Linking Social Networks on the Web with FOAF

Jennifer Golbeck
University of Maryland, College Park
2118F Hornbake Building
College Park, Maryland 20742
jgolbeck@umd.edu

Matthew Rothstein
University of Maryland, College Park
College Park, Maryland 20742
marothstein@gmail.com

ABSTRACT

One of the core goals of the Semantic Web is to store data in distributed locations, and use ontologies and reasoning to aggregate it. Social networking is a large movement on the web, and social networking data using the Friend of a Friend (FOAF) vocabulary makes up a significant portion of all data on the Semantic Web. Many traditional web-based social networks share their members' information in FOAF format. While this is by far the largest source of FOAF online, there is no information about whether the social network models from each network overlap to create a larger unified social network model, or whether they are simply isolated components. In this paper, we present a study of the intersection of FOAF data found in many online social networks. Using the semantics of the FOAF ontology and applying Semantic Web reasoning techniques, we show that a significant percentage of profiles can be merged from multiple networks. We present results on how this affects network structure and what it says about relationships and individual behavior. Finally, we discuss the implications this has for using web-based social networking data to create intelligent user interfaces and social software.

1. INTRODUCTION

One of the primary goals of the Semantic Web is to store data in distributed locations and to use ontologies and reasoning to aggregate and use it. Large team-engineered ontologies, or self contained applications are prominent examples of Semantic Web technologies, but these generally do not fully illustrate its potential. The missing component is a large set of instance data, distributed among many independent websites, where reasoning can be used to merge instances that would otherwise considered distinct.

The Friend of a Friend (FOAF) project is one of the largest projects on the Semantic Web. FOAF has become a widely accepted standard vocabulary for representing social networks, and many large social networking websites use it to produce Semantic Web profiles for their users. There are millions of FOAF profiles online, hosted at a wide range of websites. Because it is so successful in terms of use, FOAF is frequently used as an example of the success of the Semantic Web. The way it is used satisfies the goal of using an ontology to represent considerable amounts of distributed data in a standard form. However, for FOAF to truly serve as an example of the Semantic Web's full potential, reasoning over the data must lead to the discovery of connections between what are represented as distinct data sets. That means merging profiles of the same person from multiple social networking websites and creating a large, unified social network from subnetworks that evolved independently.

WWW2008, April 21–25, 2008, Beijing, China.

In addition to serving as an instantiation of Semantic Web vision, this FOAF-based profile merging is helpful to social network users. It is common for people to have accounts on several networks. If Semantic Web applications are built that use social networks (of which there are already several working examples), automated aggregation of a user's distributed social connections will give a fuller picture of their profile and improve the functioning of the applications.

In this paper, we present the first analysis of cross network linkages in FOAF. Using all of the accessible web-based social networks that generate FOAF profiles, we show the frequency of multiple profiles that a reasoner could merge, and describe the properties of those users. We found that 0.39% of users had accounts on multiple networks, serving as hubs that connected the social networks we studied. We also show that those users tend to connect to friends with multiple accounts more frequently. We conclude with a discussion of the implication of these results.

1.1 Web-Based Social Networks

Web-based social networks (WBSN) have grown quickly in number and scope since the mid-1990s. They present an interesting challenge to traditional ways of thinking about social networks. First, they are large, living examples of social networks. It has rarely, if ever, been possible to look at an actual network of millions of people without using models to fill in or simulate most of the network. The problem of gathering social information about a large group of people has been a difficult one. With WBSNs, there are many networks with millions of users that need no generated data. These networks are also much more complex with respect to the types of relationships they allow. Information qualifying and quantifying aspects of the social connection between people is common in these systems. This means there is a potential for much richer analysis of the network.

There are about 250 websites dedicated to social networking, i.e. they have explicit support for users to build and browse lists of friends. This includes websites like MySpace, Facebook, Orkut, and CyWorld but does not include many dating sites, like Match.com, and other online communities that connect users, such as Craig's List or MeetUp.com. The latter group of sites also contain social network information, but we do not consider them to be "Web-based Social Networks".

WBSNs have many purposes. We group them into the following general categories:

- Blogging
- Business
- Dating
- Pets
- Photos
- Religious
- Social/Entertainment

A list of all social networks we know of is maintained at http://trust.mindswap.org/. There is incredible diversity among the sites in all dimensions. The largest, MySpace, has over 150,000,000 members, while some sites have only a few dozen. They also have a range of expressivity about relationships between people. Some limit social connections to a basic friendship relationship while others provide many relationship types and options to carefully describe how people know each other.

There has been dramatic growth in the number and size of these networks. The number of sites almost doubled over the two year period from December 2004 to December 2006, growing from 125 to 223. Over the same period, the total number of members among all sites grew four-fold from 115 million to 490 million.

The size of individual networks ranges widely from a few dozen members to over 100 million members. In late 2006, the largest site (MySpace with more than 150 million members) is nearly an order of magnitude larger than the largest site in 2004 (Tickle with 18 million members). As would be expected with this kind of growth, the number of WBSNs with over a million members has increased sharply from eighteen in late 2004 to 41 in late 2006 [16].

