Link prediction in online social networks using group information

Lecture Notes in Computer Science, Cham, v.8584, p.31-45, 2014
http://www.producao.usp.br/handle/BDPI/46585

Link Prediction in Online Social Networks Using Group Information

Jorge Carlos Valverde-Rebaza and Alneu de Andrade Lopes

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

Abstract. Users of online social networks voluntarily participate in different user groups or communities. Researches suggest the presence of strong local community structure in these social networks, i.e., users tend to meet other people via mutual friendship. Recently, different approaches have considered communities structure information for increasing the link prediction accuracy. Nevertheless, these approaches consider that users belong to just one community. In this paper, we propose three measures for the link prediction task which take into account all different communities that users belong to. We perform experiments for both unsupervised and supervised link prediction strategies. The evaluation method considers the links imbalance problem. Results show that our proposals outperform state-of-the-art unsupervised link prediction measures and help to improve the link prediction task approached as a supervised strategy.

Keywords: Link prediction, social networks, communities, social network analysis, graph mining.

1 Introduction

Online social networks are Web platforms that offer to their users the possibility of meeting and networking individuals with similar interests and behaviors [14]. Online social networks such as Flickr, LiveJournal, Orkut and Youtube have become part of the daily life of millions of people around the world who maintain and create new social relationships and interest groups [11]. This fact implies the growth and quick changes in underlying structures (nodes and links) of the social networks over time [8].

The boundless growth of online social networks has resulted in several research directions that examine structural and other properties of large-scale social networks. One of the most relevant research in social networks is the link prediction [8], [10], [14], [13], [9].

B. Murgante et al. (Eds.): ICCSA 2014, Part VI, LNCS 8584, pp. 31–45, 2014.

Link prediction addresses the problem of predicting the existence of missing relations or new ones [8], [10]. Detection of hidden social relationships is a friendship suggestion mechanism used by some online social networks and constitute one of the main application of link prediction. In such case, hidden relationships may consist in existing social ties that have not been established yet in a social network or in social ties missed during the social network evolution [14], [8].

Several methods have been proposed to cope with the link prediction problem. These methods can be divided into two different strategies: unsupervised [8], [13], [14] and supervised [5], [9], [1]. Furthermore, the strategy employed for a specific method influences how its performance will be evaluated [15].

Unsupervised methods assign a score for each pair of nodes with base on neighborhood nodes (local) or path (global) information. The state-of-the-art unsupervised link prediction methods are compared in [8], [10]. According to these experimental results on real networks, global methods usually achieve higher accuracy than local methods. Nevertheless, global methods are very time-consuming and usually infeasible for large-scale networks.

On the other hand, methods based on supervised strategy consider the link prediction problem as a classification problem [5], [1]. Thus, network information such as the structural ones and nodes attributes are used to build a feature vector for each pair of nodes. Then, these vectors are used to train different classifiers to determine the link existence or not between a pair of nodes.

Most proposals have focused on exploiting either the local or the global structural information of the networks. However, other information, such as the behavior of users in social communities, are not properly used. Thus, with the aim of improving the accuracy of link prediction, different hybrid methods using local information and community information have been proposed [18], [13], [12], [7]. These hybrid methods consider that the existence of high concentration of links within communities, as well as the low concentration of links between these communities, is a important property to be exploit in the link prediction problem.

Hybrid methods using community information have a better performance than most of local methods. Notice that these hybrid methods consider that a node belongs to just one community. However, in online social networks users usually belongs to more than one community.

In this paper we propose three new measures for link prediction considering user participation in multiple groups. We compare experimentally the most popular link prediction local methods with our proposals in both unsupervised and supervised strategies.

The remainder of this paper is organized as follows. In Section 2, we present the link prediction problem and state-of-the-art link prediction measures. In Section 3, we present and explain our three proposals. In Section 4, we present experimental results obtained from four online social networks. Finally, in Section 5, we summarize the main findings and conclusions of this work.

2 The Link Prediction Problem

The link prediction problem can be approached in two different strategies: unsupervised and supervised. The evaluation process, i.e., the performance of the link prediction task, therefore, must consider such strategies. Next, we describe both strategies.

2.1 Unsupervised Strategy

Given a network $G = (V, E)$, where V and E are sets of nodes and links respectively. Multiple links and self-connections are not allowed. If G is a directed network, consider the universal set, denoted by U, containing all $|V|(|V| - 1)$ potential directed links between pair of nodes in V, where $|V|$ denotes the number of elements in V. If G is an undirected network, the universal set U contains $\frac{|V|(|V|-1)}{2}$ links. The fundamental link prediction task in the unsupervised context is to find out the missing links (future links) in the set $U - E$ (set of nonexistent links) assigning a score for each link in this set. The higher the score, the higher the connection probability, and vice versa [10], [17], [13], [14].

Most existing unsupervised link prediction methods use node neighborhood (local) or path (global) information. In this work, we use the undirected and directed definitions as in [15] for five local measures: Common Neighbors (CN), Adamic Adar (AA), Jaccard Coefficient (Jac), Resource Allocation (RA) and Preferential Attachment (PA). Afterwards, they are referred to as base measures.

Two standard evaluation measures are used to quantify the prediction accuracy considering the link imbalance problem [10]: AUC (area under the receiver operating characteristic curve) and precision. The AUC is interpreted as the probability that for a randomly chosen link correctly predicted is given a higher score than for a randomly chosen link wrongly predicted. Thus, for n independent comparisons, if n' times for the links correctly predicted are given higher scores than for links wrongly predicted whilst n'' times for both correctly and wrongly predicted links are given equal scores. Thus, the AUC is approximately 0.5 when all the scores are generate from an independent and identical distribution. Therefore, the degree to which the value exceeds 0.5 indicates how better than by pure chance the algorithm performs. The AUC is defined by Eq. 1.

