The Dissertation Committee for Suratna Budalakoti
certifies that this is the approved version of the following dissertation:

Authority Identification in Online Communities and Social Networks

Committee:

K. Suzanne Barber, Supervisor

Aristotle Arapostathis

Vijay Garg

Matthew Lease

Risto Miikkulainen

Authority Identification in Online Communities and Social Networks

by

Suratna Budalakoti, B.E.;M.S.E.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2013

Dedicated to my parents, Shailendra and Nivedita Budhalakoti, to Sailaja and Nishka.

Acknowledgments

I would like to express my appreciation for everyone who helped me in my research and academic life at the University. My advisor Dr. Barber patiently guided my research, while allowing me the freedom to pursue the direction of my interest, even on the many opportunities when I am sure I tried her nerves. Prof. Arapostathis, Prof. Garg, Prof. Lease and Prof. Miikkulainen greatly shaped this dissertation via their encouragement and judicious feedback. Dr Lease was very generous with his time, and conversations with him deepened my understanding and appreciation of the field of information retrieval. I'd also like to thank all the teachers I had at UT, who taught me so much, especially Dr. Barber, Dr. Garg, Dr. Ghosh, Dr. Julien, Dr. Sanghavi and Dr. Stone. Also, fellow members of the lab: David DeAngelis, whose feedback was indispensable to developing many of the ideas in this work, and also Jaesuk Ahn, Karen Fullam, Chris Jones, Rajiv Kadaba and Rick Scott, who helped me in my work, besides being good friends. And a special word of thanks to Melanie Gulick, Charlotte Harris and Stephanie Cardenas for making sure everything was always in place on the administrative side.

Outside the University, I would like to thank Ashok Srivastava at the NASA Ames Research Center, for serving as a mentor at a crucial time in my career. Ron Bekkerman at LinkedIn played an important role in this research,

and this work could not have been completed without his help. Discussions with Partha Saha at Yahoo! were also instrumental in guiding the direction of this work.

Finally I would like to thank my parents, and Sailaja, for their constant support and patience. Nishka helped more than she knew. And finally to Suvrat, thank you.

Authority Identification in Online Communities and Social Networks

Publication No. _____

Suratna Budalakoti, Ph.D.
The University of Texas at Austin, 2013

Supervisor: K. Suzanne Barber

As Internet communities such as question-answer (Q&A) forums and online social networks (OSNs) grow in prominence as knowledge sources, traditional editorial filters are unable to scale to their size and pace. This absence hinders the exchange of knowledge online, by creating an understandable lack of trust in information. This mistrust can be partially overcome by a forum by consistently providing reliable information, thus establishing itself as a reliable source. This work investigates how algorithmic approaches can contribute to building such a community of voluntary experts willing to contribute authoritative information. This work identifies two approaches: a) reducing the cost of participation for experts via matching user queries to experts (question recommendation), and b) identifying authoritative contributors for incentivization (authority estimation). The question recommendation problem is addressed by extending existing approaches via a new generative model that augments textual data with expert preference information among different questions.

Another contribution to this domain is the introduction of a set of formalized metrics to include the expert's experience besides the questioner's. This is essential for expert retention in a voluntary community, and has not been addressed by previous work. The authority estimation problem is addressed by observing that the global graph structure of user interactions, results from two factors: a user's performance in local one-to-one interactions, and their activity levels. By positing an intrinsic authority 'strength' for each user node in the graph that governs the outcome of individual interactions via the Bradley-Terry model for pairwise comparison, this research establishes a relationship between intrinsic user authority, and global measures of influence. This approach overcomes many drawbacks of current measures of node importance in OSNs by naturally correcting for user activity levels, and providing an explanation for the frequent disconnect between real world reputation and online influence. Also, while existing research has been restricted to node ranking on a single OSN graph, this work demonstrates that co-ranking across multiple endorsement graphs drawn from the same OSN is a highly effective approach for aggregating complementary graph information. A new scalable co-ranking framework is introduced for this task. The resulting algorithms are evaluated on data from various online communities, and empirically shown to outperform existing approaches by a large margin.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Expert finding [58] is the problem of identifying experts in a particular topic, based on evidence drawn from a dataset. The problem has traditionally been studied in the context of enterprise data: for example, to identify experts inside an organization [135]. However, the advent of the Web over the past decade has radically lowered the barriers to information exchange, making it possible for people to seek expertise outside their immediate peer group. For example, a computer programmer faced with a technical problem can seek advice from a volunteering expert on a question-answer forum, broadening the range of expertise available to her. Or an organization looking for an expert in a field could turn to profiles on online social networks to find a suitable candidate. These developments have considerably broadened the scope of applications for the task, while introducing several interesting challenges.

This dissertation focuses on the problem of identifying experts, and matching them with problems that match their expertise, in online communities. This work identifies two kinds of online communities where people and organizations search for expertise: community question-answer (Q&A) forums and professional online social networks.

Question-answer forums are primarily intended to connect information seekers with experts. However, as documented by Paul *et al.* [119], they also function as social networks, where users meet to increase their social reach and enhance their reputation among their peer group. Harper *et al.* [72] divide questions on such forums into two categories: informational and conversational. Informational questions are intended to seek useful information, while the aim of conversational questions is to establish and maintain social ties. Due to this dual aspect, Q&A forums are frequently classified as online social networks [70, 104, 111].

Interestingly, professional online social networks (OSNs) have a similar dual aspect. Apart from being a forum for maintaining social and professional ties, they also function as a place for showcasing a user's expertise and abilities. For example, a connection on a professional OSN such as LinkedIn [46] is often treated by third parties as a tacit endorsement. Often the OSN allows referrals to be made more explicit by writing a recommendation, or endorsing another user for a particular skill mentioned in their profile [47].

The distinction between Q&A communities and professional OSNs is thus, quite narrow in practice. In both communities, users proceed by making a claim to possessing certain skills or expertise, for example by answering a question, or via the text in their profile. These claims then may or may not be endorsed by their peers. The key challenge in identifying experts, or *authorities* (the two terms are used interchangeably in this work) is the aggregation of these endorsements in a useful way. An additional challenge

that may be encountered is identifying the topic of expertise the endorsement applies to. These two problems form the focus of this work.

1.1 Motivation

Patrick [118] defines cognitive authority as "influence on one's thoughts that one would recognize as proper". Traditional information sources rely on mechanisms which do not exist for the Web, such as editorial selection, institutional reputation and peer review, to establish the cognitive authority of information. The absence of such filters hinders the exchange of knowledge online, by creating an understandable lack of trust in information.

This lack of trust can be partially overcome by a forum for a user, by establishing itself as a reliable source through multiple positive experiences [97, 142]. A collaborative website relying on volunteers needs consistent participation from authoritative users to achieve this. This research identifies two approaches which can help a website progress towards the goal of increased participation by authoritative users: a) reducing the cost of participation, so that the website can select the highest quality contributions from a wider set of volunteers, and b) identifying authoritative users among contributors, and incentivizing them for participation. Often a combination of both is needed to encourage expert participation.

The next two sections discuss the specific problems in Q&A forums and OSNs identified by this research, that are motivated by the broad goal of encouraging behaviors that enhances the reputation of an online community.

1.1.1 Question-answer Forums

Community question and answer (Q&A) forums allow users access to expertise over the Internet, enabling them to find information when ordinary search results do not meet expectations. This could happen if a search query is hard to formulate for a question, if the information is not available online, or for questions requiring personalization or opinions. In addition to several specialized forums dedicated to specific domains, general-purpose question-answer forums are extremely popular: for example, the Q&A website Yahoo! Answers [85], one of the first popular Q&A forums, has reported having over 200 million users and over a billion resolved questions [86]. Despite this success, doubts about the quality of expertise available on these forums have been persistent [2, 6, 9]. Adamic *et al.* observed that the quality of expertise available on Yahoo! Answers is 'broad rather than deep' [2].

To avoid this problem, later forums evolved a different approach. As noted by Anderson *et al.* [8], "one direction this evolution has taken is the development and maturation of sites such as Stack Overflow and Quora built around focused communities in which a significant fraction of the participants have deep expertise". As a result, the core value of these forums resides in the voluntarily participating experts. Growing and retaining this community is essential to the success of such a forum.

On the other hand the expert finding problem, in the enterprise space and more recently in question-answer forums [70, 109], has traditionally been studied from the questioner's perspective. That is, given a question, the aim

is to identify experts in the question topic. This goal ignores the perspective of the responding experts, though due to the voluntary nature of expert participation, it is being recognized [35, 75] that retaining high-quality responders is the key to long-term success for the forum. This is a less serious problem in traditional applications such as enterprise expert search, where helping out colleagues may be part of the job description. In fact, in the enterprise we may be more concerned with the questioner's satisfaction, so that they can be more productive in the workplace.

An interesting aspect of Q&A recommendation is that recommending a responder for a question is at the same time, equivalent to a recommendation to the responder: the expectation being that the responder will be sufficiently interested in the question to provide an answer [35, 75]. In other words, responders have their own preferences among questions, and can lose interest if recommended too many irrelevant questions. This was emphasized by Horowitz and Kamvar [75] in their study of the social Q&A engine Aardvark, which took into account preferences such as frequency of contact, time of day, etc., while identifying relevant experts for a question. This research [35] formalizes this observation by introducing metrics based on the idea that Q&A forums should be motivated as much by minimizing responders' load of irrelevant questions, as by the quality of recommendations made to the questioner.

For example, consider two expert finding algorithms, both of which find an expert responder for a similar fraction of questions. However, the first

algorithm ensures that no expert sees more than three irrelevant questions for each question they answer, while the second algorithm makes no such commitment. So, even though both approaches find experts at the same rate, this work argues [35] that the first algorithm is better, as it promotes the long-term health of the forum via a lower expert load, and develops metrics to formalize this intuition.

Another way to encourage expert participation, apart from reducing the time investment in finding relevant questions, is to provide them with social recognition. Most Q&A forums do not provide any monetary benefit to contributing experts (one exception being the now-defunct Google Answers [55]). An important incentive, then, is visibility, in the form of 'points', 'badges'[7, 11], or by being highlighted as an important contributor (for example, 'contributor of the week', 'trending user', etc.). It is known that forum-wide exposure and feedback from peers makes it more likely that a content contributor will continue participating [38]. More recently, Q&A forums such as the Stack Exchange network [80] have started to showcase their best contributors to employers, thus monetizing as job forums [81] as well as providing another incentive for contribution.

A common approach to rewarding experts is to provide recognition to those who have answered a large number of questions well. This approach, as discussed by Deangelis [50] for Yahoo! Answers, while effective in retaining strongly engaged responders, fails to target new responders. In fact, it may even discourage new users, unwilling to make the time investment needed to

answer a large number of questions. Also, it does not take question quality into account, which could reward users who answer many simple questions well, instead of tackling the more difficult questions. As observed by Ghosh *et al.* [64] based on a game-theoretic analysis of behavior on user generated content forums, "without some connection between quality of a contribution and amount of exposure, such exposure motivated contributors will flood a site with low quality contributions".

This research proposes a complementary approach, where an attempt is made to identify users for recognition, in terms of quality of participating responses, instead of the quantity. So for example, we might be interested in providing recognition to users who are less prolific, as long as the questions they answer are difficult, and they answer them well. It is hypothesized that recognition will encourage a subset of them to become regular contributors. Another major advantage of this approach is that, identifying and encouraging high quality users enhances the reputation of the community, so that new users are more likely to trust answers they receive on the forum, and so start asking questions. Besides that, it cultivates a broader base of experts, lowering a forum's dependence on a small set of highly active users. To distinguish the problem of finding users who would provide a high quality response, from the broader issues involved in expert finding, this research refer to this second problem as *authority identification*, as the task being set is that of identifying cognitive authorities, individuals new users are willing to trust, even if these individuals are not highly active participants.

Note that authority is not the same as popularity. Thus, while a prolific expert who answers many questions is valuable, another expert who answers a few difficult questions is at least as valuable for establishing a website's reputation. High quality content is also valuable for forums due to its potential for monetization, both from an archival and search engine ranking perspective. Besides, by emphasizing quality alongside quantity, this approach encourages a broader expert base, reducing the likelihood of the community relying on a small subset of volunteers who may leave at any time, as is often the case.

Identifying authoritative responders is, however, not straightforward. One approach is to use a supervised learning approach, where features such as answer length, word size, etc., are used to estimate answer quality [5] . Such approaches are, however, highly vulnerable to manipulation. An alternate approach, investigated in the past [18, 151], and further developed in this work [32, 33], is to aggregate user endorsements. These endorsements are usually provided in Q&A forums to an expert when they answer a question. For example, the question may mark a response as the 'best answer', indicating that they prefer the response to the alternatives provided by other responders. Such endorsements can be aggregated simply by count in a naïve approach, but can be made more precise by taking into account the quality of the endorser, and the alternatives the endorser had [33], usually via graph-theoretic measures [32, 88, 151].

1.1.2 Professional Social Networks

Endorsement aggregation of a similar type is used by this research [34] for authority identification in professional social networks. In this case, however, endorsements may not have resulted from online activity, but from offline interactions. An endorsement may be straightforward, such as a positive recommendation, or may need to be inferred from actions such as an invitation to connect, or a series of initiated interactions. As this research shows [34], something as subtle as viewing a profile can be a powerful signal.

Returning to the definition of authority as "influence on one's thoughts that one would recognize as proper" [118], the question arises, influence *who* would recognize as proper. For example, should we focus solely on members active inside the network, or any interested individual who may be outside the social network? Another related question is, should this influence be measured in aggregate, so that a user who is less convincing in a single interaction might still be influential due to having more interactions, for example by being more active or being active for a longer period of time? Or should we attempt to estimate the outcome of a single interaction?

The problem of identifying influential users on an OSN, a well-studied problem in the social network research community [3, 41, 68, 92, 137, 138, 145] which often relies on endorsement aggregation [41, 138, 145], has traditionally not considered these distinctions. An influential user (or 'influencer') is usually defined as one who can induce other members to take certain actions, such as, take interest in some information they share, etc. Influence is, thus, a mea-

sure of the user's importance within an OSN. However, this measure does not discriminate between influence garnered across a large number of interactions with a low success rate, and persons with a high probability of influencing others in a single interaction, who interact less frequently. In other words, influence as historically measured, does not consider user *activity* levels.

In contrast, Patrick's definition given above considers as an authority a person likely to be influential in a one-to-one interaction, irrespective of whether they are active in the network or not. This is also the traditional view of authority, where an authoritative work or person is one that is convincing in its reasoning at an individual level. Identifying such individuals on an OSN or Q&A forum, referred to here as the *authority identification problem*, is the problem introduced by this research [34]. In other words, this research [32, 34] argues that a node's influence on an OSN as traditionally measured consists of two components, its authority strength, and its activity level, and develops algorithms that attempt to separate these components, for authority identification.

Influential users, as traditionally identified, are often not persuasive outside the network. Khrabov *et al.* [93] observed that many very influential users on the social network Twitter [82] are relatively, if not completely, unknown outside their online circles. On the other hand, an individual who is quite well-known in the real world may not be at all influential online. The reasons for this mismatch are largely related to user activity levels [34]. Maintaining an influential online presence can be extremely competitive, and many

authoritative people may not be willing to invest such a high degree of time and effort. Another reason could be that they represent a demographic group not yet engaged by the OSN, or the online world in general. Phenomena such as hyper-networking also allow low-authority users to garner disproportionate influence, by compensating for less authority with higher activity. In contrast, this research demonstrates that authoritative users who are more likely to influence someone in a single interaction, are much more likely to be well-known in the real world [34].

Authority identification is an important problem in OSNs for many reasons. Authoritative users have the potential, with the right incentives to be more active, of providing valuable content to a forum. Also, as noted by Paul *et al.* [119] for the Q&A forum Quora [123], having users that are well-known in the real world makes it more likely that users will trust the forum, boosting its reputation. Besides this, in many other applications, such as when an organization is looking to fill a job position[1], real world reputation is often a greater concern than influence. Even in the marketing domain, where marketers largely care about the online influence of users, Carl [39] has argued that an overwhelming majority of word-of-mouth marketing takes places offline, in which case real world reputation is a factor worth considering. Also, while the quest for influence can drive activity on an OSN in the short-term, it is also easy for an OSN to lose credibility if authoritative users get

[1]A number of websites, for example, LinkedIn [46], and Stack Exchange [80], combine professional social networking with recruitment solutions, which is a large fraction of their revenue.

'crowded out' [144] by users relying primarily on higher activity for influence. Thus, it is essential for an OSN to have an understanding of the level of engagement of authoritative users.

At a deeper level, this work argues that the separation of authority and influence provides important insights into the nature of influence on OSNs. Measuring user authority can enable a comparison of various influence measures, in terms of the extent to which they coincide with authority, or can explain the gap via activity. This is important because influence is usually defined operationally in terms of the algorithm used to measure it [41, 145], and it is often difficult to obtain external verification based on 'ground truth'. Authority estimation can be a useful sanity check in such a situation: a measure of influence that does not correlate at all with authority, and is not able to explain the gap, may well be measuring the wrong thing.

1.2 Problem Description

This dissertation investigates the problem of preference aggregation for authority identification in online communities: how individual expressions of preference can be aggregated to estimate user interests, and the level of authority in these topics of interest, in an online forum.

Online communities often provide way for users to endorse other users' activities. For example, on a Q&A forum, a questioner may be allowed to rate one of many answers to her question as the best answer; or a visitor to a photo or content sharing site may be able to 'up-vote' a photograph or article

she likes. It may also be reasonable to interpret certain actions as expressing a revealed preference. For example, sending another user an invitation to connect can often be seen as an endorsement: as observed by Leskovec *et al.* [105], invitations in OSNs often flow from lower status members to those of higher status.

The traditional way to represent user preference expressions on an OSN or Q&A forum is as a directed graph, where each user is represented as a vertex/node in the graph, and interactions are represented by edges. This research refers to such graphs as *endorsement graphs*, or *preference graphs*. The direction of the edge is often used to represent the asymmetric aspect of the interaction. For example, a directed edge from OSN member A to member B might represent that A endorsed B.

While preferences are expressed at an individual rater level, they are often aggregated across all raters to estimate the consensus about a rated user's importance or reputation in a network (these scores can then be sorted in descending order to identify the most authoritative users). The simplest aggregation measure is degree centrality [141], which is simply the in-degree of a node on the endorsement graph. It can be seen as the marginalized user preference, with the rater whose influence we are concerned about selected from a uniform distribution[2].

However, we may be able to infer additional information about each

[2]Or with a probability proportional to level of activity for each rating user, assuming multiple endorsements per rated user by a rater are allowed.

rater, concerning qualities a rating scheme should care about. A quality commonly taken into consideration is the level of expertise of the rater. Since the available information consists of mutual preference expressions, this is often done in Q&A forums using a recursive definition [87, 151], where the expertise of each user is defined as the average of the expertise of all other users who endorsed her.

In the graph-based representation described earlier, this definition translates to defining each user's expertise score as the stationary distribution of the Markov chain corresponding to the graph[3], or the fraction of time spent on the node during a random walk on the graph. This approach to identifying experts is inspired by the well-known PageRank [19, 30] algorithm for ranking pages based on authority, on the Web graph.

Closely related random walk based measures [25, 90] (pre-dating PageRank) have also traditionally been used for measuring the influence of nodes in online social networks. As mentioned earlier, measuring a node's influence and its authority are closely related problems. A node's influence attempts to measure how important a node is to a network, and how central it is to the interactions taking place on the network. The distinction between influence and authority is not very meaningful in real world social networks, for which these measures of influence [25, 90] were originally constructed. Given that most individuals are embedded in their real world community and cannot change

[3] After some adjustments [30] to ensure that the Markov chain corresponding to the graph is ergodic [91].

it at will, all members are strongly incentivized to attempt to maximize their influence in this community, and the more authoritative[4] are more likely to succeed.

In contrast, on an OSN, users are free to be inactive, or even leave the network at any time, with little or no negative consequences. The benefits of participation are often intangible, and may even be insufficient to motivate many authoritative users. Also, less authoritative users may have greater incentive to be active, as they may value online influence more highly. As a result, authority and influence are two quite different concepts for online communities. This research [32, 34] focuses on the problem of authority identification, distinguishing it from the problem of identifying influential nodes.

Another interesting aspect of OSNs and Q&A forums is the existence of multiple graphs over the same set of members, reflecting different aspects of user behavior. For example on an OSN, one graph may reflect invitations for connecting across the network, while another graph may reflect actions such as 'liking' someone else's content, etc. Historically research on OSNs has focused on ranking based on a single graph. This research extends these approaches via new algorithms that can combine information from multiple graphs to arrive at a single ranking [34], and finds that this improves the accuracy of ranking considerably.

[4]Defined, as mentioned previously, as members most likely to influence other members of a community, in a *one-to-one* interaction, if not at an aggregate level.

1.2.1 Preference Expressions in Q&A Forums

Q&A forums present some interesting challenges that are often not present on OSNs. For example, all actors in professional social networks are usually capable of the same set of actions. On the other hand, in the case of Q&A forums, preference revealing actions need to be considered from two perspectives: the questioner and the responder. Questioner activity largely consists of asking questions and rating experts based on the answers, and so is similar to the action set of users in professional OSNs. However the responder action of choosing a question to answer is also a preference revealing activity, among question topics. Aggregating this preference across experts can help group questions by topics: if multiple users have a history of selecting a common set of questions, it is likely that these questions are from the same topic of expertise. This research shows [35] that aggregating expert preferences among questions can considerably improve accuracy in expert finding.

Another interesting aspect of Q&A forums is the existence of *relative preference* graphs. A relative preference or pairwise preference is one expressed by a user for a particular user or item among multiple choices. For example, in a Q&A forum, a user A might pick an expert B's answer out of a set of responses by experts B,C,D, as the 'best answer'. This would be represented in the relative preference graph as directed edges from C and D to B, as B 'won' while competing against C and D. An *absolute preference*, on the other hand, is not in comparison to other items. So for example, a 'like' or 'up-vote' on a content-sharing forum is an absolute preference. Relative

preference graphs, while not restricted to them, are much more common on Q&A forums, and contain valuable information about the comparitive levels of expertise of competing experts. This is the first work to consider them for ranking users in Q&A forums [33].

1.3 Research Questions

This dissertation examines the following hypothesis:

> *In an online community of experts, mutual expressions of absolute and relative preference can be aggregated to yield effective estimates of an expert's topics of interest, and his/her credibility as an authority in these topics, both inside the community and in the real world.*

This hypothesis is evaluated by answering the following research questions.

1.3.1 Research Question 1: Responder Preference Aggregation for Question Recommendation in Q&A forums

> RQ1: How should information about experts' preferences among different questions, based on training data, be used to make more precise question recommendations to them in the future?

A motivating factor for this work is reducing the load on responders in Q&A forums through more precise recommendation of relevant questions.

This requires identifying for each responder, the set of topics of interest to them. The set of questions answered by a user in the past is an important signal of user interests. Ignoring a question, on the other hand, suggests a lack of interest in the question topic. So, for example, if users A, B and C always answer the same set of questions, this behavior is likely to have been motivated by interest in a common topic. How can this observation be used to recommend interesting questions to an expert?

