2007-01-05

Implicit Affinity Networks

Matthew Scott Smith
Brigham Young University - Provo

IMPLICIT AFFINITY NETWORKS

by

Matthew Smith

A thesis submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computer Science

Brigham Young University

April 2007

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Matthew Smith

This thesis has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date	Christophe G. Giraud-Carrier, Chair
Date	William A. Barrett
Date	Scott N. Woodfield

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Matthew Smith in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date	Christophe G. Giraud-Carrier Chair, Graduate Committee
Accepted for the Department	Parris K. Egbert Graduate Coordinator
Accepted for the College	Thomas W. Sederberg Associate Dean, College of Physical and Mathematical Sciences

ABSTRACT

IMPLICIT AFFINITY NETWORKS

Matthew Smith

Department of Computer Science

Master of Science

Although they clearly exist, affinities among individuals are not all easily identified. Yet, they offer unique opportunities to discover new social networks, strengthen ties among individuals, and provide recommendations. We propose the idea of Implicit Affinity Networks (IANs) to build, visualize, and track affinities among groups of individuals. IANs are simple, interactive graphical representations that users may navigate to uncover interesting patterns. This thesis describes a system supporting the construction of IANs and evaluates it in the context of family history and online communities.

ACKNOWLEDGMENTS

I thank Dr. Christophe Giraud-Carrier for magnifying his role as my graduate advisor; he has been the perfect advisor for me. I would like to thank Brent Wenerstrom and Keith Copsey for many useful discussions. I also thank Matt Brown for important collaboration about online communities. Furthermore, I thank the Data Mining Lab members for diligently providing feedback and suggestions.

I am grateful for my family and wish to thank them for the trust, love, and support that they have always provided me. Specifically, I wish to thank my father, Dr. Craig C. Smith, for his willingness to discuss and provide feedback on many different thoughts. Lastly, I graciously thank my wonderful wife, Camille, for her patience and charity throughout the journey.

Contents

1 Introduction **1**

2 Related Work **5**

2.1 Network Models 5

2.2 Social Network Analysis Methods and Applications 7

2.3 Social Influence 8

2.4 Collaborative Filtering 8

2.5 Social Bookmarking by Tagging 9

3 Implicit Affinity Network - IAN **11**

3.1 Implicit Affinity 11

3.2 Affinity Scoring 13

3.3 Local User Weights 14

3.4 Learned Community Weights 15

3.5 Affinity Network Building 16

3.6 Affinity Network Filtering 19

3.6.1 N-Affinities 19

3.6.2 Affinity Strength Threshold 20

3.6.3 User-Centralized n-Clique 21

3.7 Affinity Network Social Capital 21

4 Implementation **27**

4.1 IAN, Online Community Overview 27

4.2 IAN, Online Community Details 31
4.3 GIAN, Genealogical-IAN 32

5 Experimental Results **35**
5.1 Showing Affinities 36
5.2 Tracking Changes 38
5.3 Discoveries 40
5.4 Qualitative Assessment 40
5.5 GIAN Results 44

6 Conclusion and Future Work **53**

Chapter 1

Introduction

Individuals are complex entities that may be described by rich sets of attributes and whose behavior is prone to change over time. Life's circumstances (e.g., marriage, retirement), age, geographical location, occupation and social interactions are all important factors of change in people's interests and behaviors. Given this rich and dynamic information, typically available about individuals, it is difficult to find and keep track of communities or groups of people sharing common characteristics.

Plato once observed that "similarity begets friendship" [30]. In recent years, modern sociologists have christened this notion *homophily* and have come to describe it with the popular phrase: "birds of a feather flock together" [25]. Underlying homophily is the claim by psychologists, such as Maslow, that *belonging* is one of humans' basic needs [24]. Indeed, evidence suggests that people exhibit a strong need to belong to groups — whatever they may be — such as families, clubs, teams, work groups, religious organizations, interest groups, gangs, or music bands. Work on the *small world phenomenon*, which shows that people tend to be connected to each other by short chains of social acquaintances, lends further support to the notion of homophily [26, 31, 44, 19, 43]. Adequate tools are necessary to facilitate the discovery and maintenance of homophilous communities.

Perhaps the simplest form of a homophilous community is that available to an individual through match-making companies (e.g., Match.com, LDSSingles.com). In this scenario, the individual is able to find others who match his/her profile based on

pre-defined and fixed criteria. The community is somewhat degenerate here, however, as only the requesting individual knows of it. Individuals whose profiles have been retrieved are unaware of the connection they may share with the requesting individual, unless they issue the same query against the system. Furthermore, the results are generally shown as a ranked list, ordered by matching score, which limits the discovery of novel associations that may be slightly out of one's explicitly stated criteria. Generally, matching systems are not concerned with underlying communities. Rather their focus is on providing each individual with a list of those most like them. There is no attempt at gaining an overall view of the database in terms of affinities among individuals.

A more elaborate form of a homophilous community is found in online communities. The Internet offers a new channel for people to reunite or connect. Geographically distant friends previously unable to communicate can now do so through the Internet. People with any interest, however rare it may be, have a greater opportunity to associate with others having the same interest. An online community is a group of people on the Internet sharing a common interest in a particular topic or aspect of life. New online communities are continually being launched. These communities focus on a wide range of themes including social life, business, sports, religion, research, technology, and news. Currently, some of the more popular online communities include Facebook.com (college and high-school students), Flickr.com (image sharing), Friendster.com (friend networks), LinkedIn.com (business networks), MySpace.com (music and teen networking), and Google Inc.'s Orkut.com (friend networks). In addition, there are many smaller communities geared toward more specialized groups (e.g., CarSpace.com, Joga.com, JMerica.com, LDSMissions.com). Many schools, churches, and researchers have also started online communities. These online communities promote member interaction, often by using forums (i.e., message boards), personal messages, blogs (i.e., web logs), or instant messages, as well as by providing contact

information (e.g., an email address or phone number). Underlying membership in such communities is the assumption that one shares the interest that is the explicit reason for existence within the community. Hence, community evolution is limited to individuals joining and/or leaving the community, and many other potential affinities among them are simply left unexploited.

Much theoretical work has also been done to understand homophilous communities in the area of social network analysis and modeling [41, 3, 39]. However, it again assumes a fixed and explicit set of affinities (often limited to a single relationship among entities, such as acquaintance).

Online communities can be enhanced or even created by identifying similar clusters of people (or social groups). For various reasons (e.g., people are multi-dimensional), there are usually unexploited affinities among people. By understanding the affinity networks within the community, creators could more easily meet user needs. In addition, by identifying these networks, community members could better integrate with similar individuals, and existing ties with others might also be strengthened. Furthermore, new communities could be organized.

In an online community setting, it is also interesting to understand, and adequately respond to, the evolution of the community. As new users join and current users change their profile, affinity networks can dynamically shift. In fact, entirely new sub-communities can develop. When a dynamic set of users has a dynamic set of characteristics, the interactions among them become increasingly challenging to analyze. In such an environment, understanding how the community evolves may help in such aspects as adapting to users' changing homophilous needs, as well as identifying influential individuals, social deviants, or trend setters.

In this thesis, we describe and evaluate a method for building implicit affinity networks (IAN) and tracking the evolution. The widespread use of Internet technology provides a unique environment for such a concept.

In addition to helping individuals with their need for identifying relevant communities, this work can prove useful in several other areas. For example, identifying the affinity networks that individuals belong to and how these networks evolve is relevant in the context of families and family history. Evidence suggests that we often do not know members of our families as well as we could, sometimes forget about them, and routinely miss opportunities to become closer to them. Discovering the affinity networks among our relatives (both dead and alive) would increase our sense of belonging, allow us to draw strength from others, become more united, and build stronger family ties. All family history researchers collect basic personal, generally event-related data, such as full name, gender, birth date and location, marriage information, etc. Many routinely gather additional nuggets of data, including occupation, physical traits, special achievements, etc. For the most part, the available information is used exclusively to identify individuals, almost independent of one another, except for obvious family relationships (e.g., child, spouse). Rarely, if ever, is the information used to derive — at least, systematically — possible affinities among individuals. We conjecture that this is not for lack of interest or desire, but rather for lack of adequate tools to analyze the data.

Although not its main focus, this work could also be useful in match-making situations. In terms of efficiency, by discovering communities, one could pre-compute a prototypical profile of each community, thereafter limiting all requested matches to prototypes rather than individuals. Furthermore, by knowing the existence and size of a particular community to which a user belongs, the system, instead of a top N list, could alert the user to the existence of a larger group of interest and display the information differently (e.g., via a graph).

Chapter 2

Related Work

This section discusses the work related to the research on implicit affinity networks (IAN) presented in this thesis.

2.1 Network Models

What do social networks look like? There has been significant research in developing network models that claim to answer this question [27]. This section will give a brief overview of three of these models. The "random graph" model by Erdös and Rényi [28], the "small-world" model by Watts and Strogatz [44], and the "scale-free" model by Barabási and Albert [4] will be outlined.

First, in 1959 the "random network" of Erdös and Rényi was described [28]. This was a simplistic approach developed to model large and seemingly random networks. A random graph is generated by starting with n vertices and then randomly adding edges between vertices with a given probability p.

Second, the "small-world" model by Watts and Strogatz was introduced in 1998 [44]. This model begins a network with n vertices, each being connected to its k neighbors ($k/2$ on each side), such that a regular ring lattice is created. This initial state of the network is designed to be locally clustered since each vertex is connected with its neighbors. Next, the network undergoes a "rewiring" procedure that moves an end of each edge, with probability p, to another vertex chosen uniformly

at random from all vertices, except with the constraint that no double edges or self-edges are allowed. This model extends the Erdös and Rényi model by producing networks ranging from a regular ring lattice (when $p = 0$) to a random network (when $p = 1$). For intermediate values of p, the graph is highly clustered like a regular graph, yet having a small average path length like random graphs. Since this model exhibits these properties it was named after the analogy of the *small-world phenomenon* [26, 18, 31].

Finally, the "scale-free" model of Barabási and Albert was proposed in 1999 [4]. This model generates networks by adding vertices, each having degree m, one at a time. As each vertex is added, each of the m edges is connected to a vertex present in the graph with probability proportional to the degree of that vertex. Barabási and Albert called this way of connecting edges *preferential attachment*, since a new vertex has a greater probability of linking to vertices that have high degree. This property is often exhibited in many real-world networks, sometimes described as the "rich get richer." Additionally, this model favors linking to the vertices that have been in the model for longer periods of time. Figure 2.1 is an example of a "scale-free" network. This model produces networks that have few vertices with many edges and many vertices with few edges, thus producing a power-law degree distribution.

Although network modeling is related and useful to our work it is not the motivation. By understanding the structure and function of complex networks, however, the network can be better utilized. For example, the link structure of the Internet has been identified as a "scale-free" network where a few sites (vertices), like Google and Microsoft, have many links (edges) whereas most sites have very few links. "Scale-free" networks are extremely tolerant to random failures, but can be greatly disrupted by targeted attacks. That is, if Google were to be taken down it would have an immediate impact on much of the Internet community. Whereas, if a small less-known site were to go down, then very few people would be affected. Preliminary results

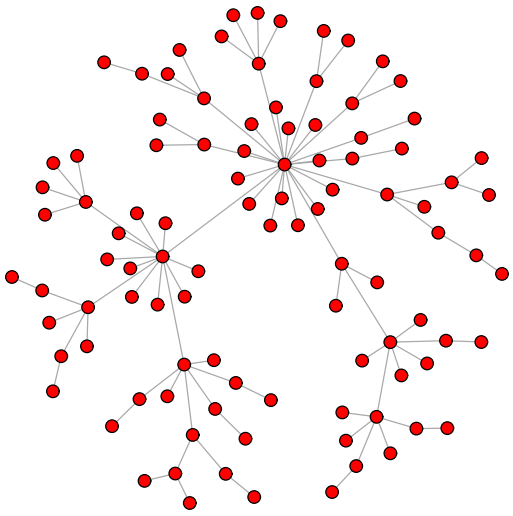

Figure 2.1: **Barabási and Albert Network**. Network data generated using the "scale-free" model (by adding 99 vertices with $m = 1$ to the network of a single vertex)

suggest that implicit affinity networks are not adequately described by any of the above models.

2.2 Social Network Analysis Methods and Applications

Social network analysis (SNA), which focuses on modeling the relationships among individuals, is highly relevant to the proposed research [41]. The data typically used for most SNA studies is somewhat different from that which we propose to use. The typical relationships used in SNA studies are measured (or collected) using an assortment of techniques including questions like "name those people with whom you are friends." The links represented in these studies are explicit relationships, whereas the links in the proposed method are based strictly on inherent similarities among the individuals, which are implicit relationships. The large body of SNA research formally defines representations for networks and provides tools for network analysis

[41, 10, 38, 45]. This notation and set of tools will be used where applicable in our work.

2.3 Social Influence

The study of social influence links the structure of social relations to attitudes and behaviors of the actors within the network [42]. Recent studies have shown that certain people have an increased ability to shape public opinion [9]. This fuels the motivation to identify such individuals as they are targets for word-of-mouth or viral marketing campaigns [37]. Additionally, it is clear that people are more influenced by those they trust rather than those they know nothing about [35]. The new perspective of viewing customers within IANs may contribute to this area of research. The relationship between social influence and IANs is reserved for future work. Furthermore, the linkage between implicit affinities and influence has yet to be shown.