Different from AUC, precision only focuses on the L links with highest scores. Thus, if for the L top-ranked links there are L_r correctly predicted links. The precision is defined by Eq. 2. Clearly, higher precision means higher prediction accuracy.

$$AUC = \frac{n' + 0.5n''}{n} \qquad (1) \qquad\qquad precision_U = \frac{L_r}{L} \qquad (2)$$

2.2 Supervised Strategy

Supervised strategy considers the link prediction problem as a classification problem. Thus, network information such as the structural ones and nodes attributes are used to build a set of feature vectors for linked and not linked pairs of nodes

[5], [1], [15]. Classifiers are able to capture important interdependence relationships between nodes since feature vectors are formed based on unsupervised link prediction measures that capture different structural information sources of networks [9].

Using the supervised strategy is possible to use different validation processes, such as k-fold cross-validation [5]. Thus, we can use the traditional evaluation measures to compare classifiers performance. In this work, we use four standard evaluation measures [4]: accuracy (acc), precision, recall and f-measure (F). These measures are defined as follows:

$$acc = \frac{|tp| + |tn|}{|tp| + |tn| + |fp| + |fn|} \quad (3) \qquad precision_S = \frac{|tp|}{|tp| + |fp|} \quad (4)$$

$$recall = \frac{|tp|}{|tp| + |fn|} \quad (5) \qquad F = \frac{2 \times precision_S \times recall}{precision_S + recall} \quad (6)$$

where $|tp|$, $|tn|$, $|fp|$ and $|fn|$ represent true positives, true negatives, false positives and false negatives rates, respectively.

It is important to notice that the precision for unsupervised strategy is calculated differently than for supervised strategy but in both cases indicates the number of existent links correctly predicted with respect to a set of analyzed links. Furthermore, unsupervised evaluation measures are applied directly on results of link prediction measures but supervised evaluation measures are applied on results of classifiers [15].

3 Proposals

For a network G, we denote by $L_{x,y}$ and $\overline{L}_{x,y}$ the class variables of link existence and nonexistence, respectively, for a pair of nodes $(x, y) \in V$. The prior probabilities of $L_{x,y}$ and $\overline{L}_{x,y}$ are calculated according to Eq. 7 and 8, respectively.

$$P(L_{x,y}) = \frac{|E|}{|U|} \quad (7) \qquad P(\overline{L}_{x,y}) = \frac{|U| - |E|}{|U|} \quad (8)$$

Furthermore, in the network G there exist $M > 1$ groups identified by different group labels $g_1, g_2, \ldots, g_M$. Each node $x \in V$ belongs to a set of node groups $\mathcal{G} = \{g_a, g_b, \ldots, g_p\}$ with size P. Thus, $P > 0$ and $P \leq M$. Each $g_i \in \mathcal{G}$ is a group of nodes, whose elements share interests and behaviors. With M groups in G is possible to form N different sets of groups $\mathcal{G}_\alpha, \mathcal{G}_\beta, \ldots, \mathcal{G}_N$. When the node x belongs to a set of node groups $\mathcal{G}_\alpha$, this node is represented as $x^{\mathcal{G}_\alpha}$. A node belongs to just one set of node groups.

Considering the structural similarity, for undirected networks, the basic structural definition for a node $x \in V$ is its neighborhood $\Gamma(x) = \{y \mid (x, y) \in E \vee (y, x) \in E\}$ which denotes the set of neighbors of x. For directed networks, the set of nodes formed by directed links from x is different from the set of nodes formed by directed links from them to x. Thus, $\Gamma_{out}(x) = \{y \mid (x, y) \in E\}$ is defined as outgoing neighborhood and $\Gamma_{in}(x) = \{y \mid (y, x) \in E\}$ is defined as incoming neighborhood [17]. Also, the set of all common neighbors of the pair of disconnected nodes (x, y) is defined as $\Lambda_{x,y} = \Gamma(x) \cap \Gamma(y)$. When there is directionality, we have $\Lambda^{in}_{x,y} = \Gamma_{in}(x) \cap \Gamma_{in}(y)$ and $\Lambda^{out}_{x,y} = \Gamma_{out}(x) \cap \Gamma_{out}(y)$ [15]. Considering these definitions and based on the approach showed in [13], we propose three new link prediction measures.

3.1 Common Neighbors Within and Outside of Common Groups

For an undirected network, according to Bayesian theory [6], the posterior probabilities of the link existence and nonexistence between a pair of nodes $(x^{\mathcal{G}_\alpha}, y^{\mathcal{G}_\beta})$, given its set of all common neighbors $\Lambda_{x,y}$, are defined by Eq. 9 and 10, respectively.

$$P(L_{x,y}|\Lambda_{x,y}) = \frac{P(\Lambda_{x,y}|L_{x,y})P(L_{x,y})}{P(\Lambda_{x,y})} \quad (9) \qquad P(\overline{L}_{x,y}|\Lambda_{x,y}) = \frac{P(\Lambda_{x,y}|\overline{L}_{x,y})P(\overline{L}_{x,y})}{P(\Lambda_{x,y})} \quad (10)$$

Considering that $\mathcal{G}_{\alpha,\beta} = \mathcal{G}_\alpha \cap \mathcal{G}_\beta$, we define the set of all common neighbors such as $\Lambda_{x,y} = \Lambda^{WCG}_{x,y} \cup \Lambda^{OCG}_{x,y}$, where $\Lambda^{WCG}_{x,y} = \{z^{\mathcal{G}_\gamma} \in \Lambda_{x,y} \mid \mathcal{G}_{\alpha,\beta} \cap \mathcal{G}_\gamma \neq \varnothing\}$ is the set of common neighbors within common groups (WCG), i.e., the common neighbors of x and y belonging to at least one group to which both x and y belong to. The complement, $\Lambda^{OCG}_{x,y} = \Lambda_{x,y} - \Lambda^{WCG}_{x,y}$ is the set of common neighbors outside of the common groups (OCG), i.e., the common neighbors of x and y belonging to any group except to one group to which both x and y belong to. Clearly, $\Lambda^{WCG}_{x,y} \cap \Lambda^{OCG}_{x,y} = \varnothing$.

