

Leadership groups on Social Network Sites based on personalized PageRank

F. Pedroche*, F. Moreno†, A. González†, and A. Valencia†

(*) Institut de Matemàtica Multidisciplinària Universitat Politècnica de València

Camí de Vera s/n. 46022 València. Espanya.

(†) Escuela de Sistemas, Universidad Nacional de Colombia, Sede Medellín

Carrera 80 No 65-223. Medellín. Colombia.

November 30, 2011

1 Introduction

Some usual centrality measures (degree, betweenness, etc.) can be used to assign importance to users in a Social Network Site (SNS). We use the concept of PageRank [1] since it has proved to be of utility in some fields, apart of being in the core of the searcher Google.

The framework presented is based on the concept of *Leadership group* recently introduced by one of the authors [2]. In particular, we show how to analyze the structure of the *Leadership group* as a function of a single parameter.

We classify the users of an SNS based on the Personalized PageRank (PPR) vector. PPR is PR when using some prescribed *personalization vector*. PPR was originally introduced to bias PR to some nodes [3]. See, e.g., [4] for an analytical formulation. An example of how to use topics of the queries to bias the PageRank can be found in [5].

We are interesting in classifying the nodes of a network considering the direct graph of the network and the features of the nodes. PPR is used to

*e-mail:pedroche@imm.upv.es. This work is supported by Spanish DGI grant MTM2010-18674.

include some features of the nodes. The method presented allows to give an extra of PR to some nodes in a controlled way. In this paper we call a *leader* a node that has higher PR than the others.

2 Preliminaries

Let $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ be the directed graph representing a Social Network Site. Users are represented by the set of nodes $\mathcal{N} = \{1, 2, \dots, n\}$ and the set of directed links is $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$. The link represented by the pair (i, j) belongs to the set \mathcal{E} if and only if there exists a link pointing from node i to node j . In this paper we assume that each node has at least one outlink.

Let $0 < \alpha < 1$ be the damping factor (that we use as $\alpha = 0.85$). Let $\mathbf{e} \in \mathbb{R}^{n \times 1}$ be the vector of all ones and let \mathbf{v} be the personalization (or teleportation) vector, i.e., $\mathbf{v} = (v_i) \in \mathbb{R}^{n \times 1} : v_i > 0$ for all $i \in \mathcal{N}$ and $\mathbf{v}^T \mathbf{e} = 1$. The Google matrix is defined as $G = \alpha P + (1 - \alpha) \mathbf{e} \mathbf{v}^T$, and is an stochastic and primitive (irreducible and aperiodic) matrix [3]. The PageRank vector is defined as the unique left Perron vector of G , that is: $\pi^T = \pi^T G$, with $\pi^T \mathbf{e} = 1$. Denoting \mathbf{e}_i the i th column of the identity matrix of order n , the PageRank of a node i is $\pi_i = \pi^T \mathbf{e}_i$. We call basic PageRank, and denote it by *basic PR* to the vector $\pi(\mathbf{e}/n)$. We call *basic leader* a node that is in the top of the *basic PR*.

3 Leadership group

The fundamentals of the model were presented in [2], where the concept of *Leadership group*, \mathcal{L} , was introduced. The following two definitions constitute the framework that allows classify users using PPR.

Definition 1. Given a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, let $0 < \epsilon \leq \frac{n-1}{n}$ and let $\mathbf{v}_i(\epsilon) = [v_{ij}] \in \mathbb{R}^{n \times 1} : v_{ii} = 1 - \epsilon, v_{ij} = \epsilon/(n - 1)$ if $i \neq j$. For each $i \in \mathcal{N}$, let $PR_i = \pi(\mathbf{v}_i)$. and we denote as $(PR_i)_j$ the j th entry of PR_i .

Note that $(PR_i)_j$ represents the value of the PR corresponding to node j when using the personalization vector $\mathbf{v}_i(\epsilon)$. This is the centrality measure that we use.

