Identifying Prominent Actors in Online Social Networks using Biased Random Walks

Frank W. Takes Walter A. Kosters

Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands

Abstract

In this paper we describe the structural characteristics of prominent actors that reside within an online social network. We will show how structural properties can be used in a classification algorithm based on biased random walks for distinguishing between prominent and regular nodes in a social network. The effectiveness of our approach is demonstrated on a large social network dataset with 8 million users and almost 1 billion links, in which the prominent users are labeled. We show that we can efficiently identify a large portion of the prominent individuals in the network using our algorithm, outperforming standard web-inspired methods such as HITS and PageRank.

1 Introduction

Over the past decade *online social networks* (*OSNs*) such as Facebook, Twitter, Hyves and LinkedIn have become extremely popular, with numbers as big as 500 million users and an estimated 50 billion friendship links (Facebook[1], 2010). The main concept of these social networks is simple: a user creates a profile with some personal attributes and then links this profile to other users, the so-called *friends*, creating a very large graph of befriended users: the *friendship graph*. In order to better understand the rich amount of information that is contained in social networks, the friendship graph is extensively being measured, modeled and mined.

A quite natural query with respect to a social network, is to ask who the most prominent, or the most important actors in the network are. Being able to identify such prominent actors has various useful applications. For example, companies nowadays frequently use social networks for their *viral marketing* [11] campaigns, in which they want to deliver a message to as many people as possible through the social network's linking structure. Prominent nodes may just be the places where such a campaign should start in order to reach a large number of people as quickly as possible.

The research question that we posed in the previous section immediately raises the issue of how prominence or importance of a node within a network should actually be defined. While various definitions may be correct, we will assume that someone is *prominent*, or *important*, or *influential*, if he or she has some celebrity status (famous politicians, soccer players, artists, movie actors, etc.) in the real world. Though this definition of prominence can be argued, we believe it can be justified based on the fact that both online and in the real world, celebrities have a certain status, or reputation. It is well-known that if a celebrity promotes a certain brand, people are far more likely to identify with that brand, compared to when a regular person would promote the brand. Within the online social network Twitter, tweets originating from people like president Obama are far more likely to be "retweeted" than when they would come from a regular person in the network. Thus, Obama could in essence be seen as more prominent than a regular person.

In this paper we first study the difference in characteristics between regular and prominent nodes within the network. We will then consider various existing methods of determining the importance of nodes in the friendship graph of an online social network, and introduce a new method which is based on the characteristics that we obtained. We will test the discussed approaches empirically on a large full crawl of an online social network of 8 million users and almost 1 billion links. For this network, we know exactly which users are considered to be prominent, allowing us to verify the obtained results against this community defined ground truth.

[1] See `https://www.facebook.com/press/info.php?statistics` (accessed June 20, 2011)

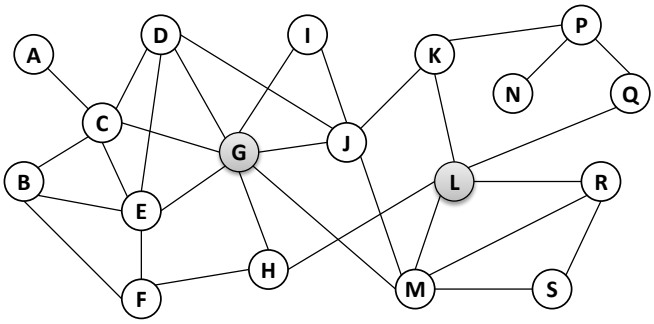

Figure 1: A graph with 18 nodes of which 2 nodes (G and L) are considered to be "prominent".

The rest of this paper is structured as follows. In Section 2 we formally define our problem, after which we discuss related and previous work in Section 3. We specifically pay attention to existing methodologies for determining the importance of a node. In Section 4 we discuss the characteristic properties of prominent nodes, and how these properties can be incorporated in a random walk algorithm. In Section 5 we will compare the performance of the discussed methods. Section 6 concludes.

