Chapter 14

Social Network Analysis on the Semantic Web: Techniques and Challenges for Visualizing FOAF

John C. Paolillo and Elijah Wright

14.1 Introduction

The Semantic Web promises to provide new applications for Internet users through the use of RDF metadata attached to various information resources on the web. Yet is somewhat unclear who will provide the metadata, or what will motivate people to provide it, let alone the exact nature of the applications the Semantic Web will ultimately support. What will the "killer app" of the Semantic Web be, and what shape will it take? An answer to this question may have already arisen, in the form of the Friend-of-a-Friend (FOAF) vocabulary. The FOAF project was begun in 1999 to explore the application of Semantic Web technologies (RDF/XML) to describing people's personal details: their professional and personal lives, their friends, interests and other social dispositions. Its main product is the FOAF vocabulary, an RDF/XML namespace with elements defined for describing an individual's social sphere (Brickley and Miller, 2003).

Recently the FOAF vocabulary has been adopted by many large web-logging ("blogging") and social networking software sites, such as LiveDoor, LiveJournal, and others. Weblogs, a recent Internet phenomenon, are diary-like sites usually consisting of entries in reverse-chronological order. Herring, et al. (2004) situate weblogs as a genre bridging between extant media technologies and new forms of computer-mediated communication. The contribution of weblogs to the Semantic Web comes from the design of the supporting software to automatically generate RDF/XML files including RSS feeds, and now FOAF.

The popularity of weblog hosting sites, and their ability to automatically generate RDF, has had a large impact on the Semantic Web. Swoogle (swoogle.umbc.edu), at present the largest fully automatic semantic web document aggregator, currently lists nineteen large web-logging sites in its top-50 index of sites with Semantic Web content (LiveDoor is top-ranked with 9473 documents in Swoogle and LiveJournal second with 7690), and collectively blogging sites are responsible for 45% of the Semantic Web documents collected by Swoogle. Even from these sites, Swoogle crawls only a small fraction of the available FOAF documents: LiveJournal, automatically generates FOAF files for each of its 4.5 million users, several times more than the number of documents in Swoogle.

The quantitative predominance of FOAF suggests that potential Semantic Web applications need to consider what kinds of additional utility FOAF can offer. On blogging sites, FOAF supplements syndication metadata (author, title, topic, date, etc.) provided in RSS 1.0 with further detail about the authors of posts (interests, instant messaging IDs, contact information, etc.). FOAF is flexible enough, however, to be used in the context of social networking sites (Friendster, Orkut, etc.), where users post information about their relationships to other people they know, as an aid to finding new social contacts, jobs, life partners, etc. (see Boyd, 2004, for further details regarding Friendster). It is not clear what effect these sites are having, or whether they are beneficial to their users, but their immense popularity (all of the sites mentioned have millions of users) suggests many users do find them beneficial for some purpose. Since the application of FOAF is new, it is also likely that the information it encodes is not being used to its greatest potential. Hence, existing FOAF documents are a good place to learn about the possible effects of blogging and social network software sites, and to start to uncover their latent utility through the Semantic Web metadata they produce.

The scale of the information available in FOAF makes it challenging to work with. Past work on visualizing FOAF focuses on exploring networks on the level of individual actors (Mika, 2004; Mika and Gangemi, 2004). Early examples of such work could employ all of the FOAF then in existence (Dan Brickley, personal communication, September 2004). This is no longer true with the large scale social networking and blogging sites that are now using FOAF. Moreover, the logic-based tools envisioned for the Semantic Web (Berners-Lee, 2001; Alferes, et al., 2003) typically have computational complexity issues that prohibit working with large actor networks on anything but the most powerful hardware. Practical use of FOAF will need to be in closer reach of the average user, requiring average hardware to accomplish. What then is the best approach to working with FOAF?

We develop here an approach to visualizing FOAF data that employs techniques of quantitative Social Network Analysis to reveal the workings of a large-scale blogging site, LiveJournal. Our analysis specifically seeks to ascertain if the interests that users express in LiveJournal are useful indicators of their social interactions on the site, as represented by their selection of friends. Information pertaining to interests and friends is extracted from a scutter crawl of FOAF and analyzed quantitatively using Principal Components Analysis to arrive at natural groupings of the users. The information is visualized in a series of reduced sociograms (Scott, 2000; Wasserman and Faust, 1994) and interpreted. Our observations reveal an interesting organization of social life on LiveJournal, and suggest modifications to the user interfaces of social networking sites that would potentially assist users in finding one another. In addition, the FOAF data explored here exemplifies the challenges encountered in Semantic Web visualization more generally. The resolutions to the problems presented in this chapter provide a model for other practical applications of semantic web visualization.

14.2 XML, the Semantic Web and FOAF

Sir Tim Berners-Lee, original architect and visionary of the semantic web, proposed a "stack" of technologies (Berners-Lee, 2001) in order to enable his vision of knowledge management through hypertext. By the mid to late 1990s it was apparent that the World-Wide Web was a success, though a chaotic one, and occasionally a syntactic nightmare. In that context, the XML project (for "eXtensible Markup Language") was initiated to provide an extensible, machine-interpretable language for storing, communicating, and interpreting information. Since 1998, work done in XML has largely coalesced around using RDF, the Resource Description Format, as a means of enabling metadata interoperability and compatibility. RDF, in a nutshell, is a language for defining metadata vocabularies. Items which have their metadata marked up using RDF may compatibly include terms from any of a variety of XML vocabularies. It is not necessary for applications to know in advance which vocabularies will be encountered, or which items may occur; the metadata defined in RDF can still be used to make inferences.

