Tournament Models for Authority Identification in Online Communities

Suratna Budalakoti
Center for Identity
The University of Texas at Austin

Razieh Nokhbeh Zaeem
Center for Identity
The University of Texas at Austin
Email: Razieh [AT]
identity.utexas.edu

K. Suzanne Barber
Center for Identity
The University of Texas at Austin

Abstract— **Authority identification is an important problem in online information-sharing communities such as question-answer (Q&A) forums and online social networks (OSNs), where users care as much about the quality of information being accessed, as its alignment with their interests. This paper investigates a tournament model based approach to authority identification, where interactions between users are modeled as generated by a Bradley-Terry model. We derive a new measure of user authority, the average winnings score, for authority identification in Q&A forums, and evaluate it on data derived from the Stack Exchange Q&A forum. We also show how the log fair bets measure, which has been successfully used for authority identification in OSNs in the past, can be derived from tournament models. We also prove some key results related to a co-ranking framework, for combining information from multiple preference expression graphs based on the same OSN. We then demonstrate the empirical effectiveness of tournament model based approaches, in conjunction with the co-ranking framework.**

Keywords- Reputation, PageRank, Authority, Trust

I. Introduction

Two popular categories of communities where people search for expertise online are community question-answer (Q&A) forums, and information-sharing online social networks (OSNs)[1]. For example, a computer programmer looking for an answer to a technical question may consult an online Q&A forum such as Stack Overflow [18]. Similarly, a person looking for interesting content related to their fields of interest might search for profiles to follow on OSNs such as Twitter [19] or LinkedIn [8]. Despite looking quite different superficially, the interaction between users on a Q&A forum, and between members of an information-sharing OSN, has many similarities. On both Q&A forums and information-sharing networks, members proceed by making a claim to possessing certain skills or expertise, for example by answering a question, or via the text in their profile, or the content they share. These claims may then be endorsed by their peer group, for example, if a question they answered is accepted as the 'best answer', or if other members are impressed by their profile, or the content they share, and send an invitation to connect. Recommendation on such communities is as concerned with the degree of trust that can be placed in a shared piece of knowledge, besides how well it aligns with a user's interests. Thus, the computer programmer would like her question answered not just by a person with similar interests, but one who is an expert and likely to provide a good answer. And a user looking for profiles to 'follow' is likely to be interested in users with *authority*, or reputed users.

II. Background

Authority is important for recommender systems, as it tries to measure how interesting a piece of information would be to an unbiased observer. It plays an important role in many recommendation tasks where users want content that is of high quality and reliable, besides being aligned with their interests, such as content recommendation [14, 15, 36] and expert recommendation in question-answer forums [37].

In the fields of library science and philosophy, a popular definition of cognitive authority was provided by Wilson [30], as "influence on one's thoughts that one would consciously recognize as proper". Cognitive authority is distinguished from performative authority, which is based on the position a source occupies in some hierarchy. For online communities, the cognitive authority of a piece of information is the value an individual would be comfortable ascribing to it, if asked for their unbiased opinion, independent of concerns such as popularity, personal affinity, etc. (which play as similar role in OSNs as hierarchies might play in the real world).

The idea of an unbiased opinion can be translated to social networks in many ways. In the context of content recommendation, Yang et al. [36] defined a user's reputation (or authority) as 'the propensity of a random user interested in the topic (not necessarily connected to the sharer) to respond positively to a typical content item shared by the sharer'. Our previous work [5] defined authoritative users on an OSN as users who were influential outside the network, in the real world, where presence and page view traffic on Wikipedia [20] were treated as measures of real world influence. An intuitive way to understand authority is thus as the quality that is intrinsic to the user, and not contingent on the structure of the network, or their history within it. This paper formalizes this idea via the Bradley-Terry model [3, 10].

However, it is often difficult to externally validate the authority of users or sources of information on online communities. Users may be wary of sharing detailed information about themselves, and certain users may overstate

[1] We use the term information-sharing online social networks to describe both content-sharing OSNs such as Twitter [19] and professional OSNs such as LinkedIn [8].

their achievements, due to the inability to verify most claims. The aggregation of member preference expressions, such as ‘follows’, ‘best question’ ratings, etc., in a useful way, to arrive at an estimate of the authority ascribed to each member in the forum, is thus an important task.

A. *Preference Graphs*

We categorize preferences in online communities as first-person and third-person. A third person preference expression is one that is made by a user comparing two other users. For example in a Q&A forum, a questioner often receives multiple answers to her question, from different responders, and then chooses one user’s answer as the ‘best answer’. In this case, the questioner is expressing a third-person preference: indicating that she prefers the user giving the best answer, compared to the other users. A first-person preference expression, on the other hand, is where the user gives an indication that another user is more authoritative when compared to him/her. For example, questioner-responder roles in a Q&A forum provide information about first-person preferences: it seems fair to assume that the person asking the question is less authoritative than the person answering it.