2 Preliminaries

In this section we will discuss some basis concepts and formulate our problem statement.

2.1 Definitions

We are given a friendship graph $G(V, E)$, where V is the set of $n = |V|$ nodes (individuals) and $E \subseteq V \times V$ is the set of $m = |E|$ edges (connections). The graph is undirected, meaning that the set of edges is symmetric, so if $(u, v) \in E$ then also $(v, u) \in E$. A *path* or *walk* from u to v is a sequence of edges, starting with an edge containing u and ending with an edge containing v. The distance $d(u, v)$ between two nodes u and v is defined as the length of a shortest possible path between these nodes. Because our graph is undirected, $d(u, v) = d(v, u)$ for all $u, v \in V$. As we will only consider the largest connected component of the graph, $d(u, v)$ is finite for all $u, v \in V$. We define the neighborhood $N(v)$ of a node v as the set of nodes at distance 1 of v, more specifically: $N(v) = \{u \in V \mid (u, v) \in E\}$. We can now define the *degree* of a node v, indicating the number of edges starting (or ending) at some node v as the size of its neighborhood: $deg(v) = |N(v)|$.

2.2 Problem Statement

Amongst the nodes in the network, there is an initially unknown set $W \subseteq V$, of size $k = |W|$, which contains the nodes that are considered to be "prominent". Logically, $k \leq n$, but in practice, k is a lot smaller than n, as only a small portion of the nodes is typically considered to be prominent. Our main goal is to find, given *only* the graph $G(V, E)$, an as small as possible subset $I \subseteq V$ such that $|I \cap W|$ is maximal, i.e., we are trying to find as many prominent nodes as possible.

In this paper we describe various existing, derived, and new methods for determining node importance. For each of these methods M we assign a normalized value $f_M(v) \in [0, 1]$ to each node $v \in V$ which determines its importance. We will assume that higher values indicate a higher level of importance. In order to determine the performance of a method M, we sort the list of nodes by their importance value $f_M(v)$ in descending order, and define I to be the top ℓ nodes of this sorted list. The *precision* and *recall*, $|W \cap I|/|I|$ and $|W \cap I|/|W|$, respectively, will ultimately determine the performance of a method M. More generally, we can say that the F-measure, $(2 \times precision \times recall)/(precision + recall)$, measures the balance between the two. Note that if $\ell = k$, precision, recall and F-measure are equal.

An example with 18 nodes of which 2 nodes (G and L) are considered to be prominent, is given in Figure 1. If some method M would determine that E and G are the prominent nodes, then $W \cap I = \{G, L\} \cap \{E, G\} = \{G\}$ and the performance of this method (in terms of both precision and recall) would be $\frac{1}{2}$, so 50%.

2.3 Online Social Networks

Over the past decade, topological properties of online social networks have been studied in great detail [13]. Social networks are usually sparse and composed of a large connected component containing the majority of the nodes. Often, there are a few smaller isolated communities, as well as various singletons [1]. Furthermore, it is well-known that online social networks greatly resemble the thoroughly studied real-life social networks [16]. Another interesting fact is that social networks generally belong to the class of *small world networks* [17]. Such networks are characterized by relatively small pairwise distances between nodes, i.e., the average distance between two nodes behaves only logarithmically compared to the total number of nodes. Online social networks are often also *scale-free*, meaning that the node degree distribution of a small world network follows a power law: there are relatively few nodes with an exceptionally high degree, and many nodes with a low(er) degree. The high degree nodes function as hubs, and are often grouped in a densely connected core, realizing the short pairwise distances between the more "peripheral" nodes.

3 Related Work

Various studies have addressed the problem of identifying the prominent actors within large (social) networks. To further refine the context of our work, we will only discuss *global importance*: the entire graph is ranked based on some measure or algorithm. Though beyond the scope of our work, a contrasting type of query is that of *relative importance* [18], where the goal is to determine the importance of only a certain node with respect to some root node. Other work deals with finding experts, or who can be trusted within some semantic social network [5, 19]. We will clearly distinguish from these methods by not considering semantics, but only structural properties of the nodes.

