Genealogical Implicit Affinity Networks

Matthew Smith and Christophe Giraud-Carrier
Department of Computer Science
Brigham Young University, Provo, UT 84602

Abstract

This paper presents a method for building networks that highlight affinities, or inherent similarities, among people, particularly family members. The content of such affinity networks can be exploited to strengthen living families and to direct family history research. Preliminary results demonstrate promise.

1 Introduction

Plato once observed that "similarity begets friendship" [5]. In recent years, modern sociologists have christened this notion *homophily* and come to describe it with the popular phrase: "birds of a feather flock together" [3]. Concurrently, others have highlighted *the small world phenomenon*, which suggests that people tend to be connected to each other by short chains of social affinities [4, 6, 11, 2, 10]. These ideas, in turn, have sparked a flurry of research in the area of social networks whose goal is to discover groups of people with particular affinities [9, 1, 7].

Now, how is this relevant to 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 what we have in common, i.e., our affinities, with 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 so for lack of interest or desire, but rather for lack of adequate tools to handle the volume of data.

In this paper, we present a method for exploiting information about individuals to build and display affinity networks within and across families. Section 2 describes how affinity networks are constructed. Section 3 shows how affinity networks can be overlayed on pedigree charts to offer additional insights to family historians. Finally, section 4 concludes the paper.

2 Affinity Network Creation

Let $A = \{A_1, A_2, \ldots, A_n\}$ be a set of attributes that characterize individuals. In practice, each A_i represents some piece of information about individuals, e.g., first name, last name, date of birth, occupation, etc. An individual x is represented by a tuple $x =< A_1 : a_1^x, A_2 : a_2^x, \ldots, A_n : a_n^x >$, where each a_j^x is the value of attribute A_j for x. The individual John Smith, for example, is represented by the tuple $< firstname : John, lastname : Smith, ... >$. We do admit the possibility that some of the A_i's be free-text fields containing a researcher's notes.

The more attributes that are available, the greater the potential is for discovering implicit affinities. For example, if there are a hundred attributes collected for individuals x and y, and only ten attributes collected for individuals z and t, then there is a greater probability of discovering implicit affinities between x and y than between z and t.

A naïve approach to discovering affinities consists of comparing attribute values across individuals using some similarity measure. Common similarity metrics include exact match, Euclidean distance, soundex, metaphone, levenstein, jaro-winkler, jaccard, and stemming.[1] Metrics generally depend on the nature of the attribute (e.g., nominal, real, string). In addition, for strings, which are common in family history, metrics vary in how they account for similarity. For example, an exact match on occupation, starting from x with occupation "Janitor" would find all and only those other individuals with the same occupation. On the other hand, a soundex comparison over first names such as "Joan" and "John" would return a match, 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.

Affinities are discovered through finding matching attributes across individuals. Consider for example Table 1, where, for simplicity, a letter is used to represent a specific attribute-value pair or characteristic (e.g., X may be birth place = Utah).

Individual	**Characteristics**
Sarah	A B C D E
Bob	A D Q R S
Jim	X Y D
Mary	X Y Z
Susan	R P Q S
Brent	Q

Table 1: Sample of Individuals and their Characteristics

Jim has characteristics X, Y, and D. In this simple example, Mary shares characteristic X with Jim so that they have an affinity that links them together. In fact, Jim and Mary also share characteristic Y, which strengthens the link between her and Jim. Note that by "sharing", we mean that the applicable similarity metric returns a match.

Through pairwise comparisons of all individuals, shared characteristics can be counted

[1] Details on these metrics are outside the scope of this paper. The reader is referred to the relevant literature.

and stored into matrix from. Table 2 shows the matrix corresponding to the individuals of Table 1.

	Sarah	**Bob**	**Jim**	**Mary**	**Susan**	**Brent**
Sarah	—	2	1	0	0	0
Bob	2	—	1	0	3	1
Jim	1	1	—	2	0	0
Mary	0	0	2	—	0	0
Susan	0	3	0	0	—	1
Brent	0	1	0	0	1	—

Table 2: Total Affinity Matrix for Individuals in Table 1

Note that here, the matrix contains only the total number of shared characteristics. This is sufficient for simple affinity analysis. For enhanced affinity analysis, the amount of similarity between each attribute could be calculated and stored.

