Linking Social Networks on the Web with FOAF: A Semantic Web Case Study

Jennifer Golbeck
College of Information Studies
University of Maryland, College Park
College Park, MD 20742, USA
jgolbeck@umd.edu

Matthew Rothstein
University of Maryland, College Park
College Park, MD 20742, USA
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, or whether they are simply isolated components. If there are intersections, it is evidence that Semantic Web representations and technologies are being used to create interesting, useful data models. 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 the success of the Semantic Web.

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 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.

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. The purpose of this study is to use social networks, a large source of data, to understand how successfully Semantic Web technologies are living up to the vision of representing and reasoning over distributed data. We found that over 16,000 accounts could be logically merged, 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.

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. In this section, we present the important details of the FOAF vocabulary and discuss how they facilitate this type of merging.

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, FOAF contains terms for describing personal information, membership in groups, and social connections.

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[1]. The property foaf:knows is used to create social links between people (i.e. one person knows another person).

Reasoning with FOAF

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

Inverse roles are used several times.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 feature is the use of owl:InverseFunctionalProperty. An inverse functional property connects an instance to a unique identifier. 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. These properties are used as unique identifiers because it is rare that two people will share the same email address, chat account, or blog address.

Any time two instances of foaf:Person have identical values for an inverse functional property, 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. In this research, we are interested *only* in merging profiles in this way. While there are other techniques for finding duplicate profiles (see the discussion below for a thorough treatment), our work is directed toward using this problem to understand the state of the Semantic Web.

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.

Data Sources and Methodology

Data Sources

We set out to understand 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, 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 1). We used all of these networks in our research. 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 the rest of the networks, a full list of members was not available, and thus we had to crawl the network. Table 1 shows the number of users in each network that we were able to use in this 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 use.

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 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. Evidence siggests that social networking behavior on blogging websites may be quite different from behavior on "pure" social networking sites with no external purposes (Golbeck 2007). Thus, if FOAF were available on a more representative set of social networking websites, the results of a study like this may be different.

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). 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.

[1] See http://foaf-project.org for the full vocabulary

Table 1: The social networks used in this study, including the average shortest path length pre and post-reasoning.

Network	Purpose	Members Studied	Avg. Degree	APL (Pre)	APL(Post)
Advogato	Business	2,778	13.51	2.17	2.15
Buzznet	Photos	208,324	1.00	4.43	2.76
DeadJournal	Blogging	9,801	3.74	3.19	3.23
eCademy	Business	61,242	3.08	2.20	2.19
FilmTrust	Social/Entertainment	1,250	1.06	3.75	3.84
GreatestJournal	Blogging	36,862	33.36	2.25	2.31
InsaneJournal	Blogging	1,410	13.36	3.19	3.26
LiveJournal	Blogging	3,563,267	8.38	2.85	2.83
Minilog.com	Blogging	119	1.63	3.66	3.66
Rossia.org	Blogging	4,180	9.65	2.33	2.36
Tribe	Social/Entertainment	218,694	9.93	2.74	2.69

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.

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 above.

Table 2 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 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. Thus, after reasoning, the network structure would change. This can be seen in the change in the average shortest path length pre and post-reasoning as shown in table 1.

Account Statistics

Our results show that 8,047 of unique people we found had accounts on multiple networks, with a total of 16,295 accounts among them. 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. 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.

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 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. It is not possible to link accounts with no mbox, 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.

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 friends of

Table 2: This table shows the number of members shared by each pair of networks.

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

members with accounts on multiple networks had an average of 1.15 accounts. 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. 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.

Implications, Discussion, and Related Work

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.