In that context, methods such as HITS [7] and PageRank [14] are known to be very successful in determining the importance of web pages [10], citation networks [3] and Wikipedia topics [6]. Therefore, we will also consider the undirected variants of these two approaches as methods for identifying the prominent actors in a social network. In [15], the NODERANKING algorithm is proposed as a method based on random walks for determining the importance of authors in a directed citation network. Random walk algorithms generally traverse the graph, moving to a random neighbor with probability $1 - p$, and jumping to a random node with probability p. The distinguishing property of the NODERANKING algorithm is that it jumps with a probability depending on the current node's degree, where a low degree indicates a high jumping probability.

Centrality measures have been popularized by social scientists in the 1970's as possible measures for the importance, or "prestige" as they call it, of a person within a social network [16]. Such measures assume that anyone who is central in the network, meaning that he or she is connected to a lot of other people via some short path, is prominent. *Degree centrality* is by far the simplest and most common measure, and is in case of an online social network simply equal to the number of friends that someone has. As we will see later on, the number of friends is a good indication of a node's prominence, but definitely not perfect. Unfortunately, the complexity of calculating other centrality measures such as betweenness centrality, closeness centrality and graph centrality is in the order of $O(mn)$ or worse [2], and therefore not considered in this work.

4 Methodology

We will now outline our approach for determining the prominent actors in an online social network. First we sketch the characteristic node properties of our target nodes. After that, we will describe an algorithm based on random walks which uses these properties to guide the walk towards the prominent nodes. We believe that a random walk algorithm is a valid choice, because such algorithms are, in case of massive graphs such as social networks, more practical than standard decision tree algorithms which require flattened data.

4.1 Properties of Prominent Nodes

The simplest intuition that we have about prominent people, is that they have a large number of connections. Therefore we expect the degree of a node to play a great role in determining the importance of a node. So we could state that the degree centrality, determining the importance of a node v based on its number of connections, could be a good first indication of importance, formally:

$$f_{deg}(v) = 1 - \frac{1}{|N(v)|}$$

However, there may be nodes in the graph with many connections, that are not prominent, or vice versa, prominent people with a smaller number of connections. Let us recall several observations regarding social networks in general, which have been described in literature. People tend to use social networks for two reasons: *social searching*, and *social browsing* [9]. These two terms refer to reconfirming real-life friendships online, and browsing for completely new relationships, respectively. Another common concept is that of *triadic closure*: the vast majority of all friendships formed within a social network takes place between two people who have at least one friend in common [8]. This probability has been shown to increase with the number of common acquaintances [8] as well as with the degree of a node (a phenomenon called *preferential attachment*). For example, in the graph in Figure 1, the connection (A, B) would be more likely to appear than the connection (A, K), as A and B have node C as a common friend, and A and K have no common friends. The connection (A, D) would in turn be more likely than (A, B), as D already has a higher degree.

Based on the above, we expect that a regular person adding someone like president Obama, is not within the circle of friends of Obama, making this friendship more like a result of social browsing instead of searching. More generally, we argue that the friends of prominent nodes have more connections in common than regular nodes. This can also be formulated as a smaller number of closed triangles amongst friends of prominent nodes. We call this concept a node's *neighborhood density* (*nd)*:

$$f_{nd}(v) = 1 - \sum_{w \in N(v)} \frac{|N(w) \cap N(v)|}{(|N(w)| - 1) * |N(v)|}$$

Here, the numerator defines the number of common connections, whereas the denominator normalizes the result so that it is independent of the degree of v or the degree of w. If $|N(v)| > 1$, $f_{nd}(v)$ is minimal in case $N(v)$ is fully connected, and becomes larger as a smaller fraction of the neighborhood is interconnected.

