Soc. Netw. Anal. Min. (2013) 3:1063–1074
DOI 10.1007/s13278-013-0142-8

ORIGINAL ARTICLE

Exploiting behaviors of communities of twitter users for link prediction

Jorge Valverde-Rebaza · Alneu de Andrade Lopes

Received: 5 May 2013 / Revised: 26 August 2013 / Accepted: 3 October 2013 / Published online: 22 October 2013

Abstract Currently, online social networks and social media have become increasingly popular showing an exponential growth. This fact have attracted increasing research interest and, in turn, facilitating the emergence of new interdisciplinary research directions, such as social network analysis. In this scenario, link prediction is one of the most important tasks since it deals with the problem of the existence of a future relation among members in a social network. Previous techniques for link prediction were based on structural (or topological) information. Nevertheless, structural information is not enough to achieve a good performance in the link prediction task on large-scale social networks. Thus, the use of additional information, such as interests or behaviors that nodes have into their communities, may improve the link prediction performance. In this paper, we analyze the viability of using a set of simple and non-expensive techniques that combine structural with community information for predicting the existence of future links in a large-scale online social network, such as Twitter. Twitter, a microblogging service, has emerged as a useful source of informative data shared by millions of users whose relationships require no reciprocation. Twitter network was chosen because it is not well understood, mainly due to the occurrence of directed and asymmetric links yet. Experiments show that our proposals can be used efficiently to improve unsupervised and supervised link prediction task in a directed and asymmetric large-scale network.

Keywords Link prediction · Community detection · Social networks · Link analysis · Social networks analysis · Social network behavior analysis · Microblogging · Social influence · Twitter

1 Introduction

In general, the use of online social networks and social media has increased in recent years (Lunden 2012; Constine 2012). Social networks offer to their users the possibility of meeting and networking individuals with similar personal and business interests. Online social networking services such as Facebook and Twitter have become part of the daily life of millions of people around the world who maintain and create new social relationships (Fire et al. 2011; Hopcroft et al. 2011). This fact implies the growth and quick changes over time in underlying structures (vertices and links) of the social networks (Liben-Nowell and Kleinberg 2007).

The boundless growth of online social networks has resulted in several research directions that examine the structural and other properties of large-scale social networks that aim to understand the basis of its social structure (Hopcroft et al. 2011). Understanding the formation of social relationships can give us insights on how an individual user influences her/his friends (Tang et al. 2009).

Detection of hidden social relationships is a friendship suggestion mechanism used by some online social networks. 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 social

J. Valverde-Rebaza (✉) · A. de Andrade Lopes
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, Campus de São Carlos, São Carlos,
SP 13560-970, Brazil
e-mail: jvalverr@icmc.usp.br; jorge.carlos14@gmail.com

A. de Andrade Lopes
e-mail: alneu@icmc.usp.br

network evolution (Fire et al. 2011; Liben-Nowell and Kleinberg 2007). The problem of predicting the existence of missing relationships or new ones is usually referred to as the link prediction problem (Liben-Nowell and Kleinberg 2007; Lü and Zhou 2011). Link prediction has many applications outside the domain of social networks, it is also used in bioinformatics to discover genetic or protein-protein interactions (Kotera et al. 2012), e-commerce to build recommendation systems (Benchettara et al. 2010; Esslimani et al. 2011), security domain to assist identifying groups of terrorist or criminals (Hasan et al. 2006), collaborative domains to assess the influence of a particular individual (Wei et al. 2013; Perez-Cervantes et al. 2013), information retrieval and information extraction domain to predict words, topics or documents in very large collections of documents (Itakura et al. 2011), etc.

Since the link prediction problem is relevant for different domains, several techniques have been proposed to cope with it. Thus, the link prediction problem can be addressed in two different strategies: unsupervised and supervised (Liben-Nowell and Kleinberg 2007; Lü and Zhou 2011; Valverde-Rebaza and de Andrade Lopes 2012; Hasan et al. 2006; Lichtenwalter et al. 2010; Davis et al. 2013). Furthermore, the strategy used for a link prediction measure indicates how its performance will be evaluated.

Unsupervised methods assign a score for each pair of nodes with base on node neighborhoods (local structural information) or path information (global structural information). The state-of-the-art unsupervised link prediction methods based on local information, such as common neighbors, Jaccard coefficient, adamic adar, resource allocation and preferential attachment, as well as based on global information, such as Katz, Rooted PageRank and SimRank are compared in (Liben-Nowell and Kleinberg 2007; Lü and Zhou 2011). According to these experimental results on real networks, methods based on global information can provide higher accuracy than measures based on local information, but their computation is 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 (Hasan et al. 2006; Benchettara et al. 2010). Thus, network information such as the structural ones and nodes' attributes are used to build a feature vector for each pair of nodes. Feature vectors are used in different classifiers, such as support vector machines (Vapnik 1995; Calderon-Niquin and Valverde-Rebaza 2012), artificial neural networks (Haykin 1998), decision trees (Quinlan 1993), etc., to assign a label to determine if a pair of nodes (a link) there exist or not.

Most of unsupervised and supervised proposals have focused on exploiting the local or global structural information of networks. However, other information, such as the behavior of users in social communities, are not used properly yet. Thus, with the aim of improving the accuracy of link prediction, Zheleva et al. (Zheleva et al. 2008), Soundarajan and Hopcroft (2012), Hoseini et al. (2012) and Valverde-Rebaza and Lopes (2012) proposed hybrid methods using local structural information and community information. In these proposals, the rationale for using community information in link prediction task is the very definition of community structure, i.e., the existence of high concentration of links within particular groups of nodes, as well as the low concentration of links between these groups. This property of community structure may convey significant information about nodes playing similar behaviors in the network (Fortunato 2010). Besides, in different experiments, Feng et al. (2012) found that accuracy of link prediction proposals based only on structural information drastically improves when community structure of networks grows.

One of the issues to be considered is that, most of the proposals for link prediction task only consider undirected networks and experiments in general do not employ large-scale networks. So, in this paper, we focus on microblogging network, which is a particular type of social service. In microblog services, such as Twitter, participants build an explicit social network by "following" (subscribing to) another user and thus they automatically receive the (short) messages generated by the target user. Different from some online social networks such as Facebook, a followed user has the option, but not the obligation to similarly follow back (Kwak et al. 2010; Yin et al. 2011; Hopcroft et al. 2011; Valverde-Rebaza and de Andrade Lopes 2012; Boutet et al. 2013).

Twitter network has been widely analyzed and several papers have been published dealing with its particularities. Kwak et al. (2010) found that it is a very asymmetric network. Romero and Kleinberg (2010) introduced the hybrid network concept and explored the directed closure process in the Twitter network. Golder et al. (2010) discussed several principles for link prediction in Twitter network, such as shared interests, shared followers and mutuality. Dawei et al. (Yin et al. 2011) and Zhang et al. (2012) explored different concepts of structural information for link prediction in Twitter. Valverde-Rebaza and Lopes (2012) show that community information can improve link prediction task in Twitter.

