Building a Robust Decentralized Recommender System for Web Credibility Evaluation

Alexandra Olteanu
LSIR, I&C, EPFL

***Abstract*—The size and diversity of today's online content makes it hard for the users to evaluate its credibility. To address this problem, we aim to build a fully-decentralized and robust recommender system to support web credibility assessment. In this proposal, we highlight three pieces of work that serve as starting points for our research, which we organize in a number of steps: (1) an extensive analysis of the factors involved in user credibility assessment, (2) designing a distributed recommender system that handles data sparsity and misrepresentation, and (3) devising incentives for users to cooperate and contribute to the system with ratings. In addition, we aim to employ a decentralized architecture that leaves users in control of their data.**

***Index Terms*—Web Credibility, Recommender Systems, Incentive Design, Peer-to-Peer Systems**

I. Introduction

TODAY'S web has shifted the way people communicate and obtain information. The online content has grown in size and diversity, becoming the primary source of information for many users. To avoid being affected by biased, misleading, or incorrect information, users need to be able to assess the *credibility* of the web pages they visit. However, this is a challenging task, since there is no systematic way of validating the web content. Users generally lack evidence about the two main factors that define credibility: *trustworthiness* (unbiased, well-intentioned, honest, good) and *expertise* (experienced, intelligent, knowledgeable, competent)[1]. Moreover, they may also be misled by factors such as the ranking in the search results, which may indicate popularity rather than content quality [2], [3].

It is therefore becoming important to produce effective tools that help users assess web page credibility. The recent research in this direction focuses on identifying the most relevant features to credibility assessments, helping users to evaluate the credibility of online information, and/or even to automate the credibility assessments [4], [5]. Their results show that augmenting the web content with additional information helps users in their evaluations.

Alas, the problem of providing the necessary infrastructure to support users in assessing credibility at the scale of today's web has not been fully tackled. In this proposal, we suggest a decentralized approach that addresses three practical challenges behind doing credibility evaluation at scale: (1) identifying the most impacting factors affecting web credibility evaluation, (2) designing a recommender system around these factors, and (3) designing incentives for users to evaluate web pages and disseminate evaluations. To support our presentation, we chose three representative papers that relate to the three aforementioned challenges.

Credibility assessment factors. The research on web credibility, mainly from social science and HCI, focuses on answering questions such as: how users evaluate credibility [6]; how appearance influences user credibility evaluation [4], [5]; whether the context (e.g., user task, involvement, beliefs, experience) affects users perception of credibility. We focus our presentation on the work done by **Schwarz and Morris [4]**, which aims to enable users to better assess credibility of search results and web pages by offering them additional information that is otherwise hard to infer (§II-A).

Recommender system for credibility assessment. In a nutshell, a recommender system aims to help users identify useful items when faced with many choices, by exploiting the community knowledge and opinions. For web credibility recommendations, the system needs to deal with large amount of items (web pages) and users (Internet users), while taking into account both user ratings and credibility impacting factors for providing personalized credibility recommendations. **TrustWalker [7]** is an example of a recommender system that takes steps in this direction, which we cover in §II-B.

Proposal submitted to committee: July 3rd, 2012; Candidacy exam date: July 10th, 2012; Candidacy exam committee: Prof. Matthias Grossglauser, Prof. Karl Aberer, Dr. Pearl Pu.

This research plan has been approved:

Date: _____

Doctoral candidate: _____
(A. Olteanu) (name and signature)

Thesis director: _____
(K. Aberer) (name and signature)

Thesis co-director: _____
(if applicable) (name and signature)

Doct. prog. director: _____
(R. Urbanke) (signature)

Incentive mechanisms for user involvement. The web involves skewed service interest and long-tail content. Moreover, the motivation impacts users in their credibility assessments [2]. As such, under different levels of involvement due to different tasks and needs, users devote different amount of time and effort to make credibility evaluations. It is therefore important to offer users proper incentives to both contribute to the system with ratings for the web pages they browse, and rate to the best of their knowledge. The **NABT incentive paradigm [8]**, which efficiently exploits the relationships among friends, represents a good starting point in this direction (§II-C).

Orthogonal to the three aforementioned challenges, the system should be supported by a decentralized P2P architecture. We motivate this design decision by the need to ensure neutrality and to leave users in control of their evaluations: Due to privacy concerns, users might be reluctant to reveal their browsing behaviour to a third party company. Additionally, a third party company storing all user data and controlling the recommender system might be interested in biasing the recommendation towards certain entities for commercial benefit. In contrast, we want to exploit the built-in trust relations between friends, by considering the recommendation of known/trusted people. In this way, we can also cast the problem as building a P2P social network in which users publish their opinions about the web pages they visit. We will discuss in more details in §III the challenges that our decentralized architecture needs to address.