Within RDF, the “FOAF” (friend-of-a-friend) initiative undertaken by the World Wide Web Consortium (W3C) has focused on developing ways to describe both the properties of human beings — date of birth, age, real name, nicknames, contact information of various sorts — and their social relationships as expressed through their interests, group affiliations, common haunts, places of employment, etc., Socially, the most interesting relationship encoded in FOAF is foaf:knows — the notion that one person “knows” another. This relationship is rather coarse, in that the properties or the quality of the relationship is not necessarily expressed. In addition, foaf:knows is unidirectional rather than bi-directional. Individuals sometimes assert that they “know” someone who would not necessarily reciprocate the assertion. Such self-reported, directed social relations are commonly employed in social network analysis research, as they are rich in information about people’s underlying social relationships, and many good methods are available for their analysis (Scott, 2000; Wasserman and Faust, 1994).

Typically, users of the FOAF schema would create a file that contains personal data, including email address, location, interests, a list of friends, etc., and place that file in a web-accessible location. Early FOAF data was produced by hand, usually by users interested in the technology. The early FOAF implementations were often brittle, due to nonstandard tag use and occasional lack of clarity in the specifications. The working schema of FOAF was extended several times to handle new problems, or when someone arrived at a new, useful conceptualization of how people could manage and share social metadata. More recently, large blogging and social networking sites have begun automatically generating FOAF directly from users’ profiles stored in their databases, and typically export them at automatically-generated addresses. Users generally edit their profiles using web-based forms. Hence, most users of FOAF today are completely unaware of the technology’s existence and their use of it.

FOAF files are sometimes indexed in what is called a “scutter plan”, typically a Wiki page containing the URLs of a large collection of FOAF URLs. Networks of FOAF-encoded actors can then be created by using a scutter, or RDF crawler/spider, to follow the network of foaf:knows tags and store the associated files for later analysis. A scutter plan works reasonably well with small quantities of data, as its scope is restricted enough to be managed with limited resources.

The issues involving the scale of FOAF data are exemplified by the data we obtained for our analysis. Our data come from a scutter dump collected by Jim Ley (jibbering.com) containing approximately 700 MB of parsed RDF files harvested between March 3 and March 7 2004, yielding more than 6.5 million RDF triples. We obtained the data as both raw RDF files and a 1GB MySQL database dump, which was modified for import into PostgreSQL. The dump consists of two main tables. The first of these is a comprehensive table of all of the RDF triples obtained; the second is a list of all of the URLs of FOAF files known to the scutter (including many that have not yet been retrieved), with information about the processing status of the file, the local cache name of the file, and a reference number used to identify the files that are the sources of individual triples. Table 14.1 summarizes the distribution of the 259,298 records in the URL table in terms of the operation of the scutter.

Table 14.1. Distribution of URLs in the scutter dump.

	LiveJournal		*Other*	
Date	*Visited*	*Not visited*	*Visited*	*Not visited*
March 3	663	0	1607	0
March 4	13940	121408	160	23
March 5	2810	17648	1279	130
March 6	11782	60844	0	6
March 7	4347	22650	0	1

The number of Live Journal FOAF files visited during scuttering was more than ten times larger than those from other sites. As the scuttering progressed the number of non-LiveJournal files visited dwindled to zero. This indicates a typical problem in using scuttering methods to characterize FOAF and other Semantic Web data generally. LiveJournal's design encourages users to elect friends who are also LiveJournal users to the exclusion of those who are not. In fact, before its user data was exposed as FOAF, LiveJournal did not permit people to designate friends that were not also LiveJournal users. Hence, a large social networking site like LiveJournal presents a kind of "black hole", from which a scutter will have little chance of escaping. Hence, scutter crawls are not representative samples of data, in a statistical sense. In addition, there is a substantial social consequence, since we can see that large social networking sites effectively control their members' social capital. Truly free association is not possible if the representation of the social sphere is dictated by interests from outside the people they concern.

Scale is also an issue for the analysis of FOAF data. Typically, social network analysis requires that one construct an actor-actor matrix. For just the LiveJournal site, which has roughly 4.5 million actors, the complete sociomatrix would require 4.5m x 4.5m cells, or 20.5 terabytes of storage, just for the link structure of the network. For our scutter dump, there are 274,305 distinct actors, yielding a sociomatrix of more than 65 billion cells. When we consider that we want to store other data, such as the interests of the actors, etc., storage requirements climb even higher.

14.3 Analyzing LiveJournal FOAF

The set of FOAF data we have obtained is too large to visualize without further reduction Consequently, we adopt a statistical approach to the analysis of the data, using Principal Components Analysis (PCA) and Hierarchical Cluster Analysis (HCA) as implemented in the statistical programming language and environment known as "R" (www.r-project.org). These techniques have a computational complexity that scales well enough to permit visualization of the global social patterns on modest dual-processor commodity machines. In addition, the analyses yield information which facilitate detailed exploration of the data down to the level of individual metadata elements. A summary of the FOAF namespace usage appears in Table 14.2.

