DOI: http://dx.doi.org/10.15446/ing.investig.v35n2.50339

An algorithm for identifying the best current friend in a social network

Un algoritmo para determinar el mejor amigo actual en una red social

F. J. Moreno[1], and S. Hernández[2]

ABSTRACT

A research field in the area of social networks (SNs) is the identification of some types of users and groups. To facilitate this process, a SN is usually represented by a graph. The centrality measures, which identify the most important vertices in a graph according to some criterion, are usual tools to analyze a graph. One of these measures is the PageRank (a measure originally designed to classify web pages). Informally, in the context of a SN, the PageRank of a user *i* represents the probability that another user of the SN is seeing the page of *i* after a considerable time of navigation in the SN. In this paper, we define a new type of user in a SN: *the best current friend*. The idea is to identify, among the friends of a user *i*, who is the friend *k* that would generate the highest decrease in the PageRank of *i* if *k* stops being his/her friend. This may be useful to identify the users/customers whose friendship/relationship should be a priority to keep. We provide formal definitions, algorithms and some experiments for this subject. Our experiments showed that the best current friend of a user is not necessarily the one who has the highest PageRank in the SN nor the one who has more friends.

Keywords: Social networks, friends, centrality measures, pagerank, graphs.

RESUMEN

Un campo de investigación en el área de las redes sociales (RSs) es la identificación de ciertos tipos de usuarios y de grupos. Para facilitar este proceso, una RS se suele representar mediante un grafo. Las medidas de centralidad, las cuales identifican los nodos más importantes en un grafo según algún criterio, suelen ser usadas para analizar un grafo. Una de estas medidas es el *PageRank* (una medida inicialmente concebida para clasificar las páginas web). Informalmente, en el contexto de las RSs, el *PageRank* de un usuario *i* representa la probabilidad de que otro usuario de la RS esté viendo la página de *i* luego de un tiempo considerable de navegación por la RS. En este artículo, se define un tipo de usuario en una RS: *el mejor amigo actual*. La idea es identificar, entre los amigos de *i*, quién es el amigo *k* que generaría el mayor decremento en el *PageRank* de *i*, si *k* dejara de ser amigo de *i*. Esto puede ser útil para identificar los usuarios/clientes cuya amistad/relación es prioritario conservar. En este artículo se presentan las definiciones formales, algoritmos y experimentos al respecto. Los experimentos demostraron que el mejor amigo actual de un usuario no es necesariamente aquel que tiene el mayor *PageRank* en la RS ni aquel que tiene más amigos.

Palabras clave: Redes sociales, amigos, medidas de centralidad, pagerank, grafos.

Received: April 30th 2014
Accepted: June 23th 2015

Introduction

Based on the relationships established by the members of a community, e.g., the users of a social network (SN), different types of users and user groups can be identified. For instance, with regard to users, leaders (Pedroche, 2010; Pedroche, 2012); best potential friends of a user (Moreno, Valencia, González, 2013); friends that show a distrust behavior (Ortega, 2012); and the efficient information spreaders (Kitsak *et al.*, 2010), among others, can be identified. With regard to groups, in (Pedroche, 2010) user groups that compete for visibility in a community are identified, and in (Masuda, Kurahashi, Onari, 2012) user groups with depressive and suicidal tendencies are analyzed.

To facilitate the identification and analysis of these types of users and user groups, the community of users and their relationships are usually represented by some mechanism. For example, a SN is usually represented by a graph. Usual tools to analyze a graph are the centrality measures (Masuda, Kurahashi, Onari, 2012), which identify the most important vertices in a graph. These include the *degree centrality* which measures the number of links of a node, the *closeness centrality* determined by the length of the shortest paths from one node to the rest of the nodes of the network, the *betweenness centrality* that is based on the total number of shortest paths that exist among all the pairs of nodes that pass through a node, and the PageRank. The PageRank is a measure originally designed to classify web pages (Page, Brin, Motwani, Winograd, 1999). Informally, the PageRank of a web page *p* represents the probability that a web surfer is visiting *p* after a long time of navigation in the web.

[1] Francisco Javier Moreno: Systems Engineer, Universidad de Antioquia. Ph.D Systems Engineering, Universidad Nacional de Colombia. Affiliation: Associate Porfessor Universidad Nacional de Colombia, Sede Medellín. E-mail: fjmoreno@unal.edu.co

[2] Santiago Hernández: Systems Engineer and Mathematician, Universidad Nacional de Colombia, Sede Medellín. E-mail: sahernandezt@unal.edu.co

How to cite: Moreno, F. J., & Hernández, S. (2015). An algorithm for identifying the best current friend in a social Network.*Ingeniería e Investigación, 35(2),* 80-88.
DOI: http://dx.doi.org/10.15446/ing.investig.v35n2.50339

In this paper, we define a new type of user in a SN based on the PageRank: the *best current friend* (BCF). Informally, our goal is to identify among the friends of a user *i*, who is the friend *k* that would generate the highest decrease in the PageRank of *i* if *k* stops being his/her friend. This may be useful to identify the users whose relationship (whether it be business, family or friends related) should be a priority to keep. These users are a key element for executives and for a company to get future customers. For instance, it is important for the sales executives to detect this type of users in order to keep and strengthen their relationships, e.g., to build customer loyalty, e.g., offering them extra benefits and customized services.

On the other hand, due to the changing relationships in a community (additions and deletions of both users and relationships), especially in a SN, the identification of the BCF (and other types of users) is a process that must be run whenever the number of changes in the users and in the relationships exceed a threshold established by the administrator of the SN. Indeed, the BCF of a node may change over time.

