An Influence Propagation View of PageRank

QI LIU, University of Science and Technology of China
BIAO XIANG, Ant Financial
NICHOLAS JING YUAN, Microsoft Corporation
ENHONG CHEN, University of Science and Technology of China
HUI XIONG, Rutgers University
YI ZHENG, University of Science and Technology of China
YU YANG, Simon Fraser University

For a long time, PageRank has been widely used for authority computation and has been adopted as a solid baseline for evaluating social influence related applications. However, when measuring the authority of network nodes, the traditional PageRank method does not take the nodes' prior knowledge into consideration. Also, the connection between PageRank and social influence modeling methods is not clearly established. To that end, this article provides a focused study on understanding 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 generalizes the PageRank-based authority computation by introducing some constraints. Then, we show that the authority computation by PageRank can be enhanced if exploiting more reasonable constraints (e.g., from prior knowledge). Next, to deal with the computational challenge of linear model with general constraints, we provide an upper bound for identifying nodes with top authorities. Moreover, we extend the proposed linear model for better measuring the authority of the given node sets, and we also demonstrate the way to quickly identify the top authoritative node sets. Finally, extensive experimental evaluations on four real-world networks validate the effectiveness of the proposed linear model with respect to different constraint settings. The results show that the methods with more reasonable constraints can lead to better ranking and recommendation performance. Meanwhile, the upper bounds formed by PageRank values could be used to quickly locate the nodes and node sets with the highest authorities.

Categories and Subject Descriptors: F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems; H.2.8 [**Database Management**]: Database Applications—*Data mining*

General Terms: Methods, Algorithms

Additional Key Words and Phrases: PageRank, social influence propagation, authority, priors, upper bounds

---

This article is a substantially extended and revised version of Xiang et al. [2013], which appeared in the proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI, oral presentation).
This research was partially supported by grants from the National Natural Science Foundation of China (Grants Nos. 61672483, 61403358, U1605251, and 71329201), the National Science Foundation for Distinguished Young Scholars of China (Grant No. 61325010) and the National Science Foundation (Grant No. IIS-1648664). Q. Liu gratefully acknowledges the support of the Youth Innovation Promotion Association of CAS (No. 2014299).
Authors' addresses: Q. Liu, E. Chen (corresponding author), and Y. Zheng, University of Science and Technology of China; emails: {qiliuql, cheneh}@ustc.edu.cn, xiaoe@mail.ustc.edu.cn; B. Xiang, AI Department, Ant Financial; email: xiangbiao.xb@antfin.com; N. J. Yuan, Microsoft Corporation; email: nicholas.yuan@microsoft.com; H. Xiong, Rutgers University, NJ; email: hxiong@rutgers.edu; Y. Yang, Simon Fraser University; email: yya119@sfu.ca.

DOI: http://dx.doi.org/10.1145/3046941

ACM Reference Format:
Qi Liu, Biao Xiang, Nicholas Jing Yuan, Enhong Chen, Hui Xiong, Yi Zheng, and Yu Yang. 2017. An influence propagation view of pageRank. ACM Trans. Knowl. Discov. Data 11, 3, Article 30 (March 2017), 30 pages.
DOI: http://dx.doi.org/10.1145/3046941

1. INTRODUCTION

Recently, a sizeable amount of network data have been accumulated in many application domains [Kwak et al. 2010; Wang et al. 2012; Zafarani et al. 2014; Yu et al. 2015; Xu et al. 2016]. They provide unparalleled opportunities for the researchers to understand the world and generate useful knowledge. Indeed, tremendous efforts have been made on these network data for node ranking by measuring node authorities [Farahat et al. 2006; Kleinberg 1999] or modeling social influence propagation [Aggarwal 2011; Garg et al. 2012; Goldenberg et al. 2001; Subbian et al. 2014; Zhang et al. 2014].

In practice, both authority and influence can be used for estimating the importance of a node. Specifically, in traditional network analysis, the term *authority* is used for measuring the endorsement that is received by the node from its inlinks. Classic models include PageRank [Page et al. 1999] and HITS [Kleinberg 1999], both of which were first proposed for ranking web pages [Jeh and Widom 2003]. However, influence (or more specifically, social influence) 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 high if many authoritative pages point to it, and an individual is most valuable if he/she influences many influential people. While authority and influence appear quite different at first glance, several researchers have sensed that they are essentially the same, e.g., we can interpret this type of cascade process [Prakash and Faloutsos 2012] as the individual earns authority by influencing others. Thus, some ranking work does not even distinguish between these two concepts deliberately [Li et al. 2011; Weng et al. 2010; Yang et al. 2013]. This is also the reason that the PageRank algorithm has been used as a solid baseline for evaluating influence-related applications [Aggarwal et al. 2011; Chen et al. 2010; Cheng et al. 2014; Goyal et al. 2011; Tang et al. 2009; Liu et al. 2012].

Nonetheless, there are still several questions, and the answers to these questions may lead to the better measurement of node importance in large-scale networks. What is the connection between PageRank and social influence modeling? Can social influence models help better understand the authority values obtained by PageRank? Can the answers to the first two questions provide a better way to address some limitations of PageRank (e.g., the traditional PageRank method does not consider the prior knowledge of the nodes when measuring the authority of a node or a set of nodes)? To answer some of these questions, in our preliminary work [Xiang et al. 2013], we propose a linear and tractable social influence model which is an approximation of the IC model (which is intractable). Then, we show that this linear model generalizes the PageRank-based authority computation by introducing some constraints, i.e., PageRank is actually a special case of this model. Therefore, we argue that the authority of each node is essentially the collection of its influence on the network. 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 constraint settings in the proposed linear model. Meanwhile, we show that the PageRank value can be used to form an upper bound for the real authority with general constraints. This upper bound is then used to develop an efficient algorithm for finding the most authoritative nodes.

In this article, we further develop an algorithm for better exploiting the computation strategy of our proposed linear influence model by the Gauss–Seidel method. More importantly, we provide a new definition of *combined influence/authority* that

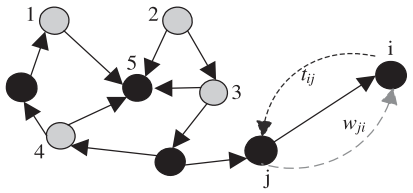


Fig. 1. An example of directed network.

measures the influence spread of a set of nodes. Meanwhile, we show how to quickly estimate these combined influences for ranking sets of nodes with the help of an important property, which indicates that the combined influence is no larger than the sum of the influence of each individual node. Finally, the extensive experimental results on four real-world network datasets prove the effectiveness of the linear model with constraints, the upper bounds and the combined influence. For instance, in terms of the constraints, we can conclude that the more reasonable constraints lead to the better ranking and recommendation performance, and when lacking of prior knowledge, it is a good choice to fix these constraints to be the same. Our contributions can be summarized as follows:

—By following the general assumptions in social influence analysis, we propose a linear influence model and reveal that PageRank is actually a special case of this model. In this way, we could connect PageRank with social influence analysis.
—We propose a solution for better ranking nodes using prior influence constraints. Then, we demonstrate that the choice of constraint terms in PageRank is exactly what is needed to get an efficiently solvable linear system. More importantly, we show the effect of these constraints on the performance of linear model.
—In conjunction with PageRank, we introduce the idea of deriving upper bounds on the authority/influence of either a node or a set of nodes for general constraints. Exploiting the upper bounds that are discovered in this article, we develop efficient algorithms to identify Top-K nodes as well as sets of nodes.

To the best of our knowledge, this is the first comprehensive study exploring the relationship between ranking and influence, which are two key topics in social network analysis. Specifically, this study focuses on understanding the traditional PageRank from an influence perspective. It could further help other types of PageRank related methods, such as TwitterRank [Weng et al. 2010], for better node ranking. Meanwhile, the methods developed in this article could be used in the application of social influence analysis (e.g., viral marketing [Chen et al. 2013; Liu et al. 2014]).

2. BACKGROUND AND RELATED WORK

In this article, we let $G = (\mathcal{V}, \mathcal{A}, \mathbf{W}, \mathbf{T})$ be a network (as shown in Figure 1), where $\mathcal{V} = \{1, 2, \ldots, n\}$ is the node set and edge set $\mathcal{A}$ represents all 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 transition matrix for influence propagation, t_{ij} represents the propagation probability 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 nonzero t_{ij} and w_{ij} [Goyal et al. 2010] is beyond the scope of this article, 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]. We present $\mathbf{W}$ and $\mathbf{T}$ simultaneously because we will study both PageRank and the influence model within the same network framework.

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

Network G is assumed to be directed, as influence propagation is directed in the most general case [Aggarwal et al. 2011]. Note that the proposed techniques can also be applied to undirected networks.

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

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

where $d \in (0, 1)$ is the damping factor, and $\mathbf{e} = [1, 1, \ldots, 1]'$. It has been proven that the above iterative process is stable and the linear system always converges [Bianchini et al. 2005]. There are also some variants of PageRank to measure nodes' authorities better by including a limited amount of domain knowledge. A typical method is to obtain a nonuniform personalization vector instead of $\frac{1}{n}\mathbf{e}$ [Haveliwala 2003]. An alternative way is to add different edge weights to get a more precise measurement $\mathbf{W}$ [Ding et al. 2009]. Going one step further, if enough ground truth labels or relations between node pairs are collected, they could be used to guide the PageRank transition matrix $\mathbf{W}$ on the network [Gao et al. 2011; Backstrom and Leskovec 2011]. For instance, Gao et al. [2011] propose a *semi-supervised PageRank*, where the transition probabilities are defined as parametric models. Then, the authors require that the ranking values should be as close to the stationary distribution of the parametric Markov process as possible. However, in this article, we consider a more general scenario where the authority values $\mathbf{x}$ are the output while $\mathbf{W}$ is given.

Actually, as an effective and efficient algorithm, PageRank model has been applied to a number of applications for authority computation, such as Web search [Page et al. 1999; Jeh and Widom 2003], bibliometrics analysis [Ding 2011], item recommendations [Liu et al. 2012], link predictions [Liben-Nowell and Kleinberg 2007] and expert finding [Zhu et al. 2011] tasks. Some works in these applications choose the ranking results of PageRank as the ground truth. For instance, in diversified ranking and recommendations, both Tong et al. [2011] and Küçüktunç et al. [2013] view the nodes' output by PageRank as the most relevant candidates (e.g., with respect to a specific query) since there is usually no ground truth in network datasets [Li and Yu 2011]. Then, they try to quantify the goodness of a given Top-K ranking list by capturing both relevance and diversity. More often, PageRank serves as a baseline method [Shi et al. 2014]. For instance, in order to identify influencers on Twitter, Kwak et al. [2010] rank users by different methods (including PageRank) and make a quantitative comparison. Further, Liben-Nowell and Kleinberg [2007] experimentally demonstrate that PageRank could beat several predictors for the task of link predictions, and Bi et al. [2011] also choose PageRank as the specific baseline for evaluating the importance of each research paper.

However, there are still some shortcomings in the traditional PageRank method. For instance, it cannot effectively consider the prior knowledge of nodes when measuring one node's authority or directly measuring the combined authority of a set of nodes [Li et al. 2011]. Though Langville and Meyer [2004] present a comprehensive survey of the research issues related to PageRank, to the best of our knowledge, most existing works use PageRank to get an overall single value for measuring the node's importance, and have limited focus in understanding PageRank by exploiting authority endorsements between nodes (this will be illustrated later).

Influence models and computation. Several models [Kimura and Saito 2006; Chen et al. 2010; Aggarwal et al. 2011; Du et al. 2013; Gomez-rodriguez et al. 2013] have been provided to describe the dynamics of influence propagation. Among them, the

IC model [Goldenberg et al. 2001] is widely used and studied [Lucier et al. 2015]. In the 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 with each link sharing the same propagation probability is called the Uniform IC model, and the one with different edge weights is called the Weighted Cascade (WC) model [Kempe et al. 2003].

An ultimate goal of social influence models is to find the most influential nodes, e.g., for viral marketing [Cheng et al. 2014]. However, most existing models are usually intractable, and a large number of Monte-Carlo simulations are needed. To improve computational efficiency, many heuristics have been proposed. For instance, Leskovec et al. [2007] have designed the cost-effective lazy forward (CELF) optimization, and Chen et al. [2009, 2010] propose both the Degree Discount heuristic and the Maximum Influence Path heuristic. Similarly, Kimura and Saito [2006] propose the shortest-path-based influence algorithm. Aggarwal et al. [2011] propose the *SteadyStateSpread* method by solving a system of nonlinear equations. Moreover, Yang et al. [2012] observe that propagation probabilities in real-world networks are usually quite small, and thus propose a quick approximation of influence spread by solving a linear system. In addition, many researchers also consider some constraints in practice, e.g., Tang et al. propose topical affinity propagation to model topic-level social influence [Tang et al. 2009].

Similar to PageRank, many existing influence propagation models follow the idea of a random walk. From this viewpoint, though influence modeling and PageRank ranking are conducted in different contexts, both of them can be viewed as estimating the importance of each node in terms of information diffusion [Yang et al. 2013]. Actually, if we simply ignore the meaning (e.g., node similarity or influence probability) of the edges, all these methods (e.g., PageRank and influence models) could be applied to this network for node ranking. Thus, Li et al. [2011] enhance PageRank through influence propagation and Weng et al. [2010] propose TwitterRank, an extension of PageRank, to measure the influence of users in Twitter. Meanwhile, PageRank has been used as a solid baseline for evaluating influence related applications, e.g., social influence maximization [Chen et al. 2010; Aggarwal et al. 2011; Goyal et al. 2011; Jung et al. 2012], estimating node's reputation [Yang et al. 2013] and finding the Trend-setters [Saez-Trumper et al. 2012]. However, as stated in Section 1, discovering the connection between PageRank and influence modeling is still an open question.

