Bimodal Invitation-Navigation Fair Bets Model for Authority Identification in a Social Network

Suratna Budalakoti
The University of Texas at Austin
suratna@utexas.edu

Ron Bekkerman
LinkedIn Corporation, USA
rbekkerman@linkedin.com

ABSTRACT

We consider the problem of identifying the most respected, authoritative members of a large-scale online social network (OSN) by constructing a global ranked list of its members. The problem is distinct from the problem of identifying influencers: we are interested in identifying members who are influential in the real world, even when not necessarily so on the OSN. We focus on two sources for information about user authority: (a) invitations to connect, which are usually sent to people whom the inviter respects, and (b) members' browsing behavior, as profiles of more important people are viewed more often than others'. We construct two directed graphs over the same set of nodes (representing member profiles): the invitation graph and the navigation graph respectively. We show that the standard PageRank algorithm, a baseline in web page ranking, is not effective in people ranking, and develop a social capital based model, called the fair bets model, as a viable solution. We then propose a novel approach, called bimodal fair bets, for combining information from two (or more) endorsement graphs drawn from the same OSN, by simultaneously using the authority scores of nodes in one graph to inform the other, and vice versa, in a mutually reinforcing fashion. We evaluate the ranking results on the LinkedIn social network using this model, where members who have Wikipedia profiles are assumed to be authoritative. Experimental results show that our approach outperforms the baseline approach by a large margin.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining

General Terms

Algorithms, Experimentation

Keywords

Social Networks,Authority,PageRank,Reputation,Influence

1. INTRODUCTION

An online social network (OSN) is an imperfect representation of real-world social interactions, as only a fraction of people's real-world activities are reflected online. It can be compared to a still camera capturing snapshots of modern society: it cannot see reality from all possible angles, however it often manages to capture the essence. This raises a fascinating question: can we use the structure of an online social network (OSN) to infer real-world social status hierarchy, in general; and – in particular – can we identify who the most well respected, prominent members of the society are? In this paper, we attempt to construct an inferred global rank of members of an OSN, according to the level of recognition they have achieved in the real world. Our model is applied to the LinkedIn professional network, which is large enough (over 100M professionals) to yield statistically meaningful results.

We distinguish between the tasks of identifying *authorities* and identifying *influencers* (as, for example, in [9]). Usually, influence is measured in terms of the degree to which a user's behavior affects her peers, as observed on the OSN. In order to influence many people, a member needs to be more than just active on an OSN. Maintaining an influential online presence can be extremely competitive, and many respected people may not be willing to invest such a high degree of time and effort. Another potential reason could be that they represent a demographic group not yet engaged by the OSN, or the online world in general. Moreover, social media influencers might not be authoritative in the real world: Khrabov *et. al.* [13] observed that the most influential Twitter users are relatively, if not completely, unknown outside their online circles. In contrast, we are interested in authority identification: inferring influence a user has in the real world, even if this influence is not highly notable online.

Since the *global* authority information is not directly obtainable, we build our model over two basic actions most OSN members perform, both of which can be viewed as a *local* endorsement from a user A to a user B: (a) user A invites user B to connect, and (b) user A views user B's profile. Our assumption is that users on a social network are more likely to send invitations to users that they respect, or at least do not disrespect. Similarly, browsing through another user's profile can be seen as a sign of interest. Obviously, the local signal is weak and noisy: there are many reasons why a user could send an invitation to another user, or view their profile. However, the signal distills with the volume: if thousands of users wish to connect with a particular person, or her profile is viewed by millions, it is a strong indication that the person is an authority outside the OSN. The signal becomes even stronger when aggregated over the entire OSN data. If user A wishes to connect with user B, and user B wishes to connect with user C, there is a certain amount

WWW 2012, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1229-5/12/04.

of implicit endorsement that goes from user A to user C. Aggregating billions of invites and profile views, the OSN produces a stream of endorsement that is directed towards a few users who may rarely send invites or view profiles, but are getting constantly approached by others who are constantly approached by others etc. The real-world influence of people on the very top of this pyramid is indubitable.

We notice that the text on a user profile is an insufficient and sometimes misleading proxy to the the real-world status of a particular OSN member. Some member profiles are too short and do not contain enough information to perform an accurate inference. Some are too verbose and may be exaggerated for marketing or search engine optimization (SEO) purposes. We do not use profile information to fit our model, however, we rely on user profiles in our model's evaluation.

The goal of this work is to answer the question "Who are the most respected $\langle \ldots \rangle$ on LinkedIn?", rather than "Is user A more respected than user B?". Despite that we build a global ranked list of LinkedIn users, it would be irresponsible for us to infer that the user ranked 30M on that list is more respected than the user ranked 31M, because the signal is too noisy. However, we'd like to be confident that the top 1% of users in the list are highly respected by many people. For example, we can infer that T. Boone Pickens[1] is among the most respected financiers on LinkedIn – he ended up being the first financier in the constructed ranked list. By creating one *global* ranked list that consists of many millions of users, we guarantee that the resulting pool of authoritative users (say, the top 1% of users in the ranked list) is large and representative enough such that it can be further refined by a specific request ("Who are the most respected *financiers* on LinkedIn?" or "Who are the most respected people from *Japan* on LinkedIn?"

Apart from the novelty of the problem being addressed, the paper makes the following technical contributions:

1. We propose a tournament-based model of user interactions, where users can be visualized as expending and accumulating social capital while respectively initiating and accepting links. In other words, user A can be modeled as making a payment in social capital when she sends user B an invitation, or views B's profile. A's authority score, in this framework, is the amount she can afford to pay per interaction initiated by her, based on the capital she has accumulated from others. Mathematically, this is represented using a variant of the *fair bets* model [5, 20].

2. We present an approach to combining authority-related information from multiple graphs, where each graph is constructed over the same set of users, but represents different aspects/modes of their behavior. This model is equivalent to simultaneously using the fair bets score vector of one graph for random restarts of walks on the other graph, and vice versa.

