

Document downloaded from:

<http://hdl.handle.net/10251/36766>

This paper must be cited as:

García, E.; Pedroche Sánchez, F.; Romance, M. (2012). On the localization of the personalized PageRank of complex networks. *Linear Algebra and its Applications*. 439(3):1-13. doi:10.1016/j.laa.2012.10.051.



The final publication is available at

<http://dx.doi.org/10.1016/j.laa.2012.10.051>

Copyright Elsevier

On the Localization of the Personalized PageRank of Complex Networks

E. García¹, F. Pedroche² and M. Romance^{1,3}

¹ Departamento de Matemática Aplicada
ESCET - Universidad Rey Juan Carlos
C/ Tulipán s/n 28933 Móstoles (Madrid), Spain.
esther.garcia@urjc.es, miguel.romance@urjc.es

² Institut de Matemàtica Multidisciplinària
Universitat Politècnica de València
Camí de Vera s/n. 46022 València, Spain.
pedroche@mat.upv.es

³ Centro de Tecnología Biomédica
Universidad Politécnica de Madrid
28223 Pozuelo de Alarcón (Madrid), Spain.

Abstract

In this paper new results on personalized PageRank are shown. We consider directed graphs that may contain dangling nodes. The main result presented gives an analytical characterization of all the possible values of the personalized PageRank for any node. We use this result to give a theoretical justification of a recent model that uses the personalized PageRank to classify users of Social Networks Sites. We introduce new concepts concerning competitiveness and leadership in complex networks. We also present some theoretical techniques to locate leaders and competitors which are valid for any personalization vector and by using only information related to the adjacency matrix of the graph and the distribution of its dangling nodes.

keywords Google matrix, PageRank, link analysis, social networking.
MSC 91D30, 15A21, 60K15, 90B15, 94C15.

1 Introduction

Much effort has been done in some aspects related to PageRank and its applications since the introduction of the PageRank algorithm to rank pages

on the web [11]. We are interested in the use of the so-called personalization vector to bias the PageRank to some nodes. We refer the reader to [10] and [3] for the theoretical basis of the PageRank algorithm.

The idea of biasing the PageRank vector using a personalization vector was, in fact, suggested originally in [11]. The first time that someone uses the personalization vector to bias to some topics appears in [6]. In [7] some different ways of biasing the PageRank with personalization vectors are summarized. In [5] the authors propose to use the personalization vector to bias the PageRank to pages that were visited more frequently by previous users. To reduce computational complexity the usual strategies consist in taking low rank approximations [15] or to decompose into subgraphs [9], [16]. Another approach consists in using Monte Carlo methods to compute only the top-k Personalized PageRank [1].

As a centrality measure, Personalized PageRank can also be used to classify users in Social Network Sites [12], [13], [14]. A way of using the Personalized PageRank to rank nodes in an SNS is by using the personalization vector to incorporate features of the users. These features can be popularity (e.g., the number of friends or followers of the user), activity (e.g., the number of actions made by the user) and recentness (e.g., the up-to-date of the actions of the user in the SNS). In [12], the fundamentals of a model that uses the Personalized PageRank to rank users in SNS were presented. This model left open some theoretical questions. One of these questions is to what extent one can use the personalization vector to modify the PageRank vector. In this paper we address this question and we show, in particular, that the component i of the PageRank vector is bounded, being this bound a sharp one, and valid for any personalization vector. We also derive some theoretical properties that let us detect nodes that compete with each other to gain PageRank. We also give some generalizations to some definitions introduced in [12].

The structure of the paper is as follows. In section 2 the basic definitions and results used in the rest of the paper are presented. In addition to this, a technical general lemma about row stochastic matrices is proved in order to provide the tools of the results of the following sections. Section 3 is devoted to prove the main result of the paper that locates all the possible values of the personalized PageRank for each node of a network. Finally, section 4 presents several applications of the localization theorem proved before. The applications include an analytical result that gives necessary and sufficient conditions for the competitiveness between nodes and a characterization of the leadership of nodes in a complex network. These analytical results give some easy algorithms to locate leaders and competitors which are valid for any personalization vector and only use information related to the adjacency matrix of the graph and the distribution of its dangling nodes. In this final section several examples are presented in order to illustrate the results proved.

2 Some definitions and a technical lemma

Let $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ be a directed graph where $\mathcal{N} = \{1, 2, \dots, n\}$ and $n \in \mathbb{N}$. Note that all the results presented in this paper deal with directed networks, but they can be straightforwardly stated for un-directed networks. The link (i, j) belongs to the set \mathcal{E} if and only if there exists a link connecting node i to node j . The *adjacency matrix* of \mathcal{G} is an $n \times n$ -matrix

$$A = (a_{ij}) \text{ where } a_{ij} = \begin{cases} 1, & \text{if } (i, j) \text{ is a link of } \mathcal{G} \\ 0, & \text{otherwise.} \end{cases}$$

A link (i, j) is said to be an *outlink* for node i and an *inlink* for node j . We denote $k_{out}(i)$ the *outdegree* of node i , i.e., the number of outlinks of a node i . Notice that $k_{out}(i) = \sum_k a_{ik}$. The graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ may have *dangling nodes*, which are nodes $i \in \mathcal{N}$ with zero outdegree.

Let $P = (p_{ij}) \in \mathbb{R}^{n \times n}$ be the *row stochastic matrix* associated to \mathcal{G} defined in the following way:

- if i is a dangling node, $p_{ij} = 0$ for all $j = 1, \dots, n$,
- otherwise, $p_{ij} = \frac{a_{ij}}{k_{out}(i)} = \frac{a_{ij}}{\sum_k a_{ik}}$.

