PageRank with Priors: An Influence Propagation Perspective

Biao Xiang[1], **Qi Liu**[1], **Enhong Chen**[1], **Hui Xiong**[2], **Yi Zheng**[1], **Yu Yang**[1]
[1]School of Computer Science and Technology, University of Science and Technology of China
{bxiang,feiniaol,xiaoe,punkcpp}@mail.ustc.edu.cn, cheneh@ustc.edu.cn
[2]Rutgers Business School, Rutgers University
hxiong@rutgers.edu

Abstract

Recent years have witnessed increased interests in measuring authority and modelling influence in social networks. For a long time, PageRank has been widely used for authority computation and has also been adopted as a solid baseline for evaluating social influence related applications. However, the connection between authority measurement and influence modelling is not clearly established. To this end, in this paper, we provide a focused study on understanding of PageRank as well as the relationship between PageRank and social influence analysis. Along this line, we first propose a linear social influence model and reveal that this model is essentially PageRank with prior. Also, we show that the authority computation by PageRank can be enhanced with more generalized priors. Moreover, to deal with the computational challenge of PageRank with general priors, we provide an upper bound for top authoritative nodes identification. Finally, the experimental results on the scientific collaboration network validate the effectiveness of the proposed social influence model.

1 Introduction

As people becoming more inextricably linked through the power of information technology, huge social network data have been collected. These network data provide unparallel opportunities for researchers to understand the human world and generate useful knowledge. Indeed, tremendous efforts have been made for measuring authority [Farahat *et al.*, 2006] and modelling social influence [Aggarwal, 2011].

Generally, in traditional social network analysis, the term *authority* is used for estimating the endorsement that is received by the node from its inlinks, and the classic models include PageRank [Page *et al.*, 1999] and HITS [Kleinberg, 1999], which were first proposed for ranking web pages. However, social influence (or influence for short) is the impact that an individual has on others (e.g., leading to the change of their opinions or behaviors) from their outlinks. The Independent Cascade (IC) model [Goldenberg *et al.*, 2001] and the Linear Threshold (LT) model [Granovetter, 1978] are two of the most popular models for describing influence propagation. In fact, a web page is ranked highly if many authoritative pages point to it, and an individual is valued most if he/she influences many influential people. While authority and influence appear quite different at a first glance, several researchers have sensed that they are essentially the same: The individual earns authority by influencing others. This is also the reason that the PageRank algorithm has been used as a solid baseline for evaluating social influence related applications [Aggarwal *et al.*, 2011; Chen *et al.*, 2010; Goyal *et al.*, 2010b; 2011; Liu *et al.*, 2012; Tang *et al.*, 2009].

Nonetheless, there are still two questions to answer. First, what is the connection between PageRank and social influence models. Second, can social influence models help better understand the authority values obtained by PageRank? To answer these two questions, in this paper, we first propose a linear and tractable social influence model which is an approximation of the IC model (which is untractable). Then, we show that this linear model is essentially PageRank with prior, i.e., the PageRank algorithm is actually a special case of this linear model. Therefore, we argue that the authority of each node is essentially the collection of its influences on the network or a specific subnetwork. Based on this finding, we reveal that many similar and effective authority computation methods, which consider more prior knowledge, can be obtained by simply changing the parameter settings in the proposed linear model. Meanwhile, we show that the PageRank value can be used to form an upper bound, which is further used to develop an efficient algorithm for finding the most authoritative nodes with general priors. Finally, we validate these discoveries by performing experiments on a real-world collaboration network. To the best of our knowledge, this is the first comprehensive attempt for exploring and building connections between the researches on PageRank and influence, with a focus on understanding both the traditional and topic-sensitive PageRanks in an influence perspective.

2 Background and Related Work

Let $G = (\mathcal{V}, \mathcal{A}, \mathbf{W}, \mathbf{T})$ be a network (as shown in Figure 1), where node set $\mathcal{V} = \{1, 2, ..., n\}$ and edge set $\mathcal{A}$ represents all the connections between nodes. $\mathbf{W} = [w_{ij}]_{n*n}$ is the PageRank matrix, w_{ij} represents the strength of the endorsement from node i to node j. $\mathbf{T} = [t_{ij}]_{n*n}$ is a transmission matrix for influence propagation, t_{ij} represents the propagation prob-

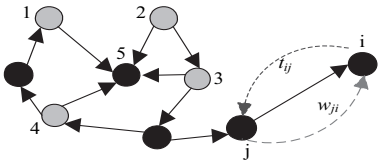


Figure 1: An example directed network.

ability from node i to node j. If there is an edge from j to i in $\mathcal{A}$ (i.e., j trusts i), then $w_{ji} > 0$ and $t_{ij} > 0$ [1], otherwise $w_{ji} = t_{ij} = 0$. Since learning the non-zero t_{ij} and w_{ij} [Goyal *et al.*, 2010a] is beyond the scope of this paper, we assume they are known and usually $\sum_{i=1}^{n} t_{ij} \leq 1$ [Yang *et al.*, 2012] and $\sum_{j=1}^{n} w_{ij} = 1$ [Bianchini *et al.*, 2005]. Notably, $\mathbf{W}'$ is actually a specification of $\mathbf{T}$ in mathematics. Here, we present $\mathbf{W}$ and $\mathbf{T}$ simultaneously because we will study both PageRank and influence model through a common network framework. G is assumed to be directed, as influence propagation is specific to direction in the most general case [Aggarwal *et al.*, 2011][2].

