Collaborative Assessment of Information Provider's Reliability and Expertise Using Subjective Logic

Konstantinos Pelechrinis*, Vladimir Zadorozhny*, Vladimir Oleshchuk‡
*University of Pittsburgh
School of Information Sciences
kpele@pitt.edu, vladimir@sis.pitt.edu
‡University of Agder, Norway
Dept. of Information & Communication Technology
vladimir.oleshchuk@uia.no

***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 her peers (on the basis of responses to questions asked by her) 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 SN. Our proposed approach is highly distributed; each user can individually estimate the expertise and the reliability of her peers using her *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 assessment in order to reach a consensus. We 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.**

Keywords: Q&A Social Networks, Subjective Logic, Expertise, Reliability

I. Introduction

Social media have intruded humans' lives during the last decade and have altered many of their social interactions. One of the aspects that have been significantly affected is the way people acquire information. 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, just to name a few of the changes. In OSNs, users 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". 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.

Nevertheless, in all these situations, there is a 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 however, increases the possibilities that their search is inadequate or less reflective and for the information obtained to be flawed.

It should be clear 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 later, there exist studies that try to assess these characteristics of a user in a Q&A SN individually. In our preliminary work [4], 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 (i.e., reliable) information.

Every question posted is related with a specific topic (e.g.,

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

COLLABORATECOM 2011, October 15-18, Orlando, United States

DOI 10.4108/icst.collaboratecom.2011.247102

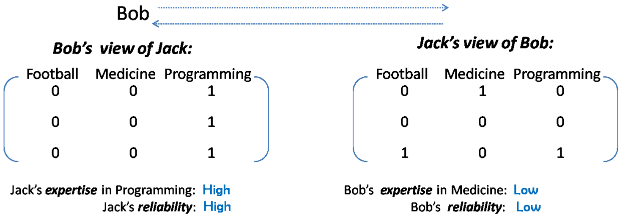

Bob
Bob's view of Jack:
Football Medicine Programming
0 0 1
0 0 1
0 0 1
Jack's expertise in Programming: High
Jack's reliability: High
Jack's view of Bob:
Football Medicine Programming
0 1 0
0 0 0
1 0 1
Bob's expertise in Medicine: Low
Bob's reliability: Low

Fig. 1. Example of Response Matrices reflecting high and low opinions

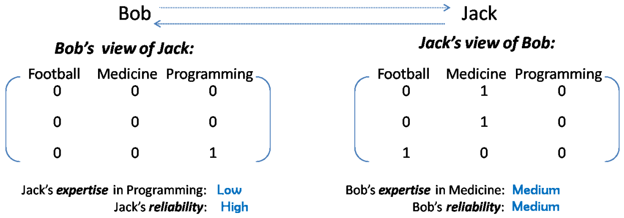

Bob
Jack
Bob's view of Jack:
Football Medicine Programming
0 0 0
0 0 0
0 0 1
Jack's expertise in Programming: Low
Jack's reliability: High
Jack's view of Bob:
Football Medicine Programming
0 1 0
0 1 0
1 0 0
Bob's expertise in Medicine: Medium
Bob's reliability: Medium

Fig. 2. Example of Response Matrices reflecting high, low and medium opinions.

"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. In this paper we further extend our local assessment framework [4]. In brief, the social network as a whole can aggregate, using subjective logic, the individual/local opinions on Jack's expertise and reliability and obtain a global opinion for his characteristics. Even just a subset of users can collaboratively estimate Jack's attributes by utilizing the subjective logic mechanism. The main advantages of our assessment system are its lightweight nature and the fact that can be applied both *locally* from every user individually or by a subset of them (or even the whole SN). 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 an 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 assessments.

The rest of the paper is organized as follows. Section II provides a simple example illustrating our system model and the basic idea of our approach. Section III briefly discusses previous related studies. Our cognitive-based assessment scheme is presented in Section IV. Section V presents our evaluations, while Section VI concluded our study.

II. Our approach in brief

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 and adequate way to estimate both (1) and (2) for any user. 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) a subjective opinion of Bob about Jack's knowledge of "Medicine" and (2) a subjective opinion of Bob about Jack's reputation. 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 of information. If a user tends to respond consistently to questions related to a particular topic, we consider her 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 her 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.