Vectors of \mathbb{R}^n will be denoted by column matrices. In particular,

$$\begin{aligned} \mathbf{e}_1 &= (1, 0, \dots, 0)^T \in \mathbb{R}^n, \\ &\dots \\ \mathbf{e}_n &= (0, \dots, 0, 1)^T \in \mathbb{R}^n, \\ \mathbf{e} &= \mathbf{e}_1 + \dots + \mathbf{e}_n = (1, \dots, 1)^T. \end{aligned}$$

The i^{th} -component of a vector $\mathbf{v} = (v_1, \dots, v_n)^T \in \mathbb{R}^n$ is given by the product $\mathbf{v}^T \mathbf{e}_i = v_i$, and the sum of the components of the vector $\mathbf{v} \in \mathbb{R}^n$ is 1 if $\mathbf{v}^T \mathbf{e} = 1$. Moreover, we will say that $\mathbf{v} > 0$ if all the components v_i of \mathbf{v} are greater than zero, i.e., $\mathbf{v}^T \mathbf{e}_i > 0$, $i = 1, \dots, n$.

We will use the *personalized PageRank vector*. The ingredients to build such PageRank vector [11] are:

- A *damping factor* $\alpha \in (0, 1)$.
- A *distribution of dangling nodes* $\mathbf{u} \in \mathbb{R}^n$ such that $\mathbf{u} > 0$ and $\mathbf{u}^T \mathbf{e} = 1$. The dangling nodes will be characterized by a vector $\mathbf{d} \in \mathbb{R}^n$ defined as $\mathbf{d} = (d_1, \dots, d_n)^T$, where

$$d_i = \begin{cases} 1, & \text{if } i \text{ is a dangling node of } \mathcal{G} \\ 0, & \text{otherwise.} \end{cases}$$

- A *personalization vector* $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{v} > 0$ and $\mathbf{v}^T \mathbf{e} = 1$.

The *Google matrix* $G = G(\alpha, \mathbf{u}, \mathbf{v})$ with dangling nodes and personalized vector \mathbf{v} is defined as

$$G = \alpha(P + \mathbf{d}\mathbf{u}^T) + (1 - \alpha)\mathbf{e}\mathbf{v}^T \in \mathbb{R}^{n \times n} \quad (2.1)$$

(this matrix G is row-stochastic, i.e., $G\mathbf{e} = \mathbf{e}$). The *PageRank vector* $\pi = \pi(\alpha, \mathbf{u}, \mathbf{v})$ is the unique eigenvector of G^T associated to eigenvalue 1 such that $\pi^T \mathbf{e} = 1$, i.e., $\pi > 0$, $\pi^T \mathbf{e} = 1$ and $\pi^T G = \pi^T$ (see [11]).

From now on we will consider a fixed damping factor α (usually $\alpha = 0.85$) and a fixed distribution of dangling nodes \mathbf{u} , so the PageRank matrix G and the PageRank vector will only depend on the personalization vector \mathbf{v} we are considering.

Since $\pi^T G = \pi^T$, from the definition of G in (2.1) and the fact that $\pi^T \mathbf{e} = 1$ we get that

$$\begin{aligned} \pi^T &= \pi^T G = \pi^T (\alpha(P + \mathbf{d}\mathbf{u}^T) + (1 - \alpha)\mathbf{e}\mathbf{v}^T) \\ &= \alpha\pi^T(P + \mathbf{d}\mathbf{u}^T) + (1 - \alpha)\pi^T \mathbf{e}\mathbf{v}^T \\ &= \alpha\pi^T(P + \mathbf{d}\mathbf{u}^T) + (1 - \alpha)\mathbf{v}^T \end{aligned}$$

so $\pi^T(I_n - \alpha P - \alpha \mathbf{d}\mathbf{u}^T) = (1 - \alpha)\mathbf{v}^T$, where $I_n \in \mathbb{R}^{n \times n}$ is the identity matrix. Therefore it was shown in [3] that

$$\pi^T = (1 - \alpha)\mathbf{v}^T(I_n - \alpha(P + \mathbf{d}\mathbf{u}^T))^{-1}. \quad (2.2)$$

We will denote by X the $n \times n$ -matrix appearing in formula (2.2) above

$$X = (1 - \alpha)(I_n - \alpha(P + \mathbf{d}\mathbf{u}^T))^{-1}, \text{ so } \pi^T = \mathbf{v}^T X.$$

Notice that formula (2.2) indicates that the PageRank of each node can be expressed as a function of the personalization vector \mathbf{v} since $\pi = \pi(\mathbf{v}) = X^T \mathbf{v}$ (see [3]). Notice that this equality makes sense for all $\mathbf{v} \in \mathbb{R}^n$ and gives the PageRank when $\mathbf{v} > 0$ and $\mathbf{v}^T \mathbf{e} = 1$.

It is easy to check that the matrix $P_{\mathbf{u}} = P + \mathbf{d}\mathbf{u}^T$ appearing in (2.2) is a row-stochastic matrix since

$$P_{\mathbf{u}}\mathbf{e} = (P + \mathbf{d}\mathbf{u}^T)\mathbf{e} = P\mathbf{e} + \mathbf{d}\mathbf{u}^T\mathbf{e} = P\mathbf{e} + \mathbf{d} = \mathbf{e}.$$

In the next section we will use the following lemma dealing with row-stochastic matrices as $P_{\mathbf{u}}$:

Lemma 2.3. *Let Q be a row-stochastic matrix and $\alpha \in (0, 1)$. Then the matrix $Y = I_n - \alpha Q$ is strictly row-diagonally-dominant, $X = (1 - \alpha)Y^{-1}$ is strictly diagonally-dominant of its column entries and the maximum of each column i of X is achieved in x_{ii} .*

Proof. Clearly $Q\mathbf{e} = I_n\mathbf{e} - \alpha Q\mathbf{e} = (1 - \alpha)\mathbf{e}$, i.e., the sum of the entries of each row of Q is $1 - \alpha$. Therefore, since $\alpha \in (0, 1)$ and $0 \leq q_{ik} \leq 1$ for all $i, k = 1, \dots, n$, we get that

$$|y_{ii}| = |1 - \alpha q_{ii}| = 1 - \alpha q_{ii} = 1 - \alpha + \alpha \sum_{k \neq i} q_{ik} > \alpha \sum_{k \neq i} q_{ik} = \sum_{k \neq i} |y_{ik}|,$$

i.e., Y is strictly row-diagonally-dominant. Now, by Theorem 2.5.12 in [8], Y^{-1} and $X = (1 - \alpha)Y^{-1}$ are strictly diagonally-dominant of their column entries and therefore for every $k \neq i$

$$|x_{ii}| > |x_{ki}|.$$

Moreover, since Y is a (nonsingular) M-matrix (see, for example, [2]) we have that $Y^{-1} \geq 0$. Hence the absolute values in the formula above can be deleted and we get

$$\max_k x_{ki} = x_{ii}.$$

□

3 Main result: Location of Personalized PageRank

The main contribution of this paper is the solution to the following problem:

Problem. Given a graph \mathcal{G} with dangling nodes indicated by some vector \mathbf{d} , a fixed damping factor $\alpha \in (0, 1)$ and fixed dangling nodes distribution \mathbf{u} , is there an easy way to locate all the possible values of the PageRank for each node i ?

