# Structural Link Prediction Using Community Information on Twitter

Jorge Valverde-Rebaza
*Instituto de Ciências Matemáticas e de Computação*
*Universidade de São Paulo - Campus de São Carlos*
*13560-970 São Carlos, SP, Brazil*
*jvalverr@icmc.usp.br*

Alneu de Andrade Lopes
*Instituto de Ciências Matemáticas e de Computação*
*Universidade de São Paulo - Campus de São Carlos*
*13560-970 São Carlos, SP, Brazil*
*alneu@icmc.usp.br*

***Abstract*—Currently, social networks and social media have attracted increasing research interest. In this context, link prediction is one of the most important tasks since it can predict the existence or missing of a future relation between user members in a social network. In this paper, we describe experiments to analyze the viability of applying the *within and inter cluster* (WIC) measure for predicting the existence of a future link on a large-scale online social network. Compared with undirected social networks, directed social networks have received less attention and still are not well understood, mainly due to the occurrence of asymmetric links. The WIC measure combines the local structural similarity information and community information to improve link prediction accuracy. We compare the WIC measure with classical measures based on local structural similarities, using real data from Twitter, a directed and asymmetric large-scale online social network. Our experiments show that the WIC measure can be used efficiently on directed and asymmetric large-scale networks. Moreover, it outperforms all compared measures employed for link prediction.**

***Keywords*-link prediction; community detection; social networks; link analysis; microblogging; social influence; Twitter**

## I. Introduction

In general, the use of online social networks and social media has increased in recent years [1], [2]. 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 [3], [4]. This fact implies the growth and quick changes over time in underlying structure (vertices and links) of the social networks [5].

Detection of hidden relationships is a friendship suggestion mechanism used by some online social networks. In such cases, 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 network evolution [3], [5]. The problem of predicting the existence of missing relationships or new ones in social networks is usually referred to as the *Link Prediction* problem [3], [5], [6]. Link prediction has many applications outside the domain of social networks, it is also used in bioinformatics to discover genetic or protein-protein interactions [7], e-commerce to build recommendation systems [8], [9], security domain to assist identifying groups of terrorist or criminals [10], information retrieval and extraction domain to predict words, topics or documents in very large collections of documents [11], etc.

Since the link prediction problem is relevant for different domains, several techniques have been proposed to solve it. Most of them are usually based on structural features and supervised methods. Measures based on structural features to assess similarity between a pair of vertices can be used for link prediction. One important type of these measures is based on the local information of the network, such as common neighbors, Jaccard coefficient, Adamic Adar, resource allocation, preferential attachment and others [5], [6]. On other hand, supervised methods consider the Link Prediction problem as a classification problem in which different network features such as the structural ones are used [9], [10].

Most of research in social network analysis has focused on exploiting the users and ties information such as friendships. However, other information, such as cluster or community information, are used as features for improving prediction performance, since the high concentration of links within particular groups of vertices, as well as the low concentration of links between these groups, may convey significant information about the network topology [12]. In different experiments, Feng et al. [13] found that accuracy of link prediction measures based on structural information drastically improves when clustering structure of networks grows. Motivated by these results, various authors have proposed some kind of clustering information for link prediction in networks from different domains [14]–[16]

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 [4], [17], [18].

Twitter network has been widely analyzed and several

978-1-4673-4794-5/12/$31.00 ©2012 IEEE

papers have been published dealing with its particularities. Kwak et al. [17] found that it is a very asymmetric network. Romero and Kleinberg [19] introduced the hybrid network concept and explored the directed closure process in the network. Golder et al. [20] discussed several principles for link prediction, such as shared interests, shared followers and mutuality. Dawei et al. [18] and Zhang et al. [21] explored different structural features for link prediction in Twitter.

In this paper, we analyze links and community structures in Twitter to deal with the link prediction task. Our contributions are two fold: 1) We extend WIC measure for link prediction. WIC, proposed in [16] for undirected networks, combines similarity between vertices and community information. We also show it is efficient on large-scale asymmetric networks such as Twitter. 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 measures based on local information with our measure.

The remainder of the paper is organized as follows. In Section II, we present the link prediction problem and the standard metrics for performance evaluation. In Section III, we present some basic measures for link prediction based on local information and their extension for directed networks. In Section IV, we present our method for link prediction and the community detection algorithm that it requires. In Section V, we present experimental results obtained from Twitter. Finally, in Section VI we summarize the main findings and conclusions of this work.