The main challenge in being able to incorporate expert feedback is that, while an expert's availability and interests are explicit for questions that they answered, the same information is not available for questions that they did not answer. Thus, it is often not possible to know whether an expert did not answer a question due to lack of interest, or because she did not read it, or read it but was too busy at the time to answer. This research overcomes this problem via two generative model based algorithms [35], that assume different models of expert behavior for questions they did not answer. These models attempt to estimate the latent behavior of the user where such information is not available.

1.3.2 Research Question 2: Preference Aggregation for Authority Identification in Q&A Forums and Social Networks

RQ2: How should information about users' absolute and relative preference for other users be aggregated to identify authoritative users in an online forum or social network?

This research question is motivated by the need to encourage participation from authoritative users via recognition within a community. It is necessary to be able to first identify authoritative users to achieve this. It is also motivated by other applications, such as finding experts for recruitment in professional OSNs, and content recommendation in content-oriented OSNs.

A naïve measure of a user's authority is her popularity, that is, the number of endorsements received by her. More sophisticated approaches include weighing these endorsements based on information that can be derived about the endorser, for example by using graph-based approaches [30] that measure a user's authority recursively. This approach has been used for ranking responders in Q&A forums [88] in the past, and for finding influential nodes in OSNs [145]. Relative preference data provides even more information in Q&A forums, not only about the user who endorsed a responder, but also who they were competing against. This data contains intrinsic information about the authority strength of various experts, as more authoritative responders will do well even when competing against other strong experts. This work is the first to use relative preference data for ranking in Q&A forums [33].

The problem of identifying influential users in social networks [68] is closely tied to authority identification. A common assumption is that the two are equivalent: this is one reason why authority analysis algorithms for web pages such as PageRank [30] are often used for identifying influencers in OSNs [145]. However, in practice, influence can deviate from authority, which is related to the likelihood of influence propagating during a single interaction,

as opposed to aggregate following, which can be achieved via increased activity.

Based on this idea, this research [33, 34] approaches the authority identification problem by positing 'authority strength' as an intrinsic property of a node on a graph. In this view, the outcome of an interaction between two nodes depends on their intrinsic level of authority 'strength': the 'stronger' node is more likely to influence the 'weaker' node, than the other way around. A less authoritative user, might however, become more influential, if it chooses which nodes to interact with wisely, or chooses to have more interactions than others. This is formalized by considering each interaction on an OSN as a game, where the player with more authority 'wins'. The final structure of an OSN or Q&A preference graph can be viewed as the result of a tournament of multiple such games.

This research finds that this model of authority is a better predictor of who will give the best answer to a given question, compared to influence [33]. It also finds [34] that authority correlates better with real world recognition, as opposed to influence as traditionally measured.

1.3.3 Research Question 3: Preference Aggregation across Multiple Graphs for Authority Identification in Q&A Forums and Online Social Networks

RQ3: How can multiple signals of user preference in an online Q&A forum or social network be combined to yield an effective consensus ranking of members by authority?

Relying on a single behavior as a signal of authority is not effective in OSNs. This is because there are often multiple modes in which these endorsements are expressed: for example, sending a user an invitation, viewing their profile, etc., can all be considered as an endorsement. Often information contained in these graphs is complementary. For example, Weng *et al.* [145] found that on the OSN Twitter [82], famous people tend to have more 'followers', but 'retweet' graphs are often dominated by individuals who tend to contribute news. Similar complementarity is demonstrated for the LinkedIn [46] OSN in this work [34]. However, traditional approaches for ranking users in Q&A forums and OSNs have been restricted to single graphs.

This research question aims to extend graph-based ranking approaches for OSNs, enabling them to aggregate information from multiple endorsement graphs into a single ranking. The goal is to combine the complementary modes in which user preferences are expressed on an OSN.

Another scenario involving multiple graphs is, where social relationships reflected in one graph, impact the interactions in another graph. An example of this is *social voting* [65], where OSN users tend to vote based on social ties, instead of objective judgments of information quality. This problem is also addressed by this work.

1.4 Contributions Outline

The contributions of this research are addressed at two problems: a) expert finding in Q&A forums, and b) authority identification in online social

networks (OSNs).

1.4.1 Expert Finding

For the expert finding problem, this research contributes:

- A new set of evaluation metrics that formalize the trade-off between questioner and responding expert's satisfaction in Q&A forums [35].
- Two new generative model-based algorithms [35] for discovering topics of discussion in a forum, and each expert's interest in these topics. These models incorporate user choices between questions as part of the topic identification process. In contrast existing approaches are largely restricted to topic-word distributions.

1.4.2 Authority Identification

This research makes the following contributions to the authority identification problem:

- Two new algorithms for authority identification in OSNs [33, 34], that overcome the chief drawback of current PageRank based approaches, sensitivity to user activity, by modeling user interactions as the outcome of a tournament-based model.
- A co-ranking algorithm [34] for combining authority or influence related information from multiple graphs representing different aspects of inter-

actions on an OSN, in a principled way. Historically, ranking approaches in OSNs have been restricted to single graphs.

- A mixture-model based algorithm [32] for balancing a questioner's preferences between authoritative sources and socially proximate experts.

1.5 Dissertation Outline

This document is organized as follows: Chapter 2 provides the technical background to this research. It briefly discusses information retrieval metrics and techniques, which are used extensively as part of the expert finding approach, outlines graph-based models for identifying important nodes in a network, and also discusses the related work. Chapter 3 is focused on the problem of expert finding in Q&A forums. It contains a discussion of metrics best suited for expert finding in online communities, and develops two new generative model based approaches that address Research Question 1.

Chapter 4 addresses Research Questions 2 and 3. For Research Question 2, a conceptual model of user authority in online networks using the Bradley-Terry model is developed, and its relationship to the concept of influence is established via tournament models. Research Question 3 is also addressed in Chapter 4, via the development of a co-ranking approach for combining information from multiple graphs. Chapter 5 presents an empirical evaluation of the ideas developed in Chapters 3 and 4. The expert finding models are evaluated on data derived from the StackExchange online com-

munity [80]. The approaches developed to address Research Questions 2 and 3 are evaluated on the professional social network LinkedIn [46] and also on data from StackExchange, Yahoo! Answers [85] and Digg [17]. Chapter 6 restates the research questions, how they have been addressed, and the research contribution made.

Chapter 2

Background

The culture of information exchange on the Web differs from traditional media in emphasizing broad-based participation as opposed to selectivity. This is a natural byproduct of the nature of the Internet, which has succeeded by breaking down the traditional barriers to communication. The trend has been further strengthened by the advent of Web 2.0 [117](or the participatory Web [23]) that relies on user generated content and collaboration, based on the idea that everyone knows something [2]. As a result, the Web is able to provide access to an unprecedented amount of constantly growing information. In the words of David Shenk [129], "putting a computer in every classroom is like putting a power plant in every home".

This informational affluence does have its drawbacks. It is generally recognized [6, 9] that user generated content unmediated by editorial discretion shows a very broad distribution of quality. In the traditional media space, information seekers rely on filters such as editorial and peer review, and guidance from recognized 'cognitive authorities' [118] such as editors, reviewers, professors, librarians, for quality control. However a combination of factors, such as the proliferation of voices, the tendency towards anonymity or pseudo-

anonymity, and the scale and speed of interactions, renders the traditional approaches impractical for the Web.

In these circumstances, users have to develop their own strategies for finding reliable information. A common strategy among users, documented by Rieh et al. [125], is to only trust content found on 'known' sources, i.e., websites they are familiar with and developed trust for, or websites recommended by their peers. Thus, a collaborative knowledge-sharing site, to succeed needs to develop a reputation for being populated by a community of experts who are credible sources of information.

This work identifies two ways in which algorithmic approaches can help with achieving the goal of encouraging an active expert community: a) *Question Recommendation*: Lower the cost of expert participation (in time), by automatically recommending questions of interest to them, and b) *Authority Identification*: Identify experts so that they can be incentivized for participation via visibility and other incentives.

2.1 Question Recommendation

The problem of matching experts to problems has traditionally been studied in the information retrieval (IR) community, in the context of large organizations. A common use case is helping employees find experts to contact via email, to meet an information need. A prominent example of this is the expert search track [61, 135] at the Text Retrieval Conference (TREC) [1], focused on the problem of finding experts to email within an organization,

given descriptions of topic of expertise, and a supporting document collection. With the growing importance of the Web as an information source, the problem scope has broadened to include expert finding in online settings such as community question-answer (Q&A) forums, a problem first addressed by Liu *et al.* [109] .

Given a query for expertise, Fang and Zhai [58] divide expert finding approaches into two categories: profile-based approaches and document-based approaches. Profile-based approaches construct a profile for each user, based on the terms (or words) that have been associated with them in the past. Experts can then be matched in terms of the proximity of their profile with the query. The alternate document-based approach users a two-step process where first, the relevance of each available document to the query is estimated. Following this, experts can be ranked based on how closely involved they are with the most relevant documents. Thus, both approaches proceed by reducing at least part of the expert finding problem to the classical information retrieval (IR) [13] problem: given a query and a dataset of documents, rank the documents in order of relevance. In the profile-based approach, this is done by explicit 'document reorganization' [61], by consolidating information related to an expert into a single document. As a result, most standard retrieval methodologies, along with evaluation metrics, can be applied to the expert finding problem in Q&A forums. Three common IR models are vector space models [130], language models [99], and cluster-based models [108]. The next section discusses standard metrics used in IR, followed by a discussion of the

	Relevant ($R = 1$)	Not Relevant ($R = 0$)
Retrieved	R^+ (True Positive)	N^+ (False Positive)
Not Retrieved	R^- (False Negative)	N^- (True Negative)

Table 2.1: General Contingency Table for Retrieval

three common retrieval methods mentioned.

2.1.1 Evaluation Metrics

In the standard IR framework, the evaluation of a document retrieval system requires a test dataset consisting of three parts [112]: a) a collection of documents D, b) a set of queries Q, and c) a binary relevance judgement for each query-document pair $P(R = 1|Q, D)$. The system is presented with each query in turn, and retrieves a set of documents it believes to be relevant. The results of this process can be represented as a contingency table, shown in Table 2.1. Two standard measures derived from this table are *precision* and *recall*, defined as:

$$\text{Precision} = \frac{R^+}{R^+ + N^+}$$
$$\text{Recall} = \frac{R^+}{R^+ + R^-}$$

Precision measures the quality of retrieval, that is, of the documents retrieved, how many were relevant. Recall, on the other hand, measures the breadth of the retrieval system, in terms of how many of the relevant documents it was able to find.

2.1.1.1 Metrics for Ranked Lists

In many cases the retrieval system returns a ranked list of documents, in decreasing order of estimated relevance, instead of an unordered set of documents. Precision and recall may still be used, by selecting a subset of documents, say the top k documents, as the relevant set. The resulting metrics are often written as Precision@k and Recall@k respectively. The choice of k depends on the application. For example, users may be relatively unlikely to look beyond the first page for search engine results, in which case the value of k is likely to be quite small. On the other hand, a researcher might be willing to look through hundreds of documents to find relevant information, in which case the value of k may be much larger.

A reasonable alternative, then, is to average across a range of values of k. This is the intuition behind *average precision* [112]. According to Robertson's probability ranking principle [126], the approach of presenting results by decreasing estimated relevance is provably optimal on many evaluation metrics, including the expected average precision.

Formally, for a given query $q \in Q$ with n relevant results, let the retrieved results include relevant documents at ranks $r_1, r_2, \ldots r_n \in \mathcal{R}$, where $r_1 \leq r_2 \leq \ldots r_n$. Then the average precision is calculated as :

$$\text{AP}(q) = \frac{1}{n} \sum_{k=1}^{r_n} \text{Precision@}k$$

If the ranked list does not cover all the documents in the dataset, certain relevant documents may be missing. The precision score for such documents

is usually set to 0. The *mean average precision* (MAP) metric is simply the mean of the average precision across a set of queries Q, given by:

$$\text{MAP}(Q) = \frac{1}{|Q|} \cdot \sum_{q \in Q} \text{AP}(q)$$

In the case of only a single relevant result (for example, if we are interested only in the rank of the 'best answer' to a question), the MAP measure becomes equivalent to the *mean reciprocal rank*, defined as:

$$\text{MRR}(Q) = \frac{1}{|Q|} \cdot \sum_{q \in Q} \frac{1}{r_1}$$

Remember that $r_1 \in \mathcal{R}$ is the first rank with a match.

Another way to mitigate the impact of an arbitrary cut-off is to analyze the system behavior at different rank levels. This is often done in IR via a precision-recall graph, where the precision is usually represented on the x-axis, and the recall on the y-axis. Generally, as the rank cut-off increases, the precision falls, while the recall increases. A curve on the graph shows the relationship. The closer the curve is to the top-right corner of the graph, the better the system is doing. This can be expressed as a single value as the *area under the curve* (AUC) of the graph. Since MAP computes the precision at each matching rank till the last match, it can be seen as an approximation of the AUC.

2.1.2 Vector Space Models

The vector space model [130] represents each document or query as a vector in an n-dimensional vector space, where n is the cardinality of the set

of all unique terms (or words) in the dataset. In other words, each term is assigned a dimension in one-to-one correspondence to it. Loosely, we can say that the terms are arbitrarily numbered 1 to n, and the k^{th} term corresponds to the k^{th} dimension in the vector space. The magnitude of a document vector d in a dimension k is usually a function of the k^{th} terms occurrence count (also called the *term frequency*). A popular and simple heuristic function is the *term frequency inverse document frequency* (TF-IDF) function [13], which is given by the product of the term frequency with the inverse document frequency (IDF), defined for the k^{th} term as:

$$\text{IDF}(k) = \log \frac{|D|}{|d \in D : k \in D|}$$

Here $|D|$ is the number of documents in the dataset D, and the denominator counts the number of documents where the k^{th} term occurs at least once. The core idea behind the IDF score is to lower the importance of terms that are common across documents in the dataset, and emphasize rare words. There are a number of other more sophisticated approaches for re-weighing the term documents based on global statistics [128].

Given two re-weighted document vectors d_1 and d_2, the similarity $\text{sim}(d_1, d_2)$ between them is calculated as the normalized dot product or cosine score between them:

$$\text{sim}(d_1, d_2) = \frac{d_1.d_2}{\|d_1\|\|d_2\|}$$

Here $\|d_1\|$ and $\|d_2\|$ represent the Euclidean norm of d_1 and d_2 respectively. As

it was found that the Euclidean norm over-penalize longer documents, Singhal *et al.* [131] have suggested an alternate approach.

The vector space model is a powerful model, and has been used in the past for expert finding in Q&A forums [67, 150]. However, it does not have a straightforward probabilistic interpretation. As a result it is difficult to extend to more complex scenarios and needs. The language model, discussed next, overcomes this drawback.

2.1.3 Language Models and Pseudo-Relevance Feedback

Given a vocabulary of terms V, a language model [99, 120][147] is formally a probability distribution over the set of all possible strings V^* of any length that can be generated by concatenating terms from V. The language modeling approach proceeds by assuming that the string of terms in a document or query was generated by an underlying probabilistic process. Based on this assumption, if the underlying generative process for a string a is known, its proximity to another string b can be estimated by the likelihood that the process that generated a could also generate b[1]. A key advantage of language models compared to the vector space model is that they can be extended via more complex generative processes, or by combining them with other probabilistic models.

A common assumption for the string generation process is that each

[1]An alternate approach is to use a statistical measure of distance, such as the Kullback-Leibler (KL) divergence, between the query and document models [149].

term is independent of the previous or subsequent terms seen. This assumption, often called the bag-of-words assumption, essentially implies that a multinomial distribution process generates each string, as the probability of any sequence of terms now depends only on the aggregate count of the number of time each term occurs, and not on the specific ordering of the terms. Another common simplification is to ignore the impact of a string's length on a language model's likelihood, given the string. This is because in practice, we are usually comparing the likelihood across different language model distributions given a string of fixed length. For this particular task, the string length distribution can be assumed to be independent of the language model, and can be ignored as constant across the different models.

Two popular language models that make different assumptions about the string generation process, are the *query likelihood model* [120] and the *document likelihood model* [99]. The query likelihood model assumes there is a language model underlying each document, and estimates the match between a document and a query based on the probability of the query being generated from the document. The document likelihood model, on the other hand, tries to estimate the language model underlying the query, and calculates the probability of each document being generated from the query's language model. Assuming the underlying distribution is multinomial, the parameters of the language model for a string are usually straightforward to calculate using a maximum likelihood estimate [63]. This estimate usually needs to be smoothed [148] to correct for data sparseness problems.

Compared to query likelihood models, document likelihood models face a steeper challenge, as language models need to be estimated for queries, which are usually much shorter than documents. This means the language model constructed for a query is more likely to be inaccurate. One way of addressing this problem, known as relevance feedback [146], is to refine the query using explicit user feedback. This can be done by providing the user with an initial set of results based on the query, allowing her to mark them as relevant or irrelevant. Terms from the relevant documents can then be used to expand the initial query. In the absence of user feedback, an alternate approach known as pseudo-relevance feedback (PRF) [99], is to make the assumption that a small subset of the highest ranked documents retrieved based on the original query are relevant, and construct the query language model using this sets of documents.

Mathematically, the steps are given next, loosely based on the description by Lease [100]. As with the description of the cosine model, let the words/terms be numbered arbitrarily from 1 to n, where n is the vocabulary size. Let the query be given by Q, so that f_i^Q is the frequency of the word marked i, in the query, and let $\boldsymbol{\theta^d} = (\theta_1^d, \theta_2^d, \ldots, \theta_n^d)$ be the multinomial (unigram) word distribution vector corresponding to the document d in the dataset $\mathcal{D}$. Then the documents are first ranked by the probability that the query was generated by the document, written as:

$$\log P(Q|d) = \sum_{i=1}^{n} f_i^Q \log \theta_i^d = (\boldsymbol{f^Q})^\top \cdot \boldsymbol{\theta^d} \tag{2.1}$$

That is, the log likelihood of the query being generated by the document is given by the dot product of the log likelihood of distribution representing the document, and the query. Each document d is often smoothed using word statistics from the entire collection C, to deal with possible sparsity problems. Thus eq. (2.1) uses instead of θ^d, $\hat{\theta^d}$, given by [100]:

$$\hat{\theta_i^d} = \lambda \frac{f_i^d}{|d|} + (1-\lambda)\frac{f_i^C}{|C|}$$

where $\lambda = \frac{|d|}{|d|+\mu}$, μ being a previously chosen hyperparameter. Here $|C|$ and $|d|$ are the number of word occurrences in C and d respectively.

Now, given the log likelihood for the query given each document, a cutoff $|R|$ is used so that only the top $|R|$ feedback documents R are used for pseudo-relevance feedback. This is done to avoid query drift, where words irrelevant to the query may become prominent in the feedback set. The feedback query $\boldsymbol{\theta^R}$ is then constructed as $\sum_{r \in R} w_r \theta^r$. The weight w_r is set to the likelihood of the query given θ^r (exponential of eq.(2.1)), normalized so that all w_r sum to 1. Following this, another cutoff κ_P is used, on the number of words in the pseudo-relevance feedback query. The words in $\boldsymbol{\theta^R}$ are ranked by weight, all but the top κ_R words are discarded, and the query is re-normalized. After this, the documents can be ranked using the feedback query, based on

equation (2.1) (the weights f_i^Q are now less than 1, as the query is normalized). This gives us the set of relevant documents $\mathcal{R}$.

We still need to translate document relevance estimates to an estimated relevance for each user. This research ranks users via their marginalized distribution across all relevant documents, using the following equation.

$$P(u|q) = \sum_{d \in \mathcal{R}} P(u|d)P(d|q)$$

Again, this research does not use all document, but a cut-off $N_{\mathcal{R}}$, so that only the top $N_{\mathcal{R}}$ are used, weighed by their normalized likelihood.

Pseudo-relevance feedback is empirically known to be a successful approach in document retrieval [147]. However, with the exception of the relatively early work of Liu *et al.* [109] on the Q&A site Wondir.com [15], the approach has not been investigated for Q&A recommendation.

2.1.4 Cluster-based Models

Cluster-based models [108] group similar documents together, based on the assumption that similar documents will be required for similar tasks. A representative model for the topic, such as a mean cluster vector or centroid, or a probabilistic distribution over words, is often created as part of the clustering process, and can be used to match to the query. There is a vast amount of literature on document clustering (surveyed in [4]), with different algorithms resulting from differing underlying assumptions. For example, a vector space representation of documents in combination with the similarity

measure is generally used for agglomerative algorithms, or partitioning algorithms such as k-means. A common probabilistic approach mathematically related to the k-means algorithm, is to assume that the dataset was generated by a mixture of multinomial topic models, which can be discovered using the expectation-maximization (EM) algorithm [22, 51]. More complex models make assumptions such as assuming a topic for each word in a document, as opposed to a single topic for each document [24], or including author information as part of the generative process [127].

Yi and Allan [147] consider two possible approaches for using clustering information during document recommendation: a) using topics for smoothing documents before matching with queries, and b) using topic models for pseudo-relevance feedback. However, for the problem of finding experts, Fang *et al.* [58] propose removing documents from consideration once topics have been identified. Thus, for a query q, the probability that an expert u is a good match can be calculated as:

$$P(u|q) = \sum_{t_i \in T} P(u|t_i)P(t_i|q)$$

Here T is the set of topics, $P(t_i|q)$ is the estimated probability that t_i is the underlying topic for the query, and $P(u|t_i)$ is the probability that expert u is the best match given topic t_i.

2.2 Authority Identification

Historically, the presence of editorial input and other filters mitigated the problem of low quality content in the information exchange landscape. The absence of these filters on the Web has broadened the scope of the retrieval problem, by emphasizing content quality or authority as an additional consideration. Given the impossibility of scaling human input to Web scale, algorithmic approaches to authority estimation on the Web were developed. These techniques usually relied on the analysis of the hyperlink structure of the Web.

2.2.1 Graph-based Authority Models for Web Pages

Authority in popular link analysis algorithms such as PageRank [30] and HITS [94] is conferred on a web page, in one way or the other, by other pages via links or references. Thus, for example, the PageRank algorithm is often interpreted via the random surfer model [42], so that a web pages authority score can be seen as the probability a web page surfer would visit the page given that she starts at a random page, and selects a random outgoing link from the page at each time step. So, it is a random walk on the Web graph, so that pages that are linked to by many others, or by a few important pages, are ranked highly. Similarly, in the HITS algorithm [94], a webpages authority score is equal to the sum of the hub scores of the pages that link to it, a hub being recursively defined as pages that tend to link to authorities. The assumption behind these approaches is that web page creators link to pages

they believe to be of high quality, and creators of higher quality web pages will link to proportionately higher quality ones.

