# Automatic Evaluation of Information Provider Reliability and Expertise

Konstantinos Pelechrinis
School of Information Sciences
University of Pittsburgh
Pittsburgh, PA
*kpele@pitt.edu*

Vladimir Zadorozhny
School of Information Sciences
University of Pittsburgh
Pittsburgh, PA
*vladimir@sis.pitt.edu*

Velin Kounev
School of Information Sciences
University of Pittsburgh
Pittsburgh, PA
*vkounev@pitt.edu*

Vladimir Oleshchuk
Department of Information & Communication Technology
University of Agder
Grimstad, Norway
*vladimir.oleshchuk@uia.no*

Mohd Anwar
Department of Computer Science
NC A&T State University
Greensboro, NC
*manwar@ncat.edu*

Yiling Lin
School of Information Sciences
University of Pittsburgh
Pittsburgh, PA
*yil54@pitt.edu*

## ABSTRACT

Q&A social media have gained a lot of attention during the recent years. People rely on these sites to obtain information due to a number of advantages they offer as compared to conventional sources of knowledge (e.g., asynchronous and convenient access). However, for the same question one may find highly contradicting answers, causing an ambiguity with respect to the correct information. This can be attributed to the presence of unreliable and/or non-expert users. These two attributes (reliability and expertise) significantly affect the quality of the answer/information provided. We present a novel approach for estimating these user's characteristics relying on human cognitive traits. In brief, we propose each user to monitor the activity of his peers (on the basis of responses to questions asked by him) and observe their compliance with predefined cognitive models. These observations lead to local assessments that can be further fused to obtain a reliability and expertise consensus for every other user in the social network (SN). For the aggregation part we use subjective logic. To the best of our knowledge this is the first study of this kind in the context of Q&A SNs. Our proposed approach is highly distributed; each user can individually estimate the expertise and the reliability of his peers using his *direct* interactions with them and our framework. The online SN (OSN), which can be considered as a distributed database, performs continuous data aggregation for users expertise and reliability assesment in order to reach a consensus. In our evaluations, we first emulate a Q&A SN to examine various performance aspects of our algorithm (e.g., convergence time, responsiveness etc.). Our evaluations indicate that it can accurately assess the reliability and the expertise of a user with a small number of samples and can successfully react to the latter's behavior change, provided that the cognitive traits hold in practice. Furthermore, the use of the consensus operator for the aggregation of multiple opinions on a specific user, reduces the uncertainty with regards to the final assessment. However, as real data obtained from Yahoo! Answers imply, the pairwise interactions between specific users are limited. Hence, we consider the aggregate set of questions as posted from the system itself and we assess the expertise and realibility of users based on their response behavior. We observe, that users have different behaviors depending on the level at which we are observing them. In particular, while their activity is focused on a few general categories, yielding them reliable, their microscopic (within general category) activity is highly scattered.

## 1. INTRODUCTION

During the last decade, advancements in computing and networking have drastically changed the way people acquire information. For example, printed sources of information and knowledge (e.g., scientific magazines, books etc.) are being supplanted by digital media, while functions of traditional libraries are being taken over by online digital libraries and search engines. In OSNs, users might seek for help in specific topics from their peers. As an example, members of the Yahoo! Answers network can post a specific question, and the rest of the users are free to provide answers. The same is possible via the most popular OSN to date, Facebook, which has introduced a new feature called “Questions”. For quick answers, such online forums, Q&A SNs, online tutoring, etc., have the advantages of being asynchronous, often without requiring face-to-face communications, and in general being more convenient.

What is common in all these situations, is the lack of vetting of these modern sources of information for their quality, correctness and accuracy, among other characteristics. For instance, in the physical world, an oculist is an eponymous source, that has been recognized as an *authority* on eye diseases. The same holds for a book that is used in a reputed medical school to train doctors; its usage in the medical school automatically attaches to it the status of infallibility. On the contrary, it is clear that for information provided by an online source, the same property does not hold. In social psychology studies, people have been found to place a higher trust on information provided from sources classified as authorities [1], even though the classification (e.g., book used in university) itself is subjective. In [2], a study with a diverse set of human participants on how they search for and appraise medical information, it was found that a “professional look" of a web site made it appear to be more authoritative. Improper banner ads affected the credibility of the site. Nevertheless, an unscrutinized source can still be preferable to humans if it is easy to access and convenient. Studies have shown that individuals may rely on less trustworthy but more accessible sources to obtain the information they need risking though the accuracy of the information itself [3]. This, increases the possibilities that their search results are inadequate or less reflective and the information obtained to be flawed.

For instance, Shachaf [4], performs content analysis of 1,522 transactions from Yahoo! Answers, Wiki Answers, Askville, and the Wikipedia Reference Desk. The goal of this study is to identify reliable answers (in terms of accuracy, completeness and verifiability) in these online information social media. The findings of this study reveal that the most popular Q&A site as captured from the number of users, questions and answers (i.e., Yahoo! Answers) provides the least accurate, complete, and verifiable information. Furthermore, there is a significant difference in answer quality among these sites. Hence, identifying high quality content is crucial in a Q&A SN.

Our position is that the reputation[1] and the expertise of the answer *provider* has a direct impact on the quality of the information obtained. As we will discuss in later section, there exist studies that try to assess these characteristics of a user individually in a Q&A SN. However, in this paper we take a novel direction by solely utilizing the human behavioral patterns. The main *fact* our scheme is based on is the **inability of a person to know everything about anything**. In other words, expertise is context dependent; Bob is a highly reliable person and an excellent Java programmer and can (with high probability) correctly answer any question with regards to this topic. However, he will not be able to answer questions about heart diseases even if he is willing to provide truthful information. Of course, depending on the *contextual distance* between two topics, there might be a correlation between the expertise values on each of them. For instance, a Java programmer might be expected to be able to answer to questions for other programming languages as well. We further study this important issue, using data from Yahoo! Answers. For now we will assume that the topics considered have a large contextual distance, that is, they are completely disjoint. Therefore, there is no correlation between the expertise attribute on these topics.

Every question posted in Q&A SN is related with a specific topic (e.g., “Java programming”, “Soccer”, etc.). Each user (e.g., Alice) keeps track of every other user’s (say Jack) activity per category with the help of the ***response matrix*** (to be defined in the following). This monitoring is **local**, in the sense that it captures the interactions between Alice and Jack. In other words, the response matrix includes information about the *reactions* of Jack on Alice’s questions. Statistical metrics that capture the compliance/deviation of Jack’s behavior with the expected profile are then defined. Their computation enables Alice to update her belief on Jack’s expertise and reliability. The social network as a whole (or even just a subset of users) can further aggregate using subjective logic, the individual/local opinions on Jack’s expertise and reliability and obtain a global opinion for his characteristics. The main advantages of our assessment system are its lightweight nature and the fact that can be applied *locally* from every user individually. The contribution of our work can be summarized in the following:

- Design of a human cognition-based, lightweight framework for simultaneously assessing the reliability and expertise of a user in a Q&A SN. Alice can use this framework to obtain a subjective opinion on Bob based on their interactions.
- Integration of our framework with subjective logic to acquire a consensus for Bob’s attributes and reduce the uncertainty that accompanies the local assessements.
- Study of the applicability of our assessment scheme to real Q&A social networks. Utilizing data gathered from Yahoo! Answers, we study the pairwise interactions between real users as well as their macro- and micro-scopic activity with regards to topic granularity. In brief, we find that the same user can appear to be both reliable and unreliable, when considering his activity with regards to general and more specific topics respectively.

The rest of the paper is organized as follows. Section 2 provides a simple example illustrating the basic idea of our approach. Section 3 discusses related studies. Our cognitive-based assessment scheme is presented in Section 4. Section 5 presents our evaluations, while Section 6 discusses the scope and limitations of our work. Section 7 concludes our study.

## 2. SYSTEM MODEL

Consider a simple scenario with two users, Bob and Jack, replying to each others questions about various topics. For our example we consider three topics of interest: “Football”, “Medicine” and “Programming”. Our objective is to enable each user to judge the *quality* of the information obtained from any other user. Assume that Bob received some information from Jack related to “Medicine”. Intuitively, the quality of this information is tightly related with (1) the knowledge of Jack about “Medicine”, and (2) the reputation of Jack. However, it would be unrealistic to assume that there is a globally consistent view of Jack from all the users of the system. Achieving global consensus in such judgments is problematic even in relatively small user communities, and it is practically impossible in large scale social networks. Instead, we propose to estimate (1) subjective opinion of Bob about Jack’s knowledge of “Medicine” and (2) subjective opinion of Bob about Jack’s reputation and then fuse them using subjective logic. As these opinions propagate via the data communication network they can be combined to reflect overall user reliability and expertise with high confidence.

In this work we introduce a scalable and automatic way to assess individual opinions as well as further fuse those opinions along information propagation routes. We utilize cognitive principles of human reactions to requests for information. If a user tends to respond consistently to questions related to a particular topic, we consider him knowledgeable in that area. Meanwhile, if the user is willing to reply to many remotely related topics, it would be safer to assume that this person is an amateur in each of those areas and his replies should be treated as less reliable. We formally capture these behavioral patterns by maintaining pairwise user views of each other in the form of **response matrices (RM)**. Columns of a response matrix correspond to topics of interests, while rows reflect history of user responses. To reiterate, the contextual distance of the topics considered is important. We consider only completely disjoint topics (e.g., “Medicine” and “Programming”). Even though this might be possible for factual categories, it might be harder for non-factual ones (e.g., “Travel”). We will come back to this important aspect in Section 5.

[1] In the following we will use the terms reputation and reliability interchangeably.

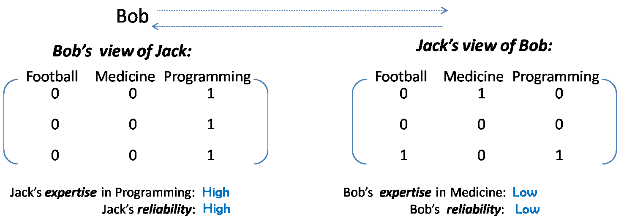

Bob
Bob's view of Jack:
Football Medicine Programming
0 0 1
0 0 1
0 0 1
Jack's expertise in Programming: High
Jack's reliability: High
Jack's view of Bob:
Football Medicine Programming
0 1 0
0 0 0
1 0 1
Bob's expertise in Medicine: Low
Bob's reliability: Low

**Figure 1: Example of Response Matrices reflecting high and low opinions**

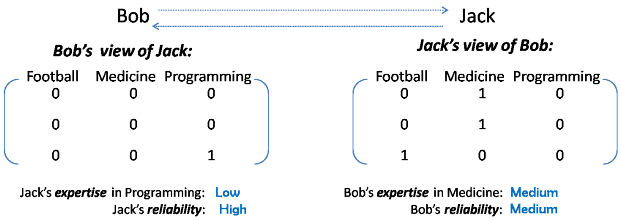

Bob
Jack
Bob's view of Jack:
Football Medicine Programming
0 0 0
0 0 0
0 0 1
Jack's expertise in Programming: Low
Jack's reliability: High
Jack's view of Bob:
Football Medicine Programming
0 1 0
0 1 0
1 0 0
Bob's expertise in Medicine: Medium
Bob's reliability: Medium

**Figure 2: Example of Response Matrices reflecting high, low and medium opinions.**

Figure 1 shows an example of two response matrices reflecting views of Bob for Jack and vice versa. In this example, Bob has posted 3 questions on each category and the same is true for Jack. For each one of Jack's questions, he assigns the value of '1' in the corresponding matrix element, if Bob replied to it; otherwise, he inputs '0'. Similar steps are followed from Bob when obtaining Jack's response matrix. In the example provided, Bob has a high opinion about knowledge of Jack in "Programming" since Jack's replies are consistently focused on this topic; Bob's opinion about Jack's reliability is also high, since Jack's responses are not spread over various remote topics. Meanwhile Jack has low opinion about Bob's knowledge in "Medicine", as well as Bob's reliability.

To sum up, user's overall reliability is reflected through spread of 1s over rows of the RM, while user's expertise in particular topics is represented as density of 1s in the corresponding columns. Figure 2 illustrates another scenario where user Bob has medium opinion about Jack and his knowledge of "Medicine". Obviously, Bob has a low opinion about knowledge of Jack in "Programming". Meanwhile Bob has a high opinion about reliability of Jack, since responses of Jack are not scattered over remotely related topics. In Section 4 we formalize our approach building on this example.