Hence, to estimate the probability of the common neighbors $\Lambda_{x,y}$ given the connection between $x^{\mathcal{G}_\alpha}$ and $y^{\mathcal{G}_\beta}$, we have to consider the number of common neighbors within common groups by the number of all common neighbors, as stated in Eq. 11. Similarly, to estimate the probability of the common neighbors $\Lambda_{x,y}$ given a disconnection between $x^{\mathcal{G}_\alpha}$ and $y^{\mathcal{G}_\beta}$, we have to consider the number of common neighbors outside of the common groups by the number of all common neighbors, as stated in Eq. 12.

$$P(\Lambda_{x,y} \mid L_{x,y}) = \frac{|\Lambda^{WCG}_{x,y}|}{|\Lambda_{x,y}|} \quad (11) \qquad P(\Lambda_{x,y} \mid \overline{L}_{x,y}) = \frac{|\Lambda^{OCG}_{x,y}|}{|\Lambda_{x,y}|} \quad (12)$$

In order to compare the existence likelihood between $x^{\mathcal{G}_\alpha}$ and $y^{\mathcal{G}_\beta}$, in Eq. 13, we define the likelihood score, $s_{x,y}$, of a node pair (x, y) as the ratio between Eq. 9 and 10.

$$s_{x,y} = \frac{P(\Lambda_{x,y} \mid L_{x,y})P(L_{x,y})}{P(\Lambda_{x,y} \mid \overline{L}_{x,y})P(\overline{L}_{x,y})} \quad (13)$$

Substituting Eq. 11 and 12, we have the final score referred to as the **common neighbors within and outside of common groups** (WOCG) measure, defined as:

$$s^{WOCG}_{x,y} = \frac{|\Lambda^{WCG}_{x,y}|}{|\Lambda^{OCG}_{x,y}|} \times \Omega \quad (14)$$

where $\Omega = \frac{P(L_{x,y})}{P(\overline{L}_{x,y})} = \frac{|E|}{|U|-|E|}$ is a constant for a network and its computation can be disregarded.

For a directed network, we consider $\Lambda_{x,y}^{WCG_{in}} = \{z^{\mathcal{G}_\gamma} \in \Lambda_{x,y}^{in} \mid \mathcal{G}_{\alpha,\beta} \cap \mathcal{G}_\gamma \neq \varnothing\}$, $\Lambda_{x,y}^{WCG_{out}} = \{z^{\mathcal{G}_\gamma} \in \Lambda_{x,y}^{out} \mid \mathcal{G}_{\alpha,\beta} \cap \mathcal{G}_\gamma \neq \varnothing\}$, $\Lambda_{x,y}^{OCG_{in}} = \Lambda_{x,y}^{in} - \Lambda_{x,y}^{WCG_{in}}$ and $\Lambda_{x,y}^{OCG_{out}} = \Lambda_{x,y}^{out} - \Lambda_{x,y}^{WCG_{out}}$. Thus, WOCG is defined based on the link direction: $s_{x,y}^{WOCG_{in}} = \frac{|\Lambda_{x,y}^{WCG_{in}}|}{|\Lambda_{x,y}^{OCG_{in}}|}$ and $s_{x,y}^{WOCG_{out}} = \frac{|\Lambda_{x,y}^{WCG_{out}}|}{|\Lambda_{x,y}^{OCG_{out}}|}$.

3.2 Common Neighbors of Groups

For an undirected network, considering a pair of nodes $(x^{\mathcal{G}_\alpha}, y^{\mathcal{G}_\beta})$, we define the set of common neighbors of groups $\Lambda_{x,y}^{\mathcal{G}} = \{z^{\mathcal{G}_\gamma} \in \Lambda_{x,y} \mid \mathcal{G}_\alpha \cap \mathcal{G}_\gamma \neq \varnothing \vee \mathcal{G}_\beta \cap \mathcal{G}_\gamma \neq \varnothing\}$. Thus, we define a score referred to as **common neighbors of groups** (CNG), as stated in Eq. 15.

$$s_{x,y}^{CNG} = |\Lambda_{x,y}^{\mathcal{G}}| \tag{15}$$

The CNG measure refers to the size of the set of common neighbors of x and y belonging to at least one group to which x or y belongs to.

For a directed network, we define the set of incoming common neighbors of groups $\Lambda_{x,y}^{\mathcal{G}_{in}} = \{z^{\mathcal{G}_\gamma} \in \Lambda_{x,y}^{in} \mid \mathcal{G}_\alpha \cap \mathcal{G}_\gamma \neq \varnothing \vee \mathcal{G}_\beta \cap \mathcal{G}_\gamma \neq \varnothing\}$ and the set of outgoing common neighbors of groups $\Lambda_{x,y}^{\mathcal{G}_{out}} = \{z^{\mathcal{G}_\gamma} \in \Lambda_{x,y}^{out} \mid \mathcal{G}_\alpha \cap \mathcal{G}_\gamma \neq \varnothing \vee \mathcal{G}_\beta \cap \mathcal{G}_\gamma \neq \varnothing\}$. Thus, CNG is defined based on the link direction: $s_{x,y}^{CNG_{in}} = |\Lambda_{x,y}^{\mathcal{G}_{in}}|$ and $s_{x,y}^{CNG_{out}} = |\Lambda_{x,y}^{\mathcal{G}_{out}}|$.