In other words, a web page's authority is the result of a weighted voting scheme, and is the consensus arrived at with respect to the quality of a page by the participants in the system. It has not been independently verified by someone, say an expert, as corresponding with external reality. For this reason, these approaches are vulnerable to manipulation, for example to 'link farms' [40], and 'sybil' strategies [44]. However, contingent trust can be placed in the results, if the basic assumptions of the algorithms are believed to be correct, and also if the results can be verified empirically by being useful to web searchers. Section 4.2.1 provides a more detailed discussion of the PageRank model, and other similar measures.

2.2.2 Influence Analysis in Social Networks

Historically, graph centrality measures such as eigenvector centrality [25] and the Katz measure [90], have been used for identifying important nodes in social networks. These methods are closely related to the PageRank model [60]. With the popularity of PageRank algorithm, it has generally been adopted for identifying influencers in online networks [41, 145].

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 users importance within an OSN, but may not be relevant to the real world. For example, Weng *et al.* [145], for

identifying influential users on Twitter [82], measure algorithm effectiveness using a measure based on the number of followers the identified users have. Ghosh and Lerman [65] measure user influence on news sharing site Digg [17] by the number of votes the stories posted by them get.

It is common for social media influencers to not be authoritative in the real world [93]. There are many reasons for this. A primary reason is the large investment of time required to maintain an influential presence on a forum. Also, given the large number of social networks, it is possible that an authoritative user might invest her time on a different forum. In fact, a good measure of a social forum's relevance might be the number of influential users who are authoritative. Moreover, there are many social norms which motivated users can use to increase their influence. Some such norms are:

1. Reciprocity: A common social norm on many forums is for users to provide a reciprocal link in response to a link. So, for example, if a user follows another users, or likes some content by her, the latter user may reciprocate by following or liking in return. This norm can be seen as a form of courtesy, but is easily exploited by some users to increase their link count. Reciprocity of links is a well-documented phenomenon on the websites Flickr [101] and Twitter [62]. This norm can be exploited via hyper-networking, where a user might follow or like content from random users, with the expectation that sufficient number of them will reciprocate to improve her influence scores.

2. Social Voting: Many content-sharing sites such as Digg and Yahoo! Answers allow users to add other users as contacts or friends. The aim is to increase engagement: the site is designed so that users find it easy to get updates on the activities of their contacts. As a side-effect, most users find interesting stories via their contacts, with the result that users with many contacts find it easier to promote their content. Social voting has been documented on the website Digg [65] as well as the photo-sharing website Flickr [103].

With some exceptions [62], influence estimation algorithms have not attempted to correct for these distortions. The main reason for this is that influence as a concept is strongly tied to activity, and in practice, it is difficult to make the distinction between social activity and abusive behavior. This is one reason why this work proposes a separate problem, that of authority identification, which separates the concepts of activity and influence. Loosely speaking, this work defines authority as the rate at which influence grows with activity.

2.2.3 Tournaments and Voting Models

In sports tournaments, contenders play against each other, and the results of these games need to be aggregated to a single ranking [114]. The same problem also exists in voting systems [106], where individual pairwise preferences among options may need to be aggregated into a single decision. It is easy to see that authority estimation is a preference aggregation problem:

users express preferences for each other, and these preferences need to be aggregated into a single ranking.

Expressed preferences may be categorized as relative or absolute. A relative or pairwise preference is one that is made in comparison to another user. For example in a Q&A forum, provided a choice between answers to a question by k different users, a user might choose one users answer as the best answer. This can be interpreted as expressing a relative preference for the user giving the best answer, compared to the other users. An absolute preference, on the other hand, does not include an implicit comparison to other users.

Preference aggregation methods can broadly be defined into two categories, referred to here as: a) parametric methods, and b) graph-based methods. Parametric methods [140] assume that there is an latent random variable associated with each option, signifying its quality. When a preference is expressed by a rater between two options, it is drawn from a distribution that depends on the quality distributions of both options. A popular parametric model is the Bradley-Terry model [26–28], which assumes that, for a comparison between two options i and j , two samples w_i and w_j are obtained from their quality random variable distributions, and then:

$$P(i \text{ preferred }) = P(w_i > w_j) = \frac{w_i}{w_i + w_j}$$

Then by writing w_i as $w_i = e^{(i/s)}$, where s is a scaling factor, we can

write:

$$P(w_i > w_j) = \frac{1}{1 + e^{\frac{\mu_i - \mu_j}{s}}}$$

That is, the probability of i being preferred can be modeled as a logistic function [22]. The parameters $_i$ can be estimated using iterative algorithms related to the EM algorithm [77], or via convex optimization techniques [139].

Graph-based methods for tournaments, on the other hand, result in models quite similar in spirit to the link analysis approaches used on the Web. Thus, for example a round-robin tournament is represented as a graph with each player as a node. If player i wins against player i, this is represented as an edge directed from j to i. In matrix form, this can be represented as a tournament matrix M, where M_{ij} represents the number of times user i lost to user j . Thus, the number of games users i and j have played against each other is given by $M_{ij} + M_{ji}$. In case of tournaments for which $M_{ij} + M_{ji} = 1$, referred to as generalized tournaments, a common recursive measure of a user's quality or ability is the power rank method [106], which is a close variant of the PageRank algorithm. A related measure is the fair bets score [48], which penalizes losses, unlike the PageRank score. Related random walk models have also been proposed by Dwork *et al.* [54]. Usually, a particular measure is chosen based on its suitability to the problem, based on an axiomatic analysis of different measures' properties [29, 133], or empirically [54].

Graph-based approaches are better suited for the Web, because they tend to scale more easily. This is because the computations can often be

reduced to eigenvalue calculations using the power method [16]. However, parametric methods have the advantage of being interpretable in terms of latent quality characteristics, while graph-based approaches can usually only be studied via analytically characterizing their properties. Daniels [48] established a relationship between Bradley-Terry models and ranking measures for round-robin tournaments, provided the node strength is not assumed to be parametrized by a particular distribution. This approach has traditionally been ignored in favor of parametric Bradley-Terry models. However, given the size of tournament graphs on OSNs, this research uses the result proved by Daniels [48], besides extending it to more general cases that reflect online user behavior, than round-robin tournaments. A similar approach was recently proposed by Oh *et al.* [116], where they used a graph-based approach to discover the underlying parameters of data generated by a Bradley-Terry process.

2.3 Research Contributions

The problem of expert finding in Q&A communities was introduced by Liu *et al.* [109]. They presented experiments using cluster-based and pseudo-relevance feedback (PRF) based approaches on the Wondir Q&A forum[2] [15]. In their experiments, cluster-based approaches outperformed PRF by a small margin. Since then, research on Q&A forums can be divided into two complementary categories: *expert finding*, usually via text analysis approaches for matching topics with responders, and *expert ranking* for ranking responders

[2]Wondir.com was one of the first community Q&A websites. It closed down in 2009.

by expertise using graph-analysis based approaches.

2.3.1 Expert Finding

For expert finding, Guo *et al.* [70] explored topic-based generative models for finding best responders. Liu *et al.* [107] investigated LDA-based [24] techniques for the task of topic identification in Yahoo! Answers [85]. Similar approaches have been explored by Riahi *et al.* [124]. Qu *et al.* [122] discuss the problem of identifying the best responder, after assuming that the set of responders is already known. Most of these methods rely on the textual content of interactions to identify clusters of similar Q&A interactions, presumably belonging to a single topic, and use this information to construct topic-user interest distributions.

This research takes an alternate approach, by augmenting textual information with patterns of common interests among experts for topic identification [35]. For example, if a subset of available experts have answered (loosely) the same set of questions, while ignoring others, this is a strong signal that the questions belong to the same topic, which is the motivation behind the common interest among the experts. This insight is incorporated via a generative model [63] based approach. This signal has generally been ignored in current research, with the exception of Guo *et al.* [71], whose approach is similar to one of the two generative models, the *pure multinomial* model, proposed in this research [35], and was proposed around the same time. The other model proposed in this work, the *extended generative* model [31], outperforms this

older approach.

The other contribution of this work to text-based expert finding is the introduction of formal metrics that measure expert satisfaction in a Q&A forum. Evaluation of expert finding has historically focused on the questioner's satisfaction, ignoring the experts' viewpoint. More recent research has tried to incorporate the expert's satisfaction as well. Dror *et al.* [53] investigate question recommendation to experts, as opposed to the traditional 'expert recommendation to questioners' perspective, as a supervised learning problem. Horowitz *et al.* [75], in describing the design of the social Q&A engine Aardwark, emphasize the importance of maintaining a high level of satisfaction within the expert community, as the community grows via peer invitations. The engine, when recommending questions to responders takes into account responder availability and the evenness with which question load is spread across the responder base, besides the topic match. However, their work did not propose any formal metrics that can be used to compare different algorithms that attempt to balance questioner and expert satisfaction. This research [35] investigates a large set of metrics for their suitability for representing both the questioner's and responder's experience. Based on this analysis, it identifies the key trade-off in Q&A recommendation, between two metrics, *questioner coverage*, the fraction of questions for which the most suitable expert is found, and *responder load*, the ratio of irrelevant to relevant questions recommended to an expert.

Another characteristic of Aardvark is the use of 'connectedness' along

with expertise while ranking responders by relevance. Connectedness is measured by the system using a function that takes into account questioner and responder similarity, as well as their social proximity. The final score of a responder for ranking is the product of these values, so that only experts who score highly on both counts are recommended.

This research [32] explores a different approach to the incorporation of social ties, by estimating for each user their preference between authoritative responders and personal connections as a probability (Section 4.6). This approach is chosen for two reasons:

1. A large number of questioners on a Q&A forum are not regular visitors to the site, so there is usually insufficient information for personalization. This approach handles such situations gracefully, by setting the preference for personalization to 0.

2. There is a common tendency, noted by Welser *et al.* [144], for expertise oriented forums to become dominated by conversations. Often this phenomenon 'crowds out' experts, who leave. The goal of estimating the preference of each user between authority and social affinity, is to allow each user to select (indirectly via behavior), the level of trade-off between the two preferences that they are most comfortable with.

This approach is experimentally evaluated only on the Yahoo! Answers dataset (Section 5.2.3), as social affinity is not believed to be a driver of interaction on the Stack Exchange dataset.

2.3.2 Expert/Authority Ranking

Complementary to text-based approaches, expert ranking models analyze the user interaction graph to identify experts. It is usually assumed that experts' interests are either described by them, or can be discovered as part of expert finding. Usually graphs constructed from questioner-responder interactions (absolute preference) are used, with the assumption that experts will ask harder questions compared to non-experts: relative preference information has generally been ignored. A majority of these approaches use a variant of PageRank [30] or the HITS algorithm [94]. HITS has been explored by Jurczyk and Agichtein [88], while Jhang *et al.* [151] explored both PageRank and HITS for Q&A forums. Bian *et al.* [18] proposed a coupled scheme, where answer quality and user reputation were alternately estimated. However, most of these approaches do not differentiate between highly active responders and experts who answer a few questions well.

Graph-based ranking algorithms are also used for identifying influential nodes in social networks [41, 95, 145]. These algorithms, however, do not take into account how user activity levels impact their influence. As a result, it is common to find non-authoritative but highly active users as highly influential on an OSN [93]. These algorithms are also vulnerable to manipulation using social norms such as reciprocity, where users play a 'number game' [145], endorsing a large number of fellow users, with the expectation that a subset of them will feel obligated to return the endorsement. Gayo-avello [62] attempts to correct for this distortion by heuristically reducing the weight of reciprocal

links. Weng *et al.* [145] attempt to differentiate users on Twitter who 'follow' (endorse) another user based on topical interests from those who 'follow' with the expectation of reciprocation, by constructing topic-specific 'follow' graphs, based on the hypothesis that reciprocation in users with common topical interests are more likely to be driven by common interests, or homophily [113]. Alternate measures of influence such as the 'follower' to 'following' ratio have also been proposed [10]. However, most of these approaches rely on heuristic techniques, and produce different ranked results. As a result, it is difficult to justify preferring one to another.

This research addresses the problem of differentiating between influence derived from activity and social norm manipulation from 'legitimate' influence, by introducing the concept of user authority. A user's authority is her intrinsic ability to influence another user in a one-to-one interaction. This is formalized by viewing each user interaction as a game, 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 [49], discussed in Section 2.2.3, is assumed to model the dynamics of a game between two players. The endorsement 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.

Based on the assumption that Q&A and OSN results are the product of an underlying Bradley-Terry model, a new tournament scoring model, called the *average winnings* model, is proposed by this research. The fair bets

measure [48], a tournament measure traditionally used for ranking players in round-robin tournaments, is shown to be a special case of the average winnings model. Based on their characteristics, the average winning model is found to be better suited to finding experts in Q&A forums, while the fair bets model is found to be a good fit for authority-based ranking in OSNs. These models naturally deal with the impact of activity. As they attempt to estimate an intrinsic value, they are less affected by user activity (though the degree of effect depends on the fidelity of underlying dynamics to the Bradley-Terry model).

While this is the first work to model OSN graph structure as a combination of node authority and activity, the idea that a graph node might have an intrinsic strength or quality, that may not be reflected in centrality measures due to 'first mover' advantage, has been explored for degree centrality by Bianconi *et al.* [20, 21]. Another related work is the unbiased web ranking approach by Cho *et al.* [45], that takes multiple snapshots of the Web over time, to take into account the rate at which a node's PageRank score grows.

Another challenge for link-based approaches in OSNs is the separate modes in which influence may be expressed. Cha *et al.* [41] identify three different modes of influence on Twitter [82]: 'indegree' (follow) influence, 'retweet' influence, and 'mention' influence, each of which can yield its own interaction graph. However, research on ranking nodes in OSNs has traditionally focused on a single graph drawn from the network. Also, no principled way to combine information from multiple such complementary graphs exists. Zhou *et al.* [152] have proposed a coupled random walk approach for combining

multiple graphs for the problem of combining citation networks of technical papers with the author social network. However, their approach is restricted to the PageRank model, and is not extensible to other tournament models proposed in this research.

This research demonstrates that combining information from multiple graphs is a powerful technique for authority identification in OSNs. To do this, it presents a new co-ranking framework for combining information from multiple graphs, based on a mutually positive reinforcement principle [152], where authority information from one graph is used to inform the authority measurement process in the other graph, and vice versa, till convergence.

However, such a process will require multiple estimations of authority for each graph, and will not scale to the large graphs found in real world OSNs. To overcome this drawback, a composite graph equivalent of the co-ranking framework is developed, by showing that the original co-ranking process is equivalent to a single authority calculation process on a specially constructed single graph. This model has the added advantage of being easily extensible to the authority models, average winnings and fair bets, described in this research.

Chapter 3

Expert Finding in Q&A Communities

This chapter presents two new generative model based algorithms, the pure multinomial model (Section 3.3.1), and the extended generative model (Section 3.3.3), for recommending experts for a question, given the question text. These algorithms augment traditional text-based retrieval models with expert preference information among questions. This chapter also presents two new measures, question coverage and responder load, for measuring the performance of expert recommendation algorithms in such scenarios (Section 3.2.2). Empirical evaluation of the algorithms is presented in Section 5.1. Algorithm performance is evaluated on a variety of metrices including question coverage and responder load, treating the pseudo-relevance feedback retrieval method [99, 100] (discussed in Section 2.1.3) as the baseline. Experimental results show statistically significant improvement on question coverage, holding responder load roughly constant, and improvement on other important metrics as well.

3.1 Introduction

A Q&A forum consists of participants in two chief roles: questioners seeking information via questions, and responders who answer these questions. The two roles are not mutually exclusive: a user may play either of the roles depending on the situation. The interests of the questioner and responder are, however, different. A questioner would like to find an answer to her question as quickly as possible. A responder would like to quickly find interesting questions, and would not want to search through too many irrelevant questions.

The expert finding problem has historically focused on metrics that measure questioner satisfaction [109]. This is a reasonable choice for the enterprise search problem, where it is more important that expertise seekers be able to find the relevant expert, and the query load is likely to be low. Also, the expert is most likely compensated for her efforts. On the other hand, the success of a voluntary online community depends largely on creating and maintaining a community of high quality experts that commit their time and efforts.

This is not only because strong experts are more likely to provide satisfactory answers, but also due to the monetary value of high quality content from an archival and search engine ranking perspective. As noted recently by Anderson *et al.* [8], "While most Q&A sites were initially aimed at providing useful answers to the question asker, there has been a marked shift towards question answering as a community-driven knowledge creation process whose end product can be of enduring value to a broad audience. As part of this

shift, specific expertise and deep knowledge of the subject at hand have become increasingly important."

As a result of this shift, there has been a recent perspective change in how the expert finding problem is studied: from a focus on questioner satisfaction, to also taking into account the responder's viewpoint, and considering the trade-offs involved where their interests diverge [35, 53, 75]. This research contributes [35] an evaluation of various information retrieval (IR) metrics that represent questioner and responder interests. In the end, it selects a subset of these metrics as suitable for comparing expert finding algorithms in subsequent work. This evaluation is presented in the next section.

3.2 Evaluation Metrics

The expert finding task in the Q&A scenario can be described as follows: each time a new question is introduced in the system by a questioner, the expert finding system contacts a subset of available responders in the system and recommends the question to them as of interest. Some (or none) of these may answer the question. One of the responses may then be selected as the 'best answer' by the questioner, indicating it most fit her needs[1]. This can be mapped to the standard IR retrieval scenario (Section (2.1.1)), consisting of: a document dataset, a set of queries, and a binary relevance judgement for each query-document pair. However, while the mapping of documents (each

[1]Certain Q&A forums allow votes from forum members other than the questioner on answers as well. However, this work limits its scope to the questioner's preferences.

Q&A interaction) and queries (questions) is obvious, the relevance judgements are actually made by two actors: questioners and responders. Based on this observation, this research [35] defines the standard retrieval metrics, precision and recall (Section (2.1.1)), from two perspectives.

Loosely speaking, from a questioner's perspective, her recall is the fraction of questions asked for which she received a satisfactory answer, while her precision is the fraction of answers that were relevant to her query. From a responder's viewpoint, her precision is the fraction of questions recommended that were recommended to her, while her recall is the fraction of questions she would have been interested in, that she was actually recommended. This is formalized in the next section.

3.2.1 Question and Responder Precision/Recall

Let X represent all the participants (questioners and responders) in the Q&A forum. Let a user x's responder precision be written as π_x^a, and responder recall as ρ_x^a. Then,

$$\pi_x^a = \frac{R_x^+}{R_x^+ + N_x^+}$$

$$\rho_x^a = \frac{R_x^+}{R_x^+ + R_x^-}$$

The right-hand side terms are as defined in Table 3.1. Table 3.1 can be understood as follows: R_x implies that the responder x believed the question

was relevant. This can be assumed to be the case if she responded to the question. N_x means that the responded x did not think the question was relevant (did not respond). A superscript of $+$ means the recommender system suggested the question to the responder x. A superscript of $-$ implies that the question was not suggested to the responder.

Thus, for a given responder, responder precision is the ratio of the number of questions that were recommended to a responder that the responder answered, to the total number of questions recommended to the responder. The recall for a responder measures how many of the questions the responder answered were recommended by the system. It is a measure of how well the recommender covers all the interests of the questioner.

Now, let the user x's questioner precision be written as π_x^q, and questioner recall as ρ_x^q. Then,

$$\pi_x^q = \frac{U_x^+}{U_x^+ + I_x^+}$$

$$\rho_x^q = \frac{U_x^+}{U_x^+ + U_x^-}$$

Here the right-hand-terms are as defined in Table 3.2. Questioner precision measures how many of the responders recommended by the system provided satisfactory answers. Questioner recall measures how many of the answers of interest to the questioner, the recommender was able to identify in advance.

There are three possible ways to count the number of correct matches (U_x^+) when calculating questioner recall (and precision). In decreasing order of strictness they are:

1. Count only 'best answer' matches: The recommender is assumed to have failed from the questioner's perspective, unless it retrieves the responder who provided the best answer to the question. In this case, the questioner recall is the fraction of questions asked by her, where the best responder was retrieved.

2. Count all matches: Precision and recall are calculated across all responders who answered the question.

3. 'Weak' Recall: The recommender is assumed to have succeeded if it retrieves at least one of the responders. In this case, each question for a questioner contributes as a single increment to U_x^+ or U_x^-.

Among these choices, options 2 and 3 are generally not good choices due to two reasons. Firstly, questions often receive multiple answers, only a subset of which are of good quality. Focusing on overall precision/recall or on 'weak' recall may cause the system to focus on users that consistently provide answers, but poor ones, or in the worst case, even spammers. Option 2 is generally less susceptible to this than option 3. However, a drawback of option 2 is that there is a lot of churn on most Q&A forums, with users often answering a few questions and then leaving. A fair subset of answers on questions are from such

users. Retrieval algorithms generally cannot be expected to make effective recommendations for such users, and it is not clear how much questioners stand to gain from these responses (unless there is specific feedback from the questioner).

For these reasons, this research focuses on option 1 (counting only best answer matches). An exception to this is conversationally oriented forums (such as forums discussing sports, relationships, etc.), where the idea of a 'best answer' is less meaningful, and users are much more interested in communicating with each other, than exchanging information.

3.2.2 Question Coverage and Responder Load

Despite being able to define four separate metrics, the action governing all of them is the same: each time a new question is introduced in the system, the expert finding contacts a subset of possible responders in the system and suggests the question to them. The decision to answer the question is made by each individual responder and cannot be controlled by the system. The decision made by the system to recommend a question has to take into account both the possible impact on questioner metrics as well as responder metrics.

While the importance of each metrics depends on the intentions of the designer, some judgements can be made about their relative significance. For example, responder precision is clearly an important measure of the quality of the recommender, as a system that recommends too many irrelevant questions to a responder will drive them away. Responder recall is important as well, as

otherwise responders may not see most questions of interest to them. However, in balance, responder precision is probably more important than responder recall.

Questioner precision, on the other hand, is not quite as important. A questioner may not usually mind extra answers so long as she gets the answer she is looking for. There may be problems in extreme cases, such as when a particular user is spammed, but this problem might be handled in other ways, such as by allowing questioners to ban specific responders from their questions, or by ranking answers based on responder quality or history. Questioner recall, however, is important: a questioner is primarily interested in finding the answer to their question, and a low questioner recall means that she did not receive a good answer.

Based on these intuitions, this work focuses on the following two metrics: questioner recall and responder precision [35]. Since we are only interested in best answers from the questioner's perspective, the questioner recall essentially measures the fraction of questions for which the recommender retrieved the responder who gave the best answer. Questioner recall can be seen as a measure of how satisfied questioners will be with the Q&A system if the responders relied entirely on the recommender to provide them with interesting questions. One way to think of this is as the *coverage* over all questions that the system is able to provide. From the responder's perspective, a good intuition for interpreting precision as the responder *load*, in terms of irrelevant questions read for each relevant question seen. The lower the value of

responder precision, the more questions a responder has to read through to find questions of interest to her.

There is an important trade-off between responder load and question coverage. For example, recommending all questions to every responder in the system will result in full question coverage. However, this will have a huge adverse impact on the responder load, as most of the questions we suggest to a responder will not be interesting to her. The fundamental challenge in the Q&A recommendation problem is to maximize question coverage, without overloading responders with irrelevant questions.

