., Zhou, T.: Link prediction in complex networks: A local naive bayes model. EPL 96(48007) (2011)
16. Lopes, A.A., Bertini Jr., J.R., Motta, R., Zhao, L.: Classification Based on the Optimal *K*-Associated Network. In: Zhou, J. (ed.) Complex 2009. LNICST, vol. 4, pp. 1167–1177. Springer, Heidelberg (2009)
17. Lorrain, F., White, H.C.: Structural equivalence of individuals in social networks. Journal of Mathematical Sociology 1, 49–80 (1971)
18. Lü, L., Zhou, T.: Link prediction in complex networks: A survey. Physica A: Statistical Mechanics and its Applications 390(6), 1150–1170 (2011)
19. Lu, Q., Getoor, L.: Link-based classification. In: ICML, pp. 496–503 (2003)
20. Motta, R., de Andrade Lopes, A., de Oliveira, M.C.F.: Centrality Measures from Complex Networks in Active Learning. In: Gama, J., Costa, V.S., Jorge, A.M., Brazdil, P.B. (eds.) DS 2009. LNCS, vol. 5808, pp. 184–196. Springer, Heidelberg (2009)
21. Neville, J., Jensen, D., Friedland, L., Hay, M.: Learning relational probability trees. In: KDD, pp. 625–630 (2003)
22. Newman, M.E.J.: The structure and function of complex networks. SIAM Review (45), 167–256 (2003)
23. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E 74, 036104 (2006)
24. Pons, P., Latapy, M.: Computing communities in large networks using random walks. J. Graph Algorithms Appl. 10(2), 191–218 (2006)
25. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. PNAS 101(9), 2658 (2004)
26. Spring, N., Mahajan, R., Wetherall, D., Anderson, T.: Measuring ISP topologies with rocketfuel. IEEE/ACM Transactions on Networking 12(1), 2–16 (2004)
27. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., Bork, P.: Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417(6887), 399–403 (2002)
28. Watts, D.J., Strogatz, S.H.: Collective dynamics of small-world networks. Nature 393(6684), 440–442 (1998)
29. Zachary, W.W.: An information flow model for conflict and fission in small groups. Journal of Anthropological Research 33(4), 452–473 (1977)
30. Zhou, T., Lü, L., Zhang, Y.-C.: Predicting missing links via local information. Eur. Phys. J. B 71, 623 (2009)