2.4 Collaborative Filtering

The objective of collaborative filtering (CF) is to recommend new items for a particular user based on the user's previous opinions and the opinions of similar users (i.e., users with similar tastes) [36, 20]. A CF algorithm typically has a list of m users $U = \{u_1, u_2, ..., u_m\}$ and a list of n items $I = \{i_1, i_2, ..., i_n\}$. Each user u_i has a list of items I_{u_i}, which the user has expressed an opinion about. Opinions can be explicitly given by a user through rating a particular item or they can be implicitly derived from actions performed by the user (e.g., buying a product, listening to a song, etc.). Thus, a CF algorithm can provide a *Top-N recommendation* for a user that shares tastes with other users [8]. Although not the main motivation of this work, IANs can be used for CF as they can provide a *Top-N recommendation* for a particular individual x by recommending the n neighboring individuals having the

strongest affinities. This approach recommends individuals having similar attributes as individual x.

2.5 Social Bookmarking by Tagging

Social bookmarking, a recent phenomenon on the Internet, has become increasingly popular [12, 23]. Social bookmarking allows people to share resources on the Internet through the practice of "tagging". A tag is a keyword or category label that is used to help people find resources in common. For example, on a tagging-based website, each user can tag, or label, resources (i.e., websites, images, songs) using the keywords of his or her choice. Each resource may typically be given an arbitrary number of tags to identify it. This can be nice for one individual to bookmark a resource for his or her easy reference, however, the real advantage comes through combining the collaborative efforts of everyone. Users may find other resources having tags that match their interests, thus communities of people sharing similar interests are discovered. There is a growing number of websites that have been built upon this idea (see Del.icio.us [7], Flickr [13], Technorati [40]). Additionally, it has been integrated into e-commerce sites including Amazon.com and Yahoo.com.

Our methodology can be applied to this growing body of tagged resources. Resources are mapped as nodes and similar tags as edges. Attributes, in our context, are richer than a tag as they consist of one or more values grouped under a semantic label, while tags are a single value with no defined semantics. Moreover, our approach extends traditional tagging by presenting a visual representation of the network created.

Chapter 3

Implicit Affinity Network - IAN

Due to the enormous amount of information available about individuals, it is increasingly challenging to exploit all of the inherent similarities, or affinities, that tie them together. Implicit Affinity Networks (IAN) are used to facilitate knowledge assimilation, uncover, and track relationships among groups of individuals.

Just as the *Apriori* algorithm only makes explicit associations that are implicit in customer transactions [1, 2], IANs make explicit affinities that are implicit about individuals. In both cases, the sheer volume of data makes it impossible for the user to "see" the associations/affinities. The algorithm is able to bring them out, but the element of discovery or interestingness is left to the user.

3.1 Implicit Affinity

Let I be an arbitrary set of individuals, such that the ith individual is represented by I_i where $i = \{1, \ldots, n\}$. Let A be an arbitrary set of attributes, such that A_j^i is the jth attribute of the ith individual, where $j = \{1, \ldots, m_i\}$. Let V_j be the arbitrary set of values for attribute A_j, such that V_{jk}^i is the kth value of attribute A_j for individual I_i, where $k = \{1, \ldots, p_{ij}\}$.

In practice, A_j represents some piece of information about individuals, e.g., name, month of birth, occupation, hobbies, etc. V_j, then is the set of all values for attribute A_j. For example, the "month of birth" attribute might have the value set

Individual	Attributes
John Smith	occupation: {teacher, carpenter} hobbies: {chess, gardening, rock climbing}
Jill Jones	occupation: {teacher} hobbies: {gardening, surfing}

Table 3.1:

$\{January, February, \ldots, December\}$; the attribute "hobbies" might have the value set $\{scrapbooking, stamp\ collecting, reading, hiking, \ldots\}$; etc.

In our context, individuals may be characterized by any number of attributes and each attribute may have any number of possible values. In terms of a database, this means that a normalized, dynamic schema should be used.

We define an *affinity* between two individuals as the overlapping of attribute-values for any common attribute. For example, the individuals John Smith and Jill Jones, from Table 3.1, share a couple of affinities: they are both teachers and they both enjoy gardening. The affinity is *implicit* because the connection made between the two individuals is not explicitly defined as in typical social networks. Computing affinities consists of comparing attribute-values for attributes across individuals using some similarity metric. Common similarity metrics include exact match, Euclidean distance, soundex, metaphone [29], levenshtein [21], jaro-winkler [15, 16], jaccard [14], and stemming [32]. A description of these metrics is outside the scope of this thesis. The reader is referred to the relevant literature for details. Metrics generally depend on the nature of the values being compared (e.g., nominal, real, string). In the case of strings, metrics vary significantly in how they account for similarity. For example, a soundex comparison over names such as "Joan" and "John" would return a high similarity score, not because the two names are identical but because they sound the same. It follows that the choice of similarity metrics has an impact on the nature of the implied affinity. For the studies performed in this research we chose to use the

strictest measure, an exact match, as the similarity metric, leaving the performance of other similarity metrics to future work.

3.2 Affinity Scoring

The raw similarity score between two individuals I_1 and I_2 on some attribute A_j is computed by counting the number of values of A_j that I_1 and I_2 share, where, again, sharing is taken with respect to the chosen similarity function. Note that $\mid S \mid$ is the cardinality or size of set S.

$$raw\_score_{A_j}(I_X, I_Y) = \mid V_j^X \bigcap V_j^Y \mid \tag{3.1}$$

Now, since the number of values of A_j in I_X and I_Y may differ, it is desirable to normalize the score such that individuals having the greatest number of attribute-values do not come to dominate the underlying affinity networks. Therefore, for a particular attribute, individuals are most strongly connected to those whose value set is most similar in ratio, rather than just counting the number of values in common. Thus, the following affinity score, incorporates this consideration:

$$score_{A_j}(I_X, I_Y) = \frac{\mid V_j^X \bigcap V_j^Y \mid}{\mid V_j^X \bigcup V_j^Y \mid} \tag{3.2}$$

As discussed above, individuals may be characterized by any number of attributes. In practice, the set of attributes that an individual is defined by will be determined by the individual, either through explicit specification via an interactive interface or implicitly by different sources of information about the individual (e.g., personal web pages). In any case, this set is biased by what is available about an individual at a given time and it changes over time. As a result, the communities we build exhibit rather dynamic affinities, since changes may be induced not only by addition/removal of individuals — which is typical of all current work on communities

— but also by addition/removal of attributes and attribute-values at the individual level.

We wish to accommodate both types of changes. Global, learned, community weights are used to let affinities evolve with changes in the behavior of individuals. Local, fixed, user weights are used to tailor affinities to specific user interests. Note that by user, we mean anyone interested in the community structure and behavior (e.g., an online community member, the "owner" of family history data, the manager of an online community, the director of an eCommerce site) in the sense that we are not just concerned about telling a given user about his/her own affinities — which is typically the intent of matching systems — but rather providing new insights about the community.

These mechanisms, namely a combined user weight ($\Delta_{A_j}^{X,Y}$) and community weight (w_{A_j}), defined and discussed in the following sections, are incorporated into the product to yield a weighted score:

$$scorew_{A_j}(I_X, I_Y) = score_{A_j}(I_X, I_Y) \cdot \Delta_{A_j}^{X,Y} \cdot w_{A_j} \tag{3.3}$$

The overall score across multiple attributes is then obtained with:

$$overall\_score(I_X, I_Y) = \frac{1}{\mid A^X \bigcap A^Y \mid} \sum_{A_j \in A^X \bigcap A^Y} scorew_{A_j}(I_X, I_Y) \tag{3.4}$$

3.3 Local User Weights

Local weights, specified by the user, allow the network to be tailored to the user's indicated interests (e.g., discovering name patterns in pedigree data, identifying clusters of individuals along well-defined characteristics); they increase the capacity of the user to indicate what is most important to him or her. Formally, they are represented

by $\delta_{A_j}^X$ where X denotes the individual and A_j denotes the attribute. The combined effect that the user weights have upon an affinity is defined as follows:

$$\Delta_{A_j}^{X,Y} = \frac{\delta_{A_j}^X + \delta_{A_j}^Y}{2} \tag{3.5}$$

The combined user weight, Δ, is set as the average weight of the compared individuals; it can also be set otherwise. Specifically, when the interaction between two high user weights is much more significant than two medium weights, this can be easily captured. In this case, the combined effect that the user weights have upon an affinity is multiplicative, rather than additive, as follows:

$$\Delta_{A_j}^{X,Y} = \delta_{A_j}^X \cdot \delta_{A_j}^Y \tag{3.6}$$

Eliciting weights from the user is somewhat obtrusive. However, local weights are particularly beneficial in an online community setting where users may have different goals, or as the focus of attention may change (yet be user-directed) from time to time. Note that if a user weight is not specified for a particular attribute, then it is simply equal to one and does not affect the affinity score (see Equation 3.4).

3.4 Learned Community Weights

As a technique to automatically present relevant affinities among entities, community weights can be used to highlight affinities. Community weights are dynamically learned and change as the community evolves. Two types of community weightings that apply to individual attributes are presented.

The first method assigns more weight to attributes used frequently within the community, without regard to the values of the attribute, only verifying that at least one exists. This weighting is rooted in the assumption that if an individual has something about a particular attribute recorded, then this data has some degree of

interest, otherwise it would not be recorded. It can also be used in the case that the data recorded is more interesting than the data not recorded. Remember, I represents the set of all individuals within the community. The weight learned for attribute A_j is given by:

$$w_{A_j} = \frac{\mid \{I_i \in I : V_j^i \neq \varnothing\} \mid}{\mid I \mid} \tag{3.7}$$

In other words, the weighting for attribute A_j is the number of individuals that have at least one value for attribute A_j divided by the total number of individuals in the community. This attribute weighting favors commonly used attributes.

The second method, a variation on the first one, reverses the community weights to favor less-common attributes within the community. The interest of this weighting mechanism is to allow the detection of rare or eccentric affinities in the community. The weight of attribute A_j, in this case, is given by:

$$w^*_{A_j} = 1 - w_{A_j} \tag{3.8}$$

3.5 Affinity Network Building

Affinities among groups of individuals can be used to build a network.

Consider the sample data in Table 3.2. For the sake of simplicity, attributes are abstracted as capital letters (i.e., A and B), lowercase letters are used to represent specific attribute-values (e.g., a_1, b_3, etc.). In this instance, A could represent the *hobbies* attribute while a_1 could be a value, such as *gardening*.

Through pairwise comparisons of all individuals, the score between individuals for each attribute can be found and stored into matrix form. Table 3.3 shows the matrix corresponding to the individuals of Table 3.2.

Individual	Attributes
Jim	A: $\{a_1, a_2, a_3\}$
Sarah	A: $\{a_1, a_2\}$
Mary	A: $\{a_3\}$
Bob	A: $\{a_1, a_2\}$ B: $\{b_1, b_2\}$
Susan	B: $\{b_1, b_2, b_3\}$
Brent	B: $\{b_1, b_2, b_3\}$

Table 3.2: Sample of Individuals and their Attributes

Sarah										
1	0	**Bob**								
$\frac{2}{3}$	0	$\frac{2}{3}$	0	**Jim**						
0	0	0	0	$\frac{1}{3}$	0	**Mary**				
0	0	0	$\frac{2}{3}$	0	0	0	0	**Susan**		
0	0	0	$\frac{2}{3}$	0	0	0	0	0	1	**Brent**
A	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	

Table 3.3: Total Affinity Matrix for Individuals in Table 3.2

The similarity matrix, in turn, can be represented as a weighted graph or network, where nodes are individuals and edges are affinities. Figure 3.1 is the implicit affinity network corresponding to the above matrix for the individuals of Table 3.2. The weight of each edge denoted by the relative thickness of the line, is the score of the affinity. It is used to indicate how strong the affinity is relative to the other affinities within the graph. For display purposes, edge widths are scaled so that the full spectrum for viewing graph edges is used. The ideal spectrum varies depending on the graphing package being used. For example, the desired maximum edge width ranges from 0 to 20 for the igraph package in R [6]. Since affinity scores range from 0 to 1 they would not directly produce ideal edge widths. Therefore the edge widths are derived from amplifying the affinity scores by dividing each score by the maximum affinity score within the entire graph (or graphs) where ω is the desired maximum edge width, as follows:

$$edge\_width = \omega \cdot \frac{score}{MAX(scores)} \tag{3.9}$$

The strongest affinity within the graph then becomes the thickest line. This, in essence, allows the affinity strengths to be most distinguishable on a particular graph, provided that ω is chosen properly for the graphing package used. Note, that when comparing graphs one to another, then scaling must be standardized. That is, the maximum affinity score must be the global maximum for all of the graphs being compared.

Furthermore, the affinity attribute is signified by the edge color. In this example, red is used to indicate the hobby affinities and blue is used to indicate the occupation affinities.