The algorithms described here, while evaluated only on the LinkedIn social graph, are easily applicable to other social networks. Having to combine data from multiple graphs is a common problem in social network analysis. Some examples of such datasets are: user 'follow' networks and 'retweeting' behavior on microblogging sites such as Twitter[2], and user social networks and voting/response behavior on Q&A or content sharing platforms.

2. BACKGROUND

User interactions on social networks lend themselves naturally to a graph-based representation. Treating each user as a vertex, we represent the invitation data as a directed graph, with an edge directed from the inviting user to the invitee. This graph is referred to as the *invitation graph*. Similarly, a *navigation graph* is constructed, with edges directed from the person who views a profile, to the person whose profile was viewed. So, in both cases, an endorsement is modeled as 'flowing' in the direction of the edge.

Hyperlinked datasets such as the WWW are often represented as directed graphs [3, 14] for the task of identifying high-quality pages, with hyperlinks interpreted as endorsement. Similar representations (with invitations, 'follows', or 'retweets' as directed edges) have been used for identifying influencers in social networks [4, 7, 22]. A variety of link analysis algorithms [3, 14] exist for identifying important nodes in such graphs, the most popular one being PageRank [2, 3].

The PageRank algorithm employs a recursive definition of authority: the authority of a vertex is a weighted sum of the authority of the vertices that point to it (ignoring, for now, the random restart aspect). This can distort user authority estimates in a number of ways:

1. An authoritative user is more likely to accept connection invitations than to send them out. This is because many non-authoritative users find a lot of value in connecting with authorities, while the opposite is not always true. More generally, link formation in OSNs is found to be consistent with a status-based model [18], where low status nodes link to those of high status. This observation does not play a part in the PageRank model.

2. While most information on the Web is publicly accessible, social networks such as LinkedIn have a variety of privacy settings, and sometimes do not allow users to access profiles more than a few degrees from their own. As a result, a user's network size and openness play a major role in the number of invitations / profile views they receive.

3. Motivated users can take advantage of behavioral norms. For example, the norm of reciprocity, i.e., users feeling obligated to return links with courtesy links, is used by unscrupulous users to increase their link count, on both Flickr[3] and Twitter [16, 22].

4. Older users can become entrenched over time, and have an indegree disproportionate with their authority level. This can discourage younger users from participating. It may not be an issue on professional networks such as LinkedIn, but is a serious factor [17] in information sharing networks such as Twitter and Digg[4].

In other words, members' PageRank scores on an OSN graph depends on two factors: a) their authority, which determines the desirability of a connection with them, and

[1] www.boonepickens.com
[2] www.twitter.com
[3] www.flickr.com
[4] www.digg.com

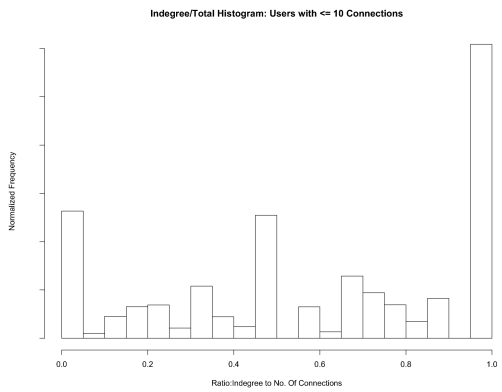

Figure 1: Indegree-Total Connections Ratio Histogram: Users with $\leq$ 10 Connections

b) their *visibility* on the graph, that is, the likelihood that they will be noticed by other users. Non-authoritative members can improve their PageRank scores by increasing their visibility, usually through increased activity (engaging other members via page views), or increased connectivity (by sending more invitations). We use the fair bets model, discussed next, to take into account this observation.

3. THE FAIR BETS MODEL

The *fair bets* model was developed [5, 20] to rank players in round-robin tournaments. The model is based on the idea that, a player is allowed to bet an amount of money per game. She forfeits this amount to her opponent if she loses the game, and if she wins, she is awarded the amount bet by her opponent. The score assigned to a player is then the amount she can afford to bet, assuming she has to bet the same amount against all players. More recently, this model has been studied by Slutzki *et al.* [21].

Mathematically, this model can be represented as follows: we can construct a graph G, with each player i as a vertex v_i, and assuming each pair of players played a maximum of one game against another player (this can be generalized), an edge directed from the loser of the game to the winner. This graph can then be represented as a matrix V, where $v_{ij} = 1$ if there is an edge directed from player i to j in the graph. Then the fair bets score a_j of player j satisfies the following property:

$$\sum_{i=1}^{N} v_{ij} a_i = \sum_{i=1}^{N} v_{ji} a_j$$

That is, the amount of money any player j pays out per game (a_j) is the amount she makes in total, divided by the number of games lost. This can can be written in matrix form as:

$$V^\top \vec{a} = C\vec{a} \tag{1}$$

where C is a diagonal matrix such that C_{ii} is equal to the sum of the ith row of V, that is, $C_{ii} = \sum_{k=1}^{N} v_{ik}$. A straightforward relationship can be established between PageRank and fair bets scores [5]. For a stochastic matrix P, the PageRank vector r corresponding to the matrix is given by the equation:

$$P^\top \vec{r} = \vec{r} \tag{2}$$

Let $P = C^{-1}V$. That is, P is the stochastic matrix constructed by normalizing all the rows of some tournament matrix V to 1.

Then, equation (2) can be written as:

$$V^\top C^{-1}\vec{r} = \vec{r} \Rightarrow C^{-1}V^\top \left(C^{-1}\vec{r}\right) = C^{-1}\vec{r}$$

Setting $\vec{a} = C^{-1}\vec{r}$ gives:

$$C^{-1}V^\top \vec{a} = \vec{a} \Rightarrow V^\top \vec{a} = C\vec{a}$$

which is the same as equation (1). That is, if the PageRank vector for a stochastic matrix P is given by $\vec{r}$, then the fair bets vector of the original graph matrix $V = CP$ is given by $\vec{a} = C^{-1}\vec{r}$. Thus, mathematically, the fair bets score of a vertex in a graph is equal to its PageRank score, divided by its outdegree.