Authority Computation by PageRank. PageRank [Page *et al.*, 1999] has been widely known as a reputable way to obtain an authority score for a node based on the network connectivity. The general PageRank values $\mathbf{x} = [x_1, x_2, ..., x_n]'$ of the nodes in a network could be formalized as:

$$\mathbf{x} = d\mathbf{W}'\mathbf{x} + \frac{(1-d)}{n}\mathbf{e} \tag{1}$$

where $d \in (0, 1)$ is a damping factor, and $\mathbf{e} = [1, 1, ...1]'$. It has been proved that the above iterative process is stable and the linear system always converges [Bianchini *et al.*, 2005]. There are also some improvements for PageRank to better measure nodes' authorities by including domain knowledge. A typical way is to add different edge weights to get a more precise $\mathbf{W}$ [Ding *et al.*, 2009]. Also, an alternative way is to use priors to obtain a nonuniform personalization vector instead of $\frac{1}{n}\mathbf{e}$ [Haveliwala, 2003]. As an effective and efficient algorithm, PageRank has been applied to a number of applications for authority computation, such as Web search [Page *et al.*, 1999], bibliometrics analysis [Ding, 2011], item recommendations [Liu *et al.*, 2012], link predictions [Liben-Nowell and Kleinberg, 2007] and expert finding [Zhu *et al.*, 2011]. Langville et al. presented a comprehensive survey of the issues related to PageRank [Langville and Meyer, 2004]. To the best of our knowledge, most of the existing works use PageRank to get an overall or topic-based single value for measuring the node's importance, and have limited focuses on understanding PageRank by exploiting the authority endorsement between nodes.

Influence Models and Computation. Several models [Kimura and Saito, 2006; Chen *et al.*, 2010; Aggarwal *et al.*, 2011] were provided to describe the dynamics of influence propagation. Among them, the IC model [Goldenberg *et al.*, 2001] is widely used. In IC model, the activated/influenced nodes have a single chance to influence their neighbors independently with a probability. This iterative propagation process will not stop until there is no newly influenced node. The IC model where each link shares the same propagation probability is called the Uniform IC Model, and the one with edge weights is called the Weighted Cascade (WC) Model [Kempe *et al.*, 2003].

[1] If j trusts i, then j will endorse i while i influences j.

[2] The proposed techniques can be applied to undirected networks.

Table 1: Several important mathematical notations.

Notations	Description
$f_{i\to j}$	influence from node i to j
$f_{i\to\mathcal{T}}$	total influence from i to the nodes in set $\mathcal{T}$
$\mathbf{f}_i$	influence vector for node i
α_i	parameter, the prior probability of node i
λ_j	parameter, the damping coefficient of node j
$\mathbf{v_i}$	vector, $\mathbf{v}_{\mathbf{i},i}$ is used to guarantee $f_{i\to i} = \alpha_i$
$\mathbf{P}$	represents both $(\mathbf{I} + \lambda\mathbf{I} - \mathbf{T}')^{-1}$ and $(\mathbf{I} + \lambda\mathbf{I} - \mathbf{W})^{-1}$, with each entry p_{ij}, each column $\mathbf{P}_{\cdot i}$
$\boldsymbol{p}$	vector, where $p_i = \sum_{j=1}^{n} p_{ji}$
$\mathbf{x}$	vector, where x_i is the PageRank value of node i
$\mathbf{x}_{\mathbf{i},j}$	similar to $f_{i\to j}$, the pairwise PageRank value

An ultimate goal of social influence models is to find the most influential nodes by computing the spread of their influences. However, most of the existing models are usually untractable. To that end, a large number of Monte Carlo simulations are needed. To improve computational efficiency, many heuristics have been proposed. For instance, Leskovec et al. [Leskovec *et al.*, 2007] designed the cost-effective lazy forward (CELF) optimization, and Chen et al. [Chen *et al.*, 2009; 2010] proposed both the Degree Discount heuristic and the Maximum Influence Path heuristic. Similarly, Kimura et al. [Kimura and Saito, 2006] proposed the shortest-path based influence algorithm. Aggarwal et al. [Aggarwal *et al.*, 2011] proposed the *SteadyStateSpread* method by solving a system of nonlinear equations for computing the influence spread under the IC model. Moreover, Yang et al. [Yang *et al.*, 2012] observed that propagation probabilities in real-world networks are usually quite small, and thus proposed a quick approximation of influence spread by solving a linear system. In addition, many researchers also consider some constraints in practice. For instance, both Chen et al. [Chen *et al.*, 2012] and Goyal et al.[Goyal *et al.*, 2010b] included time constraints into their approximation algorithms, and Tang et al. [Tang *et al.*, 2009] proposed topical affinity propagation to model the topic-level social influence.

3 Social influence modelling

In this section, we propose a linear social influence model which is both tractable and efficient. For better illustration, Table 1 shows some math notations.

Problem Formulation. In the literature of influence propagation, there are two well-known assumptions [Goldenberg *et al.*, 2001; Granovetter, 1978]: 1) if one is the initiator of something (e.g. opinion, behavior), he/she will spread that with 100% probability; 2) otherwise, this probability will depend on his/her neighbors' influence. However, in the real-world each initiator may not spread the thing with 100% probability (e.g., for lack of self-confidence), i.e., we should take prior knowledge into the first assumption for describing how much probability the node spreads influence to the neighbors. Thus, we could propose an influence model as follows:

Definition 1 *Denote the influence from i to j by $f_{i\to j}$, then*

$$f_{i\to i} = \alpha_i, \quad \alpha_i > 0 \tag{2}$$

$$f_{i\to j} = \frac{1}{1+\lambda_j}\sum_{k\in N_j} t_{kj} f_{i\to k}, \quad for\ j \neq i \tag{3}$$

where $N_j = \{j_1, j_2, ...j_m\}$ is j's trust-friends set (i.e., $\forall k \in N_j$, there is a connection $(j,k) \in \mathcal{A}$). In this definition, we assign each node i a **prior probability** value α_i. If i has a full probability to spread the information, this value should be the maximum (e.g., 1) [3]. In another direction, if i has no interest at all, it will be 0. Meanwhile, another major difference from the traditional models is that we assume the influence flowing to node j is proportional to the linear combination of the influence to j's neighbors (see Equation (3)). Thus, the computation of influence will be a linear efficient way. Here, the parameter λ_j is the damping coefficient of j for the influence propagation. It locates in range $(0, +\infty)$, and the smaller λ_j is, the less influence will be blocked by node j. For simplicity, we choose the same λ for each node, and name $\lambda\mathbf{I}$ as the damping matrix. Similarly, we denote $f_{i\to\mathcal{T}} = \sum_{j\in\mathcal{T}} f_{i\to j}$ as the influence spread from node i to a group of nodes $\mathcal{T}$; that is, it stands for the total influence to the entire network if $\mathcal{T} = \mathcal{V}$.