3.3 Common Neighbors with Total and Partial Overlapping of Groups

For an undirected network, we formulate a new proposal. Thus, according to Bayesian theory, the posterior probabilities of link existence and nonexistence between a pair of nodes $(x^{\mathcal{G}_\alpha}, y^{\mathcal{G}_\beta})$, given its set of common neighbors of groups $\Lambda_{x,y}^{\mathcal{G}}$, are defined by Eq. 16 and 17, respectively.

$$P(L_{x,y}|\Lambda_{x,y}^{\mathcal{G}}) = \frac{P(\Lambda_{x,y}^{\mathcal{G}}|L_{x,y})P(L_{x,y})}{P(\Lambda_{x,y}^{\mathcal{G}})} \tag{16}$$

$$P(\overline{L}_{x,y}|\Lambda_{x,y}^{\mathcal{G}}) = \frac{P(\Lambda_{x,y}^{\mathcal{G}}|\overline{L}_{x,y})P(\overline{L}_{x,y})}{P(\Lambda_{x,y}^{\mathcal{G}})} \tag{17}$$

Consider that $\Lambda_{x,y}^{\mathcal{G}} = \Lambda_{x,y}^{TOG} \cup \Lambda_{x,y}^{POG}$, where $\Lambda_{x,y}^{TOG} = \{z^{\mathcal{G}_\gamma} \in \Lambda_{x,y}^{\mathcal{G}} \mid \mathcal{G}_\alpha \cap \mathcal{G}_\gamma \neq \varnothing \ \wedge \ \mathcal{G}_\beta \cap \mathcal{G}_\gamma \neq \varnothing\}$ is the set of common neighbors with total overlapping of groups (TOG), i.e., the common neighbors of group of x and y belonging to at least one group of nodes to which x and y belong to. The complement, $\Lambda_{x,y}^{POG} = \Lambda_{x,y}^{\mathcal{G}} - \Lambda_{x,y}^{TOG} = \{z^{\mathcal{G}_\gamma} \in \Lambda_{x,y}^{\mathcal{G}} \mid \mathcal{G}_\alpha \cap \mathcal{G}_\gamma \neq \varnothing \veebar \mathcal{G}_\beta \cap \mathcal{G}_\gamma \neq \varnothing\}$ is the set of common neighbors with partial overlapping of groups (POG), i.e., the common neighbors of groups of x and y belonging exclusively to at least one group of nodes to which x or y belong to. Clearly, $\Lambda_{x,y}^{TOG} \cap \Lambda_{x,y}^{POG} = \varnothing$.

Using the same process presented in Section 3.1, we can estimate the probability of the common neighbors of groups $\mathit{\Lambda}^{\mathcal{G}}_{x,y}$ given the probability of link existence and nonexistence between $x^{\mathcal{G}_\alpha}$ and $y^{\mathcal{G}_\beta}$ as stated in Eqs. 18 and 19.

$$P(\mathit{\Lambda}^{\mathcal{G}}_{x,y}|L_{x,y}) = \frac{|\mathit{\Lambda}^{TOG}_{x,y}|}{|\mathit{\Lambda}^{\mathcal{G}}_{x,y}|} \qquad (18) \qquad\qquad P(\mathit{\Lambda}^{\mathcal{G}}_{x,y}|\overline{L}_{x,y}) = \frac{|\mathit{\Lambda}^{POG}_{x,y}|}{|\mathit{\Lambda}^{\mathcal{G}}_{x,y}|} \qquad (19)$$

In order to compare the existence likelihood between $x^{\mathcal{G}_\alpha}$ and $y^{\mathcal{G}_\beta}$, we define the likelihood score of a node pair (x, y) as the ratio between Eq. 16 and Eq. 17. Substituting Eq. 18 and Eq. 19, we have the final score called as the **common neighbors with total and partial overlapping of groups** (TPOG) measure, defined as:

$$s^{TPOG}_{x,y} = \frac{|\mathit{\Lambda}^{TOG}_{x,y}|}{|\mathit{\Lambda}^{POG}_{x,y}|} \times \mathit{\Omega} \qquad (20)$$

where $\mathit{\Omega} = \frac{P(L_{x,y})}{P(\overline{L}_{x,y})} = \frac{|E|}{|U|-|E|}$, in the same way that for WOCG, is a constant for a network and its computation can be disregarded.

For a directed network, we can consider $\mathit{\Lambda}^{TOG_{in}}_{x,y} = \{z^{\mathcal{G}_\gamma} \in \mathit{\Lambda}^{\mathcal{G}_{in}}_{x,y} \mid \mathcal{G}_\alpha \cap \mathcal{G}_\gamma \neq \varnothing \;\wedge\; \mathcal{G}_\beta \cap \mathcal{G}_\gamma \neq \varnothing\}$, $\mathit{\Lambda}^{TOG_{out}}_{x,y} = \{z^{\mathcal{G}_\gamma} \in \mathit{\Lambda}^{\mathcal{G}_{out}}_{x,y} \mid \mathcal{G}_\alpha \cap \mathcal{G}_\gamma \neq \varnothing \;\wedge\; \mathcal{G}_\beta \cap \mathcal{G}_\gamma \neq \varnothing\}$, $\mathit{\Lambda}^{POG_{in}}_{x,y} = \{z^{\mathcal{G}_\gamma} \in \mathit{\Lambda}^{\mathcal{G}_{in}}_{x,y} \mid \mathcal{G}_\alpha \cap \mathcal{G}_\gamma \neq \varnothing \;\veebar\; \mathcal{G}_\beta \cap \mathcal{G}_\gamma \neq \varnothing\}$ and $\mathit{\Lambda}^{POG_{out}}_{x,y} = \{z^{\mathcal{G}_\gamma} \in \mathit{\Lambda}^{\mathcal{G}_{out}}_{x,y} \mid \mathcal{G}_\alpha \cap \mathcal{G}_\gamma \neq \varnothing \;\veebar\; \mathcal{G}_\beta \cap \mathcal{G}_\gamma \neq \varnothing\}$. Thus, TPOG is defined based on the link direction: $s^{TPOG_{in}}_{x,y} = \frac{|\mathit{\Lambda}^{TOG_{in}}_{x,y}|}{|\mathit{\Lambda}^{POG_{in}}_{x,y}|}$ and $s^{TPOG_{out}}_{x,y} = \frac{|\mathit{\Lambda}^{TOG_{out}}_{x,y}|}{|\mathit{\Lambda}^{POG_{out}}_{x,y}|}$.

4 Experiments

We consider a scenario where new links of four online social networks must be predicted. Due to the fact that in online social networks users participate freely in different user groups, in each one of these social networks we use this natural group information to assign group labels to each node. We also compare the performance of our proposals to the base measures.