Definition 3.1. Given a graph \mathcal{G} with dangling nodes indicated by some vector \mathbf{d} , a fixed damping factor $\alpha \in (0, 1)$ and fixed dangling nodes distribution \mathbf{u} , for each node $i \in \mathcal{N}$ we define $\mathcal{PR}(i)$ as the set of all possible values of Personalized PageRank of node i , i.e.,

$$\mathcal{PR}(i) = \{\pi^T(\mathbf{v})\mathbf{e}_i \text{ for all } \mathbf{v} \in \mathbb{R}^n, \mathbf{v} > 0, \mathbf{v}^T\mathbf{e} = 1\} \subset (0, 1).$$

The following theorem shows that $\mathcal{PR}(i)$ coincides with an open interval whose extreme values are given by the the maximum and minimum entries of the i^{th} -column of X .

Theorem 3.2. *Given a graph \mathcal{G} with dangling nodes indicated by some vector \mathbf{d} , a fixed damping factor $\alpha \in (0, 1)$ and fixed dangling nodes distribution \mathbf{u} , for each node $i \in \mathcal{N}$*

$$\mathcal{PR}(i) = (\min_j x_{ji}, x_{ii}),$$

where $X = (x_{ij}) = (1 - \alpha)(I_n - \alpha(P - \mathbf{d}\mathbf{u}^T))^{-1}$ is the matrix appearing in formula (2.2).

Proof. We will separate the proof of the theorem in two steps:

Step 1. $\min_j x_{ji} < \mathcal{PR}(i) < x_{ii}$ for every personalization vector \mathbf{v} ;

Step 2. every x with $\min_j x_{ji} < x < x_{ii}$ can be achieved as the PageRank of node i for a certain personalization vector \mathbf{v} .

Proof of Step 1. Without loss of generality we can suppose that $i = 1$. Let $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{v}^T \mathbf{e} = 1$. Then the first component of $\pi(\mathbf{v})$ is

$$\pi^T(\mathbf{v})\mathbf{e}_1 = \mathbf{v}^T X \mathbf{e}_1 = \mathbf{v}^T \begin{pmatrix} x_{11} \\ \vdots \\ x_{n1} \end{pmatrix} = \sum_j v_j x_{j1}.$$

In particular, if \mathbf{v} is a personalization vector ($\mathbf{v} > 0$ and $\mathbf{v}^T \mathbf{e} = 1$), $\pi(\mathbf{v})$ is the PageRank corresponding to this personalization vector and the formula above gives the first component of the PageRank. Since in this case all the components of \mathbf{v} are positive and $\sum_j v_j = 1$, $\sum_j v_j x_{j1}$ is a *strict* convex combination of the entries of the first column of X and $\min_j x_{j1} < \sum_j v_j x_{j1} < \max_j x_{j1}$. Moreover, by Lemma 2.3, $\max_j x_{j1} = x_{11}$ and Step 1 is shown.

Proof of Step 2. Without loss of generality suppose again that $i = 1$. By the calculations done in Step 1, the first component of $\pi(\mathbf{v})$ of every $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{v}^T \mathbf{e} = 1$, equals $\sum_j v_j x_{j1}$. In particular, the first component of $\pi(\mathbf{e}_1)$ is x_{11} , the first component of $\pi(\mathbf{e}_2)$ is x_{21} , etc., and the extreme values of the open interval

$$\mathcal{PR}(1) = (\min_j x_{j1}, x_{11})$$

would be achieved if we admitted \mathbf{e}_1 and \mathbf{e}_{j_1} , where we denote by j_1 an index where the minimum of the first column of X is reached.

Now we define

$$\mathbf{v}_{1\varepsilon} = \begin{pmatrix} 1 - \varepsilon \\ \frac{\varepsilon}{n-1} \\ \frac{\varepsilon}{n-1} \\ \vdots \\ \frac{\varepsilon}{n-1} \end{pmatrix} \quad \mathbf{v}_{j_1\varepsilon} = \begin{pmatrix} \frac{\varepsilon}{n-1} \\ \vdots \\ 1 - \varepsilon \\ \vdots \\ \frac{\varepsilon}{n-1} \end{pmatrix} \leftarrow j_1 \text{ coordinate}$$

for every $\varepsilon \in (0, 1)$. Then it is easy to check that both $\mathbf{v}_{1\varepsilon} > 0$ and $\mathbf{v}_{j_1\varepsilon} > 0$, $\mathbf{v}_{1\varepsilon}^T \mathbf{e} = 1 = \mathbf{v}_{j_1\varepsilon}^T \mathbf{e}$, and

$$\lim_{\varepsilon \rightarrow 0^+} \pi^T(\mathbf{v}_{1\varepsilon})\mathbf{e}_1 = x_{11}, \quad \lim_{\varepsilon \rightarrow 0^+} \pi^T(\mathbf{v}_{j_1\varepsilon})\mathbf{e}_1 = x_{j_11}.$$