Figure 3 represents the general structure of information propagation and data fusion in a Q&A OSN. Individual users' opinions about their peers are continuously generated using dynamically updated (independent) response matrices. The network will utilize collective intelligence to assess a consensus reliability and expertise of the users. Subjective (local) opinions are generated and propagated automatically without explicit involvement of users. For this purpose we do not require users to evaluate quality of responses from their peers.

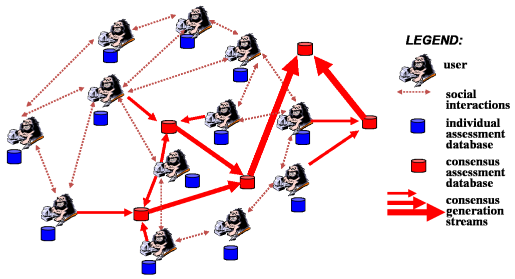

LEGEND:
user
social interactions
individual assessment database
consensus assessment database
consensus generation streams

**Figure 3: Distributed propagation and fusion o finformation about users reliabliltiy and expertise.**

## 3. RELATED WORK

In this section we will briefly discuss existing work on reputations systems and expertise inference.

**Reputation systems:** Reputation models have been primarily considered in the context of online electronic markets. Users of each specific market rate each other, and a centralized authority computes the trust value (reliability) on every single entity [5]. These computations are mainly based on simple statistics acquired from users' feedback (e.g., positive and negative feedback). Sabater *et al.* [6] design the regret system. They describe their scheme using an example borrowed from an online marketplace and they show how their system exploits the social relations among the different users. In brief, the reliability that a user (say Bob) has on any of his peers (say Jack) is based on their direct interactions as well as the interactions of *witnesses* (say Alice) with Jack and their social relation with him. Huynh *et al.* [7] further introduce the notion of certified reputation. If Bob has no interaction with Jack and he cannot find any witness to report reputation information for Jack, Jack can present certified information about his past performance. These are essentially references from other agents who have interacted with Jack. Certified reputation is very useful for open multi-agent systems, where user can leave and join the system arbitrarily in time. Wang and Singh [8] [9] follow a more rigorous approach, building on the notion of the *probability of the probability* of outcomes [10]. In particular, they use the triple of belief, disbelief and uncertainty along with different statistical measures to formally capture the trust on an agent. The same authors in [11], borrow ideas from the generalized transitive closure literature, and in particular from path algebra, to introduce two operators for propagating trust through a multi-agent system in a distributed way. This approach is in stark contrast with the centralized reputation/trust systems presented in [12] [13]. Hang *et al.* [14] further introduce a third operator that can handle cycles/dependent paths. The interested reader can find additional reputation systems in [15] [16] [17] [18] [19] [20] [21] [22] [23] [24].

**Expertise inference:** There exist studies in the literature that try to assess the expertise metric. *Referral* systems or expert finders (e.g., [25] [26] [27]) try to locate people who are most appropriate for providing the requested information. For instance, Guo *et al* [28] propose a topic-based model for finding appropriate question answerers. By discovering latent topics in the content of questions and answers as well as latent interests of users to build user profiles, they recommend a ranked list of users who are more likely to answer a new question. Similar systems account only for the expertise of an information provider, not considering her willingness to help (which is related with her reliability). For instance, ReferralWeb [29] exploits the social network within a community to identify a set of experts with regards to the information requested. It levarages the "six degrees of separation" phenomena, which manifests a small distance between two individuals in a network. Hence, one can exploit these social relations to find an expert. Nevertheless, the flexibility of similar systems is low mainly due to two reasons: (i) only the expertise of an information provider is accounted for, not considering her willingness to help (which is related with her reliability), and (ii) only binary decisions are made with respect to a user being an expert or not. However, in the majority of the situations users have some measure of expertise, thus, emerging the need to quantify the level of this expertise. Zhang *et al* [30] make a step further and not only they identify *expert* users in an online Java forum, but they also evaluate algorithms that rank these experts. They use a centralized approach that leverages social network analysis tools considering the network graph structure. ExpertRank (the core algorithm of Hermes system) [31] utilizes the main features of the PageRank algorithm [32], which ranks web pages based on their *popularity* on specific topics as seen from Web users. In our case, that of expertise ranking, it is not only imporant to know how many answers on a specific topic Jack has posted but also to whose questions he has replied. We should put less weight to answers provided to Alice who is a *newbie* as compared to asnwers provided to Eve who herself has some level of expertise. Other studies that are based on centralized graph mining algorithms and leverage social relationships can be found in [33], [34], and [35]. Nevertheless, all of them either provide binary classification (i.e., Jack is an expert or not) or they provide a relative ranking among the users, without revealing enough information for the actual expertise of the user.

Recently, Kasneci *et al* [36] designed a knowledge corroboration system for Semantic Web called CoBayes. In particular, they build a bayesian-based system that assesses the truthfulness of statements extracted from various sites. The system outsources the corroboration task to a set of assessors, whose expertise is also under question. The authors' evaluations demonstrate the applicability of their approach. However, their work is in a different context (that of semantic web and knowledge corroboration) and under the assumption that users who assess the trurth of the statemenets are indeed reliable.

**Q&A Social Networks:** Related with our work, two different types of studies on question & answers social networks exists. The first of them targets at the ***quality of the actual content of the answers***. Agichtein *et al* [37] exploit features of social media that are intuitively correlated with quality. They train a "quality" classifier to appropriately select and weight the features for each specific type of item. Three types of features are used as input to the classifier: intrinsic content quality (e.g., punctuation & typos, syntactic and semantic complexity, grammaticality), user relationships (e.g., who has answered a question from whom), and usage statistics (number of clicks on the item, dwell time, etc.). Bian et al. [38] apply a mutually reinforcing approach to learn the question and answer reputation of a user, as well as the quality of his questions and answers. They use a semi-supervised mutual reinforcement framework for calculating content quality and user reputation in Community Question Answer (CQA) systems. The same authors present a general ranking framework for factual information retrieval from social media [39] based on user interactions and community-based features. They perform content-based quality assessment without considering answerer expertise and experiment on factoid questions.

Shah and Pomerantz [40] propose a number of criteria (e.g., completeness, readability, relevance) to evaluate and predict the quality of online answers. They validate their criteria by asking Amazon Mechanical Turk workers to rate answers to some selected well-answered questions in Yahoo! Answers. With the assumption that answers are tied to the questions with various types of latent links, Tu *et al.* [41] propose an analogical reasoning-based approach which learns to measure the analogy between the new question and answer linkages. John *et al.* [42] develop a quality framework comprising social, textual, and content-appraisal features of user-generated answers in CQA services. In their study, logistic-regression analysis shows that content-appraisal features (such as comprehensiveness, truthfulness, and practicality) are the strongest predictor of quality.

Another set of studies are ***centered around the users of a Q&A system***. For example, Bouguessa *et al.* [43] propose a model to identify authoritative users based on the number of best answers provided by them. A "best answer" is selected either by the asker or by other users via a voting procedure. Jurczyk and Agichtein [44] adapt the HITS algorithm to the user interactions graph of Yahoo! Answers to discover authorities, and show a positive correlation between authority calculated with the HITS algorithm and answer quality. Golbeck and Fleischmann [45] examine the role of expertise cues in text and photo on users' trust in answers in social Q&A. Their results indicate that expertise cues in text lead to significantly higher trust among both experts and non-experts. However, expertise cues in photos increase trust among non-experts only. Suryanto *et al.* [46] propose several methods to derive quality answers using the question dependent and question independent expertise of answerers in addition to the use of answer features. Their experiment shows that expertise-based methods yield better answer quality than answer feature-based methods. Finally, Pal and Konstan [47] study the question selection bias of an answerer, that is, which questions a user would select for answering. Based on the studies, experts prefer answering questions where they have a higher chance of making a valuable contribution.

Recent literature has focused on utilizing social network and data analysis to study problems related with answering behavior and information quality in Q& A systems. Panovich *et al.* [48] studied the role of tie strength in answering questions. Through user studies they found that answers from users with whom the questioner shares stronger ties provide slightly more information than those from people with weaker ties. Furthermore, Wang *et al.* [49], studying Quora they found that people who contributed more and provide higher quality answers, tend to have more followers. These well-connected users also gain advantage by having more friends (followers) to answer their questions and upvote for their answers.

**Distinguishing our work:** Reputation systems are only interested into estimating the reliability of a network user, ignoring the context dependencies. In addition, most of these schemes are focused on different types of networks making it hard to directly apply them in the area of Q&A SNs. On the contrary, expert finder systems are focused on identifying a set of users able to reply a specific question, neglecting most of the times both the general reputation of a user as well as her *absolute* expertise. For instance, Alice might be a wonderful doctor to her regular patients but her offhand medical advice might not be completely trustful as she is not know to be entirely forthcoming. Furthermore, there are significant differencies between the architecture of our approach and that of the existing schemes. For instance, reputation systems are mainly based on feedback acquired from the users. In addition, studies that are more focused on Q&A systems, and in particular on the information quality, require the involvement of complicated functionalities (e.g., content analysis). In contrast, our approach does not require any explicit involvement from the users as mentioned in Section 2 and it is based on cognitive models for human behavior. In particular, it requires only monitoring of the users' activities, interactions and communication patterns. Most importantly, each user can apply our framework locally to obtain a subjective view of any other peer, without requiring the knowledge of the network graph structure or that of the underlying social relations[2]. To the best of our knowledge, *to date there exist no work in the literature that tries to exploit cognitive and behavioral characteristics of humans towards the* ***joint*** *estimation of a user's reliability and expertise in a Q&A social network.*

## 4. ASSESSMENT SCHEME

In this section we will present our scheme which estimates the reliability $r_i$ of user $i$ (say Jack) and his expertise $e_{i,q}$ on queries of type $q$ (say "Football"). For our presentation we build on the example of Section 2.

### 4.1 Individual estimation

**Response matrix (RM):** The participating users of a Q&A SN can be both consumers of information, as well as providers. When a consumer Bob asks a query he obtains responses directly from multiple providers (e.g., Jack). Goal of the SN is to assess the quantities $r_{Jack}$ and $e_{Jack,q}$ $\forall q \in Q$, where $Q$ is the set of different topics (in our case $Q = \{\text{``}Football\text{''}, \text{``}Medicine\text{''}, \text{``}Programming\text{''}\}$). Bob can obtain locally a *subjective* opinion about Jack's (i) reliability and (ii) expertise in $q$. He can further augment this opinion using the subjective logic consensus operator to combine views of other users (e.g., Alice) about Jack [50]. Ideally the SN can monitor all of these interactions and collect all these subjective opinions, to efficiently approximate an *objective* value for $r_{Jack}$ and $e_{Jack,q}$.

The first step is for Bob to derive the RM for Jack, $M^{Bob}_{Jack} \in \Pi^{w \times n}$; $\Pi^{w \times n}$ is the set of $w \times n$ matrices, $w$ is the number of questions per category considered (e.g., posted from Bob) during the time period $T_{RM}$ over which the matrix is calculated and $n$ is the number of different topics. For ease of presentation we assume that Bob posts the same number of questions (that is $w$) for every one of the $n$ different categories. In our example we have $w = n = 3$. Note here that, there is no actual correspondence between the actual time and the rows except that the queries were made within the time interval $T_{RM}$ corresponding to the RM. Thus, multiple "ones" in a row simply imply responses obtained to multiple queries in different topics within $T_{RM}$. A single RM can be thought of as a single snapshot of the network (with respect to Jack's activity as per Bob's view). As time elapses there are more questions posted and more snapshots for the network created. Hence, for the purposes of our study time is *measured* with regards to the number of snapshots that we have for the Q&A SN.

Before providing the details of our estimation scheme we would like to emphasize on the fact that even though our criteria are based on widely accepted human cognitive traits (that are mainly accepted as *common sense*), there are studies in the literature that support our models. For instance, Adamic *et al.* [51] analyze the activity of 41,266 active users of Yahoo! Answers. The authors find that there is a corelation between the interest entropy of a user and the "best answer" votes he obtains. In particular, users with lower entropy (e.g., users whose answers span few topics only) obtain a larger number of votes. Similar findings (e.g., question selection bias of answerer [47]) support our reliable user model; if a user tends to respond consistently to questions related to particular topic, then he is knowledgable in that topic.