4.1 Datasets

Social network graphs considered in our experiments are Flickr, LiveJournal, Orkut and Youtube. These graphs, available in [11], are among the most popular social networking sites. On the other hand, these graphs have information both of links between users and of friendship groups to which each user belongs.

Each online social network have different features. Flickr[1] is a photo-sharing network to organize images using tags and allows users to form groups of common photography interests. LiveJournal[2] is a popular blogging site whose users form a social network and create custom user groups for posting discussion. Orkut[3] is

[1] http://www.flickr.com

[2] http://www.livejournal.com

[3] http://www.orkut.com

a social networking site run by Google considered a pure social network since it has the sole purpose of friendship networking and allows users to create groups of users with similar interests. Youtube[4] is a popular video-sharing site that includes a social network that allows users to create groups of users with similar video preferences.

Table 1. High-level topological features of our four social network graphs

	Flickr	**LiveJournal**	**Orkut**	**Youtube**
Number of nodes	1,846,198	5,284,457	3,072,441	1,157,827
Number of links	22,613,981	77,402,652	223,534,301	4,945,382
Average degree per node	12.24	16.97	106.1	4.29
Fraction of links symmetric	62.0%	73.5%	100.0%	79.1%
Average path length	5.67	5.88	4.25	5.10
Diameter	27	20	9	21
Average clustering coefficient	0.313	0.330	0.171	0.136
Average assortativity coefficient	0.202	0.179	0.072	−0.033
Number of node groups	103,648	7,489,073	8,730,859	30,087
Average number of groups membership per node	4.62	21.25	106.44	0.25
Average group size	82	15	37	10
Average group clustering coefficient	0.47	0.81	0.52	0.34

High-level topological features of the four social network graphs are presented in Table 1. From this table, we observe that by the high number of nodes and links these networks are considered as large-scale networks. The average degree per node indicates the average of number of neighbors per user. The fraction of links symmetric denotes the degree in which directed links from a source to a destination have an endorsement of the destination by the source. For instance, with the exception of Orkut, which is an undirected network (with 100% of links symmetric), the other networks (directed networks) have a significant degree of symmetry, i.e., many of the target of the links reciprocate. Furthermore, independent of the causes, the symmetric nature of social links affects the structure of large scale social networks, mainly by increasing the overall connectivity of the network and reducing its diameter [11].

Also, Table 1 shows global topological features of networks. The average path length is the average number of steps along the shortest paths for all possible node pairs and the diameter is defined as the maximum shortest path between any two nodes. In absolute terms, the average path lengths and diameters for all four networks are remarkably shorter compared with average path length and diameter of the Web graph (16.12 and 905, respectively) [2]. The average clustering coefficient is the degree to which nodes in a network tend to cluster together. A high average clustering coefficient suggests the presence of strong local community structure, i.e., in friendship social networks, users tend to be introduced to other users via mutual friends. The average assortativity coefficient indicates the likelihood for nodes to connect to other nodes with similar degrees. When this coefficient tends to 1, means that nodes likely are connected to nodes with similar degrees, and the opposite when the value tends to -1.

[4] `http://www.youtube.com`

Between the group features, we observe that the four networks have a high amount of node groups and that each user belongs on average to more than four groups (except Youtube). Thus, all groups of all four networks have a minimum of 10 users. It is important to note here that for each network each user can belongs to more than one group. Also, note that users in a group do not necessarily need link to each other in the network and user groups represent tightly clustered communities of users in social networks. This can be seen from the average group clustering coefficients, which is defined as the average of clustering coefficients of the subgraphs of the network consisting of only the users who are members of each group [11].

4.2 Experimental Setup

For the network preprocessing, for a network G, the set E is divided into the training set E^T and the testing set E^P. From the set E, for selecting the links for E^P, we take randomly two-third of the links formed by nodes whose number of neighbors is two times greater than the average degree per node. The remaining links, except those formed by nodes whose number of neighbors is less than two-third of the average degree per node, constitute the training set E^T. This evaluation method is widely used in the link prediction literature [14], [15], [16], [17].

After that, the link prediction process is initiated. This process includes both unsupervised and supervised strategies. In unsupervised strategy, for each pair of nodes from E^T, the connection likelihood is calculated based on the link direction, choosing the highest score between its *in* and *out* scores as final and unique score, e.g., by vertex pair (x, y) if $s_{x,y}^{out} > s_{x,y}^{in}$ then $s_{x,y} = s_{x,y}^{out}$, otherwise, $s_{x,y} = s_{x,y}^{in}$.

In supervised strategy, we use decision tree (J48), naive Bayes (NB), multilayer perceptron with backpropagation (MLP) and support vector machine (SMO) classifiers from Weka[5]. Previously, for each network, we compute a set of feature vector formed by randomly selected pair of nodes from E^T. If the pair of nodes taken from E^T is also in E^P then the feature vector formed by this pair of nodes takes the positive class (existent link), otherwise takes the negative class (nonexistent link). To avoid the links imbalance problem, the set of feature vectors for each network have 50% with positive class and 50% with negative class. Table 2 shows the number of instances by class and the total of instances for each social network.

For each network, we create five different data sets in ARFF format. Each data set is formed by features which combine different link prediction measures. Thus, VLocal is the data set whose feature vectors are formed by CN, AA, Jac, RA and PA. VGroup is the data set whose feature vectors are formed by WOCG, CNG and TPOG. VTop is the data set whose feature vectors are formed by the three best base measures from the literature, i.e., CN, AA and RA, and the two best measures based on group information, i.e., CNG and TPOG (see Section

[5] `http://www.cs.waikato.ac.nz/ml/weka/`

Table 2. Number of instances by class for all networks

	Existent	**Non-existent**	**Total**
Flickr	500001	500001	1000002
LiveJournal	300001	300001	600002
Orkut	1500001	1500001	3000002
Youtube	20001	20001	40002

4.3). Similarly, VTop2 is the data set whose feature vectors are formed by the five overall best link prediction measures (see also Section 4.3), i.e., TPOG, CNG, AA, WOCG and CN. VTotal is the data set whose feature vectors are formed by all base measures and all our proposals, i.e., CN, AA, Jac, RA, PA, WOCG, CNG and TPOG.