In this paper, we analyze links and community structures in Twitter to deal with the link prediction task. Our contributions are twofold: (1) We use the measures proposed by Valverde-Rebaza and Lopes in (2012) but performing a more extensive analysis than the performed in (Valverde-Rebaza and de Andrade Lopes 2012), i.e., using link prediction measures in unsupervised and supervised strategies. 2) We analyze the importance of community detection in Twitter and how it improves the link prediction accuracy. Moreover, we compare experimentally the most popular link prediction methods based on local structural information with hybrid measures based on community information.

The remainder of this paper is organized as follows. In Sect. 2, we present the link prediction problem and the usual measures for performance evaluation. In Sect. 3, we present some basic link prediction measures based on local structural information and their extensions for directed networks. In Sect. 4 we present the extensions for directed networks of the measures proposed in (Valverde-Rebaza and de Andrade Lopes 2012) as well as the community detection algorithm that it requires for its efficient use in large-scale networks. In Sect. 5 we present experimental results obtained from Twitter using unsupervised and supervised link prediction strategies. Finally, in Sect. 6 we summarize the main findings and conclusions of this work.

2 Problem description

Given a directed network $G(V, E)$, where V and E are sets of nodes and links, respectively. Multiple links and self connections are not allowed. Consider the universal set, denoted by U, containing all $|V|(|V| - 1)$ potential directed links between every pair of vertices in V, where $|V|$ denotes the number of elements in V. Thus, the fundamental task of a link prediction method is to find out the missing links (future links) in the set $U - E$ (set of non-existent links) assigning a score for each link in this set. The higher the score the higher the connection probability is and vice versa (Lü and Zhou 2011; Zhang et al. 2012).

So, given a predictor, we can rank all the non-existent links according to their scores. To test the accuracy of the predictor, the set E is divided into two parts: the training set E^T is treated as known information, while the testing set (probe set) E^{P} is used for testing and no information therein is allowed to be used for prediction. Clearly, $E = E^{\mathrm{T}} \cup E^{\mathrm{P}}$ and $E^{\mathrm{T}} \cap E^{\mathrm{P}} = \varnothing$.

When the unsupervised strategy is used, two standard measures are applied to quantify the prediction accuracy (Zhou et al. 2009; Lü and Zhou 2011; Liu et al. 2011; Valverde-Rebaza and de Andrade Lopes 2012; Valverde-Rebaza and de Andrade Lopes 2012; Zhang et al. 2012): AUC (area under the receiver operating characteristic curve) (Hanley and McNeil 1982) and precision (Herlocker et al. 2004). The AUC evaluates the predictor performance according to the list of scores of all non-observed links, $U - E^{\mathrm{T}}$, and can be interpreted as the probability that for a randomly chosen missing link (a link in E^{P}) is given a higher score than for a randomly chosen non-existent link (a link in $U - E$). Let n be the number of independent comparisons, if n' times for the missing links are given higher scores than non-existent links while n'' times for both missing and non-existent links are given equal scores, AUC value is

$$\mathrm{AUC} = \frac{n' + 0.5n''}{n} \quad (1)$$

The AUC is approximately 0.5 when all the scores are generated from an independent and identical distribution. Therefore, the degree to which the value exceeds 0.5 indicates as to how better than pure chance the algorithm performs.

Different from AUC, precision only focuses on the L links with highest scores. It is defined as the ratio between relevant items (L_{r}) selected and the items selected (L). Let the top-L links, if L_{r} links are accurately predicted (i.e., there are L_{r} links in the testing set), then the precision is

$$\mathrm{Precision_U} = \frac{L_{\mathrm{r}}}{L} \quad (2)$$

Clearly, higher precision means higher prediction accuracy.

When the supervised strategy is used, different standard measures from supervised machine learning, such as accuracy, precision recall, F-measure, etc. are applied to quantify the classifiers performance (Fatourechi et al. 2008). These measures are defined as follows:

$$\mathrm{Accuracy} = \frac{|\mathrm{TP}| + |\mathrm{TN}|}{|\mathrm{TP}| + |\mathrm{TN}| + |\mathrm{FP}| + |\mathrm{FN}|} \quad (3)$$

$$\mathrm{Precision_S} = \frac{|\mathrm{TP}|}{|\mathrm{TP}| + |\mathrm{FP}|} \quad (4)$$

$$\mathrm{Recall} = \frac{|\mathrm{TP}|}{|\mathrm{TP}| + |\mathrm{FN}|} \quad (5)$$

$$F - \mathrm{value} = \frac{2 \times \mathrm{Precision_S} \times \mathrm{Recall}}{\mathrm{Precision_S} + \mathrm{Recall}} \quad (6)$$

where |TP|, |TN|, |FP| and |FN| represent true positives, true negatives, false positives and false negatives rates, respectively.

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

3 Structural link prediction measures

Structural (topological) measures use the similarity between nodes since similar nodes likely share same relations (links).

Link prediction measures based on similarity can be classified into different ways, such as the measures based on local or global structural information. Measures based on global information use all available information on the network, and thus, may lead to higher accuracy. However, the global measures computation is very time-consuming and usually infeasible for large-scale networks. On the other hand, local measures are generally faster, but provide lower accuracy compared to the global ones (Liben-Nowell and Kleinberg 2007; Lü and Zhou 2011).

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. In directed networks (e.g., Twitter) 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, $\Gamma_{\text{out}}(x) = \{y \mid (x,y) \in E\}$ is defined as outgoing neighborhood and $\Gamma_{\text{in}}(x) = \{y \mid (y,x) \in E\}$ is defined as incoming neighborhood (Fire et al. 2011; Zhang et al. 2012).

Besides these neighborhood definitions, we use the following local similarity measures:

1. Common neighbors (CN): This measure refers to the size of the set of all common friends, $\Lambda_{x,y}$, of both x and y. For an undirected network is according to Eq. 7.

$$s^{\text{CN}}_{x,y} = |\Lambda_{x,y}| = |\Gamma(x) \cap \Gamma(y)| \tag{7}$$

For a directed network, CN measure is defined based on the link direction. CN measure for incoming neighborhood is according to Eq. 8.

$$s^{\text{CN}_{\text{in}}}_{x,y} = \left|\Lambda^{\text{in}}_{x,y}\right| = |\Gamma_{\text{in}}(x) \cap \Gamma_{\text{in}}(y)| \tag{8}$$

N measure for outgoing neighborhood is according to Eq. 9.

$$s^{\text{CN}_{\text{out}}}_{x,y} = \left|\Lambda^{\text{out}}_{x,y}\right| = |\Gamma_{\text{out}}(x) \cap \Gamma_{\text{out}}(y)| \tag{9}$$

2. Adamic adar (AA) This measure refines the simple counting of common neighbors by assigning more weight to the less-connected neighbors, as defined in Eq. 10 for an undirected network.

$$s^{\text{AA}}_{x,y} = \sum_{z \in \Lambda_{x,y}} \frac{1}{\log|\Gamma(z)|} \tag{10}$$