## II. 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 pair of vertices in $V$, where $|V|$ denotes the number of elements in $V$. The fundamental task of a link prediction method is to find out the missing links (future links) in the set $U \setminus E$ (set of non-observed links) assigning a score for each link in this set. The higher the score, the higher the connection probability is, and vice versa [6], [21].

Given a predictor, we can rank all the non-observed 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, $E = E^T \cup E^P$ and $E^T \cap E^P = \varnothing$.

We apply two metrics to quantify the prediction accuracy [6]: AUC (area under the receiver operating characteristic curve) and precision. The AUC evaluates the predictor performance according to the whole list 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 nonexistent link (a link in $U \setminus E$). Let $n$ be the number of independent comparisons, if $n'$ times for the missing links are given higher scores than nonexistent links whilst $n''$ times for both missing and nonexistent links are given equal scores, AUC value is

$$AUC = \frac{n' + 0.5n''}{n} \tag{1}$$

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 pure chance the algorithm performs [6].

Different from AUC, precision only focuses on the $L$ links with highest scores. It is defined as the ratio of relevant items selected to 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

$$Precision = \frac{L_r}{L} \tag{2}$$

Clearly, higher precision means higher prediction accuracy.

## III. Structural link prediction measures

Structural measures use the similarity between vertices since similar vertices likely share same relations (links). The similarity between vertices can be classified into different ways, such as the measures based on local or global information. Measures based on global information may lead to higher accuracy, but their computation is very time-consuming and usually infeasible for large-scale networks. On the other hand, measures based on local information are generally faster but provide lower accuracy [5], [6].

The basic structural definition for a vertex $x \in V$ is its neighborhood $\Gamma(x) = \{y \mid (x, y) \in E \vee (y, x) \in E\}$ which denotes the set of friends of $x$. In directed networks (e.g., Twitter) the set of vertices formed by directed links from $x$ is different from the set of vertices 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 [3], [21].

Using these neighborhood definitions, next we present some basic standard measures based on local information [3], [5], [6].

*1) Common Neighbors (CN):* The common neighbors or common friends refer to the size of the set of all common friends, $\Lambda_{x,y}$, of both $x$ and $y$, according to Equation 3.

$$s_{x,y}^{CN} = |\Lambda_{x,y}| = |\Gamma(x) \cap \Gamma(y)| \tag{3}$$

For a directed network, CN measure can be defined based on the link direction: $s_{x,y}^{CN_{in}} = \left|\Lambda_{x,y}^{in}\right| = |\Gamma_{in}(x) \cap \Gamma_{in}(y)|$ and $s_{x,y}^{CN_{out}} = \left|\Lambda_{x,y}^{out}\right| = |\Gamma_{out}(x) \cap \Gamma_{out}(y)|$.

*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 Equation 4.

$$s_{x,y}^{AA} = \sum_{z \in \Lambda_{x,y}} \frac{1}{\log \Gamma(z)} \tag{4}$$

For a directed network, AA measure can be defined based on the link direction: $s_{x,y}^{AA_{in}} = \sum_{z \in \Lambda_{x,y}^{in}} \frac{1}{\log \Gamma_{in}(z)}$ and $s_{x,y}^{AA_{out}} = \sum_{z \in \Lambda_{x,y}^{out}} \frac{1}{\log \Gamma_{out}(z)}$.

*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 Equation 5.

$$s_{x,y}^{Jac} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{5}$$

For a directed network, Jac measure can be defined based on the link direction: $s_{x,y}^{Jac_{in}} = \frac{|\Gamma_{in}(x) \cap \Gamma_{in}(y)|}{|\Gamma_{in}(x) \cup \Gamma_{in}(y)|}$ and $s_{x,y}^{Jac_{out}} = \frac{|\Gamma_{out}(x) \cap \Gamma_{out}(y)|}{|\Gamma_{out}(x) \cup \Gamma_{out}(y)|}$.

