The Perception of Others: Inferring Reputation from Social Media in the Enterprise

Michal Jacovi, Ido Guy, Shiri Kremer-Davidson, Sara Porat, Netta Aizenbud-Reshef
IBM Research – Haifa
Mt. Carmel, Haifa – 31905
jacovi@il.ibm.com

ABSTRACT
The emergence of social media allows people to interact with others all over the world. During interaction, people leave many traces behind that can reveal things about themselves, or about how they perceive others: having many followers may indicate that one is an influencer; forum answers that gain high ranking, are likely to testify for expertise; people who gain high ranking in eCommerce sites are likely to be trustworthy. In this paper, we examine whether public online traces can be used for inferring the reputation of a person as perceived by others in relation to trustworthiness, influence, expertise, and impact. We describe a study performed on indicators of reputation that employees leave in a rich organizational social media platform. We compare different indicators, and report the results of an extensive user study with over 500 participants who provided their perception of thousands of others through a set of hypothetical scenarios.

Author Keywords
Reputation; social media; enterprise; workplace; trust; influence; expertise.

ACM Classification Keywords
H.5.3 Group and Organizational Interfaces: Computer-supported cooperative work.

INTRODUCTION
Reputation is defined in the literature as an expectation about an entity's behavior, based on information about or observations of its past behavior. Social media provides access to information and observations about people's past behavior from their past public interactions. As users of social media, we leave public traces of our online interactions in the form of tweets, talkbacks, forum questions and answers, wiki edits, and more. These traces may serve as indicators, telling others about us. Other people's reaction during interaction with us, e.g., comments, re-shares, or "likes", form special types of indicators – they may testify to how we are perceived in the eyes of others. The sociologist Steven Nock [29] defines reputation as "*shared or collective perception about a person*". That is, the traces other people leave during interaction with us may serve as indicators to our reputation.

In this paper we propose a framework for inferring reputation from such indicators. We harvest social network information – people's interaction with other people and with their content, such as blogs, forum, files, and more – and suggest that this data may serve for inferring reputation.

Our research is conducted in IBM – a global IT company, with over 400,000 employees world-wide, in which the use of social media is well established for several years. This gives us the unique opportunity to examine a very rich set of indicators. While the majority of our findings are easily extensible outside the enterprise, our focus in this work is on *reputation in the enterprise*. This study is part of a larger research where we investigate the potential value in enterprise social media. In a related study, we examine the role of network size in the enterprise as well as different user roles in terms of contributions [24].

In a series of interviews we conducted with employees – managers and regular staff members – we discovered that people have a hard time referring to the term "reputable person". Our interviewees kept asking about the specific type of reputation we referred to, offering various scenarios that would yield, in their minds, different answers. Examples covered cases such as "*if you ask me whom I would like to work with, the answer would be different than if you ask me who made the greatest impact*" or "*I might pick the person who most people turn to for assistance – you know, the 'know it all', but then that would be different from, say, the person people look up to – the best 'mentor'*".

Analyzing the responses we received, we identified four principle characteristics that employees refer to as different types of reputation: the *influential*, the *expert*, the most *trustworthy*, and the person who made the largest business *impact*. Studying the literature, we realized that most works focus on one specific reputation type – with the vast majority studying *trust*, and a substantial body of research studying *influence*. *Impact* does not seem to have been studied in the context of reputation; while *expertise* has its own research domain that is not necessarily paralleled with reputation and is often evaluated by the users' presentation of self, as reflected in their email, files, or self assessments.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
CSCW'14, February 15 - 19 2014, Baltimore, MD, USA
Copyright 2014 ACM 978-1-4503-2540-0/14/02…$15.00.
http://dx.doi.org/10.1145/2531602.2531667.

In this paper we suggest that reputation has various types, and that examining them together may shed new light on the ability to infer reputation. We set out to examine the four reputation types that we identify as relevant in the enterprise: *Trust*, *Influence*, *Expertise*, and *Impact*.

Trustworthiness is the quality that makes other people feel that they can rely on a person and feel confident during interaction and collaboration. *Influence* is the characteristic that allows a person to reach a wide audience and affect the behavior of many people, be that by changing their way of thinking, or by introducing a new technology. While *expertise* is usually examined as an objective characteristic, implying extensive experience with a domain, in this paper we refer to the subjective quality that sets a person high above others as a good source for information or knowledge. Finally, by *impact* we refer to the footprint a person leaves on the business, or in our terms, the perception of this footprint in other people's mind.

Our research sets out to address the following questions: Can implicit traces in social media be used for inferring these reputation types? Are some types more easily inferable than others? Are different indicators more useful for inferring certain types? We examine the ability of inferring each type from individual indicators out of the rich set of social indicators at our disposal. We further continue to comparing the different types with each other, so as to determine which types have better potential to be inferred, and which indicators better serve for inferring each type. It should be noted that this is a preliminary study to examine the power of individual indicators. Future research sets out to examine aggregates of indicators.

The literature does not define a standard way for evaluating reputation. Evaluation often relies on the selection of one suggested algorithm or indicator as a 'ground truth', or on the comparison of different algorithms to each other. In this work, we looked for a direct way to "unlock" how people perceive their peers. In order to do that, we suggest an innovative evaluation method, by which we define a 'ground truth' of how people are perceived by others through a survey. In our survey employees are presented with hypothetical scenarios that reflect the need for specific types of reputation, and asked to point out the most suitable from a list of employees they are familiar with.

The design of our survey takes into consideration various difficulties, such as avoiding the need to single out any one specific person – for better or for worse; and engaging participants in an enjoyable and non-time-consuming process so as to ensure a large amount of responses. By gathering responses from over 500 employees about how they perceive tens of peers they are familiar with, we were able to calculate the public perception of a set of thousands of employees under different circumstances, requiring one of the four types of reputation. We then used this public perception as the ground truth for assessing the ability of our indicators to infer reputation.

The contribution of this paper is threefold: 1) Understanding how social media can be used for inferring reputation from **other people's** interaction with a person, 2) wrapping **several reputation types** in one study and crossing them over with a rich set of individual social indicators, 3) presenting an **innovative evaluation method**.