For a directed network, AA measure for incoming neighborhood is according to Eq. 11.

$$s^{\text{AA}_{\text{in}}}_{x,y} = \sum_{z \in \Lambda^{\text{in}}_{x,y}} \frac{1}{\log|\Gamma_{\text{in}}(z)|} \tag{11}$$

AA measure for outgoing neighborhood is according to Eq. 12.

$$s^{\text{AA}_{\text{out}}}_{x,y} = \sum_{z \in \Lambda^{\text{out}}_{x,y}} \frac{1}{\log|\Gamma_{\text{out}}(z)|} \tag{12}$$

3. Jaccard coefficient (Jac) This measure indicates whether two users of a network have a significant number of common neighbors regarding their total neighbors set size. For an undirected network it is defined according to Eq. 13.

$$s^{\text{Jac}}_{x,y} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{13}$$

For a directed network, Jac measure for incoming neighborhood is according to Eq. 14.

$$s^{\text{Jac}_{\text{in}}}_{x,y} = \frac{|\Gamma_{\text{in}}(x) \cap \Gamma_{\text{in}}(y)|}{|\Gamma_{\text{in}}(x) \cup \Gamma_{\text{in}}(y)|} \tag{14}$$

Jac measure for outgoing neighborhood is according to Eq. 15.

$$s^{\text{Jac}_{\text{out}}}_{x,y} = \frac{|\Gamma_{\text{out}}(x) \cap \Gamma_{\text{out}}(y)|}{|\Gamma_{\text{out}}(x) \cup \Gamma_{\text{out}}(y)|} \tag{15}$$

4. Resource allocation (RA) This measure punishes the high-degree common neighbors more heavily than AA. The formal definition for an undirected network is according to Eq. 16.

$$s^{\text{RA}}_{x,y} = \sum_{z \in \Lambda_{x,y}} \frac{1}{|\Gamma(z)|} \tag{16}$$

For a directed network, RA measure for incoming neighborhood is according to Eq. 17.

$$s^{\text{RA}_{\text{in}}}_{x,y} = \sum_{z \in \Lambda^{\text{in}}_{x,y}} \frac{1}{|\Gamma_{\text{in}}(z)|} \tag{17}$$

RA measure for outgoing neighborhood is according to Eq. 18.

$$s^{\text{RA}_{\text{out}}}_{x,y} = \sum_{z \in \Lambda^{\text{out}}_{x,y}} \frac{1}{|\Gamma_{\text{out}}(z)|} \tag{18}$$

5. Preferential attachment (PA) This index is proportional to the number of neighbors of each vertex. The formal definition for an undirected network is according to Eq. 19.

$$s^{\text{PA}}_{x,y} = |\Gamma(x)| \times |\Gamma(y)| \tag{19}$$

For a directed network, PA measure for incoming neighborhood is according to Eq. 20.

$$s^{\text{PA}_{\text{in}}}_{x,y} = |\Gamma_{\text{in}}(x)| \times |\Gamma_{\text{in}}(y)| \tag{20}$$

PA measure for outgoing neighborhood is according to Eq. 21.

$$s_{x,y}^{\mathrm{PA_{out}}} = |\Gamma_{\mathrm{out}}(x)| \times |\Gamma_{\mathrm{out}}(y)| \tag{21}$$

In (Liben-Nowell and Kleinberg 2007; Lü and Zhou 2011; Fire et al. 2011) various local measures were systematically compared on many real networks. According to these extensive experimental results, the RA index achieves the best performance, while AA and CN indexes have the second best overall performance. On the other hand, PA has the worst overall performance.

4 Link prediction in large-scale networks using community membership information

Valverde-Rebaza and Lopes (2012) proposed two new measures for link prediction: (1) the within and inter community (WIC) measure, and (2) the set of Within-community common neighbors measures, also called as the W form measures. To use these approaches, a community detection algorithm must be previously applied on the network.

In a network G there are $M > 1$ communities represented by the labels $C_\alpha, C_\beta, \ldots, C_M$. When a node $x \in V$ belongs to a community with label C, this vertex is represented as x^C. Consider that each vertex belongs to a unique community.

Figure 1 shows a directed graph with its communities identified by a given community detection algorithm. This graph has 12 vertices and 19 links between them. Each node is labeled with its index, for easy identification. Thus, the vertices from 1 to 4 belong to the community with label C_α or color gray community. Vertices from 5 to 9 belong to the community with label C_β or color black community. Vertices from 10 to 12 belong to the community with label C_γ or color white community.

Considering that community detection process in large-scale social networks can be very expensive and in some cases intractable, the use of the WIC and W form measures should not be recommended. Thus, in this paper, we use the label propagation algorithm (LPA) (Raghavan et al. 2007) as a specific algorithm for community detection in large-scale networks. LPA, which has near-linear time complexity, enables the use of the WIC and W form measures for large-scale networks without incurring excessive computational complexity. Furthermore, according to (Valverde-Rebaza and de Andrade Lopes 2012), we modify the W form measures for considering the incoming and outgoing links of directed and asymmetric networks.

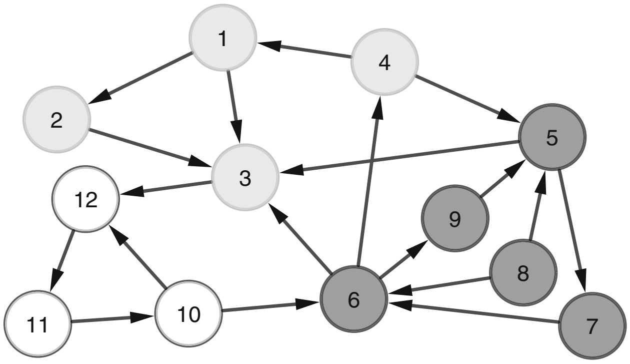

Fig. 1 Directed network in which three communities are distinguished by different gray scale. Nodes from the same color belong to the same community

4.1 Label propagation algorithm

In real networks, vertices with low degrees coexist with some vertices with large degrees. Furthermore, the distribution of links is not only globally heterogeneous but also locally heterogeneous, with high concentrations of edges within certain groups of vertices and low concentrations among these groups. This characteristic of real networks is named community structure (Fortunato 2010).

Several algorithms have been proposed to find community structures in networks (Clauset et al. 2004; Newman and Girvan 2004; Newman 2004; Pons and Latapy 2006; Almeida and de Andrade Lopes 2009). Community detection algorithms aim to find out groups with an inherent or an externally specified notion of similarity among nodes within groups. Furthermore, the number of communities in a network and their sizes are not known beforehand and they are established by a community detection algorithm (Fortunato 2010).

The LPA (Raghavan et al. 2007) is a simple and fast method for community detection. Initially, it assigns a unique label to each vertex (e.g., its vertex label). At each iteration, a pass over all vertices, in a random order, is performed: each vertex takes the label shared by the majority of its neighbors. If there is no unique majority, one of the majority labels is select at random. In this way, labels are propagated across the graph. The process converges when each vertex has the majority label of its neighborhood, or a maximum number of iteration is achieved. Communities are defined as groups of vertices having identical labels at convergence. After the process, each vertex has more neighbors in its community than in any other community.