We have verified the two intuitions mentioned above on our social network dataset (see Section 5.1 for a description of the dataset). To do this, we took $1,000$ random regular nodes and $1,000$ random prominent nodes, and compared their importance values. The degree centrality f_{deg} was on average equal to $8.9 \cdot 10^{-3}$ for regular nodes, and $3.4 \cdot 10^{-2}$ for prominent nodes. Indeed, prominent nodes appear to have on average around 3 to 4 times as many friends as regular nodes. For the neighborhood density f_{nd}, we found a value of $9.1 \cdot 10^{-1}$ for regular nodes, and $4.5 \cdot 10^{-1}$ for prominent nodes, which is consistent with our intuition of prominent nodes having fewer closed triangles within their neighborhood as compared to regular nodes.

We believe that a combination of the two measures analyzed above may be able to efficiently identify the various prominent actors. Therefore we devised an algorithm based on random walks, which has a parameterized bias towards each of these properties.

Algorithm 1 BIASEDRANDOMWALK

Input: Graph $G(V, E), N, p, \alpha$
Output: $f[\,]$, containing the importance value $f[v]$ for each node $v \in V$

for $v \in V$ **do**
  $f[v] \leftarrow 0$
end for
$i \leftarrow 0$

$v \leftarrow$ RANDOMNODEFROM(V)
while $(i < N)$ **do**
  $f[v] \leftarrow f[v] + \frac{1}{N}$
  if (**rand**$(0, 1) > p$) **then**
    $v \leftarrow$ BIASSELECTFROM$(N(v), \alpha)$
  else
    $v \leftarrow$ RANDOMNODEFROM(V)
  end if
  $i \leftarrow i + 1$
end while

return $f[\,]$

4.2 BiasedRandomWalk

Our algorithm, called BIASEDRANDOMWALK (BRW), takes as input an unweighted graph $G(V, E)$ and parameters N, p and α, and outputs a function value $f_{BRW}(v)$ for each node $v \in V$ in the graph, determining its importance. Here N is the number of steps in the random walk algorithm, p is the jumping probability, and α is used to define the focus on either one of the two measures that we discussed.

The procedure is outlined in Algorithm 1, and works as follows. After setting some initialization values in lines 3–6, the algorithm starts by selecting a random node from V (line 7). After that, for N iterations, the algorithm repeatedly increases the function value (line 9) of the current node v by $1/N$ (to keep the function value within $[0; 1]$). Then, the algorithm either selects a new node from the neighborhood $N(v)$ of v using the function BIASSELECTFROM() with probability $1-p$ (line 11), or jumps to a completely random node with probability p (line 13). If in BIASSELECTFROM() a random neighbor is selected, the algorithm would be a plain random walk algorithm. However, in our case, the function BIASSELECTFROM() selects a node with a probability dependent on different function values of our prominence measures, as we know that each of these function values tells us something about the probability of that node being prominent. So given current node v, the probability $P(w)$ of selecting a node $w \in N(v)$ is equal to:

$$P(w) = \frac{\alpha f_{deg}(w) + (1-\alpha) f_{nd}(w)}{\sum_{u \in N(v)} (\alpha f_{deg}(u) + (1-\alpha) f_{nd}(u))}$$

Setting the value of α to 1 logically resulted in roughly the same result as degree centrality, whereas a value of 0 turned out to give 0% success. We believe that this is due to the fact that even though f_{nd} is normalized, the degree plays a significant role in identifying a node's prominence, and very low degree nodes can still get a high neighborhood density score. It turned out that any value between 0.2 and 0.8 gave decent results and therefore we fixed the parameter to 0.5 to give equal focus to both measures. As for p, we fixed this value to 0.15 as suggested in literature [12]. Finally, N, the number of iterations, should be set to a value significantly larger than the number of nodes n.

5 Experiments

In this section we will compare our algorithm with various existing approaches for determining node importance in networks. Our algorithm as well as other discussed measures have been implemented in C++ and tested on a 3.2GHz machine with 10GB memory, allowing us to keep the large network dataset in memory.