RELATED WORK

Reputation types and terminology. It is worth pointing that while the use of reputation systems is widely studied, there is still a lack of common terminology and means to exchange reputation information. This led to attempts to define and analyze reputation ontologies [1,4]. The term "reputation" is used in many works interchangeably with terms such as "*trust*" [19], or "*influence*" (as done by Klout.com). These models do not refer to the same notion, yet they all use the term "reputation", as indeed they study certain types of reputation. In this work we set out to compare, for the first time, how the same rich set of data may contribute to the inference of different types of reputation: *trust*, *influence*, *expertise*, and *impact*.

Works on *trust* were done by [8,20,21,38], encouraging trustworthiness, studying with whom to interact, and motivating users to enhance the quantity and quality of their contributions. In our study we present participants with a hypothetical scenario that pinpoints the people they would most *trust* with confidential information.

Influence is measured by commercial tools such as Klout.com, providing scores of users by capturing activities made on their posts, such as retweets, mentions, comments, posts, likes, and shares. Works on *influence* were also done by [30,33,37], examining various aspects of influencing other users, showing that popularity is not necessarily an indication of *influence*. In our study we aim to identify the people who are perceived to have the strongest *influence* on others in terms of reaching out to a wide audience for the dissemination of new technology.

Expertise location is highly correlated with identifying reputable individuals, and is at the core of many works (e.g., [25,27].]). In this work, we examine the notion of *expertise* from the viewpoint of other's perception. Rather than analyzing the users' own created content (e.g., emails, blogs, file system documents), we examine if they are perceived by others as experts.

Impact is the footprint a person leaves on the business, and this is how we present it in our survey. Steinfield et al. [32] refer to it as "social capital", providing a link between the internal use of corporate social networking and social capital.

Social Media in the Enterprise. Internal use of social media in the enterprise has been widely studied in recent years, covering information-centric tools such as blogging, wikis, social bookmarking [6,13,15,26], and even investigating the use of a corporate social network site to enable a more social-centric employee communication [7].

The work by Steinfield et al. [32] takes the study of corporate social networking further ahead by providing an empirical analysis of the link between the internal use of corporate social networking and social capital.

A body of research examines the mining of different relationship types from social media in the enterprise, including familiarity [9], similarity [11], and interest [16]. In all of these works, relationships are inferred based on the user's own activity, while in this work we use others' activities that relate to the user to infer reputation. Using graph terminology, this work is concerned with examining and comparing the "in-degrees", while previous works focused on "out-degrees" in social graphs.

While there is a significant body of research on the benefits of using corporate social information, there is no prior work, to the best of our knowledge, on the link between corporate social information and reputation. Our work thus lays the foundation for future work on social media as a base to assess the reputation of employees.

Evaluating Reputation. The literature does not define a standard way to evaluate reputation as it does, for example, for evaluating search engine results. A common evaluation method relies on comparing the outcomes of various algorithms – e.g., Kuter & Golbeck [21] evaluate their trust values by comparing them against the performance of TIDALTRUST, an algorithm they studied before; while other studies suggest several forms of calculating reputation and report the comparison between these calculations [3,14,18,31]. Another evaluation method suggests splitting the data set into a training set and a testing set, examining whether the algorithm can predict reputation values of the testing set based on the training set [2,8]. Finally, another evaluation method relies on one of the reputation indicators as the 'ground truth' [34]. Some studies do not provide any evaluation [14,27]. We suggest a method for tapping into the perceptions of real users, in order to establish a ground truth for evaluating reputation.

Characteristics of reputation systems. Some reputation systems are based on collecting *explicit feedback* on users – directly asking users about other users [8,19,30,38]. As stated by McNally et al. [28], systems based on explicit feedback are vulnerable, as in many cases the information given is false, for various sociological reasons. This is even more acute in the enterprise, where accountability inhibits people from providing negative testimony. Other systems use *implicit indicators* such as number of followers, retweets and mentions in Twitter [3,18], or number of comments made on content authored by the user. Our work examines a wide variety of implicit indicators, which are side-effects of people's interaction over social media inside IBM. Our focus on implicit indicators ensures feedback that is less vulnerable, while also minimizing the effort of the users.

Resnick et al. [30] explore the principles of *external reputation systems*, where reputation scores are made public and revealed to users. In eBay, for example, the buyer and seller can rate each other after a transaction [20]. Users have feedback points visibly attached to their screen names. In contrast, *internal reputation systems* do not reveal scores to users, but use them under the scenes to support in assessing content quality or user reputation in applications such as community question answering or news forums [2,5,14]. Our work focuses on studying the quality of reputation indicators (referred to as "criteria" by Kardara et al. [18], "features" by Bian et al. [2], or "metrics" by Maresch [23]), with no references to the application that uses them. Our contribution is thus expected to be valuable for both external and internal reputation systems.

RESEARCH SETTINGS AND METHODOLOGY

We conducted our research inside IBM – a large global corporation that employs a rich social media platform – IBM Connection[1] – for internal use. The platform is deployed for over five years, and individual components of it were deployed for years before that. It is therefore well disseminated within the company.

IBM Connections supports the following social applications: a **blogging** system that allows users to write blog posts, comment, or 'like' others' posts; a **board** application that allows posting messages on other individuals' boards (sometimes called "walls" in social media) and comment on existing posts; a **file sharing** system that allows users to upload files they authored. Users can share a file with other users (even if they are not the authors), download, comment on, or like files; a **forum** system that allows users to create new topics for discussion, or to comment on others' topics; a **people tagging** application where users can tag each other with descriptive tags; and a system that allows users to create **wikis** where pages can be co-edited by multiple users. In addition, IBM Connections allows one user to **follow** all public activities of another user, and two users to reciprocally connect (or "**friend**") with each other.

We claim that many activities in these applications, either through direct interaction of people with each other or through interaction of people over artifacts, can serve as indicators of how people perceive each other – indicators of reputation. We set out to examine the potential of the in-degree of the interaction graph of the various indicators to infer various types of reputation.

For each of the eight social applications listed above, we selected activities that may serve as such indicators. For example, an activity of liking a blog post is a positive indication about the author. Similarly, downloading a person's file, following one's activities, or writing on one's

[1] http://www-03.ibm.com/software/products/us/en/conn/

board, may all serve as indications about how one is perceived. We used the applications' APIs to harvest all public activities of these types over the past two years.