Due to the volume of users in SNs such as Facebook and Twitter (1.19 billion (Facebook Inc., 2013) and 237 million (Frier, & Spears, 2013) as of September 30, 2013 respectively), the calculation of the BCF for each node in these SNs may become a expensive computational process. For example, in our experimental environment, for a SN of 769 nodes, the calculation to generate the base matrix to identify the BCF took around two hours (see more details in the experiments section). Therefore, for SNs involving a large number of users (such as Facebook and Twitter), techniques such as sampling, pre-calculated and estimated data, parallel computing, among others must be used (Leskovec, Rajaraman, & Ullman, 2011; Bahmani, Chowdhury, & Goel, 2010; Lee, 2003).

The present paper is organized as follows: in Section 2, we present the basic elements of the PageRank algorithm; in Section 3, we formally introduce the concept of the BCF based on the PageRank; in Section 4, we present and analyse some experiments; in Section 5, we review some related works; and in Section 6, we conclude the paper and outline future work.

Basic definitions

The users and their relationships in a SN may be represented by a graph. For example, consider a SN with $n = 5$ users represented by a directed graph $GSN = (N, E)$, where N represents the set of nodes {1, 2, 3, 4, 5} and E the set of edges {(1, 2), (1, 3), (2, 1), (2, 3), (2, 4), (3, 1), (3, 4), (4, 2), (4, 5), (5, 2)}, see Figure 1. An edge (i, j) indicates that i is friend of j (i points to j). Note that this representation supports both unidirectional and bidirectional relationships, e.g., in Twitter a user w follows a user z but z not necessarily follows w, i.e., they show a unidirectional relationship.

Our goal is to classify the nodes (users) of a SN applying the PageRank method (Pedroche, 2010a; Pedroche, 2012; Page, Brin, Motwani, & Winograd, 1999). Note that in a SN it is reasonable to assume that each user points to at least one friend (i.e., an *outlink*). This is a mandatory condition to apply the PageRank method to SNs analysis, i.e., there must not be *dangling nodes* (Pedroche, 2010a).

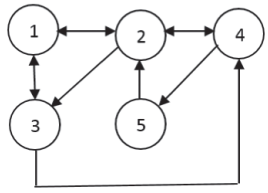

Figure 1. SN with five nodes, represented by a directed graph.

To apply the PageRank method, first we build a *connectivity matrix* $H = (h_{ij}) \in \mathbb{R}^{n\times n}$, $1 \le i, j \le n$, that represents the links of each node. If there exists a link from node i to node j, $i \neq j$, then $h_{ij} = 1$, otherwise $h_{ij} = 0$; if $i = j$ then $h_{ii} = 0$, see Figure 2.

$$H = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Figure 2. Connectivity matrix.

From H matrix we build the *row stochastic matrix* $P = (p_{ij}) \in \mathbb{R}^{n\times n}$, $1 \le i, j \le n$. A matrix is row stochastic if the sum of the elements of each of its rows is 1. P is calculated by dividing each element h_{ij} by the sum of the elements of row i of H, see Figure 3. Note that we assume that there do not exist dangling nodes then this sum (in each row) cannot be zero.

$$P = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Figure 3. Row stochastic matrix.

The PageRank method requires that the P matrix, in addition to being row stochastic, must be *primitive*. A non-negative square matrix is primitive (Varga, 2009) if the number of distinct eigenvalues of the matrix whose absolute value is equal to the spectral radius $\rho(P)$ is 1, where $\rho(P)$ is the maximum value (in absolute value) of its eigenvalues. In order to ensure this property (and still preserving the row stochastic property), we apply the following transformation (Page, Brin, Motwani, & Winograd, 1999; Pedroche, 2007): $G = \alpha P + (1-\alpha)ev^T$, where G is known as Google matrix; α is a damping factor, $0 < \alpha < 1$, and represents the probability with which the surfer of the network moves among the links of the H matrix; and (1- α) re-

presents the probability of the surfer to randomly navigate to a link which is not among the links of H. Note that if $\alpha=1$, then $G=P$, i.e., we would be working with the original P matrix. Usually, α is set to 0.85, a value that was established by Brin and Page, the creators of the PageRank method (Page, Brin, Motwani, & Winograd, 1999; Pedroche, 2007). In (Becchetti, & Castillo, 2006; Boldi, Santini, & Vigna, 2005; Boldi, 2005) the effect of several values of α is analyzed.

On the other hand, $e \in \mathbb{R}^{n\times 1}$ is the vector of all ones, and $v^T e=1$. v is called *personalization* or *teletransportation* vector and may be used to affect (to benefit or to harm) the ranking of the nodes of the network (Pedroche, 2007): $v=(v_i)\in \mathbb{R}^{n\times 1}: v_i>0$, $1\le i\le n$. Usually, $v=(1/n)$, and is known as the *basic personalization vector*. However, if we want to affect the ranking of a specific node i, v may be defined as follows: Let $0<\varepsilon<1$ then $v_i=(v_{ij})\in \mathbb{R}^{n\times 1}: v_{ij}=\varepsilon/(n-1)$ for $i\neq j$, $v_{ii}=1-\varepsilon$,. Thus, when e is close to zero, the ranking of node i tends to increase, but if ε is close to one, its ranking tends to decrease. A commonly used value in specialized literature for ε is 0.3 (Pedroche, 2010a).

Note that the constraint $v^T e=1$ allows us to define a v vector such that benefits (or harms) the ranking of several nodes simultaneously. For example, if $v=$ (7/20 7/20 1/10 1/10 1/10) then the ranking of nodes 1 and 2 tend to be benefited whereas the ranking of nodes 3, 4, and 5 tend to be harmed. We denote *PPR* (Personalized PageRank) as the PageRank of a node using some pre-scribed personalization vector v_i and we denote PR_i the PageRank vector computed using v_i.