Influence Computation. Under the above model definition, we can solve the influence spread vector $\mathbf{f}_i = [f_{i\to 1}, f_{i\to 2}, ...f_{i\to n}]'$ for each node i as follows. First, we can rewrite Equation (2) and Equation (3) as

$$\mathbf{f}_i = (\mathbf{I}+\lambda\mathbf{I})^{-1}(\mathbf{T}'\mathbf{f}_i + \nu_\mathbf{i}) = (\mathbf{I}+\lambda\mathbf{I}-\mathbf{T}')^{-1}\nu_\mathbf{i} \tag{4}$$

$$= \mathbf{P}\nu_\mathbf{i} \tag{5}$$

where $\nu_\mathbf{i} = [0, 0, ..., \nu_{\mathbf{i},i}, ...0]'$ is a vector with only the i-th entry $\nu_{\mathbf{i},i}$ is nonzero; that is, $\nu_{\mathbf{i},i}$ should be equal to a number to guarantee $f_{i\to i} = \alpha_i$ as described in Equation (2). In this equation, $(\mathbf{I}+\lambda\mathbf{I}-\mathbf{T}')$ is invertible because its transpose is strictly diagonally dominant, and $n*n$ matrix $\mathbf{P} = (\mathbf{I}+\lambda\mathbf{I}-\mathbf{T}')^{-1}$. As $\nu_\mathbf{i}$ is a vector with only $\nu_{\mathbf{i},i}$ is nonzero, Equation (5) could be rewritten as $\mathbf{f}_i = \nu_{\mathbf{i},i}\mathbf{P}_{\cdot i}$. Specifically, $f_{i\to i} = \nu_{\mathbf{i},i}p_{ii}$, with Equation (2), we could get

$$\nu_{\mathbf{i},i} = \frac{\alpha_i}{p_{ii}}, \quad and\ thus, \quad \mathbf{f}_i = \frac{\alpha_i}{p_{ii}}\mathbf{P}_{\cdot i} \tag{6}$$

Since $\mathbf{P}$ is a positive definite matrix, $p_{ii} > 0$. Then, the total influence from node i to the entire network G should be

$$f_{i\to\mathcal{V}} = \mathbf{f}_i'\mathbf{e} = \sum_{j=1}^{n} f_{i\to j} = \frac{\alpha_i}{p_{ii}}\sum_{j=1}^{n} p_{ji} \tag{7}$$

Given two types of parameters α_i and λ, and the influence transmission matrix $\mathbf{T}$, to get the influence vector $\mathbf{f}_i$, we only need to compute the i-th column of $\mathbf{P}$ ($\mathbf{P}_{\cdot i}$), which can be computed in $O(|\mathcal{A}|)$ since $\mathbf{P}^{-1}\mathbf{P}_{\cdot i} = \mathbf{e}_i$ is a linear system which satisfies the Gauss-Seidel condition.

This linear influence model has close relationship with the traditional ones. For instance, it is easy to prove that the linear approximation method for the IC model [Yang *et al.*, 2012] is actually a specialization of our linear model when $\lambda = 0$ and $\alpha_i = 1$. Also, the non-linear stochastic model [Aggarwal *et al.*, 2011] can be well approximated by our model when $\lambda_i \in [0, 1)$ and $\alpha_i = 1$. The detailed proof is omitted due to the space limit. For the same reason, in the experiments, we just provide the evaluation results on authority measurement rather than finding the most influential nodes.

[3] If initially $\alpha_i > 1$, we could normalize it into (0,1].

4 PageRank with Prior

Here, we find that this linear model is essentially PageRank with prior. Let us first solve Equation (1) algebraically:

$$\mathbf{x} = (\mathbf{I} - d\mathbf{W}')^{-1}\frac{(1-d)}{n}\mathbf{e} \overset{\frac{1}{d}=1+\lambda}{=} (\mathbf{I}+\lambda\mathbf{I}-\mathbf{W}')^{-1}\lambda\frac{\mathbf{e}}{n}$$

Since $\mathbf{W}'$ is actually a specification of influence transmission matrix $\mathbf{T}$ (Section 2), we could further replace matrix $(\mathbf{I}+\lambda\mathbf{I}-\mathbf{W}')^{-1}$ by the matrix $\mathbf{P}'$ (Equation (4)), that is

$$\mathbf{x} = \frac{\lambda}{n}\mathbf{P}'\mathbf{e}, \quad Specifically \quad x_i = \frac{\lambda}{n}\sum_{j=1}^{n} p_{ji} \tag{8}$$

Comparing with Equation (7), we find that

$$x_i = f_{i\to\mathcal{V}}, \quad when \quad \alpha_i = \frac{\lambda}{n}p_{ii} \quad for\ i = 1, 2, ..., n$$

which proves the following theorem.

Theorem 1 *The PageRank value of one node (x_i) is actually its total influence to the entire network ($f_{i\to\mathcal{V}}$) under linear influence model when 1) $\mathbf{T} = \mathbf{W}'$, 2) $\alpha_i = \frac{\lambda}{n}p_{ii}$.*

If we further use $[\mathbf{x}_{\mathbf{i},1}, \mathbf{x}_{\mathbf{i},2}, ..., \mathbf{x}_{\mathbf{i},n}]'$ to denote the authority obtained by node i from each endorsement, we can have $x_i = \sum_{j=1}^{n}\mathbf{x}_{\mathbf{i},j}$, and $\mathbf{x}_{\mathbf{i},j} = \frac{\lambda}{n}p_{ji}$. Based on the above, we know:

- PageRank is actually a special case of our linear social influence model. Thus, PageRank has strong connections with existing social influence models and this is also the reason that PageRank serves as a strong baseline in social influence related applications. Meanwhile, setting $\alpha_i = \frac{\lambda}{n}p_{ii}$ enables the computation of the PageRank values to be linear time (we will explain this later). However, is $\frac{\lambda}{n}p_{ii}$ an appropriate prior? Do there exist more accurate ones? In the following, we will present other possible priors along this line;
- When computing authority and influence, the major difference is just using w_{ji} or t_{ij}. In most of existing works, they are determined in the same way, i.e., equal to or proportional to $\frac{Weight(\mathcal{A}_{ji})}{OutWeight(j)}$ [Bianchini *et al.*, 2005; Kempe *et al.*, 2003], so the authority and influence computed are actually the same thing. In other words, the amount of authority endorsement given from node j to node i is depending on the number of influence flows from i to j ($\mathbf{x}_{\mathbf{i},j} \propto f_{i\to j}$), and vice versa. One step further, we argue that each node's authority (influence) is essentially the collection of its influence (authority) on the network or a subnetwork (e.g., topic-sensitive).