2. FOAF SYNTAX AND SEMANTICS

Many people maintain accounts at multiple social networking websites. It is desirable, for example, to keep information intended for business networking separate from personal information.At the same time, users put significant effort into maintaining information on social networks. Multiple accounts are not just for compartmentalizing parts of their lives. A person may have one group of friends who prefer MySpace, another group on Facebook, and have an account on a religious website to stay connected to that community.

From the perspective of managing an entire set of social connections that are spread across sites, it is advantageous to merge all of those connections together into one set of data. In a merged social network, friends who have multiple accounts would be represented as a single person. Information about the user that is distributed across several sites also would be merged. The Friend-of-a-Friend (FOAF) Project is a potential solution to sharing social networking data among sites, and this section introduces how that is being done.

2.1 The Vocabulary

Rather than a website or a software package, FOAF is a framework for representing information about people and their social connections. Written in OWL, the FOAF Vocabulary contains terms for describing personal information, membership in groups, and social connections. Table 1 shows the full set of classes and properties in FOAF.

People are described as instances of the foaf:Person class. There are many properties to describe attributes of people, including name, email address, and documents they produce. The property foaf:knows is used to create social links between people (i.e. one person knows another person).

2.2 Reasoning with FOAF

FOAF utilizes the semantics of the Web Ontology Language OWL. While the overall idea - describe attributes of people - is straightforward, FOAF utilizes several features of OWL so interesting inferences can be made.

Inverse properties are used several times. In table 1, these are indicated parenthetically. This allows a reasoner to infer some bi-directional relationships between instances of FOAF classes.

For the work presented in this paper, the most important semantic features is the use of owl:InverseFunctionalProperty. An inverse functional property connects an instance to a unique identifier (e.g. a US citizen is uniquely identified by their social security number). Unique identifiers in FOAF are the following:

- foaf:aimChatID
- foaf:homepage
- foaf:icqChatID
- foaf:jabberID
- foaf:mbox
- foaf:mbox_sha1sum
- foaf:msnChatID
- foaf:weblog
- and foaf:yahooChatID

The above properties are used as unique identifiers because it is rare that two separate people will share the same email address, chat account, or blog address.

Any time two instances of foaf:Person have identical values for a property in the list above, an OWL reasoner will infer that the instances represent the same person. This is the critical inference used in merging profiles that represent the same person. Fortunately, all of the social networking websites that produce FOAF include at least one foaf:mbox_sha1sum for each user. This means that we can merge profiles from different networks based on this property.

In this research, we are interested *only* in profiles that can be merged by an OWL reasoner. While there are other techniques for finding duplicate profiles (see section 5 for a thorough treatment), our work is concerned with how standard web technologies can be applied to this problem. This approach illustrates the benefits provided by Semantic Web technologies. FOAF is interesting for representing social

Table 1: FOAF Classes (in initial capitals) and properties (lower case). Full details are available at http://xmlns.com/foaf/spec/

FOAF Basics	Personal Info	Online Accounts	Projects / Groups	Documents
Agent	weblog	OnlineAccount	Project	Document
Person	knows	OnlineChatAccount	Organization	Image
name	interest	OnlineEcommerceAccount	Group	PersonalProfileDocument
nick	currentProject	OnlineGamingAccount	member	topic (page)
title	pastProject	holdsAccount	membershipClass	primaryTopic
homepage	plan	accountServiceHomepage	fundedBy	tipjar
mbox	based_near	accountName	theme	sha1
mbox_sha1sum	workplaceHomepage	icqChatID		made (maker)
img	workInfoHomepage	msnChatID		thumbnail
depiction (depicts)	schoolHomepage	aimChatID		logo
surname	topic_interest	jabberID		
family_name	publications	yahooChatID		
givenname	geekcode			
firstName	myersBriggs			
	dnaChecksum			

networks primarily because it relies on OWL reasoning for merging profiles; if other techniques were used instead, the FOAF format could likely be abandoned for a much simpler representation.

When an OWL reasoner infers that two profiles represent the same person, the inference is always *logically* correct. However, it can be the case that the inference is incorrect in the real world. For example, two people may share an email address or a user may have a typo that makes their email the same as someone else's. This potential for error is possible with every automated system, and short of having a human personally interview each member to confirm they are, in fact, the same person, there is no way to be 100% accurate. We have intentionally chosen to ignore this problem. First, we agree with the FOAF creators that it is quite rare for two people to use a shared address as their address in online social networks. Secondly, in this work we are interested *only* in the logical inferences that allow us to merge profiles. Other techniques for the entity resolution problem are applicable here, and we discuss them in section 5.

3. DATA SOURCES AND METHODOLOGY

3.1 Data Sources

The goal of our work is to show how frequently user profiles from multiple social networks can be merged using the semantics of FOAF, and to understand the impact that has on the structure of the unified social network. While it is possible to get social relationships and personal information from networks that do *not* generate FOAF - by spidering or utilizing APIs - the scope of this work is to look *only* at FOAF files produced by the networks.