While precision and recall assume binary relevance judgements are provided, in practice a system is likely to calculate scores such as a probability value, that each responder will provide the best answer for a question. A common approach used in such cases is to select an arbitrary rank cut-off, so that only the top k ranks are considered relevant [112]. A more sophisticated approach is to average across multiple rank cut-offs for responder precision, resulting in the responder mean average precision, or *responder MAP* measure (Section 2.1.1.1). Since this work focuses retrieving the 'best answer' alone, the questioner's experience can be encaspulated by the mean reciprocal rank (Section 2.1.1.1) of the best answer, referred to here as *quesioner-MRR*. The lower the best responder in the ranks, the less likely it is that she will be contacted, thus enabling the questioner to receive the 'best answer'.

Another way to understand the relationship between questioner recall and responder precision is to draw a responder precision vs. questioner recall

graph, analogous to a precision-recall graph (Section 2.1.1.1) in document retrieval, showing the trade-off different algorithms make between load and coverage.

3.2.2.1 Micro and Macro Averaging

Given multiple topics, or in our case users, macro-averaging refers to the process of first calculating the relevant metrics for each user, and then averaging these metrics over the number of users [112]. Micro-averaging, in contrast, counts all matches/non-matches individually, and then averages these values across the entire dataset. An alternate interpretation of micro-averaging is, thus, weighing each user by their frequency of occurrence in the test dataset.

There is no strong reason to use macro-averaging for questioners. However, given the large amount of churn in the dataset, micro-averaging is an attractive choice for responders, as macro-averaging will over-emphasize the transient users. However, at the same time, Q&A forums are often dominated by a few highly active users. This set might sometimes be as small as consisting of only $2-5$ highly active users. Generally a reasonable recommender should do well on these users. Also, a recommender that retrieves these users frequently will do well on questioner recall metrics, due to the large number of questions they answer, so overall the recommender will appear to perform quite well. However, given their existing investment in the community, it is unlikely that a recommender would lead to increased participation from them, though it may reduce the likelihood of their stopping participation. Thus,

	Answered question	Did not Answer
Contacted	R_x^+	N_x^+
Not Contacted	R_x^-	N_x^-

Table 3.1: Metrics table for Responder x[35]

	Liked Answer	Did not like Answer / Ignored Answer
Contacted	U_x^+	I_x^+
Not Contacted	U_x^-	I_x^-

Table 3.2: Metrics table for Questioner x [35]

while micro-averaging is a reasonable choice for measuring responder satisfaction, it does not provide the complete picture.

For these reason, this work considers both micro and macro averaging while considering responder metrics, by reporting the macro mean average precision (MAP) scores, as well as the micro-precision score@10. To deal with the churn problem during macro-averaging, a threshold is used: only users who have answered at least 20 questions in the training dataset are considered during the macro-averaging stage.

3.3 Topic Models for Recommendation

A simple approach to the expert finding problem is to build a text-based profile, for example a TF-IDF profile (Section 2.1.2), for each responder. Then, when the system receives a new question, the question text could be compared to all the responder profiles using a similarity measure such as the cosine score (Section 2.1.2), and the responders with most similar profiles could be recommended the question. This is a common approach with respect to

expertise modeling and recommendation [150].

However this approach, which treats the expertise identification problem as a document retrieval problem, suffers from two serious drawbacks. Firstly, an expert is very different from a document in the sense the exact words used by a responder are heavily contingent on the questions she chose to answer, and do not cover all the information that a responder has, or the topic she may be knowledgeable about. Secondly, data about infrequent or new responders is likely to be sparse, and insufficient for retrieval purposes. For this reason a simple text profile based approach is not sufficient for the purpose of modeling human expertise. Preliminary experiments with a cosine-TFIDF based approach are presented on the Yahoo! Answers dataset in Section 5.2.3. The approach under performed other approaches by a wide margin, and so was not pursued in the main evaluation.

An alternate approach that overcomes this drawback introduced by this work is to model user expertise in terms of *topics*, instead of words. A topic can be seen as a higher-level concept over words and responding experts, and is modeled as a distribution over words, as well as experts. Hence, two questions may belong to the same topic even though they may have no words, or responders, in common. Similarly, a expert may be recommended a question even though there is no match in terms of profile words, if the question is judged as belonging to a topic the expert is interested in.

The next section introduces two generative models for a collection of question-answers in a Q&A system. Learning the parameters of these models

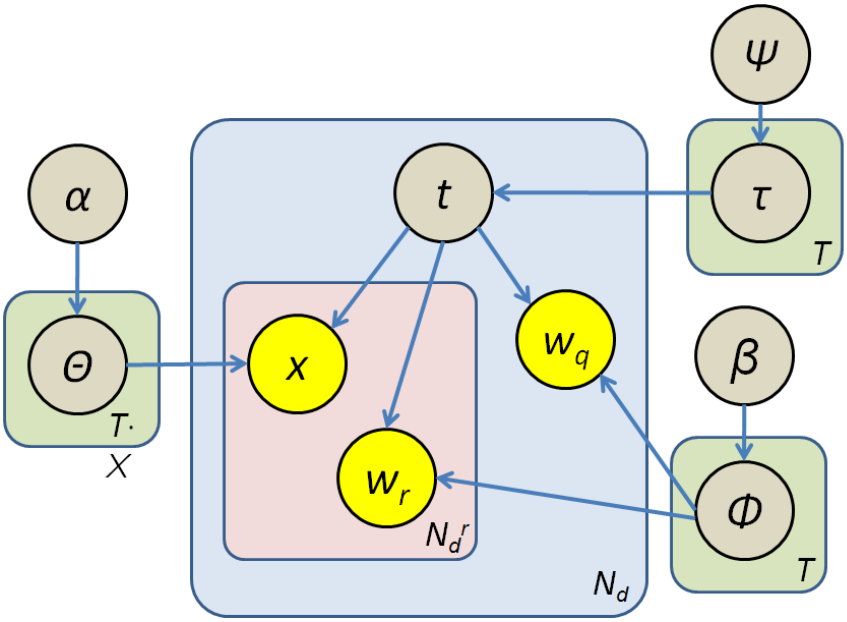

Figure 3.1: The Pure Multinomial Model [35]

enables the identification of topics of interest to various authors, as well as topic-word and topic-expert distributions.

3.3.1 The Pure Multinomial Model [35]

A basic outline for any online forum where people might gather for a discussion, or to exchange information can be constructed as follows: at each timestep, a topic is generated from a distribution over topics, which might have its own prior distribution. Then, some (question) words are generated related to the topic. The topic distribution may or may not be independent of the original author of the post, depending on how closely people stick to their topic of interest. Following this, a set of responders are chosen from a

distribution, based on the topic, and each of these responders generate further words. The words generated by the responders are related to the topic, but may or may not be seen as drawn from the same distribution as the topic. For example, if users have strong personal opinions, or try to draw the discussion in some favored direction, this might need to be modeled as each user might having its own word distribution for each topic, or the words as drawn from a mixture distribution of the original topic distribution, and a word distribution related to the user.

The model outlined above will be expensive to model due to the large number of parameters involved, but it can simplified considerably by introducing some assumptions, reducing it to a finite mixture model. These simplifications reduce the number of parameters considerable, while still providing important insight into the dataset. A more detailed outline of this model which we refer to as the *pure multinomial*[2] (PM) model, is given below:

Let the number of unique words p, the number of topics $|T|$, and the number of unique responders s be known in advance. Let the words be labeled $1, \ldots, p$ and the users $1, \ldots, s$, arbitrarily. At each timestep, a topic t is generated from a mutinomial distribution τ over topics, and a vector $\mathbf{w^q} = \{w_1, \ldots, w_p\}$ is generated, where w_i, is the count of word labeled i in the generated words. The words are generated from ϕ_t, a multinomial distribution over words corresponding to topic t. Following this a responder vector $\mathbf{x} =$

[2]Because, unlike the next model presented, this model uses multinomial models to represent both word and user distributions.

$\{x_1, \ldots, x_s\}$ is generated from θ_t, a multinomial distribution over users for topic t, where x_i is the number of time user labeled i responded. Each of the users in x in turn generates words based on the topic t. Here, an important simplifying assumption is made that the words generated by a responder as part of the answer are drawn from the same distribution ϕ_t as the topic.

This assumption can be understood as saying that the words used in the answer to a question by a responder depend only on the topic of the question, and do not depend on any attributes of the responder. This appears to be a reasonable assumption in Q&A forums where factual information is exchanged for the most part, or even in forums where personal opinions are expressed but the vocabulary used does not differ very much from user to user. It may not hold true in forums such as blogs or discussion forums, where responses to topics are much longer and more personal, and people may have favourite topics they might try to steer the topic to. But this level of model complexity is not required for Q&A forums. Figure 3.1 displays the generative model described above in plate notation. The shaded variables are the observed variables, while the unshaded variables are the hidden variables. Also, α, β and γ are symmetric Dirichlet priors. Another simplifying assumption is made that the total number of words and users generated for each question is independent of τ, θ and ϕ, and hence their randomness can be ignored in our discussion. Also, since it is assumed that the words generated by the responders depend solely on the topic, the words generated by all responders can be written as a vector $\mathbf{w^r} = \{w_1, \ldots, w_p\}$. Let $\mathbf{w} = \mathbf{w^q} + \mathbf{w^r}$.

3.3.2 Parameter Estimation

The generative model is represented in plate notation in Figure 3.1, and is essentially a mixture model with a finite number of components, where the number of components is the number of topics expected in the dataset. Let the number of such components/topics be g. Let the total number of unique Q&A interactions in the dataset D be n, where the j^{th} such interaction is referred to as d_j. Then let there be associated with each d_j a hidden vector $\mathbf{z_j}$ of length g, where $z_{ji} = 1$ if d_j is about topic i. Let $\theta = \{\theta_1, \ldots, \theta_g\}$, $\phi = \{\phi_1, \ldots, \phi_g\}$, and $\epsilon = (\theta, \phi)$. Then, assuming z_j is known for all d_j, the log likelihood of ϵ given D is given by:

$$\log_D L(\epsilon) = \sum_{i=1}^{g} \sum_{j=1}^{n} z_{ij} \{\log \tau_i + \log P(\boldsymbol{w_j}|\theta_i) + \log \left(P(\boldsymbol{x_j}|\phi_i)\right\} \tag{3.1}$$

The expectation maximization (EM) algorithm [51] is used to estimate the hidden variables z_i. The derived E-step and M-step for the algorithm are below:

E-Step: Given a guess for τ and ϵ, the expected value of z_{ij} is given by:

$$z_{ij} = \frac{\tau_i \cdot P(\boldsymbol{w_j}, \boldsymbol{x_j}|\epsilon_i)}{\sum_{h=1}^{g} \tau_h \cdot P(\boldsymbol{w_j}, \boldsymbol{x_j}|\epsilon_h)}$$

M-step:

Given expected values of z_j, τ, θ and ϕ can be estimated as:

$$\tau_i = \sum_{j=1}^{n} \frac{\psi + z_{ij}}{\psi|T| + n}$$

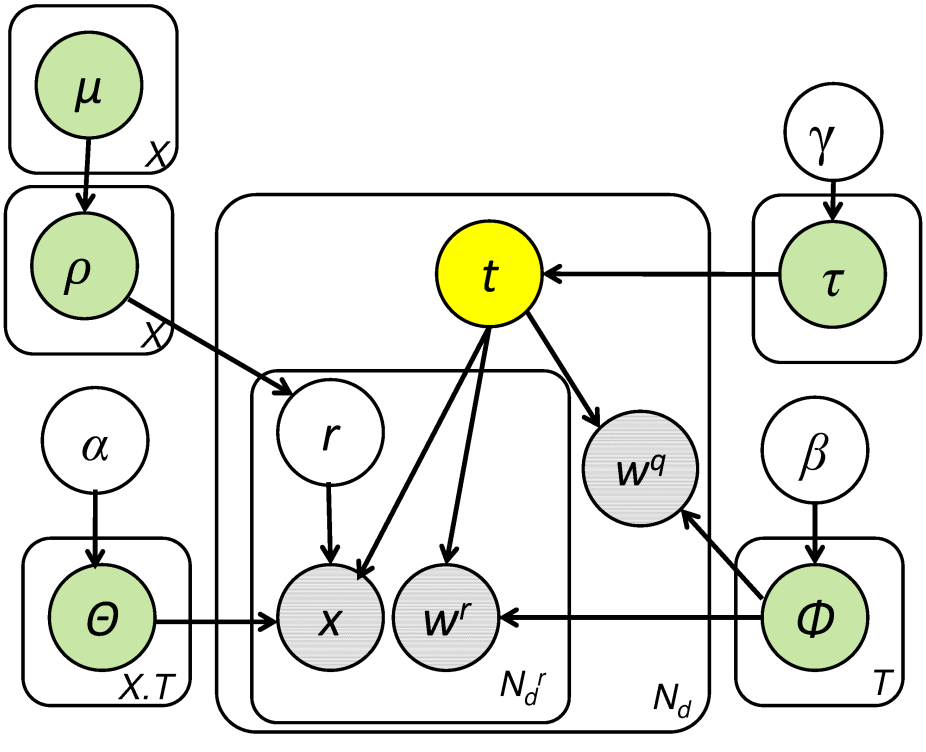

Figure 3.2: The Extended Generative model[31]

$$\theta_i = \frac{\alpha + \sum_{j=1}^{n} z_{ij} x_j}{\alpha s + \sum_{j=1}^{n} \sum_{i=1}^{g} z_{ij} x_j}$$

$$\phi_i = \frac{\beta + \sum_{j=1}^{n} z_{ij} w_j}{\beta p + \sum_{j=1}^{n} \sum_{i=1}^{g} z_{ij} w_j}$$

All θ_i and ϕ_i are normalized to 1.

3.3.3 Extended Generative Model with Responder Preferences [31]

The pure multinomial assumes that users are drawn from a multinomial distribution. On the other hand, the decision made by a user whether or not to answer a question, can be considered a much stronger preference expression. For example, a user reads a question, and then, if the topic matches her preference, she answers the question, else she does not. In this interpretation, the information provided by a user's decision to not answer a question is given weight. A natural choice to express this decision is as a Bernoulli random variable. If the user is interested in the topic, she answers the question with a probability θ, else she does not answer with a probability $1 - \theta$. Notice that we used θ for the user-topic multinomial distribution in the pure multinomial model: effectively we have removed a single multinomial model per topic with a much 'stronger' binomial model for each user-topic pair.

However, if we do not have information about whether a user saw a question, it is difficult to interpret a user's decision not to answer a question as binary (the reason for the original multinomial model). To overcome this, an extended generative model (Figure 3.2) is introduced, which uses a hidden variable-based approach is used, where for any question q and user j, a random variable r_{jq} is 1 if the user saw the question, and 0 otherwise. There is a multinomial distribution underlying r_j for each user j, which is the probability that a user saw a question, independent of the topic. ρ_j can be seen as roughly estimating the level of activity of user j on the forum. Available or heuristic information about whether a user read a question is incorporated via

a prior probability distribution μ_j, for each user. Currently, the prior simply incorporates Laplace smoothing. The aim here is to provide a lower-bound estimate of the effectiveness of this approach. In settings where more information about user browsing patterns is available, the prior estimate can be improved, or done away with altogether if the actual values of r_{iq} are known.

3.3.3.1 Parameter Estimation

A Gibbs sampling [63] based approach was used to estimate the distribution parameters. The sampling equations are straightforward to derive. Given a user i and question q, r_{iq}, the probability the user read the question even when she did not answer is estimated as:

$$\begin{aligned} P(r_{iq} = 1 | x_{iq} = 0) &\propto P(x_{iq} = 0 | r_{iq} = 1) P(r_{iq} = 1) \\ &= (1 - \theta_{it}) \rho_i \end{aligned}$$

Here θ_{it} is the Bernoulli distribution representing the user's interest in the question, based on the current best guess of its topic t. So essentially a user is more likely to have seen the question even if she did not answer it, if she is usually active (high value of ρ_q), and also not interested in the topic (high value of θ_{it}), Since a user always read a question if she answered it, $P(r_q = 1 | x_{iq} = 1)$ is always equal to 1. The topic t can be sampled as:

$$P(t|\boldsymbol{w},\boldsymbol{x}) \propto P(\boldsymbol{w},\boldsymbol{x}|t)P(t)$$
$$= P(\boldsymbol{w}|\phi_t) \prod_{r_i \in \boldsymbol{r}, r_i=1} (\theta_{it})^{x_i}(1-\theta_{it})^{1-x_i}$$

So essentially, the extended generative model incorporates responder preferences between questions much more strongly in the model via the user of a Bernoulli random variable. If a responder is believed to be active, but still does not answer a question, this is seen as strong evidence that the topic of the question does not interest the user. The aim is to arrive at clusters of questions where broadly the same set of users were active. The pure multinomial model and the extended generative model are evaluated empirically in Section 5.1.

Chapter 4

Authority Identification in OSNs

This chapter presents a new model, the average winnings model (Section 4.4.2) for ranking members in an question-answer forums. It also demonstrates (Section 4.4.3) that the fair bets model [48], a model for ranking players in round robin tournaments, is a good choice for ranking members in online social networks. This chapter also introduces a new co-ranking framework (Section 4.5), for combining information from multiple endorsement graph. Experimental results presented in the next chapter (Section 5.2), demonstrate the effectiveness of the algorithms developed here.

4.1 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. We see, so to speak, 'through a glass darkly'. However it does often capture a certain version of reality. This raises a fascinating question: can the member interactions on an OSN be used in general to draw conclusions about real-world relation-

ships and hierarchy; and – in particular – can we identify who the most well respected, prominent members of a social group are? This work attempts to construct an inferred global ranking of members of an OSN, according to the level of recognition they have achieved in the real world, or the authority[118] they may be considered to have on individuals in the real world.

4.2 Endorsement Graphs

Member interactions on an online social network (OSN) are often as a graph G, where each vertex in G represents a member, and an edge between members i and j indicates that at least one interaction took place between them. The edges may be directed if representing asymmetric interactions, such as user preference expressions. So, for example, a directed edge in graph G from member i to j might represent that i endorsed some content posted by j, for example, by 'liking' the content, or sharing it with others, etc. The resulting graph, referred to in this work as an *endorsement graph*, can be analyzed to identify the most popular (by endorsements) nodes in the graph. An endorsement graph can be represented as an adjacency matrix M such that $M_{ij} = 1$ if i endorsed j.

It is possible to have multiple endorsement graphs over the same OSN [34]. For example, an endorsement graph may be constructed from invitation data, containing information about who initiated a connection with an invitation (*invitation graph* [34]). The assumption would be that users on a social network are more likely to send invitations to users that they respect, or at

least do not disrespect. Similarly, liking content provided by another user, or even navigating to another user's profile multiple times and browsing through it can be seen as a sign of interest (*navigation graph*[34]).

A number of methods have been developed in social network analysis [141] for measuring the importance of a node in a single graph. A straightforward measure is degree centrality, which is essentially the indegree of a node, in this case the number of endorsements received. A more sophisticated measure, *eigenvector centrality* [25], weighs each endorsement by the importance of the endorser. Thus the weight of each node depends on all other nodes, in a recursive fashion. The well-known PageRank algorithm [30], used for ranking web pages by interpreting hyperlinks as endorsement, is a variant of the eigenvector centrality measure.

4.2.1 Eigenvector Centrality and PageRank

There are two main differences between eigenvector centrality and PageRank:

1. While eigenvector centrality uses the adjacency matrix, PageRank normalizes each row of the adjacency matrix M to sum to 1, by dividing each row by its outdegree c_i, so that the resulting matrix P' is stochastic.

2. Each row of the stochastic matrix P' is smoothed with a random positive stochastic vector $\boldsymbol{r}$, so that the resulting matrix is non-zero. In other words, a new matrix P is created such that, each row i of P, $P_{i*} =$

$(1-d)P'_{i*} + dr^\top$. The new matrix P used for further computation is thus a positive stochastic matrix.

Given an endorsement graph G with k vertices, the PageRank score a_i of the vertex v_i is given by:

$$a_i = \sum_{j=1}^{K} P_{ji} a_j$$

That is, the score a_i for vertex i is recursively the summation across all the vertices j that endorsed i, of the fraction of endorsements that were for P_{ji}, multiplied by their score a_j. In matrix form, this can be written as:

$$P^\top \boldsymbol{a} = \boldsymbol{a}$$
$$\Rightarrow \boldsymbol{a}^\top P = \boldsymbol{a}^\top \tag{4.1}$$

Since P is a row stochastic matrix, $\boldsymbol{a}$ in eq.(4.1) is by definition [91] the stationary distribution vector of the Markov chain defined by P. An intuitive interpretation for why $\boldsymbol{a}$ is the stationary distribution is provided by the *random surfer model*, in the context of web pages. According to this interpretation, the PageRank algorithm mimics the behavior of a web surfer, who starts surfing at a random page and at each timestep, randomly select an outgoing link (edge) from the current page. However, with a certain probability d at each timestep, the surfer gets bored with the current page and jumps to a randomly selected new page. The new page is selected from the probability distribution $\boldsymbol{r}$. Then the PageRank score of a page corresponds to the fraction of time the surfer will spend on it. Thus, pages with a high PageRank score

are pages that are linked to by a large number of web pages, or by a smaller number of other pages with high scores.

Due to this interpretation, d is often called the *random restart probability*, or the *teleportation probability*, as it is the probability with which the random walk randomly restarts. $\boldsymbol{r}$ is usually referred to as the random restart of teleportation *vector*.

A similar reasoning justifies the use of eigenvector centrality in social networks: the important nodes connect not only to a lot of other nodes, but to other important nodes as well. The PageRank vector has been used to identify influential users in OSNs [41, 62, 145].

4.3 Random Surfer Model: Drawbacks

Intuitively, the random surfer model suggests PageRank as a reasonable algorithm for identifying authorities: authorities are likely to be users who receive a large number of strong endorsements. However, the model ignores a number of social dynamics, which can distort authority estimates. Two prominent dynamics identified by this research [34]are:

4.3.1 Impact of User Activity on Endorsements

Two factors determine the number of endorsements an OSN member receives: a) their authority level, which determines the desirability of a connection with them, and 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). Some other factors that complicate the relationship between endorsements and authority are:

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 [105], where low status nodes link to those of high status. This observation does not play a part in the random surfer model.

2. In contrast to the Web where most information is publicly accessible, OSNs 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 she receives.

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 [101, 145], on both Flickr[84] and Twitter[82].

4. Older users can become entrenched over time, and have an indegree

disproportionate with their authority level. This can discourage younger users from participating.

4.3.2 Complementary Endorsement Graphs

There are many different ways in which an OSN member may choose to interact with another. The chosen type of interaction often depends on the relationship between them. A useful distinction made by this research [34] is between *symmetric* and *asymmetric* interactions. A symmetric interaction is one with an expectation of reciprocation, or at least an acknowledgement. For example, sending an invitation to connect, or sending an email, is a symmetric interaction. An asymmetric interaction, on the other hand, does not expect any response for the receiving person. Looking at a person's profile, 'following' their updates on a network such as Twitter [82], are asymmetric interactions.