**Assessment of $e^{Bob}_{Jack,\text{``}Football\text{''}}$:** The expertise of Jack is tightly related with a *specialization*. An expert on one topic is expected to be rather engaged on the related questions. Thus being *consistently* active is a sign of expertise in the corresponding category [30]. For this task Bob will use the column of $M^{Bob}_{Jack}$ that corresponds to "Football" (let it be column $j$). Column $j$ is a vector, denoted by $\overrightarrow{\Lambda}^{Bob,j}_{Jack}(t) \in \Re^{w \times 1}$, of 0s and 1s. $\overrightarrow{\Lambda}^{Bob,j}_{Jack}(t)$ can be thought as an observation vector. Its $h^{th}$element, denoted by $[\lambda_h(t)]^{Bob,j}_{Jack}$, is equal to 1 if Jack responded to the $h^{th}$ "Football" question in the snapshot $t$, otherwise it is 0. Since we currently do not consider the appropriate of the answer, we just *measure* the interest of Jack on "Football" through his active participation in the corresponding discussions; this can roughly capture his *tendency* for expertise in the field. A spammer, or a person who just posts noisy answers, can be thus falsely considered to be an expert on "Football". Later, in Section 5, we will describe scenarios where expertise is falsely inferred and how we can mitigate these occurencies.

Each one of the questions in a snapshot can be thought as a Bernoulli trial $X$. The trial is successful if Jack responds. Thus, assuming Jack is not a spammer, the probability of success $p$ of $X$ is equal to Jack's expertise on "Football", which we assume to be constant throughout the snapshot. In random variables terminology, the outcome of the $h^{th}$ trial $[\lambda_h(t)]^{Bob,j}_{Jack}$, is 0 if Jack did not respond to the $h^{th}$ "Football" question, and 1 otherwise. Therefore, the pdf of $X$ is:

$$f_h(X = \lambda_h) = p^{\lambda_h} \cdot (1-p)^{1-\lambda_h} \tag{1}$$

By replacing $p$ with $e^{Bob}_{Jack,\text{``}Football\text{''}}$, the probability density function described by Equation 1 can be thought as the formal definition of Jack's expertise. Given the expertise sample set we have collected, we use the MLE framework to obtain an estimate on parameter $p$. In particular, this estimate corresponds to the solution of the following optimization problem:

$$\max_{p} \quad \frac{1}{w} \cdot \sum_{i=1}^{w} log(f_i(\lambda_i|p)) \quad \text{subject to} \;\; p \in [0,1] \tag{2}$$

Considering one snapshot/RM of the network at time $t$ provides Bob with a single sample set. Thus by solving the MLE problem he acquires a single point estimate $\widetilde{p}(t)$. In order to compute the uncertainty on the expertise value with respect to Jack, we propose the use of $m$ snapshots in time, which will provide $m$ sample sets. Using the estimates computed from MLE for each of the above sets, Bob can compute the average estimator $\overline{\widetilde{p}}$ and its standard deviation $\widetilde{p_{sd}}$. In turn, this provides a method to obtain an expertise interval $E$ of width $\widetilde{p_{sd}}$, centered at $\overline{\widetilde{p}}$. Using an interval, rather than a single point value, allows us to capture the uncertainty embedded in the expertise estimation.

**Assessment of $r^{Bob}_{Jack}$:** Reliability is a personality trait, related with the "good will" of a user. Given its highly subjective nature, there are no clear metrics of Jack's reliability. However, as aforementioned, a reliable person (within our context) can be *roughly profiled* as follows:

1. Given that Jack cannot be an expert in a large variety of different topics, he is expected to reply to a few topics. This

[2] This is possible under the assumption of enough pairwise interactions between specific pairs of users. We discuss this issue later in the paper.

translates to the matrix $M^{Bob}_{Jack}(t)$ of a reliable person being dominated by 0s.

2. Reliable Jack is expected to consistently reply to the topics of his interest/expertise. This translates to the matrix $M^{Bob}_{Jack}(t)$ having a *minimum* number of '1' entries.

Using the above profile we can formally define the $r^{Bob}_{Jack}$. Let $R_1$ be the number of '1' entries in $M^{Bob}_{Jack}(t)$. With $\delta_{xy}$ being Kronecker's delta, $R_1 = \sum_{i=1}^{w}\sum_{j=1}^{n}\delta_{[m_{ij}]^{Bob}_{Jack},1}$. Furthermore, let vector $\overrightarrow{\Pi}^{Bob}_{Jack} = [\pi_j]^{Bob}_{Jack} = [\sum_{i=1}^{w}\delta_{[m_{ij}]^{Bob}_{Jack},1}]^{Bob}_{Jack}$. Each element of $\overrightarrow{\Pi}^{Bob}_{Jack}$ is the number of Jack's replies in each query category. Finally, let $R_2$ be the number of *modes* in the sample set $\overrightarrow{\Pi}^{Bob}_{Jack}$ (see Appendix). Then Jack is considered *reliable*, that is $r^{Bob}_{Jack} = 1$, iff:

$$\alpha \leq R_1 \leq \beta \quad \wedge \quad R_2 \leq \gamma \tag{3}$$

Here $\alpha$, $\beta$ and $\gamma$ are functions of the dimensions of $M^{Bob}_{Jack}$ $(w,\ n)$. When the first part of (3) does not hold, we need to penalize Jack. For example, if $R_1 < \alpha$, Jack can be thought as acting *selfishly*; not providing any answers at all (even at the topics of his expertise)[3]. In this case, Bob panalizes Jack based on (i) the deviation of $R_1$ from its lower bound, that is $d_1 = \alpha - R_1$, as well as (ii) the deviation of $R_2$ from $\gamma$ ($d_2 = \gamma - R_2$):

$$\begin{aligned} r^{Bob}_{Jack} &= y_1 \cdot (1 - \frac{1}{\alpha} \cdot (\alpha - R_1)) + y_2 \cdot (1 - \frac{1}{\gamma} \cdot (\gamma - R_2)) \\ &= y_1 \cdot \frac{R_1}{\alpha} + y_2 \cdot \frac{\gamma}{R_2},\ \ y_1 + y_2 = 1 \end{aligned} \tag{4}$$

If $R_1 > \beta$, Jack is unreliable. He is *talkative* and simply provides answers in many areas where he has no background. This can lead to the difusion of low quality information In this case, Bob penalizes Jack based on the (i) the deviation of $R_1$ from its upper bound, that is $d_3 = R_2 - \beta$, as well as (ii) the deviation of $R_2$ from $\gamma$ ($d_2 = \gamma - R_2$):

$$\begin{aligned} r^{Bob}_{Jack} &= x_1 \cdot \frac{((w \cdot n - \beta) - (R_1 - \beta))}{(w \cdot n - \beta)} + x_2 \cdot (1 - \frac{1}{\gamma} \cdot (\gamma - R_2)) \\ &= x_1 \cdot \frac{(w \cdot n - R_1)}{(w \cdot n - \beta)} + x_2 \cdot \frac{\gamma}{R_2},\ \ x_1 + x_2 = 1 \end{aligned} \tag{5}$$

Note here that, the coefficients $y_1$, $y_2$, $x_1$ and $x_2$, can also be functions of $R_1$ and/or $R_2$. For instance, when $R_1 < \alpha$, it might be the case that the number of modes present (i.e., $R_2$) is within the limit of $\gamma$. In this case, we should not use $d_2$ (which is negative!) to penalize Jack, since he adheres to the expected behavior. Therefore,

$$y_1 = \begin{cases} \rho_y & \text{if } R_2 > \gamma \\ 1 & \text{otherwise} \end{cases} \tag{6}$$

$$y_2 = \begin{cases} 1 - \rho_y & \text{if } R_2 > \gamma \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Similar definitions can be given for $x_1$ and $x_2$, controlled by a different parameter $\rho_x$ ($0 \leq \rho_x, \rho_y \leq 1$).

However, even if the $R_1$ is kept below $\beta$ it might be the case that this happens not because Jack focuses on his topics of expertise but because he is very little engaged to replying (spreading his low activity across a number of topics). Thus the right part of (3) needs to hold as well. In this case, Bob reduces the reliability of Jack based on the number of excessive modes present ($d_2 = \gamma - R_2$):

$$r^{Bob}_{Jack} = \begin{cases} 0 & \text{if } R_2 = w \\ \dfrac{\gamma}{R_2} & \text{otherwise} \end{cases} \tag{8}$$

## 4.2 Consensus assessment

By executing the above process, Bob has obtained a subjective view of Jack. The next natural step would be for Bob to combine different subjective opinions of Jack (e.g., that of Alice) in order to obtain a more *objective* opinion for him. The same is true for the SN as a whole; a central authority can gather all these local opinions and fuse them towards obtaining a consensus for every user. We use subjective logic consensus operators for this task. The consensus operator not only allows us to fuse the opinions on expertise and reliability of users, but it also reduces the uncertainty accompanied with the individual opinions.

In subjective logic, opinions are represented by triplets. Let $t$, $d$ and $u$ be non-negative values such that $t + d + u = 1$,$\{t, d, u\} \in [0, 1]^3$. Then the triple $\omega = \{t, d, u\}$ is called an *opinion*, where components $t$, $d$ and $u$ represent levels of trust, distrust and uncertainty. For example, high distrust with some uncertainty (0.1) could be expressed as an opinion $\omega_1 = \{0.0, 0.9, 0.1\}$, while high trust with a minor uncertainty of 0.04 could be expressed as opinion $\omega_2 = \{0.96, 0.00, 0.04\}$. In our case we have opinions for both Jack's reliability and his expertise on each different category (after deriving the triplets from the corresponding intervals as described in the following). Let $\omega_p^{Bob}$ and $\omega_p^{Alice}$ be two opinions of entities *Bob* and *Jack* about statement *p* (e.g., *p* can be Jack's reliability). Then their combined consensus opinion is defined as:

$$\omega_p^{Bob,Jack} = \omega_p^{Bob} \oplus \omega_p^{Jack} = \left\{t_p^{Bob,Jack}, d_p^{Bob,Jack}, u_p^{Bob,Jack}\right\} \tag{9}$$

where $t_p^{Bob,Jack} = \left(t_p^{Bob}u_p^{Jack} + t_p^{Jack}u_p^{Bob}\right)/k$, $u_p^{Bob,Jack} = \left(u_p^{Bob}u_p^{Jack}\right)/k$, $d_p^{Bob,Jack} = \left(d_p^{Bob}u_p^{Jack} + d_p^{Jack}u_p^{Bob}\right)/k$, and $k = \left(u_p^{Bob} + u_p^{Jack} - u_p^{Bob}u_p^{Jack}\right)$.

**Deriving opinions from the response matrices:** In order to be able to use subjective logic for consensus estimation we need to map the reliability and expertise intervals obtained locally from Bob and Alice about Jack into opinions.

Assuming that $r^{Bob}_{Jack} = [a, b]$ we generate the subjective logic opinions using the following equation (likewise, a mapping can be designed for the expertise opinion triplet $\omega^{Bob}_{Jack,``Football''}$):

[3] Of course, Jack might have no expertise at all and therefore, if reliable, he will exhibit extremely low activity. As discussed later in the current work we are not interested into distinguishing between a selfish user and a non-expert.

$$\omega^{Bob}_{Jack} = \{\frac{a+b}{2}, 1-\frac{a+b}{2}-\frac{b-a}{2}, \frac{b-a}{2}\} \tag{10}$$

# 5. EVALUATIONS

In this section we present our evaluation set up and results.

## 5.1 Experimental Setup

In the first part of our evaluations we create synthetic data. In order to obtain the RMs we emulate the behavior of an information provider. In our study we are primarily interested into identifying 4 categories of users; "Reliable expert", "Talkative expert", "Reliable amateur" and "Talkative amateur". The names are self explanatory but to give an example, a "Talkative expert" is someone who is a *real* expert on a few topics (as expected), but she is also replying to questions outside her specialization. On the contrary, a provider can be classified as "Reliable amateur" if she does not have any expertise in reality (something which can also be common) and is sincere enough not to provide any uncertain answers to any category. We would like to emphasize on the fact that we make the implicit assumption that providers are not *selfish*, and thus, a real expert will always reply to questions that fall into her specialization [30]. Otherwise it will be extremely hard, if possible at all, to distinguish between a "Selfish expert" and a "Reliable amateur".

