Same Places, Same Things, Same People? Mining User Similarity on Social Media

Ido Guy, Michal Jacovi, Adam Perer, Inbal Ronen, Erel Uziel
IBM Haifa Research Lab
Mt. Carmel, Haifa 31905, Israel
{ido, jacovi, adamp, inbal, erelu}@il.ibm.com

ABSTRACT
In this work we examine nine different sources for user similarity as reflected by activity in social media applications. We suggest a classification of these sources into three categories: *people*, *things*, and *places*. Lists of similar people returned by the nine sources are found to be highly different from each other as well as from the list of people the user is familiar with, suggesting that aggregation of sources may be valuable. Evaluation of the sources and their aggregates points at their usefulness across different scenarios, such as information discovery and expertise location, and also highlights sources and aggregates that are particularly valuable for inferring user similarity.

Author Keywords
Social networks, user similarity, social media, social software.

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

General Terms
Design, Experimentation, Human Factors, Measurement

INTRODUCTION

Millions of people use social media applications as a part of their daily online activities. Typical users may diligently comment on their peer's blogs, tag their online photos, friend their colleagues on social network sites (SNSs), and annotate their digital bookmarks. This plethora of users and content, and the diversity of tasks performed on social media applications, enable people to expand and refine their interests with people who are as passionate as they are about the topics they care about. While social media provides a new medium for inferring user relations through the rich and diverse user activity, the variety of applications and content types make mining a challenging task.

CSCW 2010, February 6–10, 2010, Savannah, Georgia, USA.

In general, relationships connecting people on social media sites come in at least two flavors: *familiarity* evidence and *similarity* evidence. Familiarity evidence items provide clues to when users may know one another, such as an explicit connection on an SNS, tight collaboration on a wiki page, or a public message exchange. Recent studies have focused on this type of social network information, exploiting familiarity relationships for scenarios such as buddylist weighting [16], people recommendation [6,17], and tie strength prediction [13]. However, there are other evidence items reflected in social media applications that provide clues to *similarity* rather than *familiarity*. These highlight similar behaviors and activities of people who may actually be strangers. Examples of similarity evidence include using the same tags, bookmarking the same web pages, or connecting with the same people.

Harvesting similarity information may be useful in various scenarios. It may be used for information discovery, by making users aware of people who share similar interests and who may be commenting on interesting blogs or bookmarking interesting articles. It may be used in expertise location scenarios where an expert is not available, but people with similar expertise may be approached [2]. Recommender systems already make use of similarity (in collaborative filtering [14]), and may gain from expanding their similarity information sources beyond their own system, and from better understanding the characteristics of different similarity sources. Promoting response for advice is another motivation for identifying and highlighting similar people: Constant et al. [8] discuss the "kindness of strangers" and argue that people are likely to provide help to people who are similar to them. Homophily, a term coined by Lazarsfeld and Merton [22], refers to the tendency of people to associate and bond with others who are similar to them. Other scenarios for leveraging user similarity information include choosing group members [18], building and maintaining a community [30], and clustering of similar users to better understand their behaviors in social applications [32].

In this paper, we set out to explore different sources of similarity information in social media. The characterization and comparison of the sources assists in choosing the best similarity source combination for various scenarios such as

information discovery and expertise location. The results of nine similarity sources for 557 users are examined through a comparison with the user's *familiarity* network and with each other, in order to expand the understanding of the characteristics of the lists of people returned by each of the sources. We observe that similarity sources belong to three categories: 1) sources related to knowing or being known by the same *people*, 2) sources related to being interested in the same *things*, and 3) sources related to being active in the same *places*. We believe that this level of abstraction can facilitate using similarity data and extending with new similarity sources. It may also ease mining and aggregation in scenarios that specifically gain from one of the categories. The nine similarity sources are examined under this hypothetical categorization and compared with the results of different aggregates (*people*, *places*, *things*, and *all*). Our experiment reveals that similarity sources are very diverse (largest overlap is 15.31%). Comparing sources to each other reveals a clear cluster of *people* sources, which also have the largest overlap with *familiarity*. The *things* sources are also overlapping, while the *places* sources do not seem to overlap each other or any other sources.

Evaluation of similarity relationships is more challenging than familiarity, as while users can easily judge whether they are familiar with someone, they might find it hard to decide whether someone is similar to them. Moreover, while users are conscious of (at least most) of their familiarity network, they might not be aware of many people who are similar to them. As similarity in general is hard to evaluate, evaluation should be narrowed to more concrete scenarios, such as "I am interested in reading this person's blogs" or "this person reflects a subset of my expertise". To this end, and in order to study the usefulness of the different sources and their aggregates in different scenarios, we devised a unique experiment in which 300 avid social media users take part. Experiment's participants receive recommendations of seven anonymized people based on different combinations of similarity relationships. Participants are presented with the evidence they have in common with each recommended person, choose which types of evidence they find most valuable, and respond to questions about the usefulness of the data for four different scenarios.

The rest of the paper is organized as follows. The next section reviews related work. We then describe our research, starting with a description of the experimental setup, then a section about our first experiment – mining characteristics of sources – and its results, followed by a section that describes our second experiment – people recommendation – and its results. We conclude by discussing our findings and suggesting future work.

RELATED WORK

There are several studies that measure user similarity through different methods and for different objectives. Schwartz and Wood [28] use graph algorithms over an email-based communication network to find people with similar interests for information discovery purposes. ReferralWeb [21] inspected the co-occurrence of names with close proximity in web documents to build a social network that is used to guide the search for people and documents. Ramanathan et al. [27] locate peers with similar interests, based on their ability to provide files to each other, in order to improve the overall performance of a peer-to-peer network. Xiao et el. [32] measure similarity of interests among web users based on different access log parameters, such as visited pages, access frequency, and order of clicks. They list different examples of how such similarity data can be leveraged to better understand web users and their behavior. In this work we make an extensive comparison between nine different similarity sources and examine their usefulness for a variety of scenarios.