3.1 Fair Bets as Social Capital

In the context of online social networks, fair bets can be viewed as a model of social capital accumulation and expenditure. Users can grow their connection graph in two ways: either by sending invitations or accepting them. As sending an invitation requires time and effort on a user's behalf, and a willingness to make the gesture, users are more likely to make this investment if they believe the new connection can help them in achieving social/professional growth. This growth can take place online: more connections increase the likelihood that someone will stumble on the person's profile, thus increasing the likelihood of invitations. Or both the original invitation, and subsequent new connections, could be side-effects of real world activity.

Thus, over time, the initial time and social capital spent in inviting connections pays off, as the user accumulates invitations in return. In this setup, highly respected users receive multiple invitations without making a significant effort, while the payoff for less authoritative users is lower. The standard fair bets model can then be visualized as follows: assuming users were paying each other to accept invitations on an OSN, then the fair bets score of a user is the amount she can afford to pay on average.

4. USER AUTHORITY EVOLUTION

For an OSN graph, the standard fair bets model discussed above assumes a linear relationship between a vertex's authority score and its outdegree. The fair bets score a_i of a vertex v_i can be written as:

$$a_i = \frac{\text{indegree}(v_i)}{\text{outdegree}(v_i)} \cdot \mu_i$$

That is, the fair bets authority score of a vertex directly proportional to a) the mean authority accumulated per incident vertex, μ_i, and, b) the indegree to outdegree ratio (*i-o ratio*). Both factors depend on the stage of evolution of the vertex. The evolution of user vertices on the invitation graph can be divided into three stages. The first stage is that of users with less than 10 connections. A normalized histogram of the indegree to number of connections (*i-t ratio*) for this group of users is shown in Figure 1. As can be seen, a majority of these users have a ratio close to 1.This is because new users are unlikely to send invitations, due to

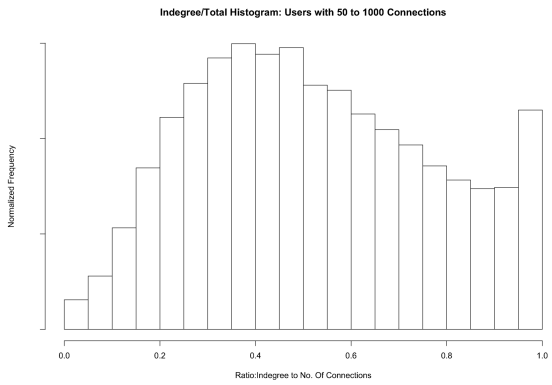

Figure 2: Indegree-Total Connections Ratio Histogram: Users with 50 to 1000 Connections

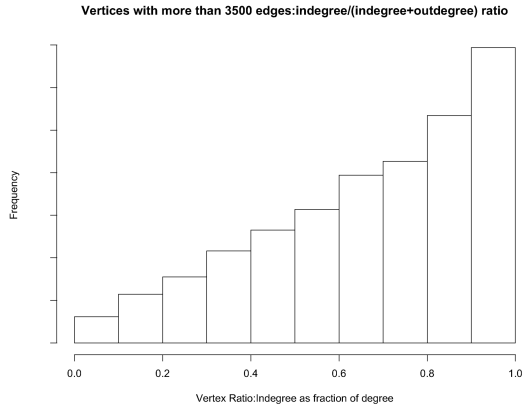

Figure 3: Indegree-Total Connections Ratio Histogram: Users with More Than 3500 Connections

being isolated by the small size of their connection graph. This can give them an artificially high i-o score. To address the skewness of the i-o ratio of poorly connected users, we use a Laplace smoothing of the outdegree value in the fair bets formula, by adding a small constant (equation 3).

The i-t ratio for users with $50-1000$ connections is shown in Figure 2. While there's still a fair number of users with an i-t ratio of more than 0.9, the ratio is relatively normally distributed, with an overwhelming majority in the 0.2-0.6 range.

On the other extreme, for users with more than 3500 connections, the graph is biased once again towards much higher ratio values, as shown in Figure 3. This is a very small subset of users, consisting largely of extremely active[5] and influential users. PageRank would rank these users near the top of the ranked list. Interestingly, a fair bets-based ranking places these users near the bottom of the list (with rare exceptions), despite their high indegree-outdegree ratio. This is because, for users with an extremely large number of incoming edges, a majority of these incoming edges have low values of authority, due to the way authority scores are usually distributed across the graph (power law). This results in a lower mean value.

4.1 Log Fair Bets

As a basic validation, we evaluated the relationship between the fair bets based rank assigned to a user, and his/her professional seniority level. The seniority level data is proprietary standardized data derived from LinkedIn profiles, that maps millions of job titles in the LinkedIn dataset to one of ten levels: from intern (0), to founder (9). A ranking by authority is more likely to be reliable if users at higher ranks, on average, hold titles of higher seniority, compared to lower ranked users. Figure 4 shows the evolution of seniority with fair bets ranks. The ranks towards the right are the highest ranks.

Interestingly, there is a dramatic jump in the seniority of people at the very top of the ranked list. However, after a certain point, users' ranks seem to bear little relationship to seniority levels. The reason is the over-steep normalization: a user with 100 connections will need to have twice the PageRank score as a user with 50 connections (assuming the same i-o ratio), to have the same fair bets score. Intuitively, this seems unlikely. PageRank scores are likely to follow a power law distribution, so that a few users would contribute most of a user's score. Assuming more active users have higher scores, users are more likely to receive their more valuable edges sooner rather than later. Also, a user's connection network grows much faster in the initial stages, as each connection makes them visible to many new users. At some point, the law of diminishing returns would set in, as most connections of a newly added connection are already part of the user's network, thus unlikely to lead to more incoming invitations. The same logic extends to page views.

