Harvesting with SONAR - The Value of Aggregating Social Network Information

Ido Guy, Michal Jacovi, Elad Shahar, Noga Meshulam, Vladimir Soroka
IBM Haifa Research Lab
Mt. Carmel, Haifa 31905, Israel
{ido, jacovi, elads, noga, vladi} @il.ibm.com

Stephen Farrell
IBM Almaden Research Center
650 Harry Road, San Jose, California
sfarrell@almaden.ibm.com

ABSTRACT
Web 2.0 gives people a substantial role in content and metadata creation. New interpersonal connections are formed and existing connections become evident. This newly created social network (SN) spans across multiple services and aggregating it could bring great value. In this work we present SONAR, an API for gathering and sharing SN information. We give a detailed description of SONAR, demonstrate its potential value through user scenarios, and show results from experiments we conducted with a SONAR-based social networking application within our organizational intranet. These suggest that aggregating SN information across diverse data sources enriches the SN picture and makes it more complete and useful for the end user.

Author Keywords
Social networks, SN, social network analysis, SNA, aggregation.

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

INTRODUCTION
Social software – software that has people as its focal point – is the core of Web 2.0. From blogs and wikis through recommender systems to social bookmarking and personal network systems – social applications proliferate. In continuation to its dominance on the internet, social software has recently emerged in organizations, as a mean of connecting employees in a better way and enhancing knowledge management and expertise location. Blogging systems [14], social bookmarking [19], and people tagging [7], are examples of social applications that became part of organizations' intranets in the purpose of promoting intra-organizational interaction.

Many of these social applications expose interesting information about people's relationships. For example, by analysing blog commenters, or bookmarking similarities, connections among people can become evident. By extracting this information and aggregating it across multiple sources, a comprehensive and often intriguing picture of individual and organizational social networks may be revealed.

Potential sources of social information are very diverse. Different users make use of different tools, and social information is scattered among many services and applications. As these applications rarely interoperate, each is typically only aware of its own social data and cannot benefit from other applications' data.

The diversity of sources of social networking data also brings a variety of semantics to interpersonal connections. While some of these semantics are straightforward and derived from the nature of the connection (brother, close friend, manager, etc.), others are more complex. Many researchers recognized that people are connected through artifacts. For example, many studies have investigated networks where two people are connected if they have co-authored a paper [22]. Newer examples of artifacts that connect people include email messages [3,28], and web pages [14]. Affiliation networks [29] present people's connections through groups in which they co-participate, such as a board of directors, or a movie cast [22]. The above examples imply that aggregating social network data presents the challenge of creating a single framework, general yet informative, to fit all types of connections.

To address the above challenges, we introduce SONAR (Social Networks Architecture) – an API for sharing social network data and aggregating it across applications to show who is related to whom and how. Applications implementing the SONAR API (SONAR *providers*) should provide internal information about how strongly people are connected and by what means. SONAR *clients* can use the SONAR API to access data from a single provider that implements the API. However, the more compelling case is

CHI 2008, April 5–10, 2008, Florence, Italy.

where an intermediate component, an *aggregator*, is used by clients, with the very same SONAR API, to consolidate the data from different providers. This way, one can choose multiple providers and assign an appropriate weight to each of them. It is expected that federating more data providers will make the resulting answers to queries more complete.

Clients of SONAR may vary from expertise miners, through network visualizers, to user interface widgets. SONAR answers questions such as "who does this person communicate with most?", "what are all the artifacts co-authored by these two individuals?", "whom should I invite to a brainstorm on a certain topic?"

SONAR includes two types of data sources: personal (private) and public. Personal sources, such as email and instant messaging (IM), are only available to their owner and reflect the owner's personal, or egocentric, social network (i.e., all nodes in the network are directly related to the owner). Public data sources, such as blogs and organizational charts, are available to all users and reflect their extended, or sociocentric, network. SONAR maintains the privacy model of its data sources: only those users who have access to a certain piece of data by the original provider will have access to the social information extracted from this data by SONAR.

As of now, we implemented the SONAR API for over ten sources, public and personal, within the IBM intranet. SONAR's ultimate goal is to be widely used by social networking and Web 2.0 services on the internet and to define a standard, analogous to RSS [11]. SONAR is based on the REST design pattern [9] and uses standard data formats such as Atom [23] and JSON [13].

When used for creating a sociocentric view of a social network, SONAR is based solely on public sources. Any user may use this view to examine publicly visible connections within any group of people. When used for creating an egocentric view of a user's network, SONAR also makes use of the user's personal sources. Enriching the egocentric network, as reflected in personal sources, with information from public sources, opens up new opportunities for learning about one's extended network (i.e., one's connections and their connections with others). Consider, for example, Alice who seeks a social connection to Cindy. Cindy may not appear at all in Alice's egocentric network based on personal sources. However, examining the extended network, Alice may discover that Bob – who appears on her egocentric network by her personal data – is related to Cindy according to public sources. Alice will then be able to discover a social path to Cindy through Bob, based on aggregation of her personal and public sources.

In order to verify the fundamental concepts on which SONAR relies, namely aggregation and the usage of public sources, we conducted three experiments. The first experiment examines different social networks derived from four of SONAR's implemented public data sources. The second experiment involves a user study in which over a hundred users evaluated different buddylists derived from 24 different combinations of SONAR public as well as personal data sources. For the third experiment we interviewed 12 users about their usage of a SONAR UI and their thoughts on the different buddylists. The experiments examine the diversity of the public data sources, their value to the user, and whether they add value over personal data sources. Finally, we checked whether there exists an ideal weighting scheme of the sources, which may serve as the system's default, and followed users as they were composing their own ideal weighting scheme.