Definition 2. Given a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, and $0 < \epsilon \leq \frac{n-1}{n}$, the Leadership group, $\mathcal{L} \subseteq \mathcal{N}$ is defined as follows: $j \in \mathcal{L}$ if, for some $i \in \mathcal{N}$ it holds that

$$(PR_i)_j \geq (PR_i)_k \text{ for all } k \neq j. \tag{1}$$

i.e. for some personalization vector $\mathbf{v}_i(\epsilon)$, node j has the greatest PageRank. The number of different indices $i \in \mathcal{N}$ for which (1) occurs is called the frequency of node j in \mathcal{L} , and we denote it as $\nu_{\mathcal{L}}(j)$.

3.1 Example

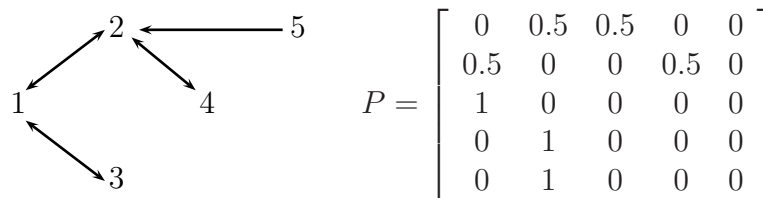


Figure 1: A graph and its corresponding row stochastic matrix

Node j	PR_1	PR_2	PR_3	PR_4	PR_5	$\nu_{\mathcal{L}}(j)$
1	0.38	0.29	0.34	0.26	0.26	2
2	0.30	0.39	0.27	0.35	0.35	3
3	0.17	0.13	0.25	0.12	0.12	0
4	0.14	0.18	0.13	0.25	0.16	0
5	0.01	0.01	0.01	0.01	0.11	0

Table 1: PR_i , and $\nu_{\mathcal{L}}(j)$ for the graph of Fig. 1, with $\epsilon = 0.3$.

In Table 1 we show the elements of the vector PR_i , for $\epsilon = 0.3$, for the graph shown in Fig. 1. Note that in this example changing $\mathbf{v}_i(\epsilon)$ we obtain a different ranking. That is why we claim that we use PPR depending on two parameters (i and ϵ) as a centrality measure. In this example we have $\mathcal{L} = \{1, 2\}$ and $\nu_{\mathcal{L}}(1) = 2$ and $\nu_{\mathcal{L}}(2) = 3$. Note also that, giving a node i , the maximum value of the PPR for that node is obtained when computing PR_i ; Note also that there are nodes such as nodes 3, 4 or 5 that do not win even though we bias the PPR to them, for the ϵ considered. Computing the same experiment for some values of ϵ we find that for $\epsilon \leq 0.68$ we have that

$\mathcal{L} = \{1, 2\}$ while for $\epsilon \geq 0.69$ we have that $\mathcal{L} = \{1\}$. Therefore, in this graph the ranking, the structure of \mathcal{L} and $\nu_{\mathcal{L}}(j)$ depend on ϵ .

In this framework we can define the probability to be a leader in the following form.

Definition 3. Given a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, and $0 < \epsilon \leq \frac{n-1}{n}$, let \mathcal{L} and $\nu_{\mathcal{L}}(j)$ given by Definition 2. We define the probability of node $i \in \mathcal{N}$ to be a leader as $P(i, \epsilon) = \frac{\nu_{\mathcal{L}}(i, \epsilon)}{\sum_{j \in \mathcal{N}} \nu_{\mathcal{L}}(j, \epsilon)}$.

4 Conclusions

We have presented a theoretical framework for classifying users in SNSs. We have shown that the Personalized PageRank can be used as a centrality measure depending on two parameters. We have shown how to analyze the Leadership group using a single parameter ϵ . We have introduced the measure $\nu_{\mathcal{L}}(j)$ that can be also used as a centrality measure. We have introduced the probability to be a leader in this framework.

References

- [1] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, Stanford Digital Library Technologies Project, 1999.
- [2] F. Pedroche, Competitivity Groups on Social Network Sites, Mathematical and Computer Modelling, 52 (2010), p. 1052-1057.
- [3] A. N. Langville, C. D. Meyer. Google's Pagerank and Beyond: The Science of Search Engine Rankings, Princeton University Press, 2006.
- [4] T. Haveliwala, S. Kamvar, G. Jeh. An Analytical Comparison of Approaches to Personalizing PageRank, Technical Report. Stanford. 2003.
- [5] T. H. Haveliwala, Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search, IEEE Transactions on knowledge and data engineering, vol. 15, No. 4. 2003.