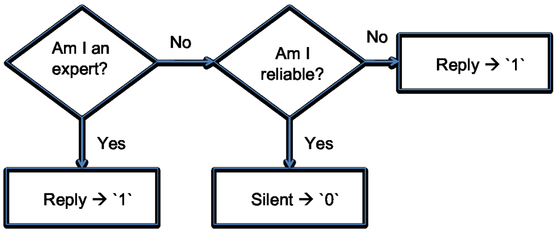


**Figure 4: Flow diagram of our user model.**

Every user in the network has an *a priori* fixed expertise on each topic (expertise vector) and a reliability value. Every element on the expertise vector as well as the a priori reliability lay in the interval $[0, 1]$. For instance, a "Reliable expert", would have a high a priori reliability value (i.e., close to 1), while his expertise will be high (close to 1) for the topics of expertise and low (close to 0) for the rest. The number of topics of expertise are sampled at random from a uniform distribution over $\{1, 2, ..., \gamma\}$, while the actual topics are picked at random. In order to construct/emulate the response matrices we use the process depicted at Figure 4. In particular, we run this process iteratively for every question emulated on every topic. In order to decide upon every decision block (e.g., "Am I an expert¿'), we further sample a uniform distribution over $[0, 1]$ and compare the sampled point with the corresponding a priori value (expertise or reliability) of the user under consideration. Furthermore, unless otherwise stated, the values of the simulation parameters used are shown in Table 1. Finally, in our experiments that involve dynamic behaviors, the notion of time is not tightly related with the absolute time (e.g., seconds). A *jiffy/time ticks* is equal to a full RM snapshot. In other words, time $t = x$, means that there exist $x$ snapshots (i.e., $x \cdot n$ questions in total) since the time we started observing the network.

We would like to emphasize on the fact that in the first part of our evaluations we are mainly interested into examining the performance of the algorithm in terms of its ability to respond to dynamic behavioral changes, converge fast etc. For this purpose, we make the assumption that users follow the cognitive traits identified earlier in this paper. In other words, we do not claim/examine the correctness of our scheme. In fact, there is no actual ground truth of these quantities for comparisson.

| w | $\alpha$ | $\beta$ | $\gamma$ | n | $\rho_x$ |
|---|---|---|---|---|---|
| 20 | 5 | 30 | 3 | 10 | 0.95 |

**Table 1: Simulation parameters.**

## 5.2 Performance under static users' behavior

Our first set of experiments focuses on scenarios where users adhere to a static behavior. For instance, a reliable user remains so throughout the whole emulation period.

**Recovering the real expertise/reliability:** Initially we opt to examine the accuracy of the individual assessment scheme. We consider a set of 10 users who we *monitor*. After obtaining the corresponding RMs, we apply our framework and obtain the corresponding opinions. We begin by examining the columns of the RMs in order to obtain an estimation for the expertise of the user with regards to each topic of interest. We then examine the structure of the whole matrix in order to assess its reliability. As one might expect, the trust value of the assessed (reliability or expertise) opinion triplet is not supposed to be exactly equal with the predefined (reliability or expertise) value. For this reason, we define some criteria in order to evaluate the quality of the estimation. Denoting the real value of the attribute (topic expertise/reliability) with $a^*$, we define to have a successful inference iff

$$a^* \in [t-u, t+u] \ \vee \ |t-a^*| \leq p \cdot a^*, \ p \in [0, 1] \tag{11}$$

The value of $p$ dictates the strictness of the convergence. Smaller values correspond to more strict convergence. In our experiments we have set $p = 0.15$, that is, the trust of the assessed opinion is at most 15% *different* than the actual value. Our results are depicted in Figure 5 where accuracy is shown for different number of snapshots used for the estimation. Accuracy is defined as the ratio of the correct inferences (based on Equation 11) over the total number of estimations. As we observe, irrespective of the number of snapshots used, our scheme is capable of indentifying the *real* reputation of all the users. Figure 6 depicts the empirical CDF for the difference between the assessed trust on reputation $t_r$ and the *real* reliability $r^*$ of the user (i.e., $t_r - r^*$). As one can see, the absolute value of this difference is always smaller than 0.05! The independence from the number of snapshots used for the estimation implies that if our cognitive model for the users holds in practice, their reliability can be restored fairly fast (i.e., small number of snapshots are required). Figure 7 depicts the (low) uncertainty $u_r$ associated with the reliability.

Despite the fact that we were able to recover the reliability for all the users, the accuracy with regards to the expertise is relatively low ($\sim 30\%$). The reason for this performance can be attributed to the fact that when applying MLE on each column of the RM, the correctness of the answer is not considered. As a result, the presence of multiple '1's in a column is considered as a sign of expertise even though it can be the result of spamming activity. In other words, a "Talkative" user will exhibit this pattern into several columns/topics (many more than the few expertise topics expected

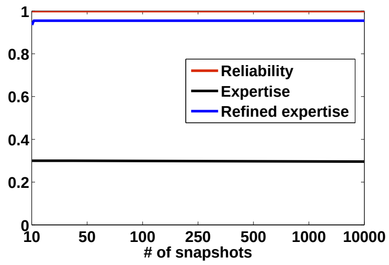


**Figure 5: Inference accuracy of our scheme.**

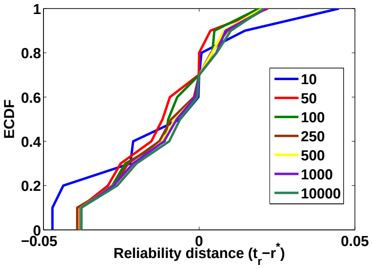


**Figure 6: Accurate reliability assessment.**

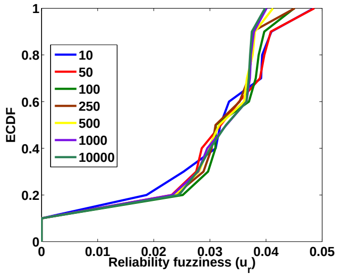


**Figure 7: Small reputation uncertainty.**

for each user). Thus, there will be an overstimation of user expertise in these topics, which results in the low accuracy. Figure 8 depicts the CDF of the difference between the trust of the expertise opinion $t_e$ and the real expertise value $e^*$ for different number of snapshot used for the estimation (i.e., $t_e - e^*$). As we can see with high probability, the infered value is much larger than the actual one. For instance, with probability greater than $40\%$ this difference is greater than 0.5. Figure 9 depicts the uncertainty $u_e$ associated with the expertise.

**Refinement phase:** The inaccurate expertise estimation can be attributed to the fact that only the column structure, and not the matrix structure, is considered for this task. In order to overcome this problem, we include a refinement phase. In brief, after using $k$ snapshots to estimate the reliability of a user (which is extremely accurate), we scale down the initial estimation of the expertise opinion (trust value) using the assessed reputation. Figure 10 illustrates the process.

Once the initial opinions for a user's (say Alice) expertise on a topic and her relibility are obtained they serve as inputs into the refinement engine, which provides a *refined* opinion for Alice's expertise, $\omega_e^{ref}$. The goal of this phase, is to scale down the expertise based on the reputation. Since reputation is estimated based on the structure of the whole matrix, it can *reduce* the instances of falsely perplexing a spammer for being an expert. In particular we use the following equation for refining the trust on the expertise:

$$t_e^{ref} = t_e \cdot t_r^2 \tag{12}$$

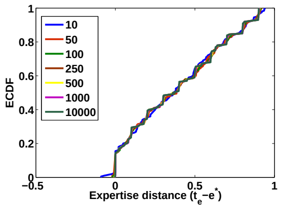


**Figure 8: Overestimating expertise.**

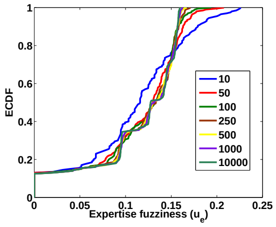


**Figure 9: Significant expertise uncertainty.**

To reiterate, when a user is less reliable, we degrade the effect of his intense activity on many topics using Equation 12. Note here that, one could possibly use another function to scale $t_e$, i.e., $t_e^{ref} = f(t_e, t_r)$. The exact shape of $f$ is essentially a design choice. For instance, if $f$ is linear to $t_r$, i.e., $t_e^{ref} = t_e \cdot t_r$, we take a more conservative approach for the reduction of $t_e$ (that is, the reduction is smaller). On the contrary, Equation 12 penalizes a user - with the same $t_r$ - more. In other words, the shape of $f$ dictates the weight we place on the reliability of a user when refining its expertise.

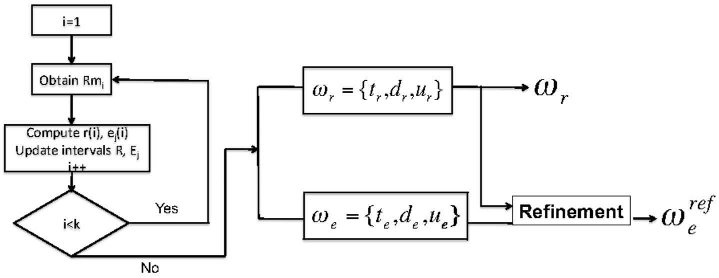


**Figure 10: Flow diagram of our assessment procedure.**

We further need to update the distrust and uncertainty associated with the expertise opinion since it must hold $t + d + u = 1$. Given that $t_e^{ref} < t_e$, if we do not update (increase) $d_e$ and $u_e$ (that is if $d_e^{ref} = d_e$ and $u_e^{ref} = u_e$), we will have $t_e^{ref} + d_e^{ref} + u_e^{ref} < 1$. Hence, we distribute the *trust degradation*, $t_{e,deg} = t_e - t_e^{ref}$, to the expertise distrust and uncertainty proportianly to their initially assessed values:

$$d_e^{ref} = d_e + \frac{d_e}{d_e + u_e} \cdot t_{e,deg} \tag{13}$$

$$u_e^{ref} = u_e + \frac{u_e}{d_e + u_e} \cdot t_{e,deg} \tag{14}$$

Care should be taken when $t_e = 1$, which means that $d_e = u_e = 0$. In this case, $t_{e,deg}$ is distributed equally across the expertise distrust and uncertainty (i.e., $d_e^{ref} = u_e^{ref} = 0.5 \cdot t_{e,deg}$).

Figure 5, depicts the accuracy of our assessment scheme when the refinement engine is used. As one can observe, the expertise accuracy is significantly increased ($\sim 95\%$). Later, we will delve into the scenarios where our scheme still fails to correctly assess the expertise of a user. In brief, this happens for the case of a "Talkative expert". The refinement phase will reduce the expertise trust, even for the topics of her actual expertise. The hit on the overall performance is not large, since based on the cognitive profile these topics are very few (at most 3 topics for each user)[4]. In addition, falsely trusting an amateur is much more critical than having less trust in

[4] We have tried to distribute the different profiles evenly across the users monitored.

the answer of an expert, since in the former case the underlying social network diffuses wrong information to its users.

Finally, Figures 11 and 12, present the ECDF of $t_e^{ref} - e^*$ and $u_e^{ref}$, respectively. It is interesting to emphasize on the increase of the fuzziness with respect to the expertise opinion. This is an artifact of Equations 13 and 14.

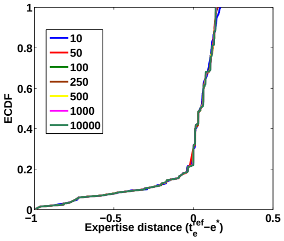


**Figure 11: Expertise distance with refinement.**

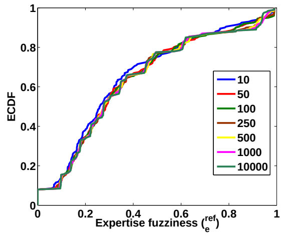


**Figure 12: Expertise uncertainty with refinement.**

| # of snapshots | 10 | 50 | 100 | 250 | 500 | 1000 | 10000 |
|---|---|---|---|---|---|---|---|
| Accuracy | 0.71 | 0.755 | 0.76 | 0.75 | 0.75 | 0.75 | 0.755 |

**Table 2: Expertise accuracy with *early* refinement.**

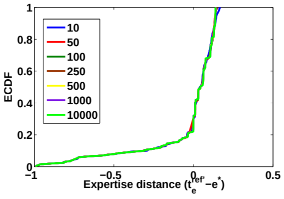


**Figure 13: Expertise distance with *early* refinement.**

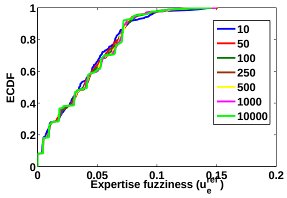


**Figure 14: Expertise uncertainty with *early* refinement.**

One could refine the assessed values for the expertise earlier in the inference engine. In particular, instead of applying the refinement on the opinions $\omega$ after $k$ snapshots, it is possible to perform this step earlier, during the computation of $r(i)$ and $e_j(i)$ (refer to Figure 10). In this case, we could have:

$$e_j^{ref}(i) = e_j(i) \cdot r^2(i) \tag{15}$$

With this *early* refinement, we manipulate the single point estimates obtained from each snapshot. Thus, we do not need to refine the opinions obtained through Equation 10 using Equations (12)-(14) . We have repeated all of our experiments with this approach. The inference accuracy results for expertise are presented in Table 2. As we see the accuracy is increased as compared with the plain approach, however it is slightly reduced as compared with the previous refinement engine (*late* refinement).