In Figure 4, we show the G matrix which was computed with $\alpha=0.85$ and the basic personalization vector.

$$G=\begin{pmatrix} 0.0300 & 0.4550 & 0.4550 & 0.0300 & 0.0300 \\ 0.3133 & 0.0300 & 0.3133 & 0.3133 & 0.0300 \\ 0.4550 & 0.0300 & 0.0300 & 0.4550 & 0.0300 \\ 0.0300 & 0.4550 & 0.0300 & 0.0300 & 0.4550 \\ 0.0300 & 0.8800 & 0.0300 & 0.0300 & 0.0300 \end{pmatrix}$$

Figure 4. G matrix computed with α = 0.85 and the basic personalization vector.

From G matrix we can compute the *PageRank vecto*r π. To compute vector π we consider the following system of equations $\pi^T=\pi^T G$, where $\pi^T=[q_1\, q_2\, q_3\, q_4\, q_5]$. In addition, to ensure that π is a probability vector, we also consider the equation: $q_1+q_2+q_3+q_4+q_5=1$. For the running example, the system of equations is

$$0.03q_1+0.3133q_2+0.455q_3+0.03q_4+0.03q_5=q_1 \quad (1)$$
$$0.455q_1+0.03q_2+0.03q_3+0.455q_4+0.88q_5=q_2 \quad (2)$$
$$0.455q_1+0.3133q_2+0.03q_3+0.03q_4+0.03q_5=q_3 \quad (3)$$
$$0.03q_1+0.3133q_2+0.455q_3+0.03q_4+0.03q_5=q_4 \quad (4)$$
$$0.03q_1+0.03q_2+0.03q_3+0.455q_4+0.03q_5=q_5 \quad (5)$$
$$q_1+q_2+q_3+q_4+q_5=1 \quad (6)$$

We solved the system using MATLAB; results are showed in Table 1. The results show that node 2 has the highest PageRank whereas node 5 has the lowest one.

Table 1. PageRank vector π.

Node	PageRank	
1	0.1972	
2	0.2944	→ Highest ranking
3	0.1972	
4	0.1972	
5	0.1138	→ Lowest ranking

As a second example, we compute vector π with the personalization vector of node 3, i.e., PR_3 with $\varepsilon=0.3$, i.e., $v_3=$(0.08 0.08 0.7 0.08 0.08). The corresponding G matrix is showed in Figure 5.

$$G=\begin{pmatrix} 0.01125 & 0.43625 & 0.53000 & 0.01125 & 0.01125 \\ 0.29458 & 0.01125 & 0.38833 & 0.29458 & 0.01125 \\ 0.43625 & 0.01125 & 0.10500 & 0.43625 & 0.01125 \\ 0.01125 & 0.43625 & 0.10500 & 0.01125 & 0.43625 \\ 0.01125 & 0.86125 & 0.10500 & 0.01125 & 0.01125 \end{pmatrix}$$

Figure 5. G matrix computed with α = 0.85 and the personalization vector v_3, ε = 0.3.

The system of equations is:

$$0.01125q_1+0.29458q_2+0.43625q_3+0.01125q_4+0.01125q_5=q_1. \quad (7)$$
$$0.43625q_1+0.01125q_2+0.01125q_3+0.43625q_4+0.86125q_5=q_2. \quad (8)$$
$$0.53q_1+0.38833q_2+0.105q_3+0.105q_4+0.105q_5=q_3. \quad (9)$$
$$0.01125q_1+0.29458q_2+0.43625q_3+0.01125q_4+0.01125q_5=q_4. \quad (10)$$
$$0.01125q_1+0.01125q_2+0.01125q_3+0.43625q_4+0.01125q_5=q_5. \quad (11)$$
$$q_1+q_2+q_3+q_4+q_5=1. \quad (12)$$

The resulting PR_3 vector is showed in Table 2.

Table 2. PR_3 vector.

Node	PageRank	
1	0.1945	
2	0.2565	
3	0.2603	→ Highest ranking
4	0.1945	
5	0.0939	→ Lowest ranking

Note that node 3 improved its ranking with regard to the PageRank vector π of Table 1 (it changed from 0.1972 to 0.2603).

The BCF

We introduce the concept of the BCF of a node in a SN. The BCF of a node i is the node k of the SN, $k\neq i$, $H[k,i]=1$, such that if k stops being friend of i (k stops pointing to i), k is the node that generates the *highest decrease* in the PageRank of i. That is, let $GSN=(N, E)$ be the original graph that represents the SN. Let $\prod_i(GSN)$ denote the i component

of the *PPR* for some personalization vector v. Given $i \in N$, let: $Q(i) = \{ j \in N: i \neq j, (j, i) \in E \}$, i.e., the set of nodes that point to i. Let $E'(j, i) = E - \{(j, i)\}$, with some $j \in Q(i)$, i.e., the original set of edges E minus the edge from j to i, and let $GSN'(j, i) = (N, E'(j, i))$. Then we say that $k \in Q(i)$ is the BCF of i if the following condition holds: $\prod_i(GSN'(k,i)) = \min(\prod_i(GSN'(j,i)))$, $j \in Q(i)$.

We define the Current Friend PageRank Vector of a node i as follows: $CFPRV_i = \prod_i(GSN'(j, i))$, $1 \leq j \leq n$, $j \neq i$. Note that if $H[j, i] = 0$ (i.e., j is not friend of i) or if j cannot be disconnected (because in doing so, the sum of the elements of each row of H would be zero) then $CFPRV_i(j)$ = N/A (not applicable).

The next algorithm computes the $CFPRV_i$. Let $CFPRV_i(k)$, $1 \leq k \leq n$, $k \neq i$, be the minimum value in $CFPRV_i$, then k is the BCF of i.