One area of research that makes extensive use of user similarity information is *recommender systems*. *Collaborative filtering* [14], one of the most common approaches, is based on similarity between users. Typically, user similarity is calculated based on input of users by rating a set of items in the system. Due to the overhead of providing such feedback, leveraging implicit interest indicators [7], such as clicks, views, or queries within the system, has become more popular. The similarity relationships mined in this work stem from different social media applications and can be used by collaborative filtering systems, especially in the social media domain, for building their underlying user similarity network rather than relying solely on in-system information.

We mine nine different sources that reflect user similarity from various social media applications. Some of the relationships we inspect were previously used in the context of user similarity. Li et al. [22] show that tags used in the social bookmarking site Delicious effectively and concisely represent users' interests. They cluster users with similar interests based on their tag usage patterns. Millen et al. [24] present an enterprise social bookmarking tool, Dogear, and show that a social network of similar people can be created by connecting users who bookmarked the same URLs (see Figure 2 in their paper). Xu et al. [1] leverage both co-bookmarking of pages and co-usage of tags to enhance personalized search. Ali-Hasan and Adamic [1] study social relationships through links and comments in blogs and state that these relationships often reflect mutual interest.

Another body of related research is around the topic of *social matching*. Terveen and McDonald [31] define a framework for social matching systems as recommender systems that suggest people to each other. Foner & Crabtree [12] present Yenta – a match making system designed to find people with similar interests and introduce them to each other. Recommendation of people within SNSs has been examined by more recent research. Guy at al. [17] present a widget that recommends people to connect to within an enterprise SNS and show its high impact on the site. Chen et al. [6] compare four algorithms for people

recommendation within an enterprise SNS and show that algorithms based on social relationships outperform ones that are based on content similarity. While their focus is on finding people to connect with, they also analyze user feedback over recommended people the user did not know before. Our experiments include recommendation of individuals, however it differs from typical social matching systems, as its focus is on the similarity evidence rather than the (anonymized) recommended person.

As part of our evaluation, we compare different similarity sources among themselves and with the user's familiarity network. Previous work has dealt with comparison of similarity networks to other types of social networks in different contexts. Jung and Euzenat [20] introduce a three-layered model that includes networks between people, between ontologies, and between concepts. They explain how relationships in one network can be extracted based on relationships in another. In an experiment based on a TV quiz show, Cosley et al. [9] show that people prefer to cooperate with others who have similar demographic background, while similarity of interests does not have an effect on cooperation. Brzozowski et al. [5] present Esembly - an ideological SNS where friends, ideological allies, and nemeses are semantically distinguished. They show that although users have greater similarity with their allies, their voting behavior is affected only by their friends (positively) and nemeses (negatively). Bonhard et al. [4] study movie recommendations and examine how familiarity and similarity with the recommending individual affects the decision maker. Similarity is examined in two different ways: profile similarity and rating overlap; while familiarity is simulated through exposure to the person's profile. In their lab experiment, familiarity did not affect participants' choices, while similarity had a significant influence.

EXPERIMENTAL SETUP

Similarity Sources

In order to understand the characteristics of different similarity sources, several social media applications within a large, global organization were analyzed (listed here in order of deployment): a **forum system** containing 2,590 forums with 433,000 overall threads and 45,500 users; a **blogging system** [19], which contains 16,300 blogs, 144,200 blog entries, with 70,000 users, and 121,750 comments; a **social bookmarking** system [24] that allows users to store and tag their favorite web pages and includes 359,300 public bookmarks with 552,000 tags by 16,300 users; a **people tagging application** [11] that allows people to tag each other and is used by 9,300 users who tagged 50,000 other people with 160,000 public tags; three enterprise **SNSs** [10,11,29] that contain altogether 250,000 public connections between 99,000 distinct users; and an **online communities** system that contains 2,800 public communities, each with shared resources and discussion forums, with a total of 120,500 members.

We examine nine different sources for similarity relationships as reflected in these social media applications. The sources are: 1) *friending* - having the same friend on one of the SNSs, 2) *tagged_by* – being tagged by the same person, 3) *tag_person* – tagging the same person, 4) *tagged_with* – being tagged with the same tag, 5) *tag_usage* – using the same tag, while tags are collected from the social bookmarking system, the people tagging application, and the blogging system, 6) *bookmarks* – bookmarking the same web page, 7) *communities* – being member of the same community, 8) *blogs* – commenting on the same blog entry, and 9) *forums* – corresponding on the same forum

Table 1. Number of users for which at least *k* similar people could be extracted based on each of the sources

k=	friending	tagged_by	tag_person	tagged_with	tag_usage	bookmarks	communities	blogs	forums
1	98,018	43,469	6,769	48,365	18,597	13,726	67,006	10,500	40,789
10	84,541	40,267	2,575	41,823	17,645	9,811	65,399	4,696	17,119
100	34,088	26,976	764	24,021	16,083	4,740	55,215	835	4,320
1000	6,299	2,597	36	8,332	12,431	799	42,044	65	395
10000	119	1	0	3	3,837	6	2,258	0	16

The number of similarity relationships that can be inferred from social media sources is diverse and depends on various factors, such as the level of adoption of the different applications within the organization, the frequency of activity that yields similarity (e.g., commenting on a blog entry vs. joining a community), and the likelihood of similar activities by other users (e.g., other users commenting on the same blog entry vs. using the same tag). As the potential for inferring similarity relationships is an important characteristic of the source, and may affect its selection for certain scenarios, we inspect the number of relationships that can be inferred for each source across the organization. Table 1 shows the number of users for which at least *k* similar people could be extracted based on each of the sources, where k=1,10,…,10,000. For example, for bookmarks and k=100, the number of people who share bookmarks with at least 100 other individuals is 4,740. For small k's, friending is the richest source – almost 100,000 users have at least one similar person based on common friends and almost 85,000 have at least 10 similar people. For higher k's communities becomes the richest source, due to a few very large communities with over 1000 users, which make all their members similar to each other, in a sense. For k=10,000, tag_usage becomes the highest, even though it's only the sixth for k=1, indicating that for people who use tags, the potential for rich similarity detection is high. Overall, we see that our similarity sources have a lot of potential spanning many relationships, rendering them suitable for use in our experiments.