In the following, we use the expression in Equation (8) to represent PageRank. Since influence spread and authority are essentially one concept and they can be distinguished from the context, we use authority/authoritative to stand for both of them. Also, as $\frac{\lambda}{n}$ is a constant and we mainly focus on finding the nodes' relative ranks rather than estimating their true PageRank value, we consider α_i to be p_{ii} in PageRank.

Implications. From the above discussions, we know that the PageRank value x_i is actually $f_{i\to\mathcal{V}}$ with a specific α_i (i.e., $\alpha_i \propto p_{ii}$). Since the traditional PageRank algorithm just considers the total authority (or influence spread) of each node,

and $\alpha_i \propto p_{ii}$ is the only way to make the computation of authority is linear time. Let $\mathbf{f} = [f_{1\to\mathcal{V}}, ..., f_{n\to\mathcal{V}}]'$, and vector $\boldsymbol{p} = \mathbf{P}'\mathbf{e} = [p_1, ..., p_n]'$, where $p_i = (\mathbf{P}_{\cdot i})'\mathbf{e} = \sum_{j=1}^{n} p_{ji}$. Then based on Equation (6) and Equation (7): $\mathbf{f} = [\frac{\alpha_1}{p_{11}}\mathbf{P}_{\cdot 1}, ..., \frac{\alpha_n}{p_{nn}}\mathbf{P}_{\cdot n}]'\mathbf{e}$.

To solve $\mathbf{f}$, if α_i is not proportional to p_{ii} (i.e., $\alpha_i \not\propto p_{ii}$), we have to compute the matrix $\mathbf{P}$ which is $O(n^2)$. Otherwise, just as PageRank does, we can get $\mathbf{f} \propto \mathbf{P}'\mathbf{e} = \boldsymbol{p}$. Then based on the Gauss-Seidel method, this linear system can be solved in $O(|\mathcal{A}|)$, and both $\mathbf{f}$ and $\boldsymbol{p}$ can be quickly computed, e.g.,

$$(\mathbf{I} + \lambda\mathbf{I} - \mathbf{T})\boldsymbol{p} = \mathbf{e} \tag{9}$$

We know why PageRank is efficient. However, it seems that it is more reasonable to set each α_i to be a positive constant when lacking of prior knowledge, or having some prior domain knowledge for guiding this value. For instance, to mine the most influential researchers, we can use the number of their publications as a prior (e.g., $\alpha_i = log(\#Publication_i)$).

However, if $\alpha_i \not\propto p_{ii}$, we have to compute each $\mathbf{P}_{\cdot i}$ to get $f_{i\to\mathcal{V}}$ (or x_i), which will take $O(|\mathcal{A}|)$ for each i. In total, it takes $O(n|\mathcal{A}|)$ to compute $\mathbf{f}$, which is the n times of the PageRank computation time. In practice, we usually are more interested in finding Top-K authoritative ones, the problem then becomes how to quickly estimate each node's authority and filter out insignificant nodes. Indeed, we find out that, for each possible α, the aforementioned $\boldsymbol{p}$, which can be computed in $O(|\mathcal{A}|)$, can be used to form an upper bound for speedup.

Upper Bound and Selection of Top-K Nodes. For a given prior α_i, node i's total authority/influence (for consistency, we note it as $f_{i\to\mathcal{V}}$ rather than x_i) is no larger than $(1 + \lambda)\alpha_i p_i$.

This upper bound can be proved in the following way. By Equation (5), we have $\mathbf{P}^{-1}\mathbf{f_i} = (\mathbf{I} + \lambda\mathbf{I} - \mathbf{T}')\mathbf{f_i} = \mathbf{v_i}$, and thus $(1 + \lambda)\alpha_i - \sum_{k\neq i} t_{ki} f_{i\to k} = \mathbf{v}_{\mathbf{i},i}$. Since both $t_{ki} \geq 0$ and $f_{i\to k} \geq 0$, we can get $\mathbf{v}_{\mathbf{i},i} \leq (1 + \lambda)\alpha_i$. Meanwhile, as $f_{i\to j} = p_{ji}\mathbf{v}_{\mathbf{i},i}$, $f_{i\to j} \leq (1 + \lambda)\alpha_i p_{ji}$. Thus, $f_{i\to\mathcal{V}} = \sum_{j=1}^{n} f_{i\to j} \leq (1 + \lambda)\alpha_i p_i$.

For finding the Top-K authoritative nodes (when $\alpha_i \not\propto p_{ii}$), we first compute all $[(1+\lambda)\alpha_i p_i]$s in $O(|\mathcal{A}|)$, and then use them to save computations. Algorithm 1 describes the proposed framework. In a nutshell, if we only have to compute the pairwise authority value for N nodes, the time complexity of Algorithm 1 is $O((N + 1)|\mathcal{A}|)$.