Based on these observations, the normalization we use, which we refer to as *log fair bets (LFB)*, is as follows:

$$f_i = \frac{indegree(v_i)}{\log\left(10 + outdegree(v_i)\right)} \cdot \mu_i \qquad (3)$$

Log fair bets can be interpreted as assuming that the arrival patterns of incoming links follows a power law distribution with respect to time (measured by outdegree). That is, the expected authority value of links received once k invites have been sent is $\frac{1}{k}$. This expected value includes both the probability of receiving a link, and the authority of the link. In this interpretation, the $\log k$ can be seen as approximating the sum $\sum_{i=1}^{k} \frac{1}{i}$. The value of 10 is the Laplace smoothing parameter, fixed based on the analysis in the previous section. A validation similar to that for fair bets results was done for log fair bets by comparing ranking results against standardized seniority data. Figure 5 shows the resulting graph. As can be seen, the log fair bets graph is much smoother, and the seniority level tracks the ranking much more closely.

5. COMBINING INVITATION AND NAVIGATION GRAPHS

We construct two separate graphs, the invitation graph and the navigation graph, to represent invitation data and browsing patterns respectively. The assumption behind this decision is that the two graphs are complementary: there is

[5]This seems paradoxical, but a user with $10,000$ connections and an i-t ratio of 0.9 has sent out 1000 invitations, a higher level of activity than most users.

authority-related information in each graph that is missing in the other. This is in contrast to a commonly accepted assumption about user browsing patterns.

According to the the *random surfer model* interpretation of the PageRank algorithm for the Web, the PageRank vector corresponds to the fraction of time a user will spend on a web page, if she were to start at a random page, and at each timestep, randomly select an outgoing edge. This posits an extremely close relationship between link structure and browsing patterns. The assumption is reasonable for web pages, but not for social network graphs, for a number of reasons. For example, while both invitations and browsing behavior are asymmetric, that is, one user takes the initiative, which reveals a greater involvement on their part than the other user, the degree of asymmetry is much lesser for invitation than navigation behavior. For an inviter's connection request to be successful, the invitee has to accept the request. On the other hand, navigation requires no activity by the person whose profile is being viewed, and is wholly asymmetric. Thus, invitation requests are more likely to be directed to people in the inviter's professional peer group, while navigation data reflects information about who the user aspires to know. A user may be more likely to connect to her immediate supervisor, but may browse her company CEO's profile more often. Also, invitation requests are guided by a number of social norms. For example, a user may feel obligated to send requests to all the people she meets at her workplace. Such obligations do not exist for navigation behavior.

Given two separate graphs over which authority ranks can be calculated, a combined rank can be arrived at in two ways:

5.1 Rank Merging via Metasearch

Use a metasearch-based approach to merge the two rankings. Borda voting [1], for example, is a simple but usually effective approach to merging two ranked lists: the rank of a user is essentially the mean of their rank in the two lists.

5.2 Bimodal Authority Models

In the random surfer interpretation of the PageRank algorithm, at each timestep, with a certain probability d (usually set to 0.85), the surfer randomly selects an outgoing link from the current page. With the remaining probability $1-d$, the surfer gets bored and jumps to a completely new page. The probability $1-d$ is referred to as the *teleportation probability*, and the vector the new page is chosen from is called the *teleportation vector*. The vector can be uniform, or biased to reflect some priorly known information. For example, the teleportation vector could be personalized [6] given sufficient information about the surfer, or be biased towards trusted vertices. Its effect is to bias the overall scores towards the preferences of the vertices with higher values in the teleportation vector.

A natural way, then, to inform one graph (say, invitation) with information from the other (say, navigation) would be to use the authority vector of one as the teleportation vector for the other. Following this, the improved results in the navigation graph can be reused to improve the results in the invitation graph, and so on till convergence. We refer to this approach as the *bimodal authority* approach. The idea behind this approach is mutual positive reinforcement: useful information in one graph can be used to improve the authority estimates of the other graph, and vice versa. The teleportation vector could be based on PageRank (*bimodal PR*), or log fair bets scores (*bimodal LFB*).

However, as the next section shows, successive alternate runs of the two algorithms are not necessary. Instead, a composite graph can be created, by merging the invitation and navigation graphs in a certain way. The invitation and navigation PageRank vectors that would result from the bimodal approach, can be obtained from the the PageRank vector of the composite graph.

5.2.1 Proof Of Equivalence: Bimodal and Composite Graph Models

We are given two graphs, $G_A = (V_A, E_A)$ and $G_N = (V_N, E_N)$, representing different aspects of user behavior. Both graphs have the same number of vertices, say, k. For each vertex $v \in V_A$, there is a corresponding twin vertex $v' \in V_N$. In our example, the vertex v for a user represents her invitation behavior, while v' represents her navigation behavior. We would like to use the PageRank vector of one graph as the teleportation vector of the other. That is, the teleportation probability for $v \in V_A$ should be equal to the PageRank score of its twin vertex $v \in V_N$, and vice versa. To do this efficiently, we prove the following result:

Construct a new graph $G = (V_A \cup V_N$, $E = E_A \cup E_N \cup E_{AN})$, where E_{AN} is a new set of directed edges, between all pair of twin vertices, and weighted d. That is, a vertex v in the invitation graph is connected edge to its twin vertex v' in the navigation graph via a directed edge of weight d. A similar directed edge of weight d connects v' to v. Then calculating the PageRank vector for graph G is equivalent to solving the problem described above.

Proof: Let the transition matrix of V_A be written as P_A and its (unknown) PageRank vector be r_A . Similarly, let the transition matrix and PageRank vector of V_N be P_N and r_N respectively. Let e be a vector such that $e_i = 1$ for all i. Then, by our recursive definition, PageRank vectors of G_A and G_N satisfy the following equations:

$$\left((1-d)\,P_A + d\vec{e}\vec{r}_N^{\top}\right)^{\top}\vec{r}_A = \vec{r}_A \tag{4}$$

$$\left((1-d)\,P_N + d\vec{e}\vec{r}_A^{\top}\right)^{\top}\vec{r}_N = \vec{r}_N \tag{5}$$

Expanding (4), we get:

$$(1-d)\,{P_A}^{\top}\vec{r}_A + d\vec{r}_N e^{\top}\vec{r}_A = \vec{r}_A \Rightarrow (1-d)\,{P_A}^{\top}\vec{r}_A + d\vec{r}_N = \vec{r}_A \tag{6}$$

since r_A sums to 1.