Finally, for every $\lambda \in (0, 1)$ we define

$$\mathbf{v}_{\lambda\varepsilon} = \lambda \mathbf{v}_{1\varepsilon} + (1 - \lambda) \mathbf{v}_{j_1\varepsilon} > 0$$

which satisfies that

$$\begin{aligned} \lim_{\varepsilon \rightarrow 0^+} \pi^T(\mathbf{v}_{\lambda\varepsilon})\mathbf{e}_1 &= \lambda x_{11} + (1 - \lambda)x_{j_11}, \text{ so} \\ \lim_{\lambda \rightarrow 1} \lim_{\varepsilon \rightarrow 0^+} \pi^T(\mathbf{v}_{\lambda\varepsilon})\mathbf{e}_1 &= x_{11} \\ \lim_{\lambda \rightarrow 0} \lim_{\varepsilon \rightarrow 0^+} \pi^T(\mathbf{v}_{\lambda\varepsilon})\mathbf{e}_1 &= x_{j_11} \end{aligned}$$

and hence for every x with $x_{j_11} < x < x_{11}$ there exists some $\varepsilon_0, \lambda_0 \in (0, 1)$ such that

$$\pi^T(\mathbf{v}_{\lambda_0\varepsilon_0})\mathbf{e}_1 = x.$$

□

Remark 3.3. Note that if $\mathbf{v} > 0$ such that $\mathbf{v}^T \mathbf{e} = 1$, then for every $i \in \mathcal{N}$ $\pi^T(\mathbf{v})\mathbf{e}_i$, is the PageRank of node i when using the personalization vector \mathbf{v} . A convenient notation for this value is $PR(i, \mathbf{v}) = \pi^T(\mathbf{v})\mathbf{e}_i$. We note here that the personalization vectors considered in [12] were of the form $\mathbf{v}_{j\varepsilon}$ and there the definitions only deal with $PR(i, \mathbf{v}_{j\varepsilon})$. The competitiveness interval in the sense of [12] is defined as

$$S_C(i, \varepsilon) = [\min_{j \in \mathcal{N}} PR(i, \mathbf{v}_{j\varepsilon}), \max_{j \in \mathcal{N}} PR(i, \mathbf{v}_{j\varepsilon})],$$

for every $\varepsilon \in (0, 1)$ and each $i \in \mathcal{N}$. Then from Theorem 3.2 it is clear that for a given $\varepsilon \in (0, 1)$ and each $i \in \mathcal{N}$ we have

$$S_C(i, \varepsilon) \subset \mathcal{PR}(i) = \bigcup_{\varepsilon > 0} S_C(i, \varepsilon).$$

4 Some applications

In addition to the intrinsic interest of the previous results, the techniques developed in the last section can be useful in order to analyze the competitiveness of nodes in a network according to their Personalized PageRank and other problems such as the localization of leaders in a complex network. It is well known that Personalized PageRank is a very remarkable tool that helps ranking the nodes of a network according to their centrality (see, for example, [11, 3, 12]). This main fact makes that in many real-life networks (such as WWW networks or social networks) it is crucial for a node i to spot other nodes that can be overcome by i in a ranking based on Personalized PageRank, since these nodes are the nodes that actually *compete* with i in the ranking based on Personalized PageRank. This problem has already been considered in the literature (see, for example, [12]). The techniques developed in the previous section can give a computationally efficient solution to the characterization of the competing nodes of a fixed vertex i . Let us start stating the basic definition of *competitiveness* between two nodes in a complex network.

Definition 4.1. Given two nodes i, j ($i \neq j$) of a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, we say that i and j are *effective competitors* if there exist two personalization vectors \mathbf{v}, \mathbf{w} ($\mathbf{v}, \mathbf{w} > 0$ and $\mathbf{v}^T \mathbf{e} = 1 = \mathbf{w}^T \mathbf{e}$) such that the i^{th} -component of the personalized PageRank with respect to \mathbf{v} is greater than the j^{th} -component of the personalized PageRank with respect to \mathbf{v} , but the i^{th} -component of the personalized PageRank with respect to \mathbf{w} is smaller than the j^{th} -component of the personalized PageRank with respect to \mathbf{w} , i.e.,

$$\begin{aligned}\pi^T(\mathbf{v})\mathbf{e}_i &> \pi^T(\mathbf{v})\mathbf{e}_j \\ \pi^T(\mathbf{w})\mathbf{e}_i &< \pi^T(\mathbf{w})\mathbf{e}_j.\end{aligned}$$

This definition means that nodes i and j appear with different rank in the personalized PageRank vector if we consider some different personalization vectors \mathbf{v} and \mathbf{w} .

Remark 4.2. Note that this definition is more restrictive than the definition of *competitiveness group* given in [12]. Furthermore note that the fact of being in the same *competitiveness group* is a necessary but not a sufficient condition to be effective competitors. Later we show some examples of this fact.

We will see in this section that the results and techniques coming from the last section give a positive answer to the following question:

Question. Is there an easy method of knowing whether two given nodes are effective competitors or not?

This question and this kind of problems have been posed in the literature in social networks, and actually in [12] a necessary condition for a couple of nodes $i, j \in \mathcal{N}$ to compete is given in terms of the so-called *competitiveness intervals*. We will see in example 4.5 that the result used in [12] only gives necessary conditions for competitiveness between nodes, while the following result gives a complete characterization of the competitors of a given node.

Theorem 4.3. *Given a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ with dangling nodes indicated by some vector \mathbf{d} , a fixed damping factor $\alpha \in (0, 1)$ and fixed dangling nodes distribution \mathbf{u} , two nodes $i, j \in \mathcal{N}$ are effective competitors if and only if there exist $k, \ell \in \{1, \dots, n\}$ such that*

$$x_{ki} > x_{kj} \quad \text{and} \quad x_{\ell i} < x_{\ell j},$$

where $X = (x_{pq}) = (1 - \alpha)(I_n - \alpha(P + \mathbf{d}\mathbf{u}^T))^{-1}$ is the $n \times n$ -matrix given in formula (2.2).