General Applications. As our linear model can generalize the PageRank based authority computation by introducing more prior knowledge (α_i), Algorithm 1 is also a general framework that will be useful in a number of scenarios. For instance, based on the finding that each node's total authority is actually a collection of its pairwise authorities, we can easily get the most authoritative ones to a specific subnetwork (e.g., the gray nodes in Figure 1) given the whole network structure (or topic/domain-sensitive authority) [Haveliwala, 2003]. Indeed, with the help of generalized authority and Algorithm 1, we can now effectively and efficiently solve this topic-sensitive authority computation as long as we collect the topic profiles (e.g., age, country) of each individual. Specifically, in Algorithm 1, we just need to change the target node set ($\mathcal{V}$) from the entire network to the ones that we are interested in (e.g., a subgroup $\mathcal{T}$) by summarizing and comparing $f_{d\to\mathcal{T}} = \sum_{j\in\mathcal{T}} f_{d\to j}$ for each authoritative candidate node d.

Algorithm 1: Top-K Nodes Selection (G, λ, α, K)

input : $G = (\mathcal{V}, \mathcal{A}, \mathbf{T}, \mathbf{W})$, λ, $[\alpha_1, ..., \alpha_n]$, K
output: **S**: the set of Top-K authoritative nodes.
S=∅;
Compute $\boldsymbol{p} = [p_1, ..., p_n]'$ in $O(|\mathcal{A}|)$ time; *//Equation (9)*
for *each node i* **do**
  $U_i = (1 + \lambda)\alpha_i p_i$; *// Upper bound*
  $IsBound_i = True$;
while $|\boldsymbol{S}| < K$ **do**
  Find node d with the biggest U_d in U;
  if $IsBound_d == True$ **then**
    Compute $f_{d\to j} = \frac{\alpha_d}{p_{dd}} p_{jd}$ for all js in $O(|\mathcal{A}|)$ time;
    //Solve $\boldsymbol{P}^{-1}\boldsymbol{P}_{\cdot d} = \boldsymbol{e}_{\cdot d}$ *by Gauss-Seidel method*
    $U_d = f_{d\to\mathcal{V}}$; *//Equation (7)*
    $IsBound_d = False$;
  else
    $\mathbf{S} = \mathbf{S} \bigcup d$;
    U_d =MINIM; *//E.g., 0*
return ***S****;*

Table 3: The selected methods with different priors.

Met. / α_i	PageRank	WPageRank	Prior(α)	Same(α)	Random(α)
α_i	p_{ii}	$p_{ii} * Conf(i)$	$Conf(i)$	1	random(0,...,1)

5 Experimental Results

We provide empirical validation on a real-world collaboration network from DBLP (http://dblp.uni-trier.de/xml/).

Experimental Setup. We focus on six research domains related to Artificial Intelligence, which are noted as "Artificial Intelligence" (AI), "Computer Vision" (CV), "Database" (DB), "Data Mining" (DM), "Information Retrieval" (IR) and "Machine Learning" (ML). We select the research papers published before January 2013 in several top-ranked journals and conferences from each domain, and the authors are used as nodes to construct the scientific collaboration network G (shown in Table 2). Specifically, an edge $\mathcal{A}_{ji}$ is added when two researchers have one co-authored paper, and the weight is accumulated by the contribution of this author pair on each of their collaborated paper; that is, their contribution for one paper with k authors is $\frac{1}{C_k^2}$. Finally, each $\mathcal{A}_{ji}$ is normalized into w_{ji} by $\frac{Weight(\mathcal{A}_{ji})}{OutWeight(j)}$. In this way, there are total 53,872 nodes and 160,968 edges in G. Meanwhile, for domain-sensitive authority, if the researcher has publications in the conferences/journals of this research domain, then this researcher is classified into the target group $\mathcal{T}$ of this research topic/domain, and the nodes' authorities on $\mathcal{T}$ are computed.

Since we focus on evaluating the effectiveness of the linear model with respect to different priors (α_i) and Algorithm 1, we choose five methods listed in Table 3 for comparison, where PageRank can be also viewed as a baseline. One of the simply designed prior $Conf(i)$ is computed by

Table 2: DBLP data statistics.

	AI	CV	DB	DM	IR	ML
Jour	AI, JAIR	PAMI, IJCV	VLDBJ, TODS	DMKD, TKDE	TOIS	ML, JMLR
Conf	AAAI, IJCAI	CVPR, ICCV	SIGMOD, VLDB, ICDE	KDD, ICDM, SDM	SIGIR, WWW, WSDM	ICML, NIPS, UAI
#Papers	14,279	13,357	10,611	8,301	6,888	11,570
#Authors	11,531	10,431	10,174	10,347	8,958	8,896

Table 5: The average H-indexes for Top-50 researchers.

	AI	CV	DB	DM	IR	ML	Total	Ave.
PageRank	42.26	43.14	51.98	42.92	40.00	37.98	55.84	44.87
WPageRank	41.32	44.10	52.56	40.16	40.56	**39.30**	56.44	44.92
Same(α)	43.06	**45.40**	52.92	**43.30**	**42.44**	38.46	**57.00**	**46.08**
Prior(α)	**43.92**	44.84	**53.28**	41.62	41.78	38.24	56.7	45.76
Random(α)	39.70	39.80	46.90	39.60	41.58	38.80	52.56	42.70

$log(1+CD_i+\sum_j w_{ji}CD_j)$, where CD_i is the observed contribution of researcher i in this specific domain denoted by $\sum_{pD_i} \frac{1}{\#Authors_{pD_i}}$ and pD_i is one publication of i in this domain. For each method, we choose the same $\lambda = 0.176$, and $d = \frac{1}{1+\lambda} = 0.85$.

Selection of Top-K Researchers. In the following, we show a performance comparison by mining top authoritative researchers in each domain. Since the methods only work on a scientific collaboration network with limited information, the results may not ideally reflect the real situation.

First, we show a case study by illustrating the names of the authoritative researchers in each research domain for K=10 in Table 4, where "Total" means the entire collaboration network G. In Table 4, we can see that the results contain influential researchers from different research domains. Even though the methods(or priors) are quite different from each other, the authoritative nodes determined are quite similar, which has been reported before [Aggarwal *et al.*, 2011]. Meanwhile, though the results obtained by *Random*(α) are comparatively different from others, its output are also well-known researchers which demonstrate that not only the prior but also the network structure contributes to the final result.

