S. Bødker, N. O. Bouvin, W. Lutters ,V. Wulf and L. Ciolfi (eds.)
ECSCW 2011: Proceedings of the 12th European Conference on Computer Supported Cooperative Work, 24-28 September 2011, Aarhus, Denmark


# Digital Traces of Interest: Deriving Interest Relationships from Social Media Interactions

Michal Jacovi[*], Ido Guy[*], Inbal Ronen[*], Adam Perer[**], Erel Uziel[*], and Michael Maslenko[*]
[*]IBM Research, Haifa, Israel, [**]IBM Research, Cambridge, USA
[*]*{jacovi, ido, inbal, erelu, mmichael}@il.ibm.com*, [**]*adam.perer@us.ibm.com*

**Abstract.** Facebook and Twitter have changed the way we consume information, allowing the people we follow to become our "social filters" and determine the content of our information stream. The capability to discover the individuals a user is most interested in following has therefore become an important aspect of the struggle against information overflow. We argue that the people users are most interested in following are not necessarily those with whom they are most familiar. We compare these two types of social relationships – interest and familiarity – inside IBM. We suggest inferring interest relationships from users' public interactions on four enterprise social media applications. We study these interest relationships through an offline analysis as well as an extensive user study, in which we combine people-based and content-based evaluations. The paper reports a rich set of results, comparing various sources for implicit interest indications; distinguishing between content-related activities and status or network updates, showing that the former are of more interest; and highlighting that the interest relationships include very interesting individuals that are not among the most familiar ones, and can therefore play an important role in social stream filtering, especially for content-related activities.

## Introduction

In the era of information overflow, feed readers have emerged as a means to aggregate syndicated web content such as news headlines, blogs, or podcasts, in a single location for easy viewing (Bergamaschi et al., 2009; Aizenbud-Reshef, Guy, & Jacovi, 2009; Samper et al., 2008). Emerging applications such as Facebook and Twitter[1] allow users to get updates from the set of people they are connected to or follow. By providing streams of news based on people, these applications have essentially become "social feed readers" that allow users to stay up-to-date through their friends or the people they follow, who serve as "social filters" (Zhao & Rosson, 2009). We refer to applications that provide news streams based on lists of people chosen by the user as *social stream applications*.

As part of the Facebook social network site (SNS), the Facebook News Feed is based on the user's set of Facebook friends – a familiarity relationship. While familiarity is probably a good indication of being interested in a person, it is not an ideal source for populating the list of people from whom the user gets news. On the one hand, not all connected people are necessarily an interesting source of news: some friending invitations are accepted for mere politeness and with no intension for a close follow-up; other connections may be with people the user anyway meets frequently and does not need to follow online. On the other hand, users may be interested in individuals who are not their friends and to whom they do not feel comfortable enough to send an invitation that needs to be reciprocated.

As opposed to Facebook, Twitter and many other social stream applications (e.g[2]., FriendFeed or Google Buzz on the web; Yammer or Chatter for the enterprise) apply an asymmetric model that allows users to follow other individuals without the need for reciprocation. When applying jump-start techniques to help new users populate the list of people they follow, these applications still typically rely on symmetric familiarity-based social network information. This information typically originates from email or instant messaging applications, reflecting the people with whom the user communicates; or from SNSs that reflect reciprocated connections. These jump-start techniques usually require that users give their passwords for accessing the third-party services, which may pose privacy issues. Moreover, communication information might be considered sensitive by users.

In this paper, we propose a novel approach for identifying the people who are of potential interest to the user. We mine social network information that reflects interest from public data sources, including commenting on another person's blog, reading someone's publicly shared file, following a person's micro-blog, or tagging another individual. We argue that mining and aggregating these interest relationships can be useful in different applications and, in particular, help improve the population process of people-following lists in social stream applications.

The mining of social network information has been previously studied. These studies have focused on mining familiarity relationships (Gilbert and Karahalios, 2009; Guy et al., CHI'08; Matsuo et al. 2006) or similarity relationships (Guy et al., 2010; Xiao et al., 2001) on the web and within enterprises. Familiarity rela-

[1] www.facebook.com, www.twitter.com

[2] www.friendfeed.com, www.google.com/buzz, www.yammer.com, www.salesforce.com/chatter

tionships are based on indications that two individuals know each other, e.g., an explicit connection on an SNS, or a tight collaboration on a wiki page. Similarity relationships are based on similar behaviors and activities of people who may actually be strangers, such as using the same tags, or commenting on the same blog posts. Having addressed mining of different sources for implicit people relationships, we came to realize that some of the sources imply a third type of relationship – **interest** – reflecting curiosity or care about another individual. Interest is different from both familiarity and similarity, as it reflects a directional type of link, while familiarity and similarity are symmetric in their nature. A possible theory connecting all three relationship types can be stated as follows: when two people are similar and one of them becomes aware of this similarity, s/he may become interested in the other person. If the other is also aware and interested, the two may become familiar. However, as both similarity and familiarity do not necessarily imply interest, there is value in distinguishing interest relationships from them. To the best of our knowledge, this is the first work to suggest such a distinction in people-to-people relationships.

There are several reasons for exploring "interest relationships", both on the web and behind the firewall. Such exploration may help point out the more interesting people in a community or a division; it may be used for identifying influence or reputation, or designing diffusion algorithms over networks; and it may facilitate attention management by allowing users to focus on news coming from the people they are most interested in, when stressed for time. In this paper, we focus on the latter scenario, inside the enterprise.

We provide an extensive evaluation of the interest relationships. First, we measure their directionality by examining how many of the people users are interested in are also interested in them. We then compare the interest relationships harvested from different sources to understand the richness of available information. We also compare the interest relationships to familiarity, which is typically used for jump-starting the list of people to follow. Familiarity is used here as a baseline that has been previously studied more thoroughly (Gilbert and Karahalios, 2009; Guy et al., CHI'08). Our main evaluation is based on a user study that combines a direct evaluation of people (rating lists of people as well as individuals) with an evaluation of content produced by these people (rating news items within an enterprise social stream application). As far as we know, this is the first study to combine both types of evaluations, for this purpose. We believe it is important to apply both evaluations, to verify that the people that seem to be more interesting indeed produce more interesting content (news items, in our case).