Table 14.2. Counts of FOAF predicate usage from the scutter sample.

count	*predicate*	*count*	*predicate*
1067568	interest	1016	surname
869916	nick	749	depicts
868538	weblog	685	jabberID
839934	knows	674	maker
40066	mbox_sha1sum	561	title
33172	page	452	workplaceHomepage
26385	dateOfBirth	403	schoolHomepage
15624	homepage	273	codepiction
13255	aimChatID	238	currentProject
6718	yahooChatID	235	img
6072	name	203	gender
4973	msnChatID	203	phone
4275	icqChatID	195	made
1603	depiction	195	regionDepicts
1436	mbox	143	workInfoHomepage
1416	thumbnail	109	lastName
1089	firstName	1213	77 items occurring 100 times or less

Our analysis begins with the extraction of the relations of interest from the PostgreSQL database of RDF triples. This is accomplished in two steps. We first used a partial string-match to identify all triples with a predicate from the FOAF namespace (http://xmlns.com/foaf/0.1/) into a new table. Since many FOAF files use Dublin Core elements (e.g. for the interests), we did the same thing with the Dublin Core relations as well. This resulted in two tables with 3.81 million and 1.16 million triples respectively. We then tabulated the frequencies of each of the elements in the FOAF namespace to identify elements of interest for further examination. The most frequent FOAF elements in our data were the foaf:interest relation (1.06 million triples) and the foaf:knows relation (839,934 triples). Users' nicknames (foaf:nick, 869,916 triples) were used to resolve participant identities, and dc:title relations (1.11 million triples) were used to identify the object of the foaf:interest relation, as this was the strategy employed for identifying interests in LiveJournal; database JOIN operations were used accomplish these steps. The modified relations were exported as two-column text files for import into the R statistical computing environment.

Within R, both relations were reconstituted as binary incidence matrices. For the foaf:knows relation, we identified the 200 most common objects, and constructed an incidence matrix involving those and the 17,305 distinct subjects they occurred with. For the foaf:interests relation, we identified the 500 most frequent interests and tabulated their occurrence with 21,506 distinct subjects. Both incidence matrices were subjected to a column-wise z-score transformation, the effect of which is to dampen the influence of very frequent foaf:interests or foaf:knows objects. These transformed matrices were the input to PCA.

PCA was performed using the LAPACK functions for Singular Value Decomposition (SVD) available in R through the function `svd()`. The output of SVD is a pair of ortho-normal matrices, representing the projections of the rows and columns of the original data into the space of its principal components, and a vector of singular values, which are weights that can be used in a matrix computation to reconstruct the original matrix or principal components scores, as desired (Basilevsky, 1994, Chapter 3). The SVD can be written in matrix notation as in (1).

(1) $$A = P\,(\lambda\, I)\, Q^T$$

Where A is the original matrix, = P and Q are the ortho-normal matrices representing the row and column associations with the principal components respectively, and λ is the vector of singular values. P and Q must have the same number of columns r as the number of singular values retained in λ. I is an r-by-r identity matrix that permits P and Q to be re-assembled into A as a matrix multiplication.

The utility of PCA is that it permits us to characterize the major dimensions of variation in the data, as determined by the correlations among foaf:interests and foaf:knows across the set of all actors. The vector λ is ordered by the size of the principal components (where size corresponds to the proportion of the variance associated with the component), and their respective values can be compared to identify a suitable number of dimensions that retains important variation in the data; methods for doing this include Bayesian, Markov Chain Monte Carlo and other computationally sophisticated methods (Basilevsky 1994, Chapter 4); we opted for the more heuristic “scree-plot” as a computationally tractable alternative; the scree plots suggest in the neighborhood of 10 to 20 principal components can be retained in both cases.

Once the principal components are found, we can calculate principal components scores for each of the columns or rows of the original data. These scores are useful in two ways. First, they can be used as input to Hierarchical Cluster Analysis (HCA) in order to identify groups of interests or users sharing similar behavior. Second, they can be plotted for visual inspection and interpretation. We proceeded by clustering the data before plotting. The projection of the data into two dimensions tends to place many unrelated points near each other, and this can be diagnosed if the clusters are color-coded in the plot. Figures 1 and 2 provide examples of this, where HCA was performed on the column scores of the foaf:interests data, to allow meaningful clusters of the 500 interests to be viewed and interpreted. A cut of nine clusters (the largest number that can be practically recognized from a rainbow palette) was selected and plotted on the first four principal components (dimensions 1 and 2 in Figure 14.1, and dimensions 3 and 4 in Figure 14.2). The data can be explored by plotting it on different principal components, as different clusters and sub-clusters of interest are revealed. On each of our plots, we have added a unit circle and axes to clearly indicate the location of the origin in each plot.

In PCA, the first principal component tends to be sensitive to overall frequency; the same is true of the second principal component, usually to a lesser extent. Subsequent principal components are more reliable indicators of correlations among the interests in the incidence matrix. In the interest data, the most frequent interests load high on the first principal component. Music, being the most frequently expressed interest, is on the extreme right, as are art, books, movies, photography, reading, cats, computers, writing and love, which are also very frequently indicated interests. The positive direction of the second principal component exhibits a large number of specific music interests (industrial, nine-inch nails, skinny puppy, marylin manson, radiohead, etc.).