We note that the experiments in this paper examine buddylists and social networks generally, while, in practice, different semantics may yield different social networks. For example, the network consisting of users' friends may be different from the network of people with whom users communicate most frequently, which may be different from the network of individuals with whom users share similar interests.

The rest of the paper is organized as follows. The next section surveys related work, followed by a more detailed description of SONAR and typical usage scenarios. We then describe our hypotheses, research method, and results. The final section discusses conclusions and future work.

RELATED WORK

The formal discipline that studies interpersonal connections is called social network analysis (SNA) [29]. A social network (SN) is a graph that represents social entities and relationships between them. SNs have often been studied through the use of social science tools such as surveys and interviews, which require a great deal of human labor. The evolution of the Web, which is often referred to as Web 2.0 [24], has introduced new possibilities for SN research.

Gathering SN Data from Computer Applications

There are several popular SN services on the Web, which help to connect friends and business acquaintances[1]. These services define an explicit SN. Users directly specify who their friends are and often manually state the nature of the connection. The manual nature of these networks is perhaps their main disadvantage. Only part of the user's actual SN will be registered in any such application, and since explicitly entering social data is tedious, even users who are registered are likely to have incomplete information about their network. These applications are useful for SONAR, providing very accurate, even if partial, social connections.

The wealth of information in computer databases and applications is a good source for automatically obtaining various kinds of SN data without burdening users with the manual management of their network. For example, Wellman views computer networks as SNs and surveys how computer networks reflect and affect traditional SNs

[1] {myspace, facebook, linkedin, orkut, friendster}.com

[30]. In addition, data mining can discover subtle details of a relationship that a person may not be able to provide accurately. For example, the rate of email communication can be used to estimate the strength of social ties [19].

Email is commonly used to extract SNs. Extraction of the sociocentric SN from email logs has been demonstrated in [28] in order to automatically identify communities. Other tools, such as ContactMap [19] and Personal Map [8], analyze emails to extract the user's egocentric SN. Email is one of the important sources for SONAR, but it is definitely not the only one. SONAR's philosophy states that there is much important social information outside the inbox.

Another common source for SN information is the Web. Adamic et al. [1] describe techniques for mining links and text of homepages to predict social relationships. The strength of a connection between two people can be estimated by querying a search engine with their names and checking for web pages where they co-occur [14,17]. Since various data sources such as papers, organizational charts, and net-news archives are available on the internet or intranet, they too could be mined by web searching [14]. SN extraction has been conducted on many other sources, including Usenet data [27] and Instant Messaging logs [25].

Public vs. Personal Data Sources

Sociocentric network approaches may encounter difficulties using personal data sources due to privacy concerns. For example, mining email, even when results are displayed in aggregated forms, might expose private information [14].

One solution is to limit data mining to public sources. For example, Aleman-Meza et al. [2] chose to aggregate only publicly available SN data due to privacy concerns. However, ignoring private information may exclude important data. An alternative solution is to have users opt-in to explicitly give permission to make certain private information public. For example, Smarr [26] suggests requiring users to opt-in to publish their information to a public Friend of a Friend (FOAF) file [10]. However, opt-in requires action and motivation on the part of users – so those publishing their FOAF files will cover only a small percentage of organization members.

SONAR can aggregate both private and public sources, without exposing private data to other people, or requiring users to opt-in.

Aggregating SN Data from Different Sources

No single archive or tool captures all our social relations with others. Aggregating SN data from several sources may improve the completeness of constructed SNs. For example, the ContactMap developers plan to extend their email mining tool to use sources such as voice mail and phone logs due to user complaints on the absence of phone-based contacts [19].

Several tools combine different sources to generate better SNs. In [5], an email database is used to extract people names and email addresses – these are then searched on the Web, to extract keywords and to find additional related people for which the search is recursively applied. Web mining, face-to-face communication, and manually entering one's SN are combined in [12] to construct the SN in a Japanese conference.

A basic problem of aggregation is deciding on the algorithm to be used for combining data from various sources. A simple approach is to compute the aggregation as a linear combination of the individual sources. Cai et al. [4] propose a method for learning the optimal linear combination given input from the user that describes the user's expectation. However, this method requires the user to specify the query in ways which may prove to be complex. Matsuo et al. [16] extract and integrate SNs from several different sources. They provide a rough sketch for integrating the networks into a single one by using a linear combination of the different networks. Our research tests whether there is a weighting scheme which is appropriate to most users, for a specific scenario.

Evaluation of Social Network Quality

A common approach for validating an automatically constructed SN is to ask the people in the network about its correctness, usually by questionnaires that require rating various aspects of the SNs and/or providing open-ended feedback [8,19,28].

Another possible approach used in SNA is to compare the automatically constructed network to external data. For example, Aleman-Meza et al. [2] detect conflict of interests between paper authors and referees by integrating data from several SNs. Their evaluation included comparison to data from an external source: a different existing system for detecting conflict of interest. The problem with this method of evaluation is that it is possible only if there exists a relevant external source, which can be used for comparison.

This paper includes three types of evaluation. In the absence of an external source to compare to when evaluating an aggregated network, our first evaluation compares the networks obtained from individual sources to the other sources, to measure their uniqueness. The second evaluation is a user evaluation, where we evaluate and compare numerous linear combinations of different sources. The third evaluation employs interviews in order to receive user feedback on individual sources and their aggregation.