The experiments were carried out in a computer with 99 GB of RAM using Linux operating system.

4.3 Results

In order to validate our results, we use the evaluation measures presented in Section 2 both for unsupervised strategy and supervised strategy. For results of our unsupervised link prediction process, we employ AUC and precision to validation. Table 3 summarizes the prediction results measured by AUC, with $n = 5000$, for the four networks. Each AUC value is obtained by averaging over 10 run over 10 independent partitions of training and testing sets.

Table 3. The prediction results measured by AUC

	WOCG	**CNG**	**TPOG**	**CN**	**AA**	**Jac**	**RA**	**PA**
Flickr	0.637 (5.0)	0.728 (1.0)	0.728 (2.0)	0.674 (3.0)	0.656 (4.0)	0.431 (8.0)	0.616 (6.0)	0.566 (7.0)
Livejournal	0.596 (4.0)	0.611 (3.0)	0.665 (1.0)	0.582 (5.0)	0.580 (6.0)	0.624 (2.0)	0.565 (7.0)	0.542 (8.0)
Orkut	0.649 (2.0)	0.621 (3.0)	0.651 (1.0)	0.572 (7.0)	0.620 (4.0)	0.575 (6.0)	0.566 (8.0)	0.602 (5.0)
Youtube	0.434 (7.0)	0.723 (5.0)	0.555 (6.0)	0.834 (4.0)	0.928 (1.0)	0.217 (8.0)	0.892 (3.0)	0.917 (2.0)
Average rank	4.50 (4.0)	3.00 (2.0)	2.50 (1.0)	4.75 (5.0)	3.75 (3.0)	6.00 (7.5)	6.00 (7.5)	5.50 (6.0)

From Table 3, each value in parentheses represents the ranking of each link prediction measure for each network. In general, our proposals perform better than the base measures in Flickr, LiveJournal and Orkut. In Youtube, TPOG and CNG have their lowest performance and WOCG has the overall worst performance. This can be explained by the fact that Youtube has the lowest values of average clustering coefficient and average group clustering coefficient, i.e., friends of a user does not necessarily become friends and user groups are weakly dense. Also, Youtube has a negative value of average assortativity coefficient, i.e., there is a tendency of friendship relations between users that share few common interests and behaviors. Among the base measures, highlighted as best CN and AA and, surprisingly, PA has a better performance than Jac and RA.

To analyze the difference between all link prediction measures, based on results of Table 3, we perform the Friedman and Nemenyi post-hoc tests [3]. The critical value of the F-statistics with 7 and 21 degrees of freedom at 95 percentile is 2.49. Thus, according to the Friedman test using the F-statistics, the null-hypothesis that all link prediction measures evaluated behave similarly should not be rejected. According to the Nemenyi statistics, the critical difference (CD) for comparing the mean-ranking of two different link prediction measures at 95 percentile is 5.25.

Results from Nemenyi test are present in Figure 1, where we show the critical difference value on the top of the diagram. In the axis of the diagram are plotted the average rank of measures (whose values are explicit in the last row of Table 3). In the axis, the lowest (best) ranks are in the left side. Thus, the null-hypothesis that all link prediction measures have a similar behavior should not be rejected, i.e., all measures analyzed have no significant difference, so they are connected by a black line in the diagram. Although there is no significant difference among them, we observe that our proposals have the first, second and fourth best overall accuracy. The base measure best positioned is AA, which is third, and CN, which is fifth.

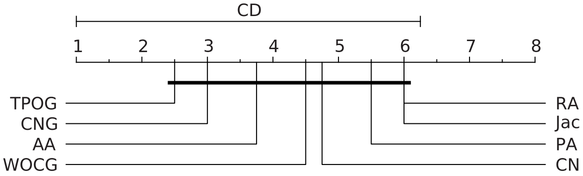
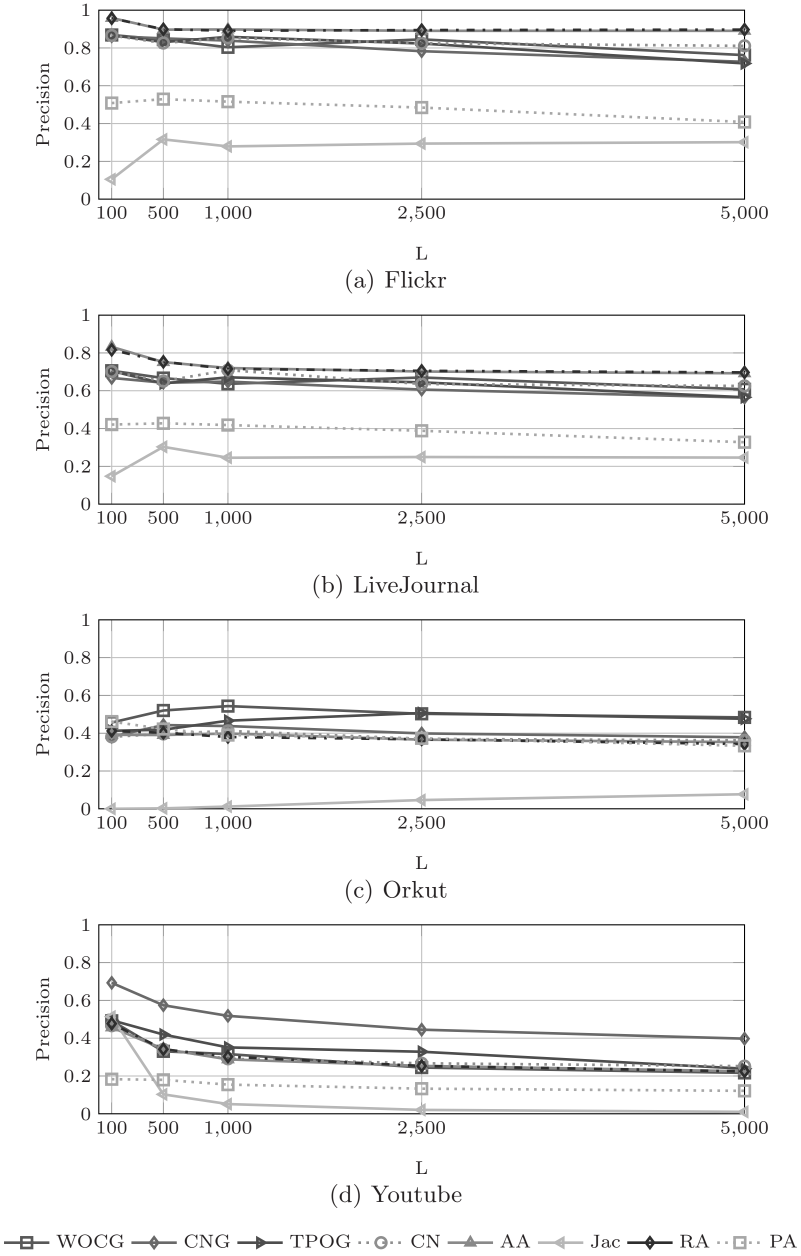