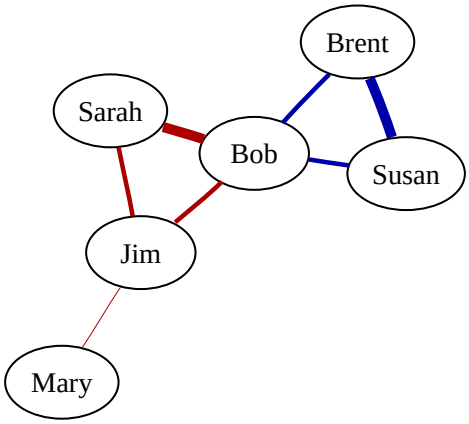

Figure 3.1: Implicit Affinity Network (for Individuals in Table 3.2)

An affinity network provides an intuitive graphical mechanism for discovering how various individuals are connected through inherent similarities. For instance, in Figure 3.1, one readily sees that Bob is directly connected with everyone except Mary, indicating that Bob has affinities with Sarah, Susan, Brent, and Jim. The network also shows that Bob's affinity with Sarah is stronger than with Jim (indicated by a thicker line), even though all three of them share the same attribute-values (i.e., $a1$

and $a2$). Jim also has $a3$, which causes him to be less connected to Sarah and Bob (see Equation 3.2), yet it enables him to be connected to Mary.

Note that one need not consider all attributes when building an affinity network. Indeed, it is possible to restrict the analysis to any subset of attributes, so that the resulting network can be specialized to only certain affinities selected by the user.

3.6 Affinity Network Filtering

The equations in the previous section are all that is needed to create an IAN. However, networks quickly become large and unmanageable. Thus, further techniques are needed to hone the network so that relevant aspects of IAN are presented. For example, if the IAN is to be beneficial for a single user, then a local view of how that user is related to the community is applicable. If the IAN is to describe the community as a whole, for a community overseer, then a more global IAN encompassing the most influential individuals and the most important affinities might be appropriate. The techniques presented need not be used exclusively, that is, they can be used together as desired.

3.6.1 *N*-Affinities

The network can be limited to show only the strongest N affinities, or edges within the graph. This technique filters the network, such that it is guaranteed to have no more than N edges and at most $2N$ individuals. Thus, intuitively bounding the size of the network. Figure 3.2 shows how the IAN in Figure 3.1 is affected when viewing the strongest 2-affinities. Note, that affinity strength ties are broken by recency (i.e., the newest affinities are shown).

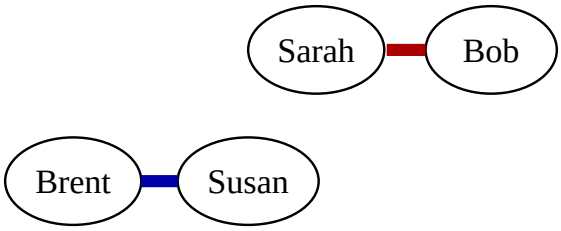

Figure 3.2: IAN, Strongest 2-Affinities (for Individuals in Table 3.2)

3.6.2 Affinity Strength Threshold

A network can be limited by showing only affinities that are above a designated threshold T. This is good if you would like to see an IAN that only shows affinities that have a specified strength. A downside to using this method alone is that the size of the network it produces is unknown, unless the user knows the size of the community and has some intuition about the threshold strength. In practice, a knob or slider that ranges through the threshold range could be used to aid the user. Figure 3.3 shows how the IAN in Figure 3.1 is affected when viewing it with an affinity strength threshold of 0.5.

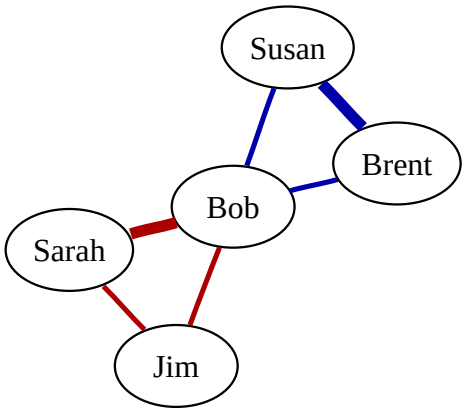

Figure 3.3: IAN, Threshold=0.5 (for Individuals in Table 3.2)

3.6.3 User-Centralized n-Clique

To produce an IAN centralized on a particular user, one method is to limit the individuals in the network to those reachable within the geodesic distance of n from the user. Note that the geodesic distance is defined as the length of a shortest path between the two nodes. The nodes in this fashion form a clique. Again, this method is useful to present an IAN from the localized perspective of the user. Also, as noted earlier, this method need not be used exclusively.

3.7 Affinity Network Social Capital

Social capital is a fundamental idea in numerous research areas including business, organizational behavior, political science, and sociology. It is explained as the advantage available through the connections between individuals within a particular network. It has been used to explain how certain individuals obtain more success through using their connections with other people. Social capital fosters reciprocity, coordination, communication, and collaboration. Harvard researcher Robert D. Putnam describes two main components of social capital, namely, bonding and bridging [33, 34].

Bonding capital refers to the network value assigned to social networks between homogeneous groups of people, whereas bridging capital refers to the network value assigned to social networks between heterogeneous groups of people. These two components of social capital are used as a basis for our measures of network strength. First, bonding strength is defined by the following, where E is the set of edges:

$$bonding\_strength = \frac{\sum_{\{I_X, I_Y\} \in E} overall\_score(I_X, I_Y)}{\mid I \mid \cdot \frac{(\mid I \mid -1)}{2}} \tag{3.10}$$

This measure explains how connected the individuals are within the network. On the other hand, bridging strength is defined, simply, as the inverse of bonding strength.

$$bridging\_strength = 1 - bonding\_strength \tag{3.11}$$

Bridging strength is how disconnected the individuals are within the network. In some sense, it is a measure of diversity within the network, potentially suggesting how individuals can further connect.

These measures of network strength are used to track the amount of bridging and bonding capital that exists within a given IAN. These metrics are based on the connectivity of the network. For example, bonding strength ranges from zero to one, or disconnected to fully-connected. A fully-connected IAN implies that every individual in the network is connected to every other individual by all possible affinities, each having maximum strength. A completely disconnected IAN implies that the individuals within the network have no affinities. For large real world networks, achieving a bonding strength of one is unlikely. However, when it occurs all profiles are perfectly homogeneous and completely bonded.

The network strength metrics calculated through time account for a dynamic set of individuals. Furthermore, they incorporate the affinity score that accounts for a dynamic set of attributes, and values (see Equation 3.4).

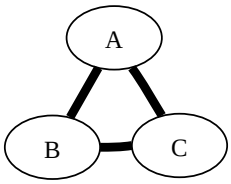

Figure 3.4: Bonding Strength: 1.00, Bridging Strength: 0.00 (t=1)

To gain some intuition about network strength a few simple examples will be shown. Initially, we will assume that the individuals share one attribute-value for

a single attribute (see Figure 3.4). In this case, the network is fully-connected thus having a bonding strength of one. That is, every individual in the network shares every possible affinity.

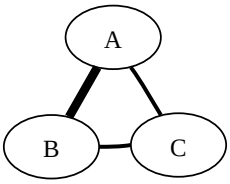

Figure 3.5: Bonding Strength: 0.67 ⇓ - Bridging Strength: 0.33 ⇑ (t=2)

As time passes, suppose that individual C, attempting to bridge out, makes it known that she has an additional attribute-value. At this point, the bonding strength has decreased (see Figure 3.5), due to the fact that individuals A and B are no longer fully-connected with individual C. Furthermore, bridging strength increases, suggesting that A and B now have the possibility of connecting with C on the new attribute-value.

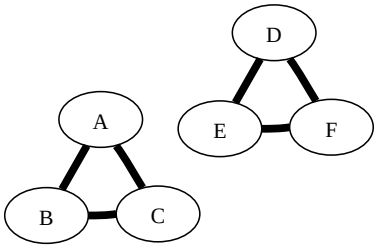

Figure 3.6: Bonding Strength: 0.60 ⇓ - Bridging Strength: 0.40 ⇑ (t=3)

Now, suppose that C makes it known that she no longer has the attribute-value mentioned previously. Additionally, three new individuals join the community, D, E, and F, sharing an attribute-value different than the one that is shared by A, B, and C. Thus, the bonding strength within the network diminishes as the bridging strength increases. Of course, this occurs because there is increased possibility for bridging.

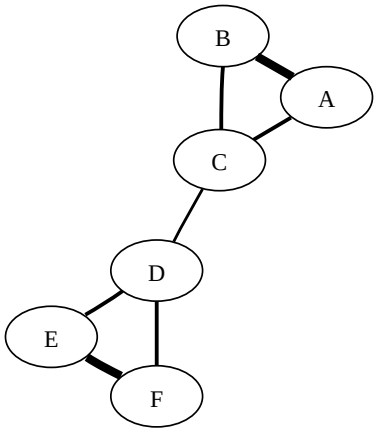

Figure 3.7: Bonding Strength: 0.43 ⇓ - Bridging Strength: 0.57 ⇑ (t=4)

After another time interval, suppose that C attempts to bridge out again, making it known that she has an additional attribute-value. Furthermore, D manifests the same additional attribute-value, thus an affinity is implied connecting the two (see Figure 3.7). Interestingly and appropriately, the bonding strength decreases as the bridging increases. This happens due to the fact that C and D have added a new attribute-value that connects only with each other and not the majority of the community. Therefore, the affinity strengths of A and B to C decrease, as do the affinity strengths from E and F to D. Thus, the bonding strength of the community as a whole decreases.

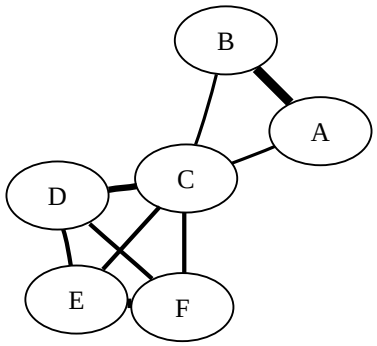

Figure 3.8: Bonding Strength: 0.47 ⇑ - Bridging Strength: 0.53 ⇓ (t=5)

Lastly, after another time interval, C adds the attribute-value that has been shared by D, E, and F, thus establishing an affinity with each of these individuals

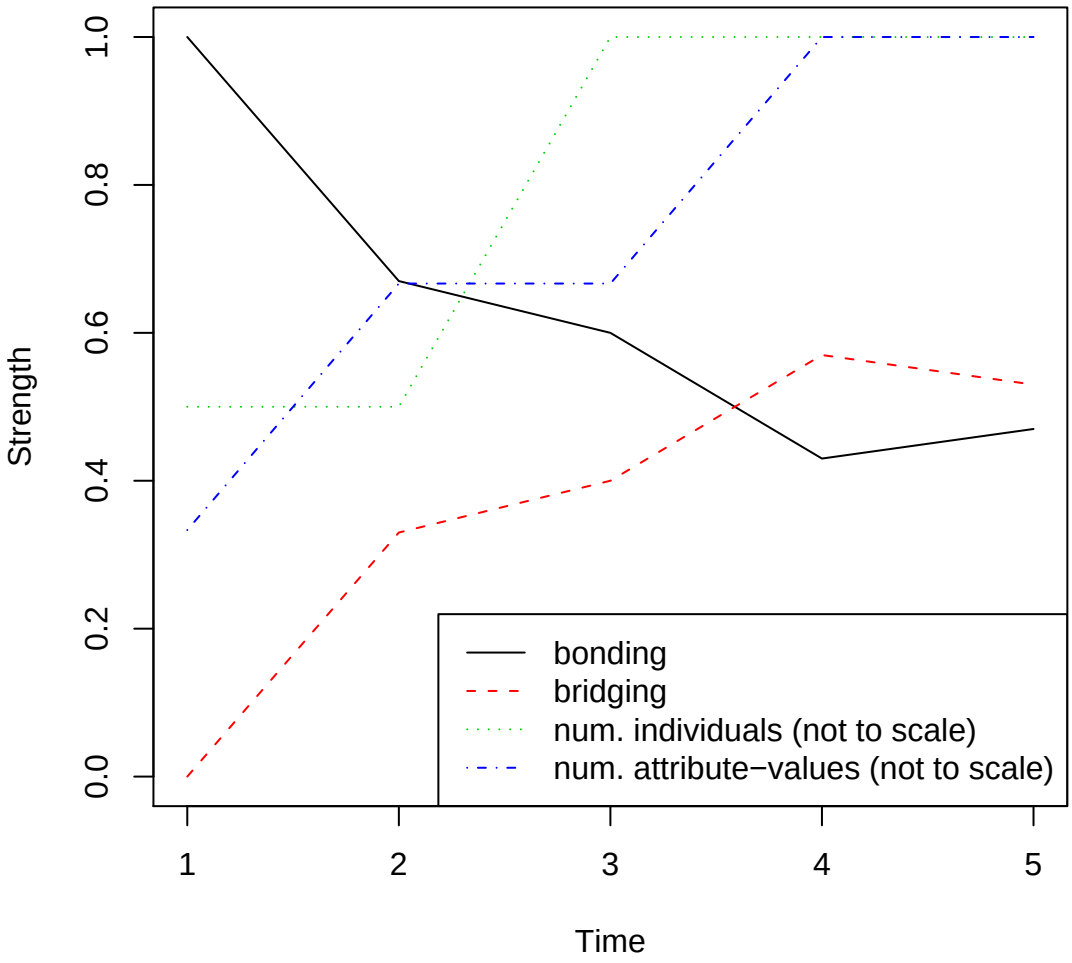

Figure 3.9: Network Strength Evolution

(see Figure 3.8). Finally, the bonding strength of the network increases, since the increase in bonding between C and [D, E, and F] is greater than the decrease in bonding between C and [A and B] and, of course, bridging decreases.