Our results indicate that the four sources for interest relationships provide a very different list of people than the user's top familiar people. While this list is noisier, it also contains individuals that are more interesting than the most familiar ones. Providing news based on this list can be useful, especially when hybridized with the familiarity list, and most prominently for news items that refer to other pieces of content, like wikis or blogs.

The structure of this paper is as follows: we open with related work. The evaluation section presents our data sources and research method, and then reports the results of an offline analysis as well as a user study. Our discussion summarizes and raises ideas for future research. We end with a conclusions section.

# Related Work

## Micro-blogging and Twitter

Micro-blogging is one of the key examples of social stream applications, allowing users to write short messages, often referred to as "status updates", describing their activities and opinions, or pointing at interesting content. Twitter is the leading micro-blogging service with over 100 million users worldwide writing real-time updates through "tweets" of up to 140 characters. Ever since its emergence in 2006, there have been numerous studies on Twitter in particular and micro-blogging in general (Huberman, Romero, & Wu, 2009; Java et al., 2007; Kwak et al., 2010; Naaman, Boase, & Lai, 2010). Twitter has also been studied as a source for recommendations, such as of interesting URLs (Chen et al., 2010) or of news stories within RSS feeds (Phelan, McCarthy, & Smyth, 2009). Several studies have examined enterprise micro-blogging, highlighting its value for enhancing information sharing, supporting information seeking, building a common ground, and sustaining a feeling of connectedness among colleagues (Ehrlich & Shami, 2010; Zhang et al., 2010; Zhao & Rosson, 2009).

From our perspective, the most interesting feature of Twitter is the option to follow other users. Following a person is an explicit indication of interest. As following does not need to be reciprocated, the Twitter network is asymmetric, in contrast to most leading SNSs. Kwak et al. (2010) find that Twitter indeed poses a low level of reciprocity: 77.9% of the user-user relationships are non-reciprocal. Zhao & Rosson (2009) argue that Twitter serves as a "people-based RSS feed", where users are able to get trustworthy and useful information from people they know personally. They also point out that often, the followed individuals are selected because they share similar interests with the subscriber, concerning either social hobbies or their professions. Bernstein & Chi (2010) interview Twitter users and point at three factors that drive satisfaction from reading individual tweets: topic relevance, tie strength (referring to the intensity of both familiarity and interest, without explicit distinction between them), and serendipity. In this work, we offer the distinction between familiarity and interest relationships, inspect the differences between the respective networks, and leverage both to yield a better list of people to follow.

## Facebook News Feed and Social Aggregators

While the Twitter stream contains solely status updates, the stream we examine contains other enterprise social media activities, such as posting or commenting on blog entries, editing wikis, joining communities, and creating bookmarks. Due to its heterogeneity, this stream is more similar to one of the most prominent features of Facebook – the News Feed (Sanghvi, 2006) – whose introduction in 2006 marked a major change on the site (Lampe, Ellison, & Steinfield, 2008). The News Feed occupies the central part of each user's Facebook homepage, showing friends' recent activities, including, apart from status updates, such other activities as group joining, page "liking", profile changing, photo sharing, application adding, and more. As opposed to Twitter, Facebook applies a symmetric model where the only people shown in the News Feed by default are the users' friends, to whom they are reciprocally connected. Research on the Facebook News Feed is sparser than studies about Twitter, and focuses mainly on privacy issues (Boyd, 2008; Hoadley, Xu, Lee, & Rosson, 2010) and diffusion models (Sun et al., 2009).

Perhaps most similar to the stream inspected in this work are the streams created by social aggregators that consolidate friends' updates across various social media sites. FriendFeed (Gupta et al. 2009) is one of the most prominent examples of such aggregators, collating activities across many popular social media services, such as blogging systems, micro-blogging services, social bookmarking services, and many others. Similarly to Twitter, users can choose whom to follow within FriendFeed without the need for reciprocation. Current literature on FriendFeed is very sparse. Celi et al. (2010) provide a descriptive analysis of the social interactions taking place. They also perform a cluster analysis of the Italian FriendFeed network that yields a distinction between weak and highly-dedicated users. Garg et al. (2009) examine the evolution of the FriendFeed network and find that membership period, proximity within the network, and subscription to common services are all factors that affect the formation of new relationships.

## Relationship Type Distinction

In this work, we distinguish between two types of social relationships: familiarity and interest. A few studies have also made the distinction between different relationship types. The most prominent example is probably the comparison between familiarity and similarity relationships in the context of recommender systems, such as for movies (Bonhard et al., 2006) or social software items (Guy et al. 2009). A few studies have suggested enhancing regular collaborative filtering, which is based on similarity between users and their tastes, with direct familiarity relationships, such as the ones articulated in SNSs (Groh & Ehmig, 2007; Lerman, 2007; Sinha & Swearingen, 2001). Hinds el al. (2000) discuss the effects of both familiarity and similarity in the context of selecting team members, while Cosley, Ludford, & Terveen (2003) compare demographic similarity with interest-based similarity in terms of affecting interaction and cooperation while performing an online task. Hogg et al. (2008) argue that supporting multiple relationship types, such as friend, fan, or colleague, can enhance the significance of the network created within an SNS. They demonstrate this through Essembly, an online political SNS that allows users to engage in content creation, voting, and discussion. Essembly semantically distinguishes between three relationship types: friends, ideological allies, and nemeses. None of these works has distinguished between familiarity and interest relationships. In this work, we directly compare the effectiveness of the familiarity and interest networks for providing newsworthy items.

# Evaluation

## Evaluation Settings

Our research is conducted inside IBM, a large, global IT organization that acknowledges the importance of social media, both for communication with its customers and for internal collaboration and knowledge sharing (Hibbard, 2010). To extract social network information from the rich set of enterprise social media inside IBM, we use SONAR, our social aggregation platform that harvests relationships between people from over 15 organizational sources. SONAR can be configured to aggregate specific sets of relationships and create weighted lists of people related to a user based on those relationship sets. Guy et al. (CHI'08) provide a detailed description of the aggregation and weighting algorithms.