SONAR IN DETAIL

The purpose of the SONAR API is to provide open interfaces to SN data, "locked up" in a multitude of systems. SONAR specifies a way to share weighted SNs as relation lists. Clients may retrieve information about how people are connected based on different parameters. Like RSS, our goal is to make a read-only interface, simple to implement by the *provider* and consume by the *client*.

The first premise behind the SONAR API is that it is necessary to present the *strength* of ties between people. In contrast to APIs for specific SN applications like Facebook [6], SONAR does not model any specific semantics of the underlying system like "friending" or "communities". Instead, it asks providers to boil down these semantics into floating point numbers between 0 and 1. SONAR aggregators combine results from multiple systems using a simple weighted average. This approach enables diverse applications from instant messaging clients through publication databases to SN sites to provide data supporting an aggregated view of relationships among people. SONAR clients are oblivious to the types of relations – when querying for strength of a relationship, all that the client sees is people and the weight of their associations.

Users of aggregated SN data frequently want to understand *how* people are connected, or *why* a connection is stronger than another. To support this need, SONAR allows queries for *evidence.* Evidence is essentially a time-ordered log of entries, originating from each of the providers. It may include comments posted by one user in the other user's blog, email messages or chat transcripts between them, or web sites that they both bookmarked. According to our privacy model, users do not have access to any private material of other people through this interface—it just organizes information they already had access to before.

SONAR API Specification

We have implemented SONAR as a REST API. The API has four methods. The first three are fetching weighted people relationships, while the fourth provides evidence for connections. The methods are summarized in Table 1.

SONAR Aggregator

SONAR is designed to enable an *aggregator* component that merges results from multiple providers. Like an HTTP proxy, the aggregator protocol is in most ways identical to the protocol for interacting with a SONAR provider. In fact, clients communicate with aggregators the exact same way they do with primary providers – through the SONAR API. The aggregator is configured to connect to one or more providers. When a request is received, the aggregator forwards it to each provider. It then processes the results by computing a weighted average and returns the result to the user. The original results from the different sources remain transparent to the user.

Name	Parameters*	Output
`Strength`	*source (user) , target (user)*	*Float (0.0 to 1.0)*
`Relations`	*user(s), limit, offset*	*<list of people>*
`Network`	*user(s), degrees, threshold*	*<graph of people>*
`Evidence`	*users(s), limit, offset*	*<list of entries>*

Table 1. SONAR API methods*

* All methods also accept parameters *since* and *until*. *Since* limits results to those after the given date, *until* limits to those before the given date.

SONAR Providers

We have implemented SONAR providers of over ten sources in our organizational intranet. The following four systems, which serve as public sources, were used in our experiments: BlogCentral (IBM's corporate blogging system [14]), Fringe, for people tagging and friending [7], Dogear, for social bookmarking [19], and the IBM organizational chart.

For the blog system, social relations are derived from the comments made to one's blog. This information is an indication of the people who leave a trace in a blog, which is likely to imply that the author is aware of them. Fringe supports extraction of social information of both friending and tagging. Friending is a reciprocal action: one person invites the other to be friends and they are defined friends only if the invitation is accepted. Tagging people is one sided, yet indicates some level of connection. The SONAR provider that extracts SNs from Dogear is based on bookmark similarity information. The connections returned by this provider are those of people who bookmark the same pages. From the organizational chart we extracted, for each user, the user's manager as well as the user's direct peers - all employees who have the same manager.

We have implemented several client-side SONAR providers that have access to the user's private data. The experiments in this paper use two of these – email and chat transcripts. The outcome of these providers is only visible to the owner, visualizing an egocentric map of connections, but not revealing any private information to others.

For the email information, our client requests the user's password and then crawls the mailbox and collects details of people the user corresponds with. The chat information is easily accessible to our SONAR client, as the client is implemented as a plugin of Lotus Sametime, IBM's chat system. We extract social information from the history of chat transcripts, as these indicate the people a person actually chats with.

SONAR Usage Scenarios

To demonstrate the potential usage of the SONAR API, we created SonarBuddies – a plugin for Lotus Sametime. The plugin presents an alternative buddylist, which consists of the people most strongly related to the user, ordered by their strength of connection (see Figure 1(a)).

Additional features include showing related people to any buddy on the list, the connection points (evidence) with a buddy (Figure 1(b)), and people who are connected to both the user and a buddy (Figure 1(c)).

The SonarBuddies extension has a preference page in which the user may choose the relative weight of each data source, the number of buddies to display, and the number of days in history to consider. When adjusting the preferences, the user may see a preview of the buddylist. This enables fine-tuning the selection of weights (see Figure 2).