The network strength metrics of bridging and bonding can be plotted over time to show the evolution of any particular network. For example, the evolution of the network given in this example is shown in Figure 3.9. The change in network strength metrics indicate when the network is bridging or bonding. Additionally, the number

of individuals and number of attribute-values are plotted for comparing the relative change with respect to the network strength. However, as noted in the legend, these two lines are not to scale. In general, as the number of individuals and/or the number of attribute-values increase the bonding strength goes down. Bridging strength, of course, comes up suggesting the increased possibility for individuals to bridge. Thus, the network strength metrics provide a nice way to view community evolution through time.

Chapter 4

Implementation

To test the methodology of IAN an online community was implemented as a Web application. Furthermore, a specialized implementation was created to cater to the genealogical domain.

4.1 IAN, Online Community Overview

The online community, also named IAN after this research, is the experimental testing grounds for analyzing community behavior and presenting IANs to users (Figure 4.1). In current form, it is a general community that enables sub-communities to emerge over a variety of topics. It is invitation-based; members can invite family and friends, allowing a diverse set of individuals to provide real data for affinity network discovery.

As described in Section 3.1, individuals describe themselves using an arbitrary number of attributes having an arbitrary number of values, thus creating a user profile (Figure 4.2). Users can view all of the affinity networks occurring within the community (see for example, Figure 4.3). The community evolves as individuals change their profile or as users join or leave the community. Concurrently, the affinity networks automatically adapt to these changes and are tracked. The network strength metrics can be used to measure the bonding and bridging of any sub-community or even the community as a whole. Additionally, visual snapshots of the network graphs

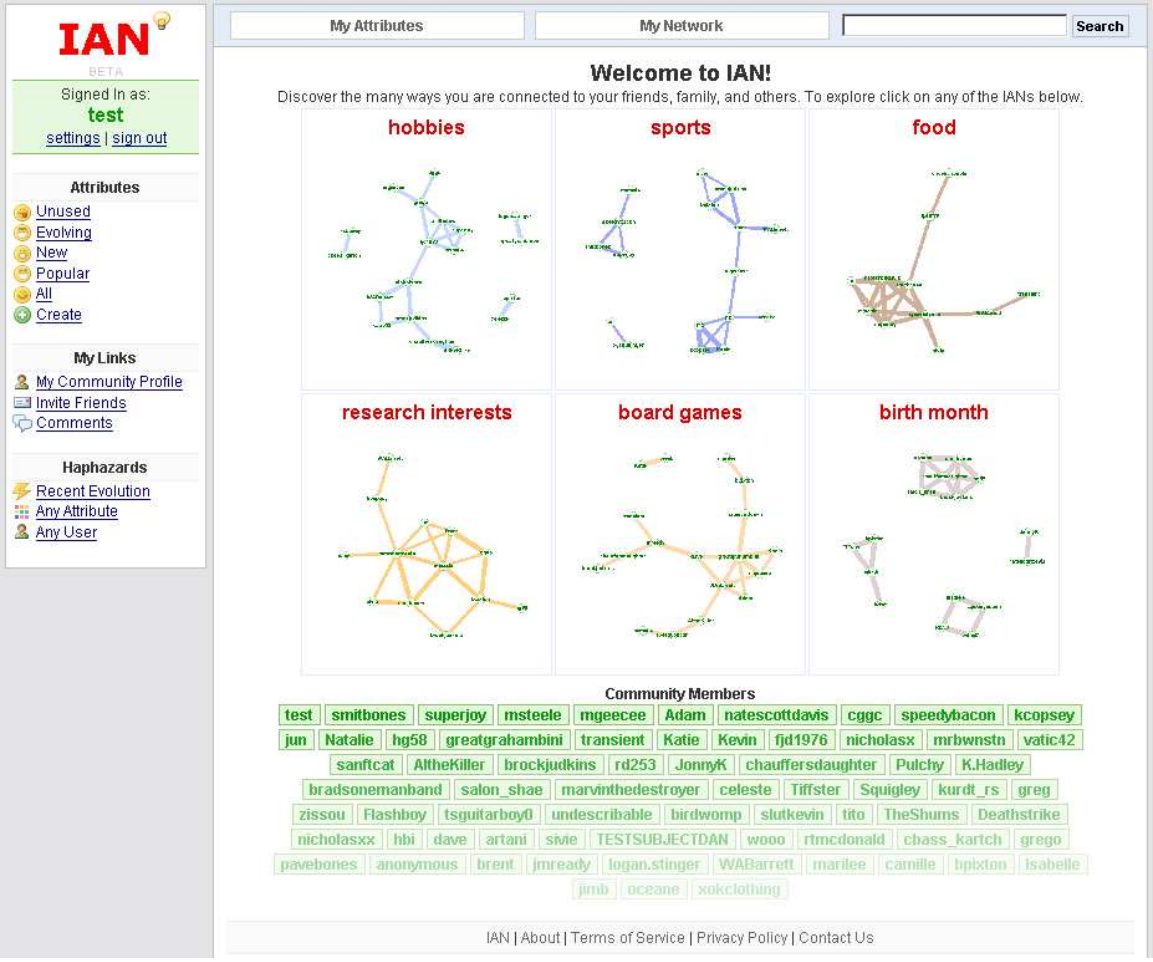

Figure 4.1: IAN Community Website Screenshot

can be taken through time, allowing general trends among the community to be discovered, as well as individuals migrating through subnetworks.

The system is designed to present relevant and useful IANs. It can be used to track the affinity networks that individuals cluster in through time. It can be used to discover new sub-communities where the affinity networks among individuals are previously unknown. It might also be particularly useful for reunions (i.e., family, high school, etc.) to gauge how people are connected and in what dimensions. Uncovering the most relevant affinity networks can be used to direct community interaction and make community decisions. There is personal value for community members through discovering how they are connected within the community, enabling members to in-

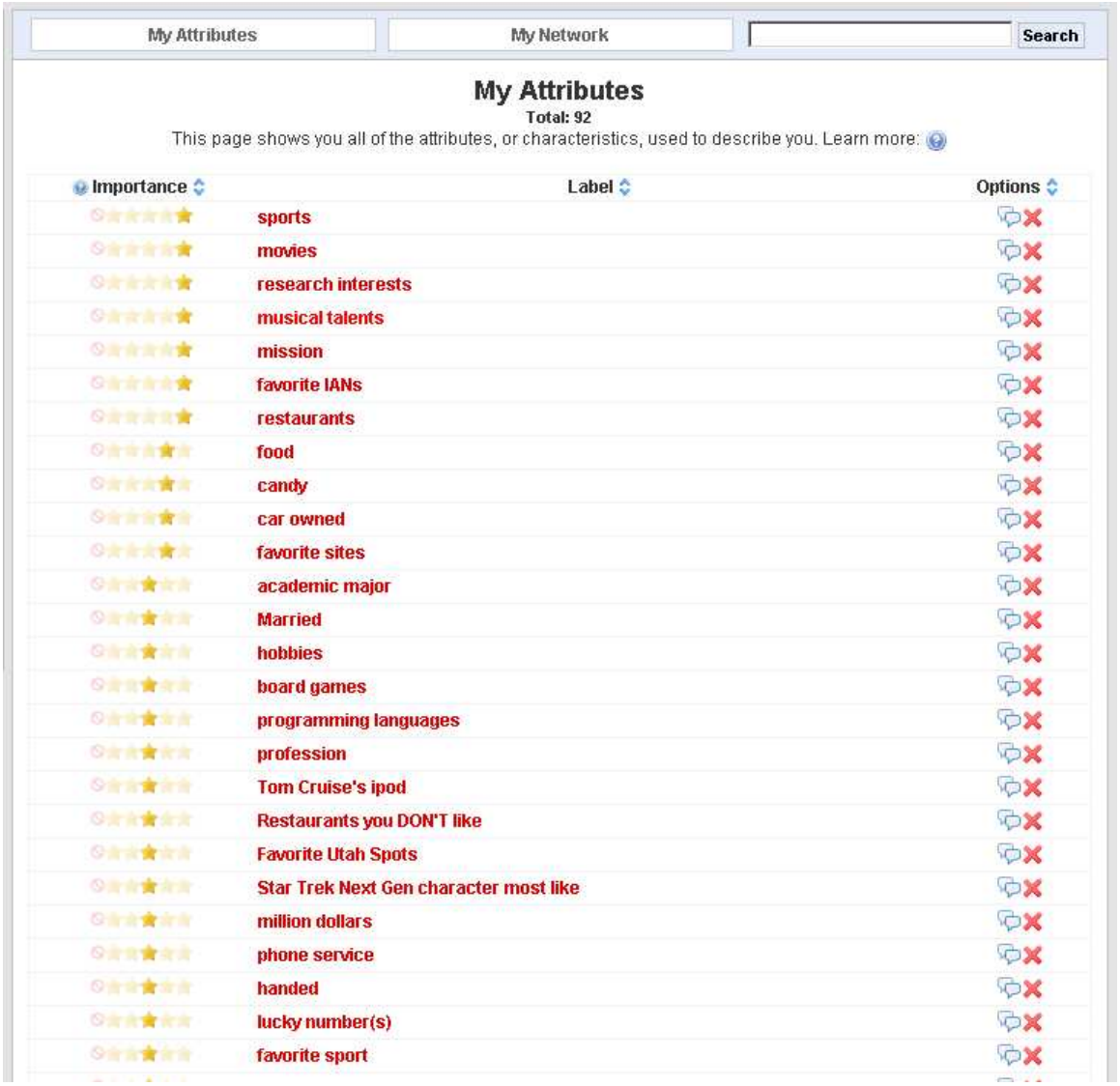

Figure 4.2: IAN Community Website - Attributes

teract on mutual interests. It also allows members to visualize where they are less connected within the community, possibly suggesting personal limitations or weaknesses. IAN community evolution is fundamentally different from that of a traditional Internet community.

One key difference is that the community is driven primarily by members rather than managers. This is accomplished in part because members can describe for themselves how they wish to be described, using the attributes they choose to use or create, rather than a static set of attributes set forth by community managers (see

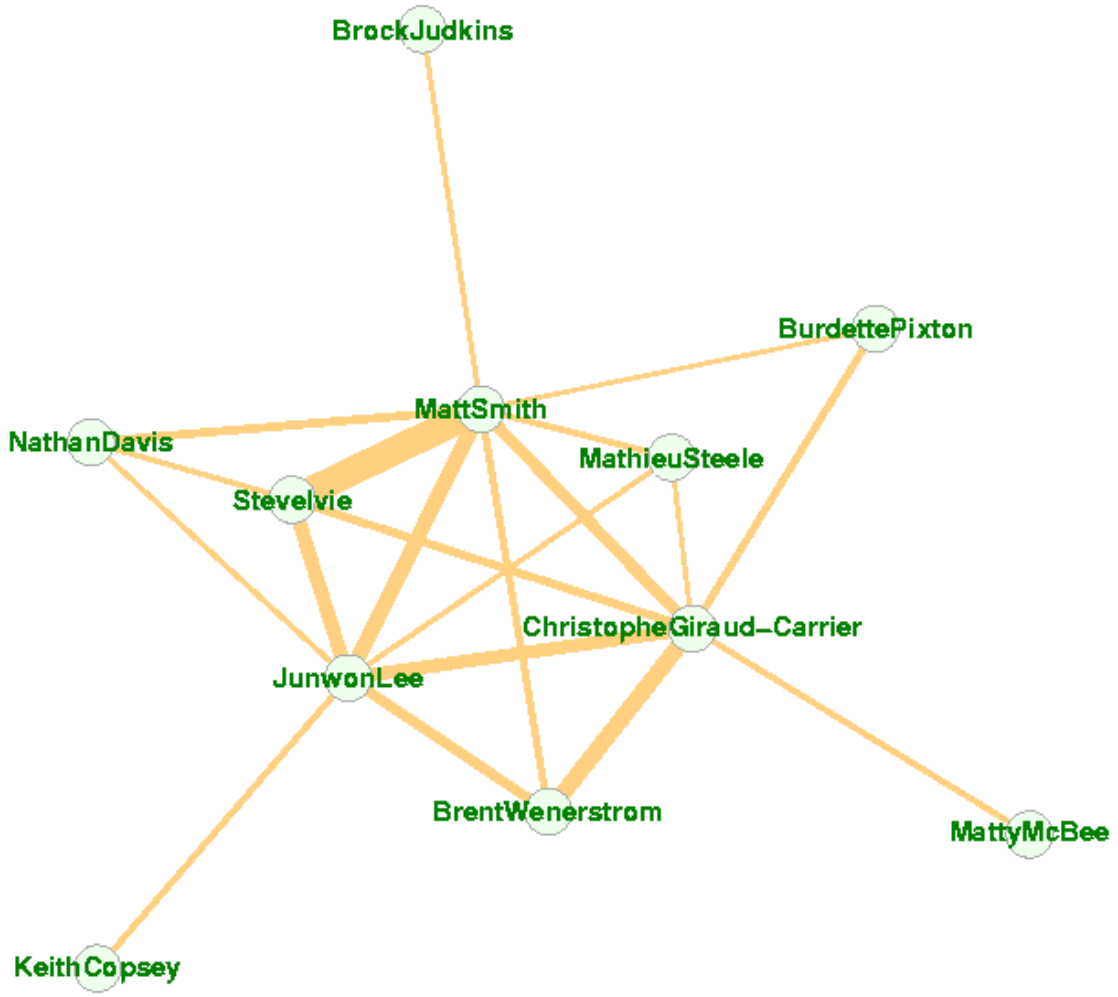

Figure 4.3: IAN Community Website - Research Interests IAN (N:25)

Figures 4.4 and 4.5). Furthermore, any change that a member makes affects all other members to some degree (see Section 3.4), particularly those in overlapping networks. These subtle differences allow for a variety of observable interactions that otherwise would not be possible. For example, a trend setter could be identified through the interaction that he/she makes by creating a new attribute which many other users subsequently use. Also, sets of attributes that certain clusters of individuals tend to use can be discovered, thus exposing community cliques (e.g., see Figure 4.3). The website makes it possible to discover and utilize the evolving affinity networks that occur over time.