In order to dig into the details of this perfromance we examine the distance between the assessed expertise value and the preconfigured one, as well as the uncertainty in the estimation. Figures 13 and 14 depict these results. As we can observe from Figure 13 the CDF of the distance between the assessed expertise and the real expertise is similar to the one obtained with the initial refining approach (Figure 11). However, the uncertainty is greatly reduced (Figures 14 and 12). As aforementioned, the *late* refinement phase needs to apply Equations 13 and 14, which greatly increases uncertainty. Consequently, this increased fuzziness makes it possible to more easily satisfy the first part of Equation 11 and thus, observe an increased accuracy. However, the actual accuracy of the two schemes is the same, as can been seen from the CDF of the difference $t_e^{ref} - e^*$. For the rest of our work we will use the *late* refinement approach, to which we will simply refer as refinement phase.

**Opinion clusters:** Before examining dynamic behaviors, we are interested into identifying possible clusters of the different user profiles to the 3D space of the reliability and expertise opinions. In particular, we consider 10 users of each category and we compute the reliability and expertise opinions for different number of snapshots. Results are presented in Figure 15. Reliability and expertise opinions for each type of users are plotted on the same axes. Considering the set of points $\{\omega_r\}$ and $\{\omega_e^{ref}\}$ we can see that for each type of user they are spread over different *areas* of $[0, 1]^3$. Hence, we can classify a user based on the estimated opinions of trust and reliability. As it is evident from the figures, these clusters are formed even when a small number of snapshots (e.g., 10) used for the assessment (top plots). Monitoring changes in these clusters for a specific user can help as *track* behavioral changes as we will see in the following.

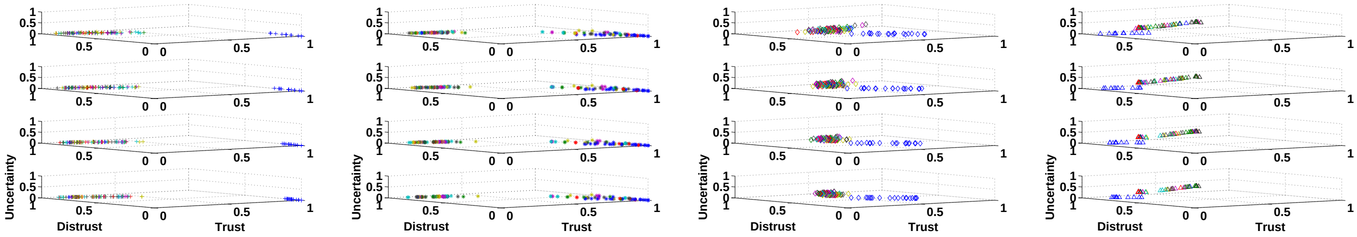
## 5.3 Performance under dynamic users' behavior

During the operations of a Q&A network, a user might change his behavior for a variety of reasons. In the simplest case, Jack can initially be a “Reliable amateur“, and after a period during which he builds his expertise, he can become a “Reliable expert“. Hence, it is important to examine the performance of our system under scenarios that involve behavioral changes. We will also study the performance of the consensus assessment and its overall effect.

**Response to dynamic behavior:** The above results correspond to static scenarios; the (real) expertise and reputation values do not change during the network evolution. However, in reality a user might change her behavior over time for various reasons. For instance, Alice is an expert in “Medicine“, but her account got compromised by Eve who is a computer scientist and knows nothing about medical questions. Alternatively, an initially amateur user can start building her expertise, just as a medical student gradually builds his/her medical specialization while attending the medical school. In this set of experiments we seek to examine the effect of similar dynamic behaviors on the assessed quantities. More specifically, we want to examine the responsiveness of our scheme to similar changes. We simulate 800 network snapshots with a behavior change every 200 snapshots. The cycle followed (we will refer to this cycle, as cycle 1) is: “Reliable amateur” → “Reliable expert” → “Talkative expert” → “Talkative amateur“. Note here that when $x$ snapshots have passed, we utilize all of them for the current assessment. In other words, the scheme currently exhibits a *full memory*.

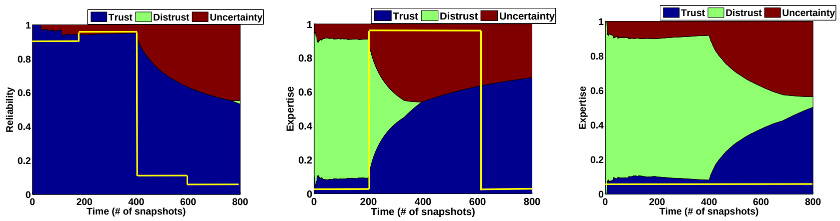
Figure 16 shows the reliability of a user (say Alice) along with her expertise (no refinement phase) with respect to two different topics. The real expertise topic corresponds to a subject for which Alice indeed has a specialization during some period in time (i.e., “Medicine“), while the false expertise topic corresponds to a category for which she is not knowledgable at all[5]. The yellow line

[5] Note here that, even for the expertise topic, there can be periods



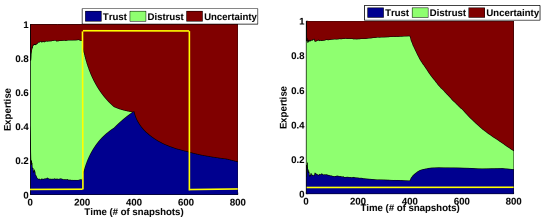
(a) Reliable amateur (b) Reliable expert (c) Talkative amateur (d) Talkative expert

**Figure 15: Opinions' clusters. Each scatter plot corresponds to different number of snapshots used (10, 20, 50, 100 from top to bottom-refinement phase is used) .**



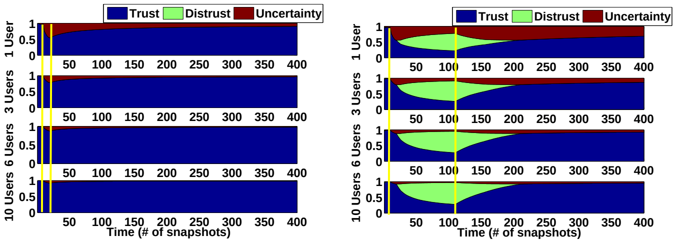
(a) Reliability (b) *Real* expertise topic (c) *False* expertise topic

**Figure 16: Dynamics with no refinement phase (cycle 1).**



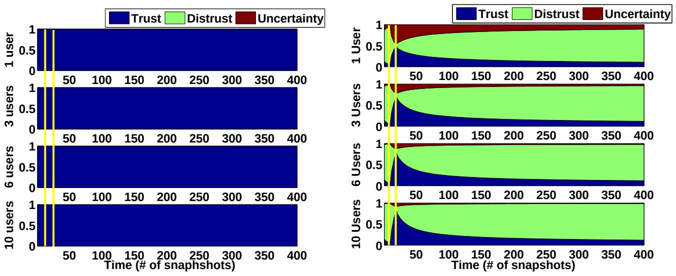
(a) *Real* expertise topic (b) *False* expertise topic

**Figure 17: Dynamics with refinement phase (cycle1).**



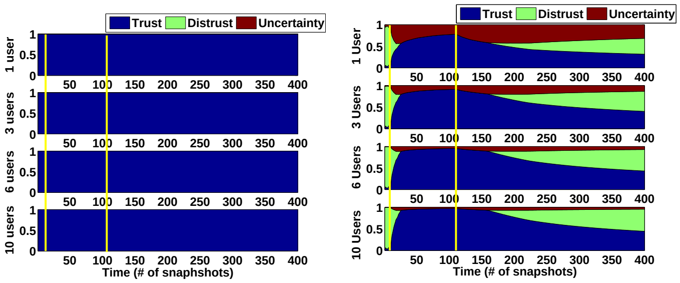
(a) Short perturbation (b) Long perturbation

**Figure 20: User reliability.**



(a) *True* expertise topic (b) *False* expertise topic

**Figure 21: User Expertise: short perturbation period.**



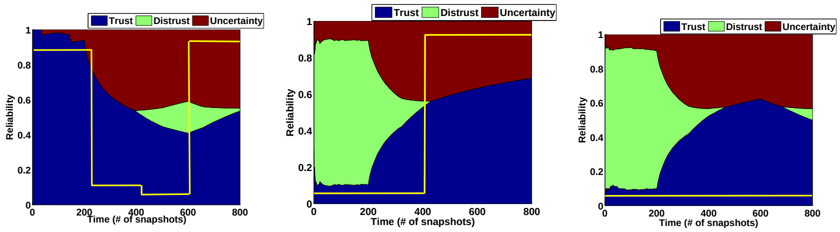
(a) *True* expertise topic (b) *False* expertise topic

**Figure 22: User Expertise: long perturbation period.**

represents the time progress of the *real* reliability/expertise of Alice as fed into our simulations. The trend observed in Alice's assessed reliability sufficiently follows the behavioral cycle we simulated. For the first 400 snapshots her reliability is high, while for the rest of the simulation period her reputation degrades. The real reputation reduces immediately to 0.1, however the degredation in the assessed value is much less steep due to the accumulated nature of the estimation. To reiterate, there are no RMs ignored, even if they correspond to old snapshots that they might have become stale.

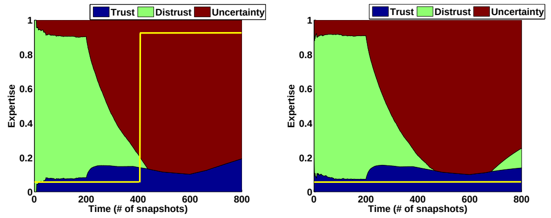
As alluded to above, the expertise assessment is more challenging. Figures 16(b) and 16(c) clearly illustrate this. When Alice becomes *talkative*, her assessed expertised is boosted in both types of topics. In the case of the false expertise topic during the period between 400-800 snapshots, Alice's expertise is falsely increased. The same holds for the expertise topic for the period between 600-800 snapshots (Figure 16(b)), during which Alice is an amateur (e.g., due to her account being misused). However, if we examine the reliability and expertise assessments in combination, we can identify the periods of false expertise assessment, due to the low reliability of Alice. This falls back to the refinement phase we introduced in the previous (static) set of experiments. Simulating the same scenario using the refinement phase, we obtain Figures 17(a) and 17(b). As it is evident, the non expertise topic does not exhibit any false assessment anymore. In particular, there is a degradation of the expertise trust for the real topic of specialization, when Alice transits from "Talkative expert" to "Talkative amateur" as it should be the case. Nevertheless, as mentioned above, there is a degradation of her ex-

for which Alice is an amateur and has no knowledge for this topic as well. As aforementioned, this can correspond to periods where she is building knowledge, her account is comprimised etc.



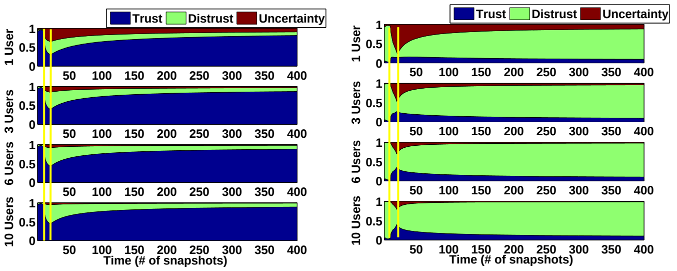
(a) Reliability (b) *Real* expertise topic (c) *False* expertise topic

**Figure 18: Dynamics with no refinement phase (cycle 2).**



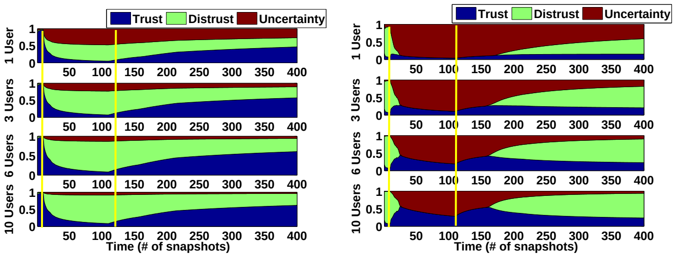
(a) *Real* expertise topic (b) *False* expertise topic

**Figure 19: Dynamics with refinement phase (cycle 2).**



(a) *True* expertise topic (b) *False* expertise topic

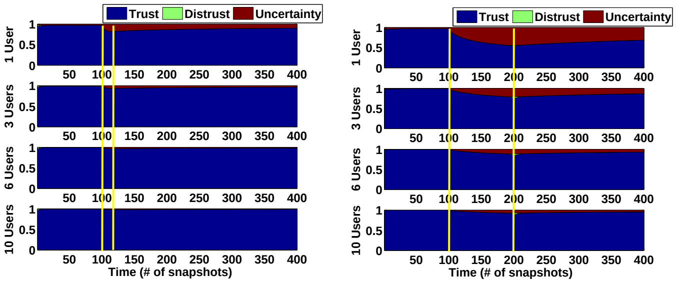
**Figure 23: User expertise: short perturbation period with refinement.**



(a) *True* expertise topic (b) *False* expertise topic

**Figure 24: User expertise: long perturbation period with refinement.**

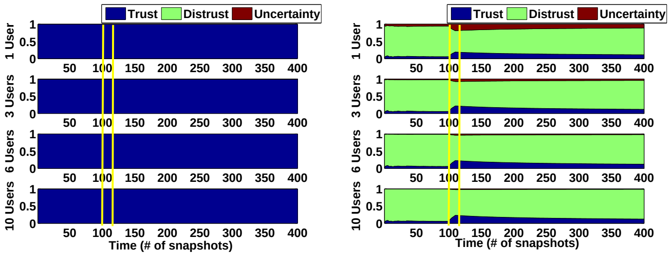
pertise during the “Talkative expert” period as well (e.g., Figure 17(b), between snapshots 400-600). This is an expected outcome of the refinement performed: the trust in user’s expertise degrades with the reduction of the user reliability. The fact that Alice is unreliable should affect our general trust on her replies.