Similarly, in information-sharing OSNs, actions such as ‘following’, or viewing a profile, or sending an invitation to connect, are asymmetric in the implied direction of authority ‘flow’. Leskovec et al. [27] in their analysis of online communities found that a status based interpretation serves as a good predictor of user endorsements. They observed that users are more likely to provide a positive rating to other users they perceive as being higher status, in terms of being more knowledgeable, etc., instead of being concerned with likability or personal relationships.

Both first and third-person preference expressions on Q&A forums and OSNs can be represented as graphs, where each vertex in the graph represents a member, and a directed edge from member i to member j indicates that, at in at least one interaction, the direction of preference expression was from i to j. So, depending on the type of graph, a directed edge in graph G from member i to j might represent that: a) i endorsed some content posted by j, for example, by ‘liking’ the content, or sharing it with others, or, b) i sent an invitation to j to connect, or, c) another user k, given a choice between users i and j for some action, chose user j over user i. In the context of Q&A forums, the edge might represent the fact that j answered a question asked by i, or j gave a higher rated answer than i for a question that both answered. The resulting graphs, referred to in this work as preference graphs, are then analyzed to identify the authoritative nodes. Often, multiple preference graphs can be constructed over the same OSN, each of which showcases different aspects of user behavior. For example, Weng et al. [35] found that on the OSN Twitter [19], famous people tend to have more ‘followers’, but ‘retweet’ graphs are often dominated by individuals who tend to contribute news. Significant differences between which profiles users send invitations to, and whose profile they view, were found [5] on LinkedIn [8]. The existence of multiple preference graphs across the same social network is an important reason why the tournament models discussed here are a good fit for the authority identification problem. As discussed below, these models are able to combine information from multiple preference graphs in a principled and scalable manner through a co-ranking framework.

B. *Preference Aggregation*

Preference aggregation models [34] usually assume that there is a latent random variable associated with each member, signifying its quality. A popular model for preference aggregation is the Bradley-Terry model [3]. Under this model, when two items i and j are compared to each other for expression of a pairwise preference, this process proceeds as follows: each item i has a distribution associated with it, from which it selects a level of quality w_i. Then, given two items *i* and *j*, with sample quality/strength w_i andw_j:

$$P(ipreferred)=P(w_i>w_j)=\frac{w_i}{w_i+w_j}$$

The distribution of w_i can be modeled in many ways, and the method of solving for w_i varies depending on the model [16, 33].

An alternate approach is to model w_i as a fixed but unknown value, as opposed to a value sampled from a distribution. Daniels [9] showed that if the outcome of a match in a round-robin tournament[2] is modeled using the Bradley-Terry model, the quality, or *strength* of a player is given by the *fair bets score*. The fair bets score of a node is defined as the PageRank [4] score of the node in the network graph, divided by the node’s outdegree. This approach, which has been known for a long time, has historically been ignored in favor of parametric Bradley-Terry models.

However, there are a number of advantages to this approach, which make it suitable for preference aggregation on the Web: a) since it is a variant of PageRank, it can take advantage of the vast amount of research into eigenvector calculation for large graphs, and b) as we have shown before [5] and explore further in this paper, it can be easily integrated into a co-ranking framework, for combining information from multiple preference graphs. This is extremely useful in social network analysis, where user preference expressions could be expressed across multiple graphs.

III. Research Contributions

This paper proposes the following approach to authority identification in Q&A forums and OSNs: a user’s authority is modeled as her intrinsic ability to influence another user in a positive way, in a single interaction involving her. This is formalized by viewing each user interaction as a match, the outcome of which, represented by the direction of the resulting edge between them, is determined by their relative authority ‘strengths’. The Bradley-Terry model [10] is assumed to model

[2] A round-robin tournament is one where each player plays against every other player exactly once.

the dynamics of a match between two players. The preference graph resulting from multiple such games can thus be seen as encapsulating the results of a tournament among users. The intrinsic 'player' strengths, or authority, can then be discovered from the tournament results.

The assumption that the preference graphs generated on Q&A forums are the product of an underlying Bradley-Terry model, and the interaction dynamics of Q&A forums, lead to a new tournament scoring model, which we call the average winnings model. We also show that a similar assumption of an underlying Bradley-Terry model for OSNs, combined with past research [26] on OSN graph evolution, leads to the log fair bets tournament model, which was empirically found [5] to be a good fit for authority identification on LinkedIn [8]. This paper provides a justification for the log fair bets model, based on existing literature on OSN graph evolution [26].

A co-ranking framework for combining authority information from multiple preference graphs was also presented before [5]. The framework relied on proving that connecting multiple preference graphs in a certain way, is equivalent to simultaneously using the authority scores of nodes in one graph as the random restart vector in the authority calculation in the other graph, and vice versa. This work proves some key results related to the co-ranking model, which provides a better understanding of the framework.

We empirically demonstrate the effectiveness of the average winnings model, in combination with a co-ranking framework, on data from the Stack Exchange Q&A forum [17]. We do not explore the empirical effectiveness of the log fair bets model in this paper, as it has already been explored for the LinkedIn social network [5].