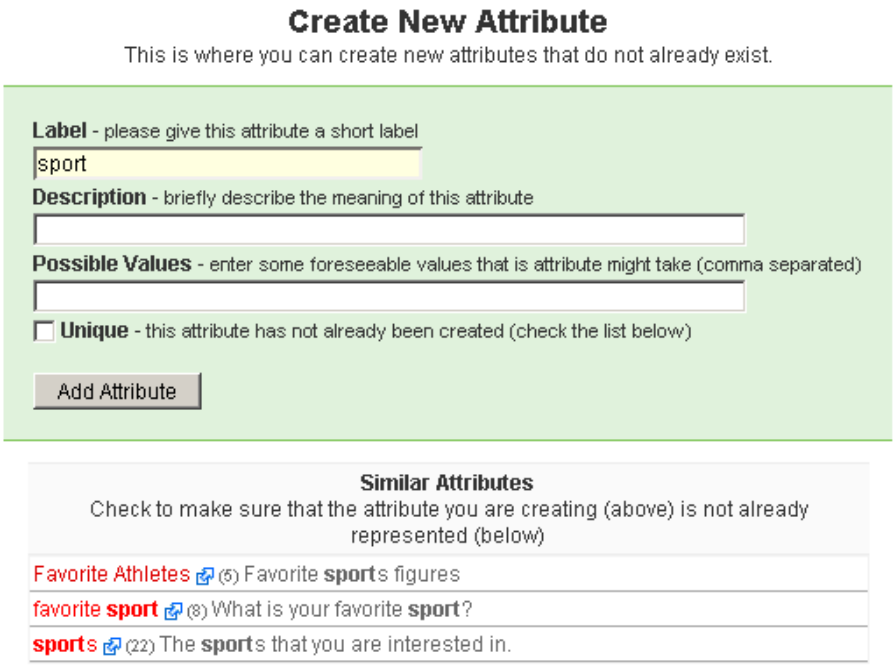

Figure 4.4: IAN Create New Attribute

The current version of the site is available online at:

`http://dml.cs.byu.edu/IAN/`

4.2 IAN, Online Community Details

IAN, the community website, is developed using PHP as the scripting language and MySQL as the database. The database is used to store individuals, attributes, and attribute-values in normalized form. User interactions are date-time stamped so that evolution can be monitored and evaluated. Furthermore HTML, CSS, and JavaScript are used to create the interface and site design.

JavaScript is primarily used to provide enhanced interactivity. Specifically, it is used to make asynchronous calls in the background so that graphs and attribute-values can be updated on-the-fly without reloading the page. Furthermore, search and auto-suggest mechanisms that help users avoid creating redundant attributes

Figure 4.5: IAN add/remove attribute-values for the *research interests* attribute

are implemented using asynchronous JavaScript calls to a PHP file that queries the database for attributes using similar keywords (See the bottom portion of Figure 4.4).

In order to associate users and profiles, login is required to authenticate. After an initial login, a cookie is set so that the user is remembered when they later return to the site on the same computer.

R, the statistical computing and graphics language, is used to dynamically create the IAN graphs that show how individuals are connected within the community [5]. Specifically, the igraph package is used to read in a graph file (i.e., NCOL graph format), create the graph layout using the Fruchterman-Reingold algorithm [11], and save the result as an image (e.g., PNG, EPS, etc.). Furthermore, R has been used to plot many of the graphs within this document.

4.3 GIAN, Genealogical-IAN

One of the interesting features of IAN, is that it can be usefully instantiated in less dynamic domains. For example, IAN can be specialized to the genealogical domain. The data generated by an online community is typically far more volatile than a researcher's family history data, so the evolution can be slower. However, more active data collection, possibly through a family website, could provide a rich source of evolving data.

Even for slowly evolving family research data, IAN provides a neat approach for the affinity networks existing in the data. For example, through our studies we have discovered naming pattern chains, event location clusters (e.g., birth, death, marriage), and multiple cross-family marriages. We would expect that further discoveries are possible as family history data becomes richer.

GIAN includes:

- The capability to import data (e.g., a GEDCOM file for family history) via a parser
- Allows for isolated communities (separate from the general community) such that family sensitive data is kept confidential and accessible only to allowed family members
- A mechanism to select any combination of attributes for display
- Network navigation and display similar to IAN, except the weights (i.e., local user and learned community) are used less frequently as the evolution is limited

These distinctions allow common people to discover the affinity networks within their family.

Chapter 5

Experimental Results

This chapter presents the results that demonstrate how effectively IAN performs at showing affinities, tracking changes, and discovering new things. Results are from the community website IAN and the specialized family history tool GIAN.

The results are largely taken from the IAN community website experiment that was conducted for 183 days, running from May 10 until November 10. The experiment was started gradually by inviting a user or two at a time, also, allowing other users to invite others as desired. During this time frame,

- 69 individual users participated in the experiment
- On average, a user was active within the community for 39 days
- On average, a user visited every 8 days
- On average, a user added 2.38 attribute-values to their profile per visit
- On average, a user had 93 attribute-values across 21 attributes

The GIAN results were performed on a one family history at a time by importing a family GEDCOM file. Under these circumstances, evolution was disregarded, since a static snapshot of the data is viewed at only one time. The results from GIAN are shown in the Discoveries section. For future work it might interesting to explore how the family evolves using the dates provided for individuals over time.

5.1 Showing Affinities

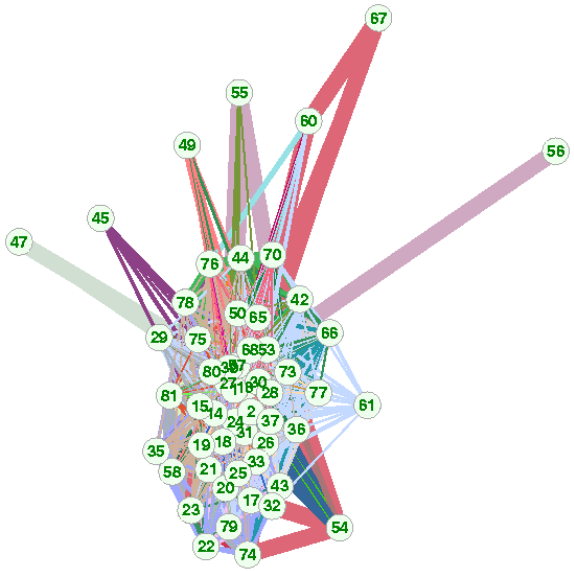

Figure 5.1: The Entire Community at 183 days - it is one connected component (no filtering)

Showing how people are related, or the affinities that connect people together is discussed in this section of the results. As motivated in the introduction, people are connected in so many ways that it is challenging to both keep track of and utilize this information. In fact, this truth is confirmed, in the online community IAN, as shown in Figure 5.1. We can see everyone is connected in some way to the community and many are connected in many ways. It is interesting that despite the differences among the users everyone is connected to the community as a whole. This graph effectively shows that the community is one connected component, yet it shows little more. Subsequent IANs are filtered so that additional interesting aspects of the community can be shown.

For example, Figure 5.2, shows a filtered version of the IAN in Figure 5.1, to illustrate only affinities that are "very important" as signified by all of the connected individuals. As mentioned previously, the local weights are provided by the user

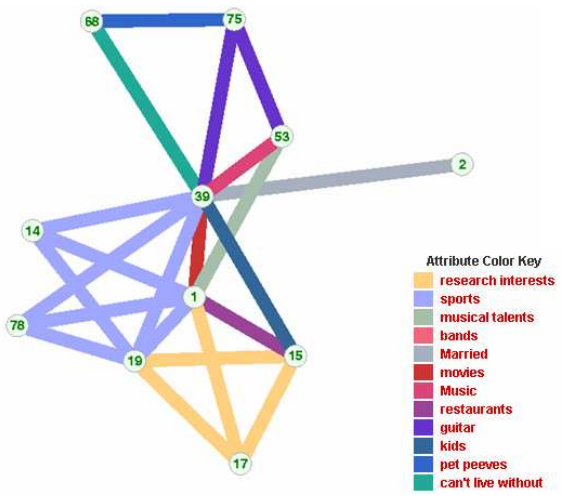

Figure 5.2: Locally weighted importance network at 180 days

according to their preference. Throughout the duration of this experiment, 30% of all community users indicated which attributes were interesting to them. Among other things, we see that there are sub-communities explicitly interested in *sports*, *research interests*, and *guitar*. We also, see that individuals having affinities within the *research interests* cluster have affinities with those in the *sports* cluster. This also occurs with the *guitar* and *sports* clusters. Individuals 1, 19, and 39 are bridging candidates between the two diverse camps.

A different perspective about the community is gained through filtering using community weights. Again, community weights are biased towards the attributes about which people tend to record their information. Figure 5.3 is a snapshot of the community with community weights, at 180 days, with a threshold of 0.2. Among other things, we see that there are implicit community connections for *hair color*, *eye color*, *food*, and *sports*.

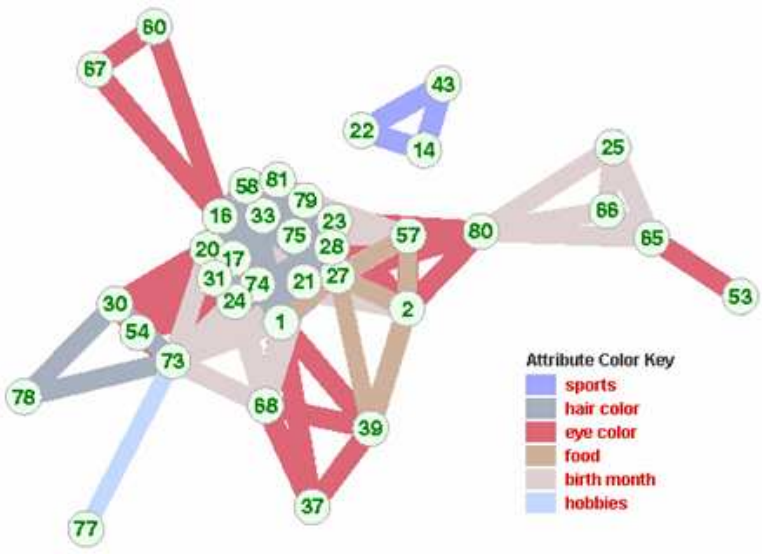

Figure 5.3: Global IAN with community weights at 180 days, T=0.2

5.2 Tracking Changes

The network strength metrics are used to track evolution of the IAN community (see Equations 3.10 and 3.11). Figure 5.4 shows the global evolution of the entire IAN community during the experiment. In this plot, it is evident that the community bonding strength as a whole decreases. Thus, bridging, or the potential for connecting with others, increases. Note that bonding slightly increases at the start of each of the upward bumps in the curve. This shows that users are bonding, just not in every possible way. In other words, users are branching out by adding new attribute-values (bridging) more than connecting to the existing attribute-values (bonding). The general downward trend is typical of a young community that is growing in new ways. Also in this plot, two significant areas are marked by red boxes. The first box (found between 50 and 100 days) contains a surge of new attribute-values that causes a small drop in bonding (an increase in bridging). The second box (found between 100 and 150 days) has a surge of new individuals that causes a major drop in bonding

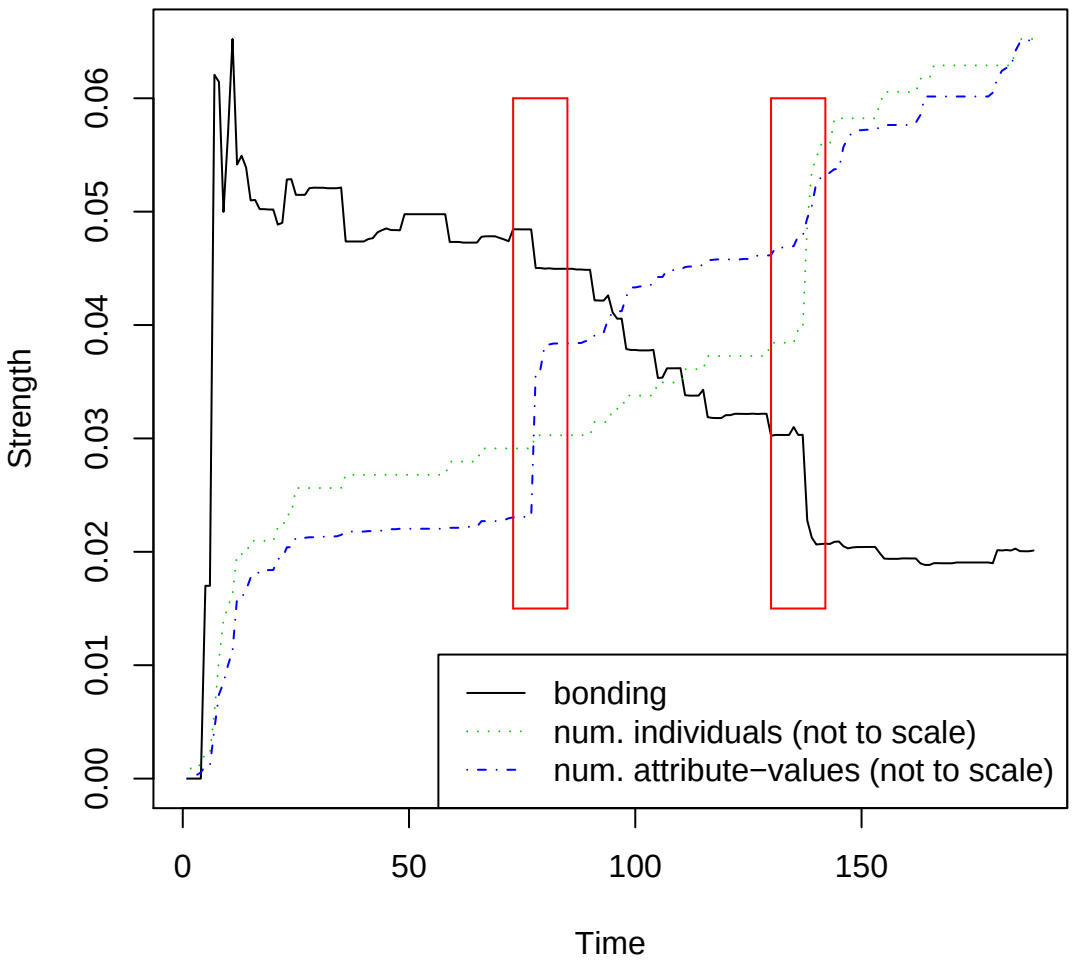

Figure 5.4: IAN Community Network Evolution

(a major increase in bridging). This occurs as the network strength metrics are more sensitive to the number of individuals than the number of attribute-values.