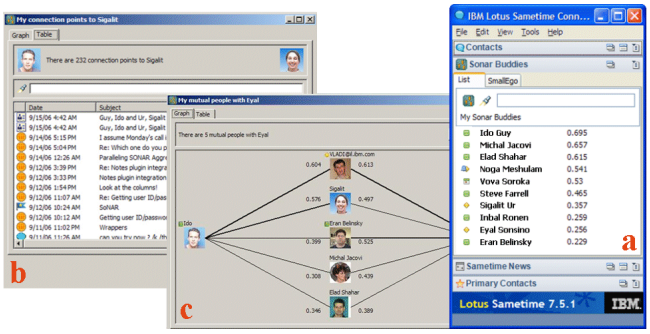

Figure 1. SONARBuddies UI

SonarBuddies is just one example of a potential SONAR client. Below, are a few examples of more advanced scenarios that SONAR may support:

- **Expertise location** (e.g., [14,18,31]) – SONAR, integrated with search, may be used for scenarios of expertise location, such as the basic "who knows about <*topic*>?", but also "who do I know that knows about <*topic*>?", and the related "who do I mostly communicate with about <*topic*>?"
- **Automatic completion** of names (e.g., [21]) and groups– completing a single string to a name may sort alternatives by strength of social ties and relevance to the context. Moreover, the completion of a whole group can be supported - e.g., if one participates in a project of 10 people, typing the names of 3 of them may automatically be completed to the entire group. This may also be useful for resolving "who's missing from the mail I'm about to send?"
- **Finding social paths** to someone who is not directly related to the user (e.g., [15])
- **Enhancing SN services** by recommending people to connect to based on other evidence and by enriching information about existing friends: connection strength, evidence, and temporal characteristics of relationships

SONAR EXPERIMENTS

SonarBuddies has been made available for download at the IBM intranet and over 1800 users downloaded and used it. SonarBuddies is mainly an egocentric application and it could be assumed that its success is due to heavy usage of private data sources by the users. However, from the usage data it is evident that public sources are explored. We envision the SONAR API as being heavily used by sociocentric applications and thus the interest in public data sources was encouraging.

This section states our hypotheses, describes our experiments, and discusses the results.

Hypotheses

Public Data Sources

The first group of hypotheses focuses on public data sources and their influence on the users' SN. Public data sources are different in nature, ranging from social bookmarking systems to blogs and therefore we assume:

(1) Public data sources provide diverse SN information. There is no single public data source that holds all SN information, and each public source makes a significant contribution to the overall SN information.

Moreover, as part of the user activities are performed "outside the mailbox", the public sources provide valuable SN information which is not reflected in private sources. We thus raise the following 2 hypotheses:

(2) Public data sources provide SN information that is valuable to the user.

(3) Public data sources enrich egocentric SN information. By combining private sources with public sources, one can potentially get a more complete picture of the SN.

Data Source Aggregation

An additional hypothesis focuses on aggregation of SN information. A key concept behind SONAR is the ability to consolidate social information from multiple data sources, assuming that there is real value in such aggregation. The following hypothesis is explored:

(4) Aggregated SN information is of greater value to users than information that originates from any single source.

A SONAR client based on this hypothesis would need to use some weight combination in order to aggregate the different sources. While the user may have control over the weights, a SONAR client should have a default weight combination that would be reasonably good for all users for the most common scenario (such as finding the people the user communicates with the most). During our experiments, we wish to study the following hypothesis:

(5) For a basic scenario, there exists a weighting scheme by which an aggregation of data sources most reasonably represents most users' SN.

If this hypothesis is correct, our experiments may reveal the weighting scheme that we should use as default.

Finally, an even more valuable aggregation of SNs can be achieved if users would share some of their private SN with others. While people are hesitant to share their private information, they may agree to share the **buddylists** created based on it, and thus allow a sociocentric view that is enhanced by private information. We hypothesize:

(6) People would be willing to share the buddylists created based on their private sources

Research Method

In order to examine our hypotheses, we conducted three experiments on SN information collected by SONAR.

Experiment 1: Information from Public Sources

For the first experiment we gathered information from the four public sources. Our goal was to compare the lists of connected people from the different sources (hypothesis (1)) and show that no source covers the others and may thus serve as a single source of information (hypothesis (4)). The collaboration tools in IBM, like many Web 2.0 services on

the internet, are still in their diffusion phase and are not yet used by all. However, they are gaining momentum and have become quite prevalent. As we wanted to compare all sources, we decided to focus on users who use all of the chosen data sources. While such users are not a statistical sample, we refer to them as the early adopters of the technologies, and use their figures as a reflection of the potential of SN information that may be extracted from public sources as Web 2.0 technologies become ubiquitous [24].

We started by locating the top 1000 heavy users of each of the tools (BlogCentral, Fringe, and Dogear). For blogs, we defined heavy users as those who received most comments in their blog. For Fringe we took the 1000 users with the largest lists of connections. For Dogear, the 1000 people with most bookmarks. Once we had these three lists, we took their union and received a list of 1761 users. We then obtained results from the different SONAR providers for all 1761 users, and examined those users who had a nonempty result in all four sources (the fourth source being the organizational chart). We ended up with a list of 273 such users. For every one of the four public sources and every one of the 273 users, we calculated the number of unique contributions of this source over the union of all other three sources.

Experiment 2: Online Questionnaire for Ranking Buddylists

For the second experiment we implemented a dedicated plugin for Lotus Sametime. The plugin was easy to install and presented a questionnaire containing three sets of up to eight buddylists that were aggregated from the different sources and tailored specifically for the user. The user could not tell what the sources of the different lists had been. Before starting the experiment, we asked the user to imagine setting up an "ideal" buddylist for communication inside IBM. By this request, we framed the experiment to a basic scenario. The user was then asked to rank the buddylists by how close they were to representing the ideal buddylist. The scale for ranking was 1-4 (where "1" is good, and "4" is bad). In addition, we asked the user to mark a single buddylist as the "best" buddylist – relative to the other lists in that set.