An outcome of a single indicator is a list of employees with a value assigned to them by the indicator (e.g., a list of all employees who have followers, along with the number of followers per each; or a list of all employees whose blog posts were liked, along with the number of people who liked one or more of their posts). The returned values may be used for calculating reputation scores; and the returned lists, when sorted, may be paralleled to the results returned by a search engine – a list of results, ranked and scored by their relevance. Our goal is to evaluate the returned lists in comparison to a ground-truth, much as search engine results are evaluated in comparison to ground-truth results, typically annotated by a small group of individuals [22], thus examining the potential value of the returned lists for inferring people's reputation of the various types.

Table I details the indicators we inspect, along with the number of people for whom information could be extracted (i.e., the number of individuals each indicator covers) in the past two years. The coverage range is high: from around 1,000 people whose files were liked, to over 100,000 with friends. Our goal is to examine the potential accuracy of the indicators, regardless of their current coverage; our analysis is thus oblivious to the number of people covered by the indicators and examines their potential to reflect reputation.

App	Indicator Semantics	#People
Blogs	Comments on one's blog posts	4,205
	"Likes" of one's blog posts	3,884
Profiles	Posts on one's board	27,027
	Replies on one's board	4,312
Files	Comments on one's files	1,439
	"Likes" of one's files	1,010
	Downloads of one's files	6,923
	Shares of one's files	2,924
	Sharing (others') files with one	10,193
Followers	People following one's activities	24,113
Forums	Replies on one's posted topics	5,935
Friends	Friends (reciprocated)	100,942
Tags	Person tags associated with one	35,048
Wikis	Edits of pages of wikis one owns	10,854

Table I. Indicators examined in our study

We observe that our indicators could be classified into two classes that may have a different interpretation to how other people are perceived. One class of indicators deals with activities done directly between **people** and thus directly reflects on how a person is perceived. These are: friending and following which are direct relations between two people; board, which allows people to communicate with a person directly through a "wall"-like page; and tagging, in which a person assigns a tag directly to another person. The other class of indicators refers to activities in which people interact over **content** artifacts. These are blogs, files, forums, and wikis, in which other people interact with a person's artifacts (blog posts, files, forum topics, and wiki pages) and this interaction indirectly reflects on how the person is perceived in their eyes.

In order to evaluate the potential of our data to serve for reputation inference, we conducted two types of analysis: on the one hand, we compared the outcome lists of the different indicators with each other, to understand the differences in their potential. The method we used in order to compare the lists is *match@k* [10], which measures the percentage of common people between the top *k* individuals in each of the lists.

On the other hand, we studied the potential of indicators to reflect a person's reputation. To this end, we defined a "ground truth" through a broad user study in which hundreds of employees provided us with their perception of thousands of other employees they are familiar with. Given a list of people returned by an indicator (e.g., the 10,854 people whose wikis were edited, ranked by the number of people who edited them), we examined the relationship of this list with the "ground truth" using the popular *normalized Discounted Cumulative Gain* (NDCG) measure [17], typically used to evaluate the results of search engines against a given ground truth.

For the user survey, we focused on employees for whom we had information from at least three of the eight applications (i.e., they were "covered" by at least one type of indicator from at least three applications) in order to ensure we had a good base to assess their reputation. Our aim was to understand how the peers of these subjects perceive them.

In order to locate peers to participate in our survey, we used the SONAR *familiarity* list, which is based on aggregated data stemming from social media applications and has been shown to produce accurate results for social media users [9]. We identified 2,474 employees who had at least 10 of our subjects in their familiarity list, covering 5,695 subjects. We chose to focus on these 2,474 as survey participants since they could provide us explicit feedback on at least 10 people they know and for whom we have substantial information from the inspected indicators.

Our goal in the survey was to understand how participants perceive their peers in the four reputation types that we set out to study: *trust*, *influence*, *expertise*, and *impact*, as well as in a general *reputation* notion. Asking a direct question about how people are perceived is a delicate matter. In order to avoid being so direct, we developed a unique type of a survey: we provided our participants with the list of their 10 peers, and asked them to choose up-to-three of them who best match a certain hypothetical scenario. The option of choosing up-to-three, freed participants from the difficulty of picking just one person, and from the need to "put down" too small a group of peers (as would have been the case, for example, in one-on-one comparisons).

The hypothetical scenarios we presented to our participants were focused around the types of reputation. Each participant was first introduced with a scenario about the general notion of reputation, followed by four other scenarios representing the four types of reputation that we study. The latter four were presented in random order.

While a more elaborated method would employ multiple scenarios per reputation type, our goal to reach out to hundreds of respondents and thus the need to simplify the survey led us to stick to a single scenario per type. In order to establish the validity of the scenarios in representing the reputation type we intended for it, we conducted a preliminary survey in which 19 participants were presented by the scenarios in random order and asked to match them to a set of eight reputation types (our five along with "leader", "visionary", and "meticulous"; choice of multiple types per scenario was enabled). The 19 validators originated from a group of 30 individuals invited to take part. They come from four different countries. 9 were females, 10 males. The results over all scenarios indicated that our intended type is indeed the highest one selected – with a consensus of 14 to 19 out of 19 participants ("*reputable*" reached a consensus of 14; "*influence*" and "*impact*", 16; and "*trust*" and "*expertise*" a straight 19 of 19 consensus). These are the scenarios:

Reputation: CNN is making a new documentary series about reputable people. Gamma[2] was asked to identify a group of the most reputable Gamma employees in various fields, to be included in a chapter about IT companies. Imagine you are sitting on the selection committee, and that the list below contains the 10 finalists. Who would you nominate?
Trust: You just got some really exciting news, but were asked to keep it under wraps until the formal announcement. You were given permission to share this news with up to three people. Suppose these people you were going to tell are ones from the list below. Who would you tell?
Influence: A group at Gamma Research has developed a new tool for internal use. The group is looking for early adopters who would be willing to use the tool and later spread it across Gamma. Imagine you are sitting on the committee responsible for picking these early adopters. The list below contains the 10 finalists. Who would you pick?
Expertise: Gamma has created a new forum of experts, and is looking for people who are experts in their field to act as the core team of the forum. Imagine you are sitting on the committee responsible for picking this core team. The list below contains 10 finalists. Who would you pick?
Impact: To celebrate its second centennial, Gamma is announcing an award for people whose work had the most impact on the company. Imagine you are sitting on the committee responsible for picking these people. The list below contains the 10 finalists. Who would you nominate for the award?