Similarly, for (5), we get:

$$(1-d)\,{P_N}^{\top}\vec{r}_N + d\vec{r}_A = \vec{r}_N \tag{7}$$

Let I_k be an identity matrix of size k. Then equations (6) and (7) can be written in matrix form as follows:

$$\begin{bmatrix}(1-d)\,{P_A}^{\top} & dI_k \\ dI_k & (1-d)\,{P_N}^{\top}\end{bmatrix}\begin{bmatrix}\vec{r}_A \\ \vec{r}_N\end{bmatrix} = \begin{bmatrix}\vec{r}_A \\ \vec{r}_N\end{bmatrix} \tag{8}$$

Let P be a matrix, such that:

$$P = \begin{bmatrix}(1-d)\,P_A & dI_k \\ dI_k & (1-d)\,P_N\end{bmatrix} \tag{9}$$

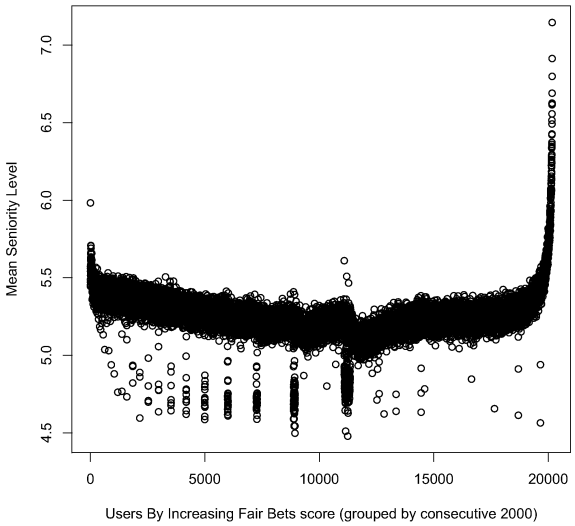

Figure 4: User Fair Bets Rank vs Mean Seniority Level (over consecutive groups of 2000 people)

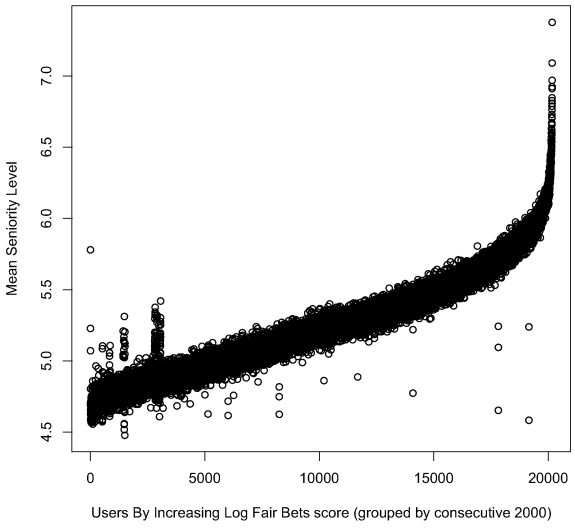

Figure 5: User Log Fair Bets Rank vs Mean Seniority Level (over consecutive groups of 2000 people)

	Invitation Graph		Navigation Graph		Hybrid			
Metric(in %)	PR	LFB	PR	LFB	Borda PR	Bimodal PR	Borda LFB	Bimodal LFB
MAP@1000	3.26	5.52(69.3%)	9.22(182.8%)	12.84(293.9%)	7.55(131.6%)	12.16(273.0%)	13.03(299.7%)	13.60(317.2%)
MAP@100K	2.45	2.53(3.2%)	1.84(−24.8%)	3.37(37.5%)	2.46(0.4%)	2.30(−6.1%)	3.76(53.4%)	3.93(61.2%)
MAP@1mil	1.23	1.36(10.5%)	0.87(−29.2%)	1.44(17.1%)	1.27(3.2%)	1.08(-12.1%)	1.84(49.6%)	1.88(52.8%)
NDCG@1000	1.48	3.08(108.1%)	3.99(169.6%)	6.80(359.5%)	2.92(97.3%)	4.59(210.1%)	4.78(222.9%)	6.35(329.7%)
NDCG@100K	3.84	4.48(16.7%)	3.64(−5.2%)	5.91(53.9%)	4.36(13.5%)	4.28(11.4%)	6.10(58.8%)	6.61(72.1%)
NDCG@1mil	8.30	9.13(10.0%)	7.65(−7.8%)	10.27(23.7%)	8.79(5.9%)	8.73(5.18%)	11.13(34.0%)	11.75(41.6%)

Table 1: MAP and NDCG Results For Invitation Graph, Navigation Graph, and Hybrid PageRank(PR) and Log Fair Bets (LFB) approaches. The values in parentheses give the percentage improvement over Invitation Graph PageRank, treated as a baseline approach.

and let $\vec{r} = \begin{bmatrix} \vec{r_A} \\ \vec{r_N} \end{bmatrix}$. Then equation (8) can be written as $P^\top \vec{r} = \vec{r}$. Then, by the definition of the Pagerank vector [3], $\vec{r}$ is the PageRank vector for P. Hence proved.