Node set mining and ranking. In this research domain, most current work focuses on mining a specific set of nodes. For instance, network community detection is one of the active research directions [Fortunato 2010; Zhang and Yu 2015]. Recently, researchers have paid more attention to community analysis with specific constraints. For instance, Tang et al. [2012] try to address the evolving group identification problem in dynamic multimode networks, and Wang et al. [2012] initiate the study of magnet community (the communities that attract significantly more peoples interests) mining problem. This type of research also includes finding Top-K nodes for diversified ranking [Tong et al. 2011] and influence maximization [Kempe et al. 2003]. In another direction, for ranking nodes in a given node set, Liu et al. [2013] propose a linear approach to compute independent influence. Though the similar way of modeling social influence prorogation is adopted, Liu et al. [2013] is significantly different from this article, for instance, it only uses the same constraint value for each node and does not try to improve the procedure of social influence modeling. To the best of our knowledge, the work of Li et al. [2011] is the most similar to the work in this article, where it also aim to rank given node sets based on their influence. However, the work of Li et al. [2011] simply views the node set as a big node and does not provide the solution for quickly identifying the Top-K influential node sets.

Table I. Several Important Mathematical Notations

Notations	Description
$\mathbf{W}$	PageRank transition matrix
$\mathbf{T}$	Transition matrix for influence propagation
$f_{i\to j}$	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 influence constraint of node i
λ_j	Parameter, the damping coefficient of node j
$\nu_{\mathbf{i}}$	Vector, $\nu_{\mathbf{i},i}$ 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} and 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
$\mathbf{x}_{i\to j}$	Similar to $f_{i\to j}$, the pairwise PageRank value
$\mathbf{e}_i$	vector, the ith entry in $\mathbf{e}_i$ ($\mathbf{e}_{i,i}$) is 1, otherwise, 0
$\mathcal{T}_t$	Node set which stores all the nodes in domain t
$f_{S\to j}$	Combined influence from node set S to j
$f_{S\to\mathcal{T}}$	Combined influence from S to the nodes in set $\mathcal{T}$
$\mathbf{f}_S$	Influence vector for node set S
$\nu_{\mathbf{S}}$	Vector, $\nu_{\mathbf{S},i}$ ($i\in S$) is used to guarantee $f_{S\to i}=\alpha_i$
$\boldsymbol{\alpha}_S$	$\lvert S\rvert$-dimensional constraint vector, $[\alpha_{s_1},\alpha_{s_2},\ldots,\alpha_{s_{\lvert S\rvert}}]'$
P_{SS}	Matrix, where the rows and columns not corresponding to the members of S are removed from $\mathbf{P}$

3. SOCIAL INFLUENCE MODELING

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

3.1. Problem Formulation

In the literature of influence propagation [Goldenberg et al. 2001; Granovetter 1978], there are two well-known assumptions on the acceptance of propagated information (e.g., opinion): (1) if someone is the original initiator, he/she will accept, while spread that information with full preference; (2) otherwise, this value will depends on his/her neighbors' influence.

However, due to many internal or external profile factors, there may be some variations on the initiator's preference that are shown to the social neighbors. Let us consider a scenario where two Twitter users are advocating for the same thing (e.g., a movie), separately. For the first user (a famous critic), he frequently uses the strong words like "great" and "fantastic" to express his opinion, while the words chosen by the second user (a normal guy) are just "good" and "fine". We can see that the first initiator shows more confidence and preference; thus, he may spread more influence on other users than the second initiator. That is, for precisely computing the influence on the social network, we should also consider the prior knowledge (e.g., confidence and preference) of the initiator. To that end, we measure initiator profiles by generally including some constraints into the first assumption, and then propose an influence model as follows.

Definition 3.1. Denote the influence from node i to node j by $f_{i\to j}$, then

$$f_{i\to j} = \frac{1}{1+\lambda_j}\sum_{k\in N_j} t_{kj}\, f_{i\to k}, \tag{2}$$

$$\text{s.t.,}\quad f_{i\to i}=\alpha_i,\quad \text{for } j=i \text{ and } \alpha_i>0, \tag{3}$$

where $N_j=\{j_1,j_2,\ldots j_m\}$ is j's trust-friend set (i.e., $\forall k\in N_j$, connection $(j,k)\in\mathcal{A}$) and t_{kj} (one entry in $\mathbf{T}$) is the given propagation probability from k to j. In this definition, we

assign each node i a constraint value α_i, which is learned from prior content or domain knowledge. Specifically, if i shows full confidence or preference to the information, this value should be the maximum (e.g., 1).[2] In another direction, if i becomes of 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 (2)). Thus, the computation of influence will be linearly efficient. Parameter λ_j is the damping coefficient of j for the influence propagation. It locates in range $(0, +\infty)$, and the smaller the λ_j is, the less the influence will be blocked. For simplicity, we choose the same λ for each node, and name $\lambda\mathbf{I}$ the damping matrix. 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 $\mathcal{T}=\mathcal{V}$.

3.2. Influence Computation

Under the above model definition, in this subsection we show the way to solve the influence spread vector $\mathbf{f}_i = [f_{i\to 1}, f_{i\to 2}, \dots f_{i\to n}]'$ for each node i.

First, we can rewrite Equations (3) and (2) as

$$f_{i\to j} = \frac{1}{1+\lambda}\left(\sum_{k\in N_j} t_{kj} f_{i\to k} + \nu_{i\to j}\right). \tag{4}$$

Here, $\nu_{i\to j}$ is the jth entry in vector $\nu_{\mathbf{i}} = [0, 0, \dots, \nu_{\mathbf{i},i}, \dots 0]'$, where only the ith entry $\nu_{\mathbf{i},i}$ is not zero; that is, $\nu_{\mathbf{i},i}$ should be equal to a number to guarantee $f_{i\to i} = \alpha_i$ as described in Equation (3). Based on Equation (4), $\mathbf{f}_i$ could be further represented by the following equations:

$$\begin{aligned}\mathbf{f}_i &= (\mathbf{I}+\lambda\mathbf{I})^{-1}(\mathbf{T}'\mathbf{f}_i + \nu_{\mathbf{i}}) \\ &= (\mathbf{I}+\lambda\mathbf{I}-\mathbf{T}')^{-1}\nu_{\mathbf{i}} \qquad (5)\\ &= \mathbf{P}\nu_{\mathbf{i}}. \qquad (6)\end{aligned}$$

In these equations, $(\mathbf{I}+\lambda\mathbf{I}-\mathbf{T}')$ is invertible because it is strictly diagonally dominant, and we denote $n*n$ matrix $\mathbf{P}$ equals $(\mathbf{I}+\lambda\mathbf{I}-\mathbf{T}')^{-1}$. As $\nu_{\mathbf{i}}$ is a vector with only $\nu_{\mathbf{i},i}$ being nonzero, Equation (6) could be rewritten as

$$\mathbf{f}_i = \nu_{\mathbf{i},i}\mathbf{P}_{\cdot i}. \tag{7}$$

Specifically, the ith entry in $\mathbf{f}_i$, i.e., $f_{i\to i}$, is $\nu_{\mathbf{i},i}\,p_{ii}$. With the help of Equation (3), we could get

$$\nu_{\mathbf{i},i} = \frac{\alpha_i}{p_{ii}}, \quad \text{and thus,} \quad \mathbf{f}_i = \frac{\alpha_i}{p_{ii}}\mathbf{P}_{\cdot i}. \tag{8}$$

Since matrix $\mathbf{P}$ is both positive definite and nonnegative, $p_{ii} > 0$. In summary, given two types of parameters α_i and λ, and the influence propagation matrix $\mathbf{T}$, to get the influence vector $\mathbf{f}_i$ for node i, we only need to compute the ith column of $\mathbf{P}$ ($\mathbf{P}_{\cdot i}$). Since $\mathbf{P}^{-1}\mathbf{P}_{\cdot i} = \mathbf{e}_i$ is a linear system which satisfies the Gauss–Seidel condition (i.e., $\mathbf{P}^{-1} = (\mathbf{I}+\lambda\mathbf{I}-\mathbf{T}')$ is strictly diagonally dominant), $\mathbf{P}_{\cdot i}$ can be computed iteratively and the iterative procedure converges to the exact solution for any initial values [Ding et al. 2008; Golub and Van Loan 1996]. In this way, we can get $\mathbf{P}_{\cdot i}$ in $O(|\mathcal{A}|)$. The computation is summarized and shown in Algorithm 1, where in each iteration the elements of $\mathbf{P}_{\cdot i}$ (e.g., $p_{ji}^{(iter+1)}$) are computed sequentially using forward substitution.[3]

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

[3]http://en.wikipedia.org/wiki/Gauss-Seidel_method.

In summary, given the influence transition matrix $\mathbf{T}$ and constraint α_i, for computing influence spread vector $\mathbf{f}_i$ for node i, we first fix $f_{i\to i}$ as α_i, then we use Algorithm 1 and Equation (8) to get other entries in $\mathbf{f}_i$ (e.g., $f_{i\to j}$). Finally, the total influence from node i to the entire network G ($f_{i\to\mathcal{V}}$) should be computed by

$$f_{i\to\mathcal{V}} = \mathbf{f}_i'\mathbf{e} = \sum_{j=1}^{n} f_{i\to j} = \frac{\alpha_i}{p_{ii}} \sum_{j=1}^{n} p_{ji}. \tag{9}$$

ALGORITHM 1: Gauss–Seidel: $(\mathbf{I} + \lambda\mathbf{I} - \mathbf{T}')\mathbf{P}_{\cdot i} = \mathbf{e}_i$ for $\mathbf{P}_{\cdot i}$

input : $\mathbf{T}$, λ, i
output: $\mathbf{P}_{\cdot i} = [p_{1i}, \ldots, p_{ni}]'$: the ith column of $\mathbf{P}$.
for *(j=0;j<n;j++)* **do**
  $p_{ji}^{(0)} = 0$; //Initialization
iter=0;
while *NOT-Converge* **do**
  for *(j=0;j<n;j++)* **do**
    $p_{ji}^{(iter+1)} = \frac{1}{1+\lambda}(\mathbf{e}_{i,j} + \sum_{k=1}^{j-1} t_{kj}\, p_{ki}^{(iter+1)} + \sum_{k=j+1}^{n} t_{kj}\, p_{ki}^{(iter)})$;
  iter++;
return $\boldsymbol{P}_{\cdot i}$

Relationship with traditional social influence models: This linear influence model is closely related to the traditional ones. In the following, we demonstrate that it approximates the IC model [Goldenberg et al. 2001]. To this end, we refer to Yang et al. [2012], where the authors prove that the influence propagation under the IC model could be well approximated by

$$f_{i\to j} = \sum_{t=0}^{\infty} p(j(t)) = \sum_{t=0}^{\infty} p(\tilde{j}(t-1))(1 - \prod_{k\in N_j}(1 - t_{kj}\, p(k(t-1)))), \tag{10}$$

where $p(j(t))$ is the probability that node j will take the action (be influenced) in step t and $p(\tilde{j}(t))$ is the probability that node j has not taken the action until step t. Therefore, under the IC model, $f_{i\to j}$ is exactly the sum of the probabilities that j will take the action in each step. Equation (10) gives a tractable-like stochastic way to represent the IC model; however, $f_{i\to j}$ is still intractable for $p(j(t))$ is intractable. Though it cannot be directly used to solve the IC model, Equation (10) helps to reveal the relationship between the linear model and the IC. Specifically, since $p(\tilde{j}(t)) \le 1$, Equation (10) can be represented by

$$\begin{aligned} f_{i\to j} &\le \sum_{t=0}^{\infty}(1 - \prod_{k\in N_j}(1 - t_{kj}\, p(k(t-1)))) \le \sum_{t=0}^{\infty}\sum_{k\in N_j} t_{kj}\, p(k(t-1)) \\ &= \sum_{k\in N_j} t_{kj} \sum_{t=0}^{\infty} p(k(t-1)) = \sum_{k\in N_j} t_{kj}\, f_{i\to k}. \end{aligned}$$

That is, the inequality of $f_{i\to j} \le \sum_{k\in N_j} t_{kj}\, f_{i\to k}$ holds for the IC model, and this inequality could also be represented by Equation (2) in Definition 3.1. Thus, we can conclude that the IC and linear models describe the influence propagation process

similarly. One step further, the influence spread result under the IC model is actually a special case of the linear model with specific λ_j and α_i settings.

We can prove that the nonlinear stochastic model [Aggarwal et al. 2011] can be also approximated by our model. The detailed proof is omitted since it is not very significant to the work in this article. Indeed, the following discussions and experiments mainly focus on exploring the relationship between PageRank and the proposed linear influence model. The aim is to precisely measure each node's importance and to get better node or node set ranking results effectively and efficiently.

4. PAGERANK WITH CONSTRAINTS

In this section, we first reveal that the linear model can generalize PageRank and introduce several interesting observations. Then, we provide an implication for the constraints automatically chosen by PageRank. Next, we discover an upper bound for quickly estimating authority/influence, and apply it to Top-K authoritative nodes identification.

4.1. General PageRank

For discovering the relationship between the linear social influence model and PageRank, let us first solve the general PageRank vector $\mathbf{x}$ (i.e., Equation (1)) algebraically:

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

Since $\mathbf{W}'$ is actually a specification of influence propagation matrix $\mathbf{T}$ (Section 2), we can further replace matrix $(\mathbf{I} + \lambda\mathbf{I} - \mathbf{W}')^{-1}$ with matrix $\mathbf{P}'$ (Equation (5)), that is

$$\mathbf{x} = \frac{\lambda}{n}\mathbf{P}'\mathbf{e}, \quad \text{Specifically} \quad x_i = \frac{\lambda}{n}\sum_{j=1}^{n} p_{ji}, \tag{11}$$

where x_i (the ith entry in vector $\mathbf{x}$) is the total PageRank value of node i. Comparing Equation (11) with Equation (9), we find that

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

which proves the following theorem.