5.1 Dataset

We will verify the various methods on an anonymized large full crawl of the local Dutch online social network HYVES. This network consists of an undirected friendship graph, of which some statistics are given in Table 1. The power law node degree distribution of this dataset in Figure 2 as well as the distance distribution (sampled over $1,000$ nodes) in Figure 3 demonstrate how the network adheres to the small-world property (see Section 2.3). Note that the network had some predefined maximum number of friends at $1,000$, $1,500$ and $2,000$, causing some slight noise in the tail of the degree distribution. Next to the friendship graph, we also have a set of nodes W of size $|W| = 4,867$ (0.06%) which is considered to be "prominent". This subset consists of various Dutch politicians, artists, athletes and actors and could be considered as a ground truth, allowing us to verify the results from each of the methods.

Property	Value
Nodes	8,113,017
"Prominent" Nodes	4,867 (0.06%)
Edges	912,120,070
Average Degree	112
Average Distance	4.75
Diameter	25
Connected Components	9,926
Nodes in Largest Component	8,083,964 (99.6%)

Table 1: Dataset statistics.

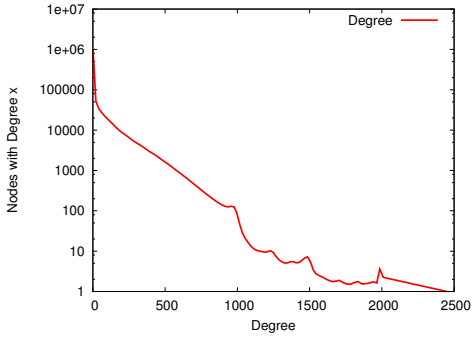

Figure 2: Dataset degree distribution (power law).

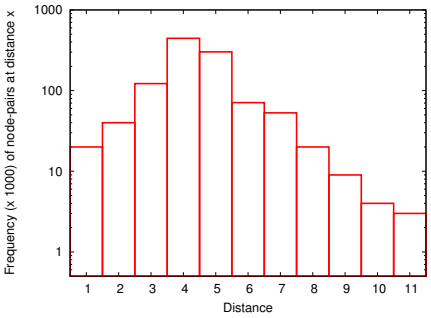

Figure 3: Dataset distance distribution.

5.2 Results & Discussion

The results of applying the various algorithms are outlined in Table 2. Here we compare the results of each of the methods based on $k = \ell$, meaning that we select exactly as many prominent people as there are in the dataset. Recall from Section 2 that we thus select the top $\ell = k$ nodes from the list of nodes sorted by their prominence function value. The Random Walk algorithm was executed with parameters $p = 0.15$ and $N = 10 * n = 80$ million iterations. Results from the Random Walk and BIASEDRANDOMWALK algorithm are averages of 10 runs in order to flatten the effect of outliers due to the inherent randomness of the approaches.

Measure	Recall
Degree Centrality	64.2%
Random Walk	63.9%
PageRank	56.4%
HITS	51.4%
BIASEDRANDOMWALK	70.1%

Table 2: Results of various importance measures with $k = \ell$.

As a baseline for comparison we could say that if we were to select ℓ random nodes to form the set W, we would on average find 0.06% of the prominent nodes in the network. Degree centrality, by far the simplest selection strategy, already greatly improves upon this by identifying 64.2% correctly. A plain random Random Walk results in slightly lower (but roughly equal) performance, whereas both HITS and PageRank perform significantly worse. Our method, BIASEDRANDOMWALK, improves another 9% upon Degree centrality, demonstrating the use of looking at the percentage of closed triangles during the walk. Unfortunately a comparison with NODERANKING (see Section 3) is not very interesting, as results produced by that algorithm are virtually identical to the simple method of degree centrality, which we expect to be due to the fact that our graph is undirected.