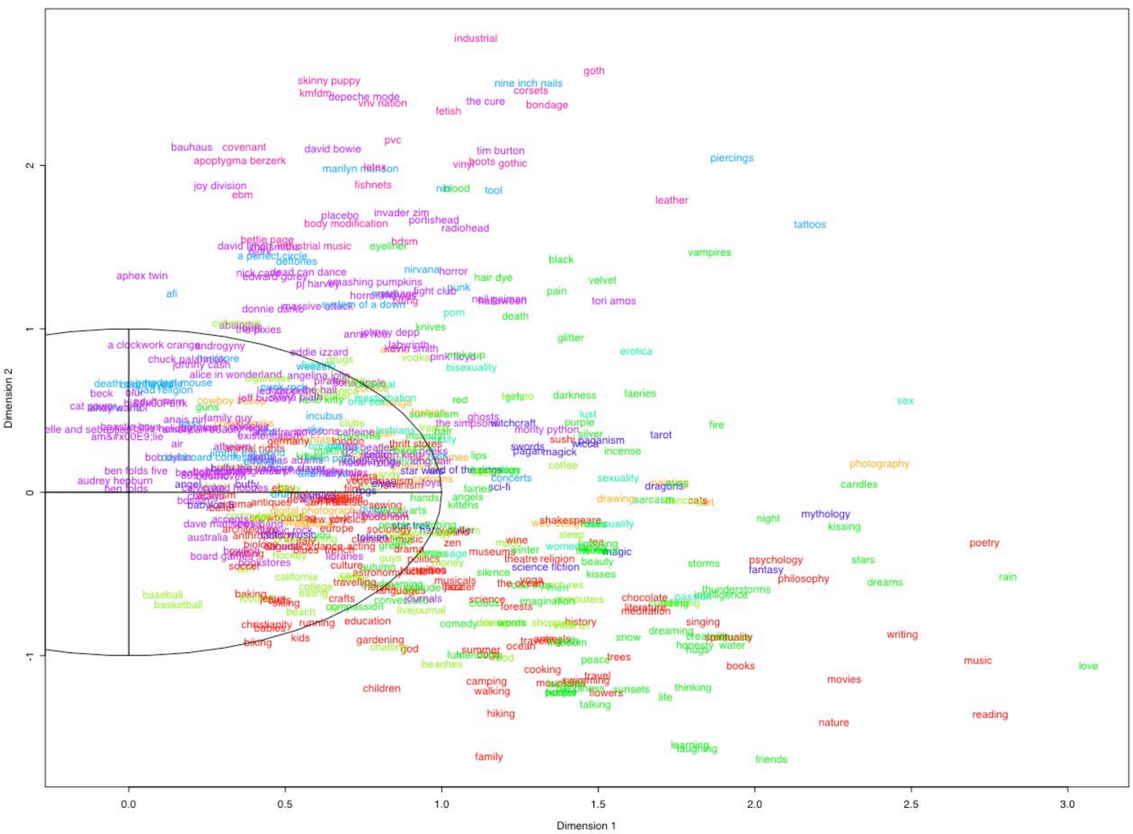

Figure 14.1. Plot of nine interest clusters on the first two principal components.

Dimensions 3 and 4 exhibit similar patterns of differentiation. The right-hand side interests concern aesthetic sensibilities (e.g. musical tastes, in the light blue and violet clusters) and general sociability, especially romance (boys, guys, kissing, cuddling, shopping, etc., in the green and yellow-green clusters). The left-hand side concerns fantasy (science fiction, paganism), philosophy (spirituality, tarot, meditation) and sexual fantasy/fetish (bondage, erotica, polyamory, etc.). Vertically, there is a split between body/sexual and romantic interests (bottom) and intellectual/aesthetic interests (top). The scatterplot reads like a map for these particular forms of user interest.

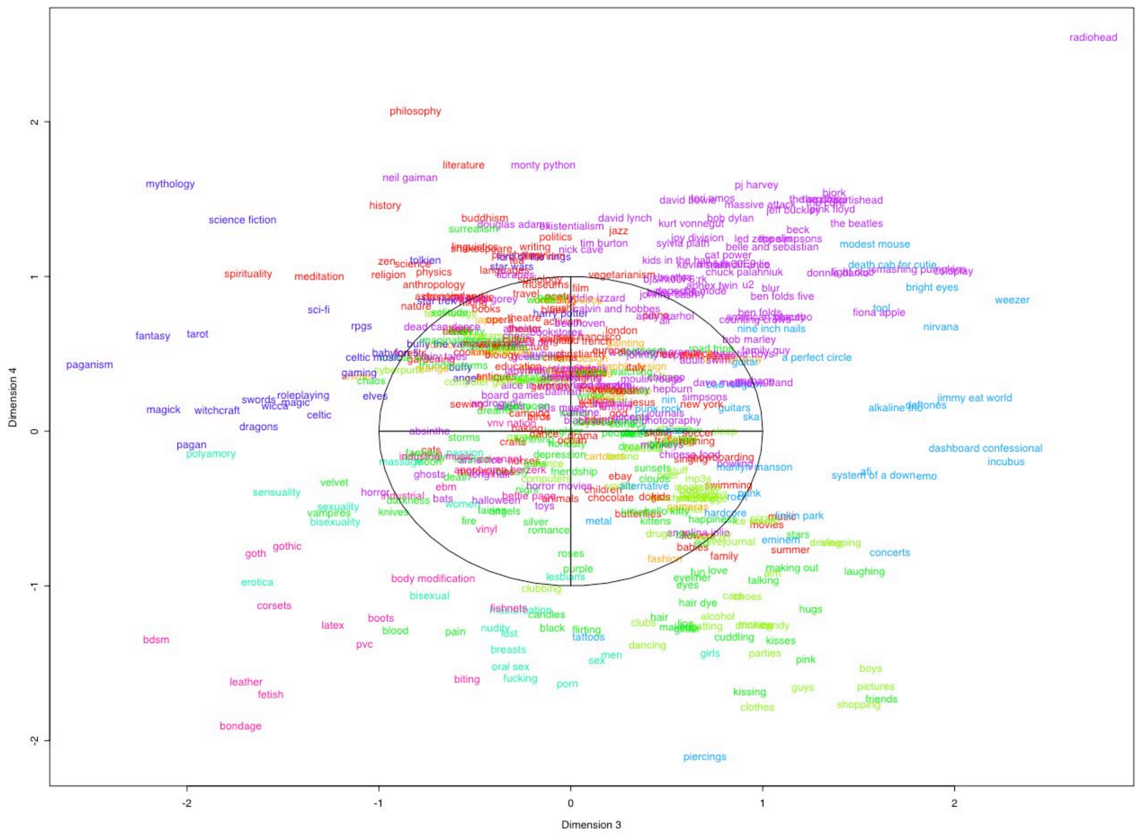

Figure 14.2. Plot of nine interest clusters on principal components 3 and 4.