The reason we opened each set of questions with the general *reputation* scenario is that we aimed to capture people's natural interpretation of reputation without being "contaminated" with the different types. The hypothetical scenarios help focus on a specific question, sparing the need to interpret abstract terms such as "*trust*" or "*influence*" which are typically hard to define.

Each survey participant received a different set of 10 people to select from, based on their computed top familiarity list. In order to collect more data, those of the 2,474 participants who happened to be familiar with more than 10 of our subjects, were presented with more than one set of 10 people. Those who were familiar with over 20 subjects, were presented with two sets – one set with their 10 most familiar subjects, and another with the next 10 familiar subjects. Those who were familiar with over 30, were presented with three sets – with the third set including subjects in positions 21-30 on their familiarity list. No one was presented with more than three sets. Participants could quit the survey at any time.

In all scenarios, we asked participants to mark those on the list who are unfamiliar to them, in order to exclude these from our analysis. We use the notation "presented" to denote the 10 subjects given in a set. Of these, the participant would mark "unfamiliar" people that we remove from the analysis. The rest of the people – those presented and found to be familiar, are denoted as "mentioned". Of these, the user would mark up-to-three "selected" subjects.

The outcomes of the survey provided us with data about people's perception of their peers in five different types. We aggregated all votes of a type into a "ground truth" ranking based on a simple measure denoted ***s/m*** – the number of participants who "selected" a subject divided by the number of participants to whom this subject was "mentioned". In our ground-truth ranking, we included only individuals who were mentioned to at least three survey participants, to make sure our ***s/m*** measure is robust.

To measure the quality of lists of people stemming from the indicators, we set out to assess the ranking (weak ordering) of people by the various indicators, in comparison to the ground-truth. We focused on the ranking of people in the lists rather than on the scoring, since different indicators use different scoring measures (e.g., number of followers vs. number of files shared). For this reason, we used the *normalized Discount Cumulative Gain* (NDCG) measure [17], which is commonly used in the Information Retrieval domain to evaluate the ranking of a search engine's results. In our case, we considered the ***s/m*** score as the graded relevance score of the person, and the lists extracted from the indicators as the search engine results. For a given p, NDCG(p) considers both the graded scores of the top p results and their relative order as retuned by the indicator.

[2] A pseudonym is used

For each indicator, we considered only people for whom we had a graded score from the survey (i.e., they were mentioned to at least three participants), relying on the fact that the survey provided over 1,000 scored individuals for each type, thus having a high coverage of the lists returned by the different indicators.

RESULTS

We sent invitations to 2,474 employees to take part in our survey. 554 people responded, going through 1,073 sets. 237 (42.78%) of the participants responded on one set, 115 (20.76%) on two sets, and 202 (36.46%) on three.

The distribution of the 1,073x10=10,730 responses between the various sets may be seen on Table II. 1,445 (13.46%) of the presented subjects were marked as unfamiliar – a rather low percentage, indicating that our chosen mechanism for locating familiar people is effective. Naturally, the percentage of unfamiliar people grows from the first set to the second and third, as familiarity with the respective individuals is weaker. For the remaining 9,285 responses (3,320 unique subjects), whose distribution among the sets may be seen on the "mentioned" row, we examined the percentage they were selected by participants for the various scenarios. The result is at the bottom of Table II.

	Set 1	Set 2	Set 3
Presented	5540	3170	2020
Unfamiliar	419 (8%)	541 (17%)	485 (24%)
Mentioned	5121	2629	1535
Selected (% out of Mentioned)			
Reputation	23%	22%	22%
Trust	23%	19%	18%
Influence	23%	23%	23%
Expertise	24%	24%	23%
Impact	18%	19%	19%

Table II. Number of people presented, unfamiliar, selected

The percentage of selected people on the first set was about 23% for almost all types, indicating a slight inclination to choose two or three people over just one. The exception is *impact* for which the average is only 18%, indicating people preferred to choose just one or two people in this scenario. These percentages remain steady in the second and third sets, with the exception of a drop in *trust* to 18%, which can be explained by a stronger tie between familiarity and *trust*, leading to fewer people being selected as the familiarity with people in these sets is weaker.

Calculating the ***s/m*** score for those who were mentioned to at least three participants resulted in 1,132 unique subjects with ***s/m*** score for *reputation* and 1,065-1,072 subjects with ***s/m*** score for the other types. We believe this is a rich basis for comparison between the indicators.

Comparison across the types. Realizing that the percentage of selected subjects is quite similar in all types, we wanted to get an indication of how similar the selections are – are the same people being chosen as top results of all types, or do participants indeed convey different semantics to different types. We were especially interested in the *reputation* type, to see which other types are close to it.

The method we used in order to compare the lists returned by the different types is *match*@100 [10], where we compare the top 100 people receiving the highest score in each type with the top 100 of all other types. The results are presented in Table III.

	Reputation	Trust	Influence	Expertise	Impact
Reputation	-	29	18	35	39
Trust		-	44	29	53
Influence			-	27	48
Expertise				-	41
Impact					-

Table III. *match@100* across the types

Overall, results are distinct: *match@100* does not surpass 53%, showing that participants selected differently across types and that there are no redundant types we can ignore – this supports our suggestion to distinguish between types, and our choice of types to be examined. *Impact* has the highest overlap with *reputation*, followed by *expertise*. *Influence* has the lowest overlap with *reputation*.

Inspecting Content Indicators. For content indicators (blogs, forums, files, wikis), we experimented with several methods for calculating a person's reputation. The differences between the methods stemmed from two factors: (1) Calculating based on counting **content** artifacts vs. **people**. E.g., for wiki editing, counting the number of edits to the person's wikis, vs. counting the number of unique people who edited them. (2) Calculating **numbers** vs. **averages**. E.g., for wiki editing, counting the overall number of edits on the person's wikis, vs. the average number of edits over them.

We first set out to examine our raw data – using *match@100* for comparing the lists of people stemming from our indicators, examining the above two factors. The results were quite similar across all indicators and we depict three representative examples in Table IV.