*4) Resource Allocation (RA):* This measure punishes the high-degree common neighbors more heavily than AA, Equation 6.

$$s_{x,y}^{RA} = \sum_{z \in \Lambda_{x,y}} \frac{1}{\Gamma(z)} \tag{6}$$

For a directed network, RA measure can be defined based on the link direction: $s_{x,y}^{RA_{in}} = \sum_{z \in \Lambda_{x,y}^{in}} \frac{1}{\Gamma_{in}(z)}$ and $s_{x,y}^{RA_{out}} = \sum_{z \in \Lambda_{x,y}^{out}} \frac{1}{\Gamma_{out}(z)}$.

*5) Preferential Attachment (PA):* This measure is proportional to the number of neighbors of each vertex, Equation 7.

$$s_{x,y}^{PA} = |\Gamma(x)| \times |\Gamma(y)| \tag{7}$$

For a directed network, PA measure can be defined based on the link direction: $s_{x,y}^{PA_{in}} = |\Gamma_{in}(x)| \times |\Gamma_{in}(y)|$ and $s_{x,y}^{PA_{out}} = |\Gamma_{out}(x)| \times |\Gamma_{out}(y)|$.

## IV. Link prediction in large networks using community membership information

In [16] Valverde-Rebaza and Lopes have proposed the *within and inter cluster* (WIC) measure for link prediction. To use the WIC measure in a network, a clustering (or community detection) algorithm must be previously applied on it. Hence the score for the existence of a link between a pair of vertices is computed taking into account whether they belong or not to the same cluster/community. In spite of the WIC measure achieving better accuracy than local similarity measures (CN, Jac, AA, and others), it is necessary previously applying the clustering process. This additional process obviously imply in additional cost of identifying communities in the network.

Considering that clustering process in large-scale social networks as Twitter can be very expensive and in some cases intractable, the use of the WIC measure should not be recommended in this case. In this paper we propose the use of a specific algorithm for community detection applying the *label propagation algorithm* (LPA) [22], which has near-linear time complexity, enabling the use of the WIC measure for large networks. Furthermore, we modify the WIC measure for considering the in and outgoing links. The previous WIC measure uses a small value $\delta$ for avoiding division by zero. Here, in this case, we simply consider the WIC value as the cardinality of the within-cluster common neighbors set (see Section IV-B).

### A. 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 [12].

Several algorithms have been proposed to find community structures in networks. The *label propagation algorithm* (LPA) [22] is a simple and fast method for community detection. Initially, it assigns a label to each vertex. 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.

The main advantage of this simplified method 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 method.

### B. WIC measure

Experiments show that for a network with low clustering structure, link prediction measures based on structural similarity perform poorly [5]. Nonetheless, as the clustering structure of the network grows, the accuracy of these measures drastically improves [13]. With this consideration and to exploit the efficiency of link prediction measures based on local information, the *within and inter cluster* (WIC) measure [16] was proposed.

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

According to the Bayesian theory, the posterior probabilities that the same or different cluster labels are assigned to a pair of vertices $(x, y)$, given their common neighbors $\Lambda_{x,y}$, are respectively

$$P(x^{C_\alpha}, y^{C_\alpha} \mid \Lambda_{x,y}) = \frac{P(\Lambda_{x,y} \mid x^{C_\alpha}, y^{C_\alpha}) P(x^{C_\alpha}, y^{C_\alpha})}{P(\Lambda_{x,y})} \tag{8}$$

$$P(x^{C_\alpha}, y^{C_\beta} \mid \Lambda_{x,y}) = \frac{P(\Lambda_{x,y} \mid x^{C_\alpha}, y^{C_\beta}) P(x^{C_\alpha}, y^{C_\beta})}{P(\Lambda_{x,y})} \tag{9}$$

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

To estimate the probability of the common neighbors of a pair vertex $(x^{C_\alpha}, y^{C_\alpha})$ given these vertices belong to the same community with label $C_\alpha$, consider the number of common neighbors with the same community label divided by the total number of common neighbors, as stated in Eq. 10.

$$P(\Lambda_{x,y} \mid x^{C_\alpha}, y^{C_\alpha}) = \frac{\left|\Lambda^W_{x,y}\right|}{\left|\Lambda_{x,y}\right|} \tag{10}$$