The first set of up to eight buddylists was composed solely from public sources. Our goal with this set was to examine the value of extracting SN information from public sources (hypothesis (2)), and to compare the quality of lists created by aggregation of different sources vs. the quality of lists created from a single source. The weight combinations of sources used in this set are displayed in the leftmost columns of Table 2 (1-8). We use the term "up to eight lists", since if two different combinations created two identical lists (in both content and order), we only presented them once. If the user voted "best" for a list that was created from more than one combination, we added a vote to all these combinations.

The second set of buddylists was also focused on public sources. The goal with this set was to learn whether a specific weight combination is preferred by most users and may serve as a default (hypothesis (5)). The weight combinations of sources used in this set are displayed in the rightmost columns of Table 2 (9-16).

The last set of buddylists introduced information gathered from the user's private sources. The goal of this set was to examine the value of information public sources add over private sources (hypothesis (3)), as well as to study the effect aggregation has on the lists (hypothesis (4)) – aggregation of private sources, and aggregation of a mix of private and public sources. The weight combinations of sources used in this set are displayed in Table 5 (17-24).

Our plugin reported the user ranking of the buddylists to a dedicated server that produced a report with all results. The results visible to us did not contain any private information nor could we see the buddylists, we only examined the ranks (1-4) and the vote for best list in each set.

Experiment 3: Sliders UI for Personal Weight Combination

A set of interviews we conducted, helped us examine hypotheses (2), (4), and (6), as well as hypothesis (5). It also gave us some insight about how people perceive their SN and what they feel about our UI.

The preferences-page of the SONAR plugin allows modifying the weight combination of different sources with sliders and simultaneously seeing a preview of the buddylist created from this combination. The user interface of this feature is shown in Figure 2.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
bookmarking	1.0				0.5		0.1	0.25	0.1	0.3	0.3	0.3	0.1	0.1	0.4	0.5
people tagging		1.0			0.5		0.2	0.25	0.3	0.1	0.3	0.3	0.2	0.1	0.3	0.3
blogs			1.0			0.5	0.3	0.25	0.3	0.3	0.1	0.3	0.3	0.3	0.2	0.1
org-chart				1.0		0.5	0.4	0.25	0.3	0.3	0.3	0.1	0.4	0.5	0.1	0.1
average score	3.65	2.63	3.38	2.36	3.01	**2.42**	**2.34**	2.59	**2.35**	2.73	2.64	2.79	**2.40**	2.47	2.87	2.97
# of "best" votes	0	23	1	**63**	6	**57**	**58**	46	**70**	48	46	47	63	61	46	43
# of score "1"	0	12	0	**18**	5	14	**17**	9	**20**	4	6	9	**20**	**20**	7	6

Table 2. Weight combinations and results of the public sources in the first two sets of buddylists

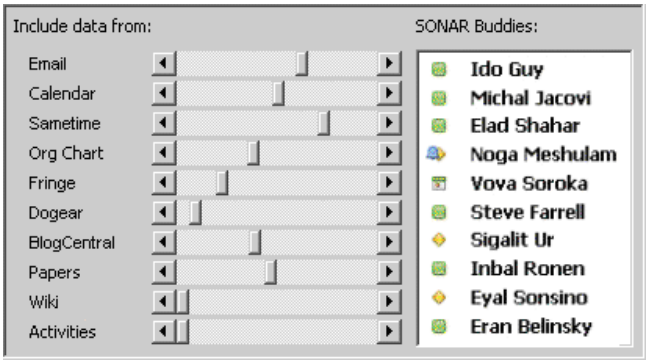

Figure 2. Weight combination user interface

We conducted personal interviews with 12 of the users who took part in experiment 2, in order to learn about their experience with the sliders and follow their line of thought while they are examining the different buddylists. Each interviewee was asked to first reset all sliders and start with an empty list. In order to examine hypotheses (2) and (4), each slider was moved separately, to compare the lists based on single sources to an ideal list (as in the framed scenario of experiment 2). Once all sources were examined, the interviewees were asked to fiddle with the sliders in order to compose a list that is closest to their ideal list. We retrieved the selected weight combinations from our logs and examined them in order to validate hypothesis (5). Finally, we posed the question about sharing the buddylists based on private sources (hypothesis (6)). The question is a multiple choice question: to share automatically vs. share after manual editing; to share with anyone, or only with friends, or only with a specific individual.

Experimental Results

Results of Experiment 1

In this experiment we examined the 273 users who had a nonempty result in all four public sources. For each of the four public sources and each of the 273 users, we calculated the number of connections extracted from the source, and the number of unique contributions of this source over the union of all other three sources. The averages of these figures appear in Table 3.

Examining the average numbers of unique contributions is interesting. One would assume that the commonalities of the lists from the different sources would be large; that a person would mostly friend with peers from the organizational group; that a person's friends would be the ones commenting in the blogging system; and even that working in the same group would imply similar interests and thus bookmarking the same web pages. However, our results reveal a different picture. It seems that for each and every source, the average number of unique contributions over the other three sources is quite close to the average number of connections, implying that the information extracted from the different sources is indeed diverse.

org chart		friending		blogs		bookmarks	
#	>	#	>	#	>	#	>
15.73	13.91	24.64	20.27	8.57	5.74	423.1	417.7

Table 3. Results of experiment 1: average number of connections (#) and unique contributions (>)

contribution	org chart	friending	blogs	bookmarks
full list	107 (39.2%)	76 (27.8%)	83 (30.4%)	55 (20.1%)
nothing	11 (4.0%)	22 (8.1%)	37 (13.6%)	0 (0.0%)