There are 11 active social networking websites that output FOAF files, with an approximate total of 13,120,000 members among them (see table 2). We used all of these networks in our research. Note that this is not just the total number of networks we used, but *all* the web-based social networks with available FOAF. LiveJournal is the largest of those, accounting for just over 75% of the total estimated membership, with approximately 10,000,000 users. We included all 11 of these websites in our survey.

For each network, we gathered as many profiles as possible. Some networks - FilmTrust, Ecademy, and Advogato - provided a full list of all of their members. In this case, we had access to all of the FOAF profiles, but this did not necessarily represent all the users. In particular, Advogato only produces FOAF for members of a certain rank. In the rest of the networks, a full list of members was not available, and we had to build a list of members by crawling the network. To do so, we chose several users as starting nodes and performed a breadth first search through the network to find all reachable members. For each user, we accessed the FOAF file, pulled URIs of their friends' FOAF files, and added those URIs to our queue. While we tried to identify the giant component of each network, there are almost certainly smaller components that our crawls did not reach. Also, users with no social connections would never be discovered on a crawl. Table 2 shows the membership of each network that we were able to use in our study. While this is not the total membership of every network, we believe that this serves as an accurate sample to illustrate inter-network connectivity. Furthermore, any applications using FOAF would need to follow the same procedures we did in this study, and thus our data set is representative of what FOAF applications would have to work with.

Earlier work [16] suggests the fraction of singleton users in the blogging websites might be very high; if users join to blog, the social network is secondary and may go unnoticed or unused. On one hand, missing these users is less significant because they do not have social connections and, thus, their profiles would not add any connections to the integrated social network. On the other hand, it is possible that a profile with no connections could contain the properties required to merge two profiles from other social networks that might otherwise be missed. However, none of these networks allowed users to have multiple email addresses, so the only way we could do this sort of merge would be through importing FOAF data from sources outside our consideration in this paper.

Six of the eleven websites on this list are blogging websites based on the open-source LiveJournal code. FOAF output is built into LiveJournal, so it is automatically produced when

Table 2: The social networks used in this study.

Network	Purpose	Members Studied	Avg. Degree
Advogato	Business	2,778	13.51
Buzznet	Photos	208,324	1.00
DeadJournal	Blogging	9,801	3.74
eCademy	Business	61,242	3.08
FilmTrust	Social/Entertainment	1,250	1.06
GreatestJournal	Blogging	36,862	33.36
InsaneJournal	Blogging	1,410	13.36
LiveJournal	Blogging	3,563,267	8.38
Minilog.com	Blogging	119	1.63
Rossia.org	Blogging	4,180	9.65
Tribe	Social/Entertainment	218,694	9.93

a website implements it. As such, blogging accounts for a disproportionate percentage of our data. Overall, blogging websites account for 19 of the 226 (or 8.4%) known social networks, and only 2.7% of the total membership. In this study, six of the 11 websites are for blogging (54.5%), and they make up 23.1% of the membership we studied. It is reasonable to think that social networking behavior on blogging websites may be quite different from behavior on "pure" social networking sites with no external purposes. That insight is supported by results in [16]. Thus, if FOAF were available on a more representative set of social networking websites, the results of a study like this may be different.

3.2 Methodology

For every member we were able to include in the study, we accessed their FOAF file. For the purpose of this work, we were interested only in the member's friends and unique identifiers (given by the inverse functional properties mentioned in section 2.2). Thus, to save space and increase efficiency, we implemented a task-specific OWL reasoner that considers only the FOAF inverse functional properties and foaf:knows property, and ignores the rest of the data.

Traditionally, a reasoner would not keep track of the sources of each axiom in the knowledge base. Since we are specifically interested in how data is repeated in multiple sources, we added a provenance tracking feature to our reasoner. This maintains a record of the document where each axiom is asserted. With this data available, it is straightforward to identify on which and how many social networks a member has accounts, as well as the sources for each friendship.

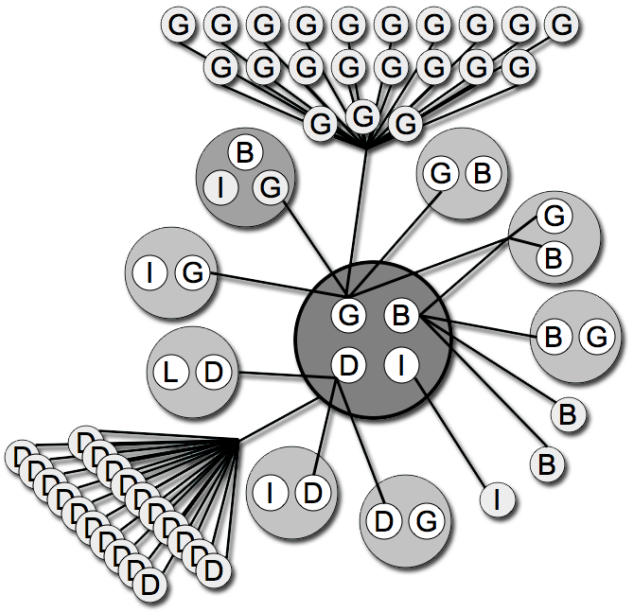

Figure 1: An egocentric network built around an individual found in our study with accounts on four WBSNs. The node labels indicate the first letter of the domain name of the WBSN.