Algorithm CFPRV($i, H: n \times n$).

Input: i: The node for which the *CFPRV* will be computed

H: Connectivity matrix

Preconditions: H is a matrix of order n, $n > 1 \wedge 1 \leq i n$
$\wedge (H[p,q] = 0 \vee H[p,q] = 1 \ \ \forall\, p, q,\ 1 \leq p, q \leq n)$
$\wedge (\forall\ p, 1 \leq p \leq n, \sum_{q=1}^{n} H[p,q] \neq 0)$

Output: *CFPRV*: Current Friend PageRank Vector of node i.

Postconditions: *CFPRV* is a vector of dimension n, $(n > 1)$ $\wedge (\forall p, 1 \leq p \leq n, 0 \leq CFPRV[p] \leq 1 \wedge CFPRV[p] = \text{N/A})$.

1. $j = 1$; //Variable to iterate through the nodes of the SN
2. **While** $j \leq n$ **Do**
3. **IF** $(H[j,i] = 1$ **AND** $numberofOutlinks(j) > 1)$ **THEN** /*$numberofOutlinks$() computes the number of outlinks of a node. If $numberofOutlinks(j) \leq 1$ then j cannot be disconnected because the sum of the elements of each row of H cannot be zero*/
4. $auxH = H$; //$auxH$ is a copy of H matrix
5. $auxH[j,i] = 0$; // j stops being friend of i
6. Compute *PageRank* vector using $auxH$ matrix
7. $CFPRV[j] = PageRank[i]$ /*Get PageRank of node i and store it in *CFPRV*/
8. **ELSE**

   $CFPRV[j] = N/A$; /*j does not point to i or cannot be disconnected from it*/
9. **END IF**
10. $j = j + 1$;
11. **END WHILE**
12. **RETURN** CFPRV

Next, we prove the correctness of our algorithm.

Loop invariant

For the j-th iteration $(1 \leq j \leq n)$, the dimension of *CFPRV* is j, and $\forall\, k,\ 1 \leq k \leq j$, $CFPRV[k]$ stores the PageRank of node i, when disconnecting node k (if possible) or *N/A* otherwise:

$(\text{CFPRV: } (j-1) \times 1) \wedge (j \leq n) \wedge (\forall\, k, 1 \leq k \leq j, \text{CFPRV}[k] = \text{PageRank}(auxH, i))$

Initialization

For $j = 1$, from the preconditions we know that $n > 1$, then $j < n$. Moreover, *CFPRV* has dimension 0.

Maintenance

For the j-th iteration $(1 < j < n)$, the dimension of *CFPRV* is $(j-1)$ and $j < n$. Considering the truth value of $(H[j,i] = 1$ AND $numberofOutlinks(j) > 1)$ there are two cases:

- If true then $CFPRV[j] = PageRank[i]$, then *CFPRV*: $j \times 1$.
- If false then $CFPRV[j]$ = N/A.

In any case $j = j + 1$, then $(CFPRV: j \times 1) \wedge (j + 1 \leq n)$.

Termination

In the last iteration (n-th), *CFPRV* has dimension $(n-1)$ and $j = n$. Again, considering the truth value of $(H[j, i] = 1$ AND $numberofOutlinks(j) > 1)$ there are two cases:

- If true then $CFPRV[j] = PageRank[i]$, then CFPRV: $j \times 1$.
- If false then $CFPRV[j]$ = N/A.

In any case $j = j + 1$, then (CFPRV: $j \times 1) \wedge (j + 1 > n)$ and because of the second condition of the invariant the loop ends.

Example. Consider the SN of Figure 1. Currently, the PageRank of node 2 (the node with the highest PageRank) is 0.2944. In Table 3, we show the change in the PageRank of this node depending on the node that has been disconnected ($CFPRV_2$). In this example, node 1 is the BCF of node 2 because if it is disconnected, it will be the node that decreases the most the PageRank of node 2. Note that node 2 currently has another connected node, node 5. However, if this node is disconnected, the PageRank method becomes inapplicable because node 5 will be left without any outlinks.

Table 3. $CFPRV_2$.

Node to be disconnected	PageRank of Node 2		
1	0.2149	→	BCF
2	N/A		
3	N/A		
4	0.2654		
5	N/A (cannot be disconnected		

Using our *CFPRV* algorithm, we can create the Current Friend PageRank Matrix (*CFPRM*), i.e., we compute the *CFPRV* for each user of the SN, as we show in the following algorithm:

Algorithm $CFPRM(H: n \times n)$.

Input: H: Connectivity matrix.

Preconditions: H is a matrix of order n, $n>1$ $(H[p,q]=0 \vee H[p,q]=1 \ \forall p,q, 1\le p,q\le n)$ $\wedge$ $(\forall p, 1\le p\le n, \sum_{q=1}^{n} H[p,q]\neq 0)$

Output: CFPRM = Current Friend PageRank Matrix

Postconditions: *CFPRM* is a matrix of order n, $\wedge(\forall p,q 1\le p,q\le n, 0\le CFPRV[p,q]\le 1 \wedge CFPRV[p,q]=$ N/A).

1. **For** $i = 1$ **to** n
2. *CFPRM* [i] = *CFPRV*(i, H)
3. **End for**
4. **Return** *CFPRM*

Since the length of this article is limited, we do not present a complete correctness proof of this algorithm. The main part is the loop invariant: For the j-th iteration ($1\le j\le n$), the order of *CFPRM* is $n\times(j-1)$, and $\forall k, 1\le k\le j$, *CFPRM*[k] is the result of computing *CFPRV*(k, H): (*CFPRM* $n\times(j-1)$) $\wedge$ ($j\le n$) $\wedge$ ($\forall k, 1\le k\le j$, *CFPRM*[k] = *CFPRV*(k, H)).