As new enterprise social applications continue to emerge, the number of data sources and relationships aggregated by SONAR has increased (Guy et al., 2010). In previous works, SONAR classified relationships into two categories – familiarity and similarity. In this paper, we identify four relationships that are likely to imply a third category – interest in a person. These relationships are: 1) following a person's tweets within an enterprise microblogging application (Ehrlich & Shami, 2010), an explicit expression of interest; 2) tagging a person within a people-tagging application (Farrell & Lau, 2006), possibly implying the wish to speedily find the person in future searches, and indicating some knowledge about the person; 3) reading someone's file[3] in a file-sharing system (Shami, Muller, & Millen, 2011), which suggests the user found that person's content of interest (reading more files of the same person indicates more interest); and 4) commenting on a person's blog post within a blogging system (Huh et al., 2007), indicating the user read the blog and felt strongly enough about the content to comment on it (more comments on a person's blog indicate a stronger interest). For each of the

[3] As we are referring to an enterprise application, most shared files are documents or presentations, hence we refer to file downloading as "file reading"

four relationships, SONAR is used for extracting the ranked list of people the user is interested in.

We observe that these four interest relationships can be classified into two categories: 1) **person-interest** – following a person's tweets and tagging a person reflect interest in a person as a self, and 2) **content-interest** – reading a file and commenting on a blog post reflect interest in content created by the person. Apart from investigating each of the relationships separately, we define an aggregated interest list (*I*), retrieved by combining the four relationships above.

We use SONAR to harvest familiar people into a familiarity list (*F*), based on 24 different relationships that indicate familiarity, such as being explicitly connected on a social network site, being connected via the organizational chart, co-editing a wiki page, co-authorship of patents and papers, and other relationships (Guy et al., 2009). Previous research has indicated that aggregating familiarity relationships this way effectively produces a list of people the user knows well.

In our analysis, we examine lists *I* and *F* above, as well as a hybrid list (*I*+*F*). This list combines people from *I* and *F* by ranking the people that appear in both lists according to the sum of their ranks in the individual lists, and then alternating between the lists for those that appear only in one of them, according to their rank. As a result, this list gives priority to people who appear in both *I* and *F* over those who appear in just one of them.

For the last phase of our study, we use an enterprise social stream aggregator that displays an activity stream of recent public news items that took place across the organization's social media applications. The news items can originate from various sources, including: (1) profiles (status updates, additions to social network, people tagging); (2) wikis (creating and editing a public wiki); (3) blogs (creating, editing, or commenting on a blog); (4) files (creating, editing, commenting, or recommending a public file); and more. Our evaluation examines the recommendation of news items to the user, originating from different lists of people that the user may be interested in.

## Research Method

The initial, offline part of our study aims to quantitatively examine the lists of people originating from the different interest relationships. We start by focusing on each separate source of interest information and studying its directionality: i.e., the match between the list of people a person is interested in, and the list of people who are interested in that person. Next, we compare the lists returned from the four interest sources to each other to examine their diversity. Finally, we compare the lists returned by the four sources to the list of familiar people, which has been extensively studied in previous papers. As social stream applications typically rely on familiarity relationships for providing news, we regard that list as a relevant baseline and seek to validate the hypothesis that the lists produced by interest relationships are indeed different from the familiarity list.

The greater part of our study is based on a personalized online survey sent as a link via email to selected participants. The survey consists of two phases: in the first phase, participants rate their direct interest in other people. In the second phase, they rate their interest in news items produced by those people. This unique combination allows us to evaluate both the direct interest in people as well as the interest in the news items they produce. We also examine the correlation between the two phases, testing whether people chosen as interesting indeed produce more interesting items.

The first phase, for rating people directly, includes two sub-phases of its own: in the first sub-phase (1a), participants are asked to rate their interest in different people **lists**, generated according to the relationships and aggregates, as described above. Each list includes 10 people. Since comparing and rating lists of people may be a complex task, we introduce a second sub-phase (1b), in which participants select individual people whom they find most interesting. We conjecture that the combination of the people list rating with individual people selection would allow us to receive a good overall picture of how interesting the people in each list are. We next describe all phases in detail.

As our goal is to identify and compare interest relationships from different sources, we focus on users who make use of at least two of the four single interest sources described above and who have at least 10 people in both their *I* list and *F* list. This ensures that participants have enough data to generate comparable lists of people and items for both phases of the survey. We identified and invited 470 such users to participate.

In phase 1a, participants are presented with up to seven lists of 10 people each. The lists are generated according to the four interest relationships described above: 1) micro-blogging following, 2) blog commenting, 3) file reading, and 4) people tagging, as well as the aggregates 5) *I,* 6) *F*, and 7) *I+F*. Each list includes the top 10 people and is not labeled according to its relationships to avoid bias. Lists that include less than 10 people are not shown (these are only lists based on a single relationship, as the selection of participants ensures at least 10 people in the aggregate lists). Thus, each participant rates between three and seven lists. In practice, the average number of lists rated by a participant was 4.89 (stdev: 1.02, median: 5).

Participants are asked to rate each list according to how much the people in the list represent a set of people from whom they would like to get news items. Rating is based on a 5-point Likert scale, ranging from “Does not represent a list of people I am interested in” to “Very much represents a list of people I am interested in”. Additionally, participants are asked to indicate the best list out of those presented. Figure 1 shows a screenshot of this part of the survey.

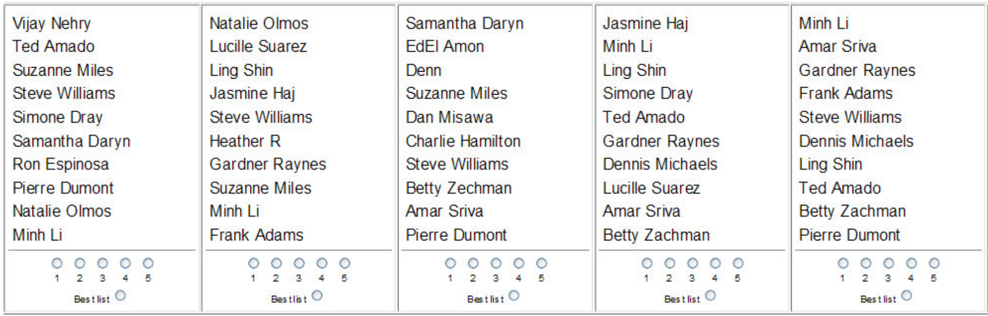


**Figure 1. Online survey, phase 1a**

In phase 1b, participants are presented with a combined list of all the people that appeared in the previous lists and are asked to select exactly five people from whom they would most like to get news items.