4. RESULTS

After aggregating and reasoning over the FOAF data, we were able to see connections between the different networks, and to analyze how the reasoning connected accounts and affected friendships.

4.1 Network Statistics

After reasoning over all the FOAF data, the distinct networks generated by each social networking website were connected when a member on multiple sites was identified as the same person. This happens when a foaf:Person is found in both networks with the same value for one of the inverse functional properties mentioned in 2.2.

Table 3 shows the networks that we were able to connect directly because they had a member in common. While every network was not directly connected to every other, every network had connections to at least four others. No network was isolated and thus the unified social network had paths connecting every network to every other. Note that LiveJournal, the largest network in this study, had members with accounts on every other network we studied.

As an example of networks are linked through users with accounts on multiple websites, consider the user shown in Figure 1. This depicts an egocentric network around one user who has accounts on four different social networking websites: Buzznet, DeadJournal, GreatestJournal, and InsaneJournal. This user had one friend with accounts on three of these networks, seven friends with accounts on two networks, and the remaining friends had accounts on only one network. We can also see that the central user is has relationships in both Buzznet and GreatestJournal with one of these friends who has two accounts.

Reasoning over the FOAF allowed us to perform analyses

Table 3: Networks linked through common members

Networks	Advogato	Buzznet	DeadJournal	eCademy	FilmTrust	GreatestJournal	InsaneJournal	LiveJournal	Rossia.org	Minilog.com	Tribe
Advogato	x	2	1	1	6			58	1		53
Buzznet		x	53	89	13	929	75	1967	5		793
DeadJournal			x			85	19	387			28
eCademy				x	8	1		22	1		161
FilmTrust					x			8			17
GreatestJournal						x	320	702	16	4	15
InsaneJournal							x	32	5	1	
LiveJournal								x	208	10	2357
Rossia.org									x		8
Minilog.com										x	3
Tribe											x

beyond points of connection between networks. By merging profiles that shared email addresses, the graph within each subnetwork changed. It was common to find many accounts sharing the same address within one website. For example, on the eCademy website, 991 users had at least two accounts with the same email address. Some people had *many* accounts with the same address; we found 38 email addresses that were each shared by five accounts. Thus, after reasoning, the network for this website would look different. Some paths would become shorter because merged nodes lead to fewer steps between people. The average path length, however, can grow *or* shrink. Certainly some paths will be shorter. However, before merging, accounts that shared addresses could be closely connected. Some were directly connected (i.e. a user makes all of his or her accounts friends with each other), and in other cases, the accounts had friends in common. When these clusters disappear as nodes are merged, many short paths disappear, which can lead to an overall longer average path length for the network. In most networks, the average shortest path was not significantly affected by reasoning. Buzznet was an exception, with the average shortest path length dropping from 4.43 in the unreasoned network to 2.76 in the reasoned network.

To compute the average shortest path length for the unified network, we selected 110 random users from each network as sources. We then selected another 110 random users pairs from each network to serve as sinks. The 110 sources from a given network were paired with 10 sinks from each network. This ensured that we used source-sink pairs that were spread throughout the network to compute the average. With this method, the average shortest path in the merged network is 3.56. However, since the networks vary widely in size and thus proportion of the population, this method is not representative of true average paths in the network. LiveJournal usres dominate the population accounting for 81% of all users. Using source,sink pairs chosen completely at random from the unified network, we can find a true average which will frequently consider users who are both members of LiveJournal. With this random approach, we found the average shortest path length was 2.94 - slightly higher than the average for LiveJournal, and close to the average for all networks considered.

Table 4: The average shortest path length (APL) in each WBSN, pre-reasoning and post-reasoning.

Network	APL (Pre)	APL (Post)
Advogato	2.17	2.15
Buzznet	4.43	2.76
DeadJournal	3.19	3.23
eCademy	2.20	2.19
FilmTrust	3.75	3.84
GreatestJournal	2.25	2.31
InsaneJournal	3.19	3.26
LiveJournal	2.85	2.83
Minilog.com	3.66	3.66
Rossia.org	2.33	2.36
Tribe	2.74	2.69
Average	*2.97*	*2.84*

4.2 Account Statistics

Our results show that 8,047 of unique people we found (approximately 0.2%) had accounts on multiple networks. While the number of members who have accounts on multiple networks is a small percentage of everyone we found, it is typical of patterns identified in social networks. The logarithmic distribution shown in figure 2 is frequently found in social networks. A small percentage of nodes serve as hubs with high centrality that connect otherwise distinct parts of the network. While most work looks at hubs connecting communities within a single network, these hubs perform the same function by connecting different social networks in the unified FOAF network.