The similarity matrix, in turn, can be represented as a weighted graph or network, where nodes are individuals and links are affinities. Figure 1 is the affinity network corresponding to the above matrix for the individuals of Table 1. The weight of each link, denoted by the relative thickness of the line, is the strength of the affinity, i.e., the number of shared attributes.

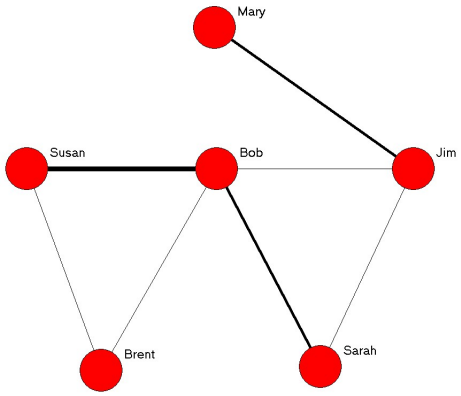

Figure 1: Affinity Network for Individuals in Table 1

An affinity network provides an intuitive graphical mechanism to discover how various individuals are connected through affinities. For instance, in Figure 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 Susan is stronger than with Sarah, Brent, Susan, or Jim (indicated by a thicker line).

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 Affinity Network Overlay

The foregoing discussion makes no assumption about any underlying family relationships. Indeed, an affinity network is unique and independent of such relationships; it is strictly based on affinities. In the context of family history, however, further insight can be obtained by overlaying the affinity network on top of a traditional pedigree chart, as illustrated in Figure 2.

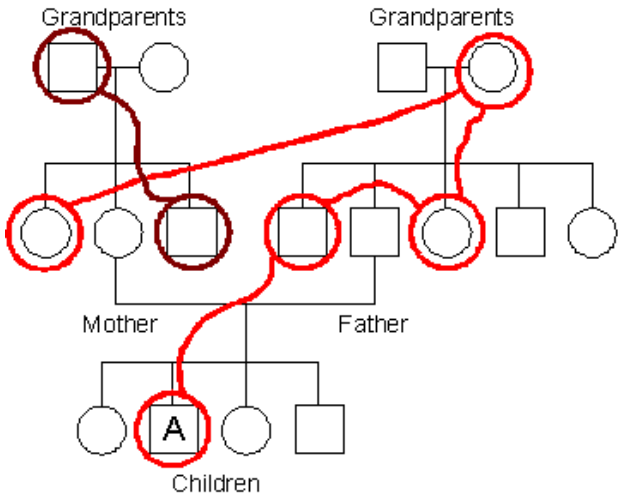

Figure 2: Pedigree Chart with Affinity Network Overlay

The content of the affinity network can then be exploited to strengthen living families and to direct genealogical research. For example, in Figure 2, we see that Child A has strong affinities with his paternal uncle. We also discover that Child A's maternal aunt has strong affinities with Child A's maternal grandmother. Furthermore, since affinity networks can be built independent of family pedigrees, affinities can be discovered across spouses' families, as well as among friends and co-workers.

Figures 3 and 4 are taken from one of the authors' pedigrees. For simplicity the actual overlay is not shown here, but it is implicit in the nature and origin of the data. Although based only on standard information, namely names and dates, the author was enthused by the discovery of things he previously ignored, such as:

- His and his spouse's maternal grandfathers share the same first and middle names.
- There are several patterns of names being consistently passed down from father to son through multiple generations.
- His brother and spouse's grandfather share the same birthday (month and day).

- Twins and/or duplicates stand out (e.g., see Jeanette Kay and Lorraine Marie in Figure 4).

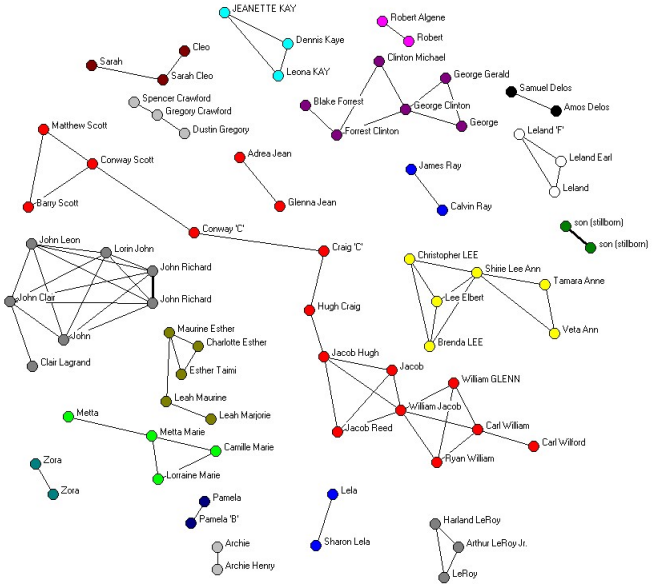

Figure 3: Family Names Network (surnames omitted)