Figure 1 shows an example of two response matrices reflecting views of Bob of Jack and vice versa. In this example, Bob has posted 3 questions for 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

IV 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.

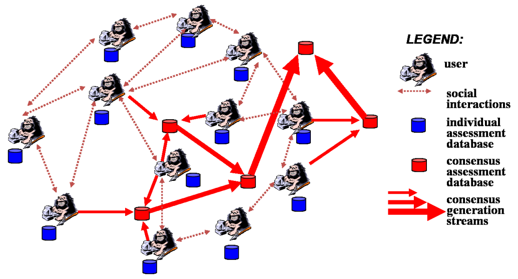

LEGEND:
user
social interactions
individual assessment database
consensus assessment database
consensus generation streams

Fig. 3. Distributed propagation and fusion of information about users reliability and expertise.

III. Related Works

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.

Expertise inference: There exist studies in the literature that try to assess the expertise metric. *Referral* systems or expert finders (e.g., [15] [16] [17]) try to locate people who are most appropriate for providing the requested information. These 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 [18] exploits the social network within a community to identify a set of experts with regards to the information requested. It leverages the "six degrees of separation" phenomena, which states that the distance between two individuals in a network is relatively small. Hence, one can possible exploit these social relations to find an expert. Nevertheless, the flexibility of similar systems is low for two main 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* [19] make a step further and not only they identify *expertise* 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) [20] utilizes the main features of the PageRank algorithm [21], 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 important to know how many answers on a specific topic Jack has posted but also to whom questions he has replied. We should put less weight to answers provided to Alice who is a *newbie* as compared to answers 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 [22], [23], and [24]. 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* [25] 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, they work in a different context (that of semantic web and knowledge corroboration) and under the assumption that users who assess the truth of the statements are indeed reliable.

Distinguishing our work: The existing studies are designed with different objectives in mind. On the one hand, 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 other hand, 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 IT consultant to her regular customers but her offhand IT advice might not be completely trustful as she is not know to be entirely forthcoming. Furthermore, there are significant differences 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 contrast, our approach does not require any explicit involvement from the users as mentioned in Section II and it is based on cognitive models for human behavior. 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.

There also exist literature that deals with closely related and interesting issues from the perspective of cognitive sciences. For instance, [26] examines the way a user builds expertise. However, to the best of our knowledge, *to date there exists no work in the literature that tries to exploit cognitive and behavioral characteristics of humans to reach the **joint** estimation of reliability and expertise*.

IV. 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 II.

A. Individual estimation

Our individual estimation scheme was presented in our initial work [4]. Here we give a brief overview for ease of further presentation.

Response matrix (RM): The Q&A SN's participating entities 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 $Q = \{``Football", ``Medicine", ``Programming"\}$). Bob can obtain locally a *subjective* opinion about Jack's (i) reliability and (ii) expertise in q. He can further augment this opinion using subjective logic consensus operator to combine views of other users (e.g., Alice) about Jack [27]. 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 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 can be *measured* with regards to the number of snapshots that we have for the Q&A SN.

Assessment of $e^{Bob}_{Jack,``Football``}$: 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 [19]. 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 though 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 V, we will describe scenarios where expertise is falsely inferred and how we can mitigate these occurrences.

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 $e^{Bob}_{Jack,``Football"}$, 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:

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

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 $\bar{\bar{p}}$ 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 $\bar{\bar{p}}$. 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 an entity. Given its highly subjective nature, there are no clear metrics for 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 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 δ_{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 $\overrightarrow{\Pi}^{Bob}_{Jack}$. The mode of a dataset is the value that occurs more often in it. In our case the sample set $\overrightarrow{\Pi}_{Jack}$ is a vector whose i^{th} element π_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:

$$S = \{i|\pi_i \geq z \cdot \max_{k\in\{1,2,\ldots n\}}\{\pi_k\}\} \tag{3}$$

we have R_2 to be equal to the cardinality of S, that is, $R_2 = |S|$.[2] Based on the above definitions, Jack is considered *reliable*, that is $r^{Bob}_{Jack} = 1$, iff:

$$\alpha \leq R_1 \leq \beta \quad \wedge \quad R_2 \leq \gamma \tag{4}$$

When these inequalities do not hold we need to update Jack's reliability value. [4] provides a detailed description of the underlying process and the corresponding penalty functions.