In addition, since it is impossible to present more researchers for manual analysis, we provide the average H-index results as an alternative. Though there are several limitations for evaluating researchers by H-index, we have two reasons to choose it as a metric. First, H-index can measure both quality and quantity of the published works of researchers based on the number of citations. As we don't include the citation information for the collaboration network, it is reasonable to use a citation metric; Second, among all the metrics for measuring researchers, H-index is well accepted, and it has been widely used in bibliometric analysis [Ding *et al.*, 2009]. Thus, Table 5 lists the average H-index results for Top-50 ranked researchers, where the H-indexes are collected simultaneously in May, 2012. In Table 5, we can see that the methods considering reasonable prior knowledge (e.g., *Same*(α), *Prior*(α) and *WPageRank*) generally perform better than those not (i.e, PageRank and *Random*(α)). Also, *Same*(α) outperforms others by setting the prior probability of each candidate as the same value. This indicates that without useful prior knowledge, it is more reasonable to assign α_i to be the same rather than p_{ii} as PageRank does.

Figure 2 shows the p_{ii} values (PageRank priors) of the Top-50 researchers (ordered by nodes' degrees), from which we cannot find any meaningful patterns, and the results again demonstrate that it is not the best choice to use these values as the prior α_i for each candidate node. For instance, it is improper to set the prior of *Charu C. Aggarwal* much lower. However, this does help us understand the results in Table 4.

Table 6: Search number (N) for finding Top-50 researchers.

	AI	CV	DB	DM	IR	ML	Total
Same(α)	156	144	112	124	168	149	176
Prior(α)	**129**	**120**	**93**	**107**	**123**	**120**	**151**
Random(α)	157	121	109	134	140	127	154

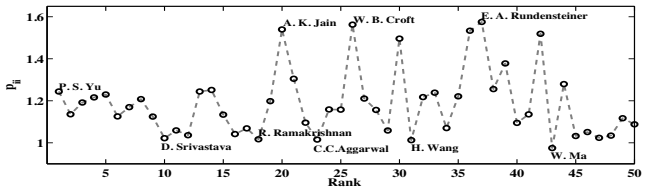
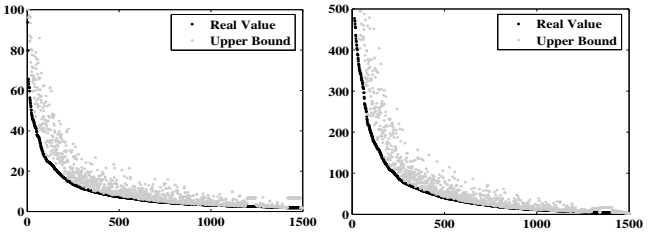


Figure 2: The p_{ii} of the Top-50 researchers in the Total data.



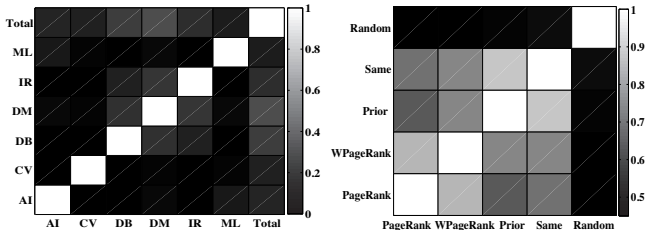
(a) Same(α) on the Total Data. (b) Prior(α) on the Total Data.

Figure 3: An illustration of upper bound.

Upper Bound Evaluation. To evaluate the upper bound in Algorithm 1, we present the number of searched candidates (N) for finding the Top-50 researchers in Table 6. We can observe that this number is quite small with respect to the entire search space (n), which indicates that Algorithm 1 is scalable. Another interesting observation is that the selected upper bound is tight for the *Prior*(α) method, as it takes the least time for *Prior*(α) to return the Top-50 researchers. For better understanding, we also illustrate the true authority value (computed by *Same*(α) and *Prior*(α) respectively) and the upper bounds for 1,500 randomly selected researchers in Figure 3, where we can observe that the upper bounds are always close to the real authority values and this is the reason that each method just has to scan a limited number of candidates for finding the most authoritative researchers (Table 6).

Table 4: An illustration of each domain's Top-10 researchers mined by the methods with different priors.