In phase 2, participants are asked to rate a set of news items which are generated in the following way: For each of the three aggregate lists (*I*, *F*, and *I*+*F*), we extract from the social stream aggregator the 25 latest news items that relate to at least one person on that list. We then randomly choose eight items out of the 25 items and mix all chosen items while removing duplicates, resulting in a list of at most 8x3=24 items. By selecting eight items at random out of 25, rather than simply the most recent eight, we aim to increase the diversity of news items, both over time and with regards to the corresponding people and sources.

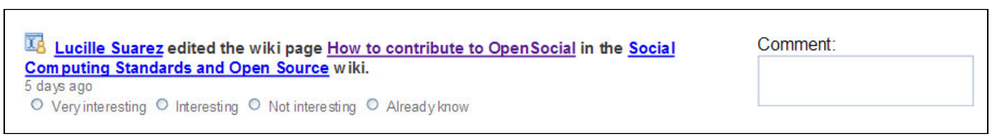


**Figure 2. Sample news item in phase 2 of the survey**

Figure 2 shows a sample news item. Each item includes an icon indicating the source it originates from, a descriptive text (e.g. “P1 commented on the blog post X” or “P2 tagged P3 with ‘hci’”, where Pi are people names, linking to their profile pages, and X is a blog post title, linking to its page). Below each news item is an indication of the time it was posted, e.g., “2 hours ago”. Participants are asked to evaluate each item as “Very Interesting”, “Interesting”, “Not Interesting”, or “Already Know”. They can also leave a comment next to each item. At the end of the survey there is another opportunity to leave a general comment.

## Results – Offline Analysis

### Asymmetry of Interest Relationships

The first analysis compared the directionality of the four interest relationships, examining the match between the list of people a person is interested in, and the list of people who are interested in that person using the Match@10[4] measure. The results are presented in Table I.

**Table I: Examination of the directionality of the sources**

| | Micro-blogging | People tagging | Blog commenting | File reading |
|---|---|---|---|---|
| **Match@10** | 4.75 | 0.42 | 1.09 | 0.51 |

For the micro-blogging system, 118 users with at least 10 people in both their ‘following’ and ‘followed’ lists were identified. The average Match@10 for these users is 4.75, indicating that almost half of the relationships are reciprocated.

The other relationships show far lower reciprocation. Blog commenting (with 282 users having at least 10 people in both their ‘comment to’ and ‘comment by’ lists) has about one matching person (1.09). File sharing (with 335 users) and people tagging (138 users) have even lower matches, around 0.5.

The reciprocation of these relationships is not inherent in the system, although it is sometimes encouraged. When people comment on a blog post, a link to their blog is automatically attached to their comment, encouraging the blog owner to visit and potentially comment. Following people on Twitter is visible to all, and an email message actually notifies the user about new followers, explicitly encouraging reciprocation, even if just out of courtesy. People-tagging and file-sharing require a higher level of involvement – you actually need to know something about a person to tag them; similarly, you actually need to have an interesting file to share so the other person can read it. These requirements may explain why these sources are the least reciprocated.

Overall, Table I shows that the four interest relationships pose a great deal of directionality – people you express an interest in do not necessarily express an interest in you, and the other way around. This shows us that these relationships may be different from the symmetric relationships that have been previously examined, raising a motivation to study them further.

## Comparing the Interest Relationships

Next, we compared the interest direction (people the user is interested in) in each of the four relationships to one another, in an attempt to understand the resemblance between them and the richness of information that can be harvested from them. For each pair of relationships, we identified the users who have at least 10 people in both relationships, and calculated the average Match@10. The results are shown in the bottom three rows of Table II.

The results show that the four relationships are very different, having less than one match, on average, out of the top 10 people they return. This tells us that har-

[4] Match@k – considers the percentage of overlapping people between the top k items in two lists (we used k=10). The match@k measure captures the similarity between the lists. It reflects how well one list can approximate the other (Guy et al. CSCW’08).

vesting the four relationships would yield different information, and that aggregating them would create a richer set of interesting people. In our user study, we examined the aggregates of the relationships as well as each one on its own.

**Table II: Match@10 between interest relationships**

| | Tweet following | Tagging | Blog commenting | File reading |
|---|---|---|---|---|
| **Familiarity** | 1.10 | 1.65 | 0.57 | 0.65 |
| **Tweet following** | | 0.12 | 0.44 | 0.71 |
| **Tagging** | | | 0.30 | 0.40 |
| **Blog commenting** | | | | 0.26 |

### Comparing Interest Relationships to Familiarity

As familiarity is commonly used as a seed list for following people, we refer to it as a baseline and compared the interest direction of all four relationships with familiarity. The results are shown in the first row of Table II above.

Tagging people has the highest average match with familiarity (1.65), with tweet following the next highest match (1.10). File reading and blog commenting have lower match rates (0.65 and 0.57, respectively). These results reflect our suggested classification of interest relationships: people-interest relationships have higher overlap with familiarity than content-interest relationships, as one may have expected.

Comparing the familiarity list (*F*) to the aggregation of the four interest relationships results (*I*) in a Match@10 of 1.49. All in all, the match between the interest relationships and familiarity is quite low; indicating that harvesting the interest relationships may enrich the set of people to follow.

# Results – Online Survey

Exactly 200 people agreed to be participants and completed phase 1 of the survey; 192 completed both phases. These 192 originated from 23 countries, spanning the organization's different divisions: 35% were from the Sales Division, 27% were from Software, 16% from Headquarters, 15% from Services, 2% from Systems, 2% from Research, and 3% from others.

## Comparing Lists of Interesting People

In Phase 1a of the online survey, participants were asked to rate up to seven lists of 10 people and select the best list. Table III shows a summary of the results. The first row shows the number of times each type of list was presented. The three aggregates were presented to all 200 users; the lists originating from individual relationships were presented when relevant (i.e., when containing at least 10 people). The second row shows the percentage of times a list was selected as the best list, relative to the number of times it was presented. List *I,* an aggregation of all interest sources, has a statistically significant lower percentage of best votes than list F (7.78% vs. 32.34% with $p$=1.25E-07 in a one-tailed paired t-test). The hybrid list (*I+F*) is perceived best (most interesting) even more times than the *F* list. Among the individual sources (listed on the right side of the table), the list based on file reading has the highest rate of best votes (13.89%), followed by people tagging (11.11%). Interestingly, these are the two lists for which more participants had data (108 and 135, respectively) – apparently they are indeed more interesting and attract more users.