This community detection algorithm does not deliver a unique solution. Due to the many relationships encountered along the detection process it is possible to derive different partitions starting from the same initial condition. Nonetheless, tests on real networks show that all partition found are similar to each other (Fortunato 2010).

The main advantage of this simplified algorithm for identifying communities is the fact that it does not need any previous information on the number and the size of the communities. The time complexity of each iteration of the algorithm is $O(m)$, where $m = |E|$. The number of iterations to convergence is independent of the graph size and can be defined as a parameter of LPA.

Furthermore, different proposals added some improvements to LPA in terms of optimization process (Barber and Clark 2009), introducing a score of labels process (Leung et al. 2009), etc. However, in this paper we use the basic LPA (Raghavan et al. 2007).

4.2 WIC measure

Experiments show that for a network with low community structure, link prediction measures based on structural similarity perform poorly (Zhou et al. 2009). Nonetheless, as the community structure of the network grows, the accuracy of these measures drastically improves (Feng et al. 2012). With this consideration and to exploit the efficiency of link prediction measures based on local information, the WIC measure (Valverde-Rebaza and de Andrade Lopes 2012) is used in large-scale networks.

The WIC is based on Bayesian theorem, i.e., given an undirected network G with M communities detected previously, the posterior probabilities that the same or different community labels are assigned to a pair of nodes (x, y), given its set of common neighbors $\Lambda_{x,y}$, are, respectively

$$P(x^{C_\alpha}, y^{C_\alpha} \mid \Lambda_{x,y}) = \frac{P(\Lambda_{x,y} \mid x^{C_\alpha}, y^{C_\alpha})P(x^{C_\alpha}, y^{C_\alpha})}{P(\Lambda_{x,y})} \tag{22}$$

$$P(x^{C_\alpha}, y^{C_\beta} \mid \Lambda_{x,y}) = \frac{P(\Lambda_{x,y} \mid x^{C_\alpha}, y^{C_\beta})P(x^{C_\alpha}, y^{C_\beta})}{P(\Lambda_{x,y})} \tag{23}$$

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

To estimate the probability of the set of all common neighbors of a pair of nodes $(x^{C_\alpha}, y^{C_\alpha})$ given these nodes belong to the same community with label C_α, consider the number of common neighbors with the same community label divided by the total number of common neighbors, as stated in Eq. 24.

$$P(\Lambda_{x,y} \mid x^{C_\alpha}, y^{C_\alpha}) = \frac{\left|\Lambda^W_{x,y}\right|}{\left|\Lambda_{x,y}\right|} \tag{24}$$

Similarly, to estimate the probability of the set of all common neighbors of a pair of nodes $(x^{C_\alpha}, y^{C_\beta})$ given these nodes belong to different communities with labels C_α and C_β, consider the number of common neighbors that may be associated with different labels C_α or C_β or with another community label C_γ divided by the total number of common neighbors, as defined by Equation 25.

$$P(\Lambda_{x,y} \mid x^{C_\alpha}, y^{C_\beta}) = \frac{\left|\Lambda^I_{x,y}\right|}{\left|\Lambda_{x,y}\right|} \tag{25}$$

To compare the likelihood of the link existence between pair of nodes (x, y) is defined the score $s_{x,y}$ as the ratio between Eqs. 22 and 23. Substituting Eqs. 24 and 25, we have:

$$s_{x,y} = \frac{\left|\Lambda^W_{x,y}\right|}{\left|\Lambda^I_{x,y}\right|} \times \frac{P(x^{C_\alpha}, y^{C_\alpha})}{P(x^{C_\alpha}, y^{C_\beta})} \tag{26}$$

Considering each node belong to a unique community and knowing that $C_\alpha \neq C_\beta$, we can assume that y^{C_γ}, and therefore $\frac{P(x^{C_\alpha}, y^{C_\alpha})}{P(x^{C_\alpha}, y^{C_\beta})} = \frac{P(x^{C_\alpha}, y^{C_\gamma})}{P(x^{C_\alpha}, y^{C_\gamma})}$ ratio can be neglected since this fraction value is 1. Thus, the final WIC measure for an undirected network is:

$$s^{\mathrm{WIC}}_{x,y} = \begin{cases} \left|\Lambda^W_{x,y}\right|, & \text{if} \quad \Lambda^W_{x,y} = \Lambda_{x,y} \\ \frac{\left|\Lambda^W_{x,y}\right|}{\left|\Lambda^I_{x,y}\right|}, & \text{otherwise} \end{cases} \tag{27}$$

It is important to notice that, in (Valverde-Rebaza and de Andrade Lopes 2012), WIC measure is extended for applying it in directed networks. Similarly to the indices showed in Sect. 3, the WIC measure can be defined based on the link direction considering the sets of incoming and outgoing within-community common neighbors: $\Lambda^{W_{\text{in}}}_{x,y} = \{z^C \in \Lambda^{\text{in}}_{x,y} \mid x^C, y^C\}$ and $\Lambda^{W_{\text{out}}}_{x,y} = \left\{z^C \in \Lambda^{\text{out}}_{x,y} \mid x^C, y^C\right\}$, and the sets of incoming and outgoing inter-community common neighbors: $\Lambda^{I_{\text{in}}}_{x,y} = \Lambda^{\text{in}}_{x,y} - \Lambda^{W_{\text{in}}}_{x,y}$ and $\Lambda^{I_{\text{out}}}_{x,y} = \Lambda^{\text{out}}_{x,y} - \Lambda^{W_{\text{out}}}_{x,y}$. These sets are used in Eq. 27 to obtain WIC measure for incoming neighborhood, as stated in Eq. 28.

$$s^{\mathrm{WIC}_{\text{in}}}_{x,y} = \begin{cases} \left|\Lambda^{W_{\text{in}}}_{x,y}\right|, & \text{if} \quad \Lambda^{W_{\text{in}}}_{x,y} = \Lambda_{x,y} \\ \frac{\left|\Lambda^{W_{\text{in}}}_{x,y}\right|}{\left|\Lambda^{I_{\text{in}}}_{x,y}\right|}, & \text{otherwise} \end{cases} \tag{28}$$

WIC index for outgoing neighborhood is computed according to Eq. 29.

$$s^{\mathrm{WIC}_{\text{out}}}_{x,y} = \begin{cases} \left|\Lambda^{W_{\text{out}}}_{x,y}\right|, & \text{if} \quad \Lambda^{W_{\text{out}}}_{x,y} = \Lambda_{x,y} \\ \frac{\left|\Lambda^{W_{\text{out}}}_{x,y}\right|}{\left|\Lambda^{I_{\text{out}}}_{x,y}\right|}, & \text{otherwise} \end{cases} \tag{29}$$

4.3 W form measures

Different measures based on local structural information are based on the set of all common neighbors, except PA index.

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 (Liu et al. 2011).