5.2.2 Bimodal Log Fair Bets

The matrix P in equation 9 can be modified to use other authority models instead of PageRank, as the teleportation vectors. For example, let the identity matrix in the first row of P be replaced by diagonal matrix R_A, whose i-th diagonal value is $\frac{1}{\log_2(o_i+10)}$, where o_i is the outdegree of vertex v_i in V_A. Similarly, replace the identity matrix in the second row of P with diagonal matrix R_N, with $R_N(j,j) = \frac{1}{\log_2(o_j+10)}$, where o_j is the outdegree of vertex v'_j in V_N. After normalizing R_A and R_N each to add to d, this results in a bimodal model, where the log fair bets vector of each graph serves as the teleportation vector of the other (the final results still need to be normalized to get log fair bets scores). This model, which we call the *bimodal log fair bets model*, outperforms other models for authority identification by a significant margin.

6. DATA DESCRIPTION

In May 2011, we were given a subset of about 50M LinkedIn members, chosen from the entire LinkedIn member base (of about 100M members) using some product business logic. We obtained all connection invitations that were sent and accepted between the members in our subset, resulting in an *invitation graph* with billions of directed edges, going from inviters to invitees. We then constructed the *navigation graph* over the same set of vertices as in the invitation graph: we draw an edge from user A to user B if user A viewed user B's profile at least twice within a certain period of time (one year). Our assumption here is that a single view of a user's profile is too weak to count as an endorsement, so two views is set as a lower bound. Unlike the invitation graph, where all edges are weighted equally, the navigation graph edges are weighed by the number of times the profile was viewed. The outgoing edge weights are normalized for both invitation and navigation graphs, so that they sum to one for each vertex.

6.1 Evaluation Dataset Construction

As the ground truth of authoritative people, we decided to use LinkedIn users who have Wikipedia[6] profiles. Wikipedia is known to be selective about allowing to create people profiles, so that only significant people tend to have Wikipedia profiles. Obviously, as any manual process, the choice of significant people is somewhat subjective. However, most well known people are likely to have Wikipedia profiles – which is a reasonable starting point for our model's evaluation. The evaluation goal is to test whether most LinkedIn users with Wikipedia profiles appear on the top of the constructed ranked list of authorities.

We built a text mining system that maps LinkedIn users to Wikipedia profiles based on matching the textual data between LinkedIn and Wikipedia profiles. Our goal was to optimize for the mapping precision trading off the recall, therefore we made quite a few assumptions that kept the resulting precision at a high level. Given a LinkedIn member li and a person wi who has a dedicated Wikipedia profile, we assume that $P(li = wi|\text{Name}_{li} \neq \text{Name}_{wi}) = 0$, that is, the probability of li and wi to be the same person is zero if li and wi do not have the same name.

We started with a list of candidate LinkedIn members whose profiles are dense enough (they contain a profile headline, at least one current position, and a reasonable number of connections). For each name of a candidate LinkedIn member, we checked if there exists a Wikipedia page with that name as a title. We extracted the first paragraph[7] of each such page, and aggregated all of them into a candidate Wikipedia profile list. From the resulting list, we filtered out disambiguation pages as well as pages that are dedicated to deceased people and to fictional characters.

We represented each LinkedIn member li from the candidate list as the Bag-of-Words BOW_{li} of his/her headline and current position information. We represented each Wikipedia personality wi from the candidate list as the Bag-of-Words BOW_{wi} of the first paragraph of his/her Wikipedia profile. We estimate the probability of li and wi to be the same person as follows:

$$P(li = wi) \propto P(li = wi|\text{Name}_{li} = \text{Name}_{wi}) \times P(li = wi|\text{Profile}_{li} \cap \text{Profile}_{wi}) \quad (10)$$

The probability of li and wi being the same person given that they share their name $P(li = wi|\text{Name}_{li} = \text{Name}_{wi})$ is inversely proportional to the commonness of the name. We estimate the name commonness over the list of all member names on LinkedIn. The probability of li and wi being the same person given the overlap in their profiles $P(li = wi|\text{Profile}_{li} \cap \text{Profile}_{wi})$ can be approximated by the cosine similarity between the two profiles, represented as TFIDF

[6] www.wikipedia.org

[7] The first paragraph of a Wikipedia page dedicated to a person usually contains the most essential biographical information about that person.

vectors of their Bags-of-Words. We estimate the IDF scores of words over the entire collection of LinkedIn member profiles.

For every person wi with a Wikipedia profile from the candidate list, and for every LinkedIn member li with the same name, we compute the right side of formula (10) and decide that $li = wi$ if the resulting value is above a preset threshold. After some hand-tuning, the final system yielded about 30K LinkedIn members who have Wikipedia profiles. We estimate the mapping's precision as very high – we spot checked a couple of hundred mappings and did not see a single instance of a wrong mapping. We cannot estimate the mapping's recall though. For our model's evaluation purposes however, the mapping's recall does not matter.

7. EXPERIMENTAL RESULTS

7.1 Evaluation Measures

We use two widely used measures, the mean average precision (MAP) score, and the normalized discounted cumulative gain (NDCG) score, to evaluate the quality of our ranked results.

Given a ranked list and a set of relevant documents (or in our case, users who have Wikipedia profiles), its Average Precision (AP) is defined as the mean of the precision scores, calculated at each rank where a relevant match was found on the list. MAP is the average of AP scores across multiple queries. Since, in our case, we are essentially evaluating a single query, the average precision score serves as the MAP. Since we are more interested in the quality of the higher ranks of our results, than the entire list, the MAP scores are given after cutting off the list at three thresholds: after 1000 ranks (MAP@100), after a hundred thousand ranks (MAP@100K), and after one million ranks (MAP@1mil).

The MAP measure treats all users on the Wikipedia list as equally relevant. The other measure we use, NDCG, enables us to differentiate between users in terms of degrees of relevance. Given a ranked list, the DCG score of the list upto n ranks is given by:

$$DCG = m_1 + \sum_{i=2}^{n} \frac{m_i}{\log_2 i} \tag{11}$$

where m_i is the estimated relevance of the i^{th} match. The NDCG score is given by normalizing this value by the *ideal* DCG (IDCG) value, that is, the maximum score that any ranking can achieve given the relevance scores.