	Num Artifacts to Avg Artifacts	Num Artifacts to Num People	Num Artifacts to Avg People	Avg Artifacts to Num People	Avg Artifacts to Avg People	Num People to Avg People
Blog Comments	19	75	18	21	81	21
Forum Replies	46	67	36	48	75	50
Wiki Edits	11	27	7	13	40	27

Table IV. Artifact-people and num-avg *match@100*

It can be seen that typically there is a higher match between lists based on number of unique people and lists based on number of artifacts as well as between those based on average number of unique people and on average number of artifacts. This gives an indication that counting people or counting artifacts typically leads to similar results. For

wikis the match is lower, possibly since the same people make many multiple edits. The match between number and average is lower in general, indicating that averaging yields quite different results, motivating the comparison of the two approaches.

Equipped with these observations, we experimented with four score calculation methods per each content indicator, using our survey results as the ground truth. We observed similar patterns for all content indicators. Figure 1 demonstrates them for blog commenting and wiki editing, across all types, for NDCG(30).

As was already hinted by the match of the people lists, number of people and number of artifacts produce similar results (two top lines in the figure); similarly, the average of people and average artifacts produce similar results as well. It seems that in all cases but *trust*, using the number (of either people or artifacts) is closer to the ground truth than using the average. We conclude that taking the number of people or artifacts rather than their average, typically improves the reputation inference capability.

The differences between considering the number of people vs. number of artifacts were minor, giving a slight preference to one or the other for different types. Based on these outcomes, we report on results stemming from counting the number of people as a unified basis for comparison.

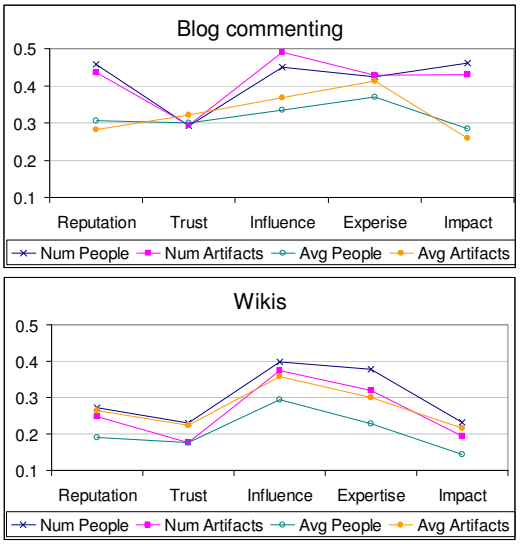

Figure 1. Four calculation methods for blog commenting and wikis

Comparing Indicator semantics in the same application. Some applications contributed more than one type of indicator, with different semantics. Blogs contributed two indicators: commenting, and liking. Similarly, boards also contributed two indicators: posting on one's board (creating a thread), and adding a comment (to a thread) on one's board. Files contributed as many as five indicators: commenting, liking, downloading, sharing (one's file with others), and being shared with (when a person shares a file with one, indicating a certain care for the one). We set out to examine the differences between the indicators of the same application.

For blogs, *match@100* between comments and likes was rather low at 32%. For board, the *match@100* between number of people who posted on one's board and number of people who commented was a high 68%. The *match@100* comparison of the file indicators are shown on Table V.

	Comm.	Like.	Down.	Share.	being Shared
Commenters		73	9	10	7
Likers			10	11	7
Downloaders				38	9
Sharers					12
being Shared					

Table V. Comparing Five Files Indicator semantics using *match@100*

As could be expected, being-shared-files has the lowest match with the rest since it represents a different semantic that is more person-centric, while the rest refer to different forms of feedback on the user's content (files). Commenters and likers have high match (i.e., people who have high number of commeners are also likely to have high number of likers and vice versa), and to some degree downloaders and sharers are also matched. Still, the match between all feedback forms is low, motivating the comparison between them using the ground truth.

Table VI compares the NDCG(30) of the different indicators within each of the source applications that had more than one indicator type. For both blogs and boards, commenting consistently performs slightly better, implying it's a slightly stronger feedback form for inferring reputation types. For files, as expected, being-shared behaves differently than the rest across the types. It has a clear edge in *reputation* and *impact*, but is very weak for *influence* and *expertise*, where downloading is the best. Liking is strongest for *trust*.

App	Indicator Type	Rep	Tru	Inf	Exp	Imp
Blogs	Commenting	**0.46**	**0.29**	**0.45**	**0.42**	**0.46**
	Liking	0.40	0.26	0.43	0.37	0.42
Board	Posting	**0.44**	**0.21**	0.36	0.34	0.36
	Commenting	0.43	**0.21**	**0.41**	**0.37**	**0.39**
Files	Commenting	0.34	0.28	0.38	0.32	0.33
	Liking	0.37	**0.34**	0.39	0.34	0.31
	Downloading	0.36	0.28	**0.42**	**0.40**	0.34
	Sharing	0.26	0.19	0.33	0.32	0.31
	Being shared	**0.44**	0.28	0.24	0.29	**0.38**

Table VI. NDCG(30) Comparison of same-application indicators

Comparing eight indicators representing eight apps. For the cross-application comparison, aiming to understand if any one application is a better indicator for any type of reputation, we chose the indicators that perform better from each application that has more than one indicator. These are the indicators we compared: board commenting, following, tagging, friending, blog commenting, file downloading, forum replying, and wiki editing. As before, we start by examining the raw data stemming from the indicators – matching the lists of people they produce. Table VII shows the results for all eight representing indicators.

	Board	Followers	Taggers	Friends	Blogs	Files	Forums	Wikis
Board	.	47	41	36	22	24	9	0
Followers		.	37	29	27	18	7	0
Taggers			.	29	18	22	6	0
Friends				.	17	21	4	1
Blogs					.	15	9	0
Files						.	3	
Forums							.	0
Wikis								.

Table VII. Comparing eight indicators representing eight apps using *match@100*

The match among the eight indicators is generally low, all under 50%, indicating that the different applications produce different lists and motivating the comparison between them. There is clearly a higher match across the four people-indicators (left side of the table), probably since they reflect a perception of the person as a whole rather than with regards to a specific content. Overall, the low match motivates comparing all to the ground truth.