In Table 4, we show the *CFPRM* matrix for the SN of Figure 1. For example, the BCF of the node 4 is the node 3.

Table 4. *CFPRM*. The BCF of each node is shaded.

Node to be disconnected (j)	Node of interes (i)				
	1	2	3	4	5
1	N/A	0.2149	0.1347	N/A	N/A
2	0.1177	N/A	0.1225	0.1371	N/A
3	0.1181	N/A	N/A	0.1067	N/A
4	N/A	0.2654	N/A	N/A	0.030
5	N/A	N/A (Cannot be disconnected)	N/A	N/A	N/A

Experiments

The algorithms were implemented in MATLAB. The experiments were carried out for a real subnetwork of 769 nodes of Facebook, Caltech-2005 (Traud, Kelsic, Mucha, & Porter, 2011). The calculation of the *CFPRM* matrix was executed in a laptop with an Intel Core I7 processor with 4 GB of RAM memory and the calculation took almost two hours.

Experiment 1: the analyses were focused on two groups of nodes: the 10 with the highest and the lowest PageRank in the SN. To determine the effect of removing the BCF from a node, the node with higher PageRank in the SN was initially chosen, the node identified with number 623. In Figure 6 we show the effect on the PageRank of the node 623 when removing its BCF (the node identified with number 104), and the following nine nodes that decrease the most its PageRank.

Results: note that if the node 623 loses its BCF (node 104), its PageRank falls from 0.006683691154 to 0.0066220198284 (e.g., a difference of 0.000061671), whereas if it loses the tenth node (node 143) that most decreases its PageRank, its PageRank falls to 0.006644868 (i.e., a difference of 0.0000388). While these differences are in the order of 1E-05, they may be significant (Leskovec, Rajaraman, & Ullman, 2011). In fact, such a decrease in the PageRank could cause that a node ceases to be the "leader" of the SN (i.e., it will no longer be the node with the highest PageRank in the SN). In Figure 7, the 10 nodes with the highest PageRank in the SN are shown, indicating their PageRank when they lose their respective BCF.

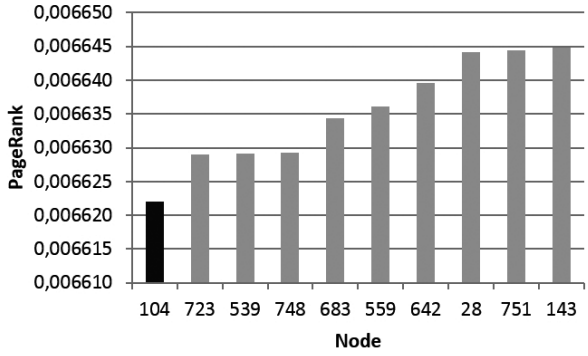

Figure 6. The 10 nodes that decrease the most the PageRank of node 623.

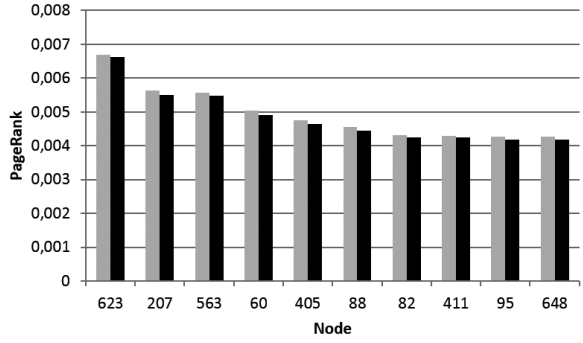

Figure 7. The 10 nodes with the highest PageRank in the SN (in gray) and their PageRank after losing their BCF (in black).

In our SN, the node with the highest PageRank (node 623) has a significant advantage regarding the PageRank of the following nodes with the highest PageRank in the SN. For example, the difference with the second node with higher PageRank (node 207) is 0.00104413; for this reason, the decrease that node 623 suffers when losing its BCF is not enough to lose its position as the leader of the SN.

However, in other nodes, the loss of their BCF may change their ranking (according to their PageRank) in the SN. For instance, consider node 82, which is currently in the seventh position in the SN according to its PageRank. If it looses its BCF, itsPageRank will be lower than the one of node 411 (which is currently in the eighth position in the SN).

Experiment 2: In a second experiment, we compared the PageRank of the 10 nodes that decrease the most the PageRank of node 623 with the PageRank of the 10 nodes that decrease the least its PageRank. The results are shown in Figure 8.

Results: It is interesting to note that the nodes that decrease the most the PageRank of node 623 have a very low PageRank compared to the ones that decrease the least their PageRank.

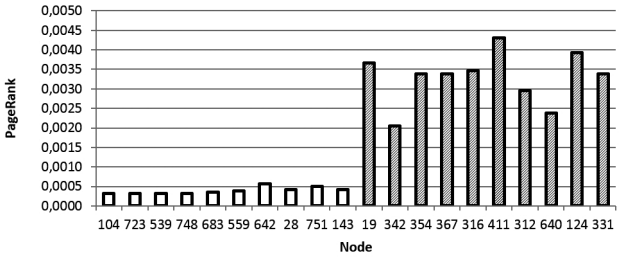

Figure 8. PageRank of the 10 most and least decreasing nodes of the PageRank of node 623, in white and grey respectively.

Experiment 3: next, we found the BCF for each node of the SN, and we determined which nodes appeared more frequently as the BCF of other nodes. In addition, the PageRank of these nodes and their ranking in the SN according to their PageRank were calculated. The results are shown in Table 5.