Thus, Valverde-Rebaza and Lopes (2012) consider that within-community common neighbors may contribute more to the connection likelihood than inter-community common neighbors because within-community common neighbors have similar behaviors. Based on that, the set of within-community common neighbors, $\Lambda_{x,y}^{W}$, is used instead the set of all common neighbors $\Lambda_{x,y}$ in the local structural measures, obtaining a set of new measures referred to as W form measures. Next, we present the respective W form for the local measures presented in Section III and their use mode for directed networks.

1. Common neighbors of W form (CN-W) Based on CN measure, its formal definition for an undirected network is according to Eq. 30.

$$s_{x,y}^{\mathrm{CN-W}} = \left|\Lambda_{x,y}^{W}\right| \tag{30}$$

For a directed network, CN-W index for incoming neighborhood is according to Eq. 31.

$$s_{x,y}^{\mathrm{CN-W_{in}}} = \left|\Lambda_{x,y}^{W_{\mathrm{in}}}\right| \tag{31}$$

CN-W index for outgoing neighborhood is according to Eq. 32.

$$s_{x,y}^{\mathrm{CN-W_{out}}} = \left|\Lambda_{x,y}^{W_{\mathrm{out}}}\right| \tag{32}$$

Notice that CN-W index is part of the formal definition of WIC index when the total set of common neighbors is the same of the set of within-community common neighbors.

2. Adamic adar of W form (AA-W) Based on AA, its formal definition for an undirected network is according to Eq. 33.

$$s_{x,y}^{\mathrm{AA-W}} = \sum_{z\in\Lambda_{x,y}^{W}} \frac{1}{\log|\Gamma(z)|} \tag{33}$$

AA-W for incoming neighborhood is according to Eq. 34.

$$s_{x,y}^{\mathrm{AA-W_{in}}} = \sum_{z\in\Lambda_{x,y}^{W_{\mathrm{in}}}} \frac{1}{\log|\Gamma_{\mathrm{in}}(z)|} \tag{34}$$

AA-W for outgoing neighborhood is according to Eq. 35.

$$s_{x,y}^{\mathrm{AA-W_{out}}} = \sum_{z\in\Lambda_{x,y}^{W_{\mathrm{out}}}} \frac{1}{\log|\Gamma_{\mathrm{out}}(z)|} \tag{35}$$

3. Jaccard coefficient of W form (Jac-W) Based on Jac measure, its formal definition for an undirected network is according to Eq. 36.

$$s_{x,y}^{\mathrm{Jac-W}} = \frac{\left|\Lambda_{x,y}^{W}\right|}{|\Gamma(x)\cup\Gamma(y)|} \tag{36}$$

Jac-W for incoming neighborhood is according to Eq. 37.

$$s_{x,y}^{\mathrm{Jac-W_{in}}} = \frac{\left|\Lambda_{x,y}^{W_{\mathrm{in}}}\right|}{|\Gamma_{\mathrm{in}}(x)\cup\Gamma_{\mathrm{in}}(y)|} \tag{37}$$

Jac-W for outgoing neighborhood is according to Eq. 38.

$$s_{x,y}^{\mathrm{Jac-W_{out}}} = \frac{\left|\Lambda_{x,y}^{W_{\mathrm{out}}}\right|}{|\Gamma_{\mathrm{out}}(x)\cup\Gamma_{\mathrm{out}}(y)|} \tag{38}$$

4. Resource allocation of W form (RA-W) Based on RA measure, its formal definition for an undirected network is according to Eq. 39.

$$s_{x,y}^{\mathrm{RAW}} = \sum_{z\in\Lambda_{x,y}^{W}} \frac{1}{|\Gamma(z)|} \tag{39}$$

RA-W measure for incoming neighborhood is according to Eq. 40.

$$s_{x,y}^{\mathrm{RA-W_{in}}} = \sum_{z\in\Lambda_{x,y}^{W_{\mathrm{in}}}} \frac{1}{|\Gamma_{\mathrm{in}}(z)|} \tag{40}$$

RA-W measure for outgoing neighborhood is according to Eq. 41.

$$s_{x,y}^{\mathrm{RA-W_{out}}} = \sum_{z\in\Lambda_{x,y}^{W_{\mathrm{out}}}} \frac{1}{|\Gamma_{\mathrm{out}}(z)|} \tag{41}$$

5 Experiments

We consider a scenario where new links of the Twitter network must be predicted. In this network, the LPA for community detection is applied to assign a community label to each vertex. Next, using unsupervised and supervised strategies, we compare the performance of the WIC and W form measures to classical link prediction measures based on local information (CN, AA, Jac, RA and PA).

5.1 Twitter network

The Twitter network used in our experiments has follower information for 40 million users and 1.4 billion links collected in June 2009 by Kwak et al. (Kwak et al. 2010). Twitter network differs from other social networks by its directed relationship nature, i.e., a Twitter user is not

Table 1 Basic features of the two graphs built from Twitter network after 7th and 15th iterations of LPA algorithm

	Twitter 7it	Twitter 15it
$\lvert V \rvert$	24617334	24617333
$\lvert E \rvert$	363565896	363565892
M	3415051	2250964
Max cluster size	1392411	10121242
Ratio of total links per user	14.77	14.77

obligated to reciprocate followers by following them back. Thus, only 22.1% of the used Twitter links are reciprocal.

In our experiments, Twitter users with more than 900 followers have been removed from the Twitter network. On this Twitter sample was employed the LPA using map-reduce formalism with 55 node Hadoop cluster. The code used to generate the community detection results is available online by Bhat (2010). Two executions of LPA were performed, in one the convergence was stipulated when 7th iteration is achieved, and in the other execution when 15th iteration is achieved. Thus, two subgraphs were generated with vertices labeled accordingly to the communities obtained at 7th and 15th iterations, Twitter 7it and Twitter 15it, respectively. Basic features of these graphs are summarized in Table 1.

5.2 Experimental setup

For our experiments, we perform two phases: the network pre-processing and the link prediction process. In the network pre-processing, the set E is divided into the training set E^T and the testing set E^P. From the set E, for select the links for E^P, we take randomly one-third of the links formed by users whose number of followers is two times greater than the ratio of total links per user[1]. The remaining links, except those formed by users whose number of followers is less than one-third of the ratio of total links per user, constitute the training set E^T. This evaluation method is widely used in the link prediction literature (Yin et al. 2011; Zhang et al. 2012; Valverde-Rebaza and de Andrade Lopes 2012).

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 $s^{\mathrm{WIC}_{\mathrm{out}}}_{x,y} > s^{\mathrm{WIC}_{\mathrm{in}}}_{x,y}$ then $s^{\mathrm{WIC}}_{x,y} = s^{\mathrm{WIC}_{\mathrm{out}}}_{x,y}$, otherwise, $s^{\mathrm{WIC}}_{x,y} = s^{\mathrm{WIC}_{\mathrm{in}}}_{x,y}$.