Proof. If we consider $\mathbf{v}_{k\varepsilon}$ and $\mathbf{v}_{\ell\varepsilon}$ as defined in the proof of theorem 3.2,

$$\begin{aligned}\lim_{\varepsilon \rightarrow 0} \pi^T(\mathbf{v}_{k\varepsilon})\mathbf{e}_i &= x_{ki}, \\ \lim_{\varepsilon \rightarrow 0} \pi^T(\mathbf{v}_{k\varepsilon})\mathbf{e}_j &= x_{kj}, \\ \lim_{\varepsilon \rightarrow 0} \pi^T(\mathbf{v}_{\ell\varepsilon})\mathbf{e}_i &= x_{\ell i}, \\ \lim_{\varepsilon \rightarrow 0} \pi^T(\mathbf{v}_{\ell\varepsilon})\mathbf{e}_j &= x_{\ell j},\end{aligned}$$

so from $x_{ki} > x_{kj}$ and $x_{\ell i} < x_{\ell j}$ there exists $\varepsilon > 0$ such that the choice of $\mathbf{v}_{k\varepsilon}$ or $\mathbf{v}_{\ell\varepsilon}$ as personalization vectors exchanges the order of nodes i and j in the PageRank vector with respect to such personalization vectors.

Conversely, suppose that i and j are effective competitors but for all $k \in \{1, \dots, n\}$ $x_{ki} \geq x_{kj}$ (similarly, $x_{ki} \leq x_{kj}$). By hypothesis, there exist some personalization vectors \mathbf{v}, \mathbf{w} such that $\pi^T(\mathbf{v})\mathbf{e}_i > \pi^T(\mathbf{v})\mathbf{e}_j$ and $\pi^T(\mathbf{w})\mathbf{e}_i < \pi^T(\mathbf{w})\mathbf{e}_j$. In particular, if $\mathbf{w}^T = (w_1, \dots, w_n)$,

$$\pi^T(\mathbf{w})\mathbf{e}_i < \pi^T(\mathbf{w})\mathbf{e}_j = \sum_{\ell} w_{\ell}x_{\ell j} \leq \sum_{\ell} w_{\ell}x_{\ell i} = \pi^T(\mathbf{w})\mathbf{e}_i$$

leading to a contradiction. \square

Remark 4.4. This theorem gives an easy way to search for effective competitors: it is enough to compare the i^{th} -column and the j^{th} -column of matrix X ; if each entry of the i^{th} -column is always greater or equal than the corresponding entry of the j^{th} -column (or if it is always smaller or equal), then nodes i and j are not effective competitors. Otherwise, some change in the sign of the difference between columns i and j provide the existence of effective competitors. Moreover, if the changes of sign occur in rows k and ℓ , we can assure that there exists $\varepsilon > 0$ such that $\mathbf{v}_{k\varepsilon}$ and $\mathbf{v}_{\ell\varepsilon}$ are personalization vectors that make nodes i and j compete.

Let us present an example of the use of the previous result and how the intersection condition presented in [12] gives less information than the corresponding one obtained from theorem 4.3.

Example 4.5. Let us consider the network $\mathcal{G}_1 = (V_1, E_1)$ given in figure 1.

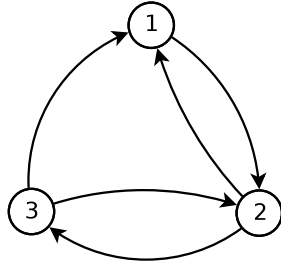


Figure 1: A directed network $\mathcal{G}_1 = (V_1, E_1)$ with 3 nodes

The adjacency matrix of \mathcal{G}_1 is

$$A_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

Then, since \mathcal{G}_1 has no dangling nodes, if we fix $\alpha = 0.85$, then we can compute the matrix $(x_{ij}) = X_1 = (1 - \alpha)(I_3 - \alpha P)^{-1}$ obtaining

$$X_1 = \begin{pmatrix} 0.4035 & 0.4186 & 0.1779 \\ 0.2982 & 0.4925 & 0.2093 \\ 0.2982 & 0.3872 & 0.3146 \end{pmatrix}.$$

Hence, by using theorem 3.2 we get that

$$\begin{aligned} \mathcal{PR}(1) &= (0.2982, 0.4035), \\ \mathcal{PR}(2) &= (0.3872, 0.4925), \\ \mathcal{PR}(3) &= (0.1779, 0.3146). \end{aligned}$$

If we use the necessary conditions obtained in [12], we get that node 1 could compete with nodes 2 and 3, but nodes 2 and 3 cannot compete between them since $\mathcal{PR}(1) \cap \mathcal{PR}(2) \neq \emptyset \neq \mathcal{PR}(1) \cap \mathcal{PR}(3)$ and $\mathcal{PR}(2) \cap \mathcal{PR}(3) = \emptyset$.

In addition to this, if we use the criterion given in theorem 4.3, we note that while 1 and 3 are actually effective competitors, nodes 1 and 2 do not compete. Indeed, by comparing on the one hand the first with the third column of X_1 we get that $x_{11} > x_{13}$ and $x_{21} < x_{33}$ while, on the other hand, by comparing the first and the second columns of X_1 we get that $x_{i1} < x_{i2}$ for all $i = 1, 2, 3$.

Another type of problems that can be solved by using the techniques introduced in the previous section deal with leadership of nodes. The leadership in complex networks has been studied in the Complex Networks Analysis from very different points of view, including (among others) the use of efficiency and robustness perspective in networks related with cryptography (see [4]) and Personalized PageRank in social networks (see [12]). Roughly speaking a node i is a leader (for the personalized PageRank-based ranking) if its personalized PageRank is maximal among all the nodes of the network for some personalization vector. This concept was studied in [12] but only considering personalization vectors of the form $\mathbf{v}_{j\varepsilon}$. As an extension of this concept we introduce the following definition.

Definition 4.6. Given a node i of a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, we say that i is a *leader* of \mathcal{G} if there exists a personalization vector $\mathbf{v} \in \mathbb{R}^n$ ($\mathbf{v} > 0$ and $\mathbf{v}^T \mathbf{e} = 1$) such that for every node $j \in \mathcal{N}$ ($j \neq i$)

$$\pi^T(\mathbf{v})\mathbf{e}_i > \pi^T(\mathbf{v})\mathbf{e}_j.$$

The set of all leader nodes of a graph \mathcal{G} is called the *leadership group* of the network.