**Table III: Selecting interesting lists of people**

| | I | F | I+F | micro-blogging | People Tagging | Blogs | Files |
|---|---|---|---|---|---|---|---|
| **# appearances** | 200 | 200 | 200 | 60 | 135 | 74 | 108 |
| **Best list selection %** | 7.78% | **32.34%** | **37.72%** | 6.67% | 11.11% | 4.05% | 13.89% |
| **Average score** | 3.14 | **3.57** | **3.81** | 3.18 | 2.98 | 2.88 | 3.10 |
| **% rated 1** | 6.50% | 7.00% | 2.50% | 8.33% | 10.37% | 9.46% | 9.26% |
| **% rated 2** | 23.50% | 14.00% | 12.00% | 15.00% | 24.44% | **31.08%** | 24.07% |
| **% rated 3** | **34.00%** | 22.00% | 17.50% | **36.67%** | **32.59%** | 28.38% | 25.00% |
| **% rated 4** | 22.00% | **29.00%** | **38.50%** | 30.00% | 22.22% | 24.32% | **30.56%** |
| **% rated 5** | 14.00% | 28.00% | **29.50%** | 10.00% | 10.37% | 6.76% | 11.11% |

The bottom part of Table III refers to the 1-5 ratings. As in the case of the best votes, the *I* list yields the lowest average rating of the three aggregates, while the average score of *I+F* is higher than that of *F*. The distribution of scores can be seen in the last five rows of the table, where we see that *I+F* not only has the best average, but in fact obtains the highest number of 4 and 5 ratings (most interesting). This is encouraging and hints that there is value in harvesting the interest relationships for composing lists for people following applications. The bottom of the table visualizes the comparison between the sources, and especially the individual sources, showing that reading files and following micro-blogs typically create more interesting lists, whereas blogs seem to yield less interesting lists.

## Selecting Interesting Individuals

Phase 1b of the study presented the participants with the same set of people they saw in phase 1a, but this time in a single list out of which they were asked to select five individual people who are of most interest.

**Table IV: Selecting interesting individuals**

| | I | F | I+F |
|---|---|---|---|
| **# of appearances** | 2000 | 2000 | 2000 |
| **# selected from group** | 424 | 329 | 565 |
| **relative percent** | 21.20% | 16.45% | **28.25%** |

The results, depicted in Table IV, are quite different from phase 1a. The *I* list yields more interesting individuals than the *F* list, despite being rated lower as a whole list in phase 1a (difference is statistically significant, $p$=1.88E-05 in a one-tailed paired t-test). The aggregated list *I+F* yields even more selected individuals. This tells us that while the *F* list, as a whole, is more interesting on average, the very interesting individuals are actually in the *I* list. This may be explained by the fact that in addition to the most interesting people, *I* also contains less interesting people, some of them unfamiliar to the user, who are considered "noise" when examined as part of a whole list.

An interesting discussion is raised here, about the tension between receiving the most interesting news along with noise, versus missing out on the most interesting news, but getting less noise. One participant wrote a related comment: "*An interesting dilemma is that there are those who are of interest for my day job, whilst others are inspirational or of interest for skill expansion!*" Another wrote: *"I think I would always want direct reports to be included in my feed. [However] if you removed them, and then presented the list, things may be more interesting*." It seems that hybridization combines the benefits of both *I* and *F*, being rated highest as a whole list as well as containing the most top-five individuals.

## Comparing Interesting Items

In phase 2 of the study, users were presented with actual news items, associated with people in the *I*, *F*, and *I+F* lists. 192 people completed this part of the survey. As a whole, they were presented with 3629 news items, and were asked to rate them as "very interesting", "interesting", "not interesting", or "already know". In our analysis, we merge the responses of "very interesting" and "interesting" under "all interesting". Figure 3 summarizes the distribution of item ratings coming from the three lists.

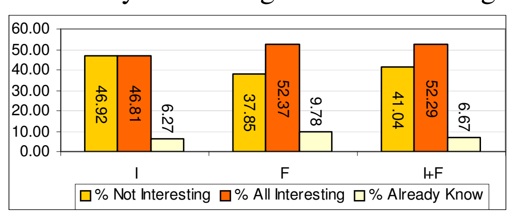


**Figure 3. Interest rate comparison of all 3629 items**

The *F* list has an advantage over the *I* list in this case: it has more items rated interesting and less of its items are considered "noise" (not interesting). These findings are statistically significant ($p$=9.13E-06 in a one-tailed unpaired t-test). The percentage of items rated interesting in *I+F* is very similar to *F*. *F* has the higher percentage of already-known items, but also the lowest percentage of non-interesting items. The *I* list evidently helps lower the expectedness of news, while at the same time increasing the noise. One participant commented: "*Some of these people I work with on a daily basis at the same location, so although I am very interested in their activities, my need to follow them on social software is minimal.*"

In order to examine the correlation between participants' ratings in phases 1 and 2, we examined whether the people selected among the top five most interesting individuals in phase 1b indeed yielded more interesting news items as rated in phase 2. Table V shows this comparison over all 192 participants who completed both phases. Individuals selected among the top five in phase 1 yielded 57.8% all interesting items vs. 35.4% non-interesting ones, while individuals who were not selected among the top five, yielded items rated 47.2% all interesting vs. 46.5% non-interesting. These differences are statistically significant ($p$=1.7E-10 in a one-tailed unpaired t-test) and indicate that participants' selection of top-five individuals indeed reflects higher likelihood that the news items they produce are interesting. While the *I* list contains more top-five individuals who yield more interesting news items, the overall rating of its news items is slightly lower than those of the *F* list, which has fewer top-five individuals. This reinforces our suspicion that alongside the highly interesting individuals, the *I* list contains more non-interesting individuals who ultimately overshadow the high rating of news items yielded by the top-five individuals.

**Table V. Interest rate in phase 2 items by people selection in phase 1b**

| | All Interest | Not Interest | Known |
|---|---|---|---|
| **top-5 ppl** | 57.8% | 35.4% | 6.8% |
| **others** | 47.2% | 46.5% | 6.3% |

## Studying the various sources

The analysis discussed above examined an aggregation of all sources of news. However, it turns out that items originating from different types of sources are very different in the interest level they yield. We next examine several different sources of news items to better understand these differences.