(a) Short perturbation (b) Long perturbation

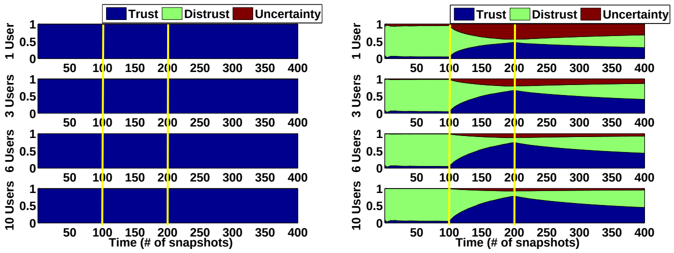
**Figure 25: User reliability for long initial “Reliable expert” period.**

In the following we examine different behavioral cycles. As another example we consider the following cycle (we will refer to this cycle, as cycle 2): “Reliable amateur” → “Talkative amateur“→ “Talkative expert” → “Reliable expert“. The difference between the two cycles is the swap between the second and fourth period (“Talkative amateur” and “Reliable expert“). Figure 18 presents the results when the refinement phase is not used. The reputation estimation successfully follows the real reliability (qualitatively). Again, it is interesting to emphasize on the fact that the trust on the reputation is increased in a slow rate after the $600^{th}$ snapshot, due to the accumulation of many observations that yield a low reliability (period between 200-600 snapshots). Another point worth of noting is the results for the real expertise topic (Figure 18(b)). In particular, there is a great similarity with the case of cycle 1 (Figure 16(b)). The reason for this is that the behavior of a “Talkative amateur” and a “Reliable expert” on the topic of expertise is similar. The former will reply because she is *talkative*, while the latter will respond because it is her specialization. As expected, the refinement engine manages to overcome the effect (Figure 19(a)). The trust on the expertise starts (slowly) increasing only after the $600^{th}$ snapshot when Alice is a “Reliable expert“. In the following we will examine the performance when considering a smaller amount of snapshots. In other words, a sliding window of fixed size will be used.



(a) *True* expertise topic (b) *False* expertise topic

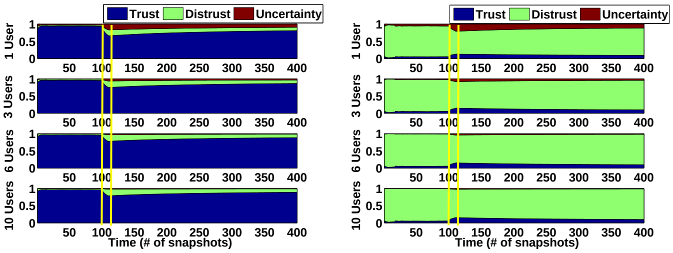
**Figure 26: User Expertise: short perturbation period and long initial “Reliable expert” period.**



(a) *True* expertise topic (b) *False* expertise topic

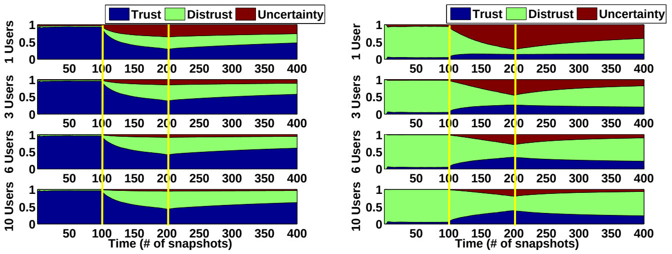
**Figure 27: User Expertise: long perturbation period and long initial “Reliable expert” period.**

**Consensus study:** Next, we consider dynamic scenarios where Alice is being monitored by a group of peers who collaborate towards obtaining a consensus on her reliability and/or expertise. In the scenarios examined, Alice is a “Reliable expert” but after some time, she perturbs for a period of time, when she acts as a “Talkative expert“. The initial “Reliable expert” period and the perturbation period are set to different values in our experiments as described below. First we consider a small initial period of 10 snapshots and two different perturbation periods; one short, 10 snapshots, and one long, 100 snapshots. Figures 20(a) and 20(b) present Alice's reliability. The vertical yellow lines mark the time points when the behavioral changes occur. As expected her reputation degrades during the perturbation period and is restored when it finishes. With a long perturbation period the degradation is higher as one might have expected. Figures 21(a) and 21(b) present Alice's estimated expertise for different numbers of monitoring peers (the vertical yellow lines identify the points of behavioral changes). Note here that, the order of opinion combining is not important, as the consensus operator is both commutative and associative [10]. Thus, in our experiments, we fix the order of users (e.g., by their ID) and in every scenario we add opinions from this sorted list.



(a) *True* expertise topic (b) *False* expertise topic

**Figure 28: User expertise: short perturbation period and long initial "Reliable expert" period with refinement.**



(a) *True* expertise topic (b) *False* expertise topic

**Figure 29: User expertise: long perturbation period and long initial "Reliable expert" period with refinement.**

When no refinement is applied we again observe the issue of false expertise assessment for the "Talkative expert" period (Figure 21(b) snapshots 10-20 and Figure 22(b) snapshots 10-110). This effect is pronounced with consensus. The reason for this is that consensus reduces uncertainty, thus, trust is increased. However, as one might anticipate from the results presented above, the refinement process eliminates the false expertise problem (Figures 23(b) and 24(b)). As mentioned in the above, expertise refinement has a slightly negative effect on the expertise assessment for a topic of real specialization. This is depicted again in Figures 23(a) and 24(a) during the perturbation period (snapshots 10-20 and 10-110 respectively). However, to reiterate, this degradation is much less important when compared with the false expertise inference. The effect is also downgraded with the increase in the number of participating peers in the consensus. For instance with 10 monitoring users we have an approximately 30% less reduction in the trust in Alice's expertise. Nevertheless, the accumulated nature of the estimation results in a slow restoration of the expertise value after the perturbation period, which ideally we would like to eliminate. As we will see later, a shorter snapshot history can help towards this direction too.

Figures (25) - (29) present the corresponding results for an initial "Reliable expert" period of 100 snapshots and two different durations of the perturbation period (10 and 100 snapshots respectively). The nature of the results is similar with the first scenarios considered, however it is interesting to observe Figure 25(a). We see that even a small perturbation period, with a *large* good past, is enough to hurt one's reputation from the standpoint of a single user. Alice's reputations is never completely restored especially when only one user is used for the estimation. Nevertheless, applying the consensus operator helps to absorb this effect.

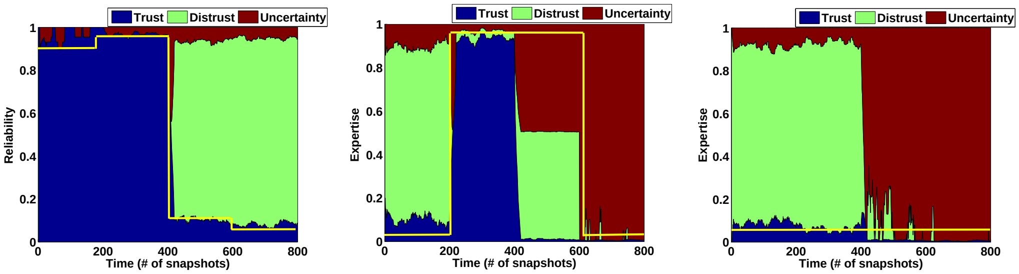
**The effect of history length:** Until now, whenever we wanted to estimate the values of Jack's attributes, we have considered the whole history up the time of assessment. However, some of these evidence might be *stale* and not accurately represent the current behavior of Jack. Keeping a long history makes the assessment scheme less responsive to dynamic changes; it might take a lot of time to restore reputation/expertise even after a relatively short *bad* period. Furthermore, as one can observe from Figure 5 our system provides similar accuracy when a small (e.g., 10) or a larger (e.g., 10000) number of snapshots is utilized for the estimation. Hence, we are interested into examining the dynamic performance of our scheme while retaining a smaller *memory*. In particular, after $x$ snapshots, instead of having observation vectors of length $x$ (from snapshot 1 to snapshot $x$), we have vectors of length $\phi$ (from snapshot $x - \phi + 1$ to snapshot $x$).

We repeat the above experiments with a history window of $\phi = 10$ snapshots (only the results with refinement are presented). Figures 30 and 31 present the results for the two behavioral "cycles" examined earlier. As one can observe, by keeping only a history of 10 snapshots our scheme is able to react faster to behavioral changes. The changing rate of the estimated values is much more steep as compared with the *smoother* changing rate when all the history was accounted for the inference (Figures 17 and 19).

Finally, we repeat our perturbation experiments with consensus computation. Aggregating the opinion of many users about Alice, through the consensus operator, resulted in a decreased uncertainty as seen above. However, even when combining the opinions of 10 users, the assessment is not very reactive to the behavioral changes (e.g., Figure 24(a)). As our simulation results in Figures (32)-(34) indicate, *forgetting* old evidence provides flexibility when aggregating opinions as well. Note here that all users whose opinions on Alice we aggregate retain the same length of history (10 snapshots in our simulations). We present our results only for an initial short "Reliable expert" period and for two different pertubation durations, however the results with other combinations of period durations are similar.

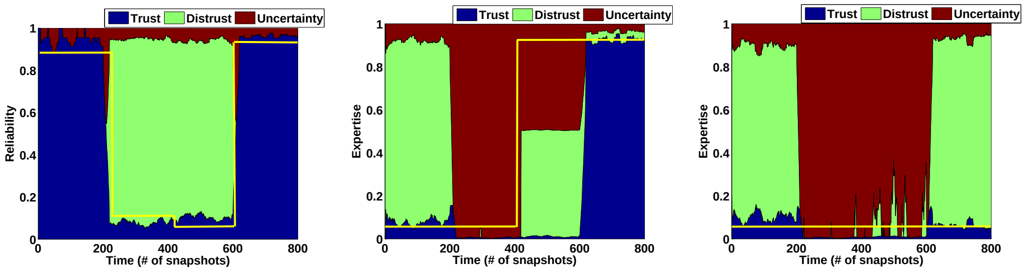
## 5.4 Real Users' Behavior

In this last set of results we are interested into studying the real behavior of users of Q&A systems, using data obtained from Yahoo! Answers. We first examine the applicability of our system by studying the pairwise interactions between the users. Furthermore, Yahoo! Answers has a hierarchical classification of question categories. In particular, there is a high level classification (e.g., travel, computers etc.), and there is a lower lever classification (subcategories), where each one of these categories map to a number of more specific ones (e.g., travel can expand to different cities such as Detroit, New York City etc.) as we will see in the following. The



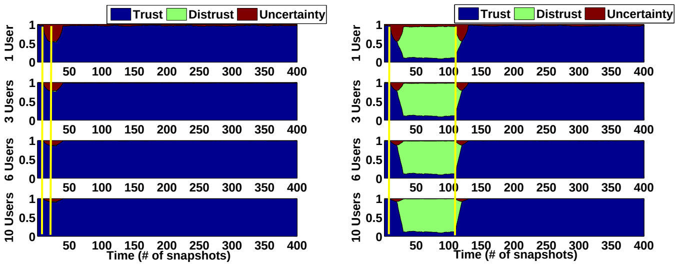
(a) Reliability (b) *Real* expertise topic (c) *False* expertise topic