For any user with a Wikipedia profile, we calculate her relevance score m_i, as the log of the mean number of page views per day received by her profile, based on two months of Wikipedia page view data[8] (May and June 2011). The relevance score for all users receiving less than three page views a day is set to 1. This gives us a relevance range of approximate 1-15, as highly trafficked profile pages on Wikipedia receive around $10,000$ page views a day.

Based on this, the idea DCG score (IDCG) can be calculated as follows: sort the Wikipedia users' list by descending order of page views, and calculate:

$$IDCG = \log_2 p_1 + \sum_{i=2}^{k} \frac{\log_2 p_i}{\log_2 i} \tag{12}$$

[8]The data was collected from the website `http://stats.grok.se`.

where p_i is the page views received by the i-th ranked user. The value of k is the cutoff limit. In our case, the maximum is approximately $30,000$, the number of Wikipedia profiles we have mapped to LinkedIn users. To ensure that ranks beyond the first few hundred impact NDCG results, we divide user ranks into buckets of 500. For the first 500 ranks, $i = 2$, $i = 3$ for the next 500, and so on, in equations (11) and (12). Thus, a user with a relevance score m_i, placed in the first 500, would add m_i to the DCG score, while the same user, placed in the 501-1000 range would add $\frac{m_i}{\log_2 3}$ to DCG.

Like MAP, we calculate NDCG after 1000 (NDCG@1000), 100,000 (NDCG@100K), and 1 million (NDCG@1mil). The IDCG score increases in value from NDCG@1000 to NDCG @100K, but then remains constant till NDCG@1million. For this reason, the NDCG score falls from the 1000 to 100,000 level, but then increases for the 1 million level.

7.2 Algorithm Comparison

All algorithms were implemented in a map-reduce framework, and run on a set of 100 Hadoop nodes. The open-source implementation of PageRank in the Pegasus software toolkit [12] was used as the original code base, and the code was modified to incorporate bimodal authority models. The results are shown in Table 1. The percentage improvements/deterioration, shown in brackets in each case, is based on treating the invitation graph based PageRank (Invitation Graph-PR) algorithm as the baseline for comparison. As can be seen from the table, the log fair bets (Log FB) model consistently performs better than the PageRank model for both the invitation and navigation graphs.

Interestingly, among the hybrid models that combine both invitation and navigation data, the best performing ones are the log fair bets models (Borda LFB and Bimodal LFB). The performance of the PageRank-based hybrid models is around the same as the single graph-based approaches. The reason for this is the large impact of user activity levels on the hybrid PageRank models. In the case of bimodal PageRank, the largest mutual reinforcement is for user who are most active, as they have higher PageRank scores on both graphs. A similar effect occurs in Borda voting based PageRank. Since Borda voting is based on mean scores, the highest ranked users on *both* graphs are people ranked highly on both graphs. These are usually highly active users. In contrast many authoritative users are not very highly ranked in one of the two graphs (for example, many people would view the profile of someone famous like Bill Gates, but very few would send an invite), and end up being ranked low on average. As a result, PageRank-based Borda voting is unable to take advantage of the best information in both graphs. In contrast, the bimodal log fair bets more (Bimodal LFB) is the only one actually able to achieve positive mutual reinforcement, and outperforms all other algorithms by a significant margin.

The only exception to this is the NDCG@1000 score, where the bimodal LFB comes in second to navigation graph LFB. The reason behind this is that there a small number of very high profile 'celebrity' users, who garner an extremely large number of page views both on Wikipedia and LinkedIn. Their high page views give them large values of m_i, which gives navigation LFB an edge at the 1000 level. This advantage, however, does not carry beyond the first 1000 or so members. Even up to the 1000 level, the actual number

Rank	Name	Affiliation
1	Barack Obama	President of the USA
2	Bill Gates	Founder of Microsoft
3	Jan Peter Balkenende	Former Prime Minister of the Netherlands
4	Sarah Palin	2008 US Vice President Nominee
5	T. Boone Pickens	Chairman of BP Capital Management
6	Hillary Clinton	US Secretary of State
7	Kevin Bacon	Actor, and 'zero-degree' of Kevin Bacon
8	Chris Brogan	Entrepreneur and Author of "Trust Agents"
9	Marc Benioff	Founder of Salesforce
10	John McCain	2008 US President Nominee
11	Michael Dell	Founder of Dell
12	Avinash Kaushik	Entrepreneur and Author of "Web Analytics"
13	Brian Solis	Entrepreneur and Author of "Engage"
14	Reid Hoffman	Founder of LinkedIn
15	Lakshmi Narayanan	Former CEO of Cognizant
16	Jeffrey Gitomer	Author of "The Little Red Book of Selling"

Table 2: Most authoritative LinkedIn users who have Wikipedia profiles.

of members matched with Wikipedia is lesser for navigation LFB than it is for bimodal LFB, as is suggested by the higher value of MAP@1000 of the latter, compared to the former.

Table 2 shows a list of the sixteen highest ranked users of LinkedIn, who have a Wikipedia profile, based on results from the bimodal log fair bets model.

8. RELATED WORK

Identifying influencers is a well-studied problem in the social network research community [4, 7, 11, 22]. Influencers are usually defined as users who can induce other members to take certain actions, such as, take interest in some information they share, etc. Influence is, thus, a measure of the user's importance within an OSN, but is not expected to contain information about their significance in the real world. This is reflected in the approach taken to empirical evaluation of the algorithms. For example, Weng *et al.* [22], for identifying influential users on Twitter, measure algorithm effectiveness using a measure based on the number of followers the identified users have. Ghosh and Lerman [7] measure user influence on Digg by the number of votes the stories posted by them get. In this paper, we propose an alternate, but related problem: given data from an OSN, can we identify people who are important in the real world, outside the network. The goal is to be able to identify such users even when they are not highly active on the OSN. To the best of our knowledge, this is the first paper to address this problem.