IV. RELATED WORK

Recommendation algorithms often rely on signals from user interactions to estimate the relative reputation, or authority, of users in the network. This is usually done by using the natural representation of interactions in social networks as a graph, and using various measures of network centrality [2] to identify the important nodes in the graph. Historically, graph centrality measures, such as eigenvector centrality [2] and the Katz measure [22], have been used for identifying important nodes in social networks. These methods are closely related to the PageRank model [11], which, along with variants [25], has played an important role in recommendation in question-answer forums [37, 21], and online social networks [15, 14].

Network centrality and related measures are, however, a better measure of the influence of a user within a network, as opposed to their level of authority. 'Influencer' is a term commonly defined, as users who can induce other members to take certain actions, such as take interest in some information they share, etc. It seems reasonable to expect authoritative users to be the most influential in an OSN, but this is often not the case. There are many reasons for this. For example, less authoritative users can become more prominent on networks due to increased level of activity, or presence on the network [5]. This creates a selection bias [36], so that their content is more likely to be noticed. PageRank and related models, while effective in identifying influential nodes, are unable to take into account these biases.

To correct for such biases, Yang et al. [36] propose the use of additional unbiased data sources, which is used to calibrate the algorithms. Gayo-avello [12] attempts to correct for such biases by heuristically reducing weights of reciprocal links, a common source of such biases, while our previous work [5] attempts to normalize PageRank score by activity levels.

While, to the best of our knowledge, this is the first work to model OSN graph structure as a product of an intrinsic strength or quality of individual nodes, similar ideas have been explored for the graph structure of the Internet [1]. Another related work is the unbiased web ranking approach by Cho et al. [7] that takes multiple snapshots of the Web over time, to take into account the rate at which a node's PageRank score grows.

V. TOURNAMENT MODELS

In sports tournaments, contenders play against each other, and the results of these games need to be aggregated to a single ranking [29]. The same problem also exists in voting systems [28], where preferences of multiple individuals among options may need to be aggregated into a single decision. Tournaments provide a broad framework which can be used to model both first-person and third-person preference expressions, by interpreting an edge in the preference graph as a match; the direction of the edge indicates the winner of the match. The challenge, then, is integrate the information provided by the results of all the matches into a single ranking for all players (or users).

A. The Bradley-Terry Model [10]

Under the Bradley-Terry model [3, 10], we can formalize this as follows: suppose the authority strengths of OSN members i and j are given by $\mathrm{a_i}$ and $\mathrm{a_j}$ respectively, where $\mathrm{a_i} \in \mathrm{R}^+$ for any i. Say we treat any preference expression between i and j as a match, where the direction of the expression decides the winner and loser. The member the edge is directed from is considered the loser, while the member it is directed to is considered to have won. Then assuming a tournament took place between i and j, the probability that i won the game is given by $\frac{\mathrm{a_i}}{\mathrm{a_i+a_j}}$. Assuming no draws, it follows that the probability that j won the game is given by $\frac{\mathrm{a_j}}{\mathrm{a_i+a_j}}$.

A tournament matrix M can be constructed to represent the result of all such 'games', with $\mathrm{M_{ij}}$, containing the number of times player i lost to j. Assuming $\mathrm{N_i}$ is the total number of games played by i, and $\mathrm{N_{ij}}$ is the number of games between i and j. Then as $\mathrm{N_{ij}} \to \infty$, $\mathrm{M_{ij}}$ converges to its expected value, $\mathrm{N_{ij}} \frac{\mathrm{a_j}}{\mathrm{a_i+a_j}}$ almost surely. The resulting matrix at convergence may be referred to as the asymptotic tournament matrix. Then:

$$E\left[\frac{M_{ij}}{N_{ij}}\right] = \frac{a_j}{a_i + a_j}$$

Based on this observation, the approach taken in this work is to treat the currently available tournament matrix at any given time, as the asymptotic tournament matrix; the assumption being that the current matrix will eventually converge to this state as time progresses.

B. The Average Winnings Model

Consider data drawn from a Q&A forum. Let N_{ij} be the number of questions that both i and j have answered. Let Z_{ij} be the number of times player i lost to player j, and let Z_i be the total number of games lost by i. This information can be represented as a tournament matrix Z, and also as a preference graph. We assume that $Z_{ij} = N_{ij}\frac{a_j}{a_i+a_j}$.Then, knowing Z_{ij} and N_{ij} for all pairs, the goal it to calculate a_i for all i.

For numerical purposes, we need to ensure that he graph is strongly connected. This can be achieved by modifying Z in two ways: a) by introducing a regularizing node in the graph [6], or, b) by adding a small probability of jumping to a randomly chosen node, as per the PageRank model [4]. Given the popularity of the PageRank algorithm [4] in the Web research community, we select the second option. Then the following proposition provides a method for calculating the authority strengths of the participating players.