Domain	Met.	Top-10 Researchers									
AI	PageRank	T. Sandholm	S. Kraus	V. R. Lesser	C. Boutilier	K. D. Forbus	M. M. Veloso	T. Walsh	M. L. Littman	R. J. Mooney	J. E. Laird
	WPageRank	T. Sandholm	V. R. Lesser	S. Kraus	C. Boutilier	R. Dechter	M. M. Veloso	T. Walsh	D. Koller	K. D. Forbus	S. Kambhampati
	Random(α)	S. Thrun	H. J. Levesque	T. Eiter	R. Greiner	P. R. Cohen	J. E. Laird	C. Boutilier	S. J. Russell	M. L. Littman	S. Zilberstein
	Same(α)	S. Kraus	T. Sandholm	V. R. Lesser	C. Boutilier	M. L. Littman	Q. Yang	D. Koller	T. Walsh	M. Tambe	S. Thrun
	Prior(α)	T. Sandholm	S. Kraus	V. R. Lesser	C. Boutilier	D. Koller	Q. Yang	T. Walsh	M. L. Littman	M. Tambe	V. Conitzer
CV	PageRank	T. S. Huang	A. K. Jain	S. K. Nayar	T. Kanade	R. Chellappa	N. Ahuja	L. J. Van Gool	L. S. Davis	A. Zisserman	G. G. Medioni
	WPageRank	T. S. Huang	S. K. Nayar	T. Kanade	A. K. Jain	N. Ahuja	R. Chellappa	L. J. Van Gool	A. Zisserman	L. S. Davis	K. Ikeuchi
	Random(α)	T. S. Huang	T. Kanade	X. Tang	A. Zisserman	G. G. Medioni	S. K. Nayar	L. S. Davis	K. Ikeuchi	C. Schmid	R. Cipolla
	Same(α)	T. S. Huang	T. Kanade	A. K. Jain	A. Zisserman	L. J. Van Gool	S. K. Nayar	R. Chellappa	N. Ahuja	R. Cipolla	L. S. Davis
	Prior(α)	T. S. Huang	T. Kanade	A. Zisserman	S. K. Nayar	L. J. Van Gool	A. K. Jain	R. Chellappa	N. Ahuja	S. Soatto	R. Cipolla
DB	PageRank	P. S. Yu	J. Han	H. Garcia-Molina	M. Stonebraker	E. A. Rundensteiner	D. J. DeWitt	C. Faloutsos	G. Weikum	M. J. Carey	R. Agrawal
	WPageRank	P. S. Yu	J. Han	H. Garcia-Molina	M. Stonebraker	D. J. DeWitt	C. Faloutsos	S. Chaudhuri	R. Agrawal	M. J. Carey	D. Srivastava
	Random(α)	D. J. DeWitt	J. F. Naughton	G. Weikum	H. Garcia-Molina	M. J. Carey	J. Han	J. M. Hellerstein	N. Koudas	D. Srivastava	U. Dayal
	Same(α)	P. S. Yu	J. Han	H. Garcia-Molina	M. Stonebraker	C. Faloutsos	D. J. DeWitt	D. Srivastava	R. Agrawal	M. J. Carey	H. V. Jagadish
	Prior(α)	P. S. Yu	J. Han	H. Garcia-Molina	D. Srivastava	D. Srivastava	D. J. DeWitt	C. Faloutsos	H. V. Jagadish	R. Agrawal	M. Stonebraker
DM	PageRank	P. S. Yu	J. Han	C. Faloutsos	M. Chen	Q. Yang	E. J. Keogh	J. Pei	K. Wang	H. Kriegel	V. Kumar
	WPageRank	P. S. Yu	J. Han	C. Faloutsos	M. Chen	C. C. Aggarwal	J. Pei	Q. Yang	K. Wang	E. J. Keogh	H. P. Kriegel
	Random(α)	P. S. Yu	C. Faloutsos	Q. Yang	K. Wang	J. Han	X. Wu	S. Chaudhuri	V. Kumar	H. Kriegel	H. Xiong
	Same(α)	P. S. Yu	J. Han	C. Faloutsos	Q. Yang	J. Pei	C. C. Aggarwal	M. Chen	K. Wang	E. J. Keogh	H. Kriegel
	Prior(α)	P. S. Yu	J. Han	C. Faloutsos	C. C. Aggarwal	J. Pei	M. Chen	Q. Yang	K. Wang	H. Wang	H. Kriegel
IR	PageRank	W. B. Croft	P. S. Yu	J. Han	H. Garcia-Molina	K. Tanaka	Q. Yang	C. Faloutsos	G. Weikum	C. T. Yu	E. Wilde
	WPageRank	W. B. Croft	P. S. Yu	J. Han	H. Garcia-Molina	Q. Yang	C. Faloutsos	G. Weikum	J. Zobel	C. Zhai	C. T. Yu
	Random(α)	P. S. Yu	C. Faloutsos	Q. Yang	C. L. Giles	W. B. Croft	W. Ma	R. Agrawal	C. T. Yu	R. Jin	K. Tanaka
	Same(α)	P. S. Yu	J. Han	W. B. Croft	H. Garcia-Molina	Q. Yang	C. Faloutsos	W. Ma	Z. Chen	G. Weikum	R. W. White
	Prior(α)	P. S. Yu	J. Han	W. B. Croft	H. Garcia-Molina	Q. Yang	W. Ma	Z. Chen	C. Faloutsos	R. W. White	G. Weikum
ML	PageRank	M. I. Jordan	T. J. Sejnowski	Y. Bengio	C. Koch	D. Koller	G. E. Hinton	B. Schölkopf	A. W. Moore	Z. Ghahramani	J. Shawe-Taylor
	WPageRank	M. I. Jordan	T. J. Sejnowski	G. E. Hinton	D. Koller	Y. Bengio	Z. Ghahramani	B. Schölkopf	A. W. Moore	A. Y. Ng	J. Shawe-Taylor
	Random(α)	M. I. Jordan	D. Koller	S. Thrun	Z. Ghahramani	J. Shawe-Taylor	D. Heckerman	A. J. Smola	C. Koch	M. Mozer	T. J. Sejnowski
	Same(α)	M. I. Jordan	B. Schölkopf	T. J. Sejnowski	D. Koller	G. E. Hinton	Y. Bengio	Z. Ghahramani	K. Müller	J. Shawe-Taylor	A. Y. Ng
	Prior(α)	M. I. Jordan	B. Schölkopf	G. E. Hinton	T. J. Sejnowski	D. Koller	Z. Ghahramani	A. Y. Ng	Y. Bengio	J. Shawe-Taylor	A. W. Moore
Total	PageRank	P. S. Yu	J. Han	C. Faloutsos	T. S. Huang	H. Garcia-Molina	M. I. Jordan	A. K. Jain	Q. Yang	W. B. Croft	T. Kanade
	WPageRank	P. S. Yu	J. Han	C. Faloutsos	T. S. Huang	H. Garcia-Molina	M. I. Jordan	Q. Yang	A. K. Jain	T. Kanade	R. Agrawal
	Random(α)	P. S. Yu	M. I. Jordan	J. Han	C. Faloutsos	K. Tan	Q. Yang	M. Stonebraker	A. Zisserman	M. J. Carey	A. K. Jain
	Same(α)	P. S. Yu	J. Han	C. Faloutsos	T. S. Huang	H. Garcia-Molina	M. I. Jordan	Q. Yang	T. Kanade	R. Agrawal	D. Srivastava
	Prior(α)	P. S. Yu	J. Han	C. Faloutsos	T. S. Huang	M. I. Jordan	H. Garcia-Molina	Q. Yang	D. Srivastava	R. Agrawal	T. Kanade



(a) Different Domains (b) Different Methods

Figure 4: The Jaccard similarity coefficient of researchers.

Correlation Demonstration. For the purpose of further understanding the research domains and the priors, Figure 4 shows the Jaccard similarity coefficient of the selected top researchers (Top-50 for each research domain by each method).