We examine the eight representative indicators in light of each of our five types. For each type we present below (in Figures 2-6) the comparison of each indicator with the ground truth as defined by the survey outcomes (using NDCG(p)). For saving space, each graph is depicted in its optimal scale.

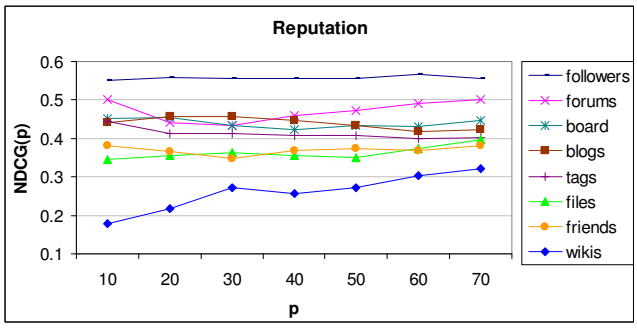

Figure 2. Comparison of the eight sources as indicators for reputation

For inferring the general notion of *reputation* (Figure 2), the followers indicator is clearly best – a person with many followers is likely to be perceived as *reputable*. Forums and the boards are also quite high – receiving responses from many people on a forum, or comments from many people on the board, are indications of being *reputable*. It seems that having many friends is not a good indicator for *reputation* and neither is having many people who edit your wikis.

The NDCG results for *trust* (Figure 3) are much lower across all indicators, suggesting that *trust* is more difficult to infer. This is an expected result as *trust* is a more personal type. Surprisingly, the friends indicator fails to be useful for inference here, too. Forums and followers seem to be the most useful for *trust* inference.

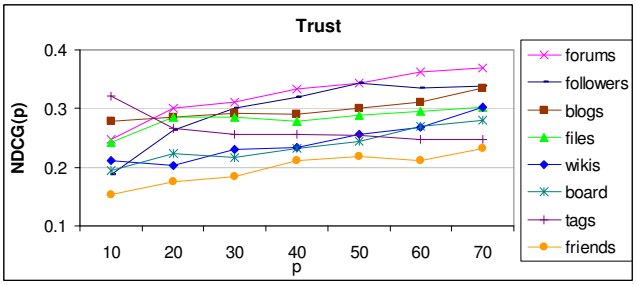

Figure 3. Comparison of the eight sources as indicators for trust

The picture for *influence* (Figure 4) is different than for *reputation* and *trust*. Followers, does not seem to be strongest for *influence*. Previous research on Twitter [3], showed that the number of followers is not a very strong indicator for likelihood of retweet (a form of *influence*), which is in line with our result. Here, content indicators such as blogs, forums, and wikis seem to be stronger.

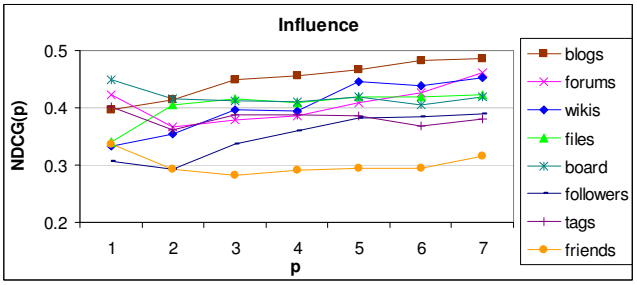

Figure 4. Comparison of the eight sources as indicators for influence

The graph for *expertise* (Figure 5) is similar to that of *influence*. Content indicators are strong here, too, as may be expected since *expertise* is associated with content. Friends are clearly last here, as in the previous three graphs, rendering it a rather insufficient indicator for inferring any type of reputation.

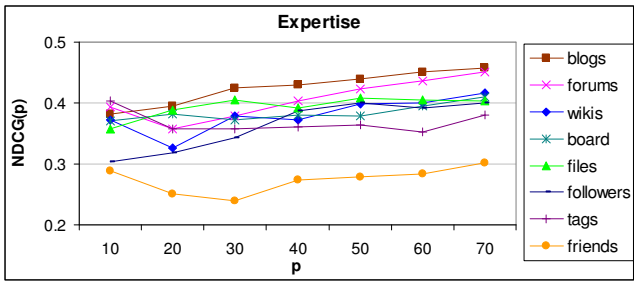

Figure 5. Comparison of the eight sources as indicators for expertise

As was already noted when examining the raw data (Table III), *impact* has the highest match to the *general reputation* type. The NDCG results are also quite similar between *impact* (Figure 6) and *reputation* (Figure 2), with the followers indicator even more dominant for *impact*.

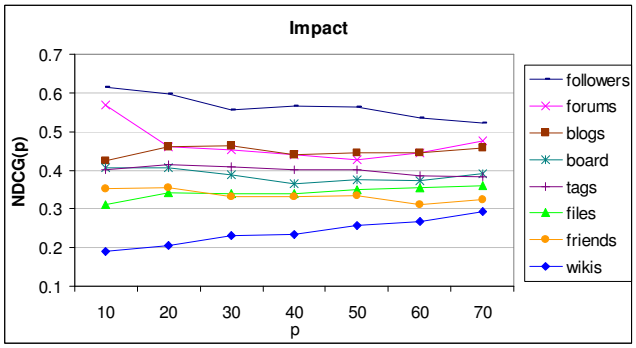

Figure 6. Comparison of the eight sources as indicators for impact

In the next section we list some limitations of our study. We then turn to discuss our method and results and present a few directions for future work.

LIMITATIONS

For assigning reputation scores to our subjects based on the survey's results we used a basic *selected/mentioned (**s/m**)* measure. The advantage of this measure lies in its simplicity and intuitiveness, however it has two main shortcomings: (1) it ignores the actual number of familiar people out of the 10 presented, and gives the same weight to selecting one or two or three subjects; and (2) it does not take into account the other individuals who appear next to a mentioned person and their own scores – e.g., one who always appear next to the CEO may never get selected despite having strong reputation. Our attempt to apply a liner programming approach did not go well since too many constraints were left unsatisfied. However, given the scale of our survey, the ***s/m*** measure gives a reliable picture of the ground truth as reflected in participants' input.

Our study was conducted inside an enterprise. As such, it may have characteristics that are specific to enterprises, as well specific to this one enterprise. However, as we used popular social media applications and rich and mature social media settings, we believe the study is suitable to serve as a first example. Generalizing may require additional future work.