Proposition 1

Let Z be a $K \times K$ tournament matrix of results based on an underlying Bradley-Terry model, so that the probability i loses to j is given by $\frac{a_j}{a_i+a_j}$. Let the number of games played between i and j be $N_{ji} = N_{ij}$, and let N_i represent the number of games played by i. Then construct a matrix P where $P_{ij} = \frac{Z_{ij}}{Z_i}$ and $Z_i = \sum_{j=1}^{K} Z_{ij}$. Then, assuming the Markov chain corresponding to Z is ergodic,$a_i = \frac{\pi_i}{Z_i}$ where $\vec{\pi}$ is the stationary distribution of the Markov chain corresponding to this matrix.

Proof:

The proof uses the property that any ergodic Markov chain that satisfies the detailed balance equations given by $\pi_i P_{ij} = \pi_j P_{ji}$ has a unique stationary distribution, given by scaling $\vec{\pi}$ to add to 1 [13]. For P, the detailed balance equation between two states i and j are given by:

$$\pi_i \cdot \frac{N_{ij}}{Z_i} \cdot \frac{a_j}{a_i + a_j} = \pi_j \cdot \frac{N_{ij}}{Z_j} \cdot \frac{a_i}{a_i + a_j}$$

Setting $\pi_i = a_i \cdot Z_i$ balances the equation. Then a_i is given by:

$$a_i = \frac{\pi_i}{Z_i}$$

In other words, under the Bradley-Terry model, the authority score of a node is given by its PageRank score, divided by the number of games it has lost. We refer to this score as the *average winnings* score.

This section presented a new method, the average winnings model, for assigning authority scores to nodes in an OSN. The next section presents the *fair bets* model [9, 29, 31], a model introduced by Daniels [9] for ranking players in round robin tournaments, where each player plays against another player exactly once.

C. The Fair Bet Model[9]

The fair bets model calculates player strength scores based on a *generalized tournament matrix* [31]. A tournament matrix M can be converted to a generalized tournament matrix V by normalizing the scores of all pairs of players, so that their total number of games played sums to 1. That is $V_{ij} + V_{ji} = 1$. Many interactions on information-sharing OSNs are generalized tournament matrices. An example is the invitation graph, consisting of information about who sent whom the first invitation to connect.

Suppose a stochastic matrix P is constructed from V by normalizing each row. That is:

$$P_{ij} = \frac{V_{ij}}{\sum_{k=1}^{K} V_{ik}} = \frac{V_{ij}}{\deg^{+}(i)}$$

Here $\deg^{+}(\mathrm{i})$ represents the out-degree of vertex i, if all V_{ij} are integers or the sum of the row otherwise. We use the notation $\deg^{+}(i)$ to represent both.

P is clearly a row stochastic matrix. Assuming, for now, that P is aperiodic, and thus ergodic, then the following proposition is true [9].

Proposition 2

Given an asymptotic generalized tournament matrix V that is ergodic, so that for any game between i and j, $P_{ij} = \frac{a_j}{a_i+a_j}$. Then $a_i = \frac{\pi_i}{\deg^{+}(i)}$, scaled by a constant factor, where π is the stationary distribution for P.

Proof:

For P, the detailed balance equations are given by:

$$\pi_i \frac{a_j}{deg^{+}(i)(a_i+a_j)} = \pi_j \frac{a_i}{deg^{+}(j)\cdot(a_i+a_j)}$$

They are satisfied by setting $\pi_i = a_i \cdot deg^{+}(i)$. Then:

$$a_i = \frac{\pi_i}{deg^{+}(i)}$$

Thus the authority vector $\vec{a}$ can be estimated from an asymptotic tournament matrix V, by calculating its stationary distribution $\vec{\pi}$, and then calculating$a_i = \frac{\pi_i}{\deg^{+}(i)}$.

1) Accounting for Unequal Win-Loss Probabilities

A key difference between Q&A forums and professional OSNs is that ‘matches’ on Q&A forums require both users to participate, and provide both users with a chance to ‘win’ the match. In contrast, on OSNs, wins or losses may be more likely at a given time for a node, simply due to network dynamics.

This fact needs to be taken into account while measuring authority scores.

Leskovec *et al.* [26] found that the out-degree of nodes grows exponentially with age (time since they joined the network). That is, for a node aged A, $\deg^{+}(i) \propto e^{A}$, or $A \propto \log(\deg^{+}(i))$. This was because the longer a member had been active on an OSN, the more active they tended to be. However, they do not provide a similar analysis for the in-degree, so we assume there is some function f(A) which relates the in-degree of a node to its age in the graph. We plan to investigate the nature of this function in future work.

Then, at any age of a node A, the win to loss ratio will approximate $\frac{f(A)}{e^{A}}$. Assuming $f(A)$ grows slower than exponentially, so that $f(A) < e^{A}$ for all A, we can interpret this in our framework, that we systematically over-sample losses compared to wins[3]. So, for a player i, we correct for this over-sampling by reweighing V_{ij} as $V'_{ij} = \frac{f(A_i)}{e^{A_i}} \cdot V_{ij}$, where A_i is the age of i. Approximating $e^{A_i} \approx \exp(\log(deg^{+}(i)))$, we can write $\frac{f(A_i)}{e^{A_i}} = \frac{f(A_i)}{deg^{+}(i)}$.