Harvesting and Aggregating Similarity Relationships

The collection and aggregation of the nine similarity relationships was enabled by SONAR, a social network aggregation system used before for extracting and aggregating familiarity relationships [15,16,17]. Here, the system is used analogously for similarity relationships. For a given person and source, SONAR extracts a ranked list of similar individuals. Rank is determined by a *similarity score*, which expresses the similarity strength between two individuals and is in the range of *[0,1]*. Similarity score is calculated for all sources using *Jaccard's index*, i.e., by dividing the number of items in the intersection set by the number of items in the union set. For example, similarity of *bookmarks* is the number of pages bookmarked by *both* users divided by the number of distinct pages bookmarked by *any* of these two users; for *tagged_with*, the number of tags *both* users are tagged with is divided by the number of distinct tags *any* of the users is tagged with.

Recent studies show that aggregation of relationships across different social media sites can provide a richer picture of the overall social network [13,16]. We hypothesize that it may be useful to classify the nine similarity sources into three categories to support different scenarios:

1) *People* sources related to knowing or being known by the same people: *friending*, *tagged_by*, and *tag_person*.
2) *Things* sources related to being interested in the same things: *tagged_with*, *tag_usage*, and *bookmarks*
3) *Places* sources related to being active in the same places: *communities*, *blogs*, and *forums.*

The borders between these categories are not definitive, and we examine the strengths and weaknesses of these borders in later sections.

SONAR supports aggregations of different sources by calculating a weighted average of their similarity scores to generate an aggregated similarity score in the *[0,1]* range. In our experiments we examine each source separately, as well as aggregates of sources according to the *people*, *things*, and *places* categories, and an aggregate of *all* nine sources. This results in 13 similarity *configurations* analyzed throughout the paper (nine separate sources, three categorical aggregates, and one full aggregate).

In addition to returning an ordered list of people who are similar to the user, SONAR also provides "evidence" for each person in the list, detailing all the intersection points of the user with that person. For example, evidence may include items such as "You both belong to the *Web 2.0 community*" or "You were both tagged with *cscw*".

As our goal is to compare sources of similarity, we focus on the *avid users* of social media in the organization – those who make use of all social media applications described above and for whom similarity relationships can be extracted based on each of the nine sources. We identified 557 such users, and our experiments described in the next two sections focus on this population.

CHARACTERIZING SIMILARITY SOURCES

Throughout this paper, we argue that mining similarity relationships from social media will have great value for a variety of scenarios. This hypothesis stems from the fact that mining familiarity relationships has shown great value in prior work [13,16]. Our second hypothesis is that different similarity sources hold different information and provide different value. Before we can examine the value of sources in various scenarios, we must show that:

(1) Similarity relationships are uniquely different from familiarity relationships (to prove that we are creating new value and not reusing old value from familiarity)

(2) Certain types of similarity sources are uniquely different from other similarity sources (to prove that similarity sources provide different results, and thus their aggregation may be useful for different tasks/scenarios)

Comparing Similarity to Familiarity

Method

In our first experiment we examine how unique the similarity lists returned from each of the sources are compared to the familiarity list. The familiarity list is retrieved from SONAR, which aggregates relationships from different social media and other public sources that reflect familiarity. The familiarity aggregate, as in [15,17], includes the following relationships with equal weights: organizational chart relationships, direct friending relationships within the three enterprise SNSs, direct tagging within the people tagging application, co-authorship in our organizations' projects-wiki, and co-authorship of papers and patents. Based on each of these relationships, a familiarity score in the range *[0,1]* is assigned to each pair of individuals and is then averaged across all sources (with an equal weight per relationship) to yield the overall familiarity score. This familiarity extraction technique has been extensively studied in previous work and was found to reliably reflect the set of people known to the user within the organization, in particular for social media avid users [15].

For each of the 557 subjects, the *match@100* measure is calculated [15], measuring the percentage of common people between the top 100 individuals in each of the lists. Other measures, such as *precision*, *coverage*, and *match* at values other than 100 (see [15] for more details), were also examined. As a high correlation between all of these measures is found, *match@100* is used solely for simplicity of reporting our results.

Results

The first row of Table 2 shows the mean *match@100* results over the 557 users for the comparison of each of the nine similarity sources with the users' familiarity list. *Friending*, by far, has the highest overlap percentage – 26.2%. This shows a correlation between the user's familiar people and the ones with whom mutual friends are shared. This correlation is already exploited by "people you may

know" widgets, as in Facebook [25] and LinkedIn [26], in order to recommend people to connect with. The top four sources, having over 9% overlap with the familiarity list, are the three *people* sources and *tagged_with*, implying the latter may have some relevance to the *people* category. Two other sources that somewhat overlap with familiarity belong to the *places* category. These are *communities* and *blogs*, followed by the remaining two *things* sources – *tag_usage* and *bookmarks*. The *forums* source has the least overlap. Generally, the overlap of the similarity sources with the familiarity network is not high (9% on average), assuring that the set of people examined through the similarity sources is uniquely different from the set of people the user knows. The *people* sources, as could be expected, have higher percentage of familiar people than *places* and *things*. One-way ANOVA indicates that average overlap with the familiarity list is significantly different across the nine sources ($F(8,5004)=399$, $p<.001$). Games-Howell post-hoc comparisons show that all differences between sources are significant, except for four pairs of sources: *tagged_by* and *tag_person*, *tag_usage* and *bookmarks*, *communities* and *blogs*, *tag_usage* and *blogs*. The fact that three of these pairs (the first three) fall within our suggested categories somewhat supports our suggested classification.