B. Consensus assessment

By executing the above process, Bob has obtained a subjective view of Jack. The next natural step would be for Bob to collaborate with other peers (e.g., Alice) and combine different subjective opinions of Jack. This will enable him to obtain a more *objective* opinion for Jack. 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 logical 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 as we will also see in our evaluations.

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 ω^{Bob}_p and ω^{Alice}_p 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:

$$\omega^{Bob,Jack}_p = \omega^{Bob}_p \oplus \omega^{Jack}_p = \left\{t^{Bob,Jack}_p, d^{Bob,Jack}_p, u^{Bob,Jack}_p\right\} \tag{5}$$

where $t^{Bob,Jack}_p = \left(t^{Bob}_p u^{Jack}_p + t^{Jack}_p u^{Bob}_p\right)/k$, $u^{Bob,Jack}_p = \left(u^{Bob}_p u^{Jack}_p\right)/k$, $d^{Bob,Jack}_p = \left(d^{Bob}_p u^{Jack}_p + d^{Jack}_p u^{Bob}_p\right)/k$, and $k = \left(u^{Bob}_p + u^{Jack}_p - u^{Bob}_p u^{Jack}_p\right)$.

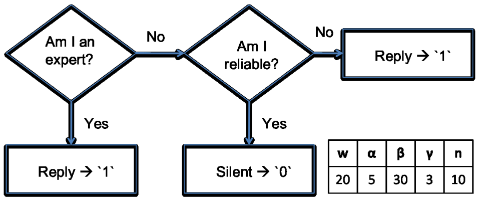
Deriving opinions from the response matrices: In order to be able to use subjective logic for consensus estimation we need to map the reliability and expertise intervals obtained locally from Bob and Alice about Jack into opinions.

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 $\omega^{Bob}_{Jack,``Football``}$):

$$\omega^{Bob}_{Jack} = \{\frac{a+b}{2}, 1 - \frac{a+b}{2} - \frac{b-a}{2}, \frac{b-a}{2}\} \tag{6}$$

V. Evaluations

In this section we present the results from our evaluations. The experimental set up is similar to the one in [4]. In brief, we create synthetic data using (i) a priori fixed expertise (on each topic) and reliability values for every user and (ii) the process depicted at Figure 4. We are primarily interested into identifying 4 categories of users; "Reliable expert", "Talkative expert", "Reliable amateur" and "Talkative amateur", with the names being self explanatory.



w	α	β	γ	n
20	5	30	3	10

Fig. 4. Flow diagram of our user model and our simulation parameters.

A. 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.

[2] In our set of experiments we have set $z = 0.8$.

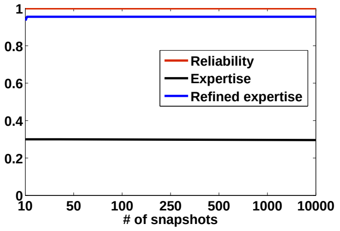


Fig. 5. Inference accuracy of our scheme.

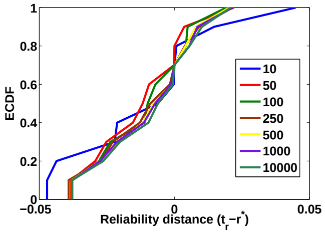


Fig. 6. Accurate reliability assessment.

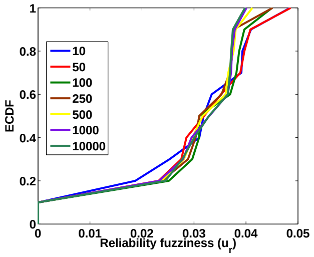


Fig. 7. Small reputation uncertainty.

We consider a set of 10 users who we *monitor*[3]. 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

$$a^* \in [t-u, t+u] \;\; \vee \;\; |t-a^*| \leq p \cdot a^*, \;\; p \in [0,1] \tag{7}$$

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 7) over the total number of estimations. As we observe, irrespective of the number of snapshots used, our scheme is capable of identifying 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 for each user). Thus, there will be an overestimation of user expertise in these topics, which results in the low accuracy. Figure 8 depicts the empirical CDF (ECDF) 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 inferred 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 reliability are obtained they serve as inputs into the refinement engine, which provides a *refined* opinion for Alice's expertise, ω_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:

$$t_e^{ref} = t_e \cdot t_r^2 \tag{8}$$

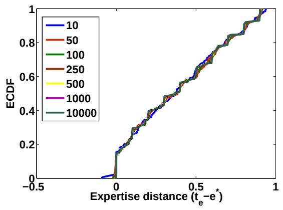


Fig. 8. Overestimating expertise.