Specifically, Figure 4.(a) demonstrates the coefficient of the top authoritative researcher set between different research domains, and Figure 4.(b) are the coefficient of the top authoritative researcher set output by different methods. In Figure 4.(a) all the five methods' output researchers for one domain are summarized together to stand for this domain, and similarly, in Figure 4.(b) all the six research domains' top researchers gained by the given method are used to stand for this specific method. From Figure 4.(a) we can see that the top researchers mined from the entire network are selected from each single domain which is supported by the similar coefficients between Total domain with six single domains. Among these six domains, CV is the most independent one and has few research connections with others, and the top researchers in AI only have limited connections with those in ML. It is also very interesting to observe that DB, DM and IR are close to each other and have the most top authoritative researchers in common. The different types of coefficient between different methods are more clear as shown in Figure 4.(b), and the most distinctive method is *Random*(α). In contrast, the the most similar two pairs of methods are (PageRank, WPageRank) and (*Same*(α), *Prior*(α)), due to the similarity of their priors.

6 Conclusion

In this paper, we provided an understanding of PageRank and authority from an influence propagation perspective. Along this line, we first developed a linear social influence model, which generalizes the authority computation of PageRank by introducing priors. Also, we revealed that the authority of each node is essentially the collection of its influence on the network or a specific subnetwork. Furthermore, we showed that many similar and effective authority computation methods, which consider more prior knowledge, can be obtained by different parameter settings in the proposed linear social influence model. Meanwhile, we found that the PageRank value can be used to form an upper bound for efficiently computing the most authoritative nodes. Finally, an empirical study was conducted on a real-world DBLP data set to show the effectiveness of the proposed social influence model. In the future, we plan to further evaluate our finding using more data and analyze the connection between authority and influence on other influence models (e.g., IC model, LT model).

Acknowledgements

The work was partially supported by grants from Natural Science Foundation of China (Grant No. 61073110), the Key Program of National Natural Science Foundation of China

(Grant No. 60933013), Research Fund for the Doctoral Program of Higher Education of China (20113402110024), and National Science Foundation (NSF) via grant numbers CCF-1018151 and IIS-1256016.

References

[Aggarwal *et al.*, 2011] C.C. Aggarwal, A. Khan, and X. Yan. On flow authority discovery in social networks. In *Proceedings of SIAM Conference on Data Mining (SDM)*, pages 522–533, 2011.

[Aggarwal, 2011] C.C. Aggarwal. *Social network data analytics*. Springer, 2011.

[Bianchini *et al.*, 2005] M. Bianchini, M. Gori, and F. Scarselli. Inside pagerank. *ACM Transactions on Internet Technology (TOIT)*, 5(1):92–128, 2005.

[Chen *et al.*, 2009] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 199–208. ACM, 2009.

[Chen *et al.*, 2010] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1029–1038. ACM, 2010.

[Chen *et al.*, 2012] W. Chen, W. Lu, and N. Zhang. Time-critical influence maximization in social networks with time-delayed diffusion process. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. AAAI Press, 2012.

[Ding *et al.*, 2009] Y. Ding, E. Yan, A. Frazho, and J. Caverlee. Pagerank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, 60(11):2229–2243, 2009.

[Ding, 2011] Y. Ding. Topic-based pagerank on author cocitation networks. *Journal of the American Society for Information Science and Technology*, 62(3):449–466, 2011.

[Farahat *et al.*, 2006] A. Farahat, T. LoFaro, J.C. Miller, G. Rae, and L.A. Ward. Authority rankings from hits, pagerank, and salsa: Existence, uniqueness, and effect of initialization. *SIAM Journal on Scientific Computing*, 27(4):1181–1201, 2006.

[Goldenberg *et al.*, 2001] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3):211–223, 2001.

[Goyal *et al.*, 2010a] A. Goyal, F. Bonchi, and L.V.S. Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 241–250. ACM, 2010.

[Goyal *et al.*, 2010b] A. Goyal, F. Bonchi, L.V.S. Lakshmanan, and S. Venkatasubramanian. Approximation analysis of influence spread in social networks. *arXiv preprint arXiv:1008.2005*, 2010.

[Goyal *et al.*, 2011] A. Goyal, F. Bonchi, and L.V.S. Lakshmanan. A data-based approach to social influence maximization. *Proceedings of the VLDB Endowment*, 5(1):73–84, 2011.

[Granovetter, 1978] M. Granovetter. Threshold models of collective behavior. *American journal of sociology*, pages 1420–1443, 1978.

[Haveliwala, 2003] T.H. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *Knowledge and Data Engineering, IEEE Transactions on*, 15(4):784–796, 2003.

[Kempe *et al.*, 2003] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[Kimura and Saito, 2006] M. Kimura and K. Saito. Tractable models for information diffusion in social networks. *Knowledge Discovery in Databases: PKDD 2006*, pages 259–271, 2006.

[Kleinberg, 1999] J.M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[Langville and Meyer, 2004] A.N. Langville and C.D. Meyer. Deeper inside pagerank. *Internet Mathematics*, 1(3):335–380, 2004.

[Leskovec *et al.*, 2007] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429. ACM, 2007.

[Liben-Nowell and Kleinberg, 2007] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[Liu *et al.*, 2012] Q. Liu, B. Xiang, E. Chen, Y. Ge, H. Xiong, T. Bao, and Y. Zheng. Influential seed items recommendation. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 245–248. ACM, 2012.

[Page *et al.*, 1999] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[Tang *et al.*, 2009] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 807–816. ACM, 2009.

[Yang *et al.*, 2012] Y. Yang, E. Chen, Q. Liu, B. Xiang, T. Xu, and S. Shad. On approximation of real-world influence spread. *Machine Learning and Knowledge Discovery in Databases*, pages 548–564, 2012.

[Zhu *et al.*, 2011] H. Zhu, H. Cao, H. Xiong, E. Chen, and J. Tian. Towards expert finding by leveraging relevant categories in authority ranking. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2221–2224. ACM, 2011.