In supervised strategy, we use decision tree (J48), naive Bayes (NB), support vector machine (SMO) and multilayer perceptron with backpropagation (MLP) classifiers from Weka 3.6 (Weka 3: data mining software in java. The University of Waikato 3.6 (2013). Previously, we compute a total of 6000002 feature vectors from E^T considering that each node generates a feature vector with just 30 % of its neighboring nodes randomly. Thus, four different data sets in ARFF format (Weka 3: data mining software in java. The University of Waikato (2013) were created. Each data set is formed by feature vector that 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 WIC, CN-W, AA-W, Jac-W and RA-W. VTop is the data set whose feature vectors are formed by the five best link prediction measures (see Sects. 5, 5.1, 5.2, 5.3), i.e., WIC, CN-W, AA-W, RA-W and RA. VTotal is the data set whose feature vectors are formed by all local and based on community information measures, i.e., CN, AA, Jac, RA, PA, WIC, CN-W, AA-W, Jac-W and RA-W. Each data set has 50% of instances (links) with positive class and 50 % of instances with negative class.

Our experiments were performed on a computer with 99 GB of RAM and using Linux as operating system. The classification task was performed in Java, and the rest of the supervised link prediction process and both network pre-processing and the unsupervised link prediction process were programmed in C++.

Figure 2 shows the execution time, in seconds, used by each link prediction measure in the unsupervised strategy. Jac-W and Jac are the most time-consuming measures. Also, all W form measures need more time than their corresponding basic form due to W form measures have to process the community information of each node pair analyzed. However, the time spent by W form measures is not excessively higher than their respective basic forms. The fastest measure is PA due to it only makes the product of the neighbors of nodes pairs analyzed. The next fastest measures are WIC and CN but with a significant difference with respect to PA.

5.3 Validating results and analysis

To validate our results, we use the evaluation measures presented in Sect. 2 both for unsupervised strategy and supervised strategy. For results of our unsupervised link prediction process, we employ AUC and Precision to validate the quality of each link prediction measure. Table 1 summarizes the prediction results measured by AUC on

[1] Ratio of total links per user is the ratio between the total number of links, |*E*|, and the total number of nodes, |*V*|. This ratio indicates the average of the size of the neighborhood for each node.

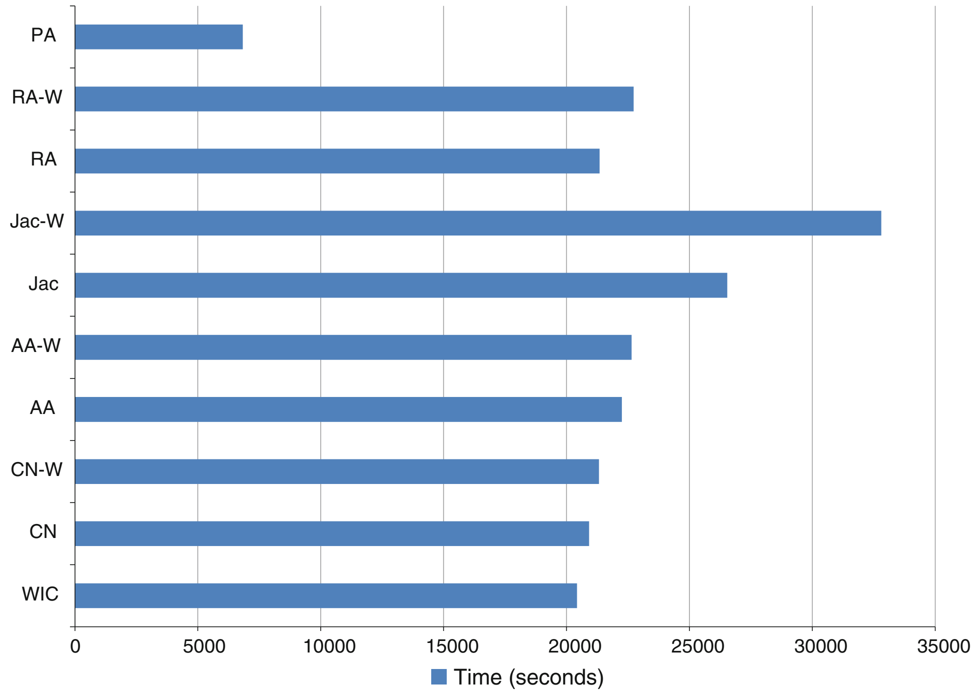

Fig. 2 Execution time, in seconds, of all link prediction measure used in our experiments in the unsupervised strategy

Twitter 7it and Twitter 15it subgraphs. Each AUC value is obtained by averaging over 10 implementations with 5 independently divisions of training and testing sets.

Looking at the results of Table 2, one should notice that the AUC performance of each link prediction measure is the same for both subgraphs. In the case of WIC and W form measures, this indicates that although both subgraphs have different number of groups (M), relations, interests or behaviors between nodes remain similar or equivalent, i.e., most of the users of Twitter network sharing similar interests or having similar behaviors are grouped in the same communities. That in turn, as commented in Sect. 4, 4.1, is justified by the fact that results from different executions performed by LPA may produce similar partitions. For the others measures, the same AUC performance to both subgraphs is justified by its similar structure with the similar number of nodes ($|V|$) and links ($|E|$).

Comparing AUC performance for all link prediction measures, WIC outperforms all of them. RA-W, RA, CN-W and AA-W are the next best measures, in that order. In addition, all W form measures outperform, with significant difference, their corresponding basic forms. Also, Jac has the worst performance and do not outperform the assignment by chance.

Figure 3 shows the prediction quality measured by precision on Twitter 7it and Twitter 15it. Similar to AUC, precision performance of each link prediction measure evaluated is the same in both Twitter 7it and Twitter 15it subgraphs. Different values of L were used. In the top-100 links, Jac-W obtains 0.52 and outperforms all other indexes. In the top-500, AA-W and RA-W obtain 0.664 and 0.652 precision values, respectively, and outperform all other measures. In the top-1,000, RA-W, AA-W and Jac obtain 0.636, 0.627 and 0.622, respectively, and outperform all other measures. In the top-5,000, WIC obtains 0.51 precision value and outperforms all other measures. In the top-10,000, WIC, CN-W and AA-W obtain 0.47, 0.46 and 0.457 precision values, respectively, and outperform all other measures.

Generally, AA-W and RA-W obtained the best precision performance. Also, it was observed that the W form measures obtained the best precision performance that their respective basic forms, except Jac-W, which performs poorly than Jac. Also, an interesting phenomenon is observed in Jac, which has a peak when $L = 1{,}000$, but this performance decreases sharply when $L = 5{,}000$. PA has the worst precision for all values of L.

For results of our supervised link prediction process, Accuracy and F-value are employed to validate the quality of the classifiers in VLocal, VGroup, VTop and VTotal data sets. Tables 3 and 4, respectively, show accuracy and F-Value average values for four different classifiers after 10-fold cross validation. For Table 3, values in parenthesis indicate the mean absolute error. For both Tables 3 and 4, values emphasized in black correspond to the highest result among the evaluated data sets for each classifier. Entries highlighted in gray indicate that a classifier get best results in data sets formed by feature vectors using measures based on community information that VLocal data set, which is formed just by feature vectors using local measures.