The stochastic matrix constructed from V' by normalizing each row is identical to P constructed for fair bets:

$$P_{ij}=\frac{\frac{f(A_i)}{\mathrm{deg}^{+}(i)}V_{ij}}{\sum_{k=1}^{K}\left(\frac{f(A_i)}{\mathrm{deg}^{+}(i)}V_{ik}\right)}=\frac{V_{ij}}{\mathrm{deg}^{+}(i)}$$

However, since $\sum_{k=1}^{K} V_{ik}=\mathrm{deg}^{+}(i)$, the normalization factor (denominator) for the i row in V' is given by $f(A_i)$, the rate at which the indegree grows with age. Plugging this change in Proposition 2 would give us the authority score of node i as $\frac{\pi_i}{f(A_i)}$.

That is, the authority score of a member in an OSN is given by their PageRank score, divided by their expected indegree, as a function of their age. The log fair bets model can then be understood as assuming that the indegree grows linearly with the age of the node, thus approximating $f(A_i)=\log(\mathrm{deg}^{+}(i))$. It may be expected that better models of OSN indegree growth with age will lead to better authority models. We plan to investigate models of indegree growth with age in future work.

D. Average Winnings and Fair Bets: Comparison

Because the fair bets model was designed for ranking in round robin tournaments, it assumes a single interaction between any two players. As mentioned earlier, this is a natural situation in many graphs on OSNs, such as for invitation graphs. In the case of multiple interactions, these interactions are normalized to add up to a single game. In comparison, the average winnings model counts each interaction separately.

The fair bets model is better suited to situations where we are only concerned with the directionality of preference, and not the quantity. Consider a situation where user A tends to re-share a lot of the content by user B. Once we have established the direction, a large quantity of re-sharing is more likely to indicate the degree of affinity or agreement, and not an extreme difference in the relative authority/status of A and B.

In contrast, in a Q&A forum, the number of times one player provided the best answer when contesting with another player contains important information about their relative level of authority. This is because the relative skills and knowledge play an important role in each interaction between them. The average winning model, by counting each win separately, is thus better suited to Q&A forums, while the fair bets model is a good model of user behavior on OSNs.

VI. Co-Ranking Preference Graphs

One key advantage of tournament-based models is that, they allow us to combine information from multiple preference graphs in a principled way. This can be done using a co-ranking model (or bimodal ranking model) first presented in our previous work [5]. The model is discussed below for completeness.

The model takes advantage of the random restart vector in the PageRank model. In the PageRank model, at each time step, with a certain probability $1 - d$ (usually set to 0.85), the random walk randomly selects an outgoing link from the current node. With the remaining probability d, the walk jumps to a randomly chosen node in the graph. The probability d is referred to as the *random restart (RSR) probability*, and the vector the new page is chosen from is called the *RSR vector*. The vector can be uniform, or biased to reflect some prior information.

Now say, for example, we have two preference expression graphs. For a Q&A forum, one graph G_A may consist of directed edges from each user who answered a given question, to the user who provided the best answer. Another graph G_N may consist of a directed edge, for each question, from the user who asked the question, to the user who provided the answer. Both graphs are over the same set of vertices (users), so that for each vertex in one graph, there is a *twin vertex* in the other graph. The co-ranking approach proposes the following algorithm:

1. Select one of the two graphs, say G_N, at random. Calculate the PageRank vector $r_N^{(1)}$ for G_N, using a uniform RSR vector z_0, and a RSR probability 0<*d*<1. Calculate the PageRank vector $r_A^{(1)}$ for G_A, using $r_N^{(1)}$ as the RSR vector.

2. Repeat till $r_A^{(t)}$ does not change: at iteration *t*, calculate the PageRank vector $r_N^{(t)}$ for G_N using $r_A^{(t-1)}$ as the

[3] If $f(A) > e^{A}$ for all A, we would be under-sampling losses.

RSR vector. Next calculate the PageRank vector $r_A^{(t)}$ for G_A, using $r_N^{(t)}$ as the RSR vector.

3. Suppose the process stops at time step *t*'. Then set final PageRank vectors $r_A = r_A^{(t')}, r_N = r_N^{(t')}$.

We prove that: a) the above algorithm converges and, b) is equivalent to simultaneously using the PageRank vector of one graph as the RSR vector of the other graph, and vice versa, in Appendix A. Proof for this claim was not provided in [5]. However, [5] proved the following proposition, which allows us to avoid doing multiple PageRank runs on both graphs. Instead a special composite graph can be constructed, and the PageRank score calculated on which gives us the same result as the multiple iteration algorithm described above.

Proposition:

Given two graphs $G_A = (V_A, E_A)$ and $G_N = (V_N, E_N)$, 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 one graph is connected edge to its twin vertex *v*' in the other graph via a directed edge of weight *d*. A similar directed edge of weight d connects *v*' to *v*. Then the PageRank vector for the graph *G*, normed so that the scores for vertices in V_A and V_N each sum to 1, is equal to simultaneously using the PageRank vector of one graph, as the RSR vector of the other graph, and vice versa.