On the other hand, the local evolution of the *research interests* sub-community shows bonding strength having a relatively continuous increase (see Figure 5.5). Note that only users that have at least one attribute-value for the particular attribute (i.e., *research interests*) are considered part of the sub-community. For instance, at around 150 days until the end, we first see the number of attribute-values and individuals

stabilize while bonding increases (bridging decreases). At the end, of course, bonding drops as a new less-connected individual joins the sub-community and attempts to bridge in new ways. In the future, we might expect the bonding to increase (and bridging to decrease), again, until a new user joins the sub-community.

Figure 5.6 shows the evolution of the *food* sub-community. This plot has some interesting features worth pointing out. For example, within the red box, all three lines are increasing. This is interesting because bonding continues to increase despite the fact that new users and new attribute-values are being added. This means that the individuals within the community are actively bonding as new possible bridges being built.

5.3 Discoveries

The graphs in Figures 5.7 and 5.8 are interesting because the two attributes allow everyone to be connected, whereas, either one of the attributes alone produces a set of disconnected segments (e.g., the *musical talents* network is three disconnected segments alone). In this instance, these two attributes complement each other to form a connected graph. This is useful to know, for example, in situations where an individual would like to extend his/her influence beyond that of the attribute they are familiar. For example, in Figure 5.8, individual 14 or 18 might consider individuals 26 and 27 as bridging candidates between their *musical talents* sub-community and the *states visited* sub-community.

5.4 Qualitative Assessment

To assess the quality of the IAN community, two surveys were sent out via email to 65 of the community users (those heavily involved in the research were excluded).

The second survey had only one question and the results from it will be presented first. It had the following question:

> Please choose one answer to complete the following sentence:
>
> IAN _____
>
> [ ] 1. did nothing for me
>
> [ ] 2. showed me things I already knew
>
> [ ] 3. highlighted things I suspected but was unsure about
>
> [ ] 4. helped me discover new things

This survey had a response rate of about 23% (15 of 65). The average response was 3.6. The complete distribution of responses is plotted in Figure 5.9.
The first survey had the following questions:

- What did you find most interesting about IAN?
- What did you find most uninteresting about IAN?
- What new discoveries did the affinity networks help to highlight?
- What, if any, thoughts or suggestions do you have?

This survey had a response rate just over 10% (7 of 65). Of the seven that responded, the general comments they made for each of the questions are presented. In response to the questions, here are the things people found interesting:

- It was great to see the different things I have in common with the community.
- I thought it was interesting to see who was most related to me, and why
- It is cool that I can add my own interests

Here are of the things people found uninteresting:

- It was sometimes arduous to sort through attributes that I cared about
- Not knowing who some of the people in the community were made the affinities less interesting
- I don't actually know most of the people in the community

Here are the discoveries that the affinity networks helped highlight:

- It was interesting to find that I actually had stronger affinities with different people than I would have expected
- I am part of a musically talentless group
- There is a good, solid, English ancestral origin
- Soccer seems to be popular among the community
- The *birth month* attribute seems surprising, with only 4 groups
- I'm surprised by the handedness affinity network. I would have thought it would have a hard split, but hasn't (most people in that sub-community are right-handed)

Here are the additional thoughts and suggestions that people had:

- It would be interesting to do a study with businesses to see if grouping people based on their attribute affinities correlates to productivity in the workplace
- Some people used lowercase letters while others capitalized the first letter words
- I misspelled a word and did not see a way I could fix it
- The graphical depiction of the network was interesting to see, but I didn't understand the way the spatial relationships were plotted.
- I think this concept could be used as a social tool for networking and friendship-building

Separate from the results above, we present the feedback received from the online community expert Matt Brown. Matt is an active member of numerous online communities and has successfully established an online community of his own. In 2003, he started the SuperjoyMusic community which has sustained growth each year and now has more than 950 users. In response to our questionnaire, his unedited response was as follows:

What did you find most interesting about IAN?

> What I find most interesting about IAN is the new form of relationship interaction, via attribute and affinity. It's more enjoyable to see how I relate to other people through more tangible evidence than a chat room or a MySpace profile. I like the graphs, and how I am able to see just where I fit in the IAN community, with each attribute.

What did you find most uninteresting about IAN?

> I think the least interesting thing – for me – would have to be specific attributes created by IAN members. For example, I myself have almost no relation to computer programming interest, yet there is an attribute which states the favorite particulars thereof. So for me, I find little, if any interest in that specific attribute. That being said, however, there is not a place on the Internet where the same does not apply, for example: I love ESPN.com, but I could care less about their hockey highlights. IAN has no downfalls, no "if onlys", and no "I'd change that's"; that's one thing I love so much about it.

What new discoveries did the affinity networks help to highlight?

> I was able to find out just how much different I am from my peers – both old and newly found through the IAN network – but I was also able to see how I relate, even when I didn't think I would. An example would

be on the BANDS Attribute, where I felt that I would be the only IAN member to like the band "Taught Me", but I was surprised to find that another IAN member did as well. Seeing that, I was able to realize that relationships can be formed through the smallest things. Without the networks, I would be unable to employ the relationship potential.

What, if any, thoughts or suggestions do you have?

I think making the IAN Member Profiles more inviting would be a great and valuable improvement to the site, as well as finding a way to bring more color to the overall design. For the most part, the site is what it is, and that is excellent. With proper exposure to the public, and a few small improvements – such as the ones I suggested – IAN could very well become the "other" reason to check the Internet every fifteen minutes.

5.5 GIAN Results

We have created sets of GIANs from GEDCOM files provided by interested individuals, and discussed the findings with them. Figures 5.10-5.12 show some of the resulting networks.

Figure 5.10 is a given name network (two attributes: first and middle name). Although the overlay with the pedigree is not shown here, there are several patterns of names being consistently passed down from father to son through multiple generations. In the same network, as a result of merging the pedigrees of a married couple, they discovered that his and his spouse's maternal grandfathers share the same first and middle names.

Figure 5.11 is another given name network, however with gender signified. Females are signified by red circles while males are signified by blue squares. Interestingly, there are certain individuals within the family that have cross-gender names.

That is, individuals whose names have both a male and female aspect to them. For example, individuals A, B, and C (marked in green) are individuals that have this quality. There given names are *Joseph Marie*, *Francois Marie*, and *Jean Marie* respectively. Each of these individuals link a male cluster of individuals with a female cluster of individuals. Another example is found near the bottom-right of the graph, it is a cluster of *Philibert*'s (three of which are males and three of which are females).

Figure 5.12 is a spouse-sibling network (two attributes: spouse and sibling). The network nicely highlights a situation where three siblings of one family married three siblings in another family. The user was aware of two of these. We have not been able to confirm whether she knew of the third one. Additionally, four other families have a similar marriage pattern.