Table 2 Link prediction results measured by AUC on two subgraphs of Twitter network. The emphasized values correspond to the highest results among the evaluated measures

Graph	WIC	CN	CN-W	AA	AA-W	Jac	Jac-W	RA	RA-W	PA
Twitter 7it	0.62	0.56	0.59	0.53	0.58	0.45	0.56	0.6	0.61	0.51
Twitter 15it	0.62	0.56	0.59	0.53	0.58	0.45	0.56	0.6	0.61	0.51

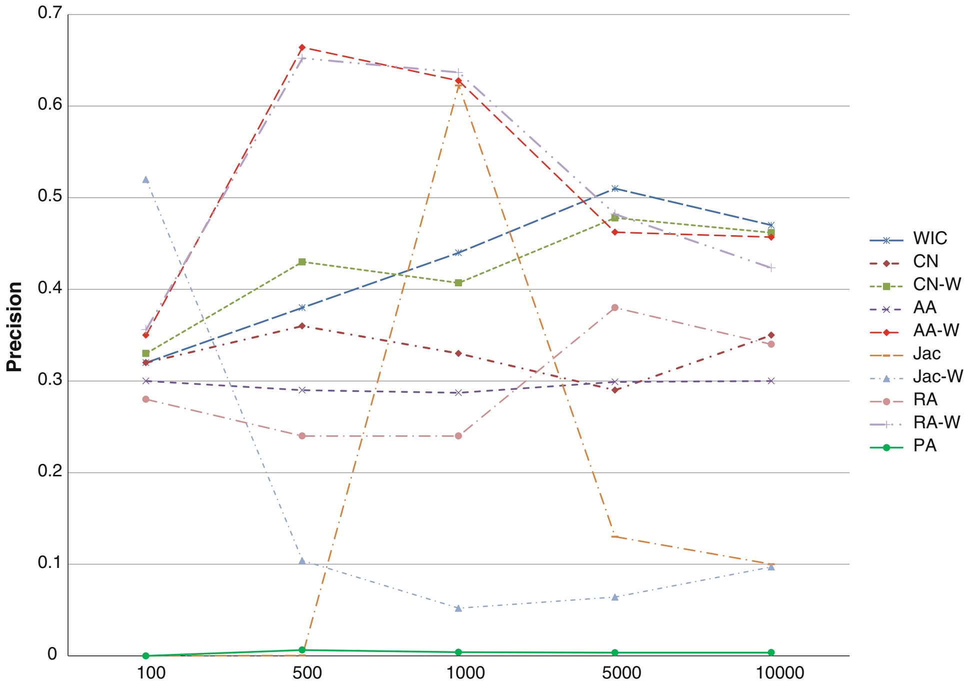

Fig. 3 Precision results on the two graphs from Twitter network. Different values of *L* are used to select the top-*L* highest scores for predicting links

Table 3 Accuracy results (in percent) on four Twitter data sets formed by feature vectors that combine different link prediction measures

Data set	J48	NB	SMO	MLP
VLocal	83.74 (0.24)	71.63 (0.28)	81.35 (0.19)	82.73 (0.25)
VGroup	82.86 (0.25)	71.70 (0.28)	80.00 (0.20)	81.88 (0.26)
VTop	83.08 (0.25)	72.12 (0.28)	80.34 (0.20)	82.01 (0.26)
VTotal	83.80 (0.24)	72.05 (0.28)	81.71 (0.18)	82.84 (0.24)

Values in parenthesis indicate the mean absolute error

Results of Table 3 show for J48, SMO and MLP classifiers the best Accuracy results are obtained in VTotal data set, i.e., the data set that combines all local measures and all measures based on community information. For NB classifier the best result is in VTop data set, i.e., the data set that uses the five best link prediction measures according to the results of Table 2. Results of Table 4 show for J48, NB, SMO and MLP classifiers the best F-measure results also are obtained in VTotal data set. Furthermore, for NB classifier the best F-measure result also is obtained in VTop data set.

Table 4 F-measure results on four Twitter data sets formed by feature vectors that combine different link prediction measures

Data set	J48	NB	SMO	MLP
VLocal	0.837	0.698	0.812	0.827
VGroup	0.829	0.699	0.798	0.819
VTop	0.831	0.703	0.801	0.820
VTotal	0.838	0.703	0.816	0.828

From entries highlighted in gray of Tables 3 and 4 we observe that classifiers perform better in data sets formed by feature vectors that include link prediction measures based on community information. This happens mainly when all local measures and all measures based on community information are combined into a feature vector, i.e., in VTotal data set.

6 Conclusion

We use different link prediction measures on two different directed graphs built from a Twitter network. For predicting a link between a pair of nodes, the WIC measure takes into account the information on which communities the pair of nodes (and their neighbors) belong to, i.e., if the nodes are in a same community or in different ones. The W form measures consider the neighborhood information of nodes belonging only to the same community of the pair of nodes analyzed.

Since for applying the WIC and W form measures is need, in a previous phase, partitioning the network into communities. Here we use a fast community detection algorithm, the LPA, to be able of applying the WIC and W form measures for large-scale networks.

When an unsupervised strategy is performed, the performance of WIC compared to other measures was better under the AUC criterion. Also, all the W form measures outperform their respective basic counterpart and also are positioned among the measures with better performance. Notice that, in the analysis by Valverde-Rebaza and Lopes (2012) on complex networks of different domains, PA obtained the worst performance, but, in the case of Twitter network analyzed in this work, Jac had the worst performance.

When analyzing precision, highlight RA-W, AA-W and WIC, which outperform the others. These three measures are characterized by uniform performance in the different L values, especially WIC, which has a sustained growth up to $L = 5{,}000$. Here, PA obtained the worst performance in all L values.

When a supervised strategy is performed, our results show that combining measures based on local information and based on community information will improve 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.

Since our analysis was performed on a online large-scale social network, such as Twitter, consider the time execution is important. Thus, for our unsupervised link prediction process, PA is the fastest measure, but has the penultimate position in AUC analysis and the ultimate position in precision analysis. WIC is the second fastest and has the best performance in AUC analysis besides being one of the first in the precision analysis. Besides, W form measures are slightly slower than their respective basic forms, but outperform them in AUC and precision analyses.

In summary, our experiments suggest that WIC and W form measures capture information from the behaviors of users into the communities which they belong, improving the link prediction performance for directed and asymmetric large-scale networks. This happens because nodes of the same communities likely have similar interests or behaviors. In the case of the Twitter network, the similar interests or behaviors between users may be a preference for following other users with the same topics of interest, following the same celebrities, or dissemination of tweets containing certain type of hashtags, among others.