One may argue that the number of prominent actors in a network is not always known in advance. Therefore we also did experiments in which we varied ℓ between $0.01*k$ and $2*k$ on a sample of our dataset, allowing us to study the precision, recall and F-measure curves of each of the approaches. These three values for the methods with the highest performance (cf. Table 2), Degree centrality and BIASEDRANDOMWALK, are shown in Figures 4 and 5, respectively. The sample that we used to create these diagrams consisted of $500,000$ nodes (6% of the original dataset). The sample was created using random walk sampling, suggested in [4] as a method for obtaining a representative sample of an online social network. Indeed, obtained statistics for the sample deviated only slightly from the numbers in the entire dataset. Though the difference between Figures 4 and 5 may appear small, improvements in terms of each of the three values can be observed.

Assuming that we want to take a number of false positives for granted, we may choose to focus solely on maximizing the recall value. Therefore we also present the recall value for each of the approaches as a function of the fraction of k in Figure 6. In Figure 7 we furthermore present a comparison of the different F-values. At $0.75 * k$, the F-value appears optimal for BIASEDRANDOMWALK, and we would have a good balance between recall and precision. This optimum lies slightly lower around $0.70 * k$ for Degree centrality.

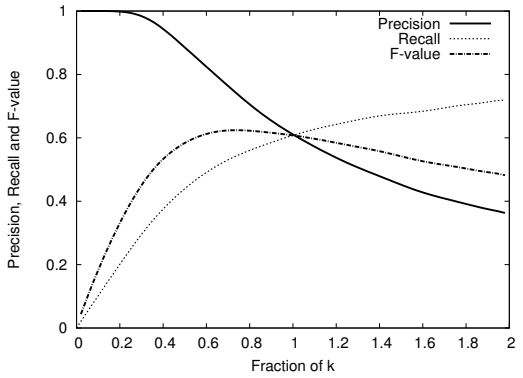

Figure 4: Degree centrality.

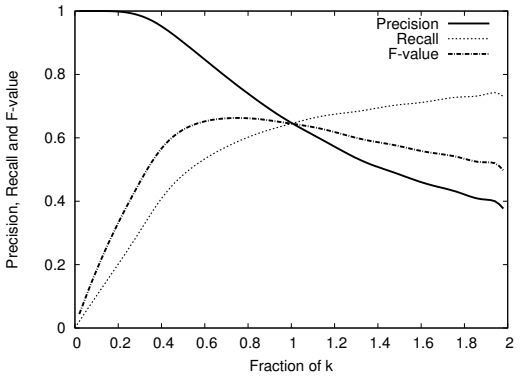

Figure 5: BiasedRandomWalk.

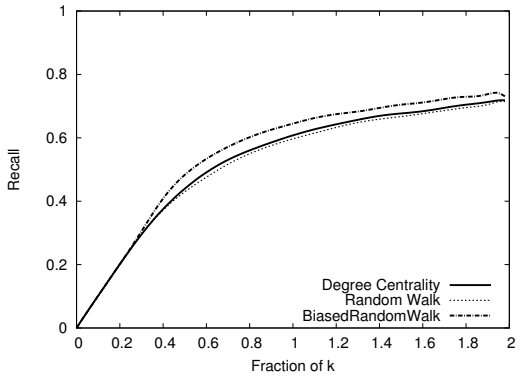

Figure 6: *Recall* for each of the methods.

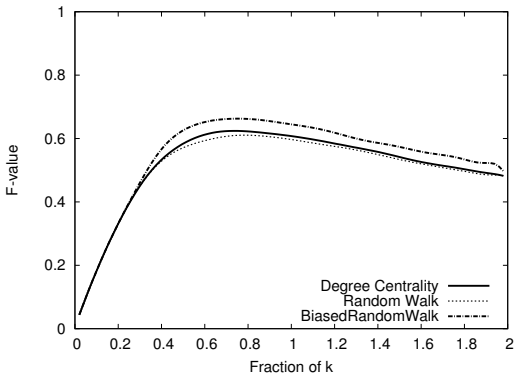

Figure 7: *F-value* for each of the methods.