There are some attributes of the data we collected that should be mentioned at this point. Of the 8,047 people with accounts on multiple social networking websites, the vast majority, 7,849 (97.5%), had accounts on only two websites. Of those, 5,473 (69.7%) had one of their accounts on LiveJournal, and one account on another network. This raises an interesting point about LiveJournal: it is the only one

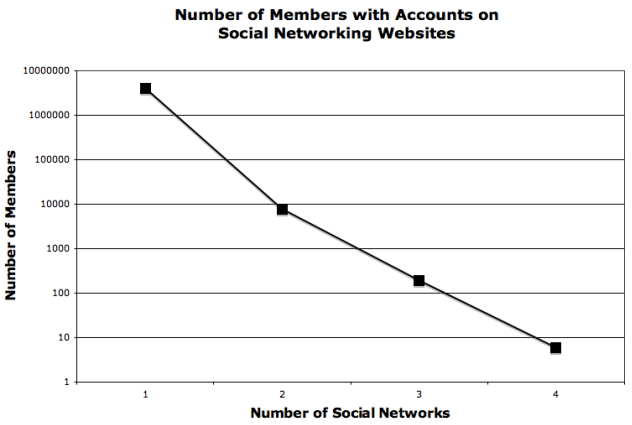

Figure 2: This chart shows the number of people with accounts on a given number of networks. Note that the y-axis is a logarithmic scale.

of the eleven networks we looked at that did *not* require users to enter an email address. In fact, only 8.8% of the LiveJournal users we found in this study had a foaf:mbox or foaf:mbox_sha1sum. This meant that it was impossible to link these accounts to any other, since every other network used the mbox_sha1sum as a unique identifier. If the LiveJournal members with email addresses are representative, we can extrapolate that just over 62,000 LiveJournal members in the population we found have accounts on other networks, which would lead to a much higher network interlinking rate with 1.5% of all users on at least two networks. A small change by LiveJournal requiring an email address so that their users' FOAF could be linked to other FOAF would make a big difference in the connections between networks, and ultimately toward taking advantage of what the Semantic Web has to offer.

There were 198 users with accounts on more than two networks, and their activity was largely centered around three WBSNs. Seven members have accounts in four networks; this is the maximum number of WBSN memberships. The remaining 191 members have accounts on three networks. The impact of LiveJournal was even similar here. Of these 198 users, 136 (68.7%) had accounts on LiveJournal. However, it was not the dominant network here. Even more users (157 or 79.3%) had accounts on Buzznet, and 129 (65.2%) users had accounts on GreatestJournal. In fact, there were no users who did not have an account on at least one of these three networks, and only three users who did not have either a Buzznet or LiveJournal account.

4.3 Friendship Statistics

Members who had accounts on multiple networks serve as hubs in our unified social network. Traditionally, hubs in social networks have more friends than average. That turned out to be the case for our network bridging members as well.

Users who have multiple accounts also tend to have more friends with multiple accounts. On average, friends of people with one account had 1.01 accounts, while members with accounts on multiple networks had an average of 1.15 friends. This difference is significant for $p < 0.05$ using a standard two-tailed t-test. To look at more specific numbers, friends of people with two accounts had an average of 1.15 accounts, and friends of people with three accounts had an average of 1.17 accounts. An ANOVA shows a significant difference within the population, and a standard 2-tailed t-test shows that friends of people with two accounts had significantly more accounts than friends of people with one. The same also holds true for friends of people with three accounts vs. friends of people with two accounts (for $p < 0.05$).

When both people in a pair of friends had multiple accounts, they were frequently friends on multiple networks. We found 15.71% of these members were friends in more than one network. On average, they were friends in 57.75% of the networks where they were both members.

These results show that a small percentage of users have multiple accounts, but they tend to be well connected with friends who also have multiple accounts. This core group is sufficient to serve as a bridge between multiple social networks and act as hubs in the aggregated FOAF network.

5. IMPLICATIONS AND DISCUSSION

Given a large unified FOAF social network where we have been able to logically merge profiles that represent the same person on different networks, opportunities for further analysis and applications become available.

5.1 Analysis

First, there are other techniques for extracting social relationship information on the web besides relying on FOAF data. For example, the HTML presentation of social networking websites can be parsed to generate personal information and social relationships. Scrapers can pull information from alternative forms of output. Some networks, like Facebook and Flickr, have APIs that grant a limited level of access to the network information. All of these methods can be used to access network data, and even to produce FOAF if that is desired. Flink [31] is a system that uses some of these approaches and others to extract, analyze (including merging), and visualize social networks on the web. The Flink-type is completely compatible with this work, and if unified social networks are to be used in applications, it will likely be necessary to extract data since FOAF availability is limited.

We have chosen to focus our work on the application of Semantic Web reasoning techniques to the Semantic Web data gathered through FOAF. This problem of merging profiles that represent the same person was done by relying on a reasoner that could handle inverse functional properties. This problem of *entity resolution* (also referred to in the literature as deduplication, object uncertainty, record linkage, and others) has been addressed extensively in the data mining community and can be handled in much more advanced ways. Traditionally, methods look at similarity in the text that describes entities to make decisions about merging (including [32, 9, 6] among many others). Some text is available from social networking websites in FOAF format; names, nicknames, and occasionally other personal information. Social relationships are always available, and entity resolution techniques that use link structure may also be applicable. These algorithms rely on relational structure [5, 3, 22] and provide a relatively computationally efficient approach to the problem. Because these techniques rely on link structure, it is critical that a first pass will have merged

people to create links between the sub-networks generated by different websites. We have shown that these cross network linkages are found in percentages expected from hubs in social networks, and this may be a suitable foundation for applying relationship-based entity resolution algorithms. One area of future work is to apply these methods to our unified network to evaluate their performance.