Proof:

See [5], Section 5.2.1.

A. Co-ranking with Other Authority Models

As described in our previous work [5], other models besides PageRank can easily be used for co-ranking. For example, weighing the edge between twin vertices by 1/log(*outdegree*) is equivalent to using the log fair bets score of graph as the RSR vector of the other graph, and vice versa. Similarly, setting the weight to 1/N(matches lost) would be equivalent to using the average winnings score. The weight need not be symmetric, so that the weight of the edge from *v* to *v*' can differ from the weight of the edge from *v*' to *v*.

VII. Experimental Results

The co-ranking based authority-identification algorithm was evaluated in the Q&A context, using the StackExchange dataset [17]. The problem was set up as follows: the co-ranking approach and other baseline algorithms were used to rank the participants in descending order of authority. Following this at each time step, the algorithms were presented with the set of all responders for a question, and were expected to recommend one of these responders, as the one most likely to provide the best answer (as rated by the questioner). The assumption was that, if a user voluntarily decided to answer a question, the question certainly matches her interests. Among a set of users whose interests match a question, an algorithm that judges authority accurately should be able to identify the user most likely to provide the best answer.

Community	**PageRank**	**Top Best Responder**	**Av. Winnings Co-ranking**
Math	59.7	58.9	67.5
Physics	58.7	58.6	65.9
Security	60.9	61.1	66.9
AskUbuntu	51.9	51.8	68.5
ServerFault	55.6	55.6	67.6
English	50.8	51.4	67.1

Table 1: Comparison of the PageRank-Average Winnings Co-Ranking Model in terms of 'best answer' prediction accuracy, given a list of responding users.

The co-ranking approach used two graphs: a first-person preference graph with edges from questioner to responder, and a third-person preference graph with directed edges from each person who 'lost' in a question, to the winner. The average winnings model was used for authority estimation over the tournament graphs. The final authority 'rank' of a node was the average of the rank of the node in each graph. However, note that the two ranking included information from each other, due to the use of the co-ranking framework.

Besides the co-ranking and the PageRank algorithm, a top best responder, which always predicted the person among the responders who has given the most 'best answers', was also used as a recommender. The results are shown in Table 1. As can be seen, the average winnings co-ranking model outperforms other models, while the PageRank model under performs simpler approaches. The average improvement of the co-ranking approach over the PageRank model was 20.1%, while the median improvement was 17.2%.

VIII. Conclusions

This paper discussed the problem of authority identification in Q&A forums and information-sharing OSNs. It presented a tournament based model of the preference expressions by users on such networks. Depending on the nature of interactions on such forums, two key authority models, average winnings and log fair bets were discussed. A key advantage of these models was the ability to combine signals from multiple preference graphs. A co-ranking model for combining signals from multiple graphs in a principled and scalable way was discussed, and key results related to this model presented.

References

[1] G. Bianconi and A.-L. Barabási. Competition and multiscaling in evolving networks. EPL (Europhysics Letters), 54(4):436, 2001.

[2] P. Bonacich. Power and centrality: A family of measures. American journal of sociology, pages 1170–1182, 1987.

[3] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika, pages 324–345, 1952.

[4] 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.

[5] S. Budalakoti and R. Bekkerman. Bimodal invitation-navigation fair bets model for authority identification in a social network. In Proceedings of the 21st international conference on World Wide Web, pages 709–718. ACM, 2012.

[6] X. Chen, P. N. Bennett, K. Collins-Thompson, and E. Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In Proceedings of the sixth ACM international conference on Web search and data mining, pages 193–202. ACM, 2013.

[7] J. Cho, S. Roy, and R. E. Adams. Page quality: In search of an unbiased web ranking. In Proceedings of the 2005 ACM SIGMOD international conference on Management of data, pages 551–562. ACM, 2005.

[8] L. Corp. Linkedin. www.linkedin.com.

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

[10] H. David. The method of paired comparisons. Griffin's statistical monographs & courses. Hafner Pub. Co., 1963.

[11] M. Franceschet. Pagerank: standing on the shoulders of giants. Communications of the ACM, 54(6):92–101, June 2011.

[12] D. Gayo-Avello. Nepotistic relationships in twitter and their impact on rank prestige algorithms. arXiv preprint arXiv:1004.0816, 2010.

[13] G. R. Grimmett and D. R. Stirzaker. Probability and random processes. Oxford University Press, USA, 2001. Chapter 6.

[14] P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh. Wtf: The who to follow service at twitter. In Proceedings of the 22Nd International Conference on World Wide Web, WWW '13, pages 505–514, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[15] L. Hong, R. Bekkerman, J. Adler, and B. D. Davison. Learning to rank social update streams. In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12, pages 651–660, New York, NY, USA, 2012. ACM.

[16] D. Hunter. MM algorithms for generalized Bradley-Terry models. The Annals of Statistics, 32(1):384–406, 2004.