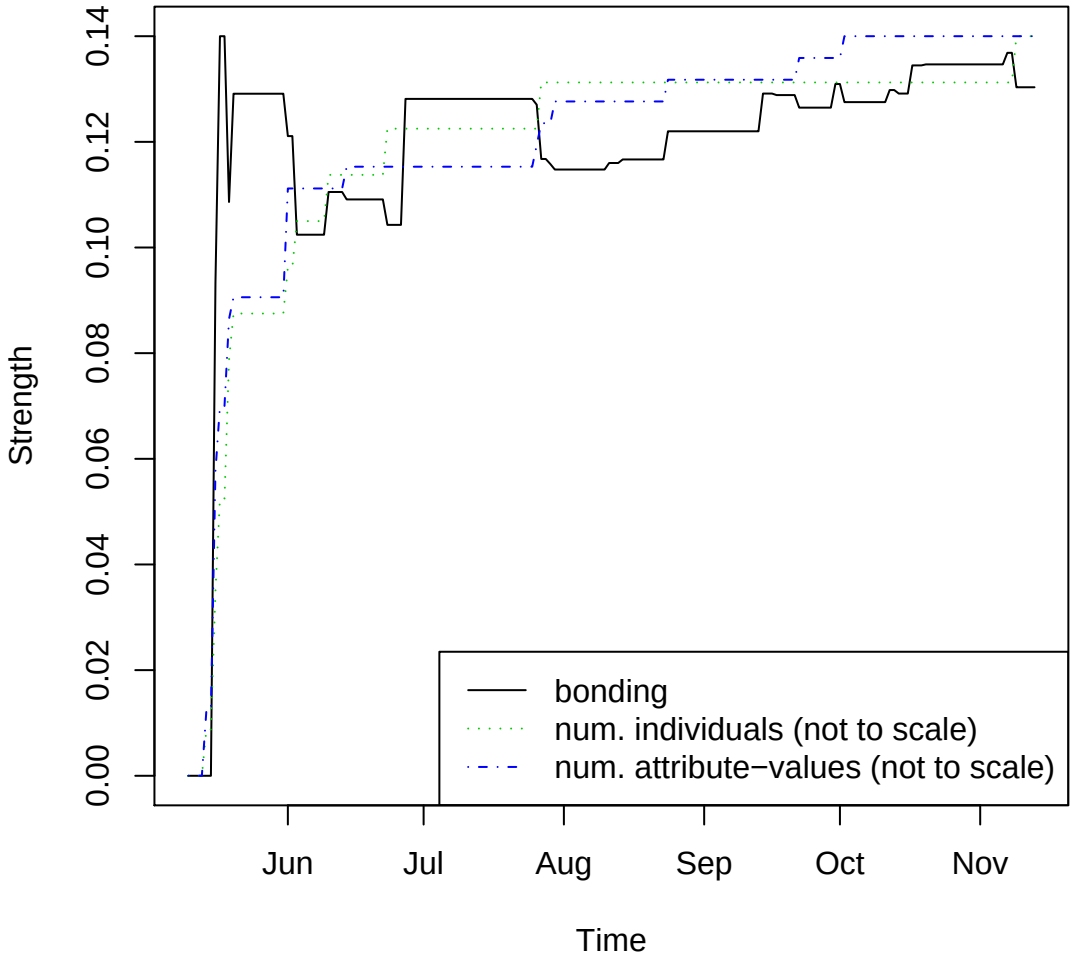

Figure 5.5: Network Evolution: Research Interests

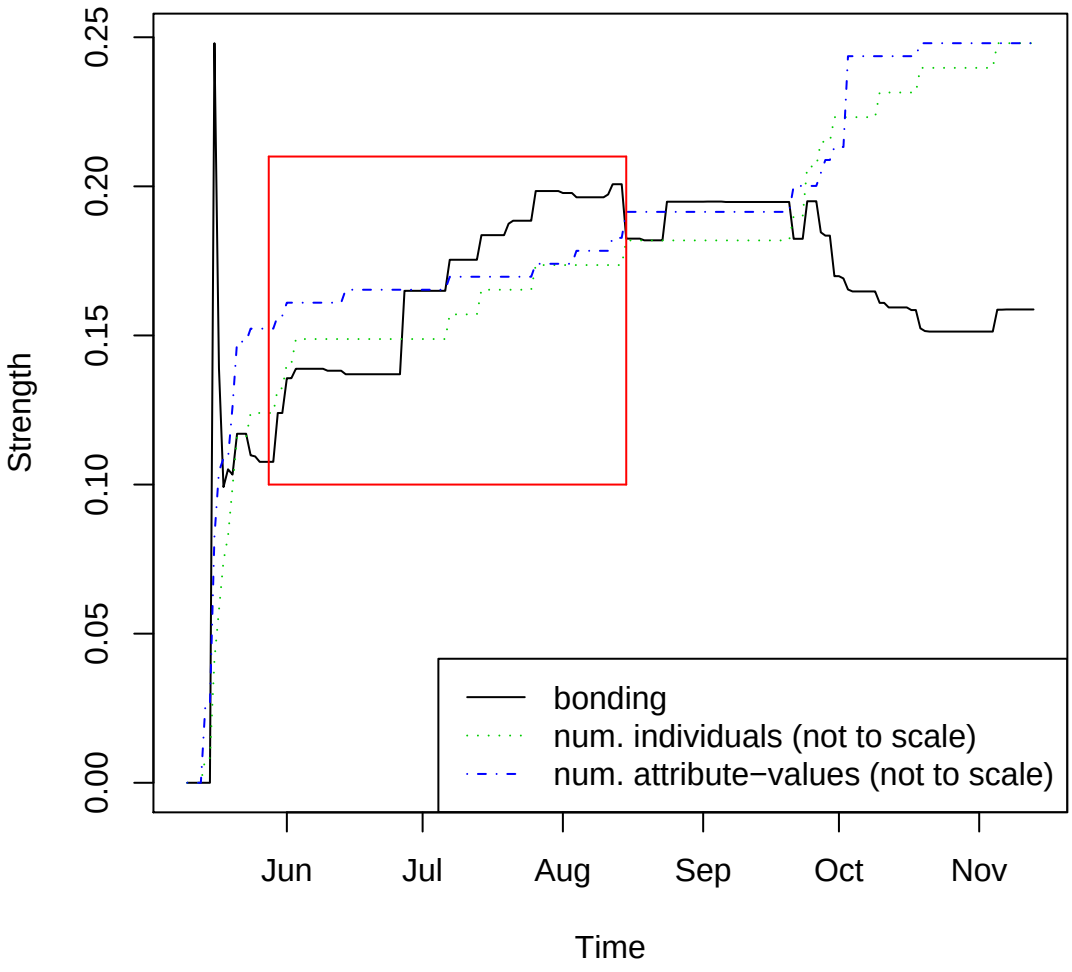

Figure 5.6: Network Evolution: Food

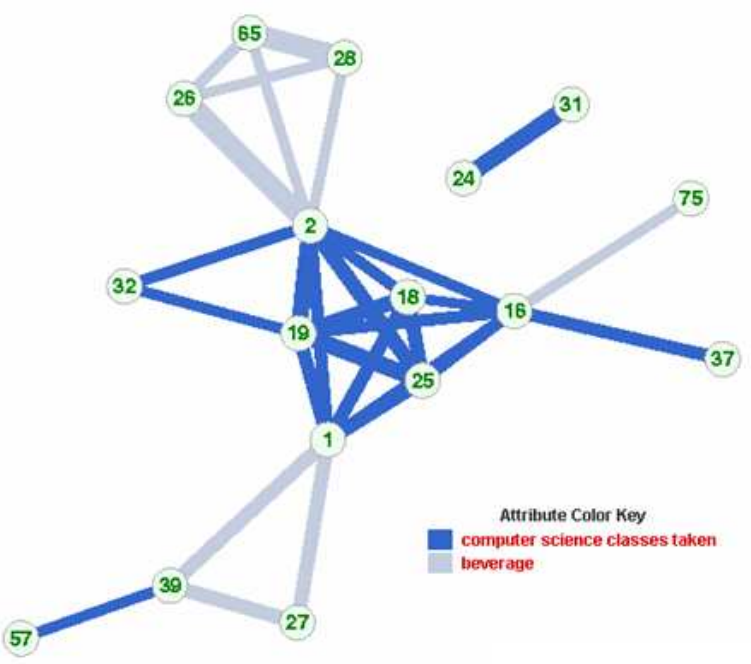

Figure 5.7: Programming Languages and Beverages

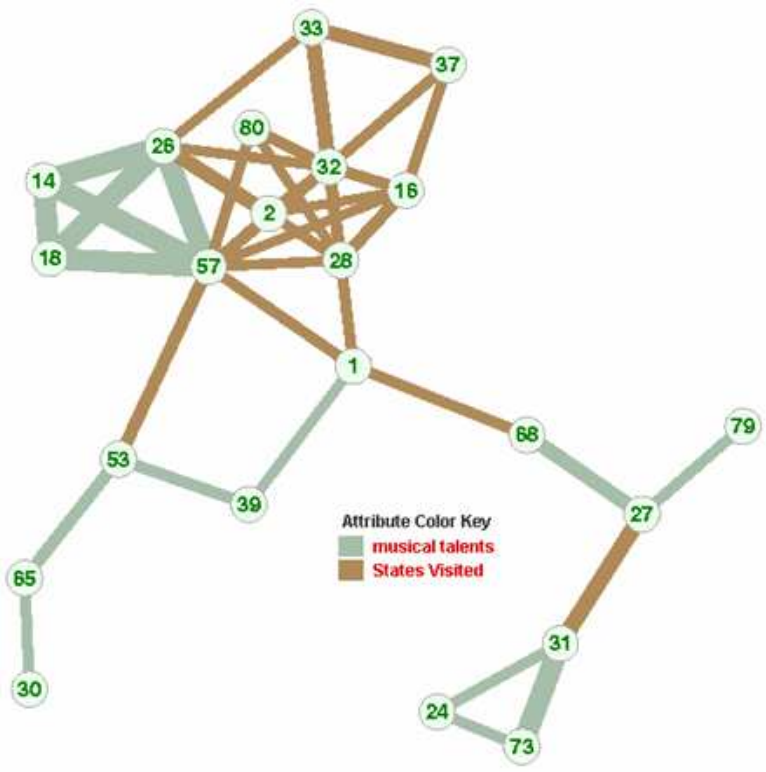

Figure 5.8: Musical Talents and States Visited

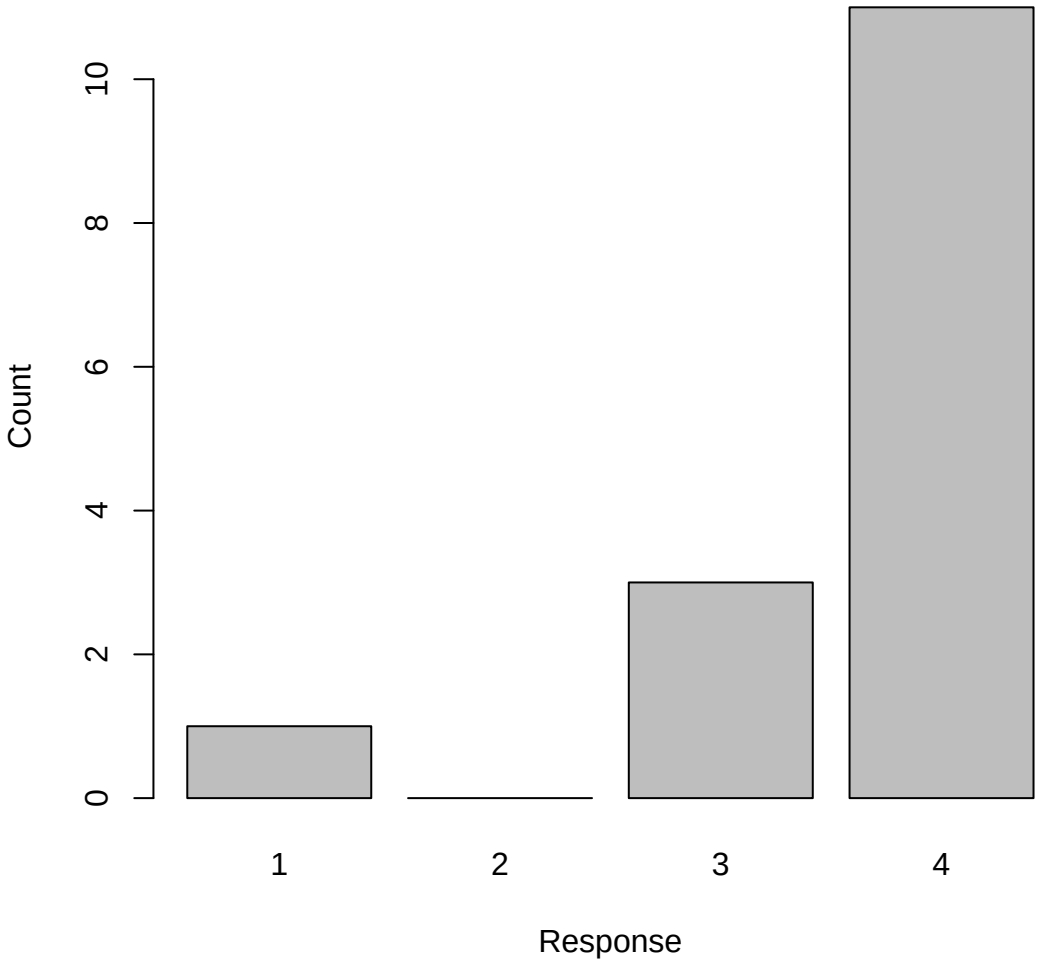

Figure 5.9: The results of the second survey. The responses were as follows: (1) did nothing for me, (2) showed me things I already knew, (3) highlighted things I suspected but was unsure about, (4) helped me discover new things.

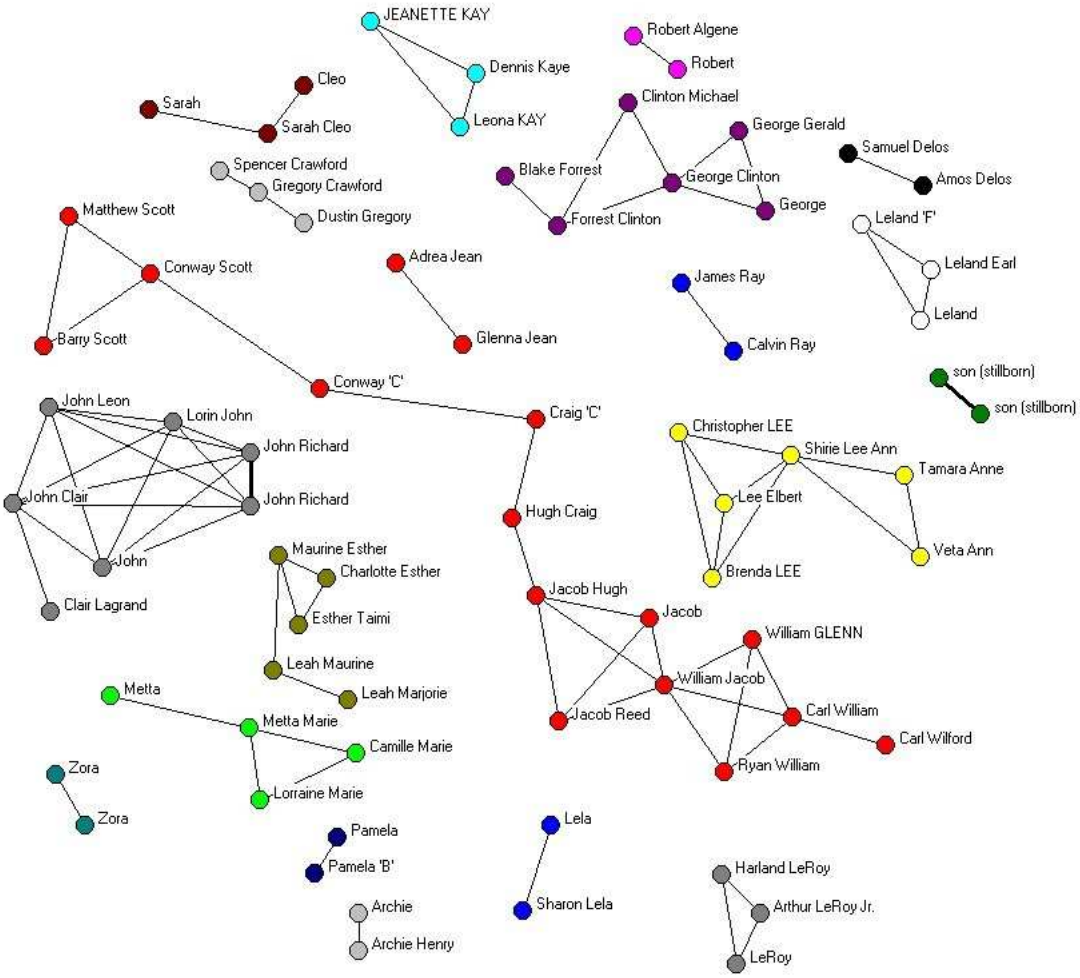

Figure 5.10: Given Name Network: Although the overlay with the pedigree is not shown here, there are several patterns of names being consistently passed down from father to son through multiple generations.