Asymmetric and symmetric endorsements are broadly complementary. Asymmetric endorsements are more aspirational in nature compared to symmetric endorsements. Thus, for example, on a professional OSN, a user may be more likely to connect to her immediate supervisor, but may browse her company CEO's profile more often. Also, symmetric endorsements are more 'exposed' to the receiver and thus more guided by social norms: a user may feel obligated to send invitations to connect to all the people she meets at her workplace. However, symmetric endorsements are less susceptible to celebrities and the vagaries of the news cycle.

Identifying user invitation (hyper-linking) and browsing behavior on

OSNs as symmetric and asymmetric respectively is an important departure of this work from the standard random surfer model, which posits a very close relationship between link structure and browsing. Essentially, for a random surfer, the PageRank vector is the result of browsing the hyperlink graph. For example, Gleich *et al.* [66] empirically learn the random restart (d) parameter over the hyperlink graph from web browsing data. On the other hand, Browserank [110] assumes that the Web hyperlink structure can be ignored given the user web browsing behavior, which it treats as the 'actual' random surfer walk. These assumptions are less true for OSNs, where the differences between linking and browsing behavior is much greater.

To address the differences between endorsement behavior on the Web, and on OSNs, this work proposes the following approach [34]:

1. A tournament model [48, 114] of member interactions, where user behavior can be interpreted as driven by an underlying Terry-Bradley model [28]. Intuitively, this can be understood as follows: each OSN member has an intrinsic authority *strength*. The direction of an interaction (edge in the graph) is a probability distribution based on the relative strength of the two member nodes.

2. A model for combining authority strength related information from multiple graphs, where each graph is constructed over the same set of members, but represents different aspects/modes of their behavior. The model is equivalent to simultaneously using the authority scores vec-

tor of one graph as the random restart vector of the other graph, and vice versa.

4.4 Tournament Models

Leskovec et al. [105] analyzed data from two online communities, epinions[57], and Wikipedia [83] moderator interaction, and 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 others users they perceive as being higher status, in terms of being more knowledgeable, etc., instead of being concerned with likability or personal relationships. Intuitively, this is plausible: say two users i and j come in contact with each other on an OSN, or in the real world, and suppose j is more authoritative than i. Then it is in i's interest to connect with j, and try to maintain that connection. Thus if we see a connection between i and j on an OSN, we should expect i to initiate that connection.

4.4.1 The Bradley-Terry Model [49]

Under the Bradley-Terry model [26–28, 49] (also discussed in Section 2.2.3), we can formalize this as follows: suppose the authority strengths of OSN members i and j are given by a_i and a_j respectively, where $a_i \in R^+$ for any i. Say we treat the interaction between i and j as a game, where the endorser is considered to have 'lost' the game, and the endorsed member to have won. Then assuming a tournament took place between i and j, the

probability that i won the game is given by $\frac{a_i}{a_i+a_j}$. Assuming no draws, it follows that the probability that j won the game is given by $\frac{a_j}{a_i+a_j}$.

A *tournament matrix* [48] M can be constructed to represent the result of all such 'games', with M_{ij}, containing the number of times player i lost to j. Assuming N_i is the total number of games played by i, and N_{ij} is the number of games between i and j. Then as $N_{ij} \to \infty$, M_{ij} converges to its expected value, $N_{ij}\frac{a_j}{a_i+a_j}$ 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. The next section presents the *average winnings model*, a new scoring model for calculating the underlying authority weights a_i of players in a tournament where each player has a different probability of engaging in a game.

4.4.2 The Average Winnings Model

Continuing our understanding of interactions on a Q&A forum or OSN as games, with the direction of the resulting edge determined by the outcome of the game, consider a tournament where the number of games two players

play against each is drawn from a distribution. Let the number of games i and j play against each other in this tournament be given by N_{ij}.

An example of this could be data drawn from a Q&A forum, where N_{ij} is the number of questions that both i and j have answered. N_{ij} could depend on a number of factors, such as their respective levels of activity, and the degree of interest they have in the same set of questions. 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 (Section 4.4.1) Z, and also as an endorsement graph. For numerical purposes, it needs to be ensured that he graph is strongly connected. This can be achieved by modifying Z in two ways:

1. By introducing a regularizing node in the graph, a technique also used by Chen *et al.* [43]. This node can be understood as a player that has won and lost against every other player exactly once. A self-loop is added to the regularizing node, to ensure that the corresponding Markov chain is aperiodic, or

2. By adding a small probability of teleportation (loss) to a uniformly chosen node, as per the PageRank model [30].

Given the popularity of the PageRank algorithm in the online social network research community, this research selects the second option. Then the following proposition provides a method for calculating the authority strengths of the participating players (OSN/Q&A members).

Proposition 4.1. *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 $N_{ij} = N_{ji}$, 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 $\boldsymbol{\pi}$ 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 $\boldsymbol{\pi}$ to add to 1 [69]. For P, the detailed balance equation between two states i and j are given by:

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

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

$$a_i = \frac{\pi_i}{Z_i} \tag{4.2}$$

□

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.

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 [48, 114, 132], a model introduced by Daniels [48] for ranking players in round robin tournaments, where each player plays against another player exactly once. This research shows that the fair bets model can be interpreted as a special case of the average winnings model, and is effective for ranking members in professional OSNs. In comparison, the average winnings model is found to be a good fit for Q&A forums.

4.4.3 The Fair Bets Model [48]

The fair bets model calculate player strength scores based on a *generalized tournament matrix* [132]. 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, $V_{ij} + V_{ji} = 1$. 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 $\deg^+(i)$ represents the outdegree of vertex i. P is a row stochastic matrix. Assuming, for now, that P is aperiodic, and thus ergodic. Then the following proposition is true [48].

Proposition 4.2. *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:

$$a_i = \frac{\pi_i}{deg^+(i)} \tag{4.3}$$

□

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

A significant advantage of the fair bets model is that, since it models each interaction as a game in a tournament, it naturally de-incentivizes reciprocation. This is because an OSN member with a higher fair bets score will see her score decrease, even if she draws with a weaker member. In comparison, there is no penalty for outgoing edges in the standard PageRank approach.

4.4.3.1 Social Capital Exchange Interpretation

For a generalized tournament matrix V, the fair bets score a_j of player j satisfies the following property:

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

Based on this equation, Slutzki *et al.* [132] provide the following interpretation of the fair bets model: 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. 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.

In the context of online social networks, this research [34] views fair bets 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[1]. 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

[1]A similar intuition can be applied to other endorsement graphs besides the invitation graph.

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. The fair bets model, is thus, intuitively a good fit for professional social networks.

4.4.4 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. 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 the likelihood of multiple interactions given the first one is high, and they are likely to follow the same pattern. Consider a professional social network: if two users interacted once, then multiple interactions are likely to follow the same pattern as the first one, in terms of edge direction. In contrast, in a Q&A forum, if one players wins one interaction by giving the 'best answer', it is still likely that the other player will win the next round. 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.

4.5 Co-ranking Complementary Graphs

In the random surfer interpretation of the PageRank algorithm [30], at each timestep, with a certain probability $1 - d$ (usually set to 0.85), the surfer randomly selects an outgoing link from the current page. With the remaining

probability d, the surfer gets bored and jumps to a completely new page. The probability d is referred to as the *teleportation probability*(also random restart (RSR) probability), and the vector the new page is chosen from is called the *teleportation vector* (or RSR vector). The vector can be uniform, or biased to reflect some priorly known information. For example, the teleportation vector could be personalized [59] 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.

This research [34] considers the following method to inform one graph (say, invitation) with information from the other (say, navigation) graph: suppose we use the authority vector of the invitation graph as the teleportation vector for the navigation graph. 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 (assuming the process converges). This approach is referred to here as the *bimodal co-ranking* approach. The idea behind this approach is mutual positive reinforcement [152]: useful information in one graph can be used to improve the authority estimates of the other graph, and vice versa. The authority vector of one graph serving as teleportation vector of the other graph could be either of the three authority scoring choices considered till now: PageRank, or fair bets, or average winnings.

As the next section shows, the process does converge for strongly connected graphs. The research also shows [34] that 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 iterative bimodal approach, can be obtained from the the PageRank vector of the composite graph.

4.5.1 The Bimodal Co-ranking Algorithm [34]

Given two strongly-connected graphs G_A and G_N with stochastic matrices P_A and P_N, the co-ranking algorithm can be defined as follows:

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

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

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

The proof for: a) the co-ranking process converges, and b) the process is equivalent to simultaneously using the PageRank vector of one graph as the teleportation vector of the other graph, and vice versa, is provided below.

4.5.1.1 Bimodal Co-ranking: Proof of Convergence [34]

To prove:

1. The algorithm described in Section (4.5.1) converges after a certain number of iterations t.

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

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

Outline We prove the result by showing that the bimodal co-ranking algorithm is equivalent to applying the power iteration eigenvalue algorithm [136] 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 (2) is shown algebraically.

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

Claim 1. *[73, equations (2–6)] For a row stochastic matrix P, construct another matrix $A = cP^\top + (1-c)\boldsymbol{r}\boldsymbol{e}^\top$, where $e_i = 1$ for all i, $0 < c < 1$ and $\boldsymbol{r}$ is a positive vector with $|\boldsymbol{r}|_1 = 1$. That is, $\boldsymbol{r}$ is a teleportation vector added to P, and $1 - c$ is the teleportation probability. Then the solution to the PageRank equation for A, $\boldsymbol{x} = A\boldsymbol{x}$, is given by $\boldsymbol{x} = (1-c)(I - cP^\top)^{-1}\boldsymbol{r}$.*

Proof.

$$\boldsymbol{x} = A\boldsymbol{x} = [cP^\top + (1-c)\boldsymbol{r}\boldsymbol{e}^\top]\boldsymbol{x} = cP^\top\boldsymbol{x} + (1-c)\boldsymbol{r}$$
$$\Rightarrow \boldsymbol{x} = (1-c)(I - cP^\top)^{-1}\boldsymbol{r}$$

□

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 [91, 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 [136]. 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 timestep, 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 $(I - cP)^{-1} = I + \sum_{k=1}^{\infty} (cP)^k$ from Claim 2.

$$S_{ij} = \begin{cases} 0 + cP_{ij} + c^2(P^2)_{ij} + c^3(P^3)_{ij} + \dots & \text{if } i \neq j \\ 1 + cP_{ii} + c^2(P^2)_{ii} + c^3(P^3)_{ii} + \dots & \text{if } i = j \end{cases} \tag{4.4}$$

Then

$$\sum_{j=1}^{K} S_{ij} = 1 + c\sum_{j=1}^{K} P_{ij} + c^2 \sum_{j=1}^{K} (P^2)_{ij} + c^3 \sum_{j=1}^{K} (P^3)_{ij} + \dots \tag{4.5}$$

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 (4.5) can be written as:

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

To show that all entries of S are positive, consider equation (4.4). 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 and hence $X = (1-c)S$ is a positive row stochastic matrix. □

Proposition 4.3. *The co-ranking process defined in Definition (4.5.1) 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 $M = (1-c)^2(I - cP_N)^{-\top}(I - cP_A)^{-\top}$, using the power iteration algorithm [96], 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:

$$\boldsymbol{z_{t+1}} = ((1-c)(I - cP_N{}^\top)^{-1}(1-c)(I - cP_A{}^\top)^{-1})\boldsymbol{z_t}$$
$$\Rightarrow \boldsymbol{z_{t+1}} = [(1-c)(I - cP_N)^{-\top}(1-c)(I - cP_A)^{-\top}]^t \boldsymbol{z_0} \tag{4.6}$$

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

Using Claim (3), it can be seen that equation (4.6) is the product of the transpose of two positive row stochastic matrices, and hence is a positive column stochastic matrix. Let this matrix be:

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

Then eq. (4.6) 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 [96], to a unique postive eigenvector corresponding to the largest eigenvalue of M [136].

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 $M^{t+1}\boldsymbol{z_0}$, 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 respective. For brevity, we

use both d and $c = 1 - d$ below. Then we can write:

$$\boldsymbol{r_N} = d(I - cP_N)^{-\top} M^t \boldsymbol{z_0} \tag{4.7}$$

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

$$\boldsymbol{r_N} = d(I - cP_N)^{-\top} M^{t+1} \boldsymbol{z_0} \tag{4.8}$$

For $\boldsymbol{r_A}$, we calculate as follows:

$$\boldsymbol{r_A} = d(I - cP_A)^{-\top} d(I - cP_N)^{-\top} M^t \boldsymbol{z_0} \tag{4.9}$$

Then using eq. (4.7) we can write:

$$\boldsymbol{r_A} = (1 - c)(I - cP_A)^{-\top} \boldsymbol{r_N} \tag{4.10}$$

Then using eq. (4.9), eq.(4.8) can be rewritten as:

$$\begin{aligned} \boldsymbol{r_N} &= d(I - cP_N)^{-\top} M M^t \boldsymbol{z_0} \\ \Rightarrow \boldsymbol{r_N} &= d(I - cP_N)^{-\top} d^2 (I - cP_A)^{-\top} (I - cP_N)^{-\top} M^t \boldsymbol{z_0} \\ \Rightarrow \boldsymbol{r_N} &= (1 - c)(I - cP_N)^{-\top} \boldsymbol{r_A} \end{aligned} \tag{4.11}$$

Based on Claim 1, eq.(4.10) is a solution to $\boldsymbol{r_A} = A_1 \boldsymbol{r_A}$, where:

$$A_1 = \left((1 - d)P_A + d\boldsymbol{e}\boldsymbol{r_N}^\top\right)^\top$$

Similarly, eq.(4.11) is a solution to $\boldsymbol{r_N} = A_2 \boldsymbol{r_N}$, where:

$$A_2 = \left((1 - d)P_N + d\boldsymbol{e}\boldsymbol{r_A}^\top\right)^\top$$

This proves the second part. □

4.5.1.2 Proof Of Equivalence: Bimodal and Composite Graph Models [34]

However, the process requires multiple runs of the PageRank algorithm, which is computationally expensive, and not feasible for very large graphs. For this reason, we develop an alternate approach by showing that computing the PageRank vector for a specially constructed composite graph is equivalent to running the co-ranking algorithm over a pair of graphs.

Assume we have 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:

Proposition 4.4. *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 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 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*

the result of the bimodal co-ranking algorithm.

Proof. Let the transition matrix of V_A be written as P_A and its (unknown) PageRank vector be $\boldsymbol{r_A}$. Similarly, let the transition matrix and PageRank vector of V_N be P_N and $\boldsymbol{r_N}$ respectively. Let $\boldsymbol{e}$ be a vector such that $e_i = 1$ for all i. Then, as proven by Theorem 2(4.3), PageRank vectors of G_A and G_N satisfy the following equations:

$$\left((1-d)\,P_A + d\boldsymbol{e}{\boldsymbol{r}_N}^\top\right)^\top \boldsymbol{r}_A = \boldsymbol{r}_A \tag{4.12}$$

$$\left((1-d)\,P_N + d\boldsymbol{e}{\boldsymbol{r}_A}^\top\right)^\top \boldsymbol{r}_N = \boldsymbol{r}_N \tag{4.13}$$

Expanding (4.12), we get:

$$(1-d)\,{P_A}^\top \boldsymbol{r}_A + d\boldsymbol{r}_N e^\top \boldsymbol{r}_A = \boldsymbol{r}_A \tag{4.14}$$

$$\Rightarrow (1-d)\,{P_A}^\top \boldsymbol{r}_A + d\boldsymbol{r}_N = \boldsymbol{r}_A \tag{4.15}$$

since r_A sums to 1.

Similarly, for (4.13), we get:

$$(1-d)\,{P_N}^\top \boldsymbol{r}_N + d\boldsymbol{r}_A = \boldsymbol{r}_N \tag{4.16}$$

Let I_k be an identity matrix of size k. Then equations (4.15) and (4.16) can be written in matrix form as follows:

$$\begin{bmatrix} (1-d)\,{P_A}^\top & dI_k \\ dI_k & (1-d)\,{P_N}^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{r}_A \\ \boldsymbol{r}_N \end{bmatrix} = \begin{bmatrix} \boldsymbol{r}_A \\ \boldsymbol{r}_N \end{bmatrix} \tag{4.17}$$

Let P be a matrix, such that:

$$P = \begin{bmatrix} (1-d)\, P_A & dI_k \\ dI_k & (1-d)\, P_N \end{bmatrix} \tag{4.18}$$

and let $\boldsymbol{r} = \begin{bmatrix} \boldsymbol{r_A} \\ \boldsymbol{r_N} \end{bmatrix}$. Then equation (4.17) can be written as $P^\top \boldsymbol{r} = \boldsymbol{r}$. Then, by the definition of the PageRank vector (Section 4.2.1), $\boldsymbol{r}$ is the PageRank vector for P. □

4.5.2 Co-ranking with Tournament Models

The matrix P in equation 4.18 can be modified to use other authority models instead of PageRank, as the teleportation vectors [34]. 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 $\frac{1}{o_i}$, 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 $R_N(j,j) = \frac{1}{o_j}$, 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 fair bets vector of each graph serves as the teleportation vector of the other (the final results still need to be normalized to get fair bets scores). A similar co-ranking models can be constructed with the average winnings model, or two different authority models can be combined in a co-ranking framework, if required.

4.6 Authority Estimation under Social Voting in Q&A Forums

A problem with a endorsement aggregation based approaches towards authority estimation is that, even in the case of authoritative users on a forum, the motivations behind a selection they made is not always clear. The reason for this is the social aspect of online forums: over time users develop social relationships with other users, and these relationships impact choices about the content they consume or favor. In other words, the reputation that users aggregate over time does not depend only on their quality, but also on many behavioral side-effects of their social network interactions. Ignoring these biases during authority identification can adversely affect the accuracy of results. Also, taking these preferences into account while recommending content for users where such information is available for them, can improve the quality of personalized recommendation.

Two common examples of such biases on OSNs are reciprocity and *social voting*. Many content-sharing sites such as Digg and Yahoo! Answers allow users to add other users as contacts or friends. The aim is to increase engagement: the site is designed so that users find it easy to get updates on the activities of their contacts. A side-effect is that since users find interesting stories via their contacts, users with many contacts find it much easier to promote their content. This phenomenon, called social voting [103] has been documented on the website Digg [65] as well as Flickr [103].

Perhaps due to the asymmetric nature of the questioner-responder re-

lationship, this research finds that reciprocation plays a smaller role on Q&A forums: if B answers a question by A, it is likely that B is more knowledgeable than A, which makes an answer by A to a question by B unlikely (unless it is in a different topic). Also, the specific nature of the information requested makes it less likely that social voting can play a role. Instead, this research identifies a new behavioral pattern, *discovered affinity*, in data derived from Q&A forums. This pattern is described in the next section.

4.6.1 Discovered Affinity

Most social network graphs demonstrate a 'small world' property [14, 115], where, even though most vertices are not adjacent, the shortest path between most vertices is small. Watts and Strogatz [143] identifed another property of small world graphs: a high value of the *clustering coefficient*. The clustering coefficient for a vertex on an undirected graph as [56]:

$$C_v(G) = \frac{\text{Number of triangles including v}}{d(v) \cdot (d(v) - 1)} \tag{4.19}$$

where a triangle is any set of three edges including v, which are all connected to each other. $d(v)$ represents the degree of vertex v. The denominator equals $\binom{d(v)}{2}$, and represents the number of possible ways in which two edges may be chosen from the edges adjacent on v. For a network based on social ties, the clustering coefficient measures the extent to which users who have common friends are also friends of each other.

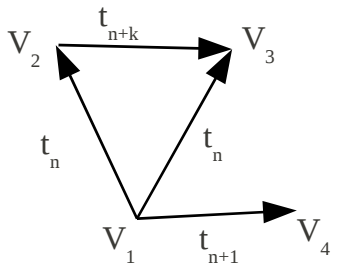

Figure 4.1: Clustering Coefficient for a Multigraph

Q&A forums and content-sharing networks (referred to here collectively as content-oriented networks of COSNs) have different dynamic, with a greater emphasis on user interests and affiliations. This can be represented as an affiliation network [98]: a bipartite graph, with edges connecting each user with her affiliation/interests. The clustering coefficient can be calculated on the induced undirected graph of the bipartite graph [115]. This approach has the drawback of assuming that all user affiliations are significant. This is a strong assumption on such networks, where an affiliation could represent something like having answered the same question on a Q&A forum, or having commented on the same article on a blog. Due to this drawback, the next section proposes a new definition of the clustering coefficient, based on a multigraph (instead of bipartite graph) model of user behavior on a COSN.

4.6.1.1 Clustering Coefficient for a Multigraph

The following observation enables the modeling of user behavior and affiliations on COSNs as a multigraph, instead of a bipartite graph: any event, or an affiliation/group, is often initiated, or 'owned' by a user who is a member of the same social network. Based on this, given a graph G consisting of all the N users on a COSN, the multigraph representing user behaviors is constructed as follows: for each event/group t, initiated by the user corresponding to vertex v_i, and attended/joined by a set of users V', for each user $v_j \in V'$, an outgoing edge is added from v_i to v_j, labeled with t. Then a *triad* on the graph is defined as three vertices v_i, v_j and v_k, such that one of them is incident on the other two via outgoing edges, and the edges share at least one common label. A *triangle* is then defined as a triad in which all vertices are connected to each other. The clustering coefficient is then defined as:

$$C(G) = \frac{\text{Number of triangles in } G}{\text{Number of triads in } G} \tag{4.20}$$

So, in Figure 4.1, vertices v_1, v_2 and v_3 form a triangle, as the $v1 - v2$ and $v1 - v3$ edges share the same label t_n. Since it is also the only triad in the graph, the clustering coefficient of the multigraph is 1.

Multigraph clustering coefficient values for three question categories, crawled from the Yahoo! Answers website, are shown in Table 4.1. The categories were selected with the expectation that they would lie at different points on the spectrum from information seeking to social behavior, with the

Space & Astronomy forum being most informational, and the Wrestling forum the least. The clustering coefficient was calculated for data crawled at different points of time, with an approximately 1.5 year gap in between. The aim was to see by how much the coefficient values changed over time.

The results show an interesting trend, where the coefficient values fall significantly for the S&A dataset, while rising significantly for the Wrestling dataset. The values for the B&A dataset are comparatively low but relatively stable. One reason for the low value, based on an analysis of the data, is that there are a large number of extremely short-term visitors to the forum, who leave the forum after a few questions or answers. However, the reason for the decline in the clustering coefficient for the S&A dataset are less clear. Preliminary analysis suggests two reasons: an increase in the number of short-term visitors, and a trend towards less discussion oriented and more factual questions. The reasons behind these trends are not known.

A high clustering coefficient indicates the following mechanism of social behavior: if users A and B attend the same event (hosted by say, C), it increases the probability that B will attend an event hosted by A, or vice versa. This research refers to this as *discovered affinity*. Identifying this phenomenon on the Yahoo! Answers website enable its use as part of the prediction model, enabling significant improvements in recommendation quality.