Table 2. Mean *Match@100* values for the nine sources

	tagged_by	friending	tagged_with	tag_person	tag_usage	bookmarks	communities	blogs	forums
familiarity	9.43	26.21	12.84	10.16	4.43	4.12	6.01	5.22	2.62
tagged_by	100	14.97	10.17	4.95	3.12	2.61	3.38	3.04	1.33
friending	14.97	100	15.31	10.52	6.21	5.10	7.50	6.25	3.05
tagged_with	10.17	15.31	100	8.28	11.06	6.56	6.54	5.86	3.18
tag_person	4.95	10.52	8.28	100	4.87	3.59	2.65	3.97	1.54
tag_usage	3.12	6.21	11.06	4.87	100	14.29	4.34	3.46	1.61
bookmarks	2.61	5.10	6.56	3.59	14.29	100	3.44	3.01	1.41
communities	3.38	7.50	6.54	2.65	4.34	3.44	100	2.52	1.53
blogs	3.04	6.25	5.86	3.97	3.46	3.01	2.52	100	2.26
forums	1.33	3.05	3.18	1.54	1.61	1.41	1.53	2.26	100
average	*5.45*	*8.61*	*8.37*	*5.05*	*6.12*	*5.00*	*3.99*	*3.80*	*1.99*

Comparing Similarity Sources

Method

In this experiment, the results from each similarity source are compared to each other in order to understand the intersections across these diverse social media activities. It is useful to understand whether different sources actually provide different data. Otherwise it may not be useful to collect and analyze each of the sources. It is also useful to examine the overlap of the sources, as clusters may provide clues about which sources to aggregate when trying to support scenarios leveraging different aspects of user similarity. Consequently, such analysis will also provide quantitative feedback on whether the *people*, *places* and *things* categorization are useful groupings. Like the previous experiment, for each of the 557 social media avid users, the *match@100* measure is used to compare the 100-person result lists from every source. This results in 36 source-to-source comparisons.

Results

The central part of Table 2 shows the mean *match@100* results over the 557 users for the comparison of each pair of similarity sources. The lists of people returned by the different nine sources are very diverse, as there is no overlap of more than 16 people (out of the top 100) between any pair of sources. This highlights the diversity of the sources and implies that aggregation would yield different results than those of a single source. *Friending* and *tagged_with* have the highest average overlap with the other sources (8.61% and 8.37% respectively), hinting that in a scenario where only a single source may be used, it may be best to choose one of them. The three *places* sources have the lowest average overlap – 3.99% for *communities*, 3.80% for *blogs*, and 1.99% for *forums*, implying that these sources encompass different information and thus including them will enrich the similarity data.

The sources that highly overlap, colored in shades of blue in Table 2, form several clusters. The largest cluster, visible on the upper left, contains the *people* sources as well as *tagged_with*. The *things* sources also have high overlaps – their smaller cluster is visible to the right of the center. *Places* do not feature this property, as they have low overlap with other sources and among themselves. As in the case of familiarity, we observe that *tagged_with* presents qualities that are typical to the *people* category, even though we classified it as a member of *things*. This is a curious result that may be explained by the fact that *tagged_with* is based on things – tags, which are given by other people. Overall, the clusters and outliers as depicted in Table 2 roughly support our initial classification of the sources into *people*, *places*, and *things,* and imply they are sensible categories for our following experiment.

PEOPLE RECOMMENDATION EXPERIMENT

In order to test the usefulness of the similarity sources and their aggregates in different scenarios, we devised an experiment, which is based on people recommendation. However, as opposed to typical people recommenders, the recommended person is anonymized and the focus is on the similarity evidence that is found for this person.

The experimental interface presents seven recommended individuals, who are found to be highly similar to the participant. For each recommended person, up to nine items are presented as evidence for the similarity relationship. In order to make sure a person's real identity does not affect the participant's judgment, the recommended person's

name and photo are kept blank. Each of the seven recommendations is based on a different configuration of similarity sources: four are based on aggregates (*people*, *things*, *places*, and *all*) and three on single sources randomly chosen out of the nine similarity sources. Participants are not asked to rate a recommended person for every single source in order to keep the experiment duration less than 10 minutes (actual average time was 8:38 minutes) and not make it too tedious. Each single source based recommendation is thus rated by approximately a third of the people. Overall, 13 different configurations are examined – nine single sources, and four aggregates. The seven configurations presented to each participant are randomly ordered, so their order does not affect the results.

For each single source based recommendation, up to nine evidence items are presented – all of the same source. For each recommendation based on a category (*people*, *places*, or *things*) up to three evidence items of each of its three sources are presented. For a recommendation based on aggregation of *all* sources, up to one evidence item is presented of each source. For a given participant and a given similarity configuration, we choose the recommended person to be the one with most evidence items from the list of top 100 similar people. Since we consider avid social media users, for many of the recommendations nine evidence items can be shown. When it is not the case, we present the maximum number of items that follow the conditions above.

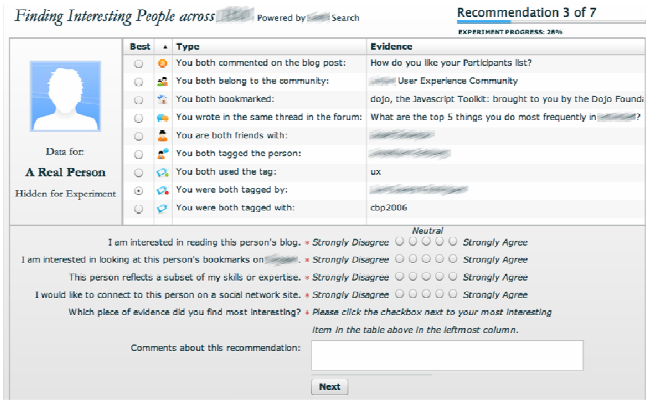

Figure 1. The experimental interface

Figure 1 shows the experimental interface. The central table lists the similarity evidence – the recommendation in this case is based on the *all* configuration and thus each item corresponds to one of the nine similarity sources. For each evidence item, a representative icon is presented, as well as a text describing the source (e.g., 'You both used the tag'), and the specific content (e.g., the tag 'ux').