Clearly, the richer the data, the more "interesting" the affinity networks. In some sense, one is only limited by one's own imagination (and, of course, what one can actually elicit from relatives and/or genealogical sources). As they go about their research, family historians who wish to take advantage of affinity networks to learn more about their ancestors and bring their family closer together ought to consider such questions as:

- What affinities would be interesting to living family members?
- Is family member geography important?
- Are family members' interests and hobbies important?
- What social aspects of life are of interest?
- What occupational data might be useful?

These questions, in turn, help determine what type of data will be needed to generate significant affinities for a particular family. Regardless of which affinities are desired, it is important to remember that "your affinities are only as good as the data you collect."

Interestingly, the latest GEDCOM standard [8] supports a very large number of tags (more than 130) for pre-defined attributes, such as Education, Occupation and Religion. Additionally, it is possible to create user-defined tags, which must begin with an underscore,

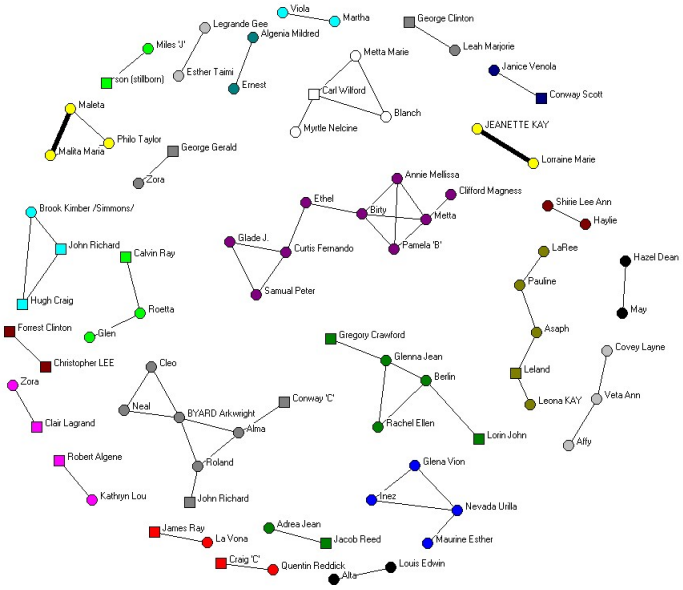

Figure 4: Birthdays Affinity Network

enabling the GEDCOM standard to be extended at will. On this same note, we recommend that family history software match this extensibility by providing functionality that allows for an arbitrary number of tags or attribute-value pairs to be stored for any individual and/or family. In the meantime, unstructured information about individuals can already be stored as free text, in the "Notes" field of, for example, Personal Ancestral File (PAF). These notes can then be mined to extract family affinity networks automatically.

4 Conclusion

This paper presents a mechanism to take advantage of research in social networks to build affinity networks from family history data. Such networks offer another perspective on family history work and often lead to new insight, that may be used to create new family ties and strengthen — possibly even mend — existing ones. In the case of ancestors, it may direct further research on specific individuals. Knowing that we have some affinity with some ancestors encourages us to find out even more about them, bringing them "closer" to us and thus effectively "turning our hearts to them."

References

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

[2] J. Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the 32nd ACM Symposium on Theory of Computing*, pages 163–170, 2000.

[3] M. McPhearson, L. Smith-Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[4] S. Milgram. The small-world problem. *Psychology Today*, 1(1):60–76, 1967.

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

[6] I. de Sola Pool and M. Kochen. Contacts and influence. *Social Networks*, 1:5–51, 1978.

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

[8] GEDCOM Team. The GEDCOM Standard Release 5.5. Family and Church History Department, The Church of Jesus Christ of Latter-day Saints, 1996.

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

[10] D.J Watts. *Six Degrees: The Science of a Connected Age*. W.W. Norton & Company, Inc., 2003.

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