Data Category	Time Period	
	Feb-Mar 2009	Oct-Nov 2010
Space and Astronomy (S&A)	10.46%	3.42%
Books and Authors (B&A)	3.47%	4.85%
Wrestling	13.75%	16.14%

Table 4.1: Clustering Coefficient Values for the Yahoo! Answers Dataset (three categories)

4.6.2 Estimating Selection Preference Distribution

Sections 3.3 and 4.6 discussed topic-based models of user authority, and some documented behavioral mechanisms that impact user behavior on OSNs. This section provides an algorithm for estimating, for each user on an OSN, the degree to which they are interested in, or influenced by these different modes of behavior: information seeking, or social. This influence is represented for each user by a latent variable, called *fairness* or *objectivity*. Raters' fairness or objectivity, represented as a vector $\mathbf{o}$, is supposed to estimate the degree to which the selections made by them are motivated by the quality of the content rated, as opposed to the influence of their. For raters motivated by content quality, $o_i = 1$, and for raters completely driven by their social network, $o_i = 0$.

The algorithm can intuitively be understood as follows: suppose the objectivity value of each user was known. Then, while performing reputation estimation, the transition probability values for the non-objective users ($o_i = 0$) should be ignored. In the random surfer model, this is equivalent to the following behavior: if the surfer visits a vertex that it knows to be non-objective, it teleports at random to a new vertex on the graph. However,

since the exact value is not known, the probability of teleportation is set in the algorithm as proportional to the expected value of the objectivity. Thus, ratings by questioners estimated as making more selections based on a personal distribution will be less influential than selections made by questioners estimated as making few or no selections based on a personal distribution.

To estimate o_i, a rater's behavior is modeled as follows: the number of ratings q_j each user provides is drawn from a distribution (this distribution need not be modeled as part of the final algorithm). Each user also has a hidden variable o_i associated with him/her. Following this, for q_j timesteps, depending on the value of o_i, the user i draws values from one of two distributions: the quality distribution (if $o_i = 1$) and his/her personal social affinity distribution (if $o_i = 0$.). The quality distribution α is the reputation distribution, calculated as defined in Section 3.3. The social affinity distribution σ_i for user i is defined as the user's social network, with all members equally likely; people who are not member are assigned a small prior, to assure nonzero likelihood. We use another prior: the prior probability of selecting from the social affinity distribution defined for each user, which is the number of times the user selected a poster from his/her social network, based on historical data. We refer to this as the affinity prior π. Then given a set of selections, the posterior probability of selecting from either of the two distributions can be calculated.

The quality distribution depends on O, as only users who are objective should be considered while calculating α. However, re-estimating α changes

the objectivity values O for all users. We use an iterative expectation maximization based algorithm, where user objectivity and the quality distribution are alternatively estimated.

The algorithm [32] is given below. An experimental evaluation of the algorithm, on data drawn from Yahoo! Answers and Digg are presented in Sections 5.2.3 and 5.2.4 respectively.

1. Initialize π_i for each user i, Q and $\mathbf{r}$. Set $\alpha = (I - OQP)^{-T}\mathbf{r}$. Repeat Step 2 to 4.

2. *Objectivity Estimation*: For each rater i in the dataset, and their ratings $\mathbf{s_i}$, estimate $o_i = \frac{\pi_i P(\mathbf{s_i}|\alpha)}{\pi_i P(\mathbf{s_i}|\alpha)+(1-\pi_i)P(\boldsymbol{s_i}|\sigma_i)}$.

3. *Likelihood Estimation*: a) Calculate $LL^{(j)} = \sum_{i=1}^{N}(1 - o_i)\log P(\boldsymbol{s_i}|\sigma_i) + o_i \log P(\boldsymbol{s_i}|\alpha)$, where j is the current iteration number.

4. *Exit Condition*: If $LL^{(j)} < LL^{(j-1)}$, exit.

5. *Reputation Estimation*: Set $\alpha = (I - OQP)^{-T}\mathbf{r}$.

Chapter 5

Experimental Results

This chapter presents an empirical evaluation of the approaches developed in Chapters 3 and 4. Section 5.1 evaluates the expert finding approaches on data derived from the online StackExchange Q&A forum [80]. Section 5.2 evaluates authority identification techniques on data derived from four online communities. Co-ranking tournament models were evaluated on the professional social network LinkedIn [46] and the Q&A forum StackExchange [80]. The algorithms for incorporating social and behavioral effects in authority calculations are evaluated on data from the Q&A forum Yahoo! answers [85], and the social content exchange forum Digg [17].

5.1 Question Recommendation in StackExchange

Due to the logistic effort involved in an online evaluation, the experimental evaluation for this research was entirely offline. There are two main consequences of an offline approach:

1. When a question is recommended to a responder, it is not possible to tell whether she saw the question or not. In the case of less active responders, it is quite likely that they did not see the question. As a

result, the evaluation metrics have unnaturally low values. It can be expected that the system will perform better on the same metrics in an online evaluation.

2. It is not possible to evaluate whether the system design decisions impacted expert retention in any way. Thus, while it is hypothesized that lower expert load should lead to higher expert satisfaction, and so retention, it cannot be empirically evaluated that this is the case.

In contrast, Horowitz *et al.* [75] and Hecht *et al.* [74] provide examples of online evaluations of Q&A recommenders.

The evaluation compared three question recommendation algorithms based on data extracted from a data dump [78] provided by the Stack Exchange network website. The dataset consisted of Q&A interactions performed on the website in August 2012. Six expert communities were selected for evaluation: two scientifically oriented communities (Mathematics and Physics), three technology and engineering communities (Security, AskUbuntu and ServerFault), and one language and discussion-oriented community (English). For each community, the dataset was chronologically ordered based on the timestamp when the question was posted, and then divided into two equal halves. For datasets consisting of more than $7,000$ questions (Mathematics and ServerFault), only the first $7,000$ questions were selected. For datasets with fewer than 4000 questions, the first 2000 documents were used for training, and the rest for testing. The reason for the $50 - 50$ division instead of the more common

$70 - 30$ one, was to mimic the fast pace at which Q&A datasets change. Each of the models was then trained on the first chronological half, and tested on the second.

5.1.1 Evaluation Metrics

The evaluation focused on the following metrics:

1. *Question Coverage*: The micro-averaged precision across questioners, where only best answers were considered. That is, the fraction of questions (expressed as a percentage) for which the responder who gave the best answer was suggested by the retrieval model.

2. *Responder Load*: The inverse of the responder micro-averaged precision. This is the number of irrelevant questions a responder will have to look through, to find a question of interest to her.

3. *Questioner Coverage and Responder Precision F_1 measure*: This F_1 measure attempts to balance the trade-off between the question coverage and the responder precision (inverse load), discussed in Section 3.2.2, to arrive at a single value for comparing different algorithms.

4. *Responder Mean Average Precision*: The mean macro-averaged precision for responders, averaged across all ranks. To deal with the large amount of churn among responders (responders who answer a few questions, then leave), only responders who answered at least 20 questions in the training dataset were considered.

5. *Best Answer Mean Reciprocal Rank*: This evaluates how high up the ranks the responder who gave the best answer was in the retrieval system's list.

The three models evaluated were, the Pseudo-Relevance Feedback (PRF) model discussed in Section 2.1.3 based on the descriptions in [99, 100], the pure multinomial model (PM) discussed in Section 3.3.1, and the extended generative (EG) model outlined in Section 3.3.3. The settings for pseudo-relevance feedback parameters is given in Table 5.1.

Table 5.2 shows the question coverage and responder load results when the top 10 ranked experts are contacted at each timestep. For all communities except for 'AskUbuntu' and 'English', the load results are in the $14-20$ range for the extended generative model, and in the $17-25$ range for the pseudo-relevance feedback approach, while the question coverage is over 30% for three of the six communities. These are reasonable values given that this is not a real run, and we have no way of knowing which questions were actually seen by users. As can be seen from the table, the extended generative model approach presented in Section 3.3.3 outperforms other approaches on both load and coverage. Noticing that the load value for the extended model outperforms the other approaches in all cases, a chi-square test was performed to test the significance of the questioner coverage improvement. The cases where the improvement was statistically significant at a 0.05 level are highlighted in bold. Thus even assuming that the responder load is equal in both cases, the

extended generative model showed a statistically significant improvement over PRF.

The responder MAP scores are shown in Table 5.4. A Fisher's randomization test as described in [134] was used to test the statistical significance of the improvement shown with the EG model at a level of 0.05, using the Average Precision results per user. The results were found to be significant in all cases, except for the English dataset. The question mean reciprocal rank (MRR) scores are shown in Table 5.3. The PM model under performs on the MRR metric. The cause of this is investigated in Figure 5.1 using data from the Mathematics community. The graph shows the cumulative number of best responders that have been matched till each rank. PM seems to do extremely well for the initial ranks, but its performance deteriorates over time. This phenomenon is reflected in the MRR metric. This may be due to its weaker model of user preference behavior, due to which it is not able to model users without extensive textual data.

Overall, the EG model out performs other approaches by a statisically significant margin for most datasets. The performance of PRF improves as the ranking cutoff is increased, as can be seen in Figure 5.1 for the Mathematics dataset. However, it still takes a fairly long time to overcome its initial disadvantage.

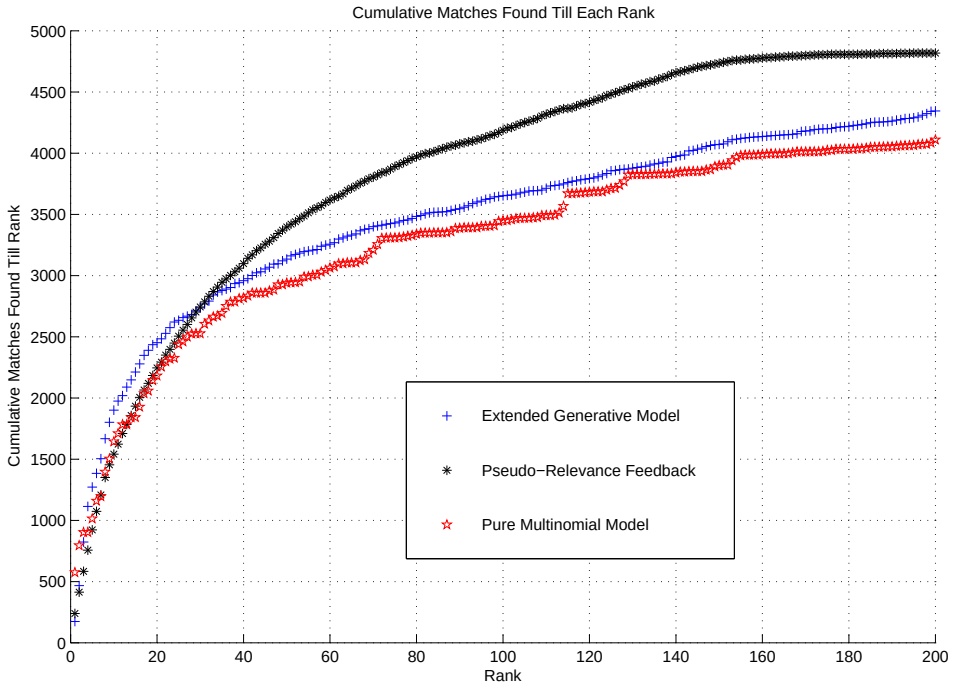

Figure 5.1: Graph showing the cumulative number of matches per rank for the StackExchange Mathematics community. The pure multinomial model performs much better in earlier ranks, while pseudo-relevance feedback outperforms others once the top 35 ranks have passed. The extended generative model is relatively consistent throughout.

Param. Label	Param. Decription	Param. Value
$\|R\|$	No. of Documents for Feedback	20
κ_R	Feedback Query Word Cutoff	10
$N_{\mathcal{R}}$	No. of Docs. Averaged Over for Users	30

Table 5.1: Parameter Settings for Pseudo-Relevance Feedback

	PRF			Pure Mult. Model			Ext. Gen. Model		
Community	Q. Cov.	R. Load	F_1	Q. Cov.	R. Load	F_1	Q. Cov.	R. Load	F_1
Math	26.3	15.64	0.102	29.3	16.39	0.1015	**35.5**	14.24	0.117
Physics	31.7	20.14	0.084	29.0	20.83	0.0824	31.5	19.55	0.088
Security	28.3	18.87	0.089	26.1	18.51	0.089	30.4	16.87	0.099
AskUbuntu	9.0	55.56	0.030	10.2	50.00	0.034	**14.2**	33.76	0.049
ServerFault	12.7	24.39	0.062	16.2	20.58	0.074	**19.1**	18.52	0.084
English	9.5	50.00	0.033	10.2	41.67	0.039	**11.3**	41.09	0.040

Table 5.2: Question Coverage (questioner recall) expressed as a percentage, Responder Load (inverse of responder precision), and Qsnr. Recall-Resp. Precision F_1 measure for six StackExchange communities, for Pseudo-Relevance Feedback (PRF), the Pure Multinomial(PM) Model, and the Extended Generative (EG) Model, with retrieval cutoff at the top 10 level. The EG model consistently outperforms the other two models on both coverage and load, and in combination in the F_1 metric. The cases where the questioner coverage improvement is statistically significant at a 0.05 level is highlighted in bold.

5.2 Authority Identification in Social Networks

5.2.1 Evaluation in the LinkedIn Social Network [34]

For authority identification in the LinkedIn social network, 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 authority-related information in each graph that is missing in the other. Given two separate graphs over which authority ranks can be calculated, a combined

Community	PRF	Pure Mult. Model	Extended Gen. Model
Math	0.076	0.102	0.121
Physics	0.129	0.069	0.126
Security	0.109	0.105	0.147
AskUbuntu	0.031	0.016	0.038
ServerFault	0.043	0.018	0.058
English	0.032	0.028	0.046

Table 5.3: The Mean Reciprocal Rank (MRR) of the best answer for the three models: pseudo-relevance feedback (PRF), the pure multinomial (PM) model, and the Extended Generative (EG) model. The Extended Generative Model outperforms the others, except for the English dataset, where PRF outperforms the other approaches by a small margin. The PM model under performs on this metric, compared to PRF. The reason for this is clearer from Figure 5.1: the PM model performs much better near the top ranks. This is because the word-based signal is effective in identifying only the top few users in any topic.

Community	PRF	Pure Mult. Model	Extended Gen. Model
Math	0.045	0.043	**0.061**
Physics	0.059	0.049	**0.084**
Security	0.047	0.049	**0.064**
AskUbuntu	0.017	0.017	**0.025**
ServerFault	0.031	0.028	**0.043**
English	0.019	0.018	0.017

Table 5.4: The Mean Average Precision (MAP) for responders with at least 20 responses each in the training dataset. By the definition of MAP, this value is macro-averaged, so all responders are weighed equally. The three models being evaluated are: pseudo-relevance feedback (PRF), the pure multinomial (PM) model, and the Extended Generative (EG) model. The EG Model outperforms the others, except for the English dataset. This may be due to the strong assumption that the EG model makes, that user responses are highly determined by the topic. In more discussion-oriented forum, this may not be the case.

rank can be arrived at in two ways:

1. *Rank Merging via Metasearch*: Use a metasearch-based approach to merge the two rankings. Borda voting [12], 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.

2. *Bimodal Authority Models*: Try to combine authority information from both graphs using a co-ranking process, as described in Section 4.5. As discussed in Section 4.4.3.1, the fair bets model is a good fit for authority estimation in professional OSNs. Thus, a natural fit for the LinkedIn graph is a fair bets based co-ranking framework.

Recall that, for a graph the fair bets score a_i of a vertex v_i, and its PageRank score r_i, can be written as:

$$a_i = \frac{r_i}{\deg^+(i)} \tag{5.1}$$

$$\Rightarrow a_i = \frac{\deg^-(v_i)}{\deg^+(v_i)} \cdot \mu_i \tag{5.2}$$

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*). Thus, the fair bets model assumes a linear relationship between a vertex's indegree and its outdegree. However, in practice, it was found that the relationship between a node's indegree and outdegree evolves as the outdegree increases. The next section discusses this relationship.

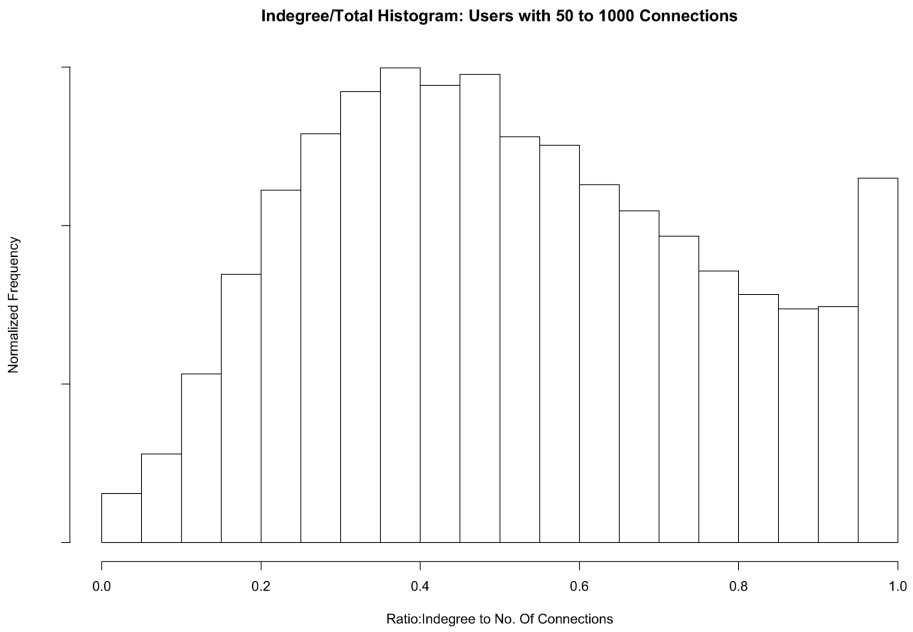

Figure 5.2: Indegree-Total Connections Ratio Histogram: Users with 50 to 1000 Connections [34]. The histogram follows an approximately normal distribution, around an indegree to total nodes ratio of 0.5. These users form the bulk of the LinkedIn social network.

5.2.1.1 Indegree Evolution with Outdegree

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 5.3. 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 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

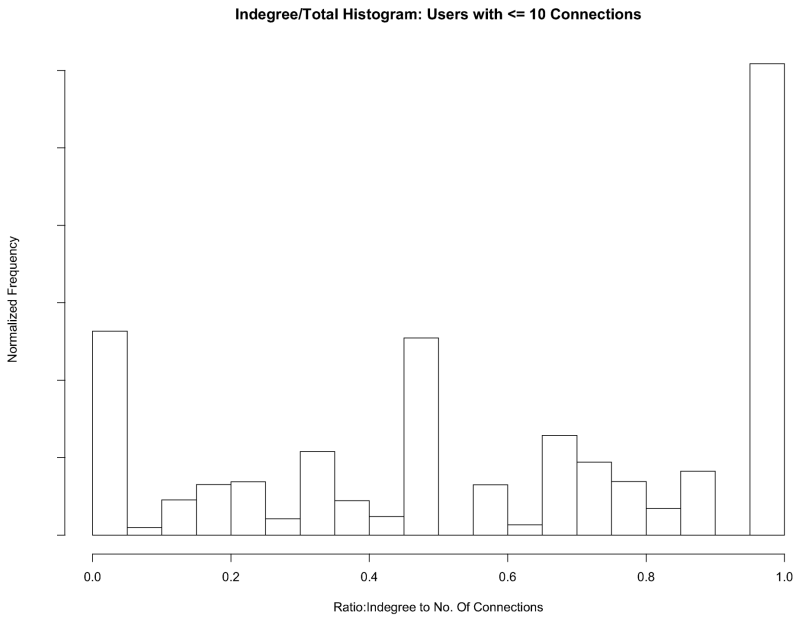

Figure 5.3: Indegree-Total Connections Ratio Histogram: Users with ≤ 10 Connections [34]. This set largely consists of new or inactive users. They have an artificially high indegree to total nodes ratio, due to their relative isolation. A subset of these nodes grow to exhibit a ratio more in line with Figure 5.2 over time.

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 5.3).

The i-t ratio for users with $50 - 1000$ connections is shown in Figure 5.2. 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 5.4. This is a very small subset of users, consisting largely of extremely active[1] and influential users. 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.

5.2.1.2 The Log Fair Bets Model [34]

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 5.5 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-

[1]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.

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:

$$f_i = \frac{\deg^-(v_i)}{\log\left(10 + \deg^+(v_i)\right)} \cdot \mu_i \tag{5.3}$$

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.

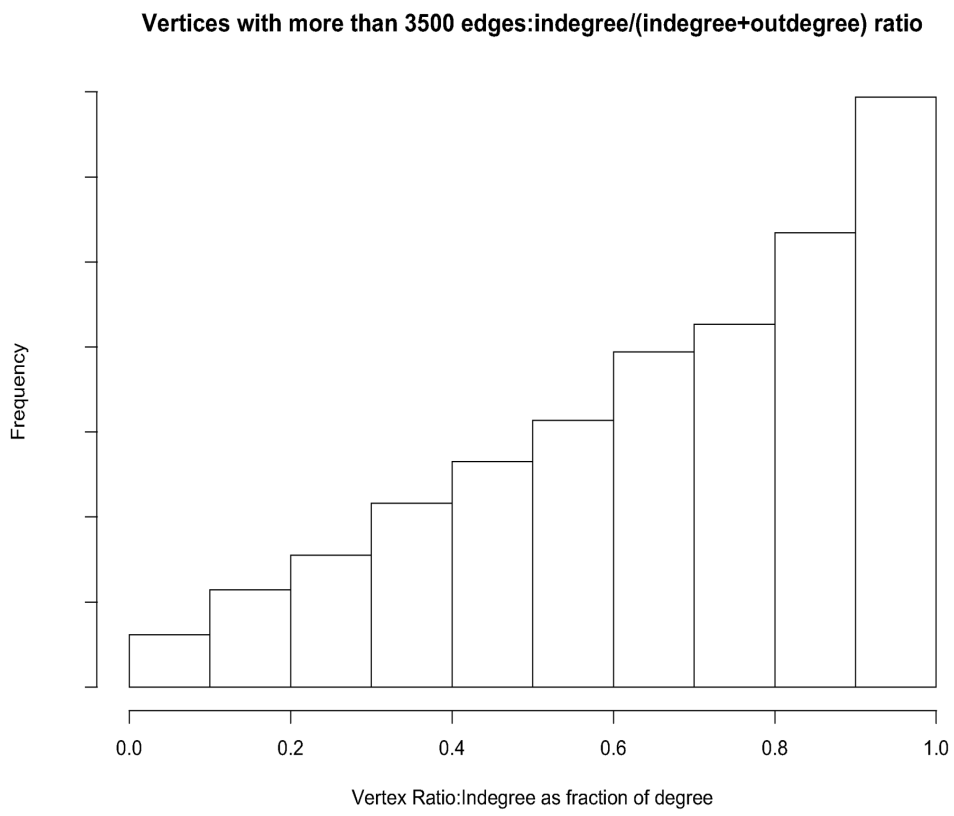

Figure 5.4: Indegree-Total Connections Ratio Histogram: Users with More Than 3500 Connections [34]. These are the outliers in the dataset. Interestingly, the PageRank algorithm ranks them near the top of the list, while the Fair Bets algorithm ranks them near the bottom, due to their large outdegree (often over $400-500$). Log Fair Bets finds a balance between the two extremes.