Similar work in link mining [13] or link prediction [7] is relevant and could be applied. While entity resolution addresses the problem of finding nodes that represent the same individual, link mining is focused on relationships in the network, including inferring the existence of links (link prediction) and link-based cluster analysis. These techniques could be used to add edges between nodes who are likely to be friends based on their other connections and properties.

5.2 Applications

A unified FOAF network can be of use to applications designed around FOAF and others that integrate social networks more generally.

Recommender systems have been a space where FOAF has been applied frequently. For example, Moleskiing [4], at http://moleskiing.it, uses FOAF as the basis for making recommendations about mountaineering ski trails in a community forum.. The subject of the website is ski mountaineering and strives to make the activity safer by collecting information from users about the conditions and quality of ski trails. Moleskiing separates information into ski routes, which are relatively static and entered by experts, and ski trips, which are dynamic comments entered by users. Ski trip information is maintained on Moleskiing-hosted blogs.Users have lists of their friends, maintained in a FOAF file, and the system can import FOAF from outside sites. The system uses the social network to compute the trust the user has in other people in the network, and the trust is used to recommend and rank ski routes.

Foafing the Music [33] is a music recommender system that uses social networks built with FOAF and other Semantic Web data to feed music information to users. The system does not store or produce FOAF files itself, but rather relies on gathering it from locations distributed across the web. User's FOAF profiles are used to determine their interests and find music that matches their tastes.

SocialBrowsing [19] is a Firefox plugin that uses social networking information, including a unified FOAF network, to add contextual information to websites as users browse. It is designed to pull social networking information from any source, with a specific emphasis on FOAF files, and adds highlights into the text or in the browser's status bar to indicate socially relevant information. This could include recommendations or ratings of the content mentioned on the page, or general information about the page, such as when the user is looking at the blog of a friend or friend of a friend.

There are recommender systems that consider the use of social networks more generally, and they could be implemented with the large, unified FOAF network as their social data source. One of the earlier descriptions of social network-based recommender systems is ReferralWeb [24]. The idea has been used for recommending collaborations [29], social connections [35, 28], and citations [30], as well as for collaborative filtering in general [26].

Email filtering is another subject where social networks can be used. Boykin and Roychowdhury [8] create a social network from the messages that a user has received. Using the structural properties of social networks, particularly the propensity for local clustering, messages are identified as spam, valid, or unknown based on clustering thresholds. Their method is able to classify about 50% of a users email into the spam or valid categories, leaving 50% to be filtered by other techniques. Extending this approach to utilize social networks drawn from online communities would bring more users into consideration, and also open up the opportunity to look for connections over longer paths. A more complex version of this approach is used in TrustMail[17]. TrustMail is a prototype email client that adds trust ratings, computed from a social network, to the folder views of a message. This allows a user to see their trust rating for each individual, and sort messages accordingly. It specifically cites FOAF networks as a potential data source.

Another interesting application of FOAF has been for detecting conflicts of interest [2]. When assigning reviewers to scientific papers, reviewers have to self report potential conflicts. For many people, this is potentially a long list. The authors present a technique for using co-authorship from DBLP and the FOAF knows relationship to automatically identify conflicts of interest, and describe how their work is applicable more generally to Semantic Web engineering problems. Access to a larger, more integrated FOAF network would improve the quality of these applications, and our results show that accessing FOAF that is automatically generated by WBSNs will be linked together after reasoning.

Social relationships, particularly trust, have been used for prioritizing and filtering within Semantic Web back-end applications as well. [23] presents a mechanism for using social relationships to prioritizing rules in default logics. [18] uses trust to rank statements in knowledge bases. [21] uses social trust to in web-based syndication systems to resolve inconsistencies that arise in knowledge bases as new publications are received.

6. RELATED WORK

Social network analysis is the study of the properties of the structure, relationships, and people in a social network. Social networks based on real world connections have been studied extensively. These studies have relied on data gathered by surveying people, studying family trees and historical documents, or extrapolating from observable behavior. Even previous work examining *online* social networks [12] recommends a survey-based approach for extracting social information about users. Similarly, the growth and activity patterns, design, and behavior [34] in online communities is the subject of a vast literature. In this work, however, we study the explicitly stated social connections, rather than social interactions of users. While that differentiates our approach from the body of work on online communities, many suggestions we present in the conclusions echo suggestions for the design of these communities in existing work [34].

The web has opened up new opportunities for social network analysis because people are providing information about themselves and their social connections in publicly accessible forums. Web-based social networking websites are growing quickly in number and size [16]. This has spawned a new set of literature studying behavior and structure of websites like Facebook [27], Cyworld [1], YouTube [15], and MySpace [14].