Participants are requested to rate the similar person according to four different scenarios: blog discovery, bookmark discovery, expertise location, and SNS connection. There are four statements below the evidence table that correspond to the four scenarios: (S1) I am interested in reading this person's blog; (S2) I am interested in looking at this person's bookmarks; (S3) This person reflects a subset of my expertise; (S4) I would like to connect to this person on a social network site. Participants are asked to rate these four statements on a 5-point Likert scale, ranging from strongly disagree to strongly agree.

If the recommendation is based on one of the four aggregates, participants are asked to select the evidence item they find most interesting by clicking a radio-button next to it. This question is phrased vaguely on purpose, so participants can make their own interpretation and choose the item that is most prominent to them. Once participants complete these tasks, they can move to the next recommendation. They may also optionally leave a comment about each recommendation and a general comment at the end of the experiment.

The rating of S1, S2, and S4 is likely to be affected by participants' interest in blogs, bookmarks, and SNSs, respectively. For example, a person who does not normally look at people's bookmarks may not be attracted to look at anyone's bookmarks, no matter what evidence is presented. To allow finer classification based on personal interests, participants are asked to rate (on a 5-point Likert scale, ranging from strongly disagree to strongly agree) three preliminary statements, presented prior to any recommendation: (P1) I enjoy reading blogs (average: 3.87, std: 1.11); (P2) I look at people's bookmarks (e.g., Delicious) (average 3.34, std: 1.23); (P3) I see value in connecting to people on SNSs (average: 4.25, std: 1.05). As we focused on social media avid users, it is not surprising that participants' tendency is towards following social media tools. Yet, there are clear differences between connecting in SNSs (most favored), reading blogs, and looking at others' bookmarks (least popular with an average of only slightly above 3). We examine the relation between rating of P1, P2, and P3 and rating of S1, S2, S4, respectively, in the next section.

We sent a request to participate in our experiment to the 557 social media avid users, of whom exactly 300 (54%) opted to participate.

Results

Average rating results by the 300 participants for the four scenarios are presented in Figure 2. For each of the 13 configurations, we calculate the average rating over all participants who were presented with this configuration. For each scenario, results are sorted by the rating each configuration yielded, in descending order. Comments to the experiment are enthusiastic, describing it as "smart" and "intriguing". One user writes *"I think this kind of people matching could be hugely and strategically valuable, not just in helping employees expand or enrich their social networks, but serving as a strong example of the ROI of*

One-way ANOVA for each of the four scenarios indicate that ratings across the 13 configurations significantly differ

for each scenario (F(12,2070)=4.867, 4.055, 7.871, and 3.148 for S1, S2, S3, and S4 respectively). Games-Howell post-hoc comparisons are marked in Figure 2 – configurations marked by '*' yield significantly higher rating than those marked by '+'. In the next subsections we analyze the results of each scenario in more detail.

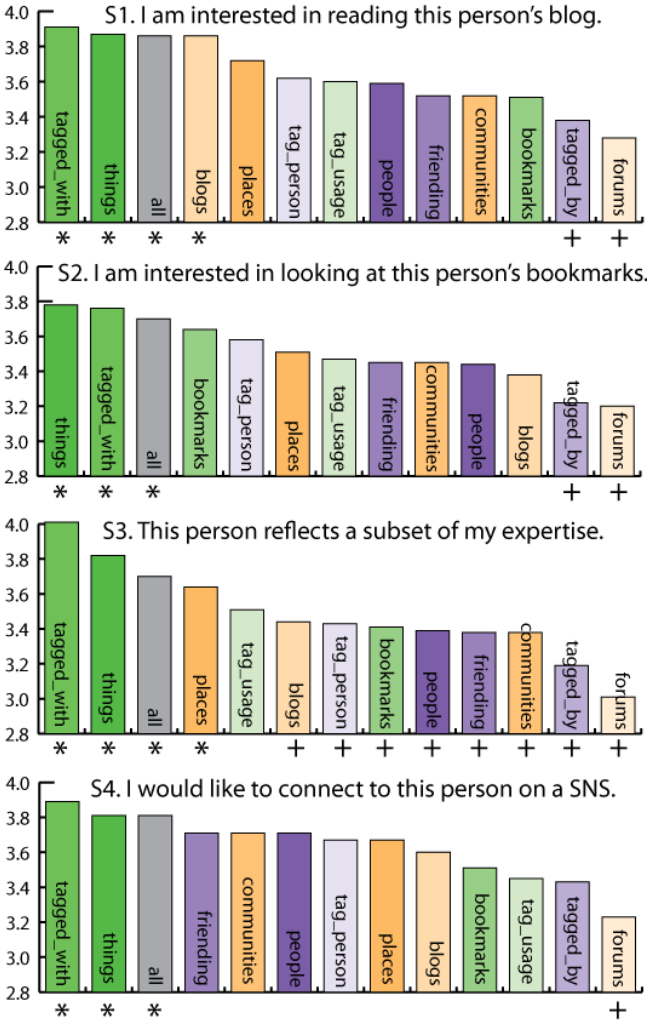

Figure 2. Average rating results for the 13 similarity configurations in each of the four scenarios (S1-S4).

S1 - I am interested in reading this person's blogs

Overall, the average response for this scenario is positive at 3.68 (std: 1.08). The response is even higher (4.13) among people who self-rate themselves as strongly enjoying blogs (answer 5 to P1, 35% of the participants).

As Figure 2-S1 illustrates, the four configurations that yield the highest results for this scenario are *tagged_with*, *things*, *all*, and *blogs*, with very small differences among them. *Places* is the fifth with somewhat lower rating than the top four. The fact that three out of the top five configurations are based on aggregation demonstrates the impact of having diverse evidence items. *Things* is the most useful category for this scenario, followed by *places*, and then *people* (only 8th). Two single sources are in the top four – *tagged_with*, which has the highest average rating, and blogs, which is expected as it is directly related to the scenario in question. Ratings by participants who self-rate themselves as strongly enjoying blogs follow the same trends, with *tagged_with* and *all* topping the list, each having a 4.41 rating.