Fig. 1. Post-hoc test for results from Table 3 with CD = 5.25

Figure 2 shows the prediction quality measured by precision on all social networks analyzed. Different values of L are used. For Flickr and LiveJournal, all link prediction measures have a similar precision performance, highlighting AA and RA as the best overall measures in all L values, but reaching their maximum performance when $L = 100$. For Orkut, also all link prediction measures have a similar precision performance, highlighting WOCG and TPOG as the best overall measures in all L values. However, WOCG reaches the highest overall performance when $L = 1,000$. For Youtube, we observe a declining performance in all link prediction measures after $L = 100$, highlighting CNG and TPOG as the best overall measures in all L values. Furthermore, CNG reaches the highest overall performance when $L = 100$. Also, we observe that PA and Jac are the worst overall performance in all the four networks.

For results of our supervised link prediction process, accuracy and f-measure are employed to validate the quality of the classifiers in VLocal, VGroup, VTop, VTop2 and VTotal data sets for each social network. Tables 4 and 5 respectively show Accuracy and F-Value average values for four different classifiers after using 10-fold cross validation. For both Tables 4 and 5, values emphasized in



(a) Flickr

(b) LiveJournal

(c) Orkut

(d) Youtube

Fig. 2. Precision results on four social networks. Different values of L are used to select the top-L highest scores for predicting links.

black correspond to the highest result among the evaluated data sets for each classifier. Results highlighted in gray indicate that a classifier get best results in data sets formed by feature vectors using our proposals than VLocal data set, which is formed by feature vectors using only the base measures.

Table 4. Correctly classified instances (in percent)

	J48	**NB**	**MLP**	**SMO**		**J48**	**NB**	**MLP**	**SMO**
Flickr Vlocal	77.70	57.94	**71.29**	66.88	Orkut VLocal	82.53	71.91	80.01	76.75
Flickr VGroup	70.66	**62.50**	69.91	67.98	Orkut VGroup	78.11	69.35	77.32	74.14
Flickr VTop	72.35	59.08	71.13	68.60	Orkut VTop	79.86	72.99	77.01	76.12
Flickr VTop2	72.20	60.96	70.89	68.08	Orkut VTop2	79.32	73.09	77.31	76.09
Flickr VTotal	**77.72**	60.37	71.20	**68.97**	Orkut VTotal	**82.59**	**74.14**	**80.13**	**77.32**
LiveJournal VLocal	**79.70**	70.66	**78.85**	**77.67**	Youtube VLocal	82.35	59.73	**73.08**	62.01
LiveJournal VGroup	76.94	71.01	76.94	75.34	Youtube VGroup	67.24	61.16	67.09	61.45
LiveJournal VTop	79.14	71.63	78.76	77.50	Youtube VTop	78.94	60.55	72.43	64.71
LiveJournal VTop2	79.12	70.92	78.15	77.46	Youtube VTop2	78.03	**63.67**	71.69	64.39
LiveJournal VTotal	**79.70**	**71.77**	78.59	77.61	Youtube VTotal	**82.66**	62.38	72.33	**65.20**

Table 4 shows results for J48, NB, MLP and SMO classifiers. In most cases, the best accuracy is obtained by VTotal data set, i.e., the data set formed by feature vectors using all our proposals. For MLP classifier the best result is by using the VLocal data set, i.e., the data set formed only by the base measures. For Orkut, the best performance of MLP classifier is using the VTotal data set. Besides, we observe that using NB, for all the four networks, and using SMO, for Flickr and Youtube, the performance of classifiers in data sets formed by feature vectors using our proposals (VGroup, VTop, VTop2 and VTotal) is markedly better than in data sets formed by feature vectors using only base measures (VLocal).

Table 5 shows for J48, NB and SMO classifiers that the best f-measure results are obtained by the VTotal data set or in any other data set whose feature vectors use our proposals. For MLP classifier the best result is using the VLocal data set.

Table 5. Average of f-measure on four social networks

	J48	**NB**	**MLP**	**SMO**		**J48**	**NB**	**MLP**	**SMO**
Flickr VLocal	**0.777**	0.507	**0.713**	0.651	Orkut VLocal	0.825	0.702	0.800	0.764
Flickr VGroup	0.706	**0.583**	0.699	0.668	Orkut VGroup	0.781	0.676	0.773	0.737
Flickr VTop	0.724	0.525	0.711	0.676	Orkut VTop	0.799	0.720	0.77	0.759
Flickr VTop2	0.722	0.558	0.709	0.669	Orkut VTop2	0.793	0.722	0.773	0.758
Flickr VTotal	**0.777**	0.548	0.712	**0.680**	Orkut VTotal	**0.826**	**0.731**	**0.801**	**0.771**
LiveJournal VLocal	**0.797**	0.687	**0.788**	**0.774**	Youtube VLocal	0.823	0.531	**0.73**	0.565
LiveJournal VGroup	0.768	0.698	0.768	0.750	Youtube VGroup	0.658	0.563	0.655	0.567
LiveJournal VTop	0.791	0.700	0.787	0.772	Youtube VTop	0.789	0.543	0.724	0.617
LiveJournal VTop2	0.79	0.691	0.781	0.772	Youtube VTop2	0.780	**0.600**	0.717	0.613
LiveJournal VTotal	**0.797**	**0.702**	0.786	**0.774**	Youtube VTotal	**0.826**	0.577	0.723	**0.623**

In general, we observe that the performance of a classifier measured by accuracy is similar that measured by f-measure. Thus, from entries highlighted in gray in Tables 4 and 5 we observe that classifiers perform better in data sets formed by feature vectors that include our proposals. This happens mainly when all base measures and all our proposals are combined into a feature vector, i.e., in the VTotal data set.