Similarly, to estimate the probability of the common neighbors of pair vertex $(x^{C_\alpha}, y^{C_\beta})$ given these vertices belong to different communities with labels $C_\alpha$ and $C_\beta$, consider the number of common neighbors that may be associated with different labels $C_\alpha$ or $C_\beta$ or with another community label $C_\gamma$ divided by the total number of common neighbors, as defined by Eq. 11.

$$P(\Lambda_{x,y} \mid x^{C_\alpha}, y^{C_\beta}) = \frac{\left|\Lambda^{IC}_{x,y}\right|}{\left|\Lambda_{x,y}\right|} \tag{11}$$

In order to compare the existence likelihood between the vertex pairs, the likelihood score of a vertex pair $(x, y)$ is defined as the ratio of Eq. 8 to Eq. 9. Substituting Eqs. 10 and 11, we have:

$$s_{x,y} = \frac{\left|\Lambda^W_{x,y}\right|}{\left|\Lambda^{IC}_{x,y}\right|} \times \frac{P(x^{C_\alpha}, y^{C_\alpha})}{P(x^{C_\alpha}, y^{C_\beta})} \tag{12}$$

Considering that each vertex belongs to only one community, i.e., $y^{C_\alpha} = y^{C_\beta}$, the $\frac{P(x^{C_\alpha}, y^{C_\alpha})}{P(x^{C_\alpha}, y^{C_\beta})}$ ratio can be neglected since this fraction value is 1. Thus, the final *within and inter cluster* (WIC) measure for an undirected network is:

$$s^{WIC}_{x,y} = \begin{cases} \left|\Lambda^W_{x,y}\right|, & if \;\; \Lambda^W_{x,y} = \Lambda_{x,y} \\ \frac{\left|\Lambda^W_{x,y}\right|}{\left|\Lambda^{IC}_{x,y}\right|}, & otherwise \end{cases} \tag{13}$$

It is important to notice that, in [16] to prevent division by zero a small constant $\delta \approx 0$ is added to the denominator of score, i.e., $s^{WIC}_{x,y} = \frac{\left|\Lambda^W_{x,y}\right|}{\left|\Lambda^{IC}_{x,y}\right| + \delta}$ . In this paper, we do not consider the use of $\delta$ because, when $\Lambda^W_{x,y} = \Lambda_{x,y}$, the score does not work according to the real precision value since it is inversely proportional to $\delta$, thus highly increasing for small values of $\delta$.

In addition, we extend the WIC measure for applying it in directed networks. Similarly to the measures showed in Section III, the WIC measure can be defined based on the link direction considering the sets of incoming and outgoing within-cluster common neighbors: $\Lambda^{W_{in}}_{x,y} = \left\{ z^C \in \Lambda^{in}_{x,y} \mid x^C, y^C \right\}$ and $\Lambda^{W_{out}}_{x,y} = \left\{ z^C \in \Lambda^{out}_{x,y} \mid x^C, y^C \right\}$, and the sets of incoming and outgoing inter-cluster common neighbors: $\Lambda^{IC_{in}}_{x,y} = \Lambda^{in}_{x,y} \setminus \Lambda^{W_{in}}_{x,y}$ and $\Lambda^{IC_{out}}_{x,y} = \Lambda^{out}_{x,y} \setminus \Lambda^{W_{out}}_{x,y}$. These sets are used in Eq. 13 to obtain $s^{WIC_{in}}_{x,y}$ and $s^{WIC_{out}}_{x,y}$, respectively.

## V. Experiments

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

### A. 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. [17]. Twitter network differs from others social networks by its directed relationship nature, i.e., a Twitter user is not obligated to reciprocate followers by following them back. So, 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 algorithm. As in [23], two subgraphs were generated with vertices labeled accordingly to the clusters obtained at 7th and 15th iterations, *Twitter 7it* and *Twitter 15it*, respectively. Basic features of these graphs are summarized in Table I.

### B. Experimental setup

In our experiments, for testing set ($E^P$) 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. The remaining links, except those formed by users

Table I
BASIC FEATURES OF THE TWO GRAPHS BUILT FROM TWITTER NETWORK AFTER 7TH AND 15TH ITERATIONS OF LPA ALGORITHM.

| | Twitter 7it | Twitter 15it |
|---|---|---|
| $\|V\|$ | 24617334 | 24617333 |
| $\|E\|$ | 363565896 | 363565892 |
| $M$ | 3415051 | 2250964 |
| max cluster size | 1392411 | 10121242 |
| ratio of total links per user | 14.77 | 14.77 |

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 [8], [18], [21].