Table 5. Nodes which appear more frequently as BCFs.

Node	Number of times that it is the BCF	PangeRank	Ranking in the SN according to its PangeRank	Number of outlinks
250	9	0.00063477	567	16
481	9	0.00059482	579	12
195	8	0.00081814	502	12
269	8	0.00081281	508	15
453	8	0.00190016	152	45
767	8	0.0004629	647	11
714	7	0.00036968	681	7
761	7	0.00045157	650	9
22	6	0.00073286	536	18
…	…	…	…	…

Results: note that the nodes which appear more frequently as BCFs of other nodes have few outlinks and their PageRank is low in comparison to other nodes in the SN. For instance, the nodes identified with numbers 250 and 481 are the BCFs of nine nodes in the SN; and according to their PageRank, these two nodes occupy the positions 567 and 579 in the SN, respectively. This suggests, at least in this SN, that the nodes that appear more frequently as BCFs tend to occupy lower positions in the SN (according to their PageRank), but they are *significant for the PageRank of other nodes*.

A possible explanation for this behavior is shown in Figure 9. As the PageRank represents the probability to reach a node *j* after a long time *t* of navigation in the SN, and if *k* is a node which is connected to *j*, then the more outlinks *k* has, the less will be the probability to reach each of them. Therefore, any path containing *k* and ending in *j* will have a lower probability *as the number of* outlinks of *k* increases due to the rule of the product for probabilities. Informally, this means that the probability to reach a node *j* from *k* will be higher if *k* has fewer outlinks.

Experiment 4: In our last experiment, we compared our proposed method (based on PageRank) to find the BCF of a node and the following alternative (based on degree centrality). Our goal was to analyze how the PageRank of a node is affected when losing its BCF or its node (friend) with the greatest number of outlinks. To this end, we selected the 10 nodes of the SN with the highest PageRank. For each of these nodes *i,* we found the node *j* (connected to *i*), which had the greatest number of outlinks. In s 6 we also show the PageRank of each node *i if it lost the friendship of j* (*j* stops being friend of *i*), the PageRank of each node *i* if it lost its BCF, and the difference between those two values.

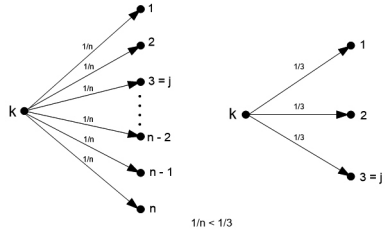

Figure 9. Probabilistic interpretation of the results from Table 5.

Table 6. BCF methods: PageRank and degree centrality.

Node (i)	j	Number of outlinks of j	PageRank of j	PageRank of i after losing j(1)	PageRank of i after losing its BCF (2)	(1) - (2)
623	563	203	0.006683691	0.006660161	0.00662202	0.00003814
207	411	172	0.005639558	0.005617879	0.00549444	0.00012344
563	623	248	0.005565427	0.005542166	0.00548848	0.00005369
60	563	203	0.00504206	0.005018149	0.00491077	0.00010738
405	563	203	0.004753378	0.004729867	0.00464953	0.00008034
88	623	248	0.004555629	0.004532099	0.00445001	0.00008209
82	623	248	0.004319919	0.004296752	0.00424259	0.00005416
411	623	248	0.004300583	0.004277415	0.00424577	0.00003164
95	623	248	0.004278484	0.004255218	0.00417605	0.00007917
648	623	248	0.004270205	0.004246928	0.00419042	0.00005651

Results: as expected, the results evidenced that from the point of view of the PageRank, the loss of a friend with many outlinks is less "severe" than the loss of the BCF. Note again that although the differences (last column of Table 6) are in the order of 1E-05, these could mean the loss of positions of a node in the SN according to their PageRank.

Related work

As for related works, we have identified the following. In (Moreno, Valencia, & González, 2013), the authors present a complementary work with regard to ours. They identified the best *potential* friend of a node *i*. Informally, the best potential friend of a node *i* is the node *k* that when linked to *i*, it provides the highest increase in the PageRank of *i*. In (Kostakos, *et al.*, 2011) the authors conduct a study that aims to answer two questions: i) given a group of users and their social graph, is it possible to predict who among them is likely to *reveal* most about the whereabouts of anyone in the group? And ii) given a user, is it possible to predict who among his/

her friends *knows* most about his/her whereabouts? In (Forestier, Stavrianou, Velcin, & Zighed, 2012), the authors present a survey about the identification of roles in the SN based on the structure of the network, the behavior of the users, and the analysis of the contents that they publish (twits, posts). This enables to identify beginner and expert users, controversial and influential users, allies and political enemies, among others. Al-Oufi, Kim, and El Saddik (2012) proposes a measure (based on the *Advogato trust metric* (http://www.advogato.org/trust-metric.html), a trust metric for attack resistance) to identify user groups based on trust levels; so that it may be possible for each user to identify, among all the friends, those who are reliable and those whom they should not share information with (e.g., untrusted users), as well as suggested users (*potential friends*) who may contribute to strengthen trust relationships. In (Ball, & Newman, 2013), the authors analyze how users select their friends. For instance, they analyzed the tendency users have to select friends who belong to their same status or category. In (Grieve, 2013), the authors developed a recommendation system for the user to find a partner and/or friends based on inferred information from his/her preferences, connections, and other aspects. (García-Barriocanal and Sicilia, 2005) propose a metric of social relevance called *PeopleRank*. Their metric is based on the explicitly social declared relationships expressed using the FOAF-like vocabulary. FOAF (Friend Of A Friend (http://www.foaf-project.org), (Golbeck, & Rothstein, 2008)) is an ontology describing persons, their activities, and their relations to other people. Then, *PeopleRank* is used as a weighting factor for the PageRank algorithm, i.e., they propose a "socially weighted" version of the original PageRank. On the other hand, Ahmedi (2012) claims that "*FOAF alone is yet insufficient to model social networks for ranking people on the Web*". He proposes a model (called AuthorRank+FOAF), which extends FOAF with PageRank and AuthorRank metrics (AuthorRank (Liu, *et al.*, 2005) is a version of PageRank which considers the weight of co-authors links when ranking) in order to compute the reputation of authors (according to Google, this is a key element for improving the ranking of pages). Ahmedi's extension relies on his earlier work (Ahmedi, Abazi-Bexheti, & Kadriu, 2011), which already extended FOAF into CO-AUTHORONTO, but aimed to capture the semantics of weighted co-authorship networks (Nascimento, Sander, & Pound, 2003). Another algorithm (that is also called AuthorRank) to rank people based on FOAF and DBLP data is presented in (Ding, *et al.*, 2006). Their work combines people ranking with co-citation analysis (Jeong, Song, & Ding, 2014). In Table 7, we present a brief overview of the related works. To the best of our knowledge, there is no work to date that defines the BCF.