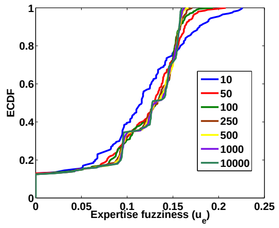


Fig. 9. Significant expertise uncertainty.

To reiterate, when a user is less reliable, we degrade the effect of his intense activity on many topics using Equation 8. We further need to update the distrust and uncertainty associated with the expertise opinion since it must hold $t+d+u=1$. Given

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

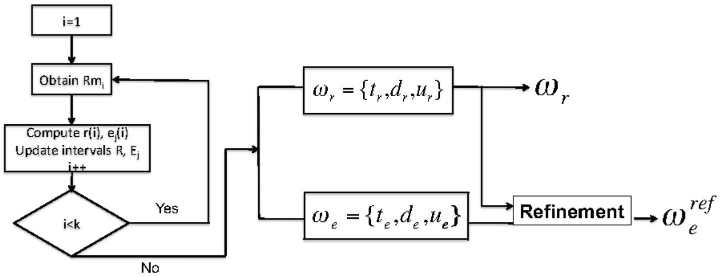

i=1
Obtain Rm$_i$
Compute r(i), e$_j$(i)
Update intervals R, E$_j$
i++
i<k
Yes
No
$\omega_r = \{t_r, d_r, u_r\}$
ω_r
$\omega_e = \{t_e, d_e, u_e\}$
Refinement
ω_e^{ref}

Fig. 10. Flow diagram of our assessment procedure.

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 proportionally to their initially assessed values:

$$d_e^{ref} = d_e + \frac{d_e}{d_e + u_e} \cdot t_{e,deg} \tag{9}$$

$$u_e^{ref} = u_e + \frac{u_e}{d_e + u_e} \cdot t_{e,deg} \tag{10}$$

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). In addition, falsely trusting an amateur is much more critical than having less trust in 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 9 and 10.

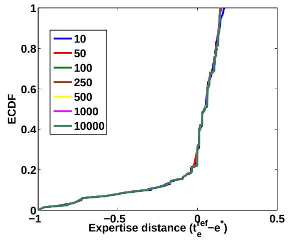


Fig. 11. Expertise distance with refinement.

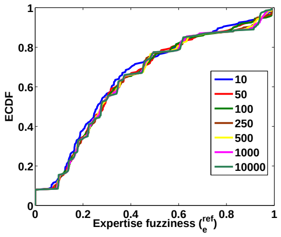
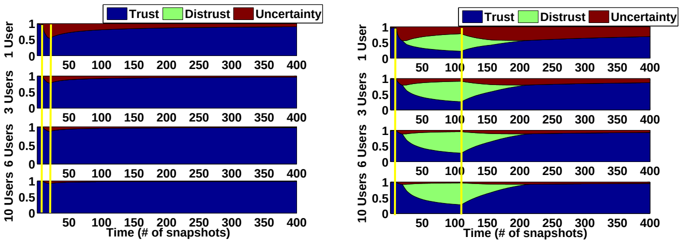


Fig. 12. Expertise uncertainty with refinement.

B. Consensus and 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.

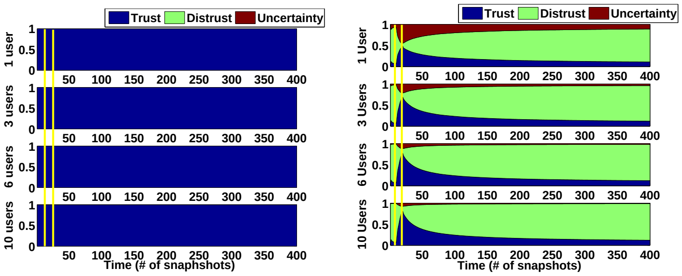
Our preliminary results [4] have shown that our scheme can follow the dynamic behavior of a user. In particular, we have seen that when users alter among the different types of behaviors, our framework can follow these changes. Recall, that when x snapshots have passed, we utilize all of them for the current assessment. In other words, the scheme as described until now exhibits a *full memory*. Later we will examine the performance using smaller number of snapshots (i.e., keeping only the most recent snapshots).