Table 4. Unique contribution over all other sources

Table 4 shows two statistics of the unique contributions of the different sources over all other sources. The first row in the table shows the number of people for whom the source contributed its full list – meaning that the lists from other sources had *no* intersection with this source. For instance, for 107 of the people (39.19%) – the lists of blog commenters, tagging friends, or similar bookmarkers did not contain *anyone* from their organizational group. As can be seen on Table 4, these numbers are rather large – 76 for Fringe, 83 for BlogCentral, and 55 for Dogear – indicating the diversity of the sources. There was not even a single person, for whom a single source covered all other sources, proving that aggregation creates a broader picture than any single source. The second row in the table shows the number of people for whom a source contributed nothing over the other lists. These figures indicate cases in which a single source may be dismissed, as the other three sources cover the information it provides. As may be seen on the table, these figures are rather small: up to 37, for BlogCentral, and as low as 0 for Dogear.

For 33 out of the 273 people examined (12.08%), there was no intersection between any of the sources – each of their public sources provided a completely different list.

The results of this experiment validate hypothesis (1) and support hypothesis (4), showing the diversity of information from public sources, that no single source holds all SN information, and thus that aggregation is likely to be of greater value than information from any single source.

Results of Experiments 2 and 3

Our questionnaire plugin collected information from 116 users who responded to all three stages of the experiment. Out of the 116 users, 65 are from the US and Canada, 49 are from Europe and the Middle East, and two are from Asia Pacific. 73 of the users who responded are using Fringe, 29 are bloggers, and 62 use Dogear. We believe these users represent a wide range of IBM employees and are thus a good test bed for our hypotheses. In addition, results collected from the in depth interviews with 12 users strengthen some of our hypotheses.

The bottom part of Table 2 shows the results of the first two sets of buddylists. On the first step, shown on the bottom left of the table, the list that got the most "best" votes (63) is the one based on a single source: the organizational chart.

The average score of this list is 2.36 (on a scale of 1=good, 4=bad). The list that got the best average score is the one based on aggregation of all four sources by weight combination 7 on Table 2. The average score of this list is 2.34 and its number of "best" votes is 58. Over 40% of the users (48) got in this aggregated list an identical list to the one based on the organizational chart. As the score of the aggregated list is better, we may conclude that when the lists were not identical, the aggregated list received better scores. Finally, note that in Table 2, the total number of "best" votes given to aggregations (i.e., combinations 5 through 8) was significantly higher than the combined number of "best" votes for all single sources (1 through 4). We consider the above findings as supporting hypothesis (4) – the value of aggregation.

The results of the second step are shown on the bottom right of Table 2. All lists of this step received scores that are higher than 2 but lower than 3, and all lists received quite a few "best" votes (over 40). It appears that no optimal weighting scheme exists that may serve a good default for most users. We were therefore unable to prove our hypothesis (5). While conducting the interviews during experiment 3, we had another chance to see how different people prefer different sources. For example, when examining the list from Fringe, one of the users said: "*Completely off. Only 7 people, out of them only 3 are familiar*" while another said: "*Accurate, very accurate actually. […] that would be my ideal buddylist*".

The results of the third step are shown on the bottom of Table 5. In this step we compared lists from private sources with lists from public sources and with lists based on a mix of private and public sources. For this step our plugin requested access to users' private data. Only 55 users granted us access to their private data. We therefore based our analysis on the responses of these 55 users only.

As can be expected, the lists based on private sources received better average scores (1.62-1.76) than those based solely on public sources. The list with most "best" votes (17) on is the list based on the (private) chat system.

	17	*18*	*19*	*20*	*21*	*22*	*23*	*24*
Bookmarking	0.25		0.16	0.1	0.2			0.1
people tagging	0.25		0.16	0.1	0.2			0.1
blogs	0.25		0.16	0.1	0.2			0.1
org-chart	0.25		0.16	0.1	0.2			0.1
email		0.5	0.16	0.3	0.1	1.0		0.4
chat		0.5	0.16	0.3	0.1		1.0	0.2
average score	2.85	**1.62**	2.40	**1.75**	2.69	**1.76**	**1.76**	**1.75**
# of "best" votes	0	**12**	2	**9**	0	**10**	**17**	**9**
# of score "1"	3	**30**	6	**23**	4	**25**	**26**	**23**

Table 5. Weight combinations and results of public and private sources in the third set of buddylists

However, 17 votes are only 30.9%, indicating that no list significantly outvoted the others. The list with best average score (1.62) is the list based on combination 18 in Table 5 (email and chat), supporting the value of aggregating information (private in this case) – hypothesis (4). Yet another supporting point for hypothesis (4) was received during the interviews, when no single user had chosen a list based solely on one source. This observation is less strong, since the experiment setting encouraged people to play with aggregations, yet they clearly had a choice to disable all other sources and stay with one, but did not.

Table 2 and Table 5 also show the number of times each of the buddylists in this experiment received score 1. List 18 (email and chat) received score 1 for the largest number of times – over 54% of the people granted it a perfect score. Other lists which obtained many high scores are list 20 (mix of public and private), list 22 (email), list 23 (chat), and list 24 (anther mix of public and private). It is obvious from the table that lists based on private sources receive score 1 more often, as expected. The value of public sources is evident from the bottom line of Table 2: the lists based solely on public sources received a perfect score a considerable amount of times, supporting hypothesis (2). In experiment 3, seven of 11 used some combination of public and private data sources (the twelfth user had no access to private data) and two of them even preferred the public sources slightly over private ones (see Table 6).