Table 7. Overview of the related works

Ref.	Metrics considered for ranking	Advantages/Disadvantages
Moreno, Valencia, & González, 2013	PageRank	They define the best potential friend of a user. They focus on predictions: who to ask regarding whereabouts of somebody.
Kostakos, et al., 2011	Degree rank (based on common ties among individuals) and trust rank (how much users know about other users).	They do not consider semantic elements. In real life, privacy issues may affect the trust rank.
Forestier, Stavrianou, Velcin, and Zighed, 2012	N.A.	They present a typology of social roles. Their work is a survey regarding the identification of roles in SNs.
Al-Oufi, Kim, and El Saddik, 2012	An extension of the Advogato trust metric.	They identify user groups based on trust levels. They do not consider other semantic elements.
Grieve, 2013	N.A.	He proposes a recommendation system: social data is sent into a neural network to predict successful connections. He does not rank users.
García-Barriocanal and Sicilia (2005)	PeopleRank	They propose a "socially weighted" version of the original PageRank, It depends on FOAF data.
Ahmedi (2012)	PageRank and AuthorRank	He focus on compute the reputation of authors. He does not consider other semantic elements.
Ding, et al., 2006	PageRank and co -citation analysis	They obtained a combined ranking of different data sets. It depends on FOAF and DBLP data.

Conclusions and future work

In this paper, a new type of user of a SN, the BCF, was formally defined. Based on the PageRank, the most important friend of a user *i* was determined. This friendship is the most important one to keep since in case it gets lost, this would seriously affect the PageRank of *i*. The identification of the BCF could be decisive for the user when keeping the visibility and influence in the SN. In addition, we presented a corresponding algorithm to identify it as well as its correctness. Although experiments with SNs with a larger number of users and with other SNs such as Twitter (it was tested with a Facebook subnetwork of 769 users) are required, the results evidenced, e.g., that the BCF of a node is not necessarily the one, among all the friends, that has the highest PageRank or the one who has more friends.

As future work, we consider the following: to define the BCF in terms of other measures (e.g., those mentioned in the related work section), to compare the results among them, and to determine correlations if there is any. For instance, if a node *k* is the BCF of a node *i* when considering a measure *c* then, how close is *k* to be the BCF of *i* if another measure were considered? Another future work could be the development of a visual tool that allows the analysts to identify, in a friendly way, the BCF of each node, and that also allows them the interactive manipulation of the SN (addition and deletion of nodes/relationships), and that shows the way the BCF of each node is affected given these changes. This could contribute to the understanding of how

the relationships of other users of the SN affect a node *i* and to its corresponding BCF. At the same time, this could lead to the identification of "the best external friendship" with regard to a node *i*, i.e., among all the *couples of friends* in a SN (couples that do not include *i*), which is the couple that generate the highest decrease in the PageRank of *i* if this couple fell out. A third possible future work is to rank users with regard to their behavior or sentiments (Liu, 2015). For example, we could define the BCF of a user *i* based on sentiments, i.e, who is the friend *k* that would generate the highest decrease in the level of happiness of *i* if *k* stops being his/her friend.

References

Ahmedi, L., Abazi-Bexheti, L., & Kadriu, A. (2011). A Uniform Semantic Web Framework for Co-authorship Networks. *IEEE 9th International Conference on Dependable, Autonomic and Secure Computing* (pp. 958-965). Sydney, Australia. DOI: 10.1109/dasc.2011.159.

Ahmedi, L. (2012). AuthorRank + FOAF: ranking for co-authorship networks on the web. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 315-321). Istanbul, Turkey.

DOI: 10.1109/asonam.2012.60.

Al-Oufi, S., Kim, H. N., & El Saddik, A. (2012). A group trust metric for identifying people of trust in online social networks. *Expert Systems with Applications*, *39*(18), 13173–13181. DOI: 10.1016/j.eswa.2012.05.084.

Bahmani, B., Chowdhury, A., & Goel, A. (2010). Fast incremental and personalized PageRank. *VLDB Endowment*, *4*(3), 173–184. DOI: 10.14778/1929861.1929864.

Ball, B., & Newman, M. E. J. (2013). Friendship networks and social status. *Network Science*, *1*(01), 16–30.

DOI: 10.1017/nws.2012.4.

Becchetti, L., & Castillo, C. (2006). The distribution of PageRank follows a power-law only for particular values of the damping factor. *15th international conference on World Wide Web* (pp. 941–942). Edinburgh, UK.

DOI: 10.1145/1135777.1135955.