S2 - I am interested in looking at this person's bookmarks

The average rating for this scenario is 3.54 (std: 1.09) for all participants and 4.05 for the 19% who strongly self-rate themselves as people who look at others' bookmarks.

Figure 2-S2 indicates that the top four configurations for this scenario are *things*, *tagged_with*, *all*, and *bookmarks*. Analogously to S1, these are the same three configurations together with the source that is directly related to the scenario (*blogs* in S1, *bookmarks* in S2). *Tag_person* is fifth for this scenario, which could be explained by its closeness to bookmarks ("bookmarking" a person rather than a web page). One user comments on the value of *tagged_with*: "*This person seems to be tagged with my job role. That says a lot. I'd definitely check this person out further [...] even his/her bookmarks*". As in S1, *things* (1st) is the highest category for this scenario, followed by *places* (6th) and then people (10th). The gap between *things* and *places* is higher than in S1, possibly due to the fact that blogs are in *places* and bookmarks are in *things*. *Blogs* receive a much lower rate for this scenario than for S1 (3.38 vs. 3.86), supporting the assumption that its high score for S1 is due to the particular relevance for the blog reading scenario. Examining the ratings by the 19% who strongly self-rate themselves as looking at others' bookmarks, *things* leads with a 4.25 rating, followed by *bookmarks* with 4.15.

S3 – This person reflects a subset of my expertise

Average rating for this scenario is 3.47 (std: 1.1) and distribution among the configurations is the most diverse, ranging from an average of 3.01 (std: 1.11) for *forums* to 4.01 (std: 1.17) for *tagged_with*.

As shown in Figure 2-S3, the top three configurations are, as in previous scenarios, *tagged_with*, *things*, and *all*, followed by *places* and *tag_usage*. This indicates that participants feel that being tagged with the same tags is the strongest indication for expertise similarity, stronger than, for example, belonging to the same communities, or using the same tags. The low rating of *forums* is surprising and may be a result of topics discussed being very narrow and the different roles of the asker and the answerer on a forum thread. As before, aggregate configurations have positive impact and three of them follow *tagged_with* at the top of the list. Knowing or being known by the same *people*, as might have been expected, is not considered a strong indication of expertise similarity. The fact that *tagged_with* and *tag_usage* are the highest two single sources supports the theory that tags are good interest or expertise indicators (see [22]). Yet, it is interesting to observe the difference between the two (4.01 for *tagged_with* vs. 3.51 for *tag_usage*) – it seems that the tags given by other people

(the "wisdom of the crowd") are more reflective of one's interests and expertise than one's own activities.

S4 – I would like to connect to this person on an SNS

Average rating for this scenario is 3.67 (std: 1.14) in general and 4.03 for people who strongly see value in connecting within SNSs. It can be seen in Figure 2-S4 that the top three sources are again *tagged_with*, *things*, and *all*. *Friending* comes only at fourth place, indicating that SNSs that recommend people to connect to, can do better than basing their recommendations on number of mutual friends. *People* are, as expected, more useful for this scenario than for the previous ones, and in fact this is the only scenario for which *people* ranks above *places*. *Tag_usage*, which is rated high for the expertise scenario, is near the bottom, better only than the regularly last two – *tagged_by* and *forums*. On the other hand, *tagged_with* is at the top of the list, indicating that tags that are given by others generate a lot of attraction to connect to a person, while the tags used by this person are less stimulating. In fact, tags given by the crowd are found to be more effective for friending recommendations than the common "mutual friends" or "mutual communities" methods [25,26]. The 55% of the participants who indicate they see strong value in SNS connections, rate *all* highest with 4.18, *tagged_with* with 4.16, *things* with 4.13, and then *blogs* with 4.12.

Most Interesting Sources

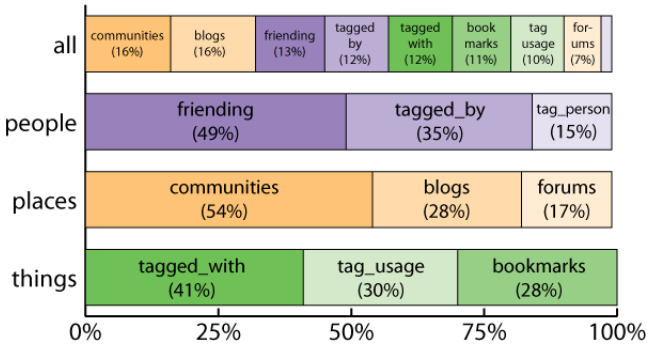

Figure 3. Distributions of sources of most interesting items for the four aggregate configurations

Figure 3 shows the distribution of answers to the "most interesting evidence item" question for the four aggregate configurations. The *all* configuration allows comparison of all nine similarity sources. Results are quite different from the ratings of the prior four scenarios, where *tagged_with* is always the top single source. *Communities* and *blogs*, of the *places* category, lead the list. Following are *friending*, *tagged_by*, and *tagged_with* with somewhat lower figures. *Forums*, as in the four scenarios, has low figures, but *tag_person* is by far the least interesting, chosen for only 2% of the recommended persons. The differences from the scenario rating are interesting, and may be explained by some comments left by participants. One indicates that "*I was looking for that something 'special' […] I think seeing people that relate to me with outlier information, relevant to me, but things I know less about, is more compelling*" and another explains that *"sometimes I selected a most interesting item because it stood out as an interest and sometimes because it stood out as a 'why is this here?'"*.

Aside from *forums* and *tag_person*, all other sources are split rather evenly (at least 10%), suggesting that aggregation is valuable as no single source always "catches the eye". Examining the results by summing the sources in each category, indicates that the chosen items are spread quite evenly across the three categories, with *places* having 38% of the selections, *things* 33%, and *people* 27%. Comparison of source types within each category, as reflected by items chosen as most interesting for each of the *people*, *places*, and *things* configurations, are also depicted in Figure 3. For the *people* aggregate, *friending* is chosen for almost half of the recommendations (49%); for *things, tagged_with* is chosen most commonly (41%); and for *places*, *communities* are the most popular, chosen for over half of the recommendations (54%).