Network Analysis

First, there are other techniques for extracting social relationship information on the web besides relying on FOAF data. Parsers, scrapers, and APIs can be used to access network data, and even to produce FOAF if that is desired. Flink (Mika 2005) 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 approach is completely compatible with this work, and if uni-

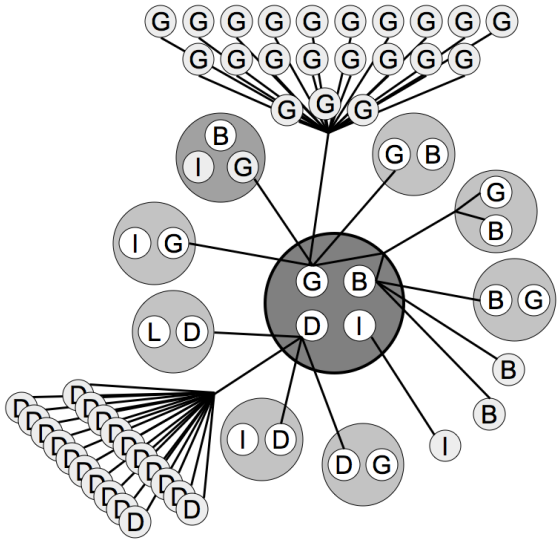

Figure 1: An egocentric network built around an individual found in our study with accounts on four WBSNs. The node labels indicate the first letter of the domain name of the WBSN.

fied social networks are to be used in applications, it will likely be necessary to use some of these techniques.

We have chosen to focus our work on the application of Semantic Web reasoning techniques for merging profiles that represent the same person. 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 (Chaudhuri *et al.* 2003; Bilenko & Mooney 2003) 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 (Bhattacharya & Getoor 2004; Kalashnikov, Mehrotra, & Chen 2005) 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.

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 (Avesani, Massa, & Tiella 2005), 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. Foafing the Music (Òscar Celma 2006) 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. FilmTrust (Golbeck 2006) also uses social networks to generate movie recommendations for users.

There are recommender systems that consider the use of social networks more generally, and they could be implemented with FOAF network as their social data source. One of the earlier descriptions of social network-based recommender systems is ReferralWeb (Kautz, Selman, & Shah 1997). The idea has been used for recommending collaborations (McDonald 2003), social connections (Terveen & McDonald 2005; Liben-Nowell & Kleinberg 2003), and citations (McNee *et al.* 2002), as well as for collaborative filtering in general (Lam 2004).

Email filtering is another subject where social networks can be used. Both (Boykin & Roychowdhury 2004) and TrustMail (Golbeck & Hendler 2004) use social networks to filter messages. A unified FOAF network, where most users are identified by email address, would provide a much richer set of social information.

Another interesting application of FOAF has been for detecting conflicts of interest (Aleman-Meza *et al.* 2006). 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.

FOAF

In (Ding *et al.* 2005), 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.

(Grimnes, Edwards, & Preece 2004) 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.

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 for applications. 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. This shows that some of the goals of the the Semantic Web are being realized in the social networking space.

Future work in this space must consider a common Semantic Web concern, particularly with social networks: privacy. The website Plink was set up to display public FOAF data, but was forced to shut down because so many people were upset at seeing this data displayed. This will certainly

continue to be a concern, especially when multiple social network profiles for a person can be merged to show even more data.

References

Aleman-Meza, B.; Nagarajan, M.; Ramakrishnan, C.; Ding, L.; Kolari, P.; Sheth, A. P.; Arpinar, I. B.; Joshi, A.; and Finin, T. 2006. 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*, 407–416. New York, NY, USA: ACM.

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

Bhattacharya, I., and Getoor, L. 2004. Iterative record linkage for cleaning and integration. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD)*.

Bilenko, M., and Mooney, R. J. 2003. 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*, 39–48. New York, NY, USA: ACM Press.

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

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

Ding, L.; Zhou, L.; Finin, T.; and Joshi, A. 2005. 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: IEEE Computer Society.

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

Golbeck, J. 2006. Generating predictive movie recommendations from trust in social networks. In *Proceedings of the Fourth International Conference on Trust Management*.

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

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

Kalashnikov, D. V.; Mehrotra, S.; and Chen, Z. 2005. Exploiting relationships for domain-independent data cleaning. In *SIAM International Conference on Data Mining (SIAM SDM)*.

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

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

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

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

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

Mika, P. 2005. 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.

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

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