THEOREM 4.1. *The PageRank value of one node* (x_i) *is equal to its total influence to the entire network* ($f_{i\to\mathcal{V}}$) *under linear influence model when* $\mathbf{T} = \mathbf{W}'$ *and* $\alpha_i = \frac{\lambda}{n}p_{ii}$.

If we further use $[\mathbf{x}_{i\to 1}, \mathbf{x}_{i\to 2}, \ldots, \mathbf{x}_{i\to n}]'$ to denote the authority obtained by node i from each endorsement, we have

$$x_i = \sum_{j=1}^{n}\mathbf{x}_{i\to j}, \quad \text{and} \quad \mathbf{x}_{i\to j} = \frac{\lambda}{n}p_{ji}, \tag{12}$$

which means PageRank value (x_i) is also a collection of pairwise authorities (e.g., $\mathbf{x}_{i\to j}$). Based on the above, we deduce the following:

—PageRank is a special case of our linear social influence model. Thus, PageRank has close connections with existing social influence models which is the reason that PageRank serves as a strong baseline in social influence related applications [Chen et al. 2010; Aggarwal et al. 2011; Goyal et al. 2011; Jung et al. 2012; Yang et al.

2013; Saez-Trumper et al. 2012]. Meanwhile, $\alpha_i = \frac{\lambda}{n} p_{ii}$ enables the computation of PageRank to be linear in time (we will explain this later). However, is $\frac{\lambda}{n} p_{ii}$ an appropriate constraint? Do there exist more accurate ones? In the following, we will present other possible constraints along this line;

—Similar to social influence, one node's authority in the network is also the collection of its authority from others. When computing authority and influence, the major difference is just using w_{ji} or t_{ij}. In most 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 depends on the number of influence flows from i to j ($\mathbf{x}_{i\to j} \propto f_{i\to j}$), and vice versa. Going 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., domain).

In the following, we use the expression in Equation (11) to represent PageRank. Since influence spread and authority are essentially one concept for measuring node's importance and they can be distinguished from the context, we use both "authority" and "influence" without distinction. Also, as $\frac{\lambda}{n}$ is a constant and could be omitted, we usually consider α_i to be p_{ii} instead of $\frac{\lambda}{n} p_{ii}$ in PageRank.

4.2. Implications

From the previous subsection, we know that node i's PageRank value x_i is actually $f_{i\to\mathcal{V}}$ with a specific α_i (i.e., $\alpha_i \propto p_{ii}$). Here, we will further demonstrate that since the traditional PageRank algorithm just considers the total authority (or influence spread) of each node, $\alpha_i \propto p_{ii}$ is the only way to finish the computation of authority in a linear time. Meanwhile, we discuss the strengths and weaknesses for other alternative settings of α_i.

We denote $\mathbf{f} = [f_{1\to\mathcal{V}}, \ldots, f_{n\to\mathcal{V}}]'$ as the vector[4] storing all the nodes' total authorities/influences, and denote vector $\boldsymbol{p} = \mathbf{P}'\mathbf{e} = [p_1, \ldots, p_n]'$, where $p_i = (\mathbf{P}_{\cdot i})'\mathbf{e} = \sum_{j=1}^{n} p_{ji}$, i.e., p_i is the sum of the values in the ith column of $\mathbf{P}$. Then, both our linear social influence model and PageRank aim to get $\mathbf{f}$. Specifically, based on Equations (8) and (9):

$$\begin{aligned}\mathbf{f} &= \left[\frac{\alpha_1}{p_{11}}\mathbf{P}_{\cdot 1}, \ldots, \frac{\alpha_n}{p_{nn}}\mathbf{P}_{\cdot n}\right]'\mathbf{e} \\ &= \left[\frac{\alpha_1}{p_{11}}p_1, \frac{\alpha_2}{p_{22}}p_2, \ldots, \frac{\alpha_n}{p_{nn}}p_n\right]'.\end{aligned}$$

We can see that for solving $\mathbf{f}$, if α_i is not proportional to p_{ii} (i.e., $\alpha_i \not\propto p_{ii}$), we have to compute the inverse matrix $\mathbf{P}$ (the inverse of $(\mathbf{I} + \lambda\mathbf{I} - \mathbf{T}')$), and the time complexity is $O(n^2)$. Otherwise (i.e., $\alpha_i \propto p_{ii}$), just as PageRank does, we can get $\mathbf{f} \propto \mathbf{P}'\mathbf{e} = \boldsymbol{p}$, and this can be further represented as

$$(\mathbf{I} + \lambda\mathbf{I} - \mathbf{T})\boldsymbol{p} = \mathbf{e}. \tag{13}$$

In this way, we only need to know the sum of these values (i.e., vector $\boldsymbol{p}$) instead of the exact value of each entry in $\mathbf{P}$. Thus, both $\mathbf{f}$ and $\boldsymbol{p}$ can be quickly computed in $O(|\mathcal{A}|)$ based on the Gauss–Seidel method (the computation process for solving Equation (13) is similar to that in Algorithm 1).

Now we know why PageRank is efficient. However, setting $\alpha_i \propto p_{ii}$ may not be a good choice for effectively measuring a node's authority since the value of p_{ii} could be

[4]Actually, PageRank vector $\mathbf{x}$ is also a specification of $\mathbf{f}$.

meaningless (as we show in Figure 3 in the experiments). It seems that it is more reasonable to set each α_i to be a positive constant (e.g., $\alpha_i = 1$) when lacking of prior knowledge, or using some prior or domain knowledge for guiding this value. For instance, to mine the most influential researchers in a scientific collaboration network, we can use the number of their publications to generate a constraint (e.g., $\alpha_i = \log(\#Publication_i)$).

However, if $\alpha_i \not\propto p_{ii}$, we need to compute the value of each entry in $\mathbf{P}_{\cdot i}$ to get $f_{i\to\mathcal{V}}$ (or x_i) as noted previously (e.g., Equation (9)), which will take $O(|\mathcal{A}|)$ for each i. In total, it takes $O(n|\mathcal{A}|)$ to compute vector $\mathbf{f}$ for all nodes, i.e., n times of the PageRank computation. We are usually more interested in finding Top-K authoritative ones [Zheng et al. 2015], the problem then becomes how to quickly estimate each node's authority and filter out insignificant ones. Indeed, we find out that, for each α, the aforementioned vector p, computed in $O(|\mathcal{A}|)$, can be used to form an upper bound for speedup.

4.3. Upper Bound and Applications

In this subsection, we first show that for a given constraint α_i, the total authority of node i under linear model (for consistency, we note it as $f_{i\to\mathcal{V}}$ rather than x_i) is no larger than $(1+\lambda)\alpha_i p_i$. Then, we design an algorithm for the quick selection of Top-K authoritative nodes. Finally, we demonstrate that this algorithm is useful in a number of application scenarios.

THEOREM 4.2. *For* $\forall\alpha_i$, $f_{i\to\mathcal{V}} \le (1+\lambda)\alpha_i p_i$, *where* $p_i = (\mathbf{P}_{\cdot i})'\mathbf{e} = \sum_{j=1}^{n} p_{ji}$.

PROOF. By Equation (6) $\mathbf{f}_i = \mathbf{P}\nu_{\mathbf{i}}$ and since $\mathbf{P} = (\mathbf{I} + \lambda\mathbf{I} - \mathbf{T}')^{-1}$, we have

$$\mathbf{P}^{-1}\mathbf{f}_{\mathbf{i}} = (\mathbf{I} + \lambda\mathbf{I} - \mathbf{T}')\mathbf{f}_{\mathbf{i}} = \nu_{\mathbf{i}},$$

and specifically the ith entry in $\nu_{\mathbf{i}}$ is

$$(1+\lambda)\alpha_i - \sum_{k\neq i} t_{ki} f_{i\to k} = \nu_{\mathbf{i},i}.$$

As both $t_{ki} \ge 0$ and $f_{i\to k} \ge 0$, we can get $\nu_{\mathbf{i},i} \le (1+\lambda)\alpha_i$.

Meanwhile, from Equation (7) $f_{i\to j} = p_{ji}\nu_{\mathbf{i},i}$, thus $f_{i\to j} \le (1+\lambda)\alpha_i p_{ji}$. In this way, $f_{i\to\mathcal{V}} = \sum_{j=1}^{n} f_{i\to j} \le (1+\lambda)\alpha_i p_i$ holds. □

For finding the Top-K authoritative nodes (when $\alpha_i \not\propto p_{ii}$), we first compute all the upper bounds $[(1+\lambda)\alpha_i p_i]$s in $O(|\mathcal{A}|)$, and then use them to save computations by only computing the real authority of the nodes with the biggest upper bounds. Algorithm 2 describes the process of the proposed framework. In a nutshell, if we only have to compute the real authority value for N nodes, the time complexity of Algorithm 2 is $O((N+1)|\mathcal{A}|)$. From the experiments, we can see that $N << n$, i.e., the upper bounds are very effective.

Since our linear model generalizes the PageRank-based authority computation by introducing constraint (α_i), Algorithm 2 is also a general framework that will be useful in a number of scenarios, for instance, the most authoritative node identification in a specific domain. Indeed, with the help of the linear model and Algorithm 2, we can now effectively and efficiently solve this domain-specific authority computation as long as we collect the domain profiles (e.g., age, country or research interest) of each individual. Specifically, in Algorithm 2, 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., the subgroup $\mathcal{T}_t$) by summarizing and comparing $f_{i\to\mathcal{T}_t} = \sum_{i\in\mathcal{T}_t} f_{i\to j}$ for each authoritative candidate node i.

ALGORITHM 2: Top-K Nodes Selection (G, λ, α, K)

input : $G = (\mathcal{V}, \mathcal{A}, \mathbf{T}, \mathbf{W})$, λ, $[\alpha_1, \ldots, \alpha_n]$, K
output: **S**: the set of Top-K authoritative nodes.
S=$\emptyset$;
Compute $\boldsymbol{p} = [p_1, \ldots, p_n]'$ in $O(|\mathcal{A}|)$ time; *//Equation (13)*
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 i with the biggest U_i in U;
  if $IsBound_i == True$ **then**
    Compute $f_{i\to j} = \frac{\alpha_i}{p_{ii}} p_{ji}$ for all j's in $O(|\mathcal{A}|)$ time;
    //Solve $\boldsymbol{P}^{-1}\boldsymbol{P}_{\cdot i} = \boldsymbol{e}_{\cdot i}$ *by Gauss–Seidel method*
    $U_i = f_{i\to\mathcal{V}}$; // *Equation (9)*
    $IsBound_i = False$;
  else
    $\mathbf{S} = \mathbf{S} \bigcup i$;
    $U_i =$ MINIM; // *e.g.,0*
return ***S;***

5. COMBINED INFLUENCE AND AUTHORITY

Considering that we are also usually interested in studying the combined influence/authority of sets of nodes (e.g., the node sets with different preferences for one product) rather than just the single nodes, in this section we extend our proposed linear influence model for measuring the influence spread of node sets. Specifically, we first give the definition of a node set's combined influence, and then show the computation process under this definition. Next, we find two upper bounds for this combined influence. Finally, we demonstrate that these two upper bounds could be applied for two possible applications.

5.1. Definition and Computation

In the following, we denote the combined influence/authority from node set S to single node j by $f_{S\to j}$, and denote $f_{S\to\mathcal{T}} = \sum_{j\in\mathcal{T}} f_{S\to j}$ as the combined influence from S to a group of nodes $\mathcal{T}$ (e.g., $\mathcal{T} = \mathcal{V}$). Let us use PageRank algorithm to introduce our definition of combined influence. Actually, PageRank does not define the authority of a set, and with respect to traditional PageRank, we could have total influence from node set S to network G as the sum of each single node's influence:

$$f_{S\to\mathcal{V}} = \sum_{i\in S} f_{i\to\mathcal{V}} = \sum_{i\in S} \frac{\lambda}{n} p_i. \tag{14}$$

However, this kind of definition does not consider the "mutual enrichment" and "influence/authority overlap" of the nodes in a set. For instance, suppose node $i \in S$ and node $j \in S$ in Figure 1, then i's influence will be enriched by node j and vice versa. If we simply sum up these two nodes' total influence (i.e., by Equation (14)), some influence will be counted more than once, making the value much higher than their true influence [Liu et al. 2013]. Notice that it is also unwise to simply remove the edges between nodes in the given set and then treat these nodes as a single node, since the graph structure will be changed in this way and some information will be lost. As the influence always spreads with the help of other nodes, it is not easy to figure out

the real influence of each node. Alternatively, we could treat the nodes in set S as a whole, then we have the following definition for the combined influence.

Definition 5.1. Denote the influence from node set S to j by $f_{S\to j}$, then

$$f_{S\to j} = \frac{1}{1+\lambda_j}\sum_{k\in N_j} t_{kj} f_{S\to k}, \tag{15}$$

$$\text{s.t.,}\quad f_{S\to i} = \alpha_i, \quad \text{for } i \in S \text{ and } \alpha_i > 0. \tag{16}$$

By fixing the influence constraints in the given node set (Equation (16)) and then treating these nodes like a single node (Equation (15)), the mutual enrichment of social influence can be addressed naturally without changing the network structure or information loss [Liu et al. 2014].