5 Conclusions

We proposed three new link prediction measures, referred to as WOCG, CNG and TPOG measures. These measures use information about the groups to which the nodes belong. Differently from the link prediction measures based on group information described in the literature, our proposals consider that a node can belong to more than one group, as usually occurs in real social networks. Thus, WOCG divides the common neighbor set of two nodes and use the neighborhood intersection information. CNG defines the set of common neighbors of two nodes belonging to at least one group to which these nodes belong and use the size of this set as a link prediction measure. TPOG uses the same schema that WOCG but using the set of common neighbors of groups defined by the CNG measure.

Since for applying measures based on group information is need, in a previous phase, partitioning the network into groups, researchers use different community detection algorithms that have low computational cost and that improve the link prediction performance [14], [13], [7], [12]. This makes the performance of link prediction measures based on group information strongly depend of clustering quality. Thus, to eliminate this dependence and spend less time executing a community detect algorithm, in social networks domain we can use the natural group information, i.e., the information from friendship groups to which users belong to.

To evaluate our proposals, we use both unsupervised and supervised strategies on four real and large-scale online social networks: Flickr, LiveJournal, Orkut and Youtube. When an unsupervised strategy is performed, the performance of our proposals compared to other measures was better under the AUC criterion. When analyzing precision, highlight AA, RA, WOCG, TPOG and CNG but there is no clear winner. It is important to note here that the performance of our proposals is influenced by the topological structure of the analyzed network.

When a supervised strategy is performed, our results show that combining measures based on local information and based on group information improves the performance of classifiers. But the improvement may not be significant because the selection processes for generating feature vectors of data sets are diverse, so how to select the most appropriate links for a supervised strategy is a challenging problem.

In summary, our experiments suggest that communities to which users belong convey relevant clues about user's interest and behavior. Hence, our proposals improve the performance of the link prediction task by considering mainly the information of common groups to which users belong to.

Acknowledgment. This work is partially supported by Grants 2011/22749-8 and 2013/12191-5 from São Paulo Research Foundation (FAPESP) and 151836/2013-2 from National Council for Scientific and Technological Development (CNPq).

References

1. Benchettara, N., Kanawati, R., Rouveirol, C.: A supervised machine learning link prediction approach for academic collaboration recommendation. In: RecSys 2010, pp. 253–256 (2010)
2. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph structure in the web. Comput. Netw. 33(1-6), 309–320 (2000)
3. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. JMLR 7, 1–30 (2006)
4. Fatourechi, M., Ward, R.K., Mason, S.G., Huggins, J., Schlogl, A., Birch, G.E.: Comparison of evaluation metrics in classification applications with imbalanced datasets. In: ICMLA 2008, pp. 777–782 (2008)
5. Al Hasan, M., Chaoji, V., Salem, S., Zaki, M.: Link prediction using supervised learning. In: SDM 2006 Workshop on Link Analysis, Counterterrorism and Security (2006)
6. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: data mining, inference and prediction, 2nd edn. Springer (2009)
7. Hoseini, E., Hashemi, S., Hamzeh, A.: Link prediction in social network using co-clustering based approach. In: WAINA 2012, pp. 795–800. IEEE (2012)
8. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. JASIST 58(7), 1019–1031 (2007)
9. Lichtenwalter, R.N., Lussier, J.T., Chawla, N.V.: New perspectives and methods in link prediction. In: ACM SIGKDD KDD 2010, pp. 243–252. ACM (2010)
10. Lü, L., Zhou, T.: Link prediction in complex networks: A survey. Physica A: Statistical Mechanics and its Applications 390(6), 1150–1170 (2011)
11. Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: ACM SIGCOMM IMC 2007, pp. 29–42. ACM (2007)
12. Soundarajan, S., Hopcroft, J.: Using community information to improve the precision of link prediction methods. In: WWW 2012, pp. 607–608. ACM (2012)
13. Valverde-Rebaza, J., de Andrade Lopes, A.: Link prediction in complex networks based on cluster information. In: Barros, L.N., Finger, M., Pozo, A.T., Gimenénez-Lugo, G.A., Castilho, M. (eds.) SBIA 2012. LNCS (LNAI), vol. 7589, pp. 92–101. Springer, Heidelberg (2012)
14. Valverde-Rebaza, J., de Andrade Lopes, A.: Structural Link Prediction Using Community Information on Twitter. In: CASoN 2012, pp. 132–137. IEEE (2012)
15. Valverde-Rebaza, J., de Andrade Lopes, A.: Exploiting behaviors of communities of Twitter users for link prediction. Social Network Analysis and Mining 3(4), 1063–1074 (2013)
16. Yin, D., Hong, L., Davison, B.D.: Structural link analysis and prediction in microblogs. In: CIKM 2011, pp. 1163–1168 (2011)
17. Zhang, Q.-M., Lü, L., Wang, W.-Q., Zhu, Y.-X., Zhou, T.: Potential theory for directed networks. PLoS ONE 8(2), e55437 (2013)
18. Zheleva, E., Getoor, L., Golbeck, J., Kuter, U.: Using friendship ties and family circles for link prediction. In: Giles, L., Smith, M., Yen, J., Zhang, H. (eds.) SNAKDD 2008. LNCS, vol. 5498, pp. 97–113. Springer, Heidelberg (2010)