Our paper encompasses a range of reputation types. Each type is typically measured with complex, multi-faceted questionnaires; while our survey dedicates a single question per type and may thus be sensitive to other factors. As our goal was to reach large participation, complex questionnaires were unsuitable. However, there is no obvious bias towards any of the indicators we examine, and we believe that the large participation we indeed achieved (554 participants, providing 10,730 responses) and the 14-19 out of 19 consensuses reached when validating the scenarios, help assure noise effects are reduced.

Finally, it should be noted that our ground truth is based on hypothetical scenarios. Although such scenarios may tell us what people think they would do in each situation, it does not necessarily shed light on their actual behavior.

DISCUSSION AND FUTURE WORK

One of the main contributions of this paper is that is refers to four different types of reputation under one umbrella. Most of these types were studied separately in the past, often being referred to simply as "reputation". Our results show that participants indeed regard different scenarios with different semantics, and that various indicators contribute differently to the inference of different reputation types. We believe this lays the ground for extensive future research, fine-tuning the terminology, identifying additional types, and studying their inference from social media.

The survey we developed is unique and deserves addressing. People are better able to formulate their preferences in a relative way, thus we opted to present a list of people to pick from, rather than directly ask for the most suitable person. Similarly, to avoid abstract terms like "reputable" or "trustworthy", we used hypothetical scenarios. The results reinforce our choice. Among the 554 who accepted our invitation and started the survey, 458 (82.67%) completed it in full, with 199 of them going through three full sets of scenarios, out of 253 who received three sets (78.66%). This suggests that the survey was clear and even enjoyable. The high familiarity level (86.54%) proves the mechanism we used was effective.

Another important contribution of our study is the distinction between people-related and content-related indicators, and the multiple types of measurements of content-related indicators (number vs. average, people vs. artifacts). Using number (of people or artifacts) seems to be closer to ground truth than using average.

In addition to the four types we examined, we included a scenario for general *reputation*, in order to understand how people interpret this concept. Results tell us that the type closest to general *reputation* is *impact*, followed by *expertise*. Trust and to a larger extent influence, the two

types most associated with reputation in the literature, are found to have lower overlap with general *reputation*.

General *reputation* and *impact* seem to be the easiest to infer – their NDCG values (showing proximity to the ground truth) were highest for most examined indicators. For both, the followers indicator is strongest. Followers does not seem to be useful for the inference of the other reputation types, though, not even for *influence*.

Expertise is best inferred through content indicators, which is expected. The blogs indicators were the strongest, and were prominent for all types but *trust*. Of all types, *trust* is the hardest to infer. Getting forums responses from many people was the best indicator for *trust*.

The friends indicator turned out to be a weak indicator for all types, even for *trust*, although one might assume that friendship implies a sense of *trust*. This is not surprising, however, as social media "friending", being a reciprocated action, does not correlate well with real-world friendship – especially in the enterprise. Wikis were the weakest indicator for *reputation* and *impact*, but were stronger for the other types, especially *expertise* and *influence*.

Inspecting the different semantics of activities within the eight applications, we learn that commenting is a better indicator than both liking (in blogs) and posting (in boards). Sentiment analysis may help fine tune the commenting indicator, by referring not only to the fact that people respond to a post, but actually inferring whether their response is positive or negative.

An interesting finding came up in the files application, where we learned that to be on the receiving end of a file sharing activity is a good indication for *impact*. This indicator is very different from the other files indicators, where liking one's files turned out to be best for indicating *trust* and downloading is best for *expertise* and *influence*

Our scenarios reflect reputation in an overall sense. In contrast, Hennis [12] introduces a method for evaluating scores per context or topic. While this is beyond the scope of our work, our findings can still be incorporated in systems that seek for contextual reputation: as social media data can be used to associate people with terms, global reputation measures can be combined with IR methods that rank people with respect to a given topic. Alternatively, for indicators with textual elements (all content indicators plus boards and tags), a topic-based reputation can be inferred by narrowing the graph to include edges that are associated with the topic. We leave these ideas for future work.

Previous studies applied link analysis algorithms, such as PageRank or HITS, to measure reputation [14,28,35,36]. In this work, we used the simpler in-degree measure as the basis for comparison across indicators. Link analysis may improve the ability to infer reputation by assigning higher weight to an incoming edge from a node whose own reputation is higher. Future studies should examine how link analysis can be effectively applied over the enterprise social graph and whether it can improve the inference capabilities for the different reputation types.

There are several other directions to pursue for further enhancement of reputation inference: while our paper focuses on reputation indicators largely from the perspective of employees. It is of interest to search for evidence of whether employers consider different indicators to be predictive of employee reputation. This would require a different definition of ground truth, focusing on employers' perspective.

We examine a wide set of indicators. As social media continues to evolve there are more opportunities to infer reputation. For example, retweet data on Twitter has been commonly used for influence analysis [3] and an analogous feature in the enterprise can be interesting to examine; applications that apply voting mechanisms, such as for ideas or for answers, can also serve as fertile ground for deriving reputation.

On the other hand, it would be interesting to study how our work is extended to the web outside the workplace. Social media applications have different usage patterns on the web, and it is interesting to understand where the differences apply. Aggregation of various reputation indicators is also a subject we leave for future work. Our results indicate that different indicators are stronger for different types and it would be interesting to examine aggregation methods for reputation indicators and to what extent they can improve reputation inference from social media.

ACKNOWLEDGEMENTS

This work has been partially funded in the 7th framework of the European scientific targeted research projects +Spaces and SocIoS which are co-financed by the European Commission. +Spaces is co-financed through theme ICT-2009.7.3 ICT for Governance and Policy Modelling under contract no. 248726 (see http://www.positivespaces.eu). SocIoS is co-financed under contract no. 257774 through theme ICT 2009.1.2: Internet of Services (for more details see http://www.sociosproject.eu).

We are grateful to the IBM employees who participated in our survey.