Once we have considered the definition of the leadership group of a graph \mathcal{G} it is natural to ask the following question:

Question. Is there an easily-computable way to determine the leadership group of a graph \mathcal{G} ?

Once more this question was considered in [12] and some results in terms of competitiveness intervals were presented, but they only gave sufficient conditions for a node i to be a leader of the network. By using our methods we can go further and prove the following result:

Theorem 4.7. *Given a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ with dangling nodes indicated by some vector \mathbf{d} , a fixed damping factor $\alpha \in (0, 1)$ and fixed dangling nodes distribution \mathbf{u} , the leadership group of \mathcal{G} is the set of nodes $i \in \mathcal{N}$ verifying that there is a value $j \in \mathcal{N}$ such that for every $k \in \mathcal{N}$ ($k \neq i$)*

$$x_{ji} > x_{jk},$$

where $X = (x_{pq}) = (1 - \alpha)(I_n - \alpha(P + \mathbf{d}\mathbf{u}^T))^{-1}$ is the $n \times n$ -matrix given in formula (2.2).

Proof. Let us denote

$$\begin{aligned} A &= \{i \in \mathcal{N}; i \text{ is a leader of } \mathcal{G}\}, \\ B &= \{i \in \mathcal{N}; \exists j \in \mathcal{N} : x_{ji} > x_{jk} \text{ for all } k \neq i\}. \end{aligned}$$

On the one hand, if we take $i \in B$, since there is a value $j \in \mathcal{N}$ such that for every $k \neq i$ we get that $x_{ji} > x_{jk}$, by using the same techniques (and notation) as in the proof of theorem 3.2, a simple continuity argument makes that there is an $\varepsilon \in (0, 1)$ such that for every $k \neq i$

$$\pi^T(\mathbf{v}_{j\varepsilon})\mathbf{e}_i > \pi^T(\mathbf{v}_{j\varepsilon})\mathbf{e}_k,$$

which makes that $i \in A$ and therefore $A \subseteq B$.

On the other hand, if $i \in A$, there is a personalization vector $v \in \mathbb{R}$ such that for every $k \neq i$ we know that $\pi^T(\mathbf{v})\mathbf{e}_i > \pi^T(\mathbf{v})\mathbf{e}_k$. As it was proved in step 1 of the proof of theorem 3.2, if we denote $v = (v_1, \dots, v_n)$ we get that for every $k \neq i$

$$\sum_j v_j x_{ji} = \pi^T(\mathbf{v})\mathbf{e}_i > \pi^T(\mathbf{v})\mathbf{e}_k = \sum_j v_j x_{jk}. \quad (4.8)$$

We are going to show that if the last expression holds, then there is a value $j \in \mathcal{N}$ such that $x_{ji} > x_{jk}$ for all $k \neq i$, otherwise for every $j \in \mathcal{N}$ it should be that $x_{ji} \leq x_{jk}$ and hence for every $k \neq i$

$$\pi^T(\mathbf{v})\mathbf{e}_i = \sum_j v_j x_{ji} \leq \sum_j v_j x_{jk} = \pi^T(\mathbf{v})\mathbf{e}_k,$$

which contradicts equation (4.8). Therefore there is a value $1 \leq j \leq |\mathcal{N}|$ such that $x_{ji} > x_{jk}$ for all $k \neq i$, which makes that $i \in B$ and hence $B \subseteq A$ and we conclude the proof. \square

Remark 4.9. Note that the last result gives an effective algorithm to locate the leadership group of a network $\mathcal{G} = (\mathcal{N}, \mathcal{E})$. It is enough to spot the maximum of each row of the $n \times n$ -matrix X given in formula (2.2) and if it takes place at the element x_{ij} , then the node j is a leader of the network. Actually, the last result ensures that the only possible leader nodes of the networks are those who fulfil this property.

Therefore the matrix $X = (x_{pq}) = (1 - \alpha)(I_n - \alpha(P + \mathbf{d}\mathbf{u}^T))^{-1}$ encapsulates a lot of useful information about the personalized PageRank of the network: The extremal values on each column i correspond to the extremal values of the set $\mathcal{PR}(i)$, the comparison between two columns $j \neq k$ gives the information about the effective competitiveness between j and k and finally, the maximum on each row gives a leader node of the graph.

Let us finish this section with some examples that illustrate the use of our results.

Example 4.10. Let us take the network $\mathcal{G}_2 = (V_2, E_2)$ given in figure 2.

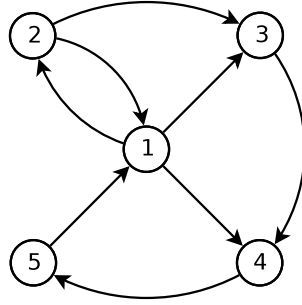


Figure 2: A directed network $\mathcal{G}_2 = (V_2, E_2)$ with 5 nodes

The adjacency matrix of \mathcal{G}_2 is

$$A_2 = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Then, since \mathcal{G}_2 has no dangling nodes, if we fix $\alpha = 0.85$, then we can compute the matrix $X_2 = (1 - \alpha)(I_5 - \alpha P)^{-1}$ obtaining

$$X_2 = \begin{pmatrix} 0.3514 & 0.0995 & 0.1419 & 0.2201 & 0.1871 \\ 0.2410 & 0.2183 & 0.1611 & 0.2052 & 0.1744 \\ 0.2158 & 0.0611 & 0.2371 & 0.2627 & 0.2233 \\ 0.2539 & 0.0719 & 0.1025 & 0.3090 & 0.2627 \\ 0.2986 & 0.0846 & 0.1206 & 0.1871 & 0.3090 \end{pmatrix}.$$

Now, theorem 3.2 determines the set of all possible personalized PageRank values of all the nodes and we get that

$$\begin{aligned}\mathcal{PR}(1) &= (0.2158, 0.3514), \\ \mathcal{PR}(2) &= (0.0611, 0.2183), \\ \mathcal{PR}(3) &= (0.1025, 0.2371), \\ \mathcal{PR}(4) &= (0.1871, 0.3090), \\ \mathcal{PR}(5) &= (0.1744, 0.3090).\end{aligned}$$

Note that the maximum of the first and second column is reached at x_{11} and x_{21} respectively. The maximum of the third and fourth column is reached at x_{34} and x_{44} respectively, and the maximum of the last row is reached at x_{55} . Therefore, by using the method given by theorem 4.7 we get that the leadership group is $\{1, 4, 5\}$.