Two of the combinations that mixed all six sources (number 20 and number 24 in Table 5) received 9 "best" votes each. Examining our results reveals that each such vote was given by a different user, implying that for 18 people, they created a buddylist that is preferred over the buddylists based solely on private sources. Both these lists received an average score of 1.75, which is identical to the average score of the list based on chat, and they both received a perfect score 23 times. This implies that for quite a few people the mix with public sources creates buddylists of high quality. Together with the fact that in experiment 3 most users chose to combine private with public data, we conclude that hypothesis (3) is true - information from public sources may provide a more complete picture of one's SN.

Finally, Experiment 3 revealed what people think about sharing lists coming from their private data sources (see Table 6). Most of the people (10) said that they will be

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
prv	85	00	100	48	63	100	100	100	69	48	54	54
pub	15	100	00	52	37	00	00	00	31	52	46	46
shr	2	3	2	2	2	3	2	2	3	3	1	0

Table 6. Experiment 3 results*

* First row shows percentage of private sources, second row percentage of public sources, third row – sharing list preference (3 – automatically with anyone, 2- after editing with anyone, 1 – after editing with friends, 0 – will not share)

willing to share their lists with anyone. Seven stated that they would want to edit first. "*I am worried that SONAR results are not accurate enough and would like to […] make sure the people who see the lists get good lists*", said one user. Another user said, at first, "*[I am] worried about what my buddies would say about me sharing their names*", but, after thinking about it he decided that it was harmless enough and said he would share his lists. Four users showed great openness by declaring that they would share their list automatically, without any editing, with anyone. Only one user said that he will not share his lists with anyone. One of the users summarized these results nicely: "*... I think it should* ***always*** *be left up to the individual as there are dangerous things about SN. [but]…within a company, you have to realize you are probably not going to be that private*". All in all, hypothesis (6) is supported to a high extent. Our results suggest that there is a good chance people will be willing to share their private-based buddylists after applying some editing to it.

Discussion

Our experiments show that information from public sources is very diverse and no single source may provide all SN information. While some sources provide communication data, others provide similarity data. While reacting to lists generated from Dogear, users said: "*Far from ideal list*", "*Over 50% are strangers*". But another user, while discussing her usage of SNs described: "*if my goal was to search for expertise, then I would lean it heavily towards social bookmarking and blogs*". It suggests that diverse sources can become valuable for diverse tasks. It also explains why our hope to locate an optimal default combination of weights for aggregation was not fulfilled - different users with different views and different needs may require different combinations of the sources.

The information extracted from public sources is shown to be of value, and while private sources provide better information, public sources do contribute additional information and create a more complete picture of SNs. The value of aggregation is proven both by the diversity of public sources and by user votes for aggregated lists.

We saw in our interviews that aggregation and collection of SN information becomes crucial in global organizations. While analyzing the organizational chart data source one of the users noted: "*It just looks at people within my world, and I deal with a lot of people outside of my function, if you will*". Another user commented: "*Names that were missing on email appear now. Top person on ideal list does not appear here*". It shows that no single data source is sufficient for creating one's ideal buddylist and thus demonstrates the value of aggregation.

CONCLUSIONS AND FUTURE WORK

In this paper, we present the motivation and challenge for aggregating SN information from multiple data sources. We describe SONAR, an API for exposing relations embedded in numerous applications or services. SONAR allows building weighted networks with evidence for each connection, showing how strongly people are connected. SONAR was implemented for various sources, public and personal, within IBM, and demonstrated through a plugin for Lotus Sametime: SONARBuddies.

Our experiments indicate that aggregation produces a more comprehensive SN. Information coming from public data sources is shown to be relevant and diverse. Public sources mainly represent new emerging social technologies on the Web. We believe that people's involvement in these technologies will continue to grow, while new technologies appear. Hence, more quality social information will be available for frameworks like SONAR. For some users, public sources make a significant contribution over private ones, which may have been considered the predominant sources. It is extremely interesting to continue examining the potential of public sources to actually replace SN information currently extracted from private sources, and thus relieving privacy issues.

This paper focuses on general SNs. In practice, there are different types of SNs, which reflect different semantics of connections, like friendship, interpersonal communication, or similarity. Such networks tend to be semantically different even for the same user. Shared bookmarks, for instance, reflect similarity between users rather than a direct connection. The scenario used in our experiments asked the users to rank buddylists, typically used for communication. It was natural that Dogear, which exposes similarity, received low scores by our users. One could think of different scenarios, such as finding potential people for a community on a specific topic, where this similarity network would be a perfect networking tool. It would be interesting to identify different scenarios and examine the contribution of different sources to them.

Another interesting direction is deriving contextual networks – networks that are related to a specific context or term. For example, the list of people with whom one communicates most frequently about Java, is likely to be different from the list of people with whom one communicates mostly about SN analysis.

Our plans for future work on the SONAR API include extending it to support different types of relations (e.g., familiarity vs. similarity). We also plan to allow a specification of a search term to support contextual queries such as: "who is most related to <person> w.r.t. <topic>?" These extensions would allow us to further explore the variety of SNs, the differences and relations among them, the value they bring to users, and the patterns of their usage.