Computation: If we further denote the influence spread vector $\mathbf{f}_S = [f_{S\to 1}, f_{S\to 2}, \ldots f_{S\to n}]'$ for node set S, $\mathbf{f}_S$ can be computed similarly to vector $\mathbf{f}_i$ (Section 3.2). Specifically, Equations (16) and (15) could be first rewritten as

$$f_{S\to j} = \frac{1}{1+\lambda}\left(\sum_{k\in N_j} t_{kj} f_{S\to k} + \nu_{\mathbf{S},j}\right). \tag{17}$$

$\nu_{\mathbf{S}} = [0, 0, \ldots, \nu_{\mathbf{S},i}, \ldots 0]'$ is a vector with only the entries $\nu_{\mathbf{S},i}$ ($i \in S$) nonzero, and $\nu_{\mathbf{S},i}$ is equal to a number to guarantee $f_{S\to i} = \alpha_i$. One step further, influence spread vector $\mathbf{f}_S$ can be represented by the following equation:

$$\mathbf{f}_S = \mathbf{P}\nu_{\mathbf{S}}. \tag{18}$$

Here, matrix $\mathbf{P}$ is also equal to $(\mathbf{I} + \lambda\mathbf{I} - \mathbf{T}')^{-1}$. Suppose $S = \{s_1, s_2, \ldots, s_{|S|}\}$, where $|S|$ is the cardinality of S, and without loss of generality we assume the node id $s_1 < s_2 < \cdots < s_{|S|}$. We denote $\nu_{\mathbf{SS}} = [\nu_{\mathbf{S},s_1}, \nu_{\mathbf{S},s_2}, \ldots, \nu_{\mathbf{S},s_{|S|}}]'$, i.e., the subvector after removing 0's from $\nu_{\mathbf{S}}$, and the matrix $\mathbf{P}_{SS}$ is cut down from $\mathbf{P}$ by removing the columns and rows not corresponding to the members of S. Then, for the nodes in S, Equation (18) could be rewritten as

$$\boldsymbol{\alpha}_S = \mathbf{P}_{SS}\nu_{\mathbf{SS}}, \tag{19}$$

where $\boldsymbol{\alpha}_S$ is a $|S|$-dimensional vector $[\alpha_{s_1}, \alpha_{s_2}, \ldots, \alpha_{s_{|S|}}]'$. As $\nu_{\mathbf{SS}} = \mathbf{P}_{SS}^{-1}\boldsymbol{\alpha}_S$, we get

$$\nu_{\mathbf{S},s_i} = \left[\mathbf{P}_{SS}^{-1}\boldsymbol{\alpha}_S\right]_i. \tag{20}$$

Indeed, $\mathbf{P}_{SS}$ is a principal submatrix of $\mathbf{P}$ (a positive definite matrix), and thus $\mathbf{P}_{SS}^{-1}$ exists and is also a positive definite matrix. Based on Equation (20), we get $\nu_{\mathbf{SS}}$ and $\nu_{\mathbf{S}}$. One step further, $\mathbf{f}_S$ can be solved by Equation (18). In this way, for computing $\mathbf{f}_S$, we have to first compute two inverse matrices, $\mathbf{P}$ (i.e., $(\mathbf{I} + \lambda\mathbf{I} - \mathbf{T}')^{-1}$) and $\mathbf{P}_{SS}^{-1}$; thus, the time complexity will be at least $O(n^2)$. Since the $\nu_{\mathbf{S},j}$ is 0, if $j \notin S$, Equation (18) can be rewritten as

$$\mathbf{f}_S = \sum_{i\in S} \nu_{\mathbf{S},i}\mathbf{P}_{\cdot i}. \tag{21}$$

From this equation, we observe that $\mathbf{f}_S$ is actually a linear combination of the corresponding columns in $\mathbf{P}$. Specifically, if S only contains one element, e.g., node i, then Equation (21) will turn into Equation (7). Thus, the total influence from node set S

to the entire network or subnetwork (domain-specific combined influence) of G (e.g., $f_{S\to\mathcal{V}}$) is

$$f_{S\to\mathcal{V}} = \mathbf{f}'_S\mathbf{e} = \sum_{j=1}^{n} f_{S\to j} = \sum_{i\in S} v_{\mathbf{s},i}\, p_i. \tag{22}$$

Meanwhile, with the representation of Equation (21), we only need to compute $|S|$ columns of matrix $\mathbf{P}$ for getting $\mathbf{f}_S$ and $f_{S\to\mathcal{V}}$. These values could be computed in $O(|S||\mathcal{A}|)$ by the Gauss–Seidel method (the computation of one column is shown in Algorithm 1). Notably, S often contains a limited number of nodes, then $|S|$ is small and the computation of $\mathbf{P}_{SS}^{-1}$ is very quick; thus, $\mathbf{f}_S$ can be computed in nearly $O(|\mathcal{A}|)$.

5.2. Upper Bounds and Applications

However, when $|S|$ is comparably large, the computation of influence spread vector $\mathbf{f}_S$ will become very time consuming (at least $O(|S|^2 + |S||\mathcal{A}|)$). Luckily, we can estimate $\mathbf{f}_S$ and $f_{S\to\mathcal{V}}$ by using two upper bounds, e.g., the sum of each single node's influence spread (i.e., $\sum_{i\in S}\mathbf{f}_i$).

Upper bound: Given $\boldsymbol{\alpha}_S$, node set S's total social influence under the linear model is no larger than the sum of each single node's influence.

THEOREM 5.2. *For* $\forall S$, $f_{S\to\mathcal{V}} \le \sum_{i\in S}\mathbf{f}_{i\to\mathcal{V}} \le \sum_{i\in S}(1+\lambda)\alpha_i p_i$.

PROOF. First, let us prove $f_{S\to\mathcal{V}} \le \sum_{i\in S}\mathbf{f}_{i\to\mathcal{V}}$ by $\mathbf{f}_S \le \sum_{i\in S}\mathbf{f}_i$ (based on Equations (22) and (9)) for $\forall S$.

Suppose $i \in S$, we rewrite the definition of linear social influence (Definition 3.1) as

$$f_{i\to j} = \begin{cases} \alpha_i, & j = i, \\ \frac{1}{1+\lambda}\sum_{k\in N_j} t_{kj} f_{i\to k} = \beta_{ij}, & j \ne i \& j \in S, \\ \frac{1}{1+\lambda}\sum_{k\in N_j} t_{kj} f_{i\to k}, & j \notin S. \end{cases}$$

Specifically, we first conclude $f_{i\to j} \ge 0$ for each k. Then, we denote function $f(\boldsymbol{\alpha}_i) = \mathbf{f}_i$, where $\boldsymbol{\alpha}_i = [\beta_{s_i s_1}, \beta_{s_i s_2}, \ldots, \beta_{s_i s_i} = \alpha_{s_i}, \ldots, \beta_{s_i s_{|S|}}]'$. According to the definition of $\mathbf{f}_i$, if $\boldsymbol{\alpha}_i \ge \boldsymbol{\alpha}_j$ (i.e., $\beta_{s_i s_k} \ge \beta_{s_j s_k}$ for $\forall k$), we have $f(\boldsymbol{\alpha}_i) \ge f(\boldsymbol{\alpha}_j)$. Similarly, we can denote $f(\boldsymbol{\alpha}_S) = \mathbf{f}_S$, where $\boldsymbol{\alpha}_S = [\alpha_{s_1}, \alpha_{s_2}, \ldots, \alpha_{s_{|S|}}]'$. Thus,

$$\sum_{i\in S}\mathbf{f}_i = \sum_{i\in S} f(\boldsymbol{\alpha}_i) = f\left(\sum_{i\in S}\boldsymbol{\alpha}_i\right). \tag{23}$$

Denote $\sum_{i\in S}\boldsymbol{\alpha}_i = [\beta_{s_1}, \beta_{s_2}, \ldots, \beta_{s_{|S|}}]'$, where $\beta_{s_l} = \sum_{i=1}^{|S|}\beta_{s_i s_l}$, i.e.,

$$\beta_{s_l} = \beta_{s_l s_l} + \sum_{i=1, i\ne l}^{|S|}\beta_{s_i s_l} = \alpha_{s_l} + \sum_{i=1, i\ne l}^{|S|}\beta_{s_i s_l} \ge \alpha_{s_l}. \tag{24}$$

That is, each value in vector $\boldsymbol{\alpha}_S$ is no bigger than the corresponding value in vector $\sum_{i\in S}\boldsymbol{\alpha}_i$; thus, $f(\boldsymbol{\alpha}_S) \le f(\sum_{i\in S}\boldsymbol{\alpha}_i)$. Combining with Equation (23), we could get $\mathbf{f}_S \le \sum_{i\in S}\mathbf{f}_i$, and one step further $f_{S\to\mathcal{V}} \le \sum_{i\in S}\mathbf{f}_{i\to\mathcal{V}}$ holds.

Second, from Theorem 4.2, we can easily prove $\sum_{i\in S}\mathbf{f}_{i\to\mathcal{V}} \le \sum_{i\in S}(1+\lambda)\alpha_i p_i$ by the fact that $\mathbf{f}_{i\to\mathcal{V}} \le (1+\lambda)\alpha_i p_i$.

In this way, $f_{S\to\mathcal{V}} \le \sum_{i\in S}\mathbf{f}_{i\to\mathcal{V}} \le \sum_{i\in S}(1+\lambda)\alpha_i p_i$ holds.

Applications: We have proposed two upper bounds for the combined influence ($f_{S\to\mathcal{V}}$). The first bound $\sum_{i\in S}\mathbf{f}_{i\to\mathcal{V}}$ is much tighter but the computation is as time consuming as the combined influence. In contrast, the second upper bound $\sum_{i\in S}(1+\lambda)\alpha_i p_i$ is

Table II. Statistics of the Four Real-World Networks

Networks	DBLP	Epinions	Flixster	WebND
#Node	53,872	405,176	787,213	325,729
#Edge/Arc	160,968	717,667	7,058,819	1,497,134
Type	Undirected	Directed	Undirected	Directed

very efficient (please refer to Section 4.3). These two upper bounds are useful in some scenarios of influence and authority analysis, and in this article, we mainly focus on two of them, i.e., one application for each bound.

First, the more the influence overlap among the nodes in set S, the bigger the difference between $\sum_{i\in S}\mathbf{f}_{i\to\mathcal{V}}$ and $f_{S\to\mathcal{V}}$. Thus, upper bound $\sum_{i\in S}\mathbf{f}_{i\to\mathcal{V}}$ can help us measure the closeness of the nodes. Specifically, we can define the overall *Influence Overlap Rate* (*IOR*) for one node set as

$$IOR = \frac{\sum\limits_{i\in S}\mathbf{f}_{i\to\mathcal{V}} - f_{S\to\mathcal{V}}}{\sum\limits_{i\in S}\mathbf{f}_{i\to\mathcal{V}}}. \tag{25}$$

For instance, we could use this *IOR* metric (in the range of [0,1]) to find the closest collaborators in the scientific collaboration network. That is, the bigger *IOR* stands for the closer relationship among researchers in set S (more influence overlap), and vice versa.

Second, given a number of node sets, the upper bound $\sum_{i\in S}(1+\lambda)\alpha_i p_i$ can be used to develop an efficient algorithm which helps us quickly identify the Top-K influential node sets. To this end, the computation process is similar to Algorithm 2. Specifically, we can first compute the upper bound ($(1+\lambda)\alpha_i p_i$) for each node's influence in $O(|\mathcal{A}|)$, then sum up the upper bounds for each node set (e.g., $\sum_{i\in S}(1+\lambda)\alpha_i p_i$ for S) and rank these node sets based on their upper bounds. Then, we only have to compute the real influence (e.g., $f_{S\to\mathcal{V}}$) for the node sets with the biggest upper bounds (e.g., $O(|S|^2+|S||\mathcal{A}|)$ for S), and if one set's real influence is still larger than other sets' upper bounds, we select this set as the candidate. The entire procedure does not stop until we have selected K candidates (i.e., Top-K influential node sets). In this scenario, if we have to compute the real social influence for N node sets, and suppose the average size of one node set is $|S|$, then the time complexity will be $O(N(|S|^2+|S||\mathcal{A}|))$.

6. EXPERIMENTAL RESULTS

We conduct experiments on four real-world networks. Specifically, we demonstrate[5]: (1) the results of the Top-K nodes selection with respect to different constraints; (2) the combined influence analysis; (3) the effectiveness of our upper bounds; (4) the performance of the combined influence computation for research committee recommendation; and (5) a study of the correlations of the output of different methods.

6.1. Experimental Setup

Datasets: The datasets are collected from different domains and platforms (Table II): *DBLP* is a scientific collaboration network[6]; *Epinions* is a given who-trust-whom social network of consumer review site Epinions.com [Massa and Avesani 2006]; *Flixster* is

[5]Some data and code used in this article could be reached from http://staff.ustc.edu.cn/~qiliuql/PageRankPriors.html.

[6]http://dblp.uni-trier.de/xml/.

Table III. Detailed 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
#Pape.	14,279	13,357	10,611	8,301	6,888	11,570
#Auth.	11,531	10,431	10,174	10,347	8,958	8,896

Table IV. The Selected Methods with Different Constraints

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

a social rating network[7] constructed from Flixster [Jamali and Ester 2010]; *WebND* (web-NotreDame[8]) is a famous web page linkage network [Albert et al. 1999].

From these networks, we can collect the prior information from DBLP data. Since the experimental results are comparatively easier for understanding [Aggarwal et al. 2011; Subbian et al. 2014; Zhou and Liu 2015], we also give several carefully designed case studies on this data so as to get more intuitive conclusions. Specifically, we focus on six research domains (six subnetworks), which are "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 treated as nodes to construct the scientific collaboration network G (the statistics can be found in Table III). An edge $\mathcal{A}_{ji}$ is added when two researchers have at least one co-authored paper, and the weight is accumulated by the contribution of this author pair on each of their collaborated paper; that is, the contribution of two researchers for one paper with k authors is $1/\binom{k}{2}$. Finally, each $\mathcal{A}_{ji}$ is normalized into w_{ji} by $\frac{Weight(\mathcal{A}_{ji})}{OutWeight(j)}$ [Bianchini et al. 2005; Kempe et al. 2003]. Meanwhile, for domain-specific authority, if the researcher has publications in the conferences/journals of this research topic/domain (e.g., AI), then this researcher is classified into the target group $\mathcal{T}$ of this domain, and the nodes' authorities on $\mathcal{T}$ are computed.