Example 4.11. Finally, let us now take the network $\mathcal{G}_3 = (V_3, E_3)$ that was introduced in [17] (see figure 3).

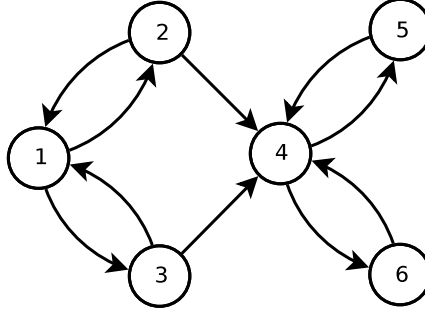


Figure 3: A directed network $\mathcal{G}_3 = (V_3, E_3)$ with 6 nodes introduced in [17]

The adjacency matrix of \mathcal{G}_3 is

$$A_3 = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Then, since \mathcal{G}_3 has no dangling nodes, if we fix $\alpha = 0.85$, then we can compute the matrix $X_3 = (1 - \alpha)(I_6 - \alpha P)^{-1}$ obtaining

$$X_3 = \begin{pmatrix} 0.2348 & 0.0998 & 0.0998 & 0.3057 & 0.1299 & 0.1299 \\ 0.0998 & 0.1924 & 0.0424 & 0.3597 & 0.1529 & 0.1529 \\ 0.0998 & 0.0424 & 0.1924 & 0.3597 & 0.1529 & 0.1529 \\ 0 & 0 & 0 & 0.5405 & 0.2297 & 0.2297 \\ 0 & 0 & 0 & 0.4595 & 0.3453 & 0.1953 \\ 0 & 0 & 0 & 0.4595 & 0.1953 & 0.3453 \end{pmatrix}.$$

In order to determine the set of all possible personalized PageRank values for all nodes, we use once more theorem 3.2 and we obtain that

$$\begin{aligned}\mathcal{PR}(1) &= (0, 0.2348), & \mathcal{PR}(4) &= (0.3057, 0.5405), \\ \mathcal{PR}(2) &= (0, 0.1924), & \mathcal{PR}(5) &= (0.1299, 0.3453), \\ \mathcal{PR}(3) &= (0, 0.1924), & \mathcal{PR}(6) &= (0.1299, 0.3453).\end{aligned}$$

From theorem 4.7 we have that the leadership group is $\{4\}$. We recall here that matrix G given by (2.1) is always an irreducible matrix since $\mathbf{v} > 0$. Nevertheless, X_3 may be a reducible matrix as it happens in this example.

Example 4.12. Finally, let us now take the network with a dangling node $\mathcal{G}_4 = (V_4, E_4)$ that was introduced in [3] (see figure 4). Note that the set of nodes is $V_4 = \{1, \dots, 10\}$ instead of $\{0, \dots, 9\}$ as it was introduced in the original paper [3] in order to keep our notation.

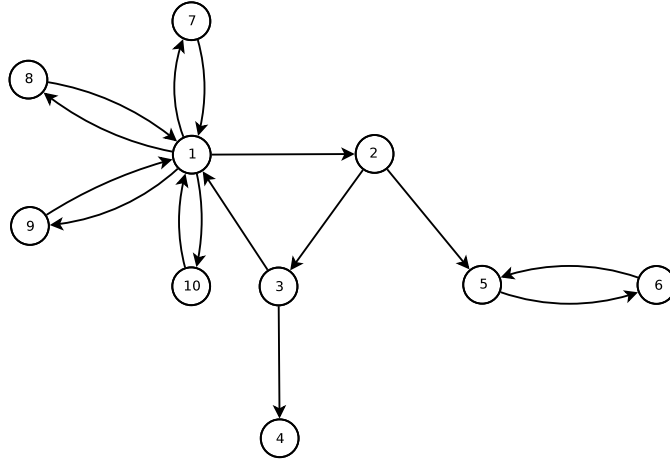


Figure 4: A network $\mathcal{G}_4 = (V_4, E_4)$ with a dangling node – node 4 – with 10 nodes introduced in [3]

The adjacency matrix of \mathcal{G}_4 is

$$A_4 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and since node 4 has no outlinks, it is a dangling node, $\mathbf{d} = \mathbf{e}_4$. As the distribution of dangling nodes we take $\mathbf{u} = \frac{1}{10}\mathbf{e}$. With these data, if we fix $\alpha = 0.85$, the matrix $X_4 = (1 - \alpha)(I_{10} - \alpha(P + \mathbf{d}\mathbf{u}^T))^{-1}$ results

$$X_4 = \begin{pmatrix} 0.399 & 0.0690 & 0.0306 & 0.0142 & 0.114 & 0.0978 & 0.0690 & 0.0690 & 0.0690 & 0.0690 \\ 0.108 & 0.171 & 0.0758 & 0.0353 & 0.282 & 0.243 & 0.0213 & 0.0213 & 0.0213 & 0.0213 \\ 0.254 & 0.0501 & 0.179 & 0.0828 & 0.124 & 0.112 & 0.0501 & 0.0501 & 0.0501 & 0.0501 \\ 0.197 & 0.0488 & 0.0362 & 0.180 & 0.177 & 0.167 & 0.0488 & 0.0488 & 0.0488 & 0.0488 \\ 0 & 0 & 0 & 0 & 0.540 & 0.459 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.459 & 0.540 & 0 & 0 & 0 & 0 \\ 0.339 & 0.0587 & 0.0260 & 0.0121 & 0.0966 & 0.0831 & 0.209 & 0.0587 & 0.0587 & 0.0587 \\ 0.339 & 0.0587 & 0.0260 & 0.0121 & 0.0966 & 0.0831 & 0.0587 & 0.209 & 0.0587 & 0.0587 \\ 0.339 & 0.0587 & 0.0260 & 0.0121 & 0.0966 & 0.0831 & 0.0587 & 0.0587 & 0.209 & 0.0587 \\ 0.339 & 0.0587 & 0.0260 & 0.0121 & 0.0966 & 0.0831 & 0.0587 & 0.0587 & 0.0587 & 0.209 \end{pmatrix}.$$