Boldi, P., Santini, M., & Vigna, S. (2005). PageRank as a function of the damping factor. *14th international conference on World Wide Web* (pp. 557–566). Chiba, Japan.

DOI: 10.1145/1060745.1060827.

Boldi, P. (2005). TotalRank: ranking without damping. *Special interest tracks and posters of the 14th international conference on World Wide Web* (pp. 898–899). Chiba, Japan. DOI: 10.1145/1062745.1062787.

Ding, Y., Scharffe, F., Harth, A., & Hogan, A. (2006). AuthorRank: Ranking Improvement for the Web. *International Conference on Semantic Web and Web Services Conference* (pp. 1-7). Las Vegas, USA.

Facebook Inc. (30, Oct, 2013). *Facebook reports third quarter 2013 results – Facebook*. Retrieved from: http://investor.fb.com/releasedetail.cfm?ReleaseID=802760. [Accessed: 14-Nov-2013].

Forestier, M., Stavrianou, A., Velcin, J., & Zighed, D. A. (2012). Roles in social networks: Methodologies and research issues. *Web Intelligence and Agent Systems: An international Journal*, *10*(1), 117–133.

Frier, S., & Spears, L. (7, Nov, 2013). *Twitter surges in trading debut after $1.82 billion share sale - Bloomberg*. Retrieved from: http://www.bloomberg.com/news/2013-11-07/twitter-raises-1-82-billion-pricier-value-than-facebook.html.

García Barriocanal, E., & Sicilia, M. A. (2005). Filtering information with imprecise social criteria: A FOAF-based backlink model. *4th conference of the European Society for Fuzzy Logic and Technology* (pp. 1094-1098). Barcelona, Spain.

Golbeck, J., & Rothstein, M. (2008). Linking Social Networks on the Web with FOAF: A Semantic Web Case Study. *23th AAAI Conference on Artificial Intelligence* (pp. 1138-1143). Chicago, USA.

Grieve, A. (2013). *Predicting likelihood of a successful connection between unconnected users within a social network using a learning network*. U.S. Patent No. 8595167 (B1). Mountain View, CA: U.S. Patent and Trademark Office.

Jeong, Y. K., Song, M., & Ding, Y. (2014). Content-based author co-citation analysis. *Journal of Informetrics*, 8(1), 197–211. DOI: 10.1016/j.joi.2013.12.001.

Kitsak, M., Gallos, L.K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., & Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature Physics*, *6(11)*, 888-893. DOI: 10.1038/nphys1746.

Kostakos, V., Venkatanathan, J., Reynold, B., Sadeh, N., Toch, E., Shaikh, S. A. & Jones, S. (2011). Who`s your best friend? Targeted privacy attacks in location-sharing social networks. *ACM Conference on Ubiquitous Computing* (pp. 177-186). Beijing, China.

Lee, C. P. (2003). A fast two-stage algorithm for computing PageRank and its extensions. [*Technical report*], 1-9.

Leskovec, J., Rajaraman, A., & Ullman, J.D. (2011). *Mining of Massive Datasets*. New York, USA: Cambridge University Press.

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge, UK: Cambridge University Press. DOI: 10.1017/CBO9781139084789.

Liu, X., Bollen, J., Nelson, M. L., & Van de Sompel, H. (2005). Coauthorship networks in the digital library research community. *Information Processing and Management*, 41(6), 1462-1480. DOI: 10.1016/j.ipm.2005.03.012.

Masuda, N., Kurahashi, I., & Onari, H. (2013). Suicide Ideation of Individuals in Online Social Networks. *PLoS ONE*, *8*(4), e62262.

Moreno, F., Valencia, A., & González, A. (2013). My best potential friend in a social network. In Tu, X.M., White, A.M. and Lu, N. (Ed), *Social Networking: Recent Trends, Emerging Issues and Future Outlook*, 113–124.

Nascimento, M. A., Sander, J., & Pound, J. (2003). Analysis of SIGMOD's co-authorship graph. *SIGMOD Record*, 32(3), 8-10. DOI: 10.1145/945721.945722.

Ortega, F.J., Troyano, J.A., Cruz, F.L., Vallejo, C.G., & Enríquez, F. (2012). Propagation of trust and distrust for the detection of trolls in a social network. *Computer Networks*, 56 (12), 2884–2895. DOI: 10.1016/j.comnet.2012.05.002.

Page, L., Brin, S., Motwani, R., & Winograd, T. (11, Nov, 1999). *The PageRank citation ranking: bringing order to the web*. Retrieved from: http://ilpubs.stanford.edu:8090/422/.

Pedroche, F. (2010a). Ranking nodes in social network sites using biased PageRank. *2º Encuentro de Álgebra Lineal Análisis Matricial y Aplicaciones, ALAMA-2010. Universidad Politécnica de Valencia* (pp. 1-7). Valencia, Expaña.

Pedroche, F. (2010b). Competitivity Groups on Social Network Sites. *Mathematical and Computer Modelling*, 52(7–8), 1052–1057. DOI: 10.1016/j.mcm.2010.02.031.

Pedroche, F. (2012). A model to classify users of social networks based on PageRank. *International Journal of Bifurcation and Chaos*, *22* (7), 1-14. DOI: 10.1142/s0218127412501623.

Pedroche, F. (2007). Métodos de cálculo del vector PageRank. *SeMA Boletín de la Sociedad Española de Matemática Aplicada*, 9(39), 7–30.

Traud, A.L., Kelsic, E.D., Mucha, P.J., & Porter, M.A. (2011). Comparing community structure to characteristics in online collegiate social networks, *SIAM Review*, *53*(3), 526-543. DOI: 10.1137/080734315.

Varga, R. S. (2009). *Matrix iterative analysis*. Berlin: Springer.