Methods: Since we focus on evaluating the effectiveness of the linear model with respect to different constraint settings (α_i) and the upper bounds (e.g., Algorithm 2), we choose five ranking methods listed in Table IV for comparison, where PageRank can also be viewed as a baseline and WPageRank is the abbreviation for weighted PageRank. Since the choice of constraint α_i is application-specific, it is hard to find an α_i that works for different applications. Generally speaking, we think there are two ways to determine an effective α_i: We can design the constraints based on the prior knowledge that we can get; Without useful prior knowledge, we could fix α_i to be the same (e.g., 1) for each node i. Following this, we design methods $Prior(\alpha_i)$ and $Same(\alpha_i)$, respectively. Specifically, the constraint α_i in $Prior(\alpha_i)$ is noted by $Const(i)$. Here, $Const(i)$ is computed by $log(1 + \#Publication_i)$ (i.e., $\#Publication_i$ is the total number of publications of node i) for DBLP. As we have no such content information for the rest of the networks, we simply set $Const(i)$ to be proportional to $\log(1 + Degree_i)$ for the nodes in Epinions, Flixster and WebND. Meanwhile, to better test the impact of different constraints, we also randomly generate some constraints and get another baseline $Random(\alpha_i)$.

Finally, from Theorem 4.2, we can determine that the upper bounds for the $Same(\alpha)$ method are linear to PageRank, and the upper bounds for the $Prior(\alpha)$ method are

[7]http://www.sfu.ca/~sja25/datasets/flixster.zip.
[8]http://snap.stanford.edu/data/web-NotreDame.html.

linear to WPageRank. Thus, the corresponding ranking results of these upper bounds should be the same as PageRank and WPageRank, respectively. For each method, we choose the same $\lambda = 0.176$, since $d = \frac{1}{1+\lambda} = 0.85$.

Evaluation metrics: Directly evaluating the results is not easy since there is no general way to get the ground truth of node ranking based on authority. As an alternative, we refer to the research of social influence analysis, where the output of the WC influence model [Kempe et al. 2003] is often chosen as the evaluation metric [Jung et al. 2012; Liu et al. 2013]. The major reason is that as a kind of the IC model [Yang et al. 2012], WC is the most widely accepted for simulating the influence propagation process. In each task, given the selected nodes (node sets) by different methods, we run a Monte-Carlo simulation under the WC model for sufficiently many (e.g., 20,000) times and sum up the influence spread (i.e., the expected number of nodes that will be influenced) as an estimation of the real influence/authority. We then evaluate the performance of each method based on the correlations between their output and the WC model. That is, we assume higher ranking correlations lead to better ranking results.

We also choose H-index to measure the ranking results on DBLP. Though there are several limitations for evaluating researchers by H-index, we do so for the following reasons: First, H-index can measure both quality and quantity of the published works of researchers, and we do not include the citation information for constructing the DBLP network; Second, H-index is well accepted and widely used, e.g., in bibliometric analysis [Ding et al. 2009] and network role analysis [Jin et al. 2014]; Third, we have observed the positive correlations between H-indexes and researcher authorities (which also proves the similar assumption in Subbian et al. [2014]). For instance, the average spearman correlation between the ranking lists based on H-indexes and PageRank values of 20 randomly selected researchers is nearly 0.4.

6.2. Selection of Top-*K* Nodes

First, we show a case study by illustrating the names of the Top-10 authoritative researchers in each research domain of DBLP network in Table V, where "Total" means the entire DBLP collaboration network. In Table V, we can see that the results contain influential researchers from different research domains. Even though the methods (or constraints) are different from each other, the authoritative nodes determined are quite similar [Aggarwal et al. 2011]. Meanwhile, the results obtained by *Random*(α) are comparatively different from others, but its outputs are also well-known researchers that demonstrate that not only the constraint but also the network structure contribute to the final result. Furthermore, Table VI lists the average H-index results for Top-50 ranked researchers, where the H-indexes are collected simultaneously in May, 2012. In Table VI, we can see that the methods considering reasonable prior knowledge (e.g., *Prior*(α), *Same*(α) and WPageRank) generally perform better than those that do not (i.e., PageRank and *Random*(α)), and *Prior*(α) could figure out the top researchers with the highest H-indexes.

Next, we evaluate each method's effectiveness based on influence spread. Since it is very time consuming to run the Monte-Carlo simulation under the WC model, we design the following evaluation strategy. We first run the PageRank algorithm on each network and get the Top-50 nodes, then rank these nodes by other methods and the WC influence model, respectively. Finally, we compute the Spearman correlations[9] and the Kendall's tau coefficients[10] between each ranking list with the output of the WC model, the results for which are shown in Figure 2. From this figure, we can see that the output

[9]http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient.
[10]http://en.wikipedia.org/wiki/Kendall_tau_rank_correlation_coefficient.

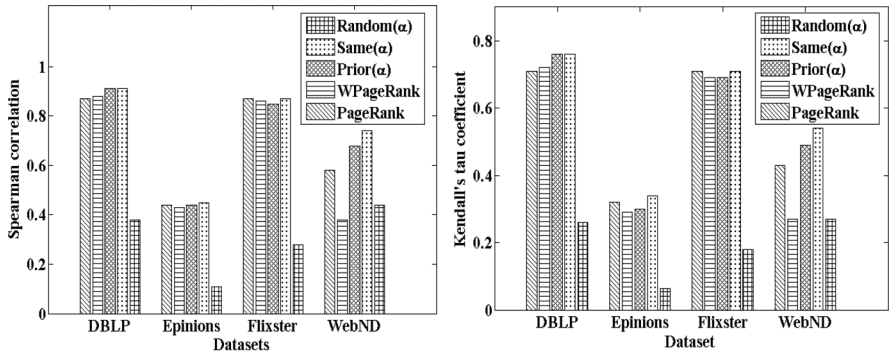
Table V. Each Domain's Top-10 Researchers in DBLP Data that are Mined by the Methods with Different Constraints

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	S. Kraus	C. Boutilier	V. R. Lesser	Q. Yang	D. Koller	R.Dechter	M. L. Littman	R. J. Mooney	T. Walsh
	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	Q. Yang	C. Boutilier	D. Koller	M. L. Littman	V. R. Lesser	T. Walsh	M. Tambe	S. Thrun
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	A. K. Jain	S. K. Nayar	T. Kanade	L. J. Van Gool	R. Chellappa	N.Ahuja	A. Zisserman	L. S. Davis	M. Shah
	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. K. Jain	A. Zisserman	L. J. Van Gool	S. K. Nayar	R. Chellappa	N. Ahuja	L. S. Davis	X. Tang
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	C. Faloutsos	M. Stonebraker	E. A. Rundensteiner	D. J.DeWitt	G. Weikum	R. Agrawal	S. Chaudhuri
	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	C. Faloutsos	M. Stonebraker	D. Srivastava	D. J. DeWitt	R. Agrawal	H. V. Jagadish	S. Chaudhuri
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	Q. Yang	M. Chen	J. Pei	E. J. Keogh	H. P. Kriegel	C. C. Aggarwal	K. Wang
	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	Q. Yang	J.Pei	C. C. Aggarwal	M. Chen	K. Wang	H. Kriegel	H. Garcia-Molina
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	P. S. Yu	W. B. Croft	J. Han	H. Garcia-Molina	C. Faloutsos	Q. Yang	G. Weikum	C. T. Yu	C. L. Giles	K. Tanaka
	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	C. Faloutsos	Z. Chen	G. Weikum	W. Ma	C. L. Giles
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	D. Koller	T. J. Sejnowski	G. E. Hinton	Y. Bengio	B. Schölkopf	A. W. Moore	Z.Ghahramani	C. Koch	A. Y. Ng
	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	D. Koller	G. E. Hinton	T. J. Sejnowski	Y. Bengio	Z. Ghahramani	A. Y. Ng	J.Shawe-Taylor	K. Müller
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	W. B. Croft
	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	H. Garcia-Molina	M. I.Jordan	Q. Yang	T. Kanade	D. Srivastava	R. Agrawal

Table VI. The Average H-Indexes for Top-50 Researchers (DBLP)

	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	44.10	46.34	52.12	**46.32**	42.42	42.18	56.60	47.15
Same(α)	43.06	45.40	52.92	43.30	42.44	38.46	**57.00**	46.08
Prior(α)	**45.24**	**47.02**	**53.08**	45.24	**44.30**	**43.64**	56.70	**47.87**
Random(α)	39.70	39.80	46.90	39.60	41.58	38.80	52.56	42.70

The values given in bold are used to highlight the algorithms with the best performance.



(a) Spearman. (b) Kendall's Tau.

Fig. 2. Ranking results comparison for Top-50 nodes.

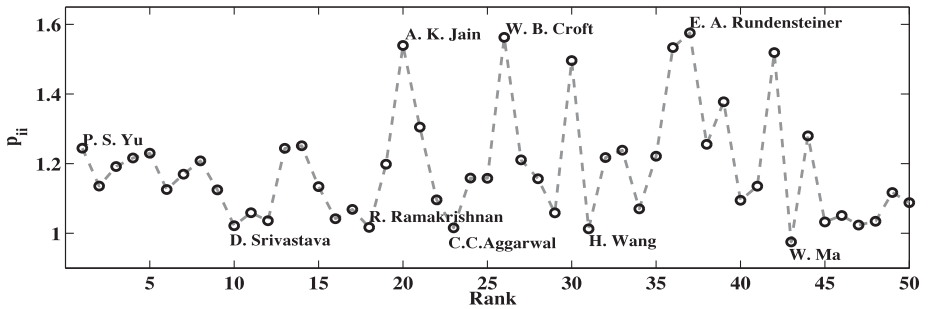


Fig. 3. The p_{ii} values of the Top-50 Researchers (DBLP).

of Spearman correlations (Figure 2(a)) and Kendall's tau coefficient (Figure 2(b)) are very similar. Generally, *Prior*(α) and *Same*(α) have the same performance for DBLP, and *Same*(α) performs best for Epinions, Flixster and WebND, indicating that without useful prior knowledge it is more reasonable to assign α_i to be the same. Though there may be bias in the constraints (i.e., the simple *Const*(i) for *Prior*(α)), *Prior*(α) still outperforms WPageRank, PageRank and *Random*(α).

To better understand the constraints automatically chosen in PageRank, Figure 3 presents the p_{ii} values (PageRank constraints) of the Top-50 researchers (ordered by nodes' degrees) from DBLP and it also gives the names of the researchers at both ends (those with the highest or lowest constraints). From this figure, it is intractable for us to find meaningful patterns, and the results again demonstrate that it is not the best choice to use these values as the constraint α_i for evaluating the importance of each candidate node. For instance, it is improper to set the constraint of Dr. *Charu C. Aggarwal* much lower. However, this observation does help us understand the results in Table V.

6.3. Combined Influence Analysis

In this subsection, we pay attention to the analysis of combined influence for node sets. Without loss of generality, we fix the size of each set (i.e., $|S|$) as 2 or 3 in the experiment.

Following Li et al. [2011], we also design a case study for influence overlap analysis. First, we manually select one famous researcher from each domain of the DBLP data (i.e., *T. Sandholm*, *T. S. Huang*, *H. Garcia-Molina*, *P. S. Yu*, *W. B. Craft* and *M. I. Jordan*), and find the Top-8 co-authors (based on the numbers of co-authored papers) for each of them. Then, we compose many node sets with these researchers and their collaborators, i.e., each node set contains one chosen researcher and one of his collaborators (e.g., $S = (T.\ Sandholm, V.\ Conitzer)$). Next, we rank these nodes sets (more specifically, the collaborators) by their authority reduction ($\sum_{i \in S} \mathbf{f}_{i \to \mathcal{V}} - f_{S \to \mathcal{V}}$, noted as *Red.*), *IOR* (Equation (25)), number of co-authored papers (CoAuthors) and number of co-neighbors (CoNeighbors), respectively. The final results are shown in Table VII, where each row represents the ranking list of one corresponding method (e.g., $Same(\alpha)$) under different metrics (Tasks, e.g., *Red.*). Notice that the scalar in each bracket (.) of the $Same(\alpha)$) output stands for the *IOR* value (e.g., the computed *IOR* value for the set of (*T. Sandholm*, *V. Conitzer*) is 0.29). Here, we put the ranking lists of CoAuthors and CoNeighbors into the task of *Red.*[11] for better comparing the Spearman correlation results between each ranking list and the ranking of the WC model (shown in the last column). For simplicity, we just give the final results (Table VII) by the $Same(\alpha)$ ranking method (Actually, $Prior(\alpha)$ and $Same(\alpha)$ perform quite similarly for this task). Note that both PageRank and WPageRank cannot be adopted for this comparison since they cannot consider the authority overlap of the nodes. From this table, we have the following observations: First, the rank of the collaborators based on $Same(\alpha)$ is very similar to the output of WC for both *Red.* and *IOR*, while the output of CoAuthors and CoNeighbors is quite different; Second, for the same method (e.g., $Same(\alpha)$, WC), the ranking lists with respect to different tasks/metrics (i.e., *Red.* or *IOR*) are also different, which implies that the node sets with the maximum authority reduction may not lead to the biggest *IOR*.

Next, we show a more comprehensive comparison on combined influence by mining top authoritative node sets. We first illustrate the Top-5 authoritative researcher sets ($|S| = 2$) from DBLP in Table VIII, where most of the researchers in each selected set are authoritative researchers in each domain. The results of PageRank and WPageRank are based on Equation (14), and we do not give the results of $Random(\alpha)$ due to limited space. Though the top researcher sets of different methods are also different, it is still difficult to directly make conclusive judgments on the performance of each method. Therefore, we refer to the average H-index results. We first compute the "H-index" of each researcher set, where the researchers in this set are viewed as a single researcher, and if two researchers have collaborated on one paper, we only count the citation of this paper once. Table IX shows the average H-index results for Top-50 researcher sets ($|S| = 2$). Similar to Table VI, the results in Table IX also demonstrate that the combined influence by $Prior(\alpha)$ can measure the real authority of each researcher set much better, as the chosen sets have higher H-indexes than that of other methods. However, compared with Table VI, the improvements of $Prior(\alpha)$ shown in Table IX are even more obvious, and here $Same(\alpha)$ also outperforms PageRank and WPageRank.