ACKNOWLEDGEMENTS

We would like to acknowledge all those who installed SONARBuddies and provided us with feedback about its usage. We are especially grateful to those who participated in our experiments and interviews. We thank James Snell

for providing us data about the usage of BlogCentral, and Jonathan Feinberg for data about Dogear. Sigalit Ur and Inbal Ronen participated in numerous discussions and provided enlightening comments, we are indebted to them.

REFERENCES

1. Adamic, L. A., and Adar, E. Friends and neighbors on the Web, *Social Networks, 25*, 3 (2003), 211-230.
2. Aleman-Meza, B., Nagarajan, M., Ramakrishnan, C., Sheth, A., Arpinar, I., Ding, L., Kolari, P., Josi, A., and Finin, T. Semantic analytics on social networks: Experiences in addressing the problem of conflict of interest detection. *Proc. WWW '06*, ACM Press (2006), 407-416.
3. Bar-Yossef, Z., Guy, I., Lempel, R., Maarek, Y. S., and Soroka, V. Cluster ranking with an application to mining mailbox networks. *Proc. ICDM '06,* IEEE Computer Society (2006), 63-74.
4. Cai, D., Shao, Z., He, X., Yan, X., and Han, J. Mining hidden community in heterogeneous social networks. In *Proc. of the 3rd International Workshop on Link Discovery (LinkKDD 2005),*ACM Press (2005), 58 – 65.
5. Culotta, A., Bekkerman, R., and McCallum, A. Extracting social networks and contact information from email and the Web. *First Conference on Email and Anti-Spam (CEAS 2004)* (2004).
6. Facebook Developers – Documentation. http://developers.facebook.com/documentation.php .
7. Farrell, S., and Lau, T. Fringe Contacts: People-Tagging for the Enterprise. *Workshop on Collaborative Web Tagging, WWW'06,* (2006).
8. Farnham, S., Portnoy, W., Turski, A., Cheng, L., and Vronay, D. Personal Map: Automatically modeling the user's online social network. *Proc. INTERACT'03*, IOS Press (2003), 567-574.
9. Fielding, R.T. *Architectural styles and the design of network-based software architectures*. PhD thesis, University of California, Irvine, CA, (2000).
10. Friend of a Friend (FOAF) project. http://www.foaf-project.org/ .
11. HammerSley, B. *Content Syndication with RSS*, (2003).
12. Hope, T., Nishimura, T., and Takeda, H. An integrated method for social network extraction. *Proc. WWW '06,* ACM Press (2006), 845-846.
13. Introducing JSON. http://www.json.org/ .
14. Jackson, A., Yates, J., Orlikowski, W. "Corporate Blogging: Building community through persistent digital talk". *Proc. 40th Annual Hawaii International Conference on System Sciences HICSS'07*, (2007), p. 80
15. Kautz, H., Selman, B., and Shah., M. ReferralWeb: Combining social networks and collaborative filtering. *Communications of the ACM 40*, 3 (1997), 63-65.
16. Matsuo, Y., Hamasaki, M. et al. Spinning multiple social networks for semantic Web. *Proc. AAAI '06* (2006).
17. Matsuo, Y., Mori, J., Hamasaki, M., Takeda, H., Nishimura, T., Hasida, K., and Ishizuka, M. POLYPHONET: An advanced social network extraction system. *Proc. WWW '06*, ACM Press (2006), 397-406.
18. McDonald D.W. and Ackerman M.S. Expertise recommender: a flexible recommendation system and architecture. *Proc. CSCW'00,* (2000), 231–240.
19. Millen, D.R., Feinberg, J., and Kerr, B. Dogear: Social Bookmarking in the Enterprise. *Proc. CHI 2006,* (2006), 111-120.
20. Nardi, B.A., Whittaker, S., Issacs, E., Creech, M., Johnson, J., and Hainsworth, J. Integrating communication and information through contact map. *Communications of the ACM 45*, 4 (2002) 89-95.
21. Neustaedter, C., Brush, A., Smith, M., and Fisher, D. The social network and relationship finder: Social sorting for email triage. *Proc. CEAS 2005*.
22. Newman, M. E. J. Scientific collaboration networks, part I. Network construction and fundamental results. *Physical Review E, 64*, 016131, (2001).
23. Nottingham, M., and Sayre, R. RFC 4287 – The Atom Syndication Format (Proposed Standard). http://tools.ietf.org/html/rfc4287
24. O'Reilly, T. *What is Web 2.0*. http://www.oreillynet.com/go/web2 .
25. Resig, J., Dawara, S., Homan, C. M., and Teredesai, A. Extracting social networks from instant messaging populations, *Proc. of the 7th ACM SIGKDD Workshop on Link KDD*, (2004).
26. Smarr, J. Technical and privacy challenges for integrating FOAF into existing applications. In *the 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web*, (2004).
27. Smith, M. Invisible crowds in cyberspace: Measuring and mapping the social structure of Usenet. In Smith, M., and Kollock, P. Eds., *Communities in Cyberspace*. Routledge Press, (1999).
28. Tyler, J.R., Wilkinson, D.M., and Huberman, B.A. Email as spectroscopy: Automated discovery of community structure within organizations. In *Communities and Technologies*, Huysman, M., Wenger, E., and V. Wulf, Eds. (2003), 81-96.
29. Wasserman, S., and Faust, K. *Social Network Analysis*. (1994).
30. Wellman, B. Computer networks as social networks. *Science 293* (2001) 2031-2034.
31. Zhang, J., Ackerman M.S. Searching for expertise in social networks: a simulation of potential strategies. *Proc. GROUP 2005*, ACM Press (2005), 71-80.