A validation similar to that for fair bets results was done for log fair bets by comparing ranking results against standardized seniority data. Figure 5.6 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.

5.2.1.3 Evaluation Dataset Construction

The analysis was done on a subset of about 50 million LinkedIn members, chosen from the entire LinkedIn member base (of about 100 million members at the time) based on some simple criteria. 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.

As the ground truth of authoritative people, we decided to use LinkedIn users who have Wikipedia[83] 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[2] 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

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

estimate the probability of li and wi to be the same person as follows:

$$\begin{aligned} P(li = wi) \quad \propto \quad & P(li = wi | \text{Name}_{li} = \text{Name}_{wi}) \times \\ & P(li = wi | \text{Profile}_{li} \cap \text{Profile}_{wi}) \end{aligned} \tag{5.4}$$

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 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 (5.4) 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.

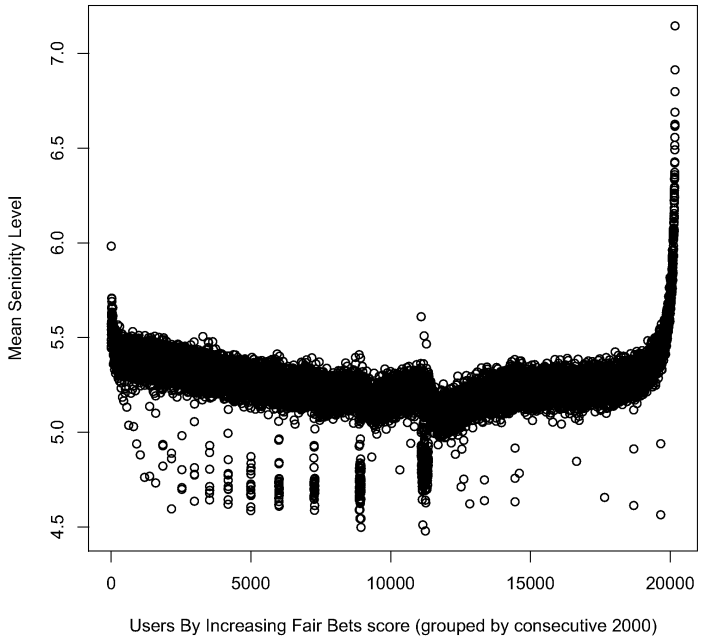

Figure 5.5: User Fair Bets Rank vs Mean Seniority Level (over consecutive groups of 2000 people)[34]. The Fair Bets model performs extremely well for the top ranks (on the right). The bottom-ranked users on the left are relatively senior but in sales and recruiting, many of whom are hyper-networkers. This is a reasonable result for the algorithm. However, in the middle, the algorithm tends to stratify by outdegree, due to its assumption of a linear relationship between indegree and outdegree. This is the reason for the repeated 'up-down' pattern. Each group in the pattern consists of people at approximately the same outdegree.

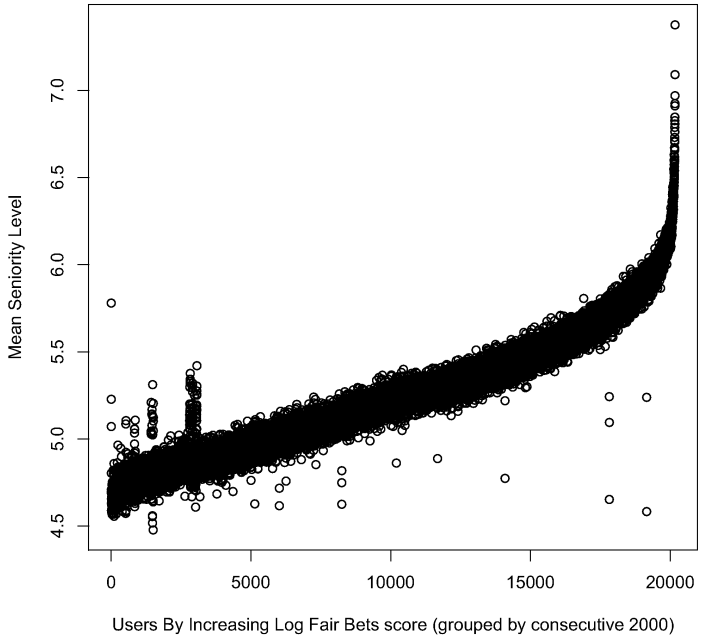

Figure 5.6: User Log Fair Bets Rank vs Mean Seniority Level (over consecutive groups of 2000 people)[34]. The log fair bets scores vary much more closely with seniority. There is a large group of low ranked relatively senior people (large spike on the left). Investigation suggested this group consists largely of sales people and recruiting professionals.

5.2.1.4 Evaluation Measures

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

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:

$$DCG = m_1 + \sum_{i=2}^{n} \frac{m_i}{\log_2 i} \tag{5.5}$$

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[3] (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. The number of levels was chosen based on discussions with subject matter experts at LinkedIn, and independent of the algorithm development.

Usually relevance scores for evaluation in the NDCG evaluation are based on explicit user feedback. A page view based approach was used due to constraints on how long the data was available. However, given the popularity of Wikipedia as a primary source of online information, and since online page views are an aggregation of the browsing preferences of millions of users, this work believes it to be a reasonable replacement.

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:

$$IDCG = \log_2 p_1 + \sum_{i=2}^{k} \frac{\log_2 p_i}{\log_2 i} \tag{5.6}$$

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 (5.5) and (5.6). Thus, a user with a relevance

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

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.

5.2.1.5 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 [89] was used as the original code base, and the code was modified to incorporate bimodal authority models. The results are shown in Table 5.6. 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 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.

	Invitation Graph		Navigation Graph	
Metric(in %)	PageRank	Log Fair Bets	PageRank	Log Fair Bets
MAP@1000	3.26	5.52(69.3%)	9.22(182.8%)	12.84(293.9%)
MAP@100K	2.45	2.53(3.2%)	1.84(−24.8%)	3.37(37.5%)
MAP@1mil	1.23	1.36(10.5%)	0.87(−29.2%)	1.44(17.1%)
NDCG@1000	1.48	3.08(108.1%)	3.99(169.6%)	6.80(359.5%)
NDCG@100K	3.84	4.48(16.7%)	3.64(−5.2%)	5.91(53.9%)
NDCG@1mil	8.30	9.13(10.0%)	7.65(−7.8%)	10.27(23.7%)

Table 5.5: MAP and NDCG Results For Invitation Graph, Navigation Graph, 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.

Metric(in %)	PageRank with Borda Voting	Bimodal PageRank	Log FB with Borda Voting	Bimodal Log FB
MAP@1000	7.55(131.6%)	12.16(273.0%)	13.03(299.7%)	13.60(317.2%)
MAP@100K	2.46(0.4%)	2.30(−6.1%)	3.76(53.4%)	3.93(61.2%)
MAP@1mil	1.27(3.2%)	1.08(-12.1%)	1.84(49.6%)	1.88(52.8%)
NDCG@1000	2.92(97.3%)	4.59(210.1%)	4.78(222.9%)	6.35(329.7%)
NDCG@100K	4.36(13.5%)	4.28(11.4%)	6.10(58.8%)	6.61(72.1%)
NDCG@1mil	8.79(5.9%)	8.73(5.18%)	11.13(34.0%)	11.75(41.6%)

Table 5.6: MAP and NDCG Results For Borda Voting and co-ranking approaches. The values in parentheses give the percentage improvement over Invitation Graph PageRank, treated as a baseline approach. The bimodal log fair best algorithm outperforms others by a wide margin.

5.2.2 Authority Identification in StackExchange

The co-ranking based authority-identification algorithm was evaluated in the Q&A context as well, using the StackExchange dataset discussed in Section 5.1. The problem was set up as follows: the co-ranking approach and other baseline algorithms were used to rank the participants in descending

Community	PageRank	Top Self-selecting	Top Best Responder	Av. Winnings
Math	40.1	42.7	40.6	44.5
Physics	43.0	40.1	39.4	47.3
Security	40.6	41.6	42.1	45.7
AskUbuntu	33.3	33.5	34.0	42.3
ServerFault	30.4	23.6	25.9	36.1
English	35.5	29.8	34.5	41.8

Table 5.7: Comparison of the PageRank-Average Winnings Co-Ranking Model in terms of 'best answer' prediction accuracy, given a list of responding users.

order of authority. Following this at each timestep, both algorithms were presented with the set of all responders for a question, and were expected to predict the responder who gave the 'best answer'. The assumption was that, if a user voluntarily decided to answer a question, if she is an expert, she should be able to get that question correct.

The co-ranking approach used two graphs: the question-answer referral model, and a tournament 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 graph. Besides the co-ranking and the PageRank algorithm, two other approaches were used for comparison: a) the top best responder, which always predicted the person among the responders who has given the most 'best answers', and b) the top self-selector: the responder who is best at selecting questions for herself, that is the one with the highest 'best answers' to answers ratio. The results are shown in Table 5.7. As can be seen, the average winnings co-ranking model outperforms other models, while the PageRank model under performs simpler approaches.

5.2.3 Incorporating Social Effects: Yahoo! Answers [32]

This section is aimed at an experimental exploration of the algorithms developed in Section 4.6 for incorporating social effects in authority identification in Q&A forums. For this purpose, two months of data (October-November 2009), for three categories was crawled from the Yahoo! Answers (YA) [85] website. The data in each category consisted of approximately $10,000$ questions, and about $15,000$ participating users.

These categories are Astronomy and Space, Books and Authors, and Wrestling. These categories were chosen to represent a broad spectrum of the type of content available on YA. After the pages were stripped of html, stopping and stemming [121] algorithms were applied to remove unimportant words and suffixes.

Table 5.8, 5.9 and 5.10 show the results of three approaches on the datasets: a cosine similarity based IR approach, a Topic-Model Based approach, and a questioner Objectivity Estimation based model. The Generative Model calculates, for each new question, the probability distribution of it belonging to each topic. It then recommends the k responders with the highest probability of responding, based on the topic-user model (ϕ) calculated by the algorithm during the training phase. The Objectivity Estimation model estimates the probability, for each questioner, that she prefers one of two components: a) an authority model, as described in Section 3.3, and b) a social influence component, based on the discovered affinity mechanism described in Section 4.6. The discovered affinity model calculates the probability of selec-

Recommender Model	Performance Measures		
	Responder Load	Questioner Coverage	Weak Coverage
IR-based (cosine)	22.41	11.87%	35.70%
Generative Topic Model	16.11	31.24%	63.83%
Mixture Model	13.68	32.72%	65.97%

Table 5.8: Astronomy and Space Dataset: Recommender Performance

Recommender Model	Performance Measures		
	Responder Load	Questioner Coverage	Weak Coverage
IR-based (cosine)	38.21	4.72%	24.72%
Generative Topic Model	22.37	16.65%	43.92%
Mixture Model	22.37	16.65%	43.92%

Table 5.9: Wrestling Dataset: Recommender Performance

tion of each responder by a questioner by assuming that, of all the responders with which the questioner has currently formed a triad, all are equally likely to be selected.

For each dataset, composed of approximately 10000 questions, the first 3500 questions were used for training, while the rest of the questions were shown chronologically to the algorithms for testing. The social influence model could be updated in real time as data became available during the test phase, but the other models did not change. The recommender performance were analyzed using the three metrics of responder load and questioner satisfaction. A simplifying assumption was made that all actual responses are satisfactory.

Recommender Model	Performance Measures		
	Responder Load	Questioner Coverage	Weak Coverage
IR-based (cosine)	37.19	1.86%	11.41%
Generative Topic Model	30.76	20.01%	30.04%
Mixture Model	27.19	20.77%	32.00%

Table 5.10: Books and Authors Dataset: Recommender Performance

Table 5.11: Correlation: Reputation and Pagerank scores vs submitter total votes on Digg

	Correlation Coefficient
Reputation Mixture Model	0.795
Pagerank	0.709

5.2.3.1 Analysis of Results

Yahoo! Answers is a noisy and unpredictable dataset. Over a set of approximately 10000 questions representing two month of activity for each category, approximately 15000 unique users participated. Of these users, for all dataset, around 45% responded to a question only once, and 40% never responded to a single question; they only asked questions. In addition, many users leave the system after a short period of time. Around 25% of the users seen during the test phase were new with no previous information available about them new users appeared in our test data set of 1000 questions. For this reason, only users who were part of at least 20 interactions (as questioners or responders) in the complete dataset were considered.

Due to these reasons, running an offline test of for a Q&A recommender is difficult. The results show that the generative and mixture model approaches

Table 5.12: Correlation: Averaged Reputation and Pagerank scores vs submitter mean votes per post on Digg

	Correlation Coefficient
Reputation Mixture Model	0.591
Pagerank	0.484

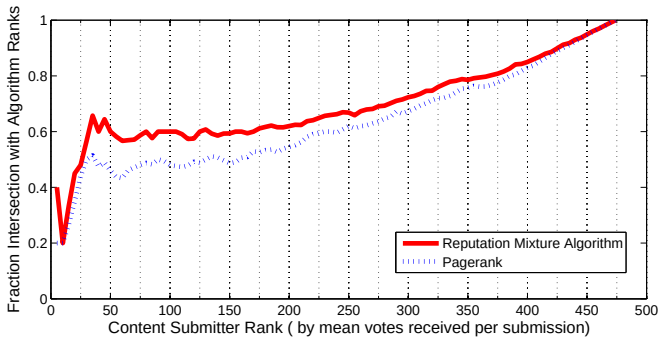

Figure 5.7: Digg Dataset: Fraction of User Ranks Predicted Correctly by Reputation algorithm and Pagerank[36]

were able to correctly recommend at least one responder more than half the time (weak coverage) while maintaining a load of $10 - 25$. This means that of every $10 - 25$ recommendations one user responded to the question. Given the bulletin board structure of YA, it is certain that responders rarely see every available question. In a live recommender test, it seems reasonable to assume a much higher percentage of responses from recommended responders because we can assume the responder sees the question and knows it has been recommended based on her expertise. Also, with a live test it would be possible to measure the questioner's satisfaction with any given response, leading to a more accurate measure of the recommender performance regarding questioner satisfaction.

The results show that the *generative model* and *mixture model* approaches outperform the information retrieval algorithm by a wide margin. Also, the multi-component approach that combines authority and behavioral models consistently outperforms the generative model, though by a smaller margin. Particularly in the case of the Wrestling dataset, the mixture model shows an improvement of 30% for coverage and 20% for weak coverage over the generative model. This results suggests that, while behavioral modeling is not essential for all forums, such as those where information seeking is emphasized, it is essential for forums with a strong social component.

5.2.4 Authority Estimation for User Generated Content: Digg [37]

Further experiments were run to test the extent to which reputation and authority identification algorithms developed for Q&A forums can be extended to other user generated content website. Digg [17] is one of the most popular content aggregators on the web, and maintains a rich, active user community and contains the necessary components for trust estimation in a content-oriented social network including: user generated content, a voting and aggregation system, and a mechanism to link users into a social network. Digg social network and endorsement data was obtained with permission from Lerman *et al.* [102].

The dataset represents one month of front page activity in 2009. For each user submitted link (story) that made it to the front page the data provides the identity of the story poster, and the identity each user who 'diggs'

the link. Additionally, for each of these users there exists asymmetric link data, indicating that a user is 'following' another, thus forming a social network. Each user has access to the activity of the users she follows, so that when a user diggs a link, all users who follow her are able to see this information. A significant portion of the votes on Digg come from this process, where users find content which their friends have endorsed, a process described as a 'cascade effect' [102]. These endorsements are driven by a mixture of two classes of motivators: similarity-based and social influence-based. Similarity-based motivation occurs when a user follows a content creator because of a preference for content by that content creator, whereas social influence-based motivation occurs when a user endorses content from a creator because of a social relationship with that creator. Because these motivations are mixed, it is difficult to identify users who submit preferred content from those who are merely socially influential.

The aim of the experiments is to test whether a mixture model based approach that attempts to model social interaction dynamics can identify users relying unfairly on their social network influence to boost their reputation. This is compared to a pagerank [30] based approach that does not take into account any information about possible social motivations of voter endorsements (diggs). As a measure of content quality, the mean number of votes received by a user once their story is promoted to the front page is used, as a large majority of votes for a front-page story come from the website's broader audience, making it difficult to rely on social affiliations. As input to the pre-

dictor, the data used for each content creator/poster, is the voting data for each story they have posted until it receives 30 votes. This information is used to calculate the reputation of each user using the mixture-model based algorithm. Following this, the correlation of the reputation scores observed with the mean number of votes received per story for each poster is calculated, and compare this value to a naïve pagerank based approach.

Table 1 shows the correlation coefficient values of the reputation and pagerank scores of each story submitter with the total votes received by his/her stories. The correlation is high in both cases, but higher for the reputation algorithm. Table 2 compares the averaged reputation and pagerank scores (obtained by dividing reputation/pagerank scores with number of submissions) with the mean votes received per submission. This to likely to be a better measure of a content creator's quality than the aggregate number of votes, as a user can be really inconsistent in quality but still receive a large number of votes in total if she submits a large number of stories. However, in this case, the correlation is weaker. But the reputation algorithm still outperforms pagerank in terms of correlation.

To compare how well the two algorithms rank users by quality, we sorted scores provided by each of them in descending order, and compared that to a ranking of posters by mean number of votes received. The comparison is shown in Figure (5.7). The y-axis of the graph shows the fraction of users in common between the ranking of users by mean vote per submission, and the ranking generated by the algorithm. The reputation algorithm identified two of the

top five ranked contributors, while the pagerank algorithm could not identify any. However, both algorithms could identify only two of the top ten. This is responsible for the initial drop in performance of the reputation algorithm from a peak. Following this the reputation algorithm consistently outperforms pagerank. This experiment was intended to demonstrate that the algorithms developed as part of this research, for the task of expertise identification, can be extended to non-reactive COSNs as well, for identifying quality content.

Chapter 6

Conclusion

The common aspects of Q&A forums and online professional social networks were established in Chapter 1. In both kinds of online communities, users make a claim to possessing expertise in various domains, either by answering questions related to the domain, or via their profile content. These claims are then validated through endorsements from their peer group. Both kinds of forums serve as a platform for people searching for expertise: for example, a programmer looking for an answer to a question, or an organization looking for an expert. As more and more people come to rely on these forums for information, it is important that the dynamics that govern these forums be understood, so that authoritative users can be identified and encouraged to participate. Enabling experts to easily find questions of interest to them, so that they are not overloaded with irrelevant questions and lose interest, is another task addressed by this research.

This research was motivated by the following hypothesis:

> In an online community of experts, mutual expressions of absolute and relative preference can be aggregated to yield effective estimates of an expert's topics of interest, and his/her credibility

as an authority in these topics, both inside the community and in the real world.

This hypothesis was investigated for two kinds of online communities: Q&A forums and professional social networks. Two kinds of probabilistic models, generative models for topic identification, and tournament models for preferences expressed in absolute and relative form, were investigated to validate this hypothesis. This chapter summarizes the work presented till now to answer each of the research questions outlined in Chapter 1, enumerating this work's contributions.

6.1 Research Question 1: Responder Preference Aggregation for Topic Identification

The first research question that this work addressed is:

RQ1: How should information about experts' preferences among different questions, based on training data, be used to make more precise question recommendations to them in the future?

An expert's preferences are expressed via the questions she chooses to answer, as opposed to the ones she ignores. The aim of RQ1 was to focus on incorporating information provided by expert preferences, to provide her with more accurate question recommendations. This problem was addressed by this work in Chapter 3, via the development of probabilistic generative models that ex-

plicitly modeled the expert's response choices as part of the content generation process.

An important challenge addressed as RQ1 was to identify the metrics that should be used to validate the model. Historically, the expert finding problem has focused on the questioner's satisfaction, ignoring the experts' interests. This is not a viable choice for voluntary forums that rely on expert goodwill. However, this shift has not been reflected in the metrics used to evaluate retrieval quality on such forums. A contribution of this research was to identify this gap, and investigate metrics that measure expert satisfaction along with the questioner's. The experimental results presented in Chapter 3 demonstrate that generative models incorporating responder preferences outperform traditional models that do not incorporate these preferences, thus providing an answer to RQ1.

6.2 Research Question 2: User Preference Aggregation for Authority Identification via Tournament Models

The second research question addressed is given below:

> RQ2: How should information about users' absolute and relative preference for other users be aggregated to identify authoritative users in a online forum or social network?

This problem was addressed in Chapter 4, by positing that the outcome of a user interaction depends probabilistically on their relative authority. This was

equivalent to assuming that the interaction graph was generated by an underlying Bradley-Terry process [28]. This reduced the problem of estimating user authority to that of finding the underlying parameters of the Bradley-Terry model. The work also showed that this result would be equivalent to finding the fair bets score [48] for each player in the graph's tournament matrix. The relationship to Bradley-Terry models also enabled the development of a new ranking model, the *average winnings* model, more suited to ranking competing responders in Q&A forums. Both the fair bets model and the average winnings model were evaluated for ranking users by estimated authority on real world social network data.

The fair bets model was evaluated on the professional OSN LinkedIn [46] (Section 5.2.1), where it outperformed the PageRank model, treated as the standard preference aggregation approach. The average winnings model was evaluated on data from the Q&A forum StackExchange [78] (Section 5.2.2). The second evaluation was done by testing the performance of the algorithm for the task of identifying who would provide the 'best answer' provided all the responders were known in advance. This was assumed to be a fair metric of user authority: as all responders voluntarily chose to provide answers to the question, so it can reasonably be believed that they saw the question as a fair test of their skills. Using this metric, the average winning model was shown to outperform currently used measures of authority such as PageRank.

6.3 Research Question 3: Combining Multiple Endorsement Graphs for Authority Identification

The third research question was:

> RQ3: How can multiple signals of user preference in professional social networks be combined to yield an effective consensus ranking of members by external authority?

Often user preferences on an OSN are often expressed in many different modes. For example, a user may send another user an invitation, an action that can be interpreted as an endorsement. Another action that might count as an endorsement is viewing another person's profile multiple times. On the Twitter [82] social network, 'following' a user profile, or 'retweeting' one of their messages is often interpreted as an endorsement of the user. Often these actions have complementary meanings, and their graphs contain complementary information. This work makes the distinction between assymetric actions such as 'following' or viewing a profile, which are more aspirational in nature, and an invitation to connect, which is more likely between peers. Each of these actions can be represented as a separate endorsement or preference graph.