For each vertex pair, 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 $s_{x,y}^{WIC_{out}} > s_{x,y}^{WIC_{in}}$ then $s_{x,y}^{WIC} = s_{x,y}^{WIC_{out}}$, otherwise, $s_{x,y}^{WIC} = s_{x,y}^{WIC_{in}}$.

### C. Validating results and analysis

AUC and Precision metrics are employed to validate the quality of each link prediction measure. Table II summarizes the prediction results measured by AUC on *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.

Table II
PRECISIONS MEASURED BY AUC ON TWO SUBGRAPHS OF TWITTER NETWORK. THE EMPHASIZED VALUES CORRESPOND TO THE HIGHEST PRECISION AMONG THE EVALUATED MEASURES.

| Graph | WIC | CN | AA | Jac | RA | PA |
|---|---|---|---|---|---|---|
| Twitter 7it | **0.62** | 0.56 | 0.53 | 0.45 | 0.6 | 0.51 |
| Twitter 15it | **0.62** | 0.56 | 0.53 | 0.45 | 0.6 | 0.51 |

Looking at the results of Table II, one should notice that the AUC performance of each link prediction measure is the same for both subgraphs. In the case of WIC measure, this indicates that although both subgraphs have different number of groups ($M$), relations and interests between nodes remain similar or equivalent, i.e., most of the users of Twitter network sharing similar interests is grouped in the same communities. In the others measures, the same AUC performance to both subgraphs is justified by its similar number of nodes ($|V|$) and links ($|E|$). Comparing AUC performance of all link prediction measures, WIC outperforms all of them. RA and CN are the next best measures. Jac has the worst performance and do not outperform the assignment by chance.

Figure 1 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, WIC, CN, AA and RA obtain 0.32, 0.32, 0.30, and 0.28 precision values, respectively, whereas Jac and PA precision values are 0.0. In the top-L links, with $L = 500$, 5000 and 10000, WIC performs better than the other measures. Just when $L = 1000$, Jac has a peak and performs better than the other measures, leaving WIC as the second best measure. When $L = 5000$ and $L = 10000$, the Jac performance decreases significantly and the WIC performance is the best. PA has the worst precision for all values of $L$.

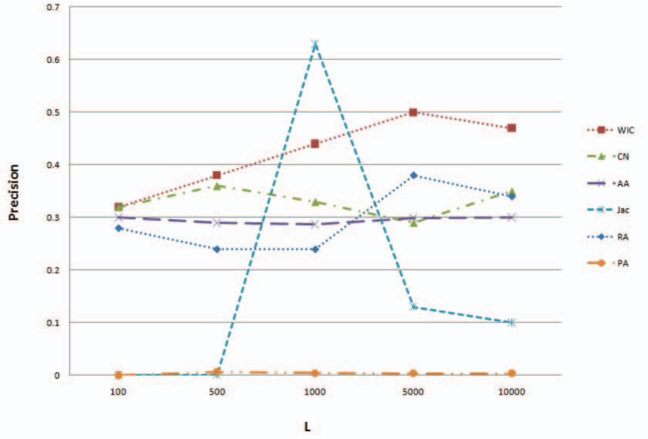

Figure 1. Precisions on the two graphs from Twitter network. Different values of $L$ are used to select the top-$L$ highest scores for predicting links.

## VI. CONCLUSION

We use the WIC measure for link prediction on two different directed graphs built from the Twitter network. For predicting link between a pair of vertices the WIC measure takes into account the information on which communities the pair of vertices (and their neighbors) belong to, i.e., if the vertices are in a same or different community. Since for applying the WIC measure is need, in a previous phase, partitioning the network into communities. Here we propose and evaluate the use of a fast community detection algorithm, the LPA algorithm, in order to be able of applying the WIC measure for large network. The performance of WIC compared to CN, AA, Jac, RA, and PA measures was better considering AUC and precision criteria. Just one case, when $L = 1000$ the precision of Jac measure outperforms the precision of WIC.

In summary, our experiments suggest that WIC measure captures information from the communities the vertices belong to, improving the link prediction task. This happens because vertices of the same communities likely have similar interests. In the case of the Twitter network, the similar interests 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.