By theorem 3.2 the sets of all possible PageRank values for each node are

$$\begin{aligned} \mathcal{PR}(1) &= (0, 0.399), & \mathcal{PR}(6) &= (0, 0.540), \\ \mathcal{PR}(2) &= (0, 0.171), & \mathcal{PR}(7) &= (0, 0.209), \\ \mathcal{PR}(3) &= (0, 0.179), & \mathcal{PR}(8) &= (0, 0.209), \\ \mathcal{PR}(4) &= (0, 0.180), & \mathcal{PR}(9) &= (0, 0.209), \\ \mathcal{PR}(5) &= (0, 0.540), & \mathcal{PR}(10) &= (0, 0.209). \end{aligned}$$

Comparing the elements of the columns of X by theorem 4.3:

- node 1 is an effective competitor of nodes 2, 5 and 6;
- node 2 is an effective competitor of nodes 1, 3, 4, 7, 8, 9 and 10;
- node 3 is an effective competitor of nodes 2, 4, 5, 6, 7, 8, 9 and 10;
- node 4 is an effective competitor of nodes 2, 3, 5, 6, 7, 8, 9 and 10;
- node 5 is an effective competitor of nodes 1, 3, 4, 6, 7, 8, 9 and 10;
- node 6 is an effective competitor of nodes 1, 3, 4, 5, 7, 8, 9 and 10;
- node 7 is an effective competitor of nodes 2, 3, 4, 5, 6, 8, 9 and 10;
- node 8 is an effective competitor of nodes 2, 3, 4, 5, 6, 7, 9 and 10;
- node 9 is an effective competitor of nodes 2, 3, 4, 5, 6, 7, 8 and 10;
- node 10 is an effective competitor of nodes 2, 3, 4, 5, 6, 7, 8 and 9;

Finally, taking into account the position of the maximum element of each row of X , by theorem 4.7 the leadership group of this network is $\{1, 5, 6\}$.

Acknowledgements

This paper was partially supported by Spanish MICINN Funds and FEDER Funds MTM2009-13848, MTM2010-16153 and MTM2010-18674, Rey Juan Carlos University Funds I3-2010/00075/001 and Junta de Andalucía Funds FQM-264. The authors would like to thank Prof. Regino Criado, who worked in a previous version of this paper, and the referee for their valuable comments and remarks.

References

- [1] K. AVRACHENKOV, N. LITVAK, D.A. NEMIROVSKY, E. SMIRNOVA, AND M. SOKOL, *Monte Carlo Methods for Top-k Personalized PageRank Lists and Name Disambiguation*, Research Report N° RR-7367, INRIA, 2010.
- [2] A. BERMAN AND R.J. PLEMMONS, *Nonnegative matrices in the mathematical sciences*, SIAM, Philadelphia, 1987.
- [3] P. BOLDI, M. SANTINI AND S. VIGNA, *PageRank: Functional Dependencies*, ACM Trans. Inf. Syst. **27** (4), 2009, pp. 19:1–19:23.
- [4] R. CRIADO, J. FLORES, M.I. GONZÁLEZ-VASCO AND J. PELLO, *Choosing a leader on a complex network*, J. Comput. Appl. Math. **204**, 2007, pp. 10–17.
- [5] M. EIRINAKI AND M. VAZIRGIANNIS *Usage-based Page Ranking for Web Personalization*, Proc. of 5th IEEE International Conference on Data Mining (ICDM 2005), November 2005, Texas, USA.
- [6] T.H. HAVELIWALA, *Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search*, IEEE Transactions on knowledge and data engineering, **15** (4), 2003.
- [7] T.H. HAVELIWALA, S. KAMVAR AND G. JEH, *An Analytical Comparison of Approaches to Personalizing PageRank*, Technical Report, Stanford University, 2003.
- [8] R.A. HORN AND C.R. JOHNSON, *Topics in Matrix Analysis*, Cambridge Univ. Press, New York, 1991.
- [9] G. JEH AND J. WIDOW, *Scaling personalized web search*, Technical Report, Stanford University, 2002.
- [10] A.N. LANGVILLE AND C.D. MEYER. *Google's Pagerank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, 2006.
- [11] L. PAGE, S. BRIN, R. MOTWANI AND T. WINOGRAD *The PageRank citation ranking: Bridging order to the Web* Tech.Rep. **66**, Stanford University. 1998.
- [12] F. PEDROCHE *Competitivity groups on Social Network Sites*, Math. Compt. Modelling **52**, 2010, pp. 1052–1057.
- [13] F. PEDROCHE *Modelling Social Network Sites with PageRank and Social Competences*, Int. J. Compl. Syst. Sci. **1** (1), 2010 pp. 65–68.

- [14] F. PEDROCHE, F. MORENO, A. GONZÁLEZ AND A.A. VALENCIA *Leadership groups on Social Network Sites based on Personalized PageRank*, Math. Compt. Modelling <http://dx.doi.org/10.1016/j.mcm.2011.12.026>.
- [15] H. TONG, C. FALOUTSOS AND J-Y. PAN, *Fast Random Walk with Restart and Its Applications*, ICDM '06 Proceedings of the Sixth International Conference on Data Mining IEEE Computer Society Washington, DC, USA, 2006.
- [16] A. VATTANI, D. CHAKRABARTI, AND M. GUREVICH, *Preserving Personalized Pagerank in Subgraphs*, Proceedings of the 28th Int. Conf. on Machine Learning, Bellevue, WA, USA, 2011.
- [17] G. YU, *Social network analysis based on BSP clustering algorithm*, Communications of the IIMA **7** (4), 2007, pp. 39–46.