Finally, note that for small values of ℓ (up to $0.3 * k$), the obtained result is always perfect: apparently the top of the list is the same for each of the measures. It turns out these nodes simply had an enormously high degree (over $2,000$), and were therefore selected by each of the methods.

6 Conclusion

We have outlined various characteristic node properties of prominent actors in a social network, and used these properties in creating an algorithm for identifying prominent actors. Our algorithm, called BIASED-RANDOMWALK, combines the measures of degree centrality and neighborhood density in a random walk algorithm by having a bias towards nodes of each of the two measures. Neighborhood density can be seen as a measure of the percentage of triadic closure, which is significantly lower for prominent actors as compared to regular nodes. Experiments show that our method works quite well, as standard methods such as degree centrality, HITS and PageRank are clearly outperformed, regardless of which measure (precision, recall or F-measure) is used to determine the method's performance.

In future work we would like to address the issue of determining relative importance other than just global importance. We also want to consider the temporal aspect of importance, and study how prominence of nodes within a social network changes over time.

Acknowledgments

This research was done as part of the NWO COMPASS project (grant #612.065.92). We thank Iris Hupkens as well as the anonymous reviewers for their input regarding this work.

References

[1] Y.Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of Topological Characteristics of Huge Online Social Networking Services. In *Proceedings of the 16th International Conference on World Wide Web*, pages 835–844, 2007.

[2] U. Brandes. A Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[3] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding Scientific Gems with Google's PageRank Algorithm. *Journal of Informetrics*, 1(1):8–15, 2007.

[4] M. Gjoka, M. Kurant, C.T. Butts, and A. Markopoulou. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. In *Proceedings of the 29th IEEE International Conference on Computer Communications*, pages 1–9, 2010.

[5] J. Golbeck and J. Hendler. Accuracy of Metrics for Inferring Trust and Reputation in Semantic Web-based Social Networks. In *Proceedings of EKAW 2004 (LNCS 3257)*, pages 116–131, 2004.

[6] A.M. Kentsch, W.A. Kosters, P. van der Putten, and F.W. Takes. Exploratory Recommendations Using Wikipedia's Linking Structure. In *Proceedings of the 20th Belgian Netherlands Conference on Machine Learning (Benelearn)*, pages 61–68, 2011.

[7] J.M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632, 1999.

[8] G. Kossinets and D.J. Watts. Empirical Analysis of an Evolving Social Network. *Science*, 311(5757):88–90, 2006.

[9] C. Lampe, N. Ellison, and C. Steinfield. A Face(book) in the Crowd: Social Searching vs. Social Browsing. In *Proceedings of the 20th Anniversary Conference on Computer Supported Cooperative Work*, pages 167–170, 2006.

[10] A.N. Langville and C.D. Meyer. *Google Page Rank and Beyond*. Princeton University Press, 2006.

[11] J. Leskovec, L.A. Adamic, and B.A. Huberman. The Dynamics of Viral Marketing. *ACM Transactions on the Web*, 1(1):5, 2007.

[12] J. Leskovec and C. Faloutsos. Sampling from Large Graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 631–636, 2006.

[13] A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, pages 29–42, 2007.

[14] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. *Technical Report, Stanford University*, 1999.

[15] J.M. Pujol, R. Sangüesa, and J. Delgado. Extracting Reputation in Multi Agent Systems by means of Social Network Topology. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 1*, pages 467–474, 2002.

[16] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[17] D.J. Watts and S.H. Strogatz. Collective Dynamics of Small-World Networks. *Nature*, 393(6684):440–442, 1998.

[18] S. White and P. Smyth. Algorithms for Estimating Relative Importance in Networks. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 266–275, 2003.

[19] J. Zhang, M.S. Ackerman, and L. Adamic. Expertise Networks in Online Communities: Structure and Algorithms. In *Proceedings of the 16th International Conference on World Wide Web*, pages 221–230, 2007.