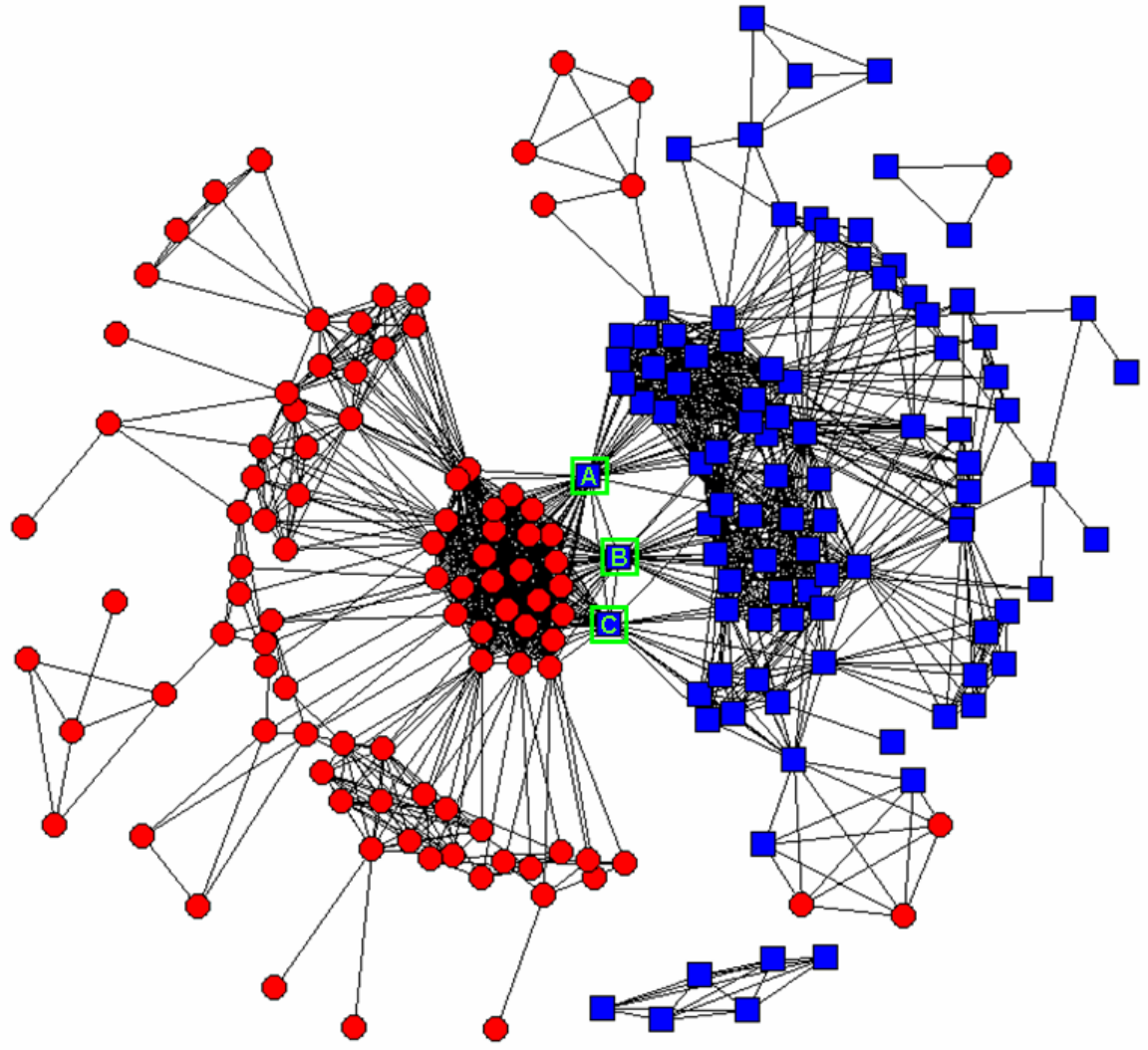

Figure 5.11: Given Name / Gender Network: Individuals A, B, and C have cross-gender names that connect male and female clusters. The given names are *Joseph Marie*, *Francois Marie*, and *Jean Marie* respectively.

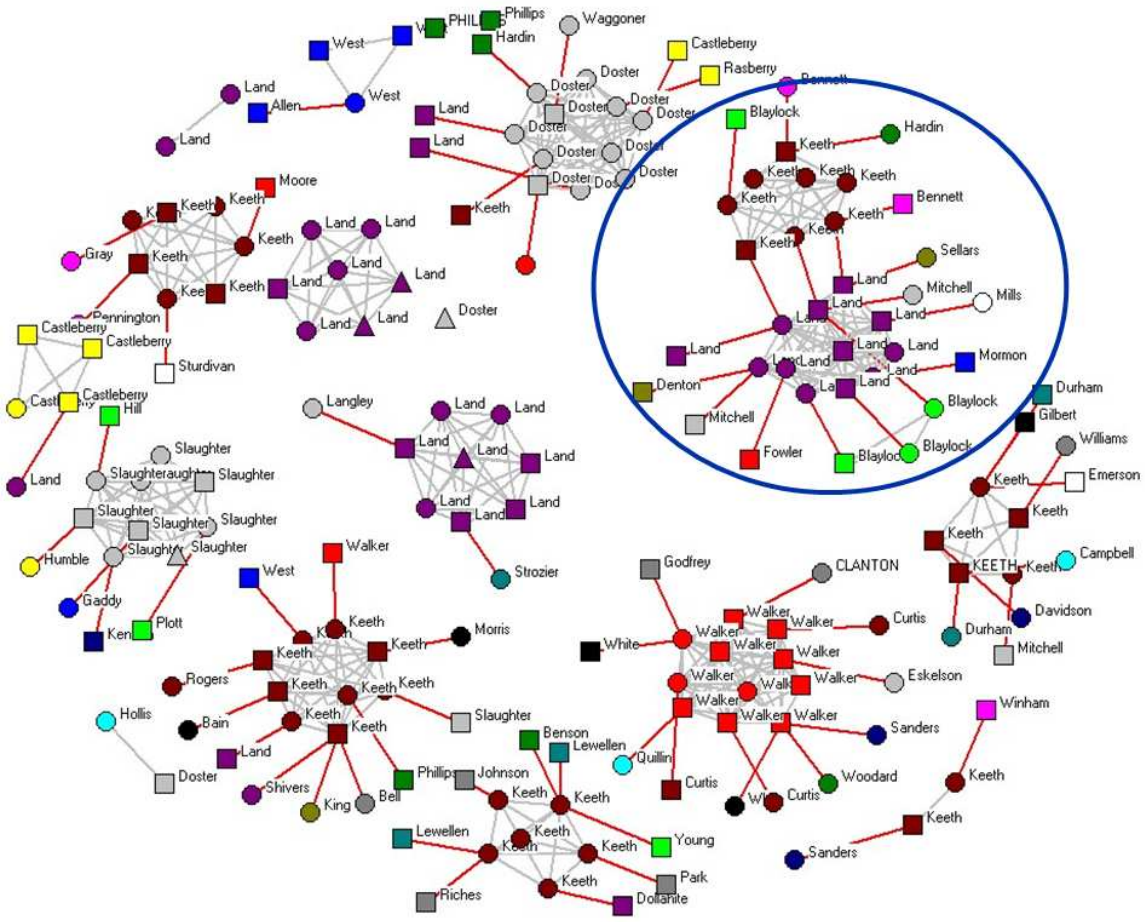

Figure 5.12: Spouse-Sibling Network: In the circled section of the graph, notice that three siblings of one family married three siblings in another family.

Chapter 6

Conclusion and Future Work

This thesis presents and evaluates the inventive methodology of creating and using Implicit Affinity Networks (IAN) for discovery. It has been shown that IANs can be used to better understand how complex entities are interconnected and how they evolve. The network strength measures of bonding and bridging can be used to measure the social capital of these communities.

IAN has been shown as a unique method for building new communities and discovering how family and friends are connected. The evolution of the community indicates levels of interest for various attributes among the members. These results provide a base for future research in IAN data mining and social networking.

In a recent article in the New York Times, renowned computer scientist Jon Kleinberg had this to say about social networking research: "We're really witnessing a revolution in measurement...This is the introduction of computing and algorithmic processes into the social sciences in a big way, and we're just at the beginning" [22]. We are encouraged by this trend. IANs offer a way to build and analyze social networks in a way that may be useful to sociologists.

IANs can be used to better understand virtually any online community. In particular, we suspect that the medical domain would benefit by the use of IAN within patient communities [17]. For example, IAN could be used to identify the affinities that patients have including symptoms and treatments. Furthermore, through the network strength measures deviating clusters might be identified more easily.

We dream of having even more interactive IAN navigation, such that users can more easily manipulate the graphs to locate and collaborate with the individuals therein, allowing users to not only discover the connections within the community, but more easily act on them.

Bibliography

[1] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami. Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C., 26–28 1993.

[2] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 12–15 1994.

[3] Albert-László Barabási. *Linked: How Everything is Connected to Everything Else*. Penguin Books Ltd., 2004.

[4] Albert-László Barabási and Albert Reka. Emergence of scaling in random networks. *Science*, pages 286:509–512, October 15 1999.

[5] R Project Contributors. The R Project for Statistical Computing. Online at: http://www.r-project.org/, October 2006.

[6] Gabor Csardi. The igraph library. Online at: http://cneurocvs.rmki.kfki.hu/igraph/, October 2006.

[7] del.icio.us. del.icio.us. Online at: http://del.icio.us, November 2006.

[8] Mukund Deshpande and George Karypis. Item-based top-n recommendation algorithms. *ACM Trans. Inf. Syst.*, 22(1):143–177, 2004.

[9] Renée Dye. The buzz on buzz. *Harvard Business Review*, November-December 2000.

[10] L.C. Freeman. *The Development of Social Network Analysis: A Study in the Sociology of Science*. Vancouver: Empirical Press, 2004.

[11] Thomas M. J. Fruchterman and Edward M. Reingold. Graph drawing by force-directed placement. *Software - Practice and Experience*, 21(11):1129–1164, 1991.

[12] Tony Hammond, Timo Hannay, Ben Lund, and Joanna Scott. Social bookmarking tools (i): A general review. *D-Lib Magazine*, 11(4), April 2005.

[13] Yahoo! Inc. Welcome to Flickr!, 2006.

[14] Jaccard. The distribution of the flora of the alpine zone. *New Phytologist*, 11:37–50, 1912.

[15] M. A. Jaro. Record linkage research and the calibration of record linkage algorithms. Technical Report 27, Statistics of Income Division, Internal Revenue Service, 1984.

[16] M.A. Jaro. Probabilistic linkage of large public health data files. *Statistics in Medicine*, 14:491–498, 1995.

[17] Grace J. Johnson and Paul J. Ambrose. Neo-tribes: the power and potential of online communities in health care. *Commun. ACM*, 49(1):107–113, 2006.

[18] Frigyes Karinthy. Chains. 1929.

[19] Jon M. Kleinberg. The Small-World Phenomenon: An Algorithmic Perspective. In *Proceedings of the 32nd ACM Symposium on Theory of Computing*, pages 163–170, 2000.

[20] Joseph A. Konstan, Bradley N. Miller, David Maltz, Jonathan L. Herlocker, Lee R. Gordon, and John Riedl. Grouplens: applying collaborative filtering to usenet news. *Commun. ACM*, 40(3):77–87, 1997.

[21] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[22] Steve Lohr. Computing, 2016: What wont be possible? The New York Times, October 2006. This has some nice quotes from Dr. Jon Kleinberg about social network analysis.

[23] Ben Lund, Tony Hammond, Martin Flack, and Timo Hannay. Social bookmarking tools (ii): A case study - connotea. *D-Lib Magazine*, 11(4), April 2005.

[24] Abraham Maslow. A theory of human motivation. *Psychological Review*, 50(4):370–396, 1943.

[25] M. McPhearson, L. Smith-Lovin, and J. Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27:415–444, 2001.

[26] S. Milgram. The Small-World Problem. *Psychology Today*, 1(1):60–76, 1967.

[27] M. E. J. Newman. The structure and function of complex networks, 2003.

[28] Erdos P. and Renyi A. On random graphs I. *Publicationes Mathematicae*, 6:290–297, 1959.

[29] Lawrence Philips. Hanging on the metaphone. *Computer Language*, 7(12 (December)):39, 1990.

[30] Plato. Phaedrus, 360 B.C.

[31] Ithiel de S. Pool and M. Kochen. Contacts and influence. *Social Networks*, 1:5–51, 1978.

[32] Martin F. Porter. An algorithm for suffix stripping. *Program 14*, pages 130–137, 1980.

[33] Robert D. Putnam. *Bowling alone: the collapse and revival of American community*. New York: Simon Schuster, 2000.

[34] Robert D. Putnam and Lewis Feldstein with Don Cohen. *Better Together: Restoring the American Community*. Simon Schuster, 2003.

[35] Frederick F. Reichheld. The one number you need to grow. *Harvard Business Review*, December 2003.

[36] P. Resnick, N. Iacovou, M. Suchak, P. Bergstorm, and J. Riedl. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, pages 175–186, Chapel Hill, North Carolina, 1994. ACM.

[37] Matthew Richardson and Pedro Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 61–70, New York, NY, USA, 2002. ACM Press.

[38] John P. Scott. *Social Network Analysis: A Handbook*. Sage, Thousand Oaks, CA, 2000.

[39] S. Staab, P. Domingos, P. Mike, J. Golbeck, Li Ding, T. Finin, A. Joshi, A. Nowak, and R.R. Vallacher. Social networks applied. *IEEE Intelligent Systems*, 20(1):80–93, 2005.

[40] Technorati, Inc. Technorati: Home, 2006.

[41] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[42] Stanley Wasserman and Joseph Galaskiewicz, editors. *Advances in Social Network Analysis: Research in the Social and Behavioral Sciences*. Cambridge University Press, Thousand Oaks, CA, 1994.

[43] Duncan J. Watts. *Six Degrees: The Science of a Connected Age*. W.W. Norton & Company, Inc., 1 edition, 2003.

[44] Duncan J. Watts and S. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.

[45] B. Wellman and S.D. Berkowitz. *Social Structures: A Network Approach*. Cambridge University Press, 1988.