The cluster analysis also reveals useful information for interpretation. The nine clusters in our chosen cut can be roughly characterized as general interests (red), art (orange), nightlife (yellow-green), romance (green), sexuality (aqua), grunge music tattoos and piercings (sky blue), science fiction/fantasy (blue), music (violet) and goth subculture, and sexual fetish (pink). Goth subculture is a primarily North American counterculture movement distinct from Punk which emphasizes morbidity, death and to some extent androgyny. Goths, both female and male, also typically put great care into their appearance, spending hours dressing and applying makeup before appearing in public.

The clusters are not entirely pure (e.g. kurt vonnegut is found in the music cluster, etc.) but the labels are general characterizations based on an inspection of the complete cluster hierarchy that can be useful in developing interpretations of the data. Clustering at a lower level in the hierarchy can also be inspected to reveal patterns of interest. For example, god, christianity and jesus form a cluster together within the red cluster of general interests; this is certainly a coherent group of interests. Its closest neighbor is a cluster consisting of two subclusters: (a) religion, philosophy, psychology, sociology, spirituality, meditation, buddhism, zen, and (b) animal rights, feminism, activism, yoga, sewing, knitting, vegetarianism, and thrift stores. Again, an overall pattern of coherence can be identified which matches subcultural social patterns in off-line behavior.

Clusters of actors on principal components scores from the foaf:knows incidence matrix indicate social positions, or equivalence relations among actors with common relations to other actors in the network. These are useful in examining the social structure of the foaf:knows network to identify patterns of centrality, power and exchange among users of LiveJournal. Figure 14.3 illustrates this set of patterns at two levels of reduction in the form of a reduced sociogram, where by “reduced” we mean that equivalent actors have been aggregated into a single node for the purposes of display.

Figure 14.3, left, shows nine clusters, which are disaggregated into 77 clusters on the right, keeping similar sub-clusters together in space and preserving the color-coding of the larger clusters they belong to (note that the spring layout algorithm does not preserve the position of the clusters in general, just their pattern of connectivity). A strong pattern of interaction is evident in both network graphs, in which members of the red cluster are seen to be most central: there are strong self-links within it, and all other clusters link strongest to the red cluster as well. In addition, the yellow-green cluster shows fairly strong links to the red cluster in the left-hand side of Figure 14.3, and ties to all of the other clusters as well. However, this pattern of linkage disappears on the right-hand side of Figure 14.3, suggesting that the yellow-green cluster’s patterns of relationship are not at the level of group-to-group in the disaggregated clusters. The threshold at which links are displayed is partly responsible for this difference in appearance, but the level of resolution, in terms of numbers of clusters across which links are aggregated is also probably responsible, as there are many more groups among which an actor’s relationships can be scattered. Hence the different levels of resolution reveal different patterns in the social structure of the network. Thus, the yellow-green cluster has a much less tight social organization than the red cluster, and other clusters are peripherally associated with the network-central red and yellow-green clusters.

These observations suggest that LiveJournal has a distinct core of social activity, at least as far as we have sampled it. Other participants are more peripheral. We do not see strong evidence of partitioning of LiveJournal into different camps, although this question bears closer scrutiny. Having found strong pattern of social structure on LiveJournal, we next

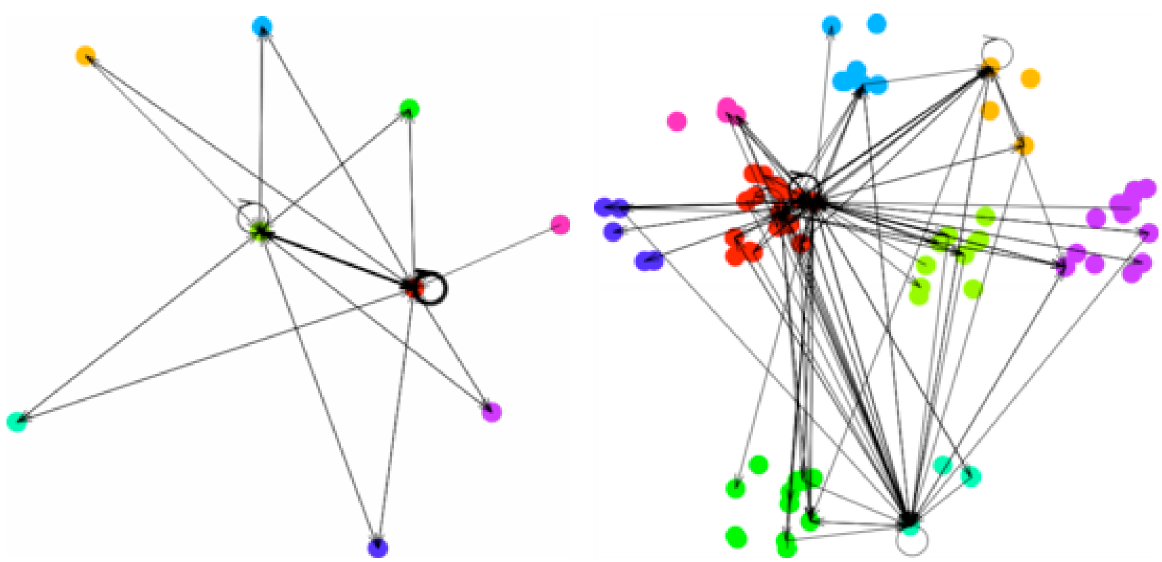

Figure 14.3. Social positions in the foaf:knows relation revealed at different levels of clustering.

ask if they are related to the users' expressed interests in any way: do the users' interests assist the social structuring of interaction on LiveJournal? To address this question, we used HCA to identify groups of users based on their principal components scores for the distribution of interests. Once identified, the foaf:knows relations among members of these groups were again aggregated. Our hypothesis is that if the interest groups are coherent social entities, we should see a predominance of self-ties within them, perhaps alongside a more global pattern of relationship and interaction among the interest groups. If not, only a global pattern of interaction would be visible.