Finally, we report the quantitative evaluations on four datasets. We first run PageRank algorithm and then get the top node sets. Next, we rank these node sets by other methods and the WC influence model, respectively. The Spearman correlations and

[11]We do not put CoAuthors and CoNeighbors into *IOR* since both the methods cannot measure influence overlap rate.

Table VII. Case Study of the Combined Influence in Terms of the Top Collaborators for Each Researcher

Researcher	Task	Methods	Top-8 co-authors								Correlation
Sandholm (AI)	*Red.*	Same(α)	V. Conitzer	V. R. Lesser	A. Gilpin	K. Larson	S. Suri	M. Andersson	C. Boutilier	M. Benisch	0.88
		WC	V. Conitzer	K. Larson	V. R. Lesser	A. Gilpin	M. Andersson	S. Suri	M. Benisch	C. Boutilier	–
		CoAuthors	V. Conitzer	A. Gilpin	K. Larson	V. R. Lesser	C. Boutilier	M. Benisch	S. Suri	M. Andersson	0.64
		CoNeighbors	V. Conitzer	C. Boutilier	A. Gilpin	K. Larson	M. Andersson	M. Benisch	S. Suri	V. R. Lesser	0.38
	IOR	Same(α)	Conitzer(0.29)	Gilpin(0.13)	Larson(0.10)	Lesser(0.10)	Suri(0.08)	Andersson(0.07)	Benisch(0.05)	Boutilier(0.03)	**0.90**
		WC	V. Conitzer	K. Larson	A. Gilpin	M. Andersson	V. R. Lesser	S. Suri	M. Benisch	C. Boutilier	–
Huang (CV)	*Red.*	Same(α)	Y. Wu	N. Ahuja	S. Yan	J. Weng	Y. Fu	L. Cao	J. Han	J. Yang	0.57
		WC	Y. Wu	Y. Fu	N. Ahuja	J. Weng	J. Yang	L. Cao	S. Yan	J. Han	–
		CoAuthors	S. Yan	Y. Wu	L. Cao	N. Ahuja	J. Yang	Y. Fu	J. Han	J. Weng	0.048
		CoNeighbors	S. Yan	L. Cao	J. Han	J. Yang	Y. Wu	Y. Fu	N. Ahuja	J. Weng	−0.69
	IOR	Same(α)	Wu(0.10)	Ahuja(0.05)	Yan(0.04)	Weng(0.04)	Fu(0.03)	Cao(0.03)	Yang(0.02)	Han(0.02)	**0.64**
		WC	Y. Wu	Y. Fu	N. Ahuja	J. Weng	J. Yang	L. Cao	S. Yan	J. Han	–
Garcia-Molina (DB)	*Red.*	Same(α)	K. Salem	J. Widom	K. Chang	A. Tomasic	W. Labio	G. Koutrika	N. Shivakumar	A. Paepcke	0.73
		WC	K. Salem	W. Labio	K. Chang	J. Widom	A. Tomasic	A. Paepcke	N. Shivakumar	G. Koutrika	–
		CoAuthors	J. Widom	A. Paepcke	G. Koutrika	W. Labio	K. Salem	K. Chang	A. Tomasic	N. Shivakumar	0
		CoNeighbors	J. Widom	G. Koutrika	A. Paepcke	N. Shivakumar	W. Labio	K. Chang	K. Salem	A. Tomasic	−0.5
	IOR	Same(α)	Salem(0.11)	Chang(0.07)	Tomasic(0.07)	Widom(0.07)	Labio(0.05)	Shivakumar(0.05)	Paepcke(0.05)	G. Koutrika(0.05)	**0.76**
		WC	K. Salem	W. Labio	K. Chang	A. Tomasic	A. Paepcke	J. Widom	N. Shivakumar	G. Koutrika	–
Yu (DM)	*Red.*	Same(α)	C. Aggarwal	M. Chen	J. Han	H. Wang	K. Wu	W. Fan	X. Yan	B. Gedik	0.43
		WC	C. Aggarwal	M. Chen	W. Fan	K. Wu	B. Gedik	H. Wang	X. Yan	J. Han	–
		CoAuthors	C. Aggarwal	J. Han	H. Wang	M. Chen	K. Wu	W. Fan	X. Yan	B. Gedik	0.19
		CoNeighbors	H. Wang	J. Han	W. Fan	C. Aggarwal	X. Yan	K. Wu	M. Chen	B. Gedik	−0.33
	IOR	Same(α)	Aggarwal(0.21)	Chen(0.13)	Han(0.09)	Wang(0.07)	Wu(0.07)	Fan(0.06)	Yan(0.05)	Gedik(0.04)	**0.43**
		WC	C. Aggarwal	M. Chen	W. Fan	K. Wu	B. Gedik	H. Wang	X. Yan	J. Han	–
Croft (IR)	*Red.*	Same(α)	D. Metzler	M. Bendersky	J. Xu	H. Turtle	J. Callan	J. Seo	X. Xue	T. Strohman	0.59
		WC	J. Xu	H. Turtle	D. Metzler	M. Bendersky	T. Strohman	J. Seo	X. Xue	J. Callan	–
		CoAuthors	M. Bendersky	D. Metzler	H. Turtle	J. Callan	J. Xu	J. Seo	X. Xue	T. Strohman	0.38
		CoNeighbors	J. Callan	D. Metzler	H. Turtle	M. Bendersky	J. Seo	T. Strohman	J. Xu	X. Xue	−0.07
	IOR	Same(α)	Metzler(0.12)	Bendersky(0.12)	Xu(0.10)	Turtle(0.10)	Seo(0.09)	Xue(0.08)	Callan(0.07)	Strohman(0.05)	**0.67**
		WC	J. Xu	H. Turtle	D. Metzler	M. Bendersky	T. Strohman	J. Seo	X. Xue	J. Callan	–
Jordan (ML)	*Red.*	Same(α)	F. Bach	A. Ng	T. Jaakkola	L. Saul	Z. Ghahramani	D. Blei	Z. Zhang	G. Lanckriet	0.60
		WC	F. Bach	D. Blei	A. Ng	L. Saul	T. Jaakkola	G. Lanckriet	Z. Zhang	Z. Ghahramani	–
		CoAuthors	F. Bach	T. Jaakkola	A. Ng	Z. Zhang	Z. Ghahramani	L. Saul	G. Lanckriet	D. Blei	0.19
		CoNeighbors	D. Blei	Z. Ghahramani	F. Bach	T. Jaakkola	A. Ng	G. Lanckriet	Z. Zhang	L. Saul	0.26
	IOR	Same(α)	Bach(0.12)	Ng(0.08)	Jaakkola(0.08)	Saul(0.08)	Ghahramani(0.06)	Blei(0.06)	Zhang(0.05)	Lanckriet(0.03)	**0.64**
		WC	F. Bach	D. Blei	A. Ng	L. Saul	T. Jaakkola	G. Lanckriet	Z. Ghahramani	Z. Zhang	–

The values given in bold are used to highlight the algorithms with the best performance.

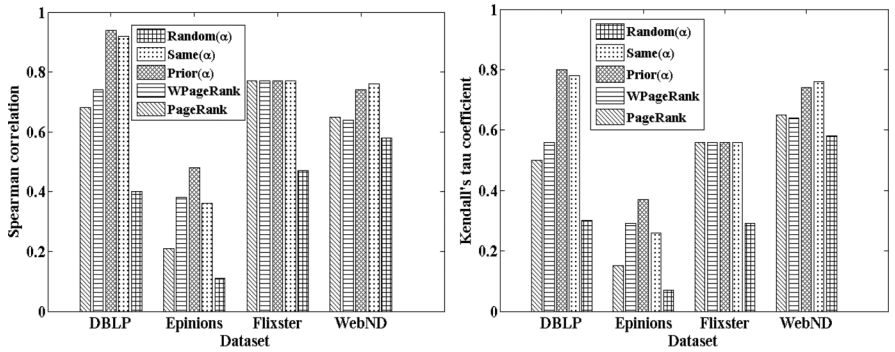
Table VIII. An Illustration of Each Domain's Top-5 Researcher Sets ($|S| = 2$) Mined by Different Methods

Domain	Met.	Top-5 Pairs				
AI	PageRank	(T. Sandholm, S. Kraus)	(T. Sandholm, V. R. Lesser)	(S. Kraus, V. R. Lesser)	(T. Sandholm, C. Boutilier)	(T. Sandholm, K. D. Forbus)
	WPageRank	(T. Sandholm, V. R. Lesser)	(T. Sandholm, C. Boutilier)	(T. Sandholm, V. R. Lesser)	(T. Sandholm, Q. Yang)	(T. Sandholm, D. Koller)
	Same(α)	(T. Sandholm, S. Kraus)	(S. Kraus, V. R. Lesser)	(S. Kraus, C. Boutilier)	(S. Kraus, M. L. Littman)	(T. Sandholm, M. L. Littman)
	Prior(α)	(T. Sandholm, S. Kraus)	(S. Kraus, Q. Yang)	(T. Sandholm, Q. Yang)	(S. Kraus, J. Han)	(S. Kraus, C. Boutilier)
CV	PageRank	(T. S. Huang, A. K. Jain)	(T. S. Huang, S. K. Nayar)	(T. S. Huang, T. Kanade)	(T. S. Huang,R. Chellappa)	(T. S. Huang, N. Ahuja)
	WPageRank	(T. S. Huang, A. K. Jain)	(T. S. Huang, S. K. Nayar)	(T. S. Huang, T. Kanade)	(T. S. Huang, N. Ahuja)	(T. S. Huang, R. Chellappa)
	Same(α)	(T. S. Huang, T. Kanade)	(T. S. Huang, A. K. Jain)	(T. S. Huang, A. Zisserman)	(T. S. Huang, L. J. Van Gool)	(T. S. Huang, S. K. Nayar)
	Prior(α)	(T. S. Huang, T. Kanade)	(T. S. Huang, A. K. Jain)	(T. S. Huang, A. Zisserman)	(T. S. Huang, L. J. Van Gool)	(T. S. Huang, S. K. Nayar)
DB	PageRank	(P. S. Yu, J. Han)	(P. S. Yu, H. Garcia-Molina)	(P. S. Yu, M. Stonebraker)	(J. Han, H. Garcia-Molina)	(P. S. Yu, E. A. Rundensteiner)
	WPageRank	(P. S. Yu, J. Han)	(P. S. Yu, H. Garcia-Molina)	(P. S. Yu, M. Stonebraker)	(P. S. Yu, C. Faloutsos)	(J. Han, H. Garcia-Molina)
	Same(α)	(P. S. Yu, H. Garcia-Molina)	(J. Han, H. Garcia-Molina)	(P. S. Yu, J. Han)	(P. S. Yu, M. Stonebraker)	(J. Han, M. Stonebraker)
	Prior(α)	(P. S. Yu, J. Han)	(P. S. Yu, H. Garcia-Molina)	(J. Han, H. Garcia-Molina)	(P. S. Yu, C. Faloutsos)	(P. S. Yu, M. Stonebraker)
DM	PageRank	(P. S. Yu, J. Han)	(P. S. Yu, C. Faloutsos)	(P. S. Yu, M. Chen)	(P. S. Yu, Q. Yang)	(J. Han, C. Faloutsos)
	WPageRank	(P. S. Yu, J. Han)	(P. S. Yu, C. Faloutsos)	(P. S. Yu, Q. Yang)	(J. Han, C. Faloutsos)	(P. S. Yu, M. Chen)
	Same(α)	(P. S. Yu, J. Han)	(P. S. Yu, C. Faloutsos)	(J. Han, C. Faloutsos)	(P. S. Yu, Q. Yang)	(P. S. Yu, J. Pei)
	Prior(α)	(P. S. Yu, J. Han)	(P. S. Yu, C. Faloutsos)	(J. Han, C. Faloutsos)	(P. S. Yu, Q. Yang)	(P. S. Yu, J. Pei)
IR	PageRank	(W. B. Croft, P. S. Yu)	(W. B. Croft, J. Han)	(P. S. Yu, J. Han)	(W. B. Croft, H. Garcia-Molina)	(W. B. Croft, K. Tanaka)
	WPageRank	(W. B. Croft, P. S. Yu)	(W. B. Croft, J. Han)	(P. S. Yu, J. Han)	(P. S. Yu, H. Garcia-Molina)	(W. B. Croft, H. Garcia-Molina)
	Same(α)	(W. B. Croft, P. S. Yu)	(W. B. Croft, J. Han)	(P. S. Yu, J. Han)	(P. S. Yu, H. Garcia-Molina)	(J. Han, H. Garcia-Molina)
	Prior(α)	(W. B. Croft, P. S. Yu)	(P. S. Yu, J. Han)	(W. B. Croft, J. Han)	(P. S. Yu, H. Garcia-Molina)	(J. Han, H. Garcia-Molina)
ML	PageRank	(M. I. Jordan, T. J. Sejnowski)	(M. I. Jordan, Y. Bengio)	(M. I. Jordan, C. Koch)	(M. I. Jordan, D. Koller)	(M. I. Jordan, G. E. Hinton)
	WPageRank	(M. I. Jordan, D. Koller)	(M. I. Jordan, T. J. Sejnowski)	(M. I. Jordan, G. E. Hinton)	(M. I. Jordan, Y. Bengio)	(M. I. Jordan, B. Schölkopf)
	Same(α)	(M. I. Jordan, B. Schölkopf)	(M. I. Jordan, T. J. Sejnowski)	(M. I. Jordan, D. Koller)	(M. I. Jordan, G. E. Hinton)	(M. I. Jordan, Y. Bengio)
	Prior(α)	(M. I. Jordan, B. Schölkopf)	(M. I. Jordan, D. Koller)	(M. I. Jordan, T. J. Sejnowski)	(M. I. Jordan, G. E. Hinton)	(M. I. Jordan, Y. Bengio)
Total	PageRank	(P. S. Yu, J. Han)	(P. S. Yu, C. Faloutsos)	(P. S. Yu, T. S. Huang)	(P. S. Yu, H. Garcia-Molina)	(J. Han, C. Faloutsos)
	WPageRank	(P. S. Yu, J. Han)	(P. S. Yu, C. Faloutsos)	(P. S. Yu, T. S. Huang)	(P. S. Yu, H. Garcia-Molina)	(J. Han, C. Faloutsos)
	Same(α)	(P. S. Yu, J. Han)	(P. S. Yu, C. Faloutsos)	(J. Han, C. Faloutsos)	(P. S. Yu, T. S. Huang)	(P. S. Yu, H. Garcia-Molina)
	Prior(α)	(P. S. Yu, J. Han)	(P. S. Yu, C. Faloutsos)	(J. Han, C. Faloutsos)	(P. S. Yu, T. S. Huang)	(P. S. Yu, H. Garcia-Molina)

Table IX. The Average H-Indexes for Top-50 Researcher Sets (DBLP)

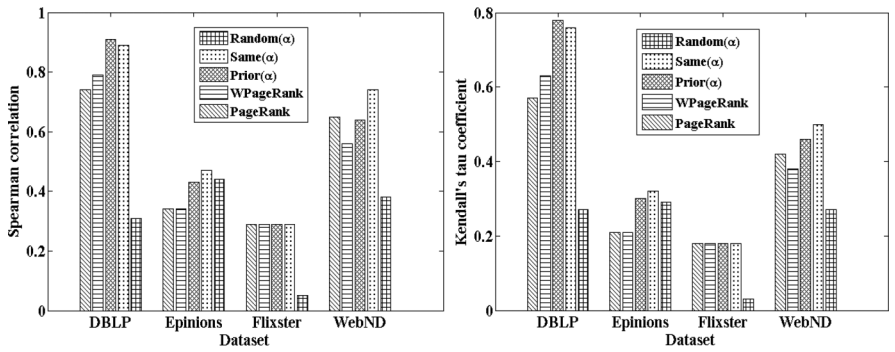
	AI	CV	DB	DM	IR	ML	Total	Ave.
PageRank	46.92	64.32	83.28	71.22	58.38	61.46	80.18	66.53
WPageRank	52.22	**64.46**	82.04	71.26	69.52	64.42	79.92	69.12
Same(α)	50.32	64.26	**85.74**	72.88	72.94	64.40	**84.58**	70.73
Prior(α)	**63.70**	63.62	84.36	**74.66**	**76.08**	**65.02**	**84.58**	**73.14**
Random(α)	48.18	59.76	83.56	71.86	63.48	53.2	80.18	65.74

The values given in bold are used to highlight the algorithms with the best performance.



(a) Spearman. (b) Kendall's Tau.

Fig. 4. Ranking comparison for Top-50 node sets ($|S| = 2$).



(a) Spearman. (b) Kendall's Tau.

Fig. 5. Ranking comparison for Top-30 node sets ($|S| = 3$).

Kendall's tau coefficients between the ranking list of each method and the ranking list of WC model are shown in both Figures 4 and 5. Compared with Figure 2, the difference between the performance of each method becomes much clearer, and *Prior*(α) and *Same*(α) still perform best for this task. Another observation is that four methods perform similarly on the Flixster data, and we find that the ranking lists of each method (PageRank, WPageRank, *Prior*(α), *Same*(α)) are almost the same. The reason is that the constraints selected using these methods on this dataset are quite similar. Meanwhile, we should note that when the network is very large or sparse, it is possible that the top nodes identified by different methods (even the selections based on the nodes' degrees) are quite similar, as the authority differences among the top nodes are usually very significant.

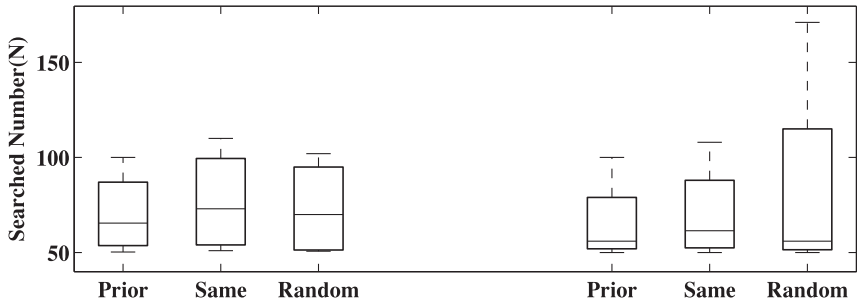
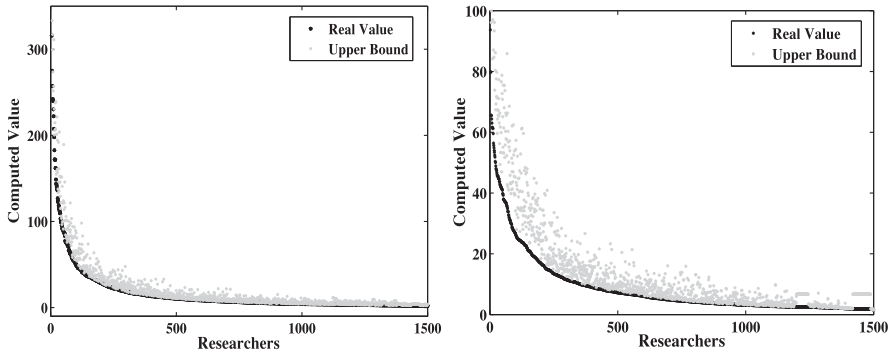


Fig. 6. Searched number (N) for finding Top-50 nodes (left) and node sets (right) from four networks.



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

Fig. 7. An illustration of the upper bounds.

6.4. Upper Bounds Evaluation

We demonstrate the effectiveness of the upper bounds, i.e., $(1+\lambda)\alpha_i p_i$ (Theorem 4.2) for single node's authority and $\sum_{i\in S}(1+\lambda)\alpha_i p_i$ (Theorem 5.2) for combined influence.

We first run $Prior(\alpha)$, $Same(\alpha)$ and $Random(\alpha)$ on four networks to quickly select the top influential nodes and node sets, respectively. We then present the searched number of the candidates (N) for finding the Top-50 nodes (node sets with $|S| = 2$) in Figure 6. We can observe that this number is usually quite small (no more than 200) with respect to the entire search space (n), which indicates that the corresponding algorithms (e.g., Algorithm 2) for searching top nodes (node sets) are scalable. For better understanding, we also illustrate the true authority value (computed by $Prior(\alpha)$ and $Same(\alpha)$ respectively) and the upper bounds for 1,500 randomly selected researchers from DBLP in Figure 7, where the researchers are ranked by their true authorities. From Figure 7, we observe that the upper bounds are always close to the real authority values. Combining the results in Figures 6 and 7, we can conclude that the upper bounds we designed are effective.

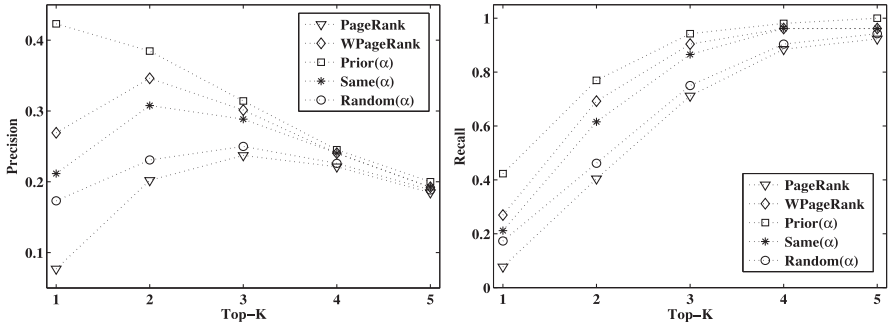
6.5. Combined Influence for Recommendation

We also evaluate the effectiveness of each ranking method and the combined influence through a task of recommending editorial/organizing committees. As is well known, the committee members for each organization should be diversified for covering the interests and requirements from various members. Take academic organizations as an example. The members in each editorial/organizing committee are better contributors who are far away from each other in the collaboration network, i.e., if two researchers

Table X. User Study Ratings

	PageRank	WPageRank	Prior(α)	Same(α)	Random(α)
Ave. rank	2.05	1.21	**0.88**	1.38	1.79
SD	104.95	61.79	**45.12**	70.62	91.22

The values given in bold are used to highlight the algorithms with the best performance.

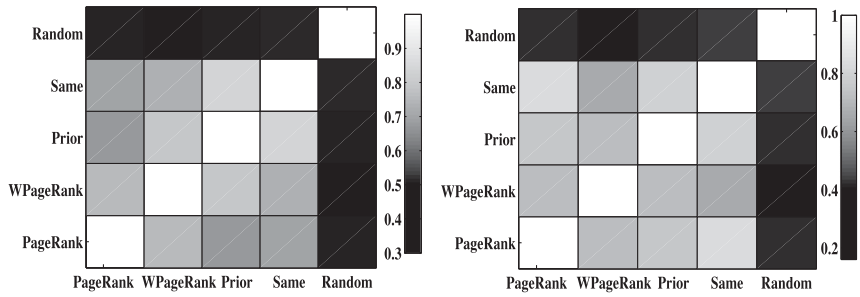


(a) Precision Result. (b) Recall Result.

Fig. 8. Results for committee recommendation.

have much authority/influence overlap, they are usually not put into the same committee. To select a committee from a number of node groups (i.e., committee candidates), the recommendation method should rank each group by measuring the combined importance of their group members. Thus, the method that can measure the combined importance much better will achieve a higher recommendation accuracy.

Along this line, we first manually collect 52 committees from the official websites of some leading journals/conferences in Table III. On average, each committee includes 18 members ($|S| = 18$, and we have made this data publicly available). Then, for each given committee, we generate many dummy groups, each of which is composed of one committee member and several of his/her neighbors, making sure that the authority/influence of the dummy groups are lower than the real committee. After that, we compute the authority of both the real committee and the dummy ones using each method, and rank these user groups according to their authority. At last, the method that can rank the real committee higher is the better one. For the evaluation, we use the average rank of the real committee as the evaluation metric (the smaller the rank, the better the method, e.g., Rank 0 means the real committee is ranked higher than all the dummy ones). This result and its standard deviations (SD) for each ranking method are shown in Table X, from which we can see that $Prior(\alpha)$ performs best and it is followed by $Same(\alpha)$ and WPageRank. Meanwhile, $Random(\alpha)$ outperforms PageRank for this task, implying that when the node set is large, the combined influence computation is still effective even with random constraints since it can handle the authority overlap issue. To better evaluate these ranking lists, we further choose "Precision" and "Recall" as the metrics, and the corresponding results are shown in Figure 8. Not surprisingly, we can observe that the ranking lists based on $Prior(\alpha)$ are much better than others. Combining the results in Table X and Figure 8, WPageRank (with constraint $p_{ii} * Const(i)$) outperforms $Same(\alpha)$ (with constraint 1), and this also demonstrates the positive effect of one reasonable prior constraint/weight. At last, by applying z-test, we find the differences between different ranking lists are statistically significant, for instance, the z-test results between $Prior(\alpha)$ and PageRank is $|z| = 3.85$ and thus $p \leq 0.05$. This experimental study once again proves that the combined influence can discover the node sets with higher importance.



(a) Top-10 Nodes. (b) Top-10 Node Sets.

Fig. 9. The Jaccard similarity between different methods.

6.6. Correlation Demonstration

For understanding the correlations of the output of different methods, Figure 9 shows the Jaccard similarity of the top nodes (Figure 9(a)) and the node sets (Figure 9(b)) output by different methods. Specifically, we first run each method on the four datasets and then select the Top-10 nodes (node sets with $|S| = 2$) for each method from one dataset. Then, all four networks' Top-10 nodes (node sets) gained by the given method are used to stand for this specific method. Thus, the Jaccard similarity is computed between each of the 40 nodes or node sets. In Figure 9, the darker the color between two methods, the smaller the similarity of their outputs. Similar results can be observed from both Figures 9(a) and (b). For instance, the results demonstrate that the output of the ranking methods (i.e., PageRank, WPageRank, $Prior(\alpha)$ and $Same(\alpha)$) are somewhat similar to each other as discovered previously; Meanwhile, the most distinctive method is $Random(\alpha)$, and the second most dissimilar two pairs of methods are (PageRank, $Prior(\alpha)$) and (WPageRank, $Same(\alpha)$), due to the difference in their constraints.

7. DISCUSSION

In this section, we discuss the significance, limitations and future research directions of this study.

From both the theoretical analysis and experimental validation, we can see that the proposed linear social influence modeling approach can effectively and efficiently measure node importance in large-scale social networks. Also, we reveal the close relationship between PageRank and social influence modeling by introducing prior influence constraint (e.g., α_i) for each node and node set. Generally, the more reasonable constraints (learnt from the prior knowledge, e.g., the number of publications for each researcher in DBLP network) lead to the better node ranking and recommendation results, and when lacking of prior knowledge, it is a good choice to fix these constraints to be the same (as shown in the experimental results on Epinions, Flixster and WebND). This finding can be further applied to help other PageRank related methods, such as TwitterRank [Weng et al. 2010], to measure the influence of a single node and the combined influence of a node set (i.e., by including prior knowledge into these methods) much better. Moreover, we propose the idea of deriving upper bounds on the social influence of either a node or a set of nodes for general constraints. With respect to different types of constraints and upper bounds, the time complexity for each computation/ranking task is shown in Table XI, where n_S is the number of candidate node sets. Since there are so many different ways to get α_i (e.g., the ones in Table IV) and

Table XI. Time Complexity for Each Task

Tasks / α_i $(i \in S)$	Influence computation		Top-K selection	
	$f_{S\to j}$	$f_{S\to\mathcal{V}}$	Single nodes	Node sets
$\alpha_i \not\propto p_{ii}$	$O(\|S\|\|\mathcal{A}\|)$	$O(\|S\|\|\mathcal{A}\|)$	$O((N+1)\|\mathcal{A}\|)$	$O(N(\|S\|^2+\|S\|\|\mathcal{A}\|))$
$\alpha_i \propto p_{ii}$	$O(\|S\|\|\mathcal{A}\|)$	$O(\|\mathcal{A}\|)$	$O(\|\mathcal{A}\|+n\log_2 K)$	$O(\|\mathcal{A}\|+n_S\log_2 K)$

for better illustration, we omit the time consumption in this procedure.[12] Specifically, in terms of both computing and applying the global influence value (e.g., computing $f_{S\to\mathcal{V}}$ and selecting Top-K nodes/node sets), the PageRank way of setting constraints (i.e., $\alpha_i \propto p_{ii}$) does help reduce the time consumption significantly. Unfortunately, we have to sacrifice effectiveness under this setting (shown in the experiments). As a tradeoff, we can exploit upper bounds for reasonable constraints (e.g., those $\alpha_i \not\propto p_{ii}$) so as to develop efficient algorithms to identify more influential nodes (node sets). These upper bounds can be further used in other applications of social influence analysis. For instance, by exploiting the properties of the linear social influence model and the upper bounds based on PageRank, Liu et al. proposed two algorithms to quickly find a set of the most influential nodes (i.e., a set of seeds) to deal with the social influence maximization problem in viral marketing [Liu et al. 2014].

However, this study still has limitations. First, we do not consider some other factors in practice. For instance, the temporal effects may be included since individuals may have diversified interests/concerns and thus can be influenced by different users at different time periods. Second, this article focuses on demonstrating the effect of different constraints on the performance of linear model, and these constraints (i.e., α_i) are simply determined. Though the experimental results have illustrated that it is reasonable and safe to assign α_i to be the same when useful prior knowledge is missing, we believe it is worth finding more effective constraints given some prior knowledge. In other words, it is possible to develop a more general solution for constraint selection in different scenarios (data). In the future, we plan to design a general approach for combining many internal (e.g., the personal interests of the user) and external (e.g., spaital and temporal dimensions) factors into the model to better understand the social influence mechanism.

8. CONCLUSION

In this article, we have provided a systematic study of PageRank and authority from an influence propagation perspective. Along this line, we first develop a linear social influence model, which generalizes the PageRank-based authority computation by introducing prior influence constraints. Also, we reveal that the authority of each node is essentially the collection of its influence on the network or a specific subnetwork. Furthermore, we show 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 find that the PageRank value can be used to form an upper bound for efficiently computing the most authoritative nodes. After that, we give the definition and computation of the combined influence of a set of nodes based on the proposed linear model, and also present the properties and applications for this combined influence. Finally, we empirically evaluate the above discoveries on real-world network datasets. Experimental results show the effect of different constraints on the performance of linear model, i.e., the more reasonable constraints lead to the better ranking results, and the proposed upper bounds can be used for quickly locating the Top-K nodes and node sets. We hope this study

[12]Actually, this time consumption could also be linearly added into the entries in Table XI.

can lead to more future work in the areas of both node ranking and social influence modeling.

REFERENCES

C. C. Aggarwal. 2011. *Social Network Data Analytics*. Springer.

C. C. Aggarwal, A. Khan, and X. Yan. 2011. On flow authority discovery in social networks. In *Proceedings of SIAM Conference on Data Mining (SDM'11)*. 522–533.

Réka Albert, Hawoong Jeong, and Albert-László Barabási. 1999. Internet: Diameter of the world-wide web. *Nature* 401, 6749 (1999), 130–131.

Lars Backstrom and Jure Leskovec. 2011. Supervised random walks: Predicting and recommending links in social networks. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*. ACM, 635–644.

Henry H. Bi, Jianrui Wang, and Dennis K. J. Lin. 2011. Comprehensive citation index for research networks. *IEEE Transactions on Knowledge and Data Engineering* 23, 8 (2011), 1274–1278.

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

Wei Chen, Laks V. S. Lakshmanan, and Carlos Castillo. 2013. Information and influence propagation in social networks. *Synthesis Lectures on Data Management* 5, 4 (2013), 1–177.

W. Chen, C. Wang, and Y. Wang. 2010. 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*. ACM, 1029–1038.

W. Chen, Y. Wang, and S. Yang. 2009. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 199–208.

Suqi Cheng, Huawei Shen, Junming Huang, Wei Chen, and Xueqi Cheng. 2014. Imrank: Influence maximization via finding self-consistent ranking. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 475–484.

Feng Ding, Peter X. Liu, and Jie Ding. 2008. Iterative solutions of the generalized Sylvester matrix equations by using the hierarchical identification principle. *Applied Mathematics & Computation* 197, 1 (2008), 41–50.

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

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

Nan Du, Le Song, Manuel Gomez-Rodriguez, and Hongyuan Zha. 2013. Scalable influence estimation in continuous-time diffusion networks. In *Proceedings of Advances in Neural Information Processing Systems Conference*. 3147–3155.

A. Farahat, T. LoFaro, J. C. Miller, G. Rae, and L. A. Ward. 2006. Authority rankings from HITS, Pagerank, and SALSA: Existence, uniqueness, and effect of initialization. *SIAM Journal on Scientific Computing* 27, 4 (2006), 1181–1201.

Santo Fortunato. 2010. Community detection in graphs. *Physics Reports* 486, 3 (2010), 75–174.

Bin Gao, Tie-Yan Liu, Wei Wei, Taifeng Wang, and Hang Li. 2011. Semi-supervised ranking on very large graphs with rich metadata. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 96–104.

Priyanka Garg, Irwin King, and Michael R Lyu. 2012. Information propagation in social rating networks. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. ACM, 2279–2282.

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

Gene H. Golub and Charles F. Van Loan. 1996. *Matrix Computations*. Johns Hopkins University Press, 392–396.

Manuel Gomez-rodriguez, Jure Leskovec, and others. 2013. Modeling information propagation with survival theory. In *Proceedings of the 30th International Conference on Machine Learning (ICML'13)*. 666–674.

A. Goyal, F. Bonchi, and L. V. S. Lakshmanan. 2010. Learning influence probabilities in social networks. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*. ACM, 241–250.

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

M. Granovetter. 1978. Threshold models of collective behavior. *American Journal of Sociology* (1978), 1420–1443.

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

Mohsen Jamali and Martin Ester. 2010. A matrix factorization technique with trust propagation for recommendation in social networks. In *Proceedings of the 4th ACM Conference on Recommender Systems*. ACM, 135–142.

Glen Jeh and Jennifer Widom. 2003. Scaling personalized web search. In *Proceedings of the 12th International Conference on World Wide Web*. ACM, 271–279.

Ruoming Jin, Victor E. Lee, and Longjie Li. 2014. Scalable and axiomatic ranking of network role similarity. *ACM Transactions on Knowledge Discovery from Data* 8, 1 (2014), 3.

Kyomin Jung, Wooram Heo, and Wei Chen. 2012. IRIE: Scalable and robust influence maximization in social networks. In *Proceedings of the 12th IEEE International Conference on Data Mining (ICDM'12)*. IEEE, 918–923.

D. Kempe, J. Kleinberg, and É. Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 137–146.

M. Kimura and K. Saito. 2006. Tractable models for information diffusion in social networks. *Knowledge Discovery in Databases 2006* (2006), 259–271.

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

Onur Küçüktunç, Erik Saule, Kamer Kaya, and Ümit V. Çatalyürek. 2013. Diversified recommendation on graphs: Pitfalls, measures, and algorithms. In *Proceedings of the 22nd International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 715–726.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is twitter, a social network or a news media?. In *Proceedings of the 19th International Conference on World Wide Web*. ACM, 591–600.

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

J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. 2007. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 420–429.

Pei Li, Jeffrey Xu Yu, Hongyan Liu, Jun He, and Xiaoyong Du. 2011. Ranking individuals and groups by influence propagation. In *Advances in Knowledge Discovery and Data Mining*. Springer, 407–419.

Rong-Hua Li and Jeffrey Xu Yu. 2011. Scalable diversified ranking on large graphs. In *Proceedings of the 11th IEEE International Conference on Data Mining (ICDM'11)*. IEEE, 1152–1157.

D. Liben-Nowell and J. Kleinberg. 2007. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology* 58, 7 (2007), 1019–1031.

Q. Liu, B. Xiang, E. Chen, Y. Ge, H. Xiong, T. Bao, and Y. Zheng. 2012. Influential seed items recommendation. In *Proceedings of the 6th ACM Conference on Recommender Systems*. ACM, 245–248.

Qi Liu, Biao Xiang, Enhong Chen, Hui Xiong, Fangshuang Tang, and Jeffrey Xu Yu. 2014. Influence maximization over large-scale social networks: A bounded linear approach. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 171–180.

Qi Liu, Biao Xiang, Lei Zhang, Enhong Chen, Chang Tan, and Ji Chen. 2013. Linear computation for independent social influence. In *Proceedings of the 13th IEEE International Conference on Data Mining (ICDM'13)*. IEEE, 468–477.

Brendan Lucier, Joel Oren, and Yaron Singer. 2015. Influence at scale: Distributed computation of complex contagion in networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 735–744.

Paolo Massa and Paolo Avesani. 2006. Trust-aware bootstrapping of recommender systems. In *Proceedings of ECAI Workshop on Recommender Systems*.29–33.

L. Page, S. Brin, R. Motwani, and T. Winograd. 1999. The pagerank citation ranking: Bringing order to the web. (1999).

B. Aditya Prakash and Christos Faloutsos. 2012. Understanding and managing cascades on large graphs. *Proceedings of the VLDB Endowment* 5, 12 (2012), 2024–2025.

Diego Saez-Trumper, Giovanni Comarela, Virgílio Almeida, Ricardo Baeza-Yates, and Fabrício Benevenuto. 2012. Finding trendsetters in information networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1014–1022.

Chuan Shi, Xiangnan Kong, Yue Huang, Philip S. Yu, and Bin Wu. 2014. Hetesim: A general framework for relevance measure in heterogeneous networks. *IEEE Transactions on Knowledge and Data Engineering* 26, 10 (2014), 2479–2492.

Karthik Subbian, Chidananda Sridhar, Charu C. Aggarwal, and Jaideep Srivastava. 2014. Scalable information flow mining in networks. In *Machine Learning and Knowledge Discovery in Databases*. Springer, 130–146.

J. Tang, J. Sun, C. Wang, and Z. Yang. 2009. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 807–816.

Lei Tang, Huan Liu, and Jianping Zhang. 2012. Identifying evolving groups in dynamic multimode networks. *IEEE Transactions on Knowledge and Data Engineering* 24, 1 (2012), 72–85.

Hanghang Tong, Jingrui He, Zhen Wen, Ravi Konuru, and Ching-Yung Lin. 2011. Diversified ranking on large graphs: An optimization viewpoint. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'11)*, Vol. 11. 1028–1036.

Guan Wang, Yuchen Zhao, Xiaoxiao Shi, and Philip S. Yu. 2012. Magnet community identification on social networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 588–596.

Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*. ACM, 261–270.

B. Xiang, Q. Liu, E. Chen, H. Xiong, Y. Zheng, and Y. Yang. 2013. Pagerank with priors: An influence propagation perspective. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI'13)*. 2740–2746.

Tong Xu, Hengshu Zhu, Xiangyu Zhao, Qi Liu, Hao Zhong, Enhong Chen, and Hui Xiong. 2016. Taxi driving behavior analysis in latent vehicle-to-vehicle networks: A social influence perspective. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1285–1294.

Jaewon Yang, Bee-Chung Chen, and Deepak Agarwal. 2013. Estimating sharer reputation via social data calibration. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 59–67.

Y. Yang, E. Chen, Q. Liu, B. Xiang, T. Xu, and S. Shad. 2012. On approximation of real-world influence spread. In *proceedings of European Conference on Machine Learning and Knowledge Discovery in Databases*. 548–564.

Zhiwen Yu, Zhu Wang, Huilei He, Jilei Tian, Xinjiang Lu, and Bin Guo. 2015. Discovering information propagation patterns in microblogging services. *ACM Transactions on Knowledge Discovery from Data* 10, 1 (2015), 7.

Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. 2014. *Social Media Mining: An Introduction*. Cambridge University Press.

Jing Zhang, Jie Tang, Juanzi Li, Yang Liu, and Chunxiao Xing. 2014. Who influenced you? Predicting retweet via social influence locality. *ACM Transactions on Knowledge Discovery from Data* 9, 3 (2014), 25.

Jiawei Zhang and Philip S. Yu. 2015. Community detection for emerging networks. In *Proceedings of SIAM Conference on Data Mining (SDM'15)*.

Kai Zheng, Han Su, Bolong Zheng, Shuo Shang, Jiajie Xu, Jiajun Liu, and Xiaofang Zhou. 2015. Interactive top-k spatial keyword queries. In *Proceedings of the 2015 IEEE 31st International Conference on Data Engineering*. IEEE, 423–434.

Yang Zhou and Ling Liu. 2015. Social influence based clustering and optimization over heterogeneous information networks. *ACM Transactions on Knowledge Discovery from Data)* 10, 1 (2015), 2.

H. Zhu, H. Cao, H. Xiong, E. Chen, and J. Tian. 2011. Towards expert finding by leveraging relevant categories in authority ranking. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. ACM, 2221–2224.

Received November 2015; revised July 2016; accepted January 2017