**Figure 30: Short memory allows for faster response to dynamic behaviors (cycle 1).**



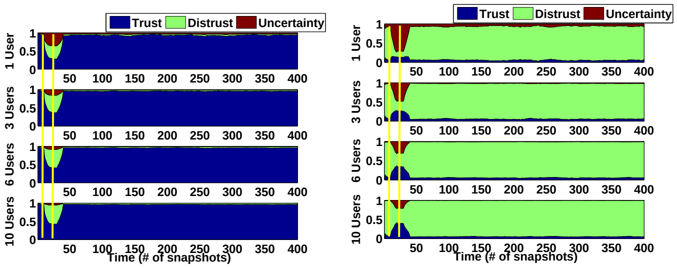
(a) Reliability (b) *Real* expertise topic (c) *False* expertise topic

**Figure 31: Reliability/Expertise can be fastly restored when maintaining short history (cycle 2).**



(a) Short perturbation (b) Long perturbation

**Figure 32: User reliability (Memory: 10 snapshots).**



(a) *True* expertise topic (b) *False* expertise topic

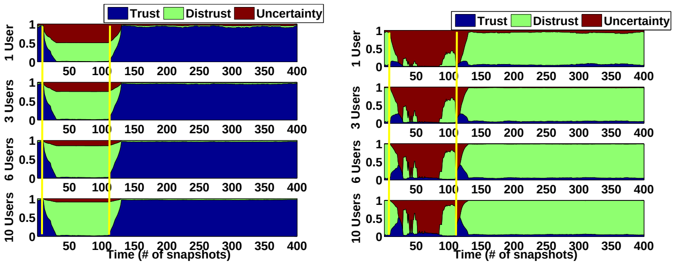
**Figure 33: User Expertise: short perturbation period (Memory: 10 snapshots).**

goal in this section is to use the data crawled in order to examine differences between the two classification levels.

**Data collection:** Yahoo! Answers groups users' questions in 24 main categories. Under the main categories there are 1,320 sub-categories distributed in multiple sub-levels. Category sub-grouping is done on wide variety of principles raging from geographic location to topic sub-partitioning. For instance, main category “Dinning out” has approximately thirty sub-categories based on geographical location, such as “United States", “Germany", “Venezuela", etc. Another example would be the main category of “Computers", which further divides into more specific topics such as “Computer Security", “Internet", “Networking” so forth. For better understanding Figure 35 visualizes part of the categories *tree* structure of Yahoo! Answers.

Yahoo! also provides two interfaces to access Q&A data. Both interfaces are accessed via the web utilizing authentication. Yahoo! uses authentication to track each query session and to limit the rate at which data is provided. This measure is in place in order to provide partial picture of the data, and therefore ensuring the privacy of the users. Furthermore, Yahoo! has no knowledge of user's real identity, making the inference of real personal data practically impossible.

There are four query types that Yahoo's API supports; (1) Query by Category, which provides question asked in the specified category, (2) Query by User, which provides questions and answers posted by a specific user, (3) Query by Question, which lists all answers to a specific question, and (4) Query for Question, which returns question that match specific search string. Any response from Yahoo! for a given query does not guarantee exhaustive answers. For instance, a query asking for all questions in given sub-category, may only return a portion of all questions and may also contain duplicate entries. This mechanism as well as the daily query rate limit makes



(a) *True* expertise topic (b) *False* expertise topic

**Figure 34: User Expertise: long perturbation period (Memory: 10 snapshots).**

deduction of the full hierarchy of categories, questions or users very difficult and time consuming. We crawled data from Yahoo! Answers for 3 months (between September 2011 and November 2011) and we were able to infer a fairly larger portion of the hierarchy tree. We collected data from 78,304 users, including 104,651 questions and 10,530 answers[6]. In what follows we only present data from 6 users with representative behavior.

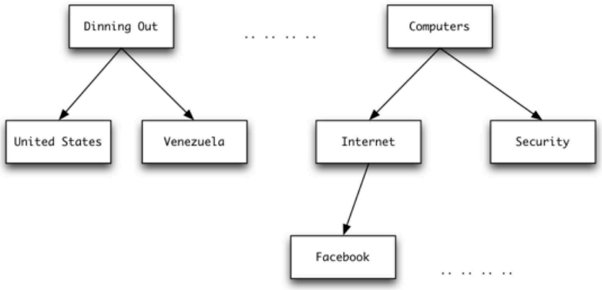


**Figure 35: Categories hierarchy of Yahoo! Answers.**

**Pairwise users' interactions:** The scheme presented in this paper requires a large number of pairwise interactions between the users, that is, pairs of user answering to each others questions. Using the data obtained from Yahoo! Answers we examine to see if users of this system exhibit this behavior. Figure 36 presents the empirical CDF of the bi-directional pairwise interactions. As we can observe the maximum number of such interactions between two users is 18, while more than 95% of the pairs have less than 5 interactions. This means that the response matrices will have less than 5 entries, rendering our system inapplicable for the case of Yahoo! Answers.

Nevertheless, the proposed scheme can still be applied in such scenarios in a centralized manner. In particular, we can consider the one "end" of each user pair (i.e., the "questioner") to be the system as a whole (i.e., Yahoo! Answers). Hence, each question posted by any user in the system, can be thought as a question originated from the system provider. Simply put, in the case of few pairwise interactions we construct the response matrix of a specific user considering the questions posted from all the rest of the users in aggregate. This provides us with a response matrix for each user for his aggregate behavior to all the questions posted in the system. Figure 37 provides an illustrative example of the above process. In this example Bob has only two Q&A interactions with Jack, while Eve has even less, that is, one. In order to obtain a larger RM we can integarate these interactions to one larger RM of a "super-user" (i.e., the system) and use the MLE framework described in Section 4 to compute a reliability and expertise value for Jack.

| User | # of *active*<br>high level categories | % of responses per category |
|---|---|---|
| 1 | 3 | 39%, 37%, 24% |
| 2 | 1 | 100% |
| 3 | 2 | 57%, 43% |
| 4 | 1 | 100% |
| 5 | 2 | 95%, 5% |
| 6 | 1 | 100% |

**Table 3: Macroscopic user's behavior.**

We would also like to note here that, other Q&A systems, more focused on specific topics (e.g., stackoverflow.com) may exhibit different behavior with regards to users' pairwise interactions, hence, allowing our system to be applicable in a distributed fashion, exactly as presented above.

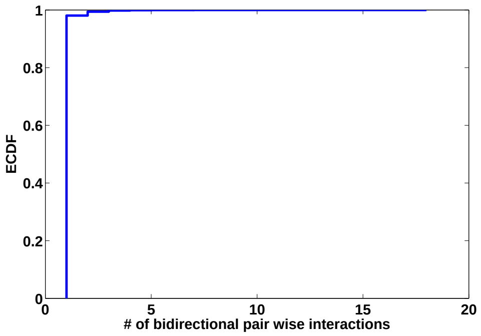


**Figure 36: Empirical CDF of the bi-directional pairwise interactions between Yahoo! Answers users.**

**Micro- vs Macro-scopic users' behavior:** We start by presenting the 6 users' behavior with regards to the high level categories. Given that in a real system the pairwise interactions during the crawling period might be few as shown previously, we consider the user that posts the questions to be the actual system, that is, we examine the aggregated response activity of a user. Table 3 presents the obtained results. In particular, we include the number of categories that each user contributed with responses at and the percentage of his activity in each one of them.

As one can observe, users highly focus on responding in particular categories. Only user 1 and 3 are (equally) active in 3 and 2 categories respectively (out of the total 24 top-level categories at Yahoo! Answers). Users 2, 4 and 6 are answering only in one particular topic, while user 5 *spends* only 5% of his activity in a second topic. Based on this activity distribution one could conclude that the above users are highly reliable since they focus on a few categories.

However, if we examine the same users' behavior in a microscopic level, zooming inside the high level categories that they are active in, the results are flipped. In particular, users spread their response activity to a large number of lower level categories, rendering them

[6] The dataset collected will be made available.

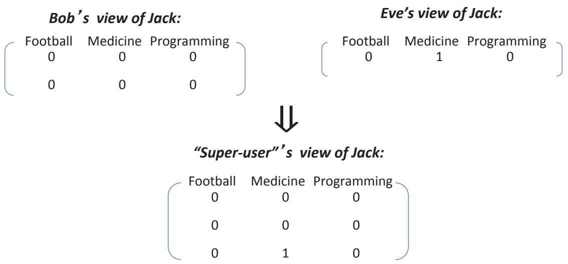

Bob's view of Jack:
Football Medicine Programming
0 0 0
0 0 0
Eve's view of Jack:
Football Medicine Programming
0 1 0
⇓
"Super-user"'s view of Jack:
Football Medicine Programming
0 0 0
0 0 0
0 1 0

**Figure 37: Aggregating pairwise user interactions to a larger RM.**

| User | # of *active* low level categories | maximum percentile activity |
|---|---|---|
| 1 | 16 | 18% |
| 2 | 11 | 14% |
| 3 | 12 | 28% |
| 4 | 13 | 15% |
| 5 | 18 | 13% |
| 6 | 12 | 7% |

**Table 4: Microscopic user's behavior.**

unreliable. Table 4 presents the results when focusing on the low level behavor of each user. In particular, we present the number of different topics that users contribute reponses to as well as the maximum percentile activity. The latter, is the activity percentage for the category in which the corresponding user is more active.

As it is evident, users appear highly unfocused when it comes to low level categories. Of course, the *contextual* distance between these topics is smaller as compared to the high level categories. For instance, the top level categories "Computers" and "Travelling" are far more distant in context than "Computer Networking" and "Internet" (which are sub-categories of "Computers"). Nevertheless, users when observed at a different level exhibit different characteristics and our framework can be applied as showed above to identify these differencies in the micro- and macro-scopic behavior of real users.

While the above analysis can reveal differences at a macro-scopic and micro-scopic level of the behavior of users, it is crucial to be able to integrate in our system the contextual distance between question categories - especially lower level ones. Assuming a distance metric[7] between categories $c_i$ and $c_j$, $\texttt{dist}(c_i, c_j)$, one possibility is to consider it in the computation of $R_2$ (details are provided in the Appendix).

The goal of the aforementioned extension is to retain an appropriately high reliability value for users that spread their activity over topics that are semantically close. Similarly, once we have inferred $e_{u,i}$ for user $u$ and category $i$, we can further aggregate the expertise values over a variety of topics that exhibit small dist. We can again use subjective logic for the aggregation or any other data fusion algorithm. This can help us obtain a view of $u$'s expertise on a category defined from the topics over which we aggregated. For instance, while the general term "Computer Science" can have a number of sub-topics, we might choose to aggregate over a subset of them (e.g., fusing expertise values over "Operating Systems", "Computer Networks" and "Embedded Systems" could possible provide us with an expertise value for "Computer Systems").

## 6. SCOPE OF OUR WORK

In this work we have focused on Q&A social networks. Our framework though, is not limited and applicable only to these type of information networks. For instance, we are currently utilizing a similar framework for assessing the reliability of conflicting data in large-scale historical data. Moreover, similar approaches can be taken for other kinds of data communication networks. As an illustrative example, let us consider a sensor infrastructure. Each mote of the sensor network is monitoring specific environmental attributes/events, and reports the corresponding information to the sink node for further processing. However, how can the sink know that the **data** provided are trustworthy? Even if the reporting device can be authenticated and is reliable (e.g., it has not been compromised), its report might not be very accurate due to its physical distance from the location of interest. This physical distance can define in this scenario the different contexts of expertise; a mote is *expert* for the events happening within a distance of $d$ meters from its own position.

Today, the amount of available data is so large that it makes the extraction of valuable knowledge extremely challenging. An automated system to filter out information of low trust will be extremely valuable. The work presented in this paper clearly aims towards this direction by enabling the design of a scalable information-centric trust system. An information consumer in a Q&A social network, needs to fast identify a trustworthy answer to her question, without the need of going through a large number of (possibly not helpful) replies. As another example, a sink in a sensor network needs to consider only the data that are of high trust. This can reduce the processing time and the computation cost. Furthermore, in tactical networks, every soldier in the battlefield needs to be aware of the trust level of the information received, which can be as critical as the position of the enemy's army.