DISCUSSION AND FUTURE WORK

Our experiment examines four very different scenarios where similarity information may be useful. However, there are many commonalities across the results of all scenarios. The three configurations that consistently get the highest rating are *tagged_with*, *things*, and *all*, while the two configurations that get the lowest ratings are *tagged_by* and *forums*. Differences between the best and the worst configurations are statistically significant. While there are other scenarios for utilizing user similarity that are not examined in our experiment, we believe that this evident consistency allows reaching a few general conclusions about the quality of similarity sources and their aggregates.

The superiority of *tagged_with* as a similarity source is noticeable throughout the experiment. It is the top rated single source for all four scenarios and is often rated higher than aggregates. The offline experiments shows that while it belongs to the *things* category, *tagged_with* also has qualities of *people* sources as reflected in the comparison with familiarity and the other similarity sources. Indeed, while the similarity inference is performed through a thing – the tag – those tags are given by the crowd. It seems that the "wisdom of the crowd" is leading to the best results in terms of mining a similar person and presenting similarity evidence. Other sources drawn from the people tagging application, like *tagged_by* and *tag_person* (and also *tag_usage* to some extent) do not produce as good results as *tagged_with*, presumably as they do not combine things and people in such a unique way. The usage of people tagging indicates that a decent amount of *tagged_with* relationships can be inferred, e.g., Table 1 shows that over *24,000* users have at least *100* similar people based on being tagged with the same tag. We note that even though people tagging applications exist both in the enterprise ([11,29]) and on the Web (for example, the Tagalag.com service or the Collabio Facebook application [3]), they are not very widespread. We hope that revealing the potential power of people tags

for mining user similarity will serve as further motivation for the promotion of people tagging applications.

Guy et al. [16] showed the value of aggregating social network information that reflects familiarity – the more sources considered, the closer the resulting network is to the user's ideal buddylist. For similarity sources, the value of aggregation is also substantial. Aggregates such as *things*, *all*, and *places* are among the top-rated configurations in all scenarios. The ratings of the aggregates are always higher than the average rating of their sub-sources and in some cases higher than any of those alone. Many comments from participants indicate that users prefer diverse evidence items over a monotonic list. These comments mostly refer to the evidence shown for the *all* configuration, which includes one item per source. For example, one participant comments that "*People I have different things in common with seem to be more interesting than those where the commonality lies only in one category*" and another adds that "*It is really the combination of these data points that is interesting"*. The distribution of the most interesting evidence items also supports the value of aggregation, as it implies that users like surprising items in their evidence list and do not favor one evidence type over the others. Finally, aggregation has value in terms of producing richer information and spanning more users than single sources. Since people tagging may not always be available, aggregation may offer an appropriate alternative, despite the additional work required for mining multiple sources.

In terms of categories, *things* is rated higher than *places*, which is generally higher than *people*. This order is exactly opposite to their order in terms of overlaps with the familiarity list, implying that the scenarios we examine are very different from the ones for which familiarity has been proven valuable. Even for the SNS connection scenario, *people* sources are not found to be particularly valuable and are outperformed by *things*. It seems that *people* is less effective for similarity detection – the fact that another person connects or tags the same people is not perceived as a great indication of similarity in any of our scenarios. One participant comments that "*In a multi-disciplinary organization like ours, it is difficult to determine skills-relevance from friendships. But the friendships are compelling, so I would be socially interested in this person*". *Things* and *places* are more valuable as similarity indicators, with a clear advantage to the former (much due to the large gap between *tagged_with* and *forums*).

One of the surprising results of our experiment is the consistently low rating for corresponding on the same forum threads. In all scenarios, the *forums* source is among the two lowest and is rarely chosen as source for the most interesting evidence. Offline experiments show that *forums* returns very different results from any other similarity source and the familiarity list. Apparently, this list of people is not very valuable. We believe that this is mainly due to the way forums are used within the organization – many users visit a forum to get answers to a question that does not necessarily represent their interests or expertise. One representative response for a forum evidence item is "*We just experienced the same problem - somewhere!*".

As mentioned in the introduction, we inspect user similarity based on common activity in social media rather than user profiles. Yet, social media also presents an opportunity to mine demographic data as more online user profiles keep popping up. Aggregating these profiles may serve as good basis for computing demographic similarity. Exploring the commonalities and differences between demographic similarity and activity-based similarity can be an interesting topic for future work.

Our future plans include implementing a non-anonymized people recommender that would recommend similar (yet unfamiliar) people to the user, based on our aggregated similarity sources. Comments in our experiment indicate that users are highly interested in people recommendations to help them extend their network (rather than of people they already know, e.g. [17]). We also wish to examine similarity sources in social media outside the enterprise, where other scenarios for exploiting user similarity become relevant (e.g., analyzing customer behavior). Our results may have been affected by the way social media is used in the organization, and may change on the Web.

Another interesting direction for future exploration, which spans beyond similarity between users, is the "transitivity" of similarity. For example, we considered using *the same* tags, which could be extended to using *similar* tags, and then to *similar* people to those who use similar tags. See [1] for an example of how such transitivity is used. The quality of the transitive similarity results should be compared with the basic results described in this work.

CONCLUSION

In this paper, we examine mining nine different sources of similarity relationships in social media applications. Inspecting the data of 557 social media avid users, the similarity sources produce lists of similar people that differ substantially from lists produced from familiarity sources. This is a notable finding, as familiarity sources are a current focus of research, and this suggests that similarity sources provide unique data that may assist a variety of scenarios for which familiarity lists are not appropriate. Furthermore, similarity sources produce very diverse lists, which suggests that aggregating them may produce richer results. Examining the similar characteristics among sources reveals that categorizing them according to *people*, *things*, and *places*, is productive.