[17] S. E. Inc. Stack Exchange Data Explorer. http://data.stackexchange.com/.

[18] S. E. Inc. The Stack Exchange Network. www.stackexchange.com.

[19] T. Inc. Twitter. http://www.twitter.com.

[20] W. Inc. Wikipedia, 2010. www.wikipedia.org.

[21] W.-C. Kao, D.-R. Liu, and S.-W. Wang. Expert finding in question-answering websites: A novel hybrid approach. In Proceedings of the 2010 ACM Symposium on Applied Computing, SAC '10, pages 867–871, New York, NY, USA, 2010. ACM.

[22] L. Katz. A new status index derived from sociometric analysis. Psychometrika, 18(1):39–43, 1953.

[23] J. G. Kemeny and J. L. Snell. Finite markov chains. Springer-Verlag New York, 1976.

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

[25] R. Lempel and S. Moran. Salsa: the stochastic approach for link-structure analysis. ACM Transactions on Information Systems (TOIS), 19(2):131–160, 2001.

[26] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '08, page 462, 2008.

[27] 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.

[28] J. Levin and B. Nalebuff. An Introduction To Vote Counting Schemes. Journal of Economic Perspectives, 9(1):3–26, 1995.

[29] J. Moon. On generalized tournament matrices. SIAM Review, 12(3):384–399, 1970.

[30] W. Patrick. Cognitive Authority. In Secondhand Knowledge: An inquiry into cognitive authority, chapter 2, pages 13–38. Greenwood Press, 1983.

[31] G. Slutzki and O. Volij. Ranking participants in generalized tournaments. International Journal of Game Theory, 33(2):255–270, 2005.

[32] G. Strang. Introduction to Linear Algebra. Wellsley-Cambrige Press, 2003.

[33] K. F. Thuesen. Analysis of Ranked Preference Data. PhD thesis, Technical University of Denmark, 2007.

[34] K. Tsukida and M. Gupta. How to analyze paired comparison data. Technical Report 206, Dept. of Electical Engineering.(No. UWEETR-2011-0004)., Washington Univ., Seattle, 2011.

[35] 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.

[36] J. Yang, B.-C. Chen, and D. Agarwal. Estimating sharer reputation via social data calibration. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13, pages 59–67, New York, NY, USA, 2013. ACM.

[37] J. Zhang, M. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms. In Proceedings of the 16th international conference on World Wide Web, page 230. ACM, 2007.

APPENDIX

A Bimodal Co-ranking

A.1 Proof of Convergence

To prove:

1. The algorithm described in Section VI converges after a certain number of iterations *t*.

2. The PageRank vectors $r_A = r_A^{(t)}$ and $r_N = r_N^{(t)}$ satisfy the following equations at convergence:

$$\left((1-d)P_A + d e r_N^{\top} \right)^{\top} r_A = r_A$$

$$\left((1-d)P_N + d e r_A^{\top} \right)^{\top} r_N = r_N$$

Outline:

We prove the result by showing that the bimodal co-ranking algorithm is equivalent to applying the power iteration eigenvalue algorithm [32] to a specially constructed positive column stochastic matrix *M*, where:

$$M = (d^2 (I - (1-d)P_N^{\top})^{-1} (I - (1-d)P_A^{\top})^{-1}$$

Since the power iteration algorithm converges for positive stochastic matrices in a finite number of steps, the process described above converges as well. Following this, part () is shown algebraically.

To simplify the notation below, *let $c=1-d$.*

Claim 1:

For a row stochastic matrix P, construct another matrix $A = cP^{\top} + (1-c) r e^{\top}$, where $e_i = 1$ for all i, $0 < c < 1$ and r is a positive vector with $|r|_1 = 1$. That is, r is a teleportation vector added to P, and $1-c$ is the teleportation probability. Then the solution to the PageRank equation for A, $x=Ax$, is given by $x=(1-c)(I-cP^{\intercal})^{-1}r$.

*Proof***:**

$$x=Ax=[cP^{\intercal}+(1-c)re^{\intercal}]x=cP^{\intercal}x+(1-c)r$$

$$\Rightarrow x=(1-c)(I-cP^{\intercal})^{-1}r$$

Claim 2:

For a row stochastic matrix P, $0<c<1$, $\sum_{k=0}^{\infty}(cP)^k=(I-cP)^{-1}$.

Proof:

This follows from the fundamental matrix theorem [23, Theorem 3.2.1], which states that, for any absorbing markov chain with transition matrix Q, $\sum_{k=0}^{\infty}(Q)^k=(I-Q)^{-1}$. This can be shown by considering the identity:

$$(I-Q)(I+Q+Q^2+Q^3+\ldots+Q^{n-1})=I-Q^n$$

$$\Rightarrow \sum_{k=0}^{n-1}(Q)^k=(I-Q)^{-1}(I-Q^n)$$

As $n\to\infty$, $(I-Q^n)\to I$. We know $(I-Q)^{-1}$ is nonsingular because $I-Q$ is diagonally dominant [32]. Thus, $\sum_{k=0}^{\infty}(Q)^k=(I-Q)^{-1}$.