Figure 14.4 visualizes, again at two levels of clustering, the social relations of groups of actors with shared interests. Although the palette is the same as for the previous visualizations, the categories and their color coding are not shared, as they represent clusters based on different information. Color coding is preserved, as well as general position in space, between the left-hand and right-hand images of Figure 14.4. Again we see a network with a strong central core and a periphery; on the left of Figure 14.4, the central group this time is yellow-green, with the strongest self-links and links to all of the other groups. Again there is a second group that is linked above the threshold level to most of the other groups, but the strong reciprocal pattern we saw on the left of Figure 14.3 is not evident. There is clearly different social information in Figures 14.3 and 14.4.

Regarding our hypothesis about the social relations and interests, we find some evidence for cohesion in the diagram on the left of Figure 14.4, in that four out of the nine groups exhibit some degree of self-linkage. This evaporates, however, in the disaggregated groups in the diagram on the right of Figure 14.4, where only the yellow-green groups retain their self-links. Note also that many foaf:knows relations appear to point outward from this central group, and relatively fewer inward. These patterns suggest a highly articulated core of interests around which social cohesion is built among a central set of participants on LiveJournal; other groups with somewhat more diffuse interests, and having lower overall social cohesion, occupy the periphery of the network.

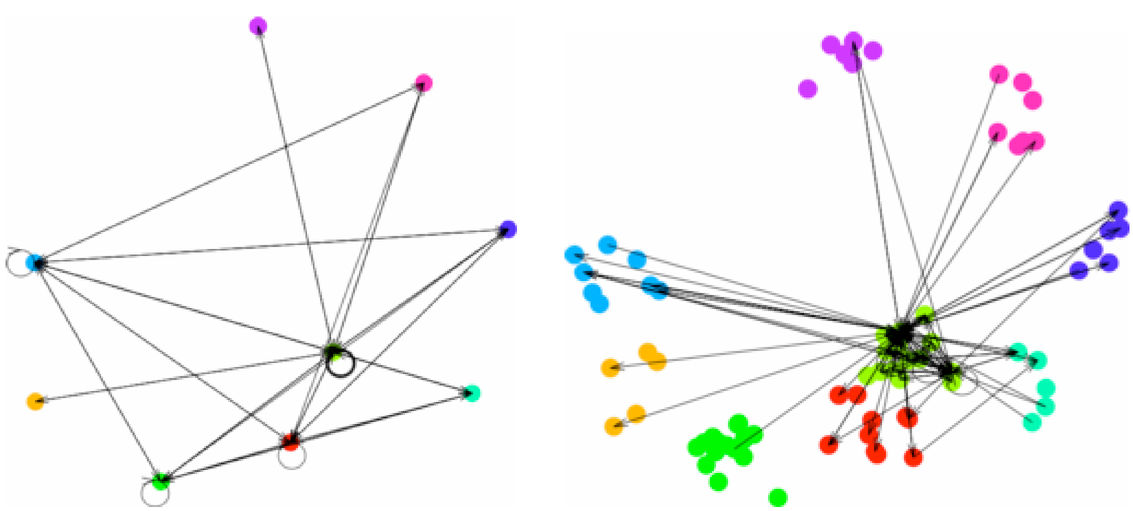

Figure 14.4. Social relations of shared interest groups at two different levels of clustering.

Having established that the social life of LiveJournal can be described meaningfully in terms of the shared interests of its users, the natural question to ask is which interests are shared, and how they contribute to the social structuring of LiveJournal. Here, we are interested if we can identify clusters or sub-clusters of relations that are associated with any of the different clusters above. To investigate this, we computed centroids of each of 76 interest clusters and each of the 77 interest-based social groups on each of the principal components scores. We then computed Euclidean distances between each interest group and interest cluster pair, from the two sets of factor scores. These distances were visualized as a bi-modal network, with the interest groups (people) represented in the same relative positions they occupy in Figure 14.4, and the interest clusters (interests) positioned in a ring outside them. The interest clusters are sorted and color-coded according to the same scheme as in Figures 14.1 and 14.2, so their relationship to these plots can be more easily identified.

This visualization is presented as Figure 14.5, where interpretive labels of the nine interest clusters have been added to assist the viewer. As suggested by our interpretation of Figure 14.4, we note a broad range of interest clusters characterizing the central