Traditionally, the quality of service provided by an information network is captured through metrics such as the amount of data delivered over a time unit, delay, packet (information) loss etc. Little, if any at all, attention has been given to the actual *helpfulness* of the information received. Without knowing the quality of the obtained data, we cannot accurately quantify the services provided by the underlying network. The above metrics cannot be used to capture the *importance* of a Q&A social network. But even more general, key to the performance of an information delivery network/system is the amount of useful/trustworthy data exchanged over it, and this is not revealed using the traditional metrics. Our work can be seen as the first step towards defining such metrics. E.g., a specific social network might consist of many " Reliable experts" on a given topic (e.g., "Medical") and we should be able have a way to capture it.

Before concluding we would like to emphasize on the ***limitations*** of our work. Even though the user model we are considering is both simple and realistic, it is not certain that every single participating peer follows it. For instance, an expert user might be selfish as well, being silent most of the time. In this case, he will rarely reply to questions, even if they fall into her expertise, leading to a false assessment of her being a "Reliable amateur". Even though such behaviors do not spread wrong information in the network, it can impact the overall quality of the underlying network (e.g.,

[7] Possibly defined from the system provider.

many questions remain unanswered). Of course, if users do not completely adhere to the cognitive model considered, the accuracy performance of the assessment scheme will get a hit. Nevertheless, even in these scenarios, our cognitive-based inference engine can still be helpful for flagging users for further examination.

In addition, despite the fact that we can identify "spammers" with the refinement phase, our scheme is not robust to the presence of **malicious** enities. Since we are not considering any feedback on the replies or their correctness, a malicious user can focus on a few categories and reply to queries of these categories, even if he does not really have the right information. Given that he adheres to the expected profile he will be classified as a "Reliable expert" and his peers will treat his responses as ones with high quality. On the positive side, this can affect only a few categories and hence, there will not be excessive wrong information diffusion. In addition, if the underlying network has many real "Reliable experts" in these categories, they can possibly isolate the malicious users and absorb the wrong information. Furthermore, Eve might respond to specific topics to Jack and to different topics to Bob. Even though, Eve will be reliable (and expert) in the eyes of Jack and Bob, it is clear that in aggregate she is spreading her responses to a large number of categories. These behaviors can be detected by the system provider in a manner similar to the one discussed in Section 5.4 (the system can compute the response matrix of each user over all the questions posted in the system) and the estimations can be refined.

## 7. CONCLUSIONS

To date the trust one has on the information delivered from a network has received very little attention. Assessing the expertise and reliability of an information provider is the first step towards a data-centric trust system. In this work we propose a cognitive-based, lightweight scheme for simultaneously assessing the expertise and reliability of a Q&A SN user. Every user can estimate locally, a subjective opinion from any other peer. These opinions can be further fused using the consensus operator borrowed from subjective logic, to obtain a more *objective* view of the users. Our simulation results show that under the assumption that users adhere to the model presented, our scheme can efficiently estimate these attributes. Table 5 summarizes three basic features of our assessment engine and the objective they accommodate.

| Feature/Module | Effect |
|---|---|
| Refinement phase | Mitigation of "False expertise" |
| Consensus | Reduction of uncertainty |
| Shorter memory | Better responsiveness to dynamic behavior |

**Table 5: Effect of the various modules of our assessment scheme.**

## 8. REFERENCES

[1] John P. Kotter. *Power and Influence: Beyond Formal Authority*. Free Press, ISBN 0-02-918330-8, 1985.

[2] G. Eysenbach and C. Kohler. How do consumers search for and appraise health information on the world wide web? qualitative study using focus groups, usability tests, and in-depth interviews. *British Medical Journal*, 324:573 – 577, 2002.

[3] B. Means, Y. Toyama, R. Murphy, M. Bakia, and K. Jones. Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies. In *U.S. Department of Education Office of Planning, Evaluation, and Policy Development Policy and Program Studies Service*, 2009.

[4] Pnina Shachaf. Answer quality on q&a sites. In *Proceedings of the Sixteenth Americas Conference on Information Systems (AMCIS '10)*. AIS Electronic Library, 2010.

[5] e-bay. http://www.auctionbytes.com/cab/abn/y06/m08/i01/s04.

[6] Jordi Sabater and Carles Sierra. Reputation and social network analysis in multi-agent systems. In *AAMAS*, 2002.

[7] Trung Dong Huynh, Nicholas R. Jennings, and Nigel R. Shadbolt. An integrated trust and reputation model for open multi-agent systems. In *Journal of Autonomous Agents and MultiAgent Systems*, pages 119–154, 2006.

[8] Y. Wang and M. P. Singh. Trust via evidence combination: a mathematical approach based on uncertainty. In *TR 2006 North Carolina State University*, 2006.

[9] Y. Wang and M. P. Singh. Formal trust model for multiagent systems. In *IJCAI*, 2007.

[10] Audun Josang. A logic for uncertain probabilities. In *Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, pages 279–311, 2001.

[11] Y. Wang and M. P. Singh. Trust representation and aggregation in a distributed agent system. In *AAAI*, 2006.

[12] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *CUAI*, 1998.

[13] P. Resnick, N. Iacovou, M. Suchak, P. Bergstorm, and J. Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *ACM CCSCW*, 1994.

[14] C. W. Hang, Y. Wang, and M. P. Singh. Operators for propagating trust and their evaluation in social networks. In *AAMAS*, 2009.

[15] M. Richardson, R. Agrawal, and P. Dominigos. Trust management for the semantic web. In *ISWC*, 2003.

[16] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *WWW*, 2003.

[17] K. Aberer and Z. Despotovic. Managing trust in a peer-2-peer information system. In *ACM CIKM*, 2001.

[18] F. Cornelli, E. Damiani, S. D. C. D. Vimercati, S. Paraboschi, and S. Samarati. Choosing reputable servents in a p2p network. In *WWW*, 2002.

[19] P. Resnick, R. Zeckhauser, E. Friedman, and K. Kuwabara. Choosing reputable servents in a p2p network. In *ACM Communications*, 2000.

[20] G. Theodorakopoulos and J. Baras. On trust models and trust evaluation metrics for ad hoc networks. In *IEEE JSAC*, 2006.

[21] S. Buchegger and J.-Y. Le Boudec. A robust reputation system for p2p and mobile ad-hoc networks. In *P2PEcon*, 2004.

[22] S. Ganeriwal and M. Srivastava. Reputation-based framework for high integrity sensor networks. In *SASN*, 2004.

[23] J. Mundinger and J.-Y. Le Boudec. Reputation in self-organized communication systems and beyond. In *Inter-Perf*, 2006.

[24] Y. Sun, W. Yu, Z. Han, and K. J. Ray Liu. Information theoretic framework of trust modeling and evaluation for ad hoc networks. In *IEEE JSAC*, pages 305–317, 2006.

[25] H. Kautz, A. Milewski, and B. Selman. Agent amplified communication. In *National Conference of Artificial Intelligence*, 1996.

[26] L.N. Foner. Yenta: A multi-agent, referral-based matchmaking system. In *Agents*, 1997.

[27] B. Krulwich and C. Burkey. The contactfinder agent: Answering bulleting board questions with referrals. In *National Conference of Artificial Intelligence*, 1996.

[28] 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 (CIKM '08)*, pages 921–930, New York, NY, USA, 2008. ACM.

[29] H. Kautz, B. Selman, and M. Shah. Referralweb: Combining social networks and collaborative filtering. In *ACM Communications, vol. 40, no. 3*, 1997.

[30] J. Zhang, M. A. Ackerman, and L. Adamic. Expertise networks in online communities: Structure and algorithms. In *WWW*, 2007.

[31] A. John and D. Seligmann. Collaborative tagging and expertise in the enterprise. In *WWW*, 2006.

[32] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *Stanford Digital Libraries Technologies Project*, 1998.

[33] L. Streeter and K. Lochbaum. Who knows: A system based on automatic representation of semantic structure. In *RIAO*, 1988.

[34] M.S. Ackerman and D.W. McDonald. Answer garden 2: merging organizational memory with collaborative help. In *CSCW*, 1996.

[35] B. Dom, I. Eiron, A. Cozzi, and Y. Zhang. Graph-based ranking algorithms for e-mail expertise analysis. In *DMKD*, 2003.

[36] G. Kasneci, J. Van Gael, D. Stern, and T. Graepel. Cobayes: Baysian knowledge corroboration with assessors of unknown areas of expertise. In *WSDM*, 2011.

[37] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. Finding high-quality content in social media. In *Proceedings of the international conference on Web search and web data mining (WSDM '08)*, pages 183–194, New York, NY, USA, 2008. ACM.

[38] 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 (WWW '09)*, pages 51–60, New York, NY, USA, 2009. ACM.

[39] Jiang Bian, Yandong Liu, Eugene Agichtein, and Hongyuan Zha. Finding the right facts in the crowd: factoid question answering over social media. In *Proceeding of the 17th international conference on World Wide Web (WWW '08)*, pages 467–476, New York, NY, USA, 2008. ACM.

[40] C. Shah and J. Pomerantz. Evaluating and predicting answer quality in community qa. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10*, pages 411–418. ACM, 2010.

[41] Xudong Tu, Xin-Jing Wang, Dan Feng, and Lei Zhang. Analogical reasoning for answer ranking in social question answering. *IEEE Intelligent Systems*, 2010.

[42] B M John, A Y K Chua, and D H L Goh. What makes a high-quality user-generated answer? *IEEE Internet Computing*, 15(1):66–71, 2011.

[43] Mohamed Bouguessa, Benoît Dumoulin, and Shengrui Wang. Identifying authoritative actors in question-answering forums: the case of yahoo! answers. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08)*, pages 866–874, New York, NY, USA, 2008. ACM.

[44] Pawel Jurczyk and Eugene Agichtein. Discovering authorities in question answer communities by using link analysis. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (CIKM '07)*, pages 919–922, New York, NY, USA, 2007. ACM.

[45] Jennifer Golbeck and Kenneth R. Fleischmann. Trust in social q&a: The impact of text and photo cues of expertise. In *Proceedings of the American Society for Information Science and Technology (ASIS&T âĂŹ10)*, pages 1–10. Wiley Subscription Services, 2010.

[46] Maggy Anastasia Suryanto, Ee Peng Lim, Aixin Sun, and Roger H. L. Chiang. Quality-aware collaborative question answering: methods and evaluation. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM '09)*, pages 142–151, New York, NY, USA, 2009. ACM.

[47] Aditya Pal and Joseph A. Konstan. Expert identification in community question answering: exploring question selection bias. In *Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10)*, pages 1505–1508, New York, NY, USA, 2010. ACM.

[48] K. Panovich, R. Miller, and D. Karger. Tie strength in question & answer on social network sites. In *CSCW*, 2012.

[49] G. Wang, K. Gill, M. Mohanlal, H. Zheng, and B. Zhao. Wisdom in the social crowd: an analysis of quora. In *WWWW*, 2013.

[50] A. Josang. Artificial reasoning with subjective logic. In *Second Australian Workshop on Commonsense Reasoning*, 1997.

[51] Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In *Proceeding of the 17th international conference on World Wide Web (WWW '08)*, pages 665–674, New York, NY, USA, 2008. ACM.

# APPENDIX

Given a data set, the mode is the value that occurs more frequently. In our case the sample set $\overrightarrow{\Pi}_{Jack}$ is a vector whose $i^{th}$ element $\pi_i$, is the number of responses from Jack with respect to category $i$. For a topic of expertise $j$ we expect to have $\pi_j = w$, which will be the mode of $\overrightarrow{\Pi}_{Jack}$ (since this is the maximum possible value). By defining the set $S$ as follows:

$$S = \{i | \pi_i \geq z \cdot \max_{k \in \{1,2,\ldots n\}} \{\pi_k\}\} \tag{16}$$

we have $R_2$ to be equal to the cardinality of $S$, that is, $R_2 = |S|$. In our set of experiments we have set $z = 0.8$.

In the case where we want to consider the contextual distance $\texttt{dist}(c_i, c_j)$ between categories $c_i$ and $c_j$ in the computation of $R_2$, we can further process set $S$. With $g$ being the step function centered at 0 (i.e., $g(x) = 1,\ iff\ x \geq 0$, otherwise $g(x) = 0$), we have:

$$f(S) = \sum_{i,j \in S} g(\texttt{dist}(i,j) - \theta) \tag{17}$$

where $\theta$ is a predefined threshold of the contextual distance metric. Then we have:

$$R_2 = \frac{f(S)}{\frac{|S| \cdot (|S| - 1)}{2}} \cdot |S| \tag{18}$$

In other words, using threshold $\theta$, we identify the fraction of all possible pairs of categories in $S$ that can be thought of being contextual close to each other and we scale accordingly the cardinality of $S$, in our computations of $R_2$. Note here that, one could possibly use a "smoother" weighting by considering the actual distance between the various category pairs (instead of using a step function at Equation 17).