As cP is an absorbing markov chain with a probability $d=1-c$ of absorption at each time step, setting $Q=cP$, the claim is correct.

Claim 3:

For an irreducible row stochastic matrix P and $0<c<1$, $X=(1-c)(I-cP)^{-1}$ is a positive row stochastic matrix.

Proof:

We show that for a row stochastic matrix P with K rows, each row of $S=(I-cP)^{-1}$ sums to $\frac{1}{1-c}$. As a result, $X=(1-c)\,S$ is a row stochastic matrix. Using the relationship .

$$S_{ij}=\begin{cases}0+cP_{ij}+c^2(P^2)_{ij}+\cdots if\ i\ \neq j\\ 1+cP_{ii}+c^2(P^2)_{ii}+\cdots if\ i\ =j\end{cases}$$

Then

- $\sum_{j=1}^{K}S_{ij}=1+c\sum_{j=1}^{K}P_{ij}+c^2\sum_{j=1}^{K}(P^2)_{ij}+c^3\sum_{j=1}^{K}(P^3)_{ij}+\ldots$ (7)

Since P is a row stochastic matrix and the product of row stochastic matrices is a row stochastic matrix, rows of P^n sum to 1 for all n. Thus the above equation can be written as:

$$\sum_{j=1}^{K}S_{ij}=1+c+c^2+\ldots=\frac{1}{1-c}$$

To show that all entries of S are positive, since P is an irreducible matrix, for some $0<k<K$ (K is the number of vertices), $c^k(P^k)_{ij}>0$. Thus S has all positive entries. Hence $X=(1-c)S$ is a positive row stochastic matrix.

Claim 4:

The co-ranking process for two graphs with irreducible stochastic matrices P_A and P_N is equivalent to calculating the eigenvector corresponding to the largest eigenvalue of a column stochastic positive matrix $M=(1-c)^2(I-cP_N)^{-\intercal}(I-cP_A)^{-\intercal}$, using the power iteration algorithm [24], and will thus converge in a finite number of steps. The result is equivalent to using the PageRank vector of one graph as the teleportation vector of the other graph, and vice versa.

Proof:

Let z_0 be a uniform stochastic vector we start with, and let the vector after t application of alternate PageRank runs be z_t. Then, using Claim 1, the co-ranking process can be written as:

$$z_{t+1}=((1-c)(I-cP_N{}^{\intercal})^{-1}(1-c)(I-cP_A{}^{\intercal})^{-1})z_t$$

$$\Rightarrow z_{t+1}=[(1-c)(I-cP_N)^{-\intercal}(1-c)(I-cP_A)^{-\intercal}]^t z_0$$

Here t is the number of steps till convergence (infinite if the process does not converge).

Using Claim 3, it can be seen that the above equation is the product of the transpose of two positive row stochastic matrices, and hence is a positive column stochastic matrix. Let this matrix be:

$$M=(1-c)^2[(I-cP_N)(I-cP_A)]^{-\intercal}$$

Then this is equivalent to applying the power iteration algorithm to the positive stochastic matrix M, and as a result is guaranteed to converge in a finite number of steps [24], to a unique positive eigenvector corresponding to the largest eigenvalue of M [32].

A.2 Proof of Equivalence

We now show that the above process is equivalent to using the PageRank vector of each graph as the other graph's teleportation vector. Assume that the co-ranking algorithm required t alternate runs of PageRank on P_N and P_A to converge, with P_N randomly chosen to be first (initially multiplied with z_0). In this case, the last PageRank calculation would be applied to P_A. The code stops at the t+1 run, when it realizes it has converged. The last run calculates $M^{t+1}z_0$, with$M^{t+1}z_0=M^t z_0$. Let the final converged values of PageRank vectors for P_A and P_N be r_A and r_N respectively. For brevity, we use both d and c=1−d below. Then we can write:

$$r_N \quad = d(I-cP_N)^{-\top}M^t z_0$$

Since$M^t z_0=M^{t+1}z_0$:

$$r_N \quad = d(I-cP_N)^{-\top}M^{t+1}z_0$$

For r_A, we calculate as follows:

$$r_A \quad = d(I-cP_A)^{-\top}d(I-cP_N)^{-\top}M^t z_0$$

Then we can write:

$$r_A \quad = (1-c)(I-cP_A)^{-\top}r_N$$

Then

$$r_N = d(I-cP_N)^{-\top}MM^t z_0$$

$$\Rightarrow r_N = d(I-cP_N)^{-\top}d^2(I-cP_A)^{-\top}(I-cP_N)^{-\top}M^t z_0$$

$$\Rightarrow r_N = (1-c)(I-cP_N)^{-\top}r_A$$

Based on Claim 1, we have the solution to $r_A=A_1r_A$, where:

$$A_1=((1-d)P_A+der_N^{\top})^{\top}$$

Similarly, we have the solution to $r_N=A_2r_N$, where:

$$A_2=((1-d)P_N+der_A^{\top})^{\top}$$

This proves the second part.