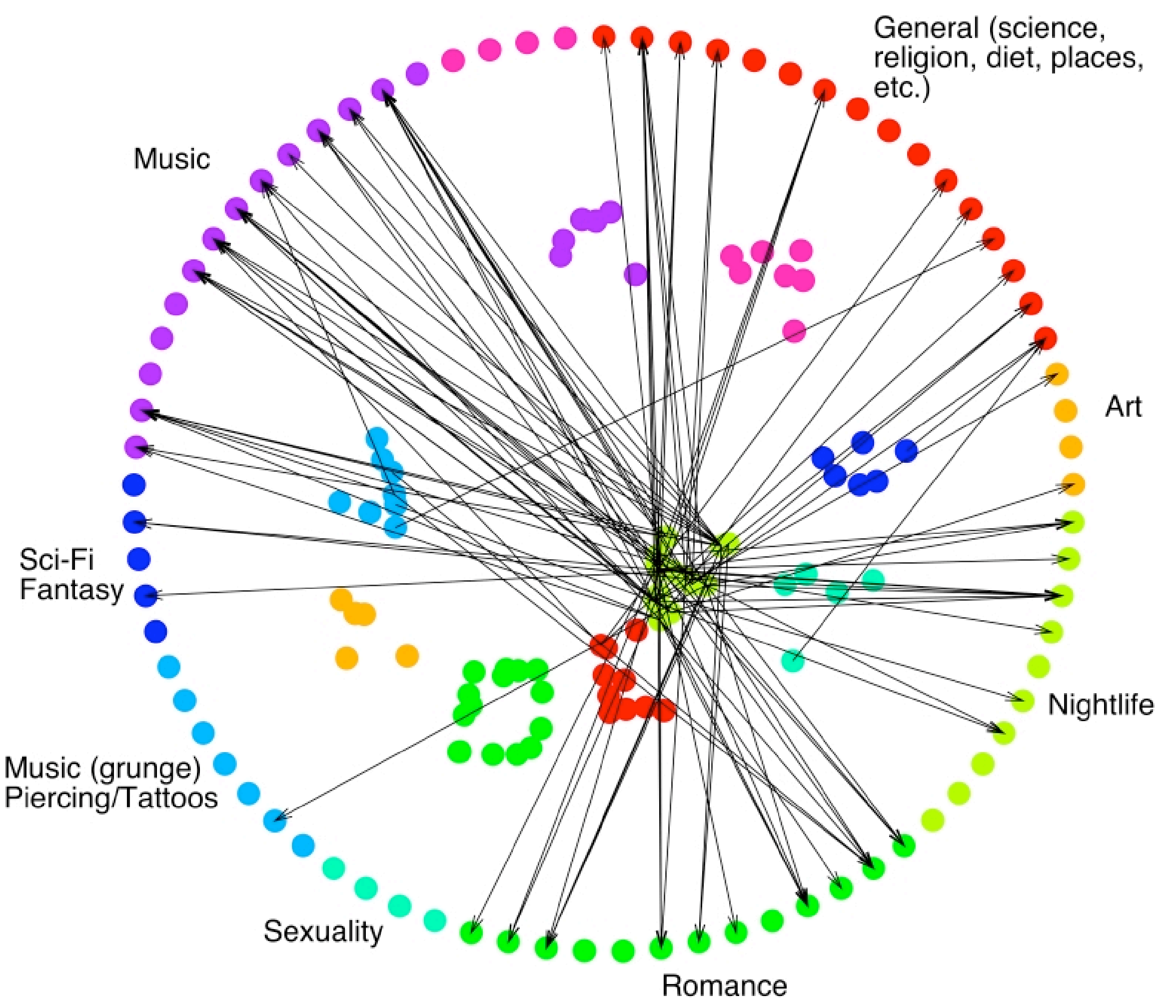

Figure 5. Relation of interest clusters to groups of actors with shared interests.

yellow-green groups, but markedly fewer interests characterizing the peripheral interest groups. Each sub-group within the yellow-green group has somewhat different links to the sub-clusters of interests, although there is substantial overlap among particular sub-interests in the music, general, nightlife and romance interests. Only a very few of the peripheral interest groups have links of any strength to any specific clusters of interests. Note that this is the first pattern to emerge from the network at any level of link threshold: any threshold low enough to increase the number of links to peripheral groups causes almost all of the interest clusters to be connected to all of the yellow-green sub-groups.

These observations confirm our impression that the yellow-green group's interests are more coherent, both in terms of relation to general spheres of interest as well as in social structuring, than the other peripheral interest groups. At the same time, the relations between interest sub-clusters and sub-groups are very highly articulated, such that different sub-groups of users can be meaningfully distinguished in term of their interests.

We do not know yet specifically which interests are responsible for which structures, or which users are members of which groups, but this information could be obtained by careful collation of our statistical analyses with the foaf:knows and foaf:interests database tables. In future research, we hope to examine in particular members of the yellow-green group to ascertain if their online behavior comports with and/or further illuminates our findings here.

14.4 Discussion and conclusions

Our visualizations of Semantic Web social metadata suggest that statistical and quantitative approaches have much to contribute to the understanding of how these new technologies are used, and what sorts of changes might improve them. While Semantic Web metadata is meant to provide standardized ways of annotating information, we see in the application of the FOAF relations studied here, a very rich structuring of the data that is not readily captured in ontologies and yet which is very close to the meaning of social life in the online environment of LiveJournal.

The foaf:interest relation is especially instructive in this regard. It is unlikely that a useful ontology of interests could be devised that would express the same social meanings and correlations as the interest clusters discovered here. These clusters are discovered within markup that is essentially uncontrolled — the object of the dc:title relation, which accepts any literal value. Moreover the empirically-discovered interest clusters are subject to change, as people re-organize their social relations and re-orient themselves to the new social realities they create. An interesting research project in the Semantic Web applications of FOAF would be to try to expose some of this dynamic, changing and emergent structure for use in inferences. This will require means beyond the sort of logical deduction envisioned by the architects of the Semantic Web's current form (Berners-Lee, 2001; Alferes, et al., 2003).

Our visualizations also suggest that LiveJournal's user profile interface, which allows users to select their interests in a fairly open-ended way, nonetheless permits users to usefully organize themselves according to their interests. Patterns of interest and user interaction are intricately inter-woven, suggesting that the design of the site is at least partly successful in its goal of fostering online, interest-based communities of users.