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

References

Almeida LJ, de Andrade Lopes A (2009) An ultra-fast modularity-based graph clustering algorithm, Aveiro, Portugal 1–9

Barber MJ, Clark JW (2009) Detecting network communities by propagating labels under constraints. Phys Rev E Stat Phys 80(2): 026129

Benchettara N, Kanawati R, Rouveirol C (2010) A supervised machine learning link prediction approach for academic collaboration recommendation. In: Proceedings of RecSys Vol 10, pp 253–256

Bhat AU (2010) Twitter community detection. Community detection for Twitter follower network. Available: https://github.com/AKSHAYUBHAT/TwitterCommunityDe-tection

Boutet A, Kim H, Yoneki E (2013) Whats in Twitter, i know what parties are popular and who you are supporting now!. Soc Netw Anal Min

Calderon-Niquin MA, Valverde-Rebaza J (2012) Multiple kernel learning based on local and nonlinear combinations. In: Informatica (CLEI), XXXVIII Conferencia Latinoamericana, pp 1 –7

Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70. doi:10.1103/PhysRevE.70.06611110.1103/PhysRevE.70.066111

Constine J (2012) How big Is Facebook's data? 2.5 billion pieces of content and 500+ terabytes ingested every day. Techcrunch. Available:http://techcrunch.com/2012/08/22/how-big-is-facebooks-data-2-5-billion-pieces-of-content-and-500-terabytes-ingested-every-day/

Davis D, Lichtenwalter R, Chawla R (2013) Supervised methods for multi-relational link prediction. Soc Netw Anal Min 3: 127–141

Esslimani I, Brun A, Boyer A (2011) Densifying a behavioral recommender system by social networks link prediction methods. Soc Netw Anal Min 1:159–172

Fatourechi M, Ward R, Mason S, Huggins J, Schlogl A, Birch G (2008) Comparison of evaluation metrics in classification applications with imbalanced datasets. In: Machine learning and applications. ICMLA '08. Seventh International Conference on, pp 777–782

Feng X, Zhao J, Xu K (2012) Link prediction in complex networks: a clustering perspective. Eur Phys J B 85(1): 3

Fire M, Tenenboim L, Lesser O, Puzis R, Rokach L, Elovici Y (2011) Link prediction in social networks using computationally efficient topological features. In: Privacy, security, risk and trust, 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SOCIALCOM), pp 73 –80

Fortunato S (2010) Community detection in graphs. CoRR abs/0906.0612v2

Golder SA, Yardi S (2010) Structural predictors of tie formation in twitter: transitivity and mutuality. In: Proceedings of SOCIALCOM '10, pp 88–95

Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1): 29–36

Hasan MA, Chaoji V, Salem S, Zaki M (2006) Link prediction using supervised learning. In: Proceedings of SDM 06 workshop on link analysis, counterterrorism and security

Haykin S (1998) Neural networks: a comprehensive foundation, 2nd ed. Prentice Hall PTR, Upper Saddle River

Herlocker JL, Konstan JA, Terveen LG, Riedl JT (2004) Evaluating collaborative filtering recommender systems. ACM Trans Inf Syst 22(1):5–53

Hopcroft J, Lou T, Tang J (2011) Who will follow you back?: reciprocal relationship prediction. In: Proceedings of CIKM ’11, pp 1137–1146

Hoseini E, SHashemi E, Hamzeh A (2012) Link prediction in social network using co-clustering based approach. In: Proceedings of the 2012 26th international conference on advanced information networking and applications workshops, ser. WAINA ’12. IEEE Computer Society, pp 795–800

Itakura KY, Clarke CLA, Geva S, Trotman A, Huang WC (2011) Topical and structural linkage in wikipedia. In: Proceedings of ECIR’11, pp 460–465

Kotera M, Yamanishi Y, Moriya Y, Kanehisa M, Goto S (2012) Genies: gene network inference engine based on supervised analysis. Nucleic Acids Res 40: 162–167

Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media? In: Proceedings of WWW ’10, pp 591–600

Leung I, Hui P, Lio P, Crowcroft J (2009) Towards real-time community detection in large networks. Phys Rev E 79(6): 066107

Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. JASIST 58(7): 1019–1031

Lichtenwalter RN, Lussier JT, Chawla NV (2010) New perspectives and methods in link prediction. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, ser. KDD’10. ACM, New York, pp 243–252

Liu Z, Zhang Q-M, Lü L, Zhou T (2011) Link prediction in complex networks: a local naïve bayes model. Europhys Lett 96(48007)

Lü L, Zhou T (2011) Link prediction in complex networks: a survey. Phys A Stat Mech Appl 390(6): 1150–1170

Lunden I (2012) Analyst: Twitter passed 500M users in June 2012, 140M of them in US; Jakarta ’biggest tweeting’ city. Techcrunch. Available: http://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/

Newman MEJ (2004) Fast algorithm for detecting community structure in networks. Phys Rev E 69(6): 066133

Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2): 026113

Pons P, Latapy M (2006) Computing communities in large networks using random walks. J Graph Algorithms Appl 10(2):191–218

Perez-Cervantes E, Mena-Chalco JP, de Oliveira MCF, Cesar-Jr RM (2013) Using link prediction to estimate the collaborative influence of researchers. In: IEEE 9th International Conference on e-Science 2013, Beijing, pp 1–8

Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc, San Francisco

Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E 76: 036106

Romero DM, Kleinberg JM (2010) The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In: ICWSM

Soundarajan S, Hopcroft J (2012) Using community information to improve the precision of link prediction methods. In: Proceedings of the 21st international conference companion on World Wide Web, ser. Proceedings of WWW ’12 Companion, pp 607–608

Tang J, Sun J, Wang C, Yang Z (2009) Social influence analysis in large-scale networks. In: Proceedings of KDD ’09, pp 807–816

Valverde-Rebaza J, de Andrade Lopes A (2012) Link prediction in complex networks based on cluster information. In: Advances in artificial intelligence, SBIA 2012, 21th Brazilian symposium on artificial intelligence, ser, Vol 7589. Lecture Notes in Computer Science, Springer 92–101

Valverde-Rebaza J, de Andrade Lopes A (2012) Structural link prediction using community information on twitter. In: Computational aspects of social networks (CASoN), 2012 Fourth International Conference on, Nov 2012, pp 132–137

Vapnik VN (1995) The nature of statistical learning theory. Springer-Verlag New York, Inc., New York

Wei D, Deng X, Zhang X, Deng Y, Mahadevan S (2013) Identifying influential nodes in weighted networks based on evidence theory. Phys A Stat Mech Appl 392(10): 2564–2575

Weka 3: Data mining software in java (2013) The University of Waikato (2013). Available: http://www.cs.waikato.ac.nz/ml/weka/

Yin D, Hong L, Davison BD (2011) Structural link analysis and prediction in microblogs. In: Proceedings of CIKM ’11, pp 1163–1168

Zhang Q-M, Lü L, Wang W-Q, Zhu Y-X, Zhou T (2012) Potential theory for directed networks. CoRR abs/1202.2709

Zheleva E, Getoor L, Golbeck J, Kuter U (2008) Using friendship ties and family circles for link prediction. In: Proceedings of the 2nd international conference on advances in social network mining and analysis, ser. SNAKDD’08, pp 97–113

Zhou T, Lü L, Zhang Y-C (2009) Predicting missing links via local information. Eur Phys J B 71(4): 623–630