This work presents an approach to combining information from multiple such complementary graphs, using a mutually re-enforcing process, where the authority scores on one graph are used as the random restart vector of the other graph, and vice versa. A scalable analogue to this process is also presented, by showing that this process is equivalent to calculating the PageRank

vector of a specially constructed composite matrix. This result is then extended to generate a new set of bimodal co-ranking models using different underlying tournament models such as fair bets, average winning, etc. The bimodal fair bets algorithm is evaluated on the LinkedIn social network, where it outperforms existing approaches by a wide margin.

6.4 Summary

As online communities gain prominence as information sources, to ensure their credibility and reliability, it is essential to develop mechanisms that identify authoritative individuals and encourage their participation. The past few years have seen a lot of interest in the problem of identifying influential users in social networks [68, 92, 138]. The idea of influence is usually defined operationally in most of this research: that is, influence is defined in terms of the procedure to measure it. As a result, this work is not in a position to provide a principled understanding of phenomena such as the frequent disconnect between online and offline reputation [93], and the 'million follower fallacy' [41], where users with millions of online followers, have almost no influence when an alternate measure is used.

This research takes a different approach, by starting with a conceptual definition, of *authority*, defined as an intrinsic quality of a node that governs its interactions on an OSN. This understanding is empirically validated via experiments on real world social networks, and Q&A communities. In this model, the influence of a node in an OSN graph is seen as an emergent property

of the node's authority, and its interactions with other nodes. This model paves the way to a better understanding of the nature of influence in online networks, and how it can be manipulated via increased activity, or by using social norms to one's advantage. This mapping of online influence to authority is an essential tool for clarity in a world where online information is being used more and more to make judgements about the real world.

6.5 Future Work

This section describes some interesting problems and directions for further development of the research presented here.

Online Evaluation of Question Recommendation: The evaluation of topic identification and question recommendation models introduced in this work has been offline, due to the logistic difficulties in setting up a large scale online evaluation. The nature of offline evaluation means that the hypothesis that lower responder load would lead to greater expert participation, while intuitively reasonable, has not been empirically verified. An online evaluation that verifies the efficacy of the algorithms presented here would be an important extension of this work.

Improved models of indegree-outdegree ratio: Authority strength of a node, as defined in this work, can be divided into two components: the mean strength of an incoming edge, and the indegree to outdegree ratio. This

work presents a basic analysis of how indegree grows with outdegree in an OSN graph, and approximates the relationship as the indegree growing exponentially with the outdegree. Further analysis, either theoretical, as has been done for Web graphs [14, 21], simulation-based, or empirical, may result in better models for the ratio, and as a result better estimates of authority.

Using authority identification to inform 'gamification' in Q&A Forums: Online Q&A forums such as Stack Exchange [80] rely on 'gamification' [52] inspired approaches, such as badges, etc., to incentivize user participation. Stack Exchange offers around eighty badges to users, based on various criteria [79]. Currently the design and usage of badges is more of an art form, though there have been recent attempts to rigorously study their impact on users [7]. Assuming that Q&A forums are interested in incentivizing authoritative users, it would be interesting to study the degree to which such users are successful in attaining different badges, and which badges are the best predictors of authority.

Extending co-ranking to affiliation networks: The co-ranking model presented in this work is restricted to scenarios where there is an exact one-to-one correspondence between nodes across multiple graphs. An important extension of the work would be to more general cases of bipartite and multipartite graphs such as affiliation networks [98]. Affiliation networks in social network analysis are bipartite graphs, where one set of nodes consist of indi-

viduals, while the second set consists of organizations or groups. Co-ranking across such graphs would enable authority estimation algorithms to take into account information such as organizations/institutions belonged to, events attended, etc., which will result in a more comprehensive ranking.

Bibliography

[1] Text Retrieval Conference. `http://trec.nist.gov/`.

[2] Lada A. Adamic, Jun Zhang, Eytan Bakshy, and M.S. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. *Proceeding of the 17th international conference on World Wide Web*, pages 665–674, 2008.

[3] Nitin Agarwal, Huan Liu, Lei Tang, and Philip S. Yu. Identifying the influential bloggers in a community. *Proceedings of the International Conference on Web search and Web Data Mining - WSDM '08*, page 207, 2008.

[4] Charu C. Aggarwal and ChengXiang Zhai. A survey of text clustering algorithms. In *Mining Text Data*, pages 77–128. Springer, 2012.

[5] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. Finding high-quality content in social media. *Proceedings of the international conference on Web search and web data mining - WSDM '08*, page 183, 2008.

[6] Eugene Agichtein, E Gabrilovich, and H Zha. The Social Future of Web Search: Modeling, Exploiting, and Searching Collaboratively Generated

Content. *IEEE Data Engineering Bulletin, Volume 32, Issue 2, p.52-61 (2009)*, pages 1–10, 2009.

[7] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. Steering user behavior with badges. *Proceedings of the 22nd international Conference on World Wide Web - WWW '13 (to appear).*

[8] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. Discovering value from community activity on focused question answering sites: a case study of stack overflow. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 850–858. ACM, 2012.

[9] Chris Anderson. *The Long Tail: Why the Future of Business Is Selling Less of More.* Hyperion, 2006.

[10] Isabel Anger and Christian Kittl. Measuring influence on twitter. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, page 31. ACM, 2011.

[11] J. Antin and E. F. Churchill. Badges in Social Media: A Social Psychological Perspective. In *ACM conference on Human Factors in Computing Systems 2011, ACM, Vancouver, BC*, 2011.

[12] Javed A. Aslam and Mark 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.

[13] Ricardo Baeza-Yates and Berthier Ribiero-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[14] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[15] John Battelle. Wondir Launches Revamped Site, 2005. `http://battellemedia.com/archives/2005/04/wondir_launches_revamped_site.php`.

[16] Pavel Berkhin. A survey on pagerank computing. *Internet Mathematics*, 2(1):73–120, 2005.

[17] Betaworks. Digg. `http://www.digg.com`.

[18] Jiang Bian, Yandong Liu, Ding Zhou, Eugene Agichtein, and Hongyuan Zha. Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In *Proceedings of the 18th international conference on World wide web*, pages 51–60. ACM, 2009.

[19] Monica Bianchini, Marco Gori, and Franco Scarselli. Inside pagerank. *ACM Transactions on Internet*, 5(1):92–128, 2005.

[20] Ginestra Bianconi and Albert-László Barabási. Bose-einstein condensation in complex networks. *Physical Review Letters*, 86(24):5632–5635, 2001.

[21] Ginestra Bianconi and Albert-László Barabási. Competition and multiscaling in evolving networks. *EPL (Europhysics Letters)*, 54(4):436, 2001.

[22] Christopher M Bishop and Others. *Pattern recognition and machine learning*. Springer New York, 2006.

[23] Grant Blank and Bianca C. Reisdorf. The Participatory Web: A User Perspective on Web 2.0. *Information, Communication and Society*, 15(4):537–554, June 2012.

[24] D.M. Blei, Andrew Y. Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

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

[26] Ralph Allan Bradley. Rank analysis of incomplete block designs: Ii. additional tables for the method of paired comparisons. *Biometrika*, 41(3/4):502–537, 1954.

[27] Ralph Allan Bradley. Rank Analysis of Incomplete Block Designs: III Some Large-Sample Results on Estimation and Power for a Method of Paired Comparisons. *Biometrika*, pages 502–537, 1954.

[28] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, pages 324–345, 1952.

[29] Felix Brandt and Felix Fischer. Pagerank as a weak tournament solution. In *Internet and Network Economics*, pages 300–305. Springer, 2007.

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

[31] Suratna Budalakoti. An extended generative model for expert finding in question-answer forums. Technical report, University of Texas at Austin, 2013.

[32] Suratna Budalakoti and K Suzanne Barber. Authority vs affinity: Modeling user intent in expert finding. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 371–378. IEEE, 2010.

[33] Suratna Budalakoti and K Suzanne Barber. Tournament based reputation models for aggregating relative preferences. In *Trust in Agent Societies, International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2013.

[34] Suratna Budalakoti and Ron 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.

[35] Suratna Budalakoti, David DeAngelis, and K Suzanne Barber. Expertise modeling and recommendation in online question and answer forums. In *Social Computing (SocialCom), 2009 IEEE Second International Conference on*. IEEE, 2009.

[36] Suratna Budalakoti, David DeAngelis, and K Suzanne Barber. Unbiased trust estimation in content-oriented social networks. In *Trust in Agent Societies, International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2011.

[37] Suratna Budalakoti, David DeAngelis, and K Suzanne Barber. Unbiased trust estimation in content-oriented social networks. *Trust in Agent Societies (Trust-2011), International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2011.

[38] Moira Burke, Cameron Marlow, and Thomas Lento. Feed me: motivating newcomer contribution in social network sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 945–954. ACM, 2009.

[39] W. J. Carl. What's All The buzz about?: Everyday Communication and

the Relational Basis of Word-of-Mouth and Buzz Marketing Practices. *Management Communication Quarterly*, 19(4):601–634, May 2006.

[40] Carlos Castillo and Brian D Davison. *Adversarial web search*, volume 4. Now Publishers Inc, 2011.

[41] Meeyoung Cha, H Haddadi, F Benevenuto, and K.P. 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.

[42] Prasad Chebolu and P. Melsted. PageRank and the random surfer model. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1010–1018, 2008.

[43] Xi Chen, Paul N Bennett, Kevyn Collins-Thompson, and Eric 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.

[44] Alice Cheng and Eric Friedman. Manipulability of pagerank under sybil strategies. In *Proceedings of the First Workshop on the Economics of Networked Systems(NetEcon06)*. NetEcon, 2006.

[45] Junghoo Cho, Sourashis Roy, and Robert 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.

[46] LinkedIn Corp. Linkedin. `www.linkedin.com`.

[47] LinkedIn Corp. Linkedin: Skill endorsements overview. `www.slideshare.net/linkedin/introducing-linkedin-endorsements`.

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

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

[50] David DeAngelis. *Encouraging Expert Participation in Online Communities*. PhD thesis, University of Texas at Austin, 2011.

[51] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):pp. 1–38, 1977.

[52] Sebastian Deterding, Miguel Sicart, Lennart Nacke, Kenton O'Hara, and Dan Dixon. Gamification: using game-design elements in non-gaming contexts. In *PART 2 – Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, CHI EA '11, pages 2425–2428, New York, NY, USA, 2011. ACM.

[53] Gideon Dror, Yehuda Koren, Yoelle Maarek, and Idan Szpektor. I want to answer; who has a question?: Yahoo! answers recommender system. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1109–1117. ACM, 2011.

[54] Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank aggregation methods for the Web. *Proceedings of the tenth international conference on World Wide Web - WWW '01*, pages 613–622, 2001.

[55] Benjamin Edelman. Earnings and Ratings At Google Answers. *Economic Inquiry*, 50(2):309–320, April 2012.

[56] N. Eggemann and S. D. Noble. The clustering coefficient of a scale-free random graph. *Discrete Appl. Math.*, 159:953–965, June 2011.

[57] Epinions. Epinions. `http://www.epinions.com`.

[58] Hui Fang and ChengXiang Zhai. Probabilistic models for expert finding. In *Advances in Information Retrieval*, pages 418–430. Springer, 2007.

[59] Dániel Fogaras, Balázs Rácz, Károly Csalogány, and Tamás Sarlós. Towards scaling fully personalized pagerank: Algorithms, lower bounds, and experiments. *Internet Mathematics*, 2(3), 2005.

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

[61] Yupeng Fu, Wei Yu, Yize Li, Yiqun Liu, Min Zhang, and Shaoping Ma. Thuir at trec 2005: Enterprise track. *Proceedings of the Fourteenth Text Retrieval Conference (TREC 2005)*, 2005.

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

[63] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2003.

[64] Arpita Ghosh and Preston McAfee. Incentivizing high-quality user-generated content. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 137–146, New York, NY, USA, 2011. ACM.

[65] Rumi Ghosh and Kristina Lerman. Predicting influential users in online social networks. In *Proceedings of KDD workshop on Social Network Analysis (SNA-KDD)*, July 2010.

[66] David F Gleich, Paul G Constantine, Abraham D Flaxman, and Asela Gunawardana. Tracking the random surfer: empirically measured teleportation parameters in pagerank. In *Proceedings of the 19th international conference on World wide web*, pages 381–390. ACM, 2010.

[67] Hasrat Godil. *Finding Experts by Modeling Domain Expertise*. Master's thesis, University of Texas at Austin, 2006.

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

[69] Geoffrey R Grimmett and David R Stirzaker. *Probability and random processes*. Oxford University Press, USA, 2001. Chapter 6.

[70] Jinwen Guo, Shengliang Xu, Shenghua Bao, and Yong Yu. Tapping on the potential of Q&A community by recommending answer providers. In *Proceeding of the 17th ACM conference on Information and knowledge management*, pages 921–930. ACM, 2008.

[71] Lei Guo, Enhua Tan, Songqing Chen, Xiaodong Zhang, and Yihong (Eric) Zhao. Analyzing patterns of user content generation in online social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, page 369, New York, New York, USA, 2009. ACM Press.

[72] F. Maxwell Harper, Daniel Moy, and Joseph A Konstan. Facts or friends?: distinguishing informational and conversational questions in social Q&A sites. In *Proceedings of the 27th international conference on Human factors in computing systems*, 2009.

[73] Taher Haveliwala, Sepandar Kamvar, and Glen Jeh. An analytical comparison of approaches to personalizing pagerank. Technical Report 2003-35, Stanford InfoLab, June 2003.

[74] Brent Hecht, Jaime Teevan, Meredith Ringel Morris, and Dan Liebling. Searchbuddies: Bringing search engines into the conversation. *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2012.

[75] Damon Horowitz and Sepandar D. Kamvar. The anatomy of a large-scale social search engine. *Proceedings of the 19th international conference on World wide web - WWW '10*, page 431, 2010.

[76] Zan Huang. Link prediction based on graph topology: The predictive value of the generalized clustering coefficient. In *Workshop on Link Analysis: Dynamics and Static of Large Networks (LinkKDD2006)*, 2006.

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

[78] Stack Exchange Inc. Stack Exchange Data Explorer. `http://data.stackexchange.com/`.

[79] Stack Exchange Inc. Stack Exchange List of Badges. `http://stackoverflow.com/badges`.

[80] Stack Exchange Inc. The Stack Exchange Network. `www.stackexchange.com`.

[81] Stack Exchange Inc. StackExchange Blog: Career 2.0 Launches, 2011. `http://blog.stackoverflow.com/2011/02/careers-2-0-launches/`.

[82] Twitter Inc. Twitter. http://www.twitter.com.

[83] Wikipedia Inc. Wikipedia, 2010. www.wikipedia.org.

[84] Yahoo! Inc. Flickr. http://www.flickr.com.

[85] Yahoo! Inc. Yahoo! answers. http://answers.yahoo.com.

[86] Yahoo! Inc. Yahoo! answers hits 200 million visitors, worldwide!, December 2009. http://yanswersblog.com/index.php/archives/2009/12/14/\\yahoo-answers-hits-200-million-visitors-worldwide/.

[87] Jian Jiao, Jun Yan, Haibei Zhao, and Weiguo Fan. Expertrank: An expert user ranking algorithm in online communities. In *New Trends in Information and Service Science, 2009. NISS'09. International Conference on*, pages 674–679. IEEE, 2009.

[88] Pawel Jurczyk and Eugene Agichtein. Hits on question answer portals: exploration of link analysis for author ranking. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 845–846. ACM, 2007.

[89] U Kang, Charalampos E Tsourakakis, and Christos Faloutsos. Pegasus: mining peta-scale graphs. *Knowledge and information systems*, 27(2):303–325, 2011.

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

[91] John G Kemeny and James Laurie Snell. *Finite markov chains*. Springer-Verlag New York, 1976.

[92] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[93] Alexy Khrabrov and G. Cybenko. Discovering Influence in Communication Networks Using Dynamic Graph Analysis. In *IEEE International Conference on Social Computing*, pages 288–294. IEEE, 2010.

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

[95] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

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

[97] R. David Lankes. Credibility on the internet: shifting from authority to reliability. *Journal of Documentation*, 64(5):667–686, 2008.

[98] Silvio Lattanzi and D Sivakumar. Affiliation networks. In *Proceedings of the 41st annual ACM symposium on Theory of computing*, pages 427–434. ACM, 2009.

[99] Victor Lavrenko and W. Bruce Croft. Relevance based language models. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '01*, pages 120–127, 2001.

[100] Matthew Lease. Incorporating Relevance and Psuedo-relevance Feedback in the Markov Random Field Model: Brown at the TREC'08 Relevance Feedback Track. In *Proceedings of the 17th Text Retrieval Conference (TREC'08)*, 2009.

[101] J.G. 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.

[102] Kristina Lerman and Rumi Ghosh. Information contagion: an empirical study of spread of news on digg and twitter social networks. In *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*, May 2010.

[103] Kristina Lerman and Laurie Jones. Social browsing on flickr. In *Proceedings of 1st International Conference on Weblogs and Social Media (ICWSM-07)*, 2007.

[104] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew 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.

[105] Jure Leskovec, Daniel 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.

[106] Jonathan Levin and Barry Nalebuff. An Introduction To Vote Counting Schemes. *Journal of Economic Perspectives*, 9(1):3–26, 1995.

[107] Mingrong Liu, Yicen Liu, and Qing Yang. Predicting best answerers for new questions in community question answering. In *Proceedings of the 11th international conference on Web-age information management*, WAIM'10, pages 127–138, Berlin, Heidelberg, 2010. Springer-Verlag.

[108] Xiaoyong Liu and W. Bruce Croft. Cluster-based retrieval using language models. *Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR '04*, page 186, 2004.

[109] Xiaoyong Liu, W. Bruce Croft, and Matthew Koll. Finding experts in community-based question-answering services. *Proceedings of the 14th ACM international conference on Information and knowledge management - CIKM '05*, page 315, 2005.

[110] Y. Liu, Bin Gao, T.Y. 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.

[111] J. Lussier, Troy Raeder, and N. Chawla. User Generated Content Consumption and Social Networking in Knowledge-Sharing OSNs. *Advances in Social Computing*, pages 228–237, 2010.

[112] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

[113] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

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

[115] M E J Newman, D J Watts, and S H Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99 Suppl 1:2566–72, February 2002.

[116] Sewoong Oh, Sahand Negahban, and Devavrat Shah. Iterative Ranking from Pair-wise Comparisons. *Advances in Neural Information Process-*

ing Systems 25: 26th Annual Conference on Neural Information Processing Systems, pages 1–27, 2012.

[117] T OReilly. What is Web 2.0: Design patterns and business models for the next generation of software. *Communications & strategies*, (1):17–37, 2007.

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

[119] SA Paul, Lichan Hong, and EH Chi. Who is Authoritative? Understanding Reputation Mechanisms in Quora. *arXiv preprint arXiv:1204.3724*, 2012.

[120] Jay M Ponte and W Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281. ACM, 1998.

[121] M. F. Porter. *An algorithm for suffix stripping*, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[122] Mingcheng Qu, Guang Qiu, Xiaofei He, Cheng Zhang, Hao Wu, Jiajun Bu, and Chun Chen. Probabilistic question recommendation for question answering communities. In *Proceedings of the 18th international*

conference on World wide web - WWW '09, number 2, page 1229, New York, New York, USA, 2009. ACM Press.

[123] Quora. Quora. `www.quora.com`.

[124] Fatemeh Riahi, Zainab Zolaktaf, Mahdi Shafiei, and Evangelos Milios. Finding expert users in community question answering. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, pages 791–798, New York, NY, USA, 2012. ACM.

[125] Soo Young Rieh and Nicholas J Belkin. Understanding Judgment of Information Quality and Cognitive Authority in the WWW. *Journal of the American Society for Information Science and Technology*, 53(2):145–161, 2002.

[126] Stephen E Robertson. The probability ranking principle in ir. *Journal of documentation*, 33(4):294–304, 1977.

[127] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004.

[128] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[129] David Shenk. *Data Smog: Surviving The Information Glut*. HarperOne, 1997.

[130] Amit Singhal. Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4):35–42, 2001.

[131] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 21–29. ACM, 1996.

[132] Giora Slutzki and Oscar Volij. Ranking participants in generalized tournaments. *International Journal of Game Theory*, 33(2):255–270, 2005.

[133] Giora Slutzki and Oscar Volij. Scoring of web pages and tournaments – axiomatizations. *Social choice and welfare*, 26(1):75–92, 2006.

[134] Mark D Smucker, James Allan, and Ben Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 623–632. ACM, 2007.

[135] Ian Soboroff, AP de Vries, and Nick Craswell. Overview of the trec 2006 enterprise track. *TREC 2006 Working Notes*, 2006.

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

[137] M.R. Subramani and Balaji Rajagopalan. Knowledge-sharing and influence in online social networks via viral marketing. *Communications of the ACM*, 46(12):307, 2003.

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

[139] Kristine Frisnfeldt Thuesen. *Analysis of Ranked Preference Data*. PhD thesis, Technical University of Denmark, 2007.

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

[141] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*. Cambridge University Press, 1994.

[142] C. Nadine Wathen and Jacquelyn Burkell. Believe it or not: Factors influencing credibility on the Web. *Journal of the American Society for Information Science and Technology*, 53(2):134–144, 2002.

[143] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.

[144] Howard T. Welser, Eric Gleave, Vladimir Barash, Marc Smith, and Jessica Meckes. Whither the experts? social affordances and the cultivation of experts in community q&a systems. In *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04*, CSE '09, pages 450–455, Washington, DC, USA, 2009. IEEE Computer Society.

[145] Jianshu Weng, E.P. Lim, Jing Jiang, and Qi 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.

[146] J Xu and WB Croft. Query expansion using local and global document analysis. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–11, 1996.

[147] Xing Yi and James Allan. A comparative study of utilizing topic models for information retrieval. *Advances in Information Retrieval*, pages 29–41, 2009.

[148] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. *ACM SIGIR Conference on Research and Development in Information Retrieval*, page 403, 2001.

[149] Chengxiang Zhai and John Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 403–410. ACM, 2001.

[150] Jun Zhang, Mark S. Ackerman, Lada Adamic, and Kevin Kyung Nam. Qume: a mechanism to support expertise finding in online help-seeking communities. In *Proceedings of the 20th annual ACM symposium on User interface software and technology*, UIST '07, pages 111–114, New York, NY, USA, 2007. ACM.

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

[152] Ding Zhou, Sergey A Orshanskiy, Hongyuan Zha, and C Lee Giles. Co-ranking authors and documents in a heterogeneous network. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 739–744. IEEE, 2007.

Vita

Suratna Budalakoti was born in Gorakhpur, India, to his parents Shailendra Budhalakoti and Nivedita Budhlakoti. He completed his secondary education at Army School, Ambala Cantt., India, in 1997. In 2001, Suratna earned the degree of Bachelor of Engineering in Computer Science at the National Institute of Technology, Rourkela, India. He started at the Ph.D. program in Electrical and Computer Engineering at the University of Texas at Austin in 2006, where he received the Master of Science in Engineering degree in 2009.

Permanent address: 7117 Wood Hollow Dr Apt 1916
Austin, Texas 78731

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.