The source that provides the most interesting news is the file-sharing application. 311 items originated from it, notifying of file creations and edits. Figure 4 shows that in this case, the *I* list has a slightly higher percentage of interesting items than *F*. The *I+F* list outperforms both the *I* and *F* lists, with over 80% of interesting items. It even has less already-known items than the *I* list. The *F* list has the highest percentage of already-known items – 9%.

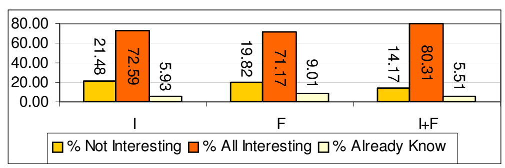


**Figure 4. Comparison of 311 file news items**

Wikis turn out to be another source for relatively interesting items. 282 items originated from wikis, notifying of wiki creation and editing. Their rates are shown in Figure 5. Here, *I* outperforms *F* by an even more noticeable gap, while *I+F* is slightly better than *I*. The *I* list has substantially less already-known items than the *F* and *I+F* lists.

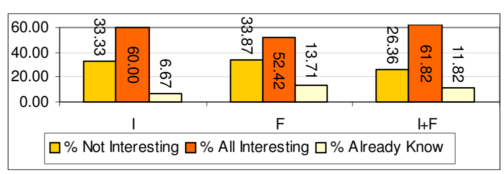


**Figure 5. Comparison of 282 wiki news items**

Blogs are another example of a source that produces mostly interesting items – 57% out of 270 blog updates were rated as interesting or very interesting. The figures of the *I*, *F*, and *I+F* lists are very similar for blog updates with a slight advantage of *I* over *F* (58.82% vs. 57.14% items rated interesting; 57.52% for *I+F*.)

Next, we examine the 518 status updates (Figure 6). The general interest rate for these items is around 54%, which is lower than files, wikis, and blogs. *F* outperforms *I* here, both in the percentage of interesting items, and in the noise level. *I+F* yields similar ratings to *F*, but does not outperform it.

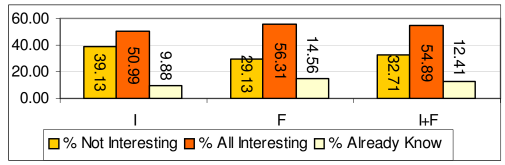


**Figure 6. Comparison of 518 status updates**

Network addition news items and updates about related people adding or being added as friends, are examples of items of less general interest (Figure 7). One participant commented: "*NEVER NEVER NEVER show me someone else's network additions! This is useless.*" Indeed, the interest rates for these types of news items are the lowest. *F* again performs better than *I* here, with both more interesting items and less noise. *I+F* performs quite similarly to *F*.

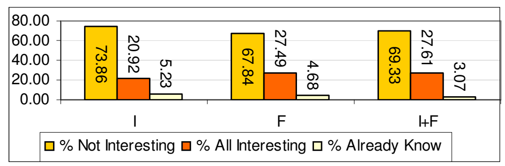


**Figure 7. Comparison of 401 network additions**

Another source of less general interest is people tagging. As can be seen in Figure 8, while the *F* list contains an even rate of interesting vs. not-interesting items, the *I* list contains mostly non-interesting items.

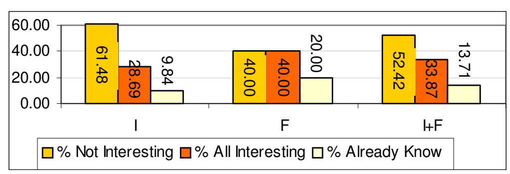


**Figure 8. Comparison of 232 people tagging items**

It seems that item types of less general interest have a better interest rate when coming from familiar people (the *F* list). This may be explained by the fact that even if the item's content is not interesting, one still gains something when learning about familiar people. All in all, it seems that the variance of interest among the sources feeding the news feed is large and deserves dedicated research.

Inspecting these results, we observe that the news sources can be categorized into three categories: (1) **content-related activities**, such as activities related to files, wikis, and blogs; (2) **micro-blogging** messages, such as status updates; and (3) **network activity**, such as people tagging and friend addition. The general interest in each category is quite different. News concerning content generated the most interest: from about 57% for blogs to nearly 75% for files. Status updates generate slightly less interest, with about 54% of the news items rated interesting. Network activities are far less interesting, with around 34% general interest in people tagging and only 25% interest in network additions.

Even more interestingly, for the content-related sources, the *I* list produces slightly better results than the *F* list (as in files, wikis, and blogs). For status updates, *F* outperforms *I*, while for network activities the gap slightly grows. Thus, the *I* list is especially productive for the most interesting news types, the content-related activities. This is a substantial outcome when considering which news items should ultimately be on the user's news feed.

Consistently across almost all news item types, *F* yields more already-known items, while *I* yields more noise. In many of the cases, the hybridization combines both advantages – it has less known items than *F* and less noise than *I,* often also leading to a slightly higher percentage of interesting items. Several participants commented that they wished to get hybrid news from both close colleagues and people outside their organization circle. One wrote "*I like a mix of people, with a few from different organizations and geographies … I specifically don't want only people in my own organization*," while another stated "*Most interested either up my management chain or outside my organization, for strategic networking*".

## Discussion, Limitations, and Future Work

The previous section provides diverse analysis of the interest relationships. First, we show that interest relationships are indeed often asymmetric – the people you are interested in are not necessarily interested in you. This is reflected in a very low level of symmetry across all interest relationships, apart from following people on micro-blogging, for which reciprocity is explicitly encouraged. In this work, we focus on one direction of the interest relationships – the people a user is interested in. Comparing the lists of people that are returned from four public sources reveals that the overlap is low, as each brings a rather different list of people. As a result, aggregating them produces a richer picture of one's "network of interest". The overlap with familiarity is generally low, indicating that the people you are most interested in are not necessarily the ones you are most familiar with. This overlap also reflects the distinction suggested for the interest relationships: person-interest relationships (following and tagging) have a higher overlap with familiarity than content-interest sources (file reading and blog commenting).