FOAF and the websites used in this study have been ad-

dresses previously in the literature as well. [25] looks at structural patterns of web-based social networks, including an analysis of LiveJournal, a network we use in this study. The authors are primarily interested in predicting the participation of users in different communities within the social networks, a topic not addressed in this article. We believe that their work fits well with the results we present, describing finer-grained internal behavior within web-based social networks, one level more specific than we analyze here.

In [11], the authors presented a survey of how FOAF was being used online. Their interests were primarily in which parts of the vocabulary were utilized, and they presented some basic statistics on the structural features of the network. The structural analysis, however, explicitly excluded FOAF generated from blogging websites which are responsible for the vast majority of FOAF documents on the web.

[20] uses learning techniques with FOAF data to infer characteristics of people in the network. The authors used a small set of approximately 9,000 people with profiles and generated a set of rules for adding properties to users found to be in a set of clusters. Their work is similar in spirit to a simple version of the link mining described above.

7. CONCLUSIONS

FOAF is one of the most popular and widely discussed uses of Semantic Web technologies. Work is appearing that discusses the possibility of using a FOAF social network as a backend. Large web-based social networks are also starting to share some of their members information and social connections in FOAF format, making millions of profiles available. However, up to this point, no work has shown to what extent users are making connections *between* those social networks.

We gathered FOAF profiles from a number of social networks with over 4 million total users. Using a customized Semantic Web reasoner, we have shown that thousands of users have accounts on multiple WBSNs, linking their subgraphs in the unified social network. This means that large collections of automatically generated FOAF contribute to a connected, distributed social network that can feed into a variety of applications. We also present results on the impact these links have on the structural properties of the unified network, and relationship patterns.

7.1 Data Challenges

Due to the nature of web spiders, our data collection was limited to the connected components of each graph of which the seed users were part. If the websiteshad provided full user lists, acquiring complete data would have been trivial. Unfortunately, many users would consider publishing data of this kind to be a significant violation of privacy. A practical example is the group of users who participate in both business and socially oriented social networking sites. Many of these users would probably prefer that their accounts remain unassociated.

We only collected data from networks that publish FOAF data for their users. This meant excluding many social networking web sites, including some of the most popular sites like MySpace or Facebook. The addition of popular sites like these would provide a substantially larger pool of users, most likely magnifying our current results.

FOAF equipped networks on the web today are primarily comprised of blogging sites, which tend to have less social activity than more pure social networking sites such as MySpace. We expect that the data we collected may be an underestimate of what happens in such purer social networks. In the future, data could be collected from web sites with a more diverse focus, to ensure a more accurate depiction of cross-network connectivity.

7.2 Future Work

One of the biggest challenges to working with a large, integrated FOAF network is scalability. Running a single breadth first search to compute a shortest path between two people in a network with several million nodes will exceed the memory capacity of most desktop computers and small servers. We were forced to shift our analysis up to larger clusters for this experiment, and we were working with less than 4 million nodes. If we had access to data for even tens of millions of nodes, let alone the hundreds of millions available, it would be very difficult to process. Aside from memory requirements, the complexity of many analysis algorithms makes handling tens of millions of nodes exceptionally difficult.

Any applications that will utilize a large, unified FOAF social network will need to access it in a way that can handle the size of the data and the complexity of the computations. We are in the early stages of collaboration using a cloud computing paradigm with Hadoop[10] to address problems of scalability and analysis of these vast amounts of data.

8. ACKNOWLEDGMENTS

This work, conducted at the College of Information Studies and MINDSWAP group was funded by Fujitsu Laboratories of America – College Park, Lockheed Martin Advanced Technology Laboratory, NTT Corp., Kevric Corp., SAIC, the National Science Foundation, the National Geospatial-Intelligence Agency, DARPA, US Army Research Laboratory, NIST, and other DoD sources.

9. REFERENCES

[1] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 835–844, New York, NY, USA, 2007. ACM.

[2] B. Aleman-Meza, M. Nagarajan, C. Ramakrishnan, L. Ding, P. Kolari, A. P. Sheth, I. B. Arpinar, A. Joshi, and T. Finin. Semantic analytics on social networks: experiences in addressing the problem of conflict of interest detection. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 407–416, New York, NY, USA, 2006. ACM.

[3] R. Ananthakrishna, S. Chaudhuri, and V. Ganti. Eliminating fuzzy duplicates in data warehouses. In *Proceedings of the 28th International Conference on Very Large Databases (VLDB 2002)*, 2002.

[4] P. Avesani, P. Massa, and R. Tiella. Moleskiing.it: a trust-aware recommender system for ski mountaineering. *International Journal for Infonomics*, 2005.

[5] I. Bhattacharya and L. Getoor. Iterative record linkage for cleaning and integration. In *ACM*

SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD), 2004.

[6] M. Bilenko and R. J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 39–48, New York, NY, USA, 2003. ACM Press.

[7] M. Bilgic, G. M. Namata, and L. Getoor. Combining collective classification and link prediction. In *Workshop on Mining Graphs and Complex Structures at the IEEE International Conference on Data Mining (ICDM-2007)*, 2007.