(a) Short perturbation

(b) Long perturbation

Fig. 13. User reliability.

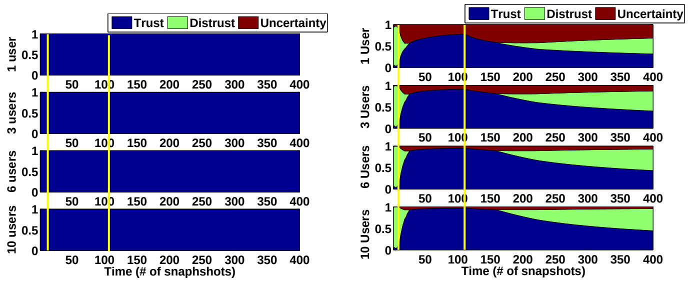


(a) *True* expertise topic

(b) *False* expertise topic

Fig. 14. User Expertise: short perturbation period.

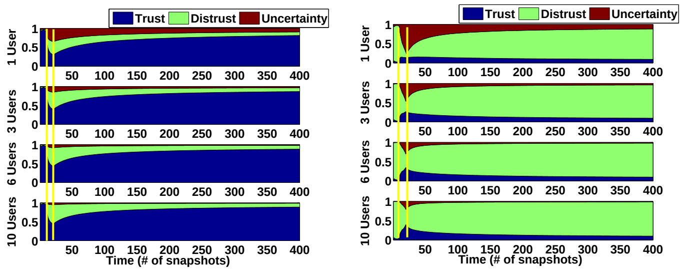
Consensus study: We consider dynamic scenarios where Alice is being monitored by a group of peers who collaborate towards obtaining a consensus on her reliability and/or expertise. In the scenarios examined, Alice is a “Reliable expert” but after some time, she perturbs for a period of time, when she 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 13(a) and 13(b) present Alice’s reliability for different



(a) *True* expertise topic

(b) *False* expertise topic

Fig. 15. User Expertise: long perturbation period.



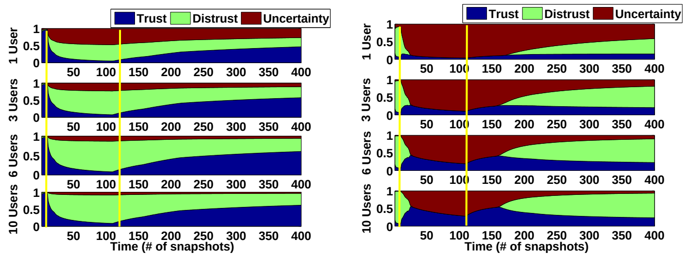
(a) *True* expertise topic

(b) *False* expertise topic

Fig. 16. User expertise: short perturbation period with refinement.

number of monitoring peers. 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 14 (short perturbation) and 15 (long perturbation) present Alice's estimated expertise for different number of monitoring peers (the vertical yellow lines identify the points of behavioral changes). 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 knowledgeable at all[4]. 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.

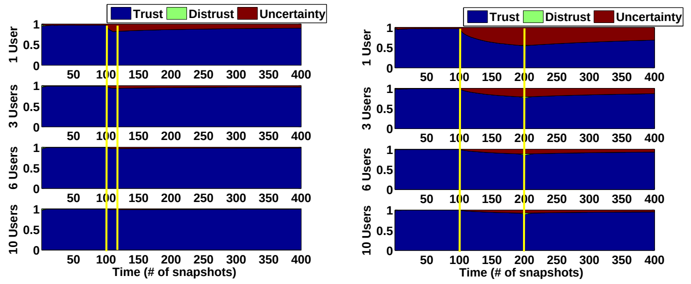
When no refinement is applied we observe that when Alice becomes *talkative* her assessed expertise is boosted in both types of topics (Figure 14(b) snapshots 10-20 and Figure 15(b) snapshots 10-110). This effect is pronounced with consensus. The reason for this is that consensus reduces uncertainty, thus, trust is increased. However, as one might anticipate from the results presented above, the refinement process eliminates the false expertise problem (Figures 16(b) and 17(b)). In other words, if we examine the reliability and expertise assessments in combination, we can identify the periods of false expertise



(a) *True* expertise topic

(b) *False* expertise topic

Fig. 17. User expertise: long perturbation period with refinement.



(a) Short perturbation

(b) Long perturbation

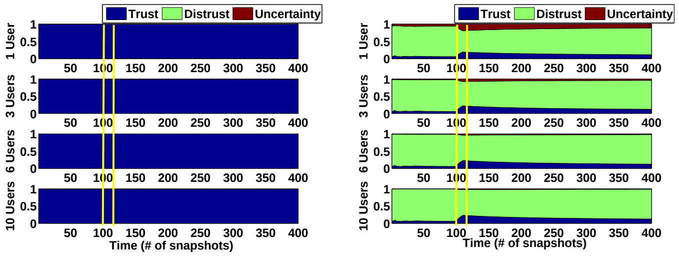
Fig. 18. User reliability for long initial "Reliable expert" period.

assessment, due to the low reliability of Alice. 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 16(a) and 17(a) during the perturbation period (snapshots 10-20 and 10-110 respectively). Nevertheless, 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 th 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 (18) - (22) 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 18(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 Alice'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 Alice. Keeping a long history makes the assessment scheme less responsive

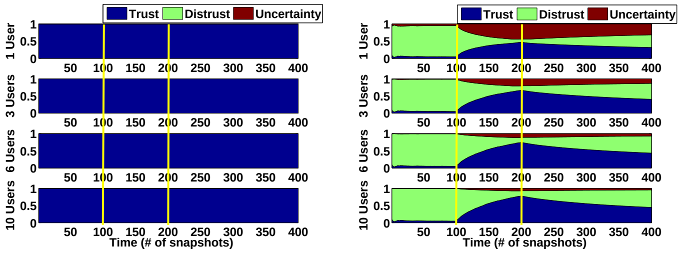
[4] Note here that, even for the expertise topic, there can be periods for which Alice is an amateur and has no knowledge for this topic as well. As aforementioned, this can correspond to periods where she is building knowledge, her account is compromised etc.



(a) *True* expertise topic

(b) *False* expertise topic

Fig. 19. User Expertise: short perturbation period and long initial "Reliable expert" period.



(a) *True* expertise topic

(b) *False* expertise topic

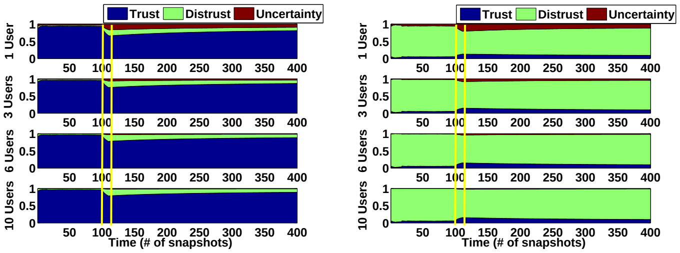
Fig. 20. User Expertise: long perturbation period and long initial "Reliable expert" period.

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 ϕ (from snapshot $x - \phi + 1$ to snapshot x).

We repeat our perturbation experiments with consensus computation with a history window of $\phi = 10$ snapshots (only the results with refinement are presented). 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 (in terms of speed of reaction) to the behavioral changes (e.g., Figure 17(a)). As our simulation results in Figures (23)-(25) 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 perturbation durations, however the results with other combinations of period durations are similar.

VI. Discussion and Conclusions

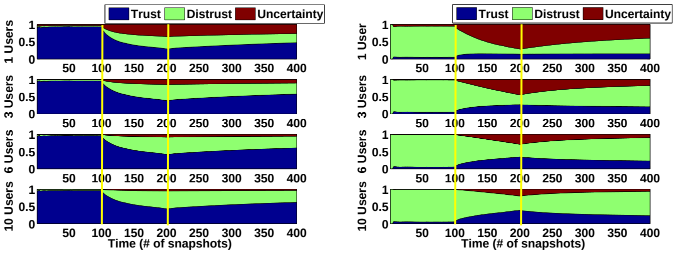
Before concluding we would like to emphasize on the ***limitations*** of our work. Even though the user model we are considering is both simple and realistic, it is not certain that every single participating peer follows it. For instance, an expert user might be selfish as well, being silent most of the time.



(a) *True* expertise topic

(b) *False* expertise topic

Fig. 21. User expertise: short perturbation period and long initial "Reliable expert" period with refinement.



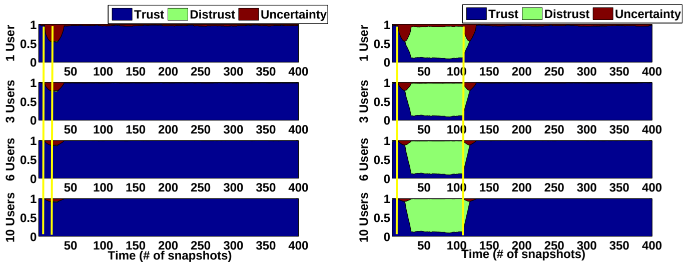
(a) *True* expertise topic

(b) *False* expertise topic

Fig. 22. User expertise: long perturbation period and long initial "Reliable expert" period with refinement.

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., many questions remain unanswered). In addition, despite the fact that we can identify "spammers" with the refinement phase, our scheme is not robust to the presence of **malicious** entities. 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.

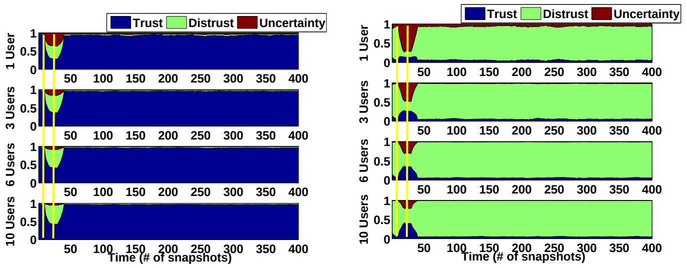
We would like to reiterate that currently we are only considering the presence of a reply or not, assuming that users strictly adhere to the cognitive profiles defined. Nevertheless, in reality the quality of answer is not binary. In the near future, we seek to utilize real data to perform the following necessary tasks: examine (i) the compliance of real users with the traits presented and (ii) any improvements possible by incorporating (assured) expert knowledge information and/or users' feedback. Note here also that, in a real Q&A network the pairwise interactions between users can be sparse. Assuming that all the questions are posted from a single *user* (i.e., the



(a) Short perturbation

(b) Long perturbation

Fig. 23. User reliability (Memory: 10 snapshots).



(a) *True* expertise topic

(b) *False* expertise topic

Fig. 24. User Expertise: short perturbation period (Memory: 10 snapshots).

network), we opt to examine modifications of our proposed scheme applicable to Q&A SNs with scattered interactions between users.

To conclude, 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 successfully estimate these attributes. Table I 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

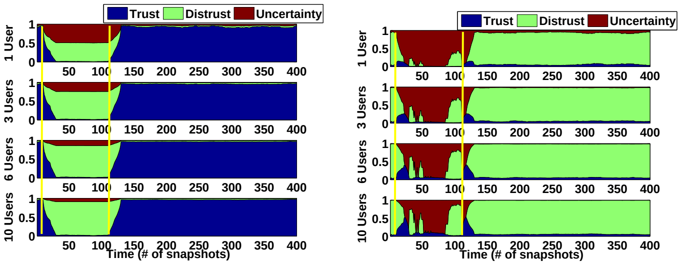
TABLE I
EFFECT OF THE VARIOUS MODULES OF OUR ASSESSMENT SCHEME.

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.



(a) *True* expertise topic

(b) *False* expertise topic

Fig. 25. User Expertise: long perturbation period (Memory: 10 snapshots).

[4] K. Pelechrinis, V. Zadorozhny, and V. Oleshchuck. A cognitive-based scheme for user reliability and expertise assessment in q&a social networks. In *WICSOC*, 2011.

[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. Hand, Y. Wang, and M. P. Singh. Operators for propagating trust and their evaluation in social networks. In *AAMAS*, 2009.

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

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

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

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

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

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

[21] 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.

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

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

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

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

[26] M. T. H. Chi, R. Glaser, and M. Farr. The nature of expertise. *Occasional Paper No. 107. Ohio State University, Colombus. National Center for Research in Vocational Education*, 1988.

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