## ACKNOWLEDGMENT

This work is partially supported by CNPq and FAPESP agencies from Brazil.

## REFERENCES

[1] I. Lunden. (2012, July) Analyst: Twitter Passed 500M Users In June 2012, 140M Of Them In US; Jakarta Biggest Tweeting City. Techcrunch. [Online]. Available: http://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/

[2] J. Constine. (2012, August) How Big Is Facebook's Data? 2.5 Billion Pieces Of Content And 500+ Terabytes Ingested Every Day. Techcrunch. [Online]. Available: http://techcrunch.com/2012/08/22/how-big-is-facebooks-data-2-5-billion-pieces-of-content-and-500-terabytes-ingested-every-day/

[3] M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici, "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)*, oct. 2011, pp. 73 –80.

[4] J. Hopcroft, T. Lou, and J. Tang, "Who will follow you back?: reciprocal relationship prediction," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, ser. CIKM '11. New York, NY, USA: ACM, 2011, pp. 1137–1146.

[5] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *JASIST*, vol. 58, no. 7, pp. 1019–1031, May 2007.

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

[7] M. Kotera, Y. Yamanishi, Y. Moriya, M. Kanehisa, and S. Goto, "Genies: gene network inference engine based on supervised analysis," vol. 40.

[8] Z. Yin, M. Gupta, T. Weninger, and J. Han, "Linkrec: a unified framework for link recommendation with user attributes and graph structure," in *Proceedings of the 19th International Conference on World Wide Web*, 2010.

[9] N. Benchettara, R. Kanawati, and C. Rouveirol, "A supervised machine learning link prediction approach for academic collaboration recommendation," in *Proceedings of the fourth ACM conference on Recommender systems*, ser. RecSys '10. New York, NY, USA: ACM, 2010, pp. 253–256.

[10] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki, "Link prediction using supervised learning," in *Proceedings of SDM 06 Workshop on Link Analysis, Counterterrorism and Security*, 2006.

[11] K. Y. Itakura, C. L. A. Clarke, S. Geva, A. Trotman, and W. C. Huang, "Topical and structural linkage in wikipedia," in *Proceedings of the 33rd European conference on Advances in information retrieval*, ser. ECIR'11, 2011, pp. 460–465.

[12] S. Fortunato, "Community detection in graphs," *CoRR*, vol. abs/0906.0612v2, 2010.

[13] X. Feng, J. Zhao, and K. Xu, "Link prediction in complex networks: a clustering perspective," *Eur. Phys. J. B*, vol. 85, no. 1, p. 3, 2012.

[14] S. Soundarajan and J. Hopcroft, "Using community information to improve the precision of link prediction methods," in *Proceedings of the 21st International Conference Companion on World Wide Web*, ser. WWW '12 Companion. New York, NY, USA: ACM, 2012, pp. 607–608.

[15] E. Hoseini, S. Hashemi, and A. Hamzeh, "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. Washington, DC, USA: IEEE Computer Society, 2012, pp. 795–800.

[16] J. Valverde-Rebaza and A. Lopes, "Link prediction in complex networks based on cluster information," in *XXI Brazilian Symposium on Artificial Intelligence*, ser. SBIA 2012, 2012, pp. 92–101.

[17] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proceedings of the 19th international conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 591–600.

[18] D. Yin, L. Hong, and B. D. Davison, "Structural link analysis and prediction in microblogs," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, ser. CIKM '11. New York, NY, USA: ACM, 2011, pp. 1163–1168.

[19] D. M. Romero and J. M. Kleinberg, "The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter," in *ICWSM*, 2010.

[20] S. A. Golder and S. Yardi, "Structural predictors of tie formation in twitter: Transitivity and mutuality," in *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, ser. SOCIALCOM '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 88–95.

[21] Q.-M. Zhang, L. Lü, W.-Q. Wang, Y.-X. Zhu, and T. Zhou, "Potential theory for directed networks," *CoRR*, vol. abs/1202.2709, 2012.

[22] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Phys. Rev. E*, vol. 76, p. 036106, Sep 2007.

[23] A. U. Bhat. (2010) Scalable community detection using label propagation & map-reduce. [Online]. Available: http://www.akshaybhat.com/LPMR/