[8] P. O. Boykin and V. Roychowdhury. Personal email networks: An effective anti-spam tool. *IEEE Computer*, 38:61, 2004.

[9] S. Chaudhuri, K. Ganjam, V. Ganti, and R. Motwani. Robust and efficient fuzzy match for online data cleaning. In *SIGMOD '03: Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 313–324, New York, NY, USA, 2003. ACM Press.

[10] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. In *Proceedings of the 6th Symposium on Operating System Design and Implementation (OSDI)*, page 137150, 2004.

[11] L. Ding, L. Zhou, T. Finin, and A. Joshi. How the semantic web is being used: An analysis of foaf documents. In *HICSS '05: Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 4*, Washington, DC, USA, 2005. IEEE Computer Society.

[12] L. Garton, C. Haythornthwaite, and B. Wellman. Studying online social networks. *Journal of Computer Mediated Communication*, 3, 1997.

[13] L. Getoor. Link mining: A new data mining challenge. *SIGKDD Explorations, volume*, 5(1):85– –89, 2003.

[14] R. Gibson. Who's really in your top 8: network security in the age of social networking. In *SIGUCCS '07: Proceedings of the 35th annual ACM SIGUCCS conference on User services*, pages 131–134, New York, NY, USA, 2007. ACM.

[15] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube traffic characterization: a view from the edge. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 15–28, New York, NY, USA, 2007. ACM.

[16] J. Golbeck. The dynamics of web-based social networks: Membership, relationships, and change. *First Monday*, 12(11), 2007.

[17] J. Golbeck and J. Hendler. Reputation network analysis for email filtering. In *Proceedings of the First Conference on Email and Anti-Spam*, 2004.

[18] J. Golbeck and B. Parsia. Trust network-based filtering of aggregated claims. *International Journal of Metadata, Semantics, and Ontologies*, 1(1):58–65, 2006.

[19] J. Golbeck and M. M. Wasser. Socialbrowsing: integrating social networks and web browsing. In *CHI '07: CHI '07 extended abstracts on Human factors in computing systems*, pages 2381–2386, New York, NY, USA, 2007. ACM Press.

[20] G. A. Grimnes, P. Edwards, and A. Preece. Learning meta-descriptions of the foaf network. In *Proceedings of the International Semantic Web Conference*, 2004.

[21] F. C. Halaschek-Wiener. *Expressive Syndication on the Web Using a Description Logic Approach*. PhD thesis, University of Maryland, College Park, MD, USA, November 2007.

[22] D. V. Kalashnikov, S. Mehrotra, and Z. Chen. Exploiting relationships for domain-independent data cleaning. In *SIAM International Conference on Data Mining (SIAM SDM)*, Newport Beach, CA, USA, April 21–23 2005.

[23] Y. Katz and J. Golbeck. Social network-based trust in prioritized default logic. *Proceedings of the Twenty-First National Conference onArtificial Intelligence (AAAI-06)*, 2006.

[24] H. Kautz, B. Selman, and M. Shah. Referral Web: Combining social networks and collaborative filtering. *Communications of the ACM*, 40(3):63–65, 1997.

[25] R. Kumar, J. Novak, and A. Tomkins. Group formation in large social networks: membership, growth, and evolution. In *KDD 0́6: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006.

[26] C. Lam. Snack: incorporating social network information in automated collaborative filtering. In *EC '04: Proceedings of the 5th ACM conference on Electronic commerce*, pages 254–255, New York, NY, USA, 2004. ACM Press.

[27] C. Lampe, N. Ellison, and C. Steinfield. A face(book) in the crowd: social searching vs. social browsing. In *CSCW '06: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 167–170, New York, NY, USA, 2006. ACM.

[28] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 556–559, New York, NY, USA, 2003. ACM Press.

[29] D. W. McDonald. Recommending collaboration with social networks: a comparative evaluation. In *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 593–600, New York, NY, USA, 2003. ACM Press.

[30] S. M. McNee, I. Albert, D. Cosley, P. Gopalkrishnan, S. K. Lam, A. M. Rashid, J. A. Konstan, and J. Riedl. On the recommending of citations for research papers. In *CSCW '02: Proceedings of the 2002 ACM conference on Computer supported cooperative work*, pages 116–125, New York, NY, USA, 2002. ACM Press.

[31] P. Mika. Flink: Semantic web technology for the extraction and analysis of social networks. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2-3):211–223, October 2005.

[32] A. E. Monge and C. Elkan. The field matching problem: Algorithms and applications. In *Knowledge Discovery and Data Mining*, pages 267–270, 1996.

[33] Òscar Celma. Foafing the music: Bridging the semantic gap in music recommendation. In *Proceedings of the International Semantic Web Conference*, volume 4273 of *LNCS*, pages 927–934. Springer, 2006.

[34] J. Preece. *Online Communities: Designing Usability and Supporting Socialbilty*. John Wiley & Sons, Inc., New York, NY, USA, 2000.

[35] L. Terveen and D. W. McDonald. Social matching: A framework and research agenda. *ACM Trans. Comput.-Hum. Interact.*, 12(3):401–434, 2005.