Our online survey focuses on comparing the interest list with the familiarity list and a hybridization of the two. The people-based evaluation (phase 1) indicates that the interest list consists of a few very interesting people (who are often selected among the top five most interesting) but also some "noise" – people who are not interesting at all (and may not even be familiar to the user). Hence, when evaluating lists of people, most users would prefer the more "solid" familiarity list; however when picking the top interesting individuals, more would come from the interest list. The hybrid *I+F* list is found to be the best performing list in all aspects of this phase: it has the highest average rating, marked best most times, and its individuals are most often among the top five. These results indicate that mining interest relationships and combining them with familiarity can enhance automatic inference of the people who are of the most interest to the user.

The news item evaluation (phase 2) shows that the *F* and *I+F* lists produce very similar percentage of interesting items (slightly over 52%). *I+F* produces less already-known items, while *F* produces less noise (non-interesting items). The *I* list produces the least interesting items (about 47%) and has the most noise (also about 47%). These results point at the tension between getting the most interesting items that span beyond your close network, but with some noise; and getting a more solid list of interesting yet more expected items, and missing out on the most interesting ones. Users of social stream applications should be allowed to choose between these two options. Hybridization offers a way to mitigate the noise while maintaining the non-expectedness level. In this work, we examined one specific hybridization method. Further techniques for hybridization and their potential to improve the results should be examined in future research.

Breaking down the results by item type reveals that interest diversity across types is high: while content-related activities are very interesting (e.g., files yield 75% interesting items), network-related activities are mostly not (e.g., additions to network yield only 25% interesting items). Status updates are in-between with slightly over 50% of interest. In terms of the comparison between *I* and *F*, the *I* list yields better results for the content-related activities (files, wikis, blogs), while *F* substantially outperforms *I* for status updates and network-related activities.

Some of the comments we received help understand these differences. One participant wrote "*Someone else's network additions is of no interest to me … unless it is someone I know well, and even then mainly for gossip,*" and another commented "*I'm only interested in updates on tagging if it's one of my close colleagues who has been tagged or used the tag.*" On the other hand, items that relate to other content are interesting beyond a user's close social circle and in general are considered more interesting news. It can also be that for content-related activities, people have other channels (like face-to-face meetings or email) to get the updates from their close colleagues; hence the results for the *F* list are lower than for the other lists for these item types. In any case, these findings indicate that the *I* list has the most substantial effect on the category of most interesting news.

Following a person on an enterprise micro-blogging system is the most explicit expression of interest, analogous to a connection on an SNS for familiarity relationships. One might assume that having micro-blogging as one of the sources in the *I* aggregation has the strongest influence on its performance. However, inspecting the results for the 108 participants who do not use the micro-blogging application and do not have it as a source, reveals that ratings of items from their *I* list are not lower than for the entire population. This tells us that a good interest list can be composed even if there is no explicit following information.

Our survey respondents are not a representative sample of the organization's employees, but rather avid users of our enterprise social media, for whom interest lists can be produced. While this is not an optimal choice, trying to identify a representative sample would result in too little data for most participants, as social media is still not prevalent enough. However, we believe that the potential population who can benefit from automatic mining of their interest network will grow in the years to come, as social media becomes more popular in organizations, and as web users get used to exposing more information in public. File reading within an enterprise file-sharing application (Shami, Muller, & Millen, 2011) is an example of a new public source of information, which exposes very valuable data about people's interests. In most of our results, file reading is indicated to be the most effective interest indicator, even if by a small margin, over the other three.

We compare an initial set of aggregated interest relationships (four overall) with a well-established aggregation of familiarity relationships (24 overall). While this combination of familiarity relationships has been shown effective in producing a list of people the user knows, we show that for producing a list of people the user is interested in, the four interest sources are important. Combining both lists can be beneficial in yielding a final list that is diverse and contains very interesting people outside the close workgroup of the user. As enterprise social media becomes more popular, new sources for mining interest may become available. Aggregating a richer set of interest relationships can reduce noise and ultimately make the interest network even more representative.

Claypool, et al. (2001) discuss *implicit interest indicators* in items, such as movies or web pages. Such indicators include clicking, viewing, or searching for the item. In this work, we essentially propose *implicit interest indicators* in **people**. These indicators are based on public data, such as commenting on a blog, tagging, or following. Our future plans include the investigation of private implicit interest indicators, such as viewing a person's profile or searching for the person. Such private indicators, however, involve sensitive data that might raise privacy issues.

Other future directions include exploring harvesting of interest relationships outside the enterprise, where the variety of social media applications and thus potential sources for mining interest relationships is larger. The interest relationships we examined in this work have counterparts on the web (following, commenting, reading, tagging), while the potential richness of sources further grows outside the firewall (e.g., "liking" another person's content as on Facebook). Finally, we plan to inspect more scenarios for leveraging the interest relationships. For example, collaborative filtering, the most popular technique for recommender systems, recommends items, such as movies or books, based on the preferences of similar people. Executing collaborative filtering based on interest relationships can potentially further boost its performance.

# Conclusions

Social stream applications rely mainly on familiarity relationships to filter news items or jump-start the list of people from whom users get news. In this work, we suggest mining a new type of social relationships – interest. Interest relationships reflect directional curiosity or care about another individual and are more asymmetric in nature than previously studied familiarity or similarity relationships.

Our evaluation examines four sources for mining interest relationships inside a large global enterprise, showing each source to yield a rather different set of individuals. The aggregated interest network is found to be very different from the familiarity network. In spite of being based on solely four relationships, the interest network is found to include very interesting people beyond the user's closest workgroup (in parallel with some noise). Hybridizing this network with the familiarity network can be highly valuable in producing interesting and diverse news items for users of social stream applications.

# References


1. Aizenbud-Reshef, N., Guy, I., and Jacovi, M. 2009. Collaborative feed reading in a community. *Proc. GROUP '09*, 277-280.
2. Bergamaschi, S., Guerra, F., Orsini, M., Sartori, C., and Vincini, M. 2009. Relevant News: a semantic news feed aggregator. *Proc. Italian Semantic Web Workshop 2007*.
3. Bernstein, M., Kairam, S., Suh, B., Hong, L., and Chi, E.H. 2010. A torrent of tweets: managing information overload in online social streams. *Workshop on Microblogging, CHI '10*.
4. Bonhard, P., Harries, C., McCarthy, J., & Sasse, M. A. 2006. Accounting for taste: using profile similarity to improve recommender systems. *Proc. CHI' 06*, 1057-1066.
5. Boyd, D. 2008. Facebook's privacy trainwreck: exposure, invasion, and social convergence. *Convergence* 14 (1).
6. Celi, F., Di Lascio, F.M.L., Magnani, M., Pacelli, B., and Rossi, L. 2010. Social network data and practices: the case of Friendfeed**.** *Advances in Social computing*, 346-353.
7. Chen, J., Nairn, R., Nelson, L., Bernstein, M., and Chi, E.H. 2010. Short and tweet: experiments on recommending content from information streams. *Proc. CHI '10*, 1185-1194.
8. Claypool, M., Le, P., Wased, M., & Brown, D. 2001. Implicit Interest Indicators. Proc. IUI '01, 33–40.
9. Cosley, D., Ludford, P., and Terveen, L. 2003. Studying the effect of similarity in online task-focused interactions. *Proc. Group '03*, 321-329.
10. Ehrlich, K. and Shami, N.S. 2010. Microblogging inside and outside the workplace. *Proc. ICWSM '10*.
11. Farrell, S., & Lau T. 2006. Fringe Contacts: People Tagging for the Enterprise. *Workshop on Collaborative Web Tagging*, WWW' 06.
12. Garg, S., Gupta, T., Carlsson, N., and Mahanti, A. 2009. Evolution of an online social aggregation network: an empirical study. *Proc. IMC '09*, 315-321.\
13. Gilbert, E. and Karahalios, K. 2009. Predicting tie strength with social media. *Proc. CHI '09*, 211-220.
14. Groh, G., and Ehmig, C. 2007. Recommendations in Taste Related Domains: Collaborative Filtering vs. Social Filtering. *Proc. GROUP'07*, 127-136.
15. Gupta, T., Garg, S., Carlsson, N., Mahanti, A., and Arlitt, M. 2009. Characterization of Friendfeed: A web based social aggregation service. *Proc. ICWSM '09*.
16. Guy, I., Jacovi, M., Meshulam, N., Ronen, I., Shahar, E. Public vs. private: comparing public social network information with email. *Proc. CSCW '08,* 393-402.

17. Guy, I., Jacovi, M., Perer, A., Ronen, I., and Uziel, E. 2010. Same Places, Same Things, Same People? Mining User Similarity on Social Media. *Proc. CSCW '10*, 41-50.
18. Guy, I., Jacovi, M., Shahar, E., Meshulam, N., Soroka, V., & Farrell, S. 2008. Harvesting with SONAR: the value of aggregating social network information. *Proc. CHI'08*, 1017-1026.
19. Guy, I., Zwerdling, N., Carmel, D., Ronen, I., Uziel, E., Yogev, S., and Ofek-Koifman S. 2009. Personalized Recommendation of Social Software Items based on Social Relationships. *Proc. RecSys '09*, 53-60.
20. Hibbard, C. 2010. How IBM Uses Social Media to Spur Employee Innovation. Social Media Examiner online Magazine (Feb. 2010). http://www.socialmediaexaminer.com/how-ibm-uses-social-media-to-spur-employee-innovation/
21. Hinds, P. J., Carley, K. M., Krackhardt, D., & Wholey, D. 2000. Choosing work group members: Balancing similarity, competence, and familiarity. *OBHDP 81 (2)*, 226–251.
22. Hoadley, C.M., Xu, H., Lee, J.J, Rosson, M.B. 2010. Privacy as information access and illusory control: The case of the Facebook News Feed privacy outcry. *Electronic Commerce Research and Applications 9, 1 (Jan. 2010)*, 50-60.
23. Hogg, T., Wilkinson, D., Szabo, G, and Brzozowski M. J. 2008. Multiple relationship types in online communities and social networks. *Proc. AAAI Symposium on Social Information Processing*.
24. Huberman, B., Romero, D., and Wu, F. 2009. Social networks that matter: Twitter under the microscope. *First Monday* 14, 1 (Jan. 2009).
25. Huh, J., Jones, L., Erickson, T., Kellogg, W. A., Bel-lamy, R. K., and Thomas, J. C. 2007. BlogCentral: the role of internal blogs at work. *Proc. CHI '07*, 2447-2452.
26. Java, A., Song, X., Finin, T., and Tseng, B. 2007. Why we twitter: understanding microblogging usage and communities. *Proc. WebKDD '07*, 56-65.
27. Kwak, H., Lee, C., Park, H., and Moon, S. 2010. What is Twitter, a social network or a news media? *Proc. WWW '10*, 591-600.
28. Lampe, C., Ellison, N. B., and Steinfield, C. 2008. Changes in use and perception of Facebook. *Proc. CSCW '08*, 721-730.
29. Lerman, K. 2007. Social networks and social information filtering on Digg. *Proc. ICWSM'07*.
30. Matsuo, Y., Hamasaki, M. et al. Spinning multiple social networks for semantic Web. *Proc. AAAI '06* (2006).
31. Naaman, M., Boase, J., and Lai, C. 2010. Is it really about me?: message content in social awareness streams. *Proc. CSCW '10*, 189-192.
32. Phelan, O., McCarthy, K., and Smyth, B. 2009. Using twitter to recommend real-time topical news. *Proc. RecSys '09*, 385-388.
33. Samper, J. J., Castillo, P. A., Araujo, L., Merelo, J. J., Cordón, í., and Tricas, F. 2008. NectaRSS, an intelligent RSS feed reader. *J. Netw. Comput. Appl.* 31, 4 (Nov. 2008), 793-806.
34. Sanghvi, R. (2006, September 5). Facebook Gets a Facelift. Retrieved January 17, 2010, from http://blog.facebook.com/blog.php?post=2207967130
35. Shami, S.N., Muller, M.J., and Millen, D.R. 2011. Browse and discover.: social file sharing in the enterprise. *Proc. CSCW '11*.
36. Sun, E., Rosenn, I., Marlow, C., Lento, T. 2009. Gesundheit! Modeling contagion through Facebook News Feed. *Proc. ICWSM '09*.
37. Xiao, J., Zhang, Y., Jia, X., & Li, T. 2001. Measuring similarity of interests for clustering web-users. *Proc. ADC'01*, 107-114.
38. Zhang, J., Qu, Y., Cody, J., and Wu, Y. 2010. A case study of micro-blogging in the enterprise: use, value, and related issues. *Proc. CHI '10*, 123-132.
39. Zhao, D. and Rosson, M. 2009. How and why people Twitter: the role that micro-blogging plays in informal communication at work. *Proc. GROUP '09*, 243-252