At the same time, it must be noted that these effects are most pronounced for a central core of the users in our sample; outside of this core, both social relationship and the coherence of interests become more diffuse. Consequently we must regard the vast majority of the ties indicated through foaf:knows or foaf:interests as "weak ties" in the sense of Granovetter (1973). Strong ties, such as those between people who know each other well or who spend a great deal of time together, are only found in the central core of the network. This pattern closely resembles that observed on the Internet Relay Chat (IRC) channel #india (Paolillo 2001), where a predominance of weak ties was also observed. In fact, a central aim of computer-mediated communications systems like weblogs and IRC is, in essence, to amplify weak ties.

The consequences of this for the structure of social interaction and its potential outcomes bear considering. The nature of social interaction through weblogs is different from that of face-to-face communication. Since the subscribership of community weblog and social networking websites is booming, this suggests a more general change in the way that social relations are enacted, at least among the users of such sites, if only because of the amount of social interaction they have in the online context as opposed to in more traditional contexts.

Moreover, as is evident from the nature of the expressed interests of the users, the majority of LiveJournal users are young (between the ages of 15 and 30), a life-stage in which people are at their most sociable, and a great deal of social development occurs (Degenne and Forse, 1999; Forse, 1981). It is important to ask what these changes in social behavior might mean for the future of social life — a change in the social development of even a few million youth could have an impact on future trends. Data-mining and visualization of the social metadata made available through FOAF is a valuable way in which we can study and begin to understand such issues.

Another possible application of these results would be to use them to inform the design of user interfaces of community weblogs and social networking sites. User interfaces could present visualizations of a site's social metadata in visualizations similar to those employed here. By manipulating the visualization, its level of resolution, the arrangement of clusters, etc., users would be able to locate themselves within the social life of the site, navigate through it, and find other users or even interests they were not previously aware of by following links among the clusters and groups. While such visualizations are more complex and harder to generate than social visualizations proposed for other communication modes by Viegas and Smith (2004), Donath, et al. (1999) and others, they complement those approaches by providing both detailed and manageable access to the complex link structure of an online social space. Such interfaces might do more to strengthen the ties among members of peripheral groups, whose structuring around interests is currently less coherent than the core.

We must keep in mind, however, that the young users of blogging and social networking sites like LiveJournal are also vulnerable to social manipulation — through sexual harassment, pornography, commercial marketing, surveillance by law enforcement, legal intimidation and other means — by forces that are arguably in a better position to exploit Semantic Web metadata than the users of the weblog services that generate it. The observations made here — ethically permissible as research because of the public nature of the data — could easily be used to target people for marketing purposes (many of the interests have a commercial aspect), for lawsuits about copyright violation (because of associations between certain musical tastes and illegal downloading) or for political surveillance (because of expressed anti-war, pro-drug or otherwise minority views). Whether our aim is to provide improved communication and community technologies, or to conduct research and advance our knowledge of the interaction of social and technical systems, we must not lose sight of the human experiences behind the statistics and metadata that we employ. It is our responsibility to wield these tools carefully as we explore the world of social metadata through visualization.

14.5 Acknowledgements

The authors would like to thank other members of the Blog Research on Genre project at Indiana University, and participants at the FOAF Workshop, Galway 2004, for their interest and support of this research.

14.6 References

Alferes J, C Damásio, L Pereira (2003) Semantic web logic programming tools. *Lecture Notes in Computer Science* 2901: 16-32.

Basilevsky A (1994) *Statistical Factor Analysis and Related Methods: Theory and Application*. Wiley, New York.

Berners-Lee T, Hendler J, Lassila O (2001) The Semantic Web. *Scientific American* 284, May: 34-43.

Boyd D (2004) Friendster and Publicly Articulated Social Networks. In *Proceedings of CHI 2004, Connect: Conference on Human Factors and Computing Systems*. ACM Press, New York.

Brickley D, Miller L (2003) FOAF Vocabulary Specification. Technical report, RDFWeb FOAF Project.

Degenne A, Forse M (1999) *Introducing Social Networks*. Sage Publications, Thousand Oaks, California.

Donath J, Karahalios K, Viegas F (1999) Visualizing conversation. In *Proceedings of the 32nd Hawaii International Conference on System Sciences*. IEEE Computer Society, Los Alamitos, California.

Forse M (1981) La sociabilite. *Economie et Statistique* 132:39-48.

Golbeck J, Hendler J (2004) Inferring Reputation on the Semantic Web. In: Proceedings of the Thirteenth International World Wide Web Conference (WWW2004). ACM Press, New York.

Granovetter M (1973) The strength of weak ties. *American Journal of Sociology*, 78.6: 1360-1380.

Herring S, Scheidt L, Bonus S, Wright E (2004) Bridging the Gap: A Genre Analysis of Weblogs. In *Proceedings of the Thirty-seventh Hawaii International Conference on System Sciences* (HICSS-37). IEEE Computer Society, Los Alamitos, California.

Mika P (2004) Bootstrapping the FOAF-Web: An Experiment in Social Network Mining. *First International Workshop on FOAF, Social Networks and the Semantic Web*, Galway, Ireland, 2004.

Mika P, Gangemi A (2004) Descriptions of Social Relation. *First International Workshop on FOAF, Social Networks and the Semantic Web*, Galway, Ireland, 2004.

Paolillo, J (2001) Language variation on Internet Relay Chat: A social network approach. *Journal of Sociolinguistics*, 5:180-213.

Scott, J (2000) *Social Network Analysis: A Handbook*, *Second Edition*. Sage Publications, Thousand Oaks, California.

Viegas F, and Smith M (2004) Newsgroup crowds and authorlines: Visualizing the activity of individuals in conversational cyberspace. In *Proceedings of the 37th Hawaii International Conference on System Sciences*. IEEE Computer Society, Los Alamitos, California.

Wasserman S, Faust K (1994) *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge.