The Social Semantics of LiveJournal FOAF: Structure and Change from 2004 to 2005

John C. Paolillo[1,2], Sarah Mercure[2] and Elijah Wright[2]

[1]School of Informatics, Indiana University
[2]School of Library and Information Science, Indiana University
Bloomington, Indiana, 47405 USA
{Paolillo, smercure, ellwrigh}@indiana.edu

Abstract. Social Network Analysis methods hold substantial promise for the analysis of Semantic Web metadata, but considerable work remains to be done to reconcile the methods and findings of Social Network Analysis with the data and inference methods of the Semantic Web. The present study develops a Social Network Analysis for the foaf:knows and foaf:interests relations of a sample of LiveJournal user profiles. The analysis demonstrates that although there are significant and generally stable structural regularities among both types of metadata, they are largely uncorrelated with each other. Also there are large local variations in the clusters obtained that mitigate their reliability for inference. Hence, while information useful for semantic inference over user profiles can be obtained in this way, the distributional nature of user profile data needs closer study.

1 Introduction

The Semantic Web is largely characterizable as an enterprise with one principal goal: the formalization and standardization of online metadata. Over the past few years, the purpose of this effort has been variously characterized as assisting interoperability of web information systems [1], facilitating Knowledge Management [2], and permitting the development of future web applications based on artificial intelligence techniques [3]. For this reason, it is somewhat surprising that social networking metadata, using the FOAF (Friend-of-a Friend) vocabulary, has emerged as possibly the single most prevalent use of Semantic Web technologies so far [4,5]. Numerous weblog and journal-hosting sites now export their user data using FOAF, alongside their content serialization in RSS.

On the surface, the two types of information are very different. For example, businesses needing to communicate about the stock, manufacturing conditions, materials and specifications of metal fasteners have very different communication needs from private individuals looking to maintain or develop a circle of friends and acquaintances. Moreover the word "ontology", used to describe a set of interlocking metadata definitions, suggests a static Platonic conception of objects in the world, whose existence is to be identified ("marked up") through the application of labels.

Social network information defies this sort of conception: is so-and-so really your "friend", or are you simply name dropping to enhance your appeal in the present conversation? Or are you more than friends, but happen to be downplaying your association with someone I may not approve of? Would two people agree on the meaning of friendship, even a shared friendship? And would this agreement be enough to permit other inferences to be drawn, as is the purpose of the Semantic Web? The parameters of social information are much harder to determine than Platonic semantics would allow, and therefore pose major challenges to the Semantic Web project.

The tools of Social Network Analysis (SNA) have developed to cope with this sort of issue. Typically, network analysts employ large amounts of readily-collected information — mutual naming patterns, event participation, helping relations, desirability ratings, etc. — to identify social positions, roles and relationships in aggregate, either through statistical means or by interpreting geometric representations of the information [6,7]. These methods offer a powerful window into social functions and processes from the most mundane (e.g. the pronunciation of English vowels by Detroit teenagers [8]) to the most rarified (e.g. US Supreme Court rulings, [9]). At the same time, observations obtained through network analysis are necessarily time-bound: they may not be true for any other time or context of observation. Hence, social network information must always be interpreted cautiously.

For SNA, the emergence of FOAF is a happy coincidence, in that it provides an inexpensive source of large amounts of reasonably rich social network data. The utility of FOAF for SNA is helped by its deliberate vagueness: relations such as foaf:knows are not explicitly tied to terms like "know" or "friend" in other ontologies like WordNet, and hence can be allowed to vary somewhat according to context. Furthermore, the availability of social network metadata makes it possible to consider the relationship of semantics to social structure, i.e. "emergent semantics", from the dynamic nature of social relations and their role in fostering and communicating semantic change. Many questions pertinent to the Semantic Web effort arise in this context. Do the ontologies we encode represent the semantic relations that people actually use? Can existing Semantic Web technologies encompass dynamic, context-bound meanings? How stable are these meanings over time? And to what extent can ontologies be adapted to such changing, context-bound meanings, to provide useful inferences? Hence, the Semantic Web also potentially benefits form the application of SNA methods, which might provide templates for evaluating the Semantic Web metadata in contexts where it is actually used.

It is this set of issues that motivates the current study, which examines the change in a set of LiveJournal profiles for friends (represented by foaf:knows) and interests (foaf:interests) of eighteen thousand users. By examining these data, we can potentially learn about the way in which people's online social spheres evolve alongside their changing interests. This provides a view onto the emergent semantic categories of interests and their relation to the social milieu.

2 Methods

For this study, we obtained two samples of LiveJournal profiles, collected one year apart. The first set of profiles consisted of a subset of the FOAF files collected by Jim Ley in March 2004 which happened to be LiveJournal profiles. Ley's purpose in obtaining the profiles was to provide a database of FOAF files upon which to develop some demonstration applications. The data were provided both in the form of raw RDF/XML files, and as a MySQL database of RDF triples. RDF statements pertaining to LiveJournal user accounts were readily identified through having common kinds of information (especially nick-names), and through their overall social coherence (the LiveJournal profile interface only permits users to friend other LiveJournal users). These users were identified and analyzed in depth in earlier work [5].

2.1 LiveJournal

LiveJournal is a service which permits users to create online personal journals. Created in 1999 by Brad Fitzpatrick as a way to keep in touch with his own friends, LiveJournal has since snowballed into a dynamic community of individual and group journals. In the spring of 2004 participation in LiveJournal reached nearly three million users and in just over one year's time since then, the number of users had ballooned to over seven million, according to the site's front page.

Users of LiveJournal include a wide range of individuals from all over the world. At this time most registered users are from the United States, Canada, the United Kingdom, the Russian Federation, and Australia, from most to least common. Of the users from the US, the most report that they live in California, with users from Florida, New York, Michigan, and Texas following closely behind. Over 60% of users are female, and the highest distribution of users fall within the 15-20 age range. Considering the broad range of user backgrounds we can only speculate as to the general reasons that users participate. Given the personal nature of data required for user registration, as well as the features provided by LiveJournal, it seems clear the site is intended for presenting personal journal-like information. Previous research supports this notion with evidence that journal-type blogs are the most common use of blogging tools [10]. Typical user pages contain personal accounts of the user's day or details of recent life events.

Upon registering for an account LiveJournal users are presented with a form which requests optional information from the user, such as birthdate, gender, geographic location, email address, etc. In addition the form provides space for the user to post an image and/or brief biography, set privacy controls for their content, and list interests. The interface for interests is a free-text box, so users are not at all limited in what or how many interests they declare. User information is exported as an automatically generated FOAF RDF/XML file [11], made available at a pre-determined URL. Users can update this information at any time.

Common interests listed by users are presented on a users "user info" page and hyperlinked to a list of all the Livejournal communities and users who list the same

interest, thereby facilitating the social engagement of users who share similar interests. The user info page also lists a user's list of "friends", who are typically other LiveJournal users. "Friending" another LiveJournal user allows one to do more than simply declare affiliation; it is also tantamount to entering a subscription to the content they produce. Users can take advantage of an automatically generated link on their main page that brings up the past several entries for each of their listed friends. The process of designating interests and friends thus creates a complex interlinked network of users, facilitating easy access to social groups and to people with common interests.

2.2 Data Handling

The RDF triples from the 2004 sample were imported into a PostgreSQL database, from which foaf:knows and the foaf:interest relations were extracted using database operations, along with other supporting relations, such as dc:title and foaf:nick, which allowed us to identify meaningful content with the relations. For foaf:interests, we identified the 500 most popular interests, and a subset of 21,506 user profiles mentioning these interests. For foaf:knows, we identified 500 most popular recipients of the foaf:knows relation, and 11,818 users' profiles stating that the associated person foaf:knows at least one of those people. Each relation was arranged into a binary incidence matrix, I_{2004} for foaf:interests and K_{2004} for foaf:knows, with cell values $I_{i,j}$ and $K_{i,j}$ being 1 or 0, indicating for each profile i whether a relation to a given interest or popular user j is present. These incidence matrices were used in the subsequent statistical analysis.

In June 2005, we took the same list of 21,506 LiveJournal users and retrieved current FOAF profiles for them by means of an automatic script. These were imported into SWI-Prolog, where Prolog rules were used in place of the database operations to obtain the foaf:interests and foaf:knows relations for the 500 most popular interests and users identified in the 2004 sample. These were arranged in a second pair of incidence matrices, I_{2005} and K_{2005}, as had been done for the 2004 profiles.

The four incidence matrices were carefully collated to produce a final incidence matrix for the complete profile data P, containing information from both the foaf:interests and foaf:knows relations, for both years. This resulted in a matrix of 18,725 rows and 2000 columns. FOAF profiles in the 2004 foaf:interests matrix that were in none of the three other matrices were dropped from consideration at this point. The foaf:knows relations showed the largest change over the two years, with nearly half of the users in each year not found in the other year's data. Such instances were handled by filling in the appropriate cells of P with zeros.

A key observation made at this point is that users' social relations — as represented by their orientation toward the most popular users — appear to fluctuate more than do their interests, but this appearance may be due to the greater sparsity of the foaf:knows relation, with respect to the foaf:interests relation.

2.3 Statistics

Statistical data handling was accomplished using R, the GPL statistical environment and programming language. Specifically, we conducted a Principal Components Analysis (PCA) of the matrix P, to identify the structure of inter-correlation among the interests and popular users (columns) of P. PCA is a statistical technique common in machine learning, and Latent Semantic Analysis [12] is a related technique. Co-citation analysis [13,14,15] is an alternative approach widely used in bibliometric studies which employs co-citation, a similar measure of relationship similar to correlation, which PCA is based on. Correlation has the advantage of being centered about the overall mean of the data, and scaled according to its overall variance. These treatments result in a projection of the data into a space whose greatest dimensions are the dimensions of greatest (co-)variation in the data itself, rather than an artificial set of dimensions representing an artifact of the analysis or observational procedures. Moreover, when distributional assumptions are met, the PCA supports recognized statistical inferences [16]. For these reasons, we felt that PCA would be an appropriate technique to allow us to compare the relations among different interests and popular users in our set of user profiles.

PCA of the matrix P is accomplished by Singular Value Decomposition (SVD) of a z-score standardized version of P, namely $Z = UdV^T$. The output matrices U and V are sets of eigenvectors representing the rows (U) and columns (V) of Z, and d is a diagonal matrix of singular values of Z. U and V are rectangular matrices with r columns, and d is an $r \times r$ square matrix, where r is the rank of matrix Z, or the number of interpretable principal components. Various methods are used to choose r, the most common being a scree plot, showing the rank-size relation among the singular values d. For interpretation, we use the factor scores of Z, where the column eigenvectors V are scaled using d: $F = Vd$. These vectors are treated as representing points in r-dimensional space, allowing us to measure their distances, for clustering the interests and popular users. Because of the size of P, R's built-in functions `prcomp()` and `princomp()` were unsuitable, and we wrote our own more lightweight wrappers for the function `svd()`, R's interface to the LAPACK routine for SVD.

A scree-plot was used to identify 20 dimensions of variation as potentially significant. These Principal Components were visualized as factor score plots. The factor scores were then split into four groups for foaf:knows and foaf:interest relations in 2004 and 2005, and these groups were submitted to hierarchical cluster analysis, using Euclidean distances and Ward's method. These cluster analyses were visualized as dendrograms, in order to identify the content and organizational structure of each cluster. Finally, the clusters were cross-tabulated for 2004 and 2005, and those cross-tabulations were visualized as networks, so that the dynamic re-organization of the foaf:knows and foaf:interest relations could be studied. The results of these analyses are presented below.

3 Results

The first set of results concerns the Principal Components of the correlation matrix of most popular users and interests. Figure 1 presents the first two Principal Components, while Figure 2 presents Principal Components 3 and 4. Interests are represented by circles, while friends are represented by squares. Likewise, 2004 data are represented by open symbols, while 2005 data are represented by filled symbols. Both figures are enhanced with four ellipses indicating 95% confidence intervals for the distribution of each of the four year/relation categories.

From Figure 1, it is immediately apparent that Principal Components 1 and 2 primarily separate out interests (projecting leftward from the origin) and friends (projecting upward from the origin). Remarkably, there is very little correlation between friends and interests of either year, whereas, across years, friends correlate closely with friends, and interests with interests. The coincidence of the pairs of ellipses for interests and friends further indicates this tendency. Moreover, the locations of interests or friends in a particular year tend to be close to their corresponding locations in the other year. The 2005 locations tend to be a bit closer to the origin overall, indicating weaker correlations among the interests and friends in 2005.

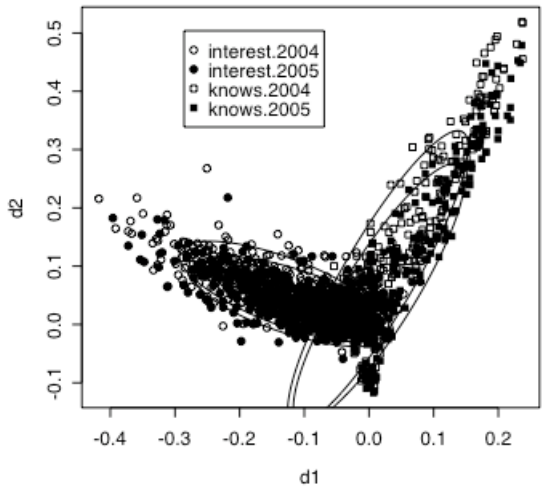

Figure 1. Principal Components 1 and 2 of the interest and friends data of 18,725 LiveJournal users.

Hence, from Figure 1 we learn that the manner in which people elect friends and interests in their LiveJournal profiles is sharply different. This is all the more surprising in that, although popular interests tend to be more common than popular friends, the two sets of relations nonetheless overlap substantially in their frequencies. Conse-

quently we must conclude that these differences are not merely artifacts of the method, but represent fundamentally different social behaviors.

Principal components 3 and 4 confirm and expand on this finding, as seen in Figure 2. Here, there are also two main branches of data points, and again 2005 data points tend to be a bit closer to the origin from their corresponding locations in 2004. This time, the two branches consist entirely of distinct intra-correlated sets of friends. There is a tendency for interests to spread in a similar pattern, but much closer to the origin. The confidence ellipses for the four sets of relations show that the 2004 and 2005 data are again largely coincident in their range of variation. However, neither interests nor friends is significantly stretched along either axis to the exclusion of the other. Further principal components show a more-or-less normal distribution about the origin, with interests clustering closer to the origin than friends.

Hence, popular interests exhibit less variation in their overall distribution than popular friends. At the same time, the election of interests and friends by users is not strongly inter-correlated, although strong intra-correlations of certain sets of users, at least, can be found.

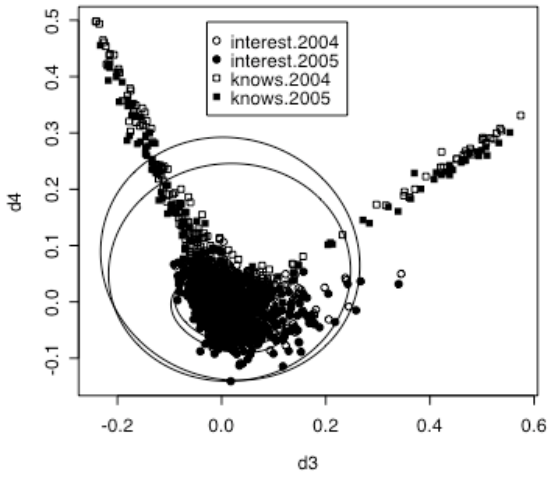

Figure 2. Principal Components 3 and 4 of the interest and friends data of 18,725 LiveJournal users.

Following this, we conducted hierarchical cluster analyses of the four subsets of variables. The clusters found among the interests largely confirm the earlier analysis [5]. At the same time, there is considerable variation from one year to the next in the content of the clusters. For example, Figure 3 shows dendrograms for two roughly corresponding clusters from the foaf:knows relations in 2004 and 2005. The 2004 cluster is much tighter than the 2005 cluster, as indicated by the height scale along the top, while the 2005 cluster is considerably larger, containing more members. Popular users in the 2004 cluster are readily located in the 2005 cluster, although their relative proximities are generally not maintained, and they are intermingled with popular users

not present in the 2004 cluster. Apparently, this cluster became somewhat looser from 2004 to 2005, and expanded into a region of space where other clusters of popular users were located.

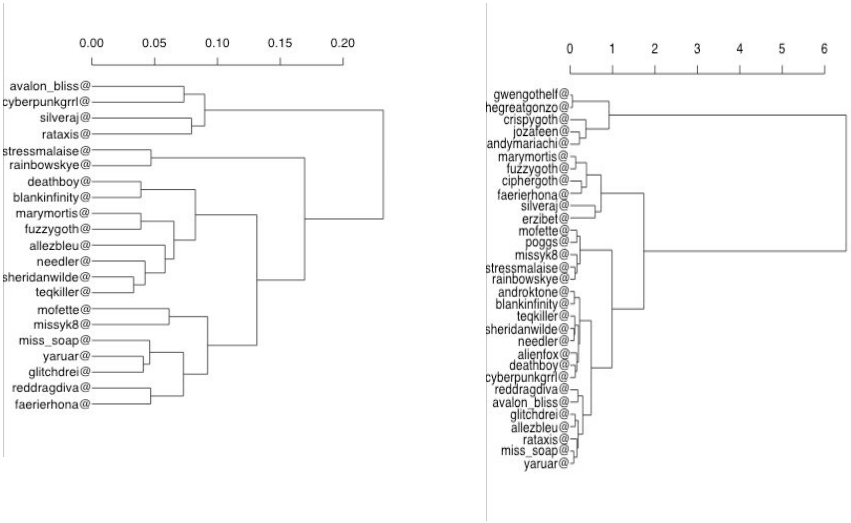

Figure 3. Dendrograms for corresponding clusters of popular users (foaf:knows) in 2004 and 2005.

Because of this movement at the finer levels of structure in the dendrograms, we felt that the analysis would benefit from examining the structures present at a coarser level of granularity. This was done by using a "cut" at a height that would partition the interests and popular users into a small number of clusters. The interests clusters were then interpreted; the knows clusters were not interpreted, since that would have required reading hundreds of weblogs to form a suitable interpretation. Using a cut with five clusters, the 2004 interests can be partitioned as follows:

(1) Science Fiction, Fantasy, Celtic, and graphic arts,
(2) General interests
(3) Sex, goth subculture, body modification and fetish
(4) A variety of contemporary music interests (The Cure, Joy Division, David Bowie, Radiohead, Pink Floyd, Led Zeppelin, Ani DiFranco, Modest Mouse, etc.)
(5) Industrial and alternative rock music.

The 2005 interests are somewhat different, partitioning into clusters as follows:

(1) Social, natural, and mystical interests
(2) General interests
(3) Ska, punk and alternative rock music
(4) Science Fiction, Fantasy, Celtic, and graphic arts
(5) Sex, goth subculture, body modification and fetish

There is clearly a broad correspondence among the categories of both years, with some of the clusters (such as Cluster 1 in 2004 and Cluster 4 in 2005) nearly identical across the two years. At the same time, some differences are evident. To ascertain the extent and nature of the reorganization of interest clusters, we cross-tabulated the analyses for the two years and re-arranged rows and columns to maximize the diagonal. Thus, we equate clusters from the two years that have maximal common membership. We then visualized this as a network diagram, using line weight to represent the strength of a link. Self-links (loops) indicate the size of the membership retained from 2004 to 2005. This network is presented in Figure 4.

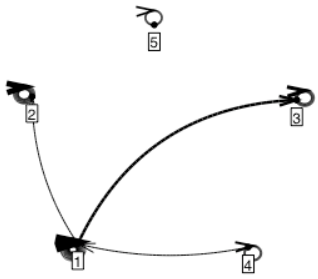

Figure 4. Exchange of members of five interest clusters from 2004 to 2005.

It is evident that most of the clusters have a fairly constant membership, although there is some movement from the largest cluster, general interests (1) into the second and third largest clusters. At the same time, there is some movement of membership from the fourth group into the first. This is because Cluster 1 is something of a grab-bag cluster, whose members lie reasonably close to the origin of the Principal Components of the interests. Hence, we see what might be a strengthening of at least certain interest clusters from 2004 to 2005, whose membership increases at the expense of the largest category of relatively undifferentiated interests.

In earlier work, we also relied on partitions into much larger numbers of interest clusters to develop interpretations [5]. However, in comparing the 2004 and 2005 cluster analyses, we noted a great deal of movement within most of the five clusters. For example, in the cluster containing sexual interests, we observed that most of the sub-clusters of sexual interests had completely re-organized between the 2004 and 2005 analyses. We consider it unlikely that the social or semantic categories of sexual (or other) interests has truly changed to this extent over the course of a single year, for this set of LiveJournal users. Rather, we propose that a user's nomination (or modification) of interests is subject to some random variation, which we observe in these intra-category shifts. Data from additional years would likely help in ascertaining the extent of this variation.

We conducted a similar analysis of the foaf:knows clusters, taking a cut that yielded six clusters of popular users. It is much harder to characterize the commonality among a cluster of users — what they represent is a group of weblogs that are visited and read by a similar set of users, so they cannot be interpreted as readily as lists of interests. However, it does appear that at least some of the clusters have a social coherence, for example, one cluster appears to be a network of goth/erotic models and photographers based in Canada and Virginia, and another appears to be a group of Spanish-speaking photographers.

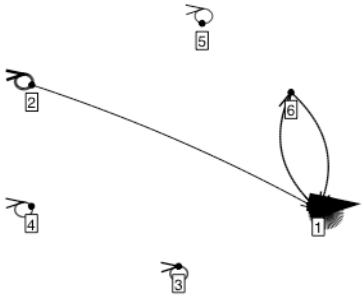

Figure 5. Exchange of members of six friends clusters from 2004 to 2005.

The 2004 and 2005 friends clusters were submitted to a network analysis in the manner of the preceding analysis for interests; this is represented in Figure 5. What is notable about Figure 4 is its relative stability. There is some accretion of popular users from the second-largest group into the largest group, but beyond this, all the groups but one largely maintain their membership. The only exception is Cluster 6, which has an apparent 100% turnover in membership. This is readily explained as an unstable cluster identification that is not distinct from Cluster 1. Had the cut been set at a higher threshold, only five clusters would have been found, with the membership of Cluster 6 for both years included in Cluster 1.

In addition to these two analyses, we made a number of attempts to illustrate the relations among the clusters of interests and clusters of friends, none of which produced a revealing set of patterns, regardless of the similarity measure used (e.g. Euclidean distance or cosine correlation). For example, a bi-modal network containing the six interest and friends clusters as the two node classes simply shows every interest cluster connected to every friends cluster at approximately the same level of strength. This is to be expected from the lack of correlation of interests and friends found in Figure 1, and in the greater concentration of interests around the origin.

4 Discussion

Among the clusters of popular users, there has been minimal change from 2004 to 2005. Among interests there is somewhat greater change, although still substantial stability, and the differentiation of the interest clusters appears to be strengthening somewhat. These results suggest that, with some caveats, it should be possible to identify clusters of users and interests that could lead to useful inferences for Semantic Web applications. For example, it is not hard to imagine a journal recommender system, or an interest or community weblog recommender system, based on the principles of analysis laid out in this research. This is in fact one way that recommender systems like those of Amazon.com operate.

At the same time, there is a great deal of re-organization of relations among interests at finer levels of distinction, suggesting that there is a limit to the efficacy of inferences we can draw from the clusters. For example, regardless of the strength of correlation between two interests, it would be incorrect to conclude that there is a necessary relation between them, or that in subsequent years the same relationship would obtain. This is important in two ways. First, in semantic domains where ontologies are not available or are unreliable, such as in the categorization of popular music, it will be necessary to supplement or replace logical modes of inference with something else. Statistical analyses of large, socially relevant user preferences is a promising source for additional information. Second, it is probable that other system developers will make use of clustering or related methods for "unsupervised learning" of Semantic Web ontologies or inference rules in these domains.

In either scenario, we depend upon statistical means of inference, which are probabilistic and subject to variation. Hence it is critical that we correctly estimate the expected variation in category assignments arising from different profile data sets. This requires careful consideration; while the literature on evaluating cluster analyses can be a helpful guide here, we also need to understand better the nature of the choice that goes into the editing of user profiles. These considerations are different for friends and interests, and lead to strikingly different distributions, even though they are represented identically in the user interface, in the metadata markup and in the incidence matrix used in our analysis.

Finally, since social and semantic relations do not correlate, it is not obvious that the semantics of interest clusters is emergent from the social relations experienced on LiveJournal. This may be because the 500 most popular friends are something like the local equivalent to celebrities. Users friending such LiveJournal celebrities need not expect to have a direct social relationship with them. Rather, their products (journal entries), like the cultural products of real-life celebrities, are passively consumed. On the other hand, this would suggest that popular users should pattern more like interests. The fact that they do not means that there is something different about these sorts of celebrities and the music celebrities that form the basis of many interest clusters. Hence the social dimensions of these user behaviors, and the semantics associated with them, need further study.

5 Conclusions

This study illustrates several ways in which the application of SNA methods to Semantic Web metadata holds much promise. First, we can use SNA methods to reveal highly structured relations among markup elements that are not themselves part of a controlled vocabulary or RDF vocabulary. Hence, we can use SNA and its statistical methods to extend the inference mechanisms already envisioned for the Semantic Web. In addition, by making diachronic observations, we can examine the extent to which naturally inter-correlated groups of interests or friends are stable over time. In this analysis, we discover distinct patterns of stability and flux for these two types of relation. Hence we recognize a need for caution in basing inferences on the clusters we discover this way, noting the need to develop a more complete understanding of the processes of nominating friends and interests.

References

[1] Passin, T (2004) *Explorers Guide to the Semantic Web*. Manning, Greenwich CT.

[2] Davies, J; Fensel, D; and van Harmelen, F, eds. (2003) *Towards the Semantic Web: Ontology-Driven Knowledge Management*. Wiley, New York.

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

[4] Ding, L; Zhou, L; and Finin, T (2005) FOAF Discovery: A Structural Perspective of the Semantic Web, *Proceedings of the 38th Hawaii International Conference on System Sciences*. IEEE Computer Society, Los Alamitos, California.

[5] Paolillo, J, and Wright, E (2005) Social network analysis on the Semantic Web: Techniques and Challenges for Visualizing FOAF. In A. Geroimenko and C. Chen, eds., *Visualizing the Semantic Web, Second Edition*. Springer, Berlin.

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

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

[8] Eckert, P (2000) *Linguistic Variation as Social Practice*. Blackwell, Oxford.

[9] Doreian, P; and Fujimoto, K (2002) Structures of Supreme Court Voting. *Connections 25(3)*.

[10] 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.

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

[12] Landauer, T; Foltz, P; and Laham, D (1998) Introduction to Latent Semantic Analysis. *Discourse Processes,* 25: 259-284.

[13] Bayer, A; Smart, J; McLaughlin G (1990) Mapping intellectual structure of a scientific subfield through author cocitations. *Journal of the American Society for Information Science*, 41(6): 444-452.

[14] White, H (1986) Cocited author retrieval. Information Technology and Libraries, 5(2):93-99.

[15] White, H; and Griffith, B (1981) Core journal networks and cocitation maps in the marine sciences: Tools for information management in interdisciplinary research. *Journal of the American Society for Information Science*, 32(3):163-171.
[16] Basilevsky, A (1994) *Statistical Factor Analysis and Related Methods: Theory and Application*. Wiley, New York.