REFERENCES

1. Alnemr, R. & Meinel, C. From reputation models and systems to reputation ontologies. Proc. IFIPTM'11, 99-116.
2. Bian, J., Liu, Y., Zhou, D., Agichtein, E., & Zha, H. Learning to recognize reliable users and content in social media with coupled mutual reinforcement. Proc. WWW'09, 51-60.
3. Cha, M., Haddadi, H., Benevenuto, F. & Gummadi, K.P. Measuring User Influence in Twitter: The Million Follower Fallacy. Proc. ICWSM'10, 10-17

4. Chang, E., Hussain, F.K. & Dillon, T.: Reputation ontology for reputation systems. OTM Workshops. LNCS, 4278, 2006, 1724–1733.
5. Chen, B.C., Guo J., Tseng, B., & Yang, J. User reputation in a comment rating environment. Proc. KDD'11, 159-167.
6. Dencheva, S., Prause, C.R., & Prinz, W. 2011. Dynamic Self-moderation in a Corporate Wiki to Improve Participation and Contribution Quality. Proc. ECSCW'11, 24-28.
7. DiMicco, J.M., Millen, D.R., Geyer, W., Dugan, C., Brownholtz, B. & Muller, M. Motivations for social networking at work. Proc. CSCW'08, 711-720
8. DuBois, T., Golbeck, J., and Srinivasan, A. Predicting Trust and Distrust in Social Networks. Proc. IEEE 3rd SocialCom'11, 418-424
9. Guy, I., Jacovi, M., Shahar, E., Meshulam, N., Soroka, V., Farrel. S. Harvesting with SONAR – The Value of Aggregating Social Network Information. CHI'08: 1017-1026
10. Guy, I., Jacovi, M., Meshulam, N., Ronen, I., Shahr, E. Public vs. private: comparing public social network information with email. CSCW 2008: 393-402
11. Guy, I., Jacovi, M., Perer, A., Ronen, I., Uziel, E. Same places, Same Things, Same People?: mining user similarity on social media. CSCW'10: 41-50
12. Hennis, T. Monitoring contributions online: a reputation system to model expertise in online communities. Proc. UMAP'11, 422-425
13. Hasan, H. & Pfaff, C.C. The Wiki: an environment to revolutionise employees' interaction with corporate knowledge. Proc OZCHI'06, 377-380
14. Hong, L., Yang, Z., & Davison, B.D. Incorporating Participant Reputation in Community-Driven Question Answering Systems. Proc. CSE'09, 475-480.
15. Jackson, A., Yates, J. & Orlikowski, W. Corporate Blogging: Building community through persistent digital talk. Proc HICSS'07, 80
16. Jacovi, M., Guy, I., Ronen, I., Perer, A., Uziel, E., & Maslenko, M. 2011. Digital Traces of Interest: Deriving Interest Relationships from Social Media Interactions. Proc ECSCW'11, 21-40
17. Jarvelin, K. & Kakalainen, J. Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. 20, 4, 2002, 422-446.
18. Kardara, M., Papadakis, G., Papaoikonomou, T., Tserpes, K., & Varvarigou, T. Influence Patterns in Topic Communities of Social Media. Proc. WIMS'12.
19. Kiefhaber, R., Hammer, S., Savs, B., Schmitt, J., Roth, M., Kluge, F., Andre, E., & Ungerer, T. The Neighbor-Trust Metric to Measure Reputation in Organic Computing Systems. Proc. SASOW'11. 41-46.
20. Kollock, P. The production of trust in online markets. Advances in Group Processes, vol. 16, 1999.
21. Kuter, U., & Golbeck, J. SUNNY: a new algorithm for trust inference in social networks using probabilistic confidence models. Proc. AAAI'07, 1377-1382
22. Manning, C., Raghavan, P., & Schtze, H. 2008. Introduction to information retrieval. Cambridge University Press.
23. Maresch, O.M. Reputationsbasierte Trust Metriken im Kontext des Semantic Web. Master's thesis, Technische Universit at Berlin, 2005.
24. Mark, G., Guy, I., Kremer-Davidson S., Jacovi, M. Most Liked, Fewest Friends: Patterns of Enterprise Social Media Use. Proc. CSCW'2014.
25. Martín-Vicente, M.I., Gil-Solla, A., Ramos-Cabrer, M., Blanco-Fernández, Y. & López-Nores, M. Semantic inference of user's reputation and expertise to improve collaborative recommendations. Expert Syst. Appl. 39, 9, 2012, 8248-8258
26. Millen, D.R., Feinberg, J. & Kerr, B. Dogear: Social bookmarking in the enterprise. Proc CHI'06, 111-120.
27. McDonald, D.W. & Ackerman, M.S., Expertise Recommender: A Flexible Recommendation System and Architecture. Proc. CSCW'00, 231-240.
28. McNally, K., O'Mahony, M.P., Smyth, B., Coyle, M., & Briggs, P. Towards a reputation-based model of social web search. Proc. IUI'10. 179-188.
29. Nock, S.L. The Costs of Privacy: Surveillance and Reputation in America. Aldine, New York (1993)
30. Resnick, P., Kuwabara, K., Zeckhauser, R. & Friedman. E. 2000. Reputation systems. Com. ACM 43, 12, 2000, 45-48.
31. Romero, D.M., Galuba, W., Asur, S. & Huberman, B.A.Influence and Passivity in Social Media. Proc. WWW'11.
32. Steinfield, C., DiMicco, J.M., Ellison, N.B. & Lampe, C. Bowling online: social networking and social capital within the organization. Proc. C&T'09, 245-254
33. Sun, J. & Tang, J. A survey of models and algorithms for social influence analysis. Charu C. Aggarwal, editor, Social Network Data Analytics, chapter 7, 2011, 177-214.
34. Tausczik, Y.R. & Pennebaker, J.W. Predicting the perceived quality of online mathematics contributions from users' reputations. Proc. CHI'11, 1885-1888.
35. Weng, J., Lim, E., Jiang, J. & He, Q. 2010. TwitterRank: finding topic-sensitive influential twitterers. Proc. WSDM'10, 261-270
36. Zhang, J., Ackerman, M., & Adamic, L. Expertise networks in online communities: structure and algorithms. Proc. WWW'07.
37. Zhu, H., Huberman, B. & Luon, Y. 2012. To switch or not to switch: understanding social influence in online choices. Proc. CHI'12.
38. Ziegler, C.N. & Lausen G. Spreading Activation Models for Trust Propagation. Proc. IEEE International Conference on e-Technology, e-Commerce, and e-Service, 2004.