An experiment featuring 300 avid users of social media extends the mining results. The user experiment evaluates similarity sources for four different scenarios: blog discovery, bookmark discovery, expertise location, and social network site connections. For each of the four scenarios, the experiment conclusively shows that similarity evidence generate positive user interest in these social

media tasks. The experiment highlights a particularly rich similarity source that is most useful across all of the scenarios – being tagged with the same tag within a people tagging application. The experiment also shows that aggregating similarity sources yields analogous richness. This is particularly useful since people-tagging applications have not yet gathered widespread adoption and aggregates may be an appropriate alternative. Among the categorical aggregates examined, *things* is most effective, followed by *places*, and then *people* – in opposite order of their overlap with familiarity. This indicates that users value *things* like tags and bookmarks for the four social media scenarios.

Both the mining results and the user experiment show that user similarity in social media applications has great value. It is clear that users have much interest in similar people that share their tags, bookmarks, friends, blogs, and communities. This conclusion opens a door to many research directions, as users' regular activities on social media sites may be leveraged to provide new ways to unite similar people and interests across the Internet.

REFERENCES

1. Ali-Hasan, N., & Adamic, L. 2007. Expressing social relationships on the blog through links and comments. *Proc. ICWSM'07*.
2. Balog, K. & de Rijke, M. 2007. Finding similar experts. *Proc. SIGIR '07*, 821-822.
3. Bernstein, M., Tan, D., Smith, G., Czerwinski, M., & Horvitz, E. 2009. Collabio: A game for annotating people within social networks. *Proc. UIST'09*.
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. Brzozowski, M. J., Hogg, T., & Szabo, G. 2008. Friends and foes: ideological social networking. *Proc. CHI '08*, 817-820.
6. Chen, J., Geyer, W., Dugan, C., Muller, M., & Guy, I. 2009. Make new friends, but keep the old: recommending people on social networking sites. *Proc. CHI '09*, 201-210.
7. Claypool, M., Le, P., Wased, M., & Brown, D. 2001. Implicit interest indicators. *Proc. IUI '01*, 33–40.
8. Constant, D., Sproull, L., & Kiesler, S. 1996. The kindness of strangers: the usefulness of electronic weak ties for technical advice. Organization Science 7 (2), 119–135.
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. DiMicco, J.M., Geyer, W., Millen, D.R., Dugan, C., & Brownholtz, B. 2009. People sensemaking and relationship building on an enterprise social network site. *Proc. HICSS'09*, 1-10.
11. Farrell, S., Lau, T., Nusser, S., Wilcox, E., and Muller, M. Socially augmenting employee profiles with people-tagging. In *Proc. UIST '07* (2007), 91-100.
12. Foner, L. N. & Crabtree, I. B. 1997. Multi-Agent Matchmaking. *Software Agents and Soft Computing: Towards Enhancing Machine intelligence, Concepts and Applications*. Springer-Verlag, London, 100-115.
13. Gilbert, E. & Karahalios, K. 2009. Predicting tie strength with social media. *Proc. CHI '09*, 211-220.
14. Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. Using collaborative filtering to weave an information tapestry. *Commun. ACM 35, 12* (Dec. 1992), 61-70.
15. Guy, I., Jacovi, M., Meshulam, N., Ronen, I., & Shahar, E. 2008. Public vs. private – comparing public social network information with email. *Proc. CSCW'08*, 393-402.
16. 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.
17. Guy I., Ronen I., & Wilcox E. 2009. Do you know? recommending people to invite into your social network. *Proc. IUI'09*, 77-86.
18. Hinds, P. J., Carley, K. M., Krackhardt, D., & Wholey, D. 2000. Choosing work group members: Balancing similarity, competence, and familiarity. *OBHDP* 81, 2 (2000), 226–251.
19. Huh, J., Jones, L., Erickson, T., Kellogg, W. A., Bellamy, R. K., & Thomas, J. C. 2007. BlogCentral: the role of internal blogs at work. *Proc. CHI'07*, 2447-2452.
20. Jung, J. J. & Euzenat, J. 2007. Towards Semantic Social Networks. *Proc. ESWC'07*, 267-280.
21. Kautz, H., Selman, B., & Shah, M. ReferralWeb: Combining social networks and collaborative filtering. Commun. ACM 40, 3 (Mar. 1997), 63-65.
22. Lazarsfeld, P. F. & Merton, R. K. Friendship as a social process: A substantive and methodological analysis. Freedom and Control in Modern Society (1954), 18-66.
23. Li, X., Guo, L., & Zhao, Y. E. 2008. Tag-based social interest discovery. *Proc. WWW '08*. 675-684.
24. Millen, D.R., Feinberg, J., & Kerr, B. 2006. *Dogear*: Social bookmarking in the enterprise. *Proc. CHI'06*, 111-120.
25. Offcial Facebook blog: http://blog.facebook.com/blog.php?post=15610312130
26. Official LinkedIn Blog: http://blog.linkedin.com/blog/2008/04/learn-more-abou.html.
27. Ramanathan, M. K., Kalogeraki, V., & Pruyne, J. 2002. Finding Good Peers in Peer-to-Peer Networks. *Proc. IPDPS'02*, 24-31.
28. Schwartz, M. F. & Wood, D. C. Discovering shared interests using graph analysis. *Commun. ACM* 36, 8 (Aug. 1993), 78-89.
29. Social Software for Business: http://www-01.ibm.com/software/lotus/products/connections/.
30. Sumi, Y. & Mase, K. 2000. Supporting awareness of shared interests and experiences in community. *Proc. Group'00*, 35-42.
31. Terveen, L. and McDonald, D. W. Social matching: A framework and research agenda. *ACM Trans. Comput.-Hum. Interact. 12, 3* (2007), 401-434.
32. Xiao, J., Zhang, Y., Jia, X., & Li, T. 2001. Measuring similarity of interests for clustering web-users. *Proc. ADC'01*, 107-114.
33. Xu, S., Bao, S., Fei, B., Su, Z., & Yu, Y. 2008. Exploring folksonomy for personalized search. *Proc. SIGIR'08*, 155-162.