II. Survey of the Selected Papers

A. Augmenting Web Content to Support Credibility Assessments

A thorough understanding of the factors that impact credibility evaluation is a first step in designing a practical system. In this regard, this section discusses the work of Schwarz and Morris [4], which provides important insights about what features are effective in helping users to better evaluate web content credibility. They target web page features that are hard to acquire by users, but which give valuable signals about web pages credibility. Assuming Definition 1. for a credible web page, they test whether these features improve the accuracy of user credibility evaluation (Hypothesis 1.)

Definition 1: ***Credible web page:*** A web page *"whose information one can accept as the truth without needing to look elsewhere. If one can accept information on a page as true at face value, then the page is credible; if one needs to go elsewhere to check the validity of the information on the page, then it is less credible."* [4]

Hypothesis 1: ***Hidden Features:*** Users would be better at assessing credibility if they had access to web page features that are non-obvious or harder to acquire, in addition to the immediate and conspicuous features.

Features Starting from the previous definition, features falling under three categories have been identified and collected, in order to validate the hypothesis:

- *On-page features* - features that can be computed solely based on the web site content, but are hard for the users to notice, e.g., spelling errors, ads, domain suffix.
- *Off-page features* - public information available only at third-party entities: e.g., awards, page rank, social popularity.
- *Aggregate features* - information owned by Internet companies, but not readily available to users, e.g., page visitors and their location, dwell time, revisitation patterns. Of a particular interest is *expert popularity*, which is heuristically obtained by labeling as topical experts those users who visited more than 10 times a whitelisted page in a certain topic.

Type	Features
On-page	domain suffix
Aggregate	overall popularity, expert popularity, geographic reach
Off-page	awards and certifications, PageRank

TABLE I
FEATURES WITH STRONGEST CORRELATION WITH GROUND TRUTH CREDIBILITY EVALUATIONS, AS SELECTED FOR THE FINAL VISUALIZATIONS.

Dataset To evaluate to what extent the collected features relate to credibility and how much they improve user credibility assessments, a dataset consisting of 1000 URLs was created. The authors first opted for five different topics exhibiting of both credible and non-credible online information. Then, based on popular search trends at the time being, they defined five queries per topic, and, for each of these queries, they considered the top 40 web pages returned by a popular search engine. Then, one of the authors rated the web pages using a five-point Likert scale rating. To ensure reliability, they collected ratings from topical experts for a small subset of web pages from the dataset, and confronted them with the original ones. By observing high agreement between expert ratings and the original ratings, the original ratings were accepted as "ground truth" throughout the study.

Content Augmentation To test whether the previously described features help users in their credibility evaluations, the web content needed to be augmented with visuals showcasing them. However, displaying all the features was undesirable. Since each feature would become less prominent due to information clutter, the augmentation risks having no impact, or being harmful altogether, as the Prominence-Interpretation theory [9] suggests. As such, only the features having the strongest correlation with the ground truth were selected for the final visualizations (Table I). To ensure that the features are salient to users, the authors opted for rich visualizations thematically grouped by interrogative questions: "when", "who" and "where", for web pages, and a more compacted view of "who" features for search results.

Results Schwarz and Morris conducted a user study in which participants were asked to rate several URLs from the dataset that fall under three topics: politics, health, and finance. Each URL was presented in search results and as a web page, both in the original form (basic), and augmented. While the politics URLs occurred in all 4 conditions, health URLs were used only for search, and finance URLs only for web pages. The data collected was used to test several hypotheses. We highlight the more general two:

1) *"Users' credibility ratings will be more accurate when*

the visualizations are available to them."

2) *"Users will feel more confident in the accuracy of their credibility ratings when the visualizations are available to them."* [4]

The obtained results suggest that people tend to overestimate the credibility of non-credible web pages, while underestimating the credibility of the credible ones. From the features used to augment either search results or web pages, users identified *expert popularity* as being the most informative, and declared to be more confident in their ratings when the visualizations are available. However, while search results augmentation proved to bring *significant benefit* (increasing both users confidence and accuracy), augmenting web pages brought little or no benefit.

Discussion Schwarz and Morris [4] suggested that offering users data about web page features not readily available to them improves the accuracy of their credibility assessments. It also gives insights about what additional information helps users to better assess the credibility of a web page or search result. For instance, one important finding for the design of our system is that information about experts behaviour or opinion may help users in better assessing the credibility of a web page. As such, our system should be able to identify experts and collect information from them. Yet, gathering information about the revisitation patterns of users in a decentralized environment is not straightforward. Moreover, their definition of expertise might lead to many false positives when labeling users as experts. For instance, someone could visit a whitelisted website only due to a temporary interest (e.g., having a cold). Thus, while we know that expert opinion is useful, finding experts and providing the right incentives for them to participate in the system is still an open question.

The authors note that assessing credibility is to some extent a subjective process, yet, they do not analyze how user involvement, task, knowledge, context, etc., affects users assessments. Several questions remain open: Are the provided visualizations more helpful in the context of a topic that in other? Are people evaluating credibility in different ways depending on the topic? Flanagin and Metzger indicate that credibility assessments vary by information type [2] and we believe that this should be further explored. Our preliminary results show that features correlate differently with users ratings in different topical contexts (§III). These give important insights for understanding if the credibility should be assessed differently depending on the context.

B. Combining Trust-based and Item-based Recommendation

Our system aims to support users when judging the credibility of web pages (for simplicity, we will consider that all web pages are static) by providing appropriate recommendations. A popular approach to building recommender systems is collaborative filtering [10], [11]. However, relying solely on this approach may be ineffective when dealing with a large number of items (i.e., billions of web pages), due to the sparsity of the user-item ratings matrix. "Cold start" users are particularly affected, and they constitute an important percentage (about 50% in Epinions [12]). In order to address this issue, trust-based recommender systems have been proposed [12]. Alas, even if these systems achieve better coverage, they also suffer due to sparse ratings, which may force them to consider ratings of weakly trusted users, thus affecting the recommendation accuracy. In this section, we cover TrustWalker [7], a system that addresses these issues by defining a random walk model, which integrates the item-based collaborative filtering approach with a trust-based one.

Recommendation Problem In the general case, the task of a recommender system is to predict ratings for unknown items for a given user. To do so, a recommender system considers a set of items $I = \{j = 1..M|i_j\}$, a set of users $U = \{k = 1..N|u_k\}$ and, for each user u, a set of items $R_u = \{k = 1..n|i_{u_k}\}$ for which the user assigned ratings r_{u_k,i_j} (e.g., a value on a likeart scale from 1 to 5). Additionally, when the recommender system exploits an underlying trust network between users, a set of trust relations $\{t_{u,v}\}$ is also considered (e.g., a real number between 0 and 1 – the higher the better – that expresses how much the user u trusts user v, or just a binary value that expresses either trust or lack of trust).

The problem is how to combine collaborative and trust-based approaches in order to overcome the "cold start" users problem and increase the coverage, while also achieving high accuracy, in a system with a large number of items and users, and a relatively small number of rated items per user.

TrustWalker Model The goal of the model is to find a good tradeoff between precision and coverage (i.e., offer recommendations for as many ⟨user, item⟩ pairs as possible). To achieve this, TrustWalker bases its design on the key observation that *ratings of trusted users on similar items are more reliable than ratings given for the target item by weakly trusted users*. As such, TrustWalker employs a random walk that combines trust-based and item-based collaborative filtering approaches.

In a nutshell, TrustWalker's random walk first tries to exploit the trust network by looking for the ratings for the target item at trusted nodes (*trust-based approach*). By design, as the length of the walk increases, the probability of using the ratings for similar items, instead of ratings for the target item, increases, as well (*item-based approach*). However, to accurately predict a rating on a target item for a given user, multiple random walks on the trust network are carried out, and their results are aggregated.

Next, we explain in more detail how the random walks are performed, the probabilistic item selection and how the ratings obtained from multiple walks are aggregated. We will use the same notations as in [7].

One Random Walk The random walk on the trust network starts from a user u_0, which needs the prediction of a rating for an item i_0. After k steps, the random walk is at some node u, from where there are three alternatives:

1) stop at the current node u, if it has a rating for the target item i_0 and return this rating r_{u,i_0};
2) otherwise, if $k <= max_{depth}$, where max_{depth} is the maximum number of steps that the random walk can do:
    a) with a probability $\phi_{u,i,k}$, the random walk stops; then, an item i similar with the target item i_0

which is rated by u is randomly selected, and its rating $r_{u,i}$ is returned. The item i is selected with a probability proportional with the similarity between i_0 and i:

$$P(Y_{u,i_0} = i) = \frac{sim(i_0, i)}{\sum_{j \in R_u} sim(i_0, j)} \quad (1)$$

where Y_{u,i_0} is the random variable for choosing i from all the items that u has rated. Note that the probability of selecting items that are not rated by u is set to 0.

b) with probability $1-\phi_{u,i,k}$, the random walk makes another step to a user v that is trusted by u. Node v is selected by the random walk with a probability of $\frac{1}{n_u}$, where n_u is the size of the set of users that are trusted by u.

3) if $k > max_{depth}$, the random walk is stopped in order to avoid considering ratings of weakly trusted users. max_{depth} is set to 6, so as to take into account the "six-degrees of separation" [13]. Note that, in this case, the random walk returns no rating for the target item i_0.

Items Similarity As previously discussed, when the random walk stops at a user that has not rated the target item i_0, the probability of selecting a certain item i, rated by user u and similar with i_0, is chosen to be proportional with the similarity between i_0 and i. Here, two items are considered similar if they are rated in a similar way by the same users. The size of the set of common users $UC_{i,j}$ that rated both items is important, since it gives more confidence in the obtained similarity, and reduces the chances for the similarity to be an artifact of the lack of ratings from common users. As such, the size of the common users is used in the similarity measure, but without favoring it too much:

$$sim(i, j) = \frac{1}{1 + e^{-\frac{UC_{i,j}}{2}}} \times corr(i, j) \quad (2)$$

where $corr(i, j)$ is the positive Pearson Correlation between the ratings expressed for both items.

Random Walk Stopping Probability The probability of stopping the random walk at user u is dependent on the *maximum similarity* between the items rated by u and the target item i_0: the higher the similarity between the target item i_0 and the items rated by u, the more likely it is to stop the random walk at user u. Additionally, the further the random walk goes from the source user, the less trustworthy the found ratings are. As such, the stopping probability should increase with the number of steps k that the random walk already preformed:

$$\phi_{u,i,k} = \max_{j \in R_u} sim(i, j) \times \frac{1}{1 + e^{-\frac{k}{2}}} \quad (3)$$

Recommendation The actual recommendation $\hat{x}_{u,i}$ is the aggregation of ratings returned by multiple random walks, and it is represented by the expected value of these ratings:

$$\hat{x}_{u,i} = \sum_{\{(v,j)\}} P(XY_{u,i} = (v, j)) \times r_{v,j} \quad (4)$$

where $XY_{u,i}$ is the random variable associated to stopping the random walk at node v, and selecting item j rated by v, when the random walk was started at user u, interested in item i.

Termination TrustWalker preforms several random walks until the estimate of the returned recommendation $\hat{x}_{u,i}$ is precise enough. To decide when this happens, the variance σ^2 in the results of all the preformed random walks is computed. If the difference between the variance σ_i^2 obtained after i steps and the one σ_{i+1}^2 obtained after $i+1$ steps is smaller than a given threshold ϵ, TrustWalker terminates. Note that, if the number of unsuccessful random walks surpass a certain threshold T, the pair $\langle$u, i$\rangle$ is considered uncovered.

[7] also mentions how the TrustWalker algorithm can be modeled using a matrix-based notation. However, we omit this discussion here, since we are interested in how TrustWalker can be applied in a decentralized manner. While the matrix model requires a global view of the network, the random walk can be applied in a local manner.

TrustWalker Properties By design, TrustWalker exhibits desirable properties:

1) *generality* – in its extreme cases, TrustWalker acts either as a pure item-based collaborative filtering recommender, when the random walk never starts ($\phi_{u,i} = 1$), or as a pure trust-based recommender, when the random walk stops only if it finds a rating for the target item ($\phi_{u,i} = 1$);
2) *confidence information* – the variance σ^2 in the results of different walks can be used to compute a confidence measure; the lower the variance, the highest the confidence should be:

$$confidence = 1 - \frac{\sigma^2}{\max \sigma^2} \quad (5)$$

where $\max \sigma^2$ is the maximum possible variance for the results;
3) *explainability* – TrustWalker can explain to users that a certain prediction was made based on a subset of trusted users and a subset of similar items with the target item.

Results The experiments were done on an Epinions dataset which has both the trust network and the ratings expressed by users, using different methods: user-based and item-based collaborative filtering, trust-based approaches and a few variants of TrustWalker. The methods are compared using four metrics: (1) *root mean squared error (rmse)* – to measure the recommendation error, (2) *precision* – a value between 0 and 1 that increases when rmse decreases, (3) *coverage* – the percentage of pairs of $\langle$user, item$\rangle$ for which a rating can be predicted, and (4) *FMeasure* – to test the prediction accuracy. The evaluation is done both for all users, and for "cold start" users only.

TrustWalker outperforms all the other methods in terms of coverage and FMeasure, while maintaining a good precision. Only a variant of TrustWalker, which performs a one-step random walk ($\phi_{u,i,k} = 0$), has a lower error rate than TrustWalker for "cold start" users, yet it has a much lower coverage for these users (i.e., more than five times lower than what TrustWalker achieves). We also note that when the

defined confidence measure exhibits higher values, the error rate is lower, thus proving the significance of the measure.

Discussion Given the huge amount and fast-paced growth rate of information available online, most of the web pages will probably not be rated by any user at all. Yet, we want to automatically evaluate the unrated web pages as well. TrustWalker partially addresses this problem since it also considers the ratings of similar items. Alas, in item-based collaborative filtering, only the common ratings are considered when measuring the similarity of two items. This means that TrustWalker is unable to provide recommendation for unrated web pages. This issue could be addressed if a content-based approach (in which the similarity between items can be computed based on their features) would be integrated in the TrustWalker model.

An important aspect of TrustWalker is that it is based on random walks, which can be implemented in a decentralized environment such as a P2P network. This is relevant for us, due to our choice of building a P2P-based system.

C. P2P trading in Social Networks: The Value of Staying Connected

It is unrealistic to assume that, in a real deployment of a credibility assessment system, users would play by the rules without deviation. Instead, they would find unequal benefit in using the system, which may also vary in time. Moreover, their interest in content is skewed ("long-tailed" content) and imbalanced (asymmetric and asynchronous over time).

This highlights the need for designing proper incentives that lead to a good tradeoff between the short-term need for quality, and long-term desire for the system's persistence. To address this need, we consider NABT (*Networked Asynchronous Bilateral Trading*) [8] an incentive paradigm that requires no global currency. NABT is built on top of a *social network* and requires friend users to keep track of a credit balance between them. It allows users to trade services (e.g., expert advice, ratings) with each other asynchronously, and even with remote peers via friends. This way, NABT addresses two issues of the traditional bartering systems: (1) *asyncronicity over users' needs* and (2) *asyncronicity over time*.

NABT: An Incentive Paradigm In NABT, users can benefit from services whenever they need them, without being constrained to provide back a service at the same time. When users receive services, they increase their debts at the friends who provide these services, and keep track of the credit balance between them. This way, NABT deals with asynchronous service requests over time. When a user needs a specific service that none of his friends can provide, he can still benefit from it through intermediary nodes (e.g., friends), provided there is a viable path to the node that can offer this service (i.e., all the intermediary nodes have enough credit along the path) – dealing with asymmetry of interest over nodes. When the debts reach some imposed limits, the services stop being provided. In order to pay the debts, the users need to provide back services. The NABT paradigm lies as a trade-off in an economic dimension, which has at its ends:

- *global currency systems* – in which users gain currency units when participating and conforming with the global rules. However, this incurs a high overhead for coordination, and relies on the good functioning of separate entities such as global reputation systems.
- *synchronous bilateral trading (bartering) systems* – in which users synchronously exchange services between each other, and which have no global currency and mechanisms to support it.

NABT Elements. NABT is built on top of a *social network*, and takes advantage of the trust built into the social links. It avoids using a global currency infrastructure, by requiring each user to keep track of a *credit balance* for each friend. This permits *asynchronous trading* between users, since whenever user A provides user B with a service, user A charges user B with a number of credits, increasing this way her credit balance. Additionally, user B can consume credits in form of services from user A only up to a *credit limit* (i.e., the amount of trust between two friends). If user B wants a certain service, e.g., expert advice, from user C, which is not a friend of B, B can still benefit from C's advice via intermediaries if there is at least one viable path between B and C, i.e., *trading via intermediaries*.

Handling Disputes. As in the real life, it is assumed that if two friends successfully provide services to each other multiple times, the trust (i.e., credit limit) between each other increases. Alas, even if NABT assumes users to be nodes in a social network (i.e., thus having friendship relations with other users) uncooperative users (e.g., free-riders) should be assumed as well. To overcome a potential impact these users might have on the good functioning of the system, an example solution is offered in the evaluated scenario, by adjusting the credit limit through the *Additive Increase Multiplication Decrease (AIMD)* algorithm. When a service was successfully provided, AIMD increases by α the credit limit up to a given maximum limit, and decreases it β times otherwise, which may lead to lack of connectivity for cheating peers.

Efficiency of NABT Let U be the set of users in a social network $G_S = (U, L)$, where L is the set of social links between users in U, weighted by credit limit $c_{i,j}$, $(i, j) \in L$. Let $G_D = (U, D)$ be the demand graph, where D contains weighted demand links, e.g., if user j request a service from user i, which i charges as $d_{i,j}$ units of credit we say that there is a demand link between i and j with an associated weight of $d_{i,j}$. We further denote the credit balance between i and j as $b_{i,j}$. Friends can exchange asynchronously services if the credit limit allows it:

$$-c_{j,i} \leq b_{i,j} \leq c_{i,j} \tag{6}$$

Additionally, users can trade via intermediaries if there is a viable path in the social network (i.e., a path with sufficient credit available – equal with the minimum available credit on links within this path – for a given demand). If the trade is possible for each pair of friends in the path, a credit transfer is done. The amount of credit that is transferred for a demand d between users i and j should satisfy *flow conservation* for

all nodes $u \in U$:

$$\sum_{i:(i,u)\in L} d_{i,u} - \sum_{j:(u,j)\in L} d_{u,j} = \begin{cases} -d \text{ if } u = i \\ d \text{ if } u = j \\ 0 \text{ otherwise} \end{cases} \quad (7)$$

We next define the balanced and unbalanced demands, and then we discuss NABT efficiency under both balanced and unbalanced demands.

Definition: ***Balanced Demand:*** We say that we have a balanced demand if the total cost of services demanded by any user $i \in U$ equals the total cost of the services requested from i:

$$\sum_{i:(i,u)\in L} d_{i,u} - \sum_{j:(u,j)\in L} d_{u,j} = 0 \quad (8)$$

Definition: ***Unbalanced Demand:*** We say that we have an unbalanced demand if there is at least one user $i \in L$ for which the total cost of services demanded by i does not equal the total cost of the services requested from i:

$$\sum_{i:(i,u)\in L} d_{i,u} \neq \sum_{j:(u,j)\in L} d_{u,j} \quad (9)$$

It has been showed that in a social network G_S, a balanced demand can be executed simultaneously if there is a direct link or a path between the users involved in the demand in G_S [8]. The set of demands can be represented as a graph, with links between each pair of requester-provider. The balanced demand execution is then obtained by decomposing the demand into same-cost finite loops in the demand graph. Executing each demand loop satisfies both the credit balance (6) and the flow conservation (7) constraints. As such, under a balanced demand, NABT behaves as a bartering system that supports *networked tit-for-tat*. In fact, NABT generalizes the synchronous bilateral trading, since it can be considered, given the balanced demand definition, that users play *tit-for-tat* with the whole network, rather than with a specific peer.

When the service demands between users are dynamic and unbalanced at a given moment in time, the credit limits absorb to some extent the temporary imbalance between peers. Let us denote all the nodes for which the service requests are bigger than the service demands as *service providers*, and those for which the service demands are bigger than service requests as *service consumers*. Let G'_S be the extended social graph of G_S with two new nodes s – *source* node connected to all the service providers, and t – *sink* node connected with all service consumers.

An unbalanced demand can be executed in a social network G_S if and only if the total number of credits for the overall demand is smaller or equal than the max-flow from s to t in the extended social network [8]. As such, the maximum service imbalance allowed by NABT is determined by the social network structure and the social link weights (i.e., the credit limit). Additionally, the executability of a demand at a certain moment t is independent of how the credit flows have been set up for past demands.

Results When compared with *Synchronous Trading (ST)* and *Global Currency Trading (GCT)*, *two-hop NABT* outperforms ST in terms of both *request success ratio* and the *waiting time* for fulfilling the service request, while being very close to GCT. The request success ratio is defined as the ratio of fulfilled service requests out of the total number of filled requests.

On the one hand, in GTC, users will have a high success ratio if they have enough credit to pay for the services they require. On the other hand, in NABT, a good connectivity increases the chance of having a high success ratio. This makes GTC and NABT obtain higher request success ratios than ST, but close to each other.

Nevertheless, in NABT, weakly connected peers wait more to find an available path through which to obtain the service. This is not the case in GCT, since once the supplying peer is located, if the requesting peer has enough credit, the service can be immediately obtained. As a result, GTC obtains a shorter waiting time.

In NABT, nodes that contribute more have a bigger chance to have a higher success ratio. Similarly, peers that act more frequently as relay nodes are more likely to have higher success ratio. Additionally, in NABT, free-riders have been found to obtain a much worse service compared with cooperative peers, while the impact on cooperative peers that end up in disputes with free-riding peers is limited.

Discussion NABT offers an efficient way to motivate users interested in receiving the service to cooperate and contribute back. Yet, as discussed in §II-A, users find particularly useful the service provided by experts (e.g., ratings, opinions). We expect the experts to be users that rarely need such a service in return. In this context, some questions arise: How to incentivize the expert users to participate in the network, if their benefit is much lower than the value of the provided service? Should experts have a different status in the network? In NABT, experts may choose to charge more for their service, increasing this way their balance with their friends and, accordingly, the probability to have a better request success ratio. However, this will not increase their need for services provided by others.

In the NABT evaluation, the peers are assumed to be connected by the underlying social network. However, in practice, a one-to-one mapping of the trading links (social connections) onto P2P links will overwhelm high-degree nodes (scalability issues) and leave unconnected the low degree ones (robustness to high-churn rate) [14]. For instance, this highlights the need for low-degree users to consider direct trading with non-friend users. If all their friends were offline, the low-degree users would have no intermediary nodes for trading.

III. Discussions and Future research directions

Our end goal is to build a robust decentralized recommender system for web credibility evaluation. We presented three relevant pieces of work that relate with the main steps that we plan to focus on:

1) identifying the factors that impact web credibility assessments;
2) defining a distributed recommender system model;
3) designing incentives for users to contribute with ratings;
4) employing a decentralized architecture that leaves users in control of their data and ensures neutrality.

We detail these steps below, in the context of our current work, and the research directions that we plan to take.

We argue that the first step towards achieving our goal is **identifying the most impacting factors** that affect users in evaluating credibility, and understanding how they impact the design of recommender systems and the underlying P2P infrastructure. In this context, we have discussed the work of Schwarz and Morris [4], who indicate that additional information about web pages, which otherwise would be hard to obtain, help users in better assessing their credibility. Moreover, users found the opinion of topical experts particularly useful.

Despite these findings being important for the design of our system, key questions still need to be addressed. First, we need a way to find expert users in our system. This problem is dual with expert finding in online social networks, since we also consider an underlying social network for our system. In our case, instead of producing content, users produce ratings for web pages, which can be exploited in a similar way.

To answer this question, we started devising algorithms for inferring online social network users expertise. In our first prototype, we infer topical experts from user profile data and produced content. In a nutshell, we independently infer the topic both from profile data, using the Alchemy API, and from user content, using LDA (Latent Dirichlet Allocation) and then test the topic confidence values. Our primary results show that this approach achieves high precision (up to 100%), but low recall (down to 14%) when applied on our dataset (i.e., a Twitter dataset that contains users profile data along with their tweets).

Additionally, we want to understand whether users assess credibility differently depending on the topic. We took steps in answering this question by first identifying a broad type of features that could affect the credibility evaluations. We have then applied statistical tests (e.g., Spearman correlation, one way Anova) to determine the subset of features that have the highest impact on user credibility assessments overall and per-topic. We did our analysis on the Microsoft dataset released by Schwarz and Morris [4].

Our preliminary analysis found popularity features to be the most informative for credibility assessments, followed by text-based features[1]. However, when looking at each topic, we found that while some features seem to be important for all topics, others are informative only for some topics (e.g., while the number of exclamation marks seems to be a good indicator for non-credible web pages that discuss health issues, it is not informative for web pages about politics). These results suggest that people assess credibility in a different way depending on the topic. This highlights another important question that we did not tackle yet: How important is the context and how is it defined? Our findings suggest that topical context matter, yet we have no indication on how, for instance, user experience and knowledge affect credibility assessments.

[1]For instance, popularity features include Google PageRank, Alexa Rank, number of Facebook comments and likes, number of tweets, and number of bookmarks, while some text-based features are the number of question and exclamation marks, SMOG score, and text informativeness.

A subsequent step is **defining a distributed recommender system model** that leverages the factors that impact the credibility evaluation. TrustWalker [7] is a promising approach, since it is based on a random walk that is suitable for a distributed environment. However, TrustWalker does not offer recommendations for items with no ratings, due to the item-based collaborative filtering approach that compares items based on the received ratings. Since we expect this to be a common case in web credibility evaluation, we plan to integrate content-based recommendation using a similar random walk model on a social network. Additionally, the recommender should be able to adapt its model for each topic context (e.g., by targeting users knowledgeable in this topic).

Another step is the **design of incentives** for users to evaluate web pages and disseminate evaluations, while ensuring the system robustness to uncooperative users. NABT represents a good starting point in this direction. However, in NABT, all users are treated uniformly, so there is no systematic way to incentivize expert users to actively contribute to the system. While their ratings are considered more relevant (and hence more valuable), expert users are unlikely to need services in return. As such, we plan to devise incentive models that will offer expert users a higher payoff when they provide ratings for web pages in their area of expertise. In this respect, one viable solution may be the implementation of a global P2P reputation system [15].

Finally, we want to design a P2P infrastructure to support the system, ensuring neutrality and leaving users in control of their data. We want to exploit the social relations between users, to offer more reliable recommendations and appropriate incentives to contribute to the system. As a result, the underlying infrastructure is expected to support workloads similar to a P2P social network. The unique set of challenges posed by the social network to its underlying P2P infrastructure is well studied [14], [16]: large population, targeted destinations, high churn rates, long-tailed content, high availability, timely information, connectivity, etc. In our previous work, we have showed how the P2P network can exploit the "small world" properties of the social graph in order to ensure its scalability and robustness to high churn rates [14].

REFERENCES

[1] B. J. Fogg and H. Tseng, "The elements of computer credibility," in *Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit*, ser. CHI '99, 1999, pp. 80–87.

[2] M. J. Metzger, A. J. Flanagin, K. Eyal, D. R. Lemus, and R. M. McCann, *Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment*. Lawrence Erlbaum, 2003, vol. 27, pp. 293–335.

[3] M. J. Metzger, "Understanding how internet users make sense of credibility: A review of the state of our knowledge and recommendations for theory, policy, and practice," 2005.

[4] J. Schwarz and M. Morris, "Augmenting web pages and search results to support credibility assessment," in *Proceedings of the 2011 annual conference on Human factors in computing systems*, ser. CHI '11, 2011, pp. 1245–1254.

[5] Y. Yamamoto and K. Tanaka, "Enhancing credibility judgment of web search results," in *Proceedings of the 2011 annual conference on Human factors in computing systems*, ser. CHI '11, 2011, pp. 1235–1244.

[6] B. J. Fogg, C. Soohoo, D. R. Danielson, L. Marable, J. Stanford, and E. R. Tauber, "How do users evaluate the credibility of web sites?: a study with over 2,500 participants," in *Proceedings of the 2003*

conference on Designing for user experiences, ser. DUX '03, 2003, pp. 1–15.

[7] M. Jamali and M. Ester, "Trustwalker: a random walk model for combining trust-based and item-based recommendation," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 397–406.

[8] Z. Liu, H. Hu, Y. Liu, K. W. Ross, Y. Wang, and M. Mobius, "P2p trading in social networks: the value of staying connected," in *Proceedings of the 29th conference on Information communications*, 2010, pp. 2489–2497.

[9] B. J. Fogg, "Prominence-interpretation theory: explaining how people assess credibility online," in *CHI '03 extended abstracts on Human factors in computing systems*, ser. CHI EA '03, 2003, pp. 722–723.

[10] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the 10th international conference on World Wide Web*, ser. WWW '01, 2001, pp. 285–295.

[11] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Commun. ACM*, vol. 35, no. 12, pp. 61–70, Dec. 1992.

[12] P. Massa and P. Avesani, "Trust-aware recommender systems," in *Proceedings of the 2007 ACM conference on Recommender systems*, 2007, pp. 17–24.

[13] S. Milgram, "The Small World Problem," *Psychology Today*, vol. 2, pp. 60–67, 1967.

[14] A. Olteanu and G. Pierre, "Towards robust and scalable peer-to-peer social networks," in *Proceedings of the Fifth Workshop on Social Network Systems*, 2012.

[15] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina, "The eigentrust algorithm for reputation management in p2p networks," in *Proceedings of the 12th international conference on World Wide Web*, ser. WWW '03. New York, NY, USA: ACM, 2003, pp. 640–651. [Online]. Available: http://doi.acm.org/10.1145/775152.775242

[16] S. Buchegger and A. Datta, "A case for p2p infrastructure for social networks - opportunities & challenges," in *Proceedings of the Sixth international conference on Wireless On-Demand Network Systems and Services*, ser. WONS'09, 2009, pp. 149–156.