The most common approach [15, 22] to identifying influentials are variations of popular link analysis algorithms [3]. The links are constructed based on actions such as invitations, which usually take place once between a pair of users, or more dynamic data such as browsing patterns, 'retweets', 'likes', etc. However, to the best of our knowledge, this is the first paper to propose a principled approach towards combining these two disparate signals in the context of OSNs. A number of approaches, however, have been proposed towards combining invitation and browsing data for identifying quality web pages on the Internet. While the traditional approach [3, 14] has been based on hyperlink analysis, more recent approaches include Browserank [19] based on a Markov chain model based on browsing data alone, and recent work by Gleich *et al.* [8] for empirically learning teleportation parameters over the hyperlink graph from navigation data. However, as discussed in Section 5, these approaches rely on the much closer relationship between hyperlink and browsing data present on the Internet. These assumptions are less true for OSNs, where the differences between linking and browsing behavior is much greater. Also, many of these approaches don't make use of user-specific browsing data, something which is often available for OSNs, as opposed to the Web.

Another related problem domain to ours is that of citation analysis [10] for evaluating scientific publications. Zhou *et al.* [23] proposed a co-ranking based approach towards the problem of combining publication citation and author social networks, that has similarities to ours. However, there are many significant differences. Besides working with a much smaller dataset in a different problem domain, the paper does not explore the mathematical implications of its coupled random walk model. This limits the applicability of the work, particularly towards extensions such as log fair bets.

9. CONCLUSION

This paper investigates a fundamental problem of sociology – identifying the most important people of the society – given the partial observation of the real-world social interactions as represented on a large-scale online social network (OSN). We note that the most important people, while definitely being influential in the real world, are not necessarily active or influential in an OSN. Nevertheless, given implicit signals of the OSN member endorsement, and aggregating those signals over the whole network of many million people, we are able to come up with a large list of authorities.

We can cut off the list at any level, leaving it with 10 people or with 10 million people. However, it is probably disadvantageous to cut off too early or too late: a short list (up to, say, a thousand people) can potentially be composed manually as many of the authorities are well known

both inside and outside the OSN. A long list of 10 million people or so would be too noisy as its length is comparable with the size of the OSN. A reasonable cutoff might be at around a hundred thousand people, based on the guidance given in Figures 4 and 5: people's seniority level drops down dramatically after the first few hundreds of thousands.

Our model is straightforwardly generalizable to any network with multiple types of endorsements between nodes. For example, in the social search domain, nodes are Web pages and edges are hyperlinks between them, while another type of edges can be the social network of the pages' creators. Similarly, for many social networks, users' invitation and activity ('likes','retweets') behavior can be modeled as two different graphs.

We note that the implicit signals of endorsement, such as sending a connection invite, or viewing someone's profile, are more useful than explicit signals of endorsement, such as, for example, writing an online recommendation. The reason for this is two-fold. First, the explicit data is significantly sparser than the implicit data. Second, there are multiple incentives for people to endorse someone publicly: for example, this can bring visibility to the endorser. Careful analysis of online users' behavior, coupled with the large scale and richness of the raw data, is the key to answering such sociological questions, that our predecessors have been trying to answer for hundreds of years.

10. REFERENCES

[1] J. A. Aslam and M. Montague. Models for metasearch. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 276–284, New York, NY, USA, 2001. ACM.

[2] M. Bianchini, M. Gori, and F. Scarselli. Inside PageRank. *ACM Transactions on Internet*, 5(1):92–128, 2005.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, Apr. 1998.

[4] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 10–17, 2010.

[5] H. Daniels. Round-robin tournament scores. *Biometrika*, 56(2):295–299, 1969.

[6] D. Fogaras, B. Rácz, K. Csalogány, and T. Sarlós. Towards scaling fully personalized pagerank: Algorithms, lower bounds, and experiments. *Internet Mathematics*, 2(3), 2005.

[7] R. Ghosh and K. Lerman. Predicting influential users in online social networks. In *SNA-KDD: Proceedings of KDD Workshop on Social Network Analysis*, 2010.

[8] D. F. Gleich, P. G. Constantine, and A. D. Flaxman. Tracking the Random Surfer : Empirically Measured Teleportation Parameters in PageRank. *Human Factors*, 2010.

[9] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1019–1028, 2010.

[10] A. A. Goodrum, K. W. McCain, S. Lawrence, and C. L. Giles. Scholarly publishing in the internet age: a citation analysis of computer science literature. *INFORMATION PROCESSING and MANAGEMENT*, 37(5):661–675, 2001.

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

[12] U. Kang, C. E. Tsourakakis, and C. Faloutsos. Pegasus: mining peta-scale graphs. *Knowl. Inf. Syst.*, 27(2):303–325, 2011.

[13] A. Khrabrov and G. Cybenko. Discovering influence in communication networks using dynamic graph analysis. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, SOCIALCOM '10, pages 288–294, Washington, DC, USA, 2010. IEEE Computer Society.

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

[15] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 591–600, New York, NY, USA, 2010. ACM.

[16] J. Lee, P. Antoniadis, and K. Salamatian. Faving Reciprocity in Content Sharing Communities: A Comparative Analysis of Flickr and Twitter. In *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*, pages 136–143. IEEE, 2010.

[17] K. Lerman. Social information processing in social news aggregation. *IEEE Internet Computing: special issue on Social Search*, 2007.

[18] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1361–1370. ACM, 2010.

[19] Y. Liu, B. Gao, T. Liu, Y. Zhang, Z. Ma, S. He, and H. Li. BrowseRank: letting web users vote for page importance. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 451–458. ACM, 2008.

[20] J. Moon and N. Pullman. On generalized tournament matrices. *SIAM Review*, 12(3):384–399, 1970.

[21] G. Slutzki and O. Volij. Ranking participants in generalized tournaments. *International Journal of Game Theory*, 33:255–270, 2005.

[22] J. Weng, E. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM, 2010.

[23] D. Zhou, S. A. Orshanskiy, H. Zha, and C. L. Giles. Co-ranking authors and documents in a heterogeneous network. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, pages 739–744, Washington, DC, USA, 2007. IEEE Computer Society.