Social Capital in the Blogosphere: A Case Study

Matthew Smith, Nathan Purser and **Christophe Giraud-Carrier**
Department of Computer Science
Brigham Young University
{smitty, npurser}@byu.net, cgc@cs.byu.edu

Abstract

Online communities are forming in the Blogosphere as people connect online with friends and those they have *affinities*, or inherent similarities with. We explore the social capital that exists within these communities by deriving an effective mathematical formulation of social capital based on implicit and explicit connections. We illustrate these concepts by conducting a case study on an active segment of the Blogosphere. We focus our discussion on only blogs that have significant relationships among each other. Topics are generated using Latent Dirichlet Allocation (LDA) to form an implicit affinity network (IAN) that highlights potential sub-communities that would result through increased bonding.

Introduction

Social capital is a fundamental idea in numerous research areas including business, organizational behavior, political science, and sociology. "Unlike other forms of capital, social capital is not possessed by individuals, but resides in the relationships individuals have with one another." (FAST 2006). Social capital fosters reciprocity, coordination, communication, and collaboration. It has been used to explain how certain individuals obtain more success through using their connections with other people. In an interesting study about CEO compensation, for example, Belliveau and colleagues show that social capital plays a significant role in the level of compensation offered to CEOs (Belliveau, O'Reilly, & Wade 1996).

Two forms of social capital, known as bonding social capital and bridging social capital, have recently been proposed to allow finer-grained analyses of social networks (Putnam 2000; Putnam & Feldstein 2003). Bonding social capital refers to the value assigned to social networks among homogeneous groups of people, whereas bridging social capital refers to the value assigned to social networks among socially heterogeneous groups of people. Associations and clubs typically create bonding social capital, whereas neighborhoods and choirs tend to create bridging social capital.

To better understand social capital and derive an effective mathematical formulation thereof, we find it useful to distinguish between two types of connections among individuals, as follows.

- An *explicit* connection links individuals together based on a well-defined relationship, such as "is a friend of" or "collaborates with." Individuals thus linked are aware of the explicit connections among them.
- An *implicit* connection links individuals together based on loosely defined affinities, or inherent similarities, such as similar hobbies or shared interests. Individuals thus linked may not be aware of the similarities in attitudes and behaviors that exist among them.

We call *explicit social networks* (ESNs), social networks built from explicit connections and *implicit affinity networks* (IANs), social networks built from implicit connections. We have shown elsewhere how to build IANs from individuals represented as collections of attributes and associated value sets, where links are created whenever two individuals share an attribute whose value sets overlap (Smith 2007; Smith, Giraud-Carrier, & Judkins 2007). For example, the characterizations of Table 1 give rise to the IAN marked by dotted lines in Figure 1. The solid lines correspond to possible explicit connections that make up an ESN over the same set of individuals.

From the perspective of social capital, ESNs and IANs are complementary. Indeed, "social capital can be viewed as based on social similarity, the shared affiliations or activities that indicate *how* one knows someone." (Belliveau, O'Reilly, & Wade 1996) (emphasis added). In this sense, social capital is naturally interested in implicit connections. On the other hand, social capital really only accrues when individuals are aware of it, that is when they establish explicit connections among themselves.

In this paper, we first show how to exploit the complementarity of IANs and ESNs to derive an effective mathematical formulation of social capital. We then report on the construction of a large hybrid social network in the blogosphere and show how social capital may be used to highlight important properties of the network, as well as influence its behavior.

Hybrid Networks: An Effective Basis to Compute Social Capital

Let a *hybrid social network* consist of an implicit affinity network (IAN) and an explicit social network (ESN) defined over the same set of individuals. Hybrid networks can

Individual	Attribute Value Sets
Amy	{Cancer (C), Smoke (S)}
Bob	{Cancer (C), Bald (B)}
Cheryl	{Cancer (C), Smoke (S)}
Dan	{Smoke (S)}
Ed	{Bald (B)}

Table 1: Sample Individuals and Attributes

be visualized by overlaying ESNs onto corresponding IANs. Hence, in social network analysis terminology, a hybrid network is a multigraph having an explicit and implicit relation among actors (e.g., see Figure 1).

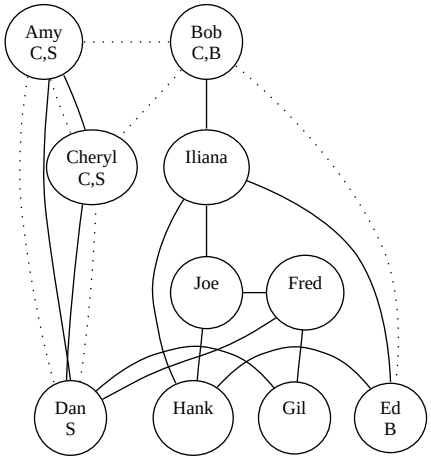

Figure 1: Sample Hybrid Network

Potential vs. Actual Social Capital

Because individuals are complex entities whose attitudes and behaviors are many, small changes to one individual's profile may have many (unexpected) effects on the overall structure of the IAN. Every time an individual's profile changes (e.g., by adding a new attribute or a new value to an existing attribute) the corresponding update creates an opportunity for new implicit connections to arise. Some are created immediately with individuals who share aspects of the updated profile, while others are established later as other individuals undergo related changes. In that sense, IANs capture the *potential* for social capital.

On the other hand, changes to an ESN are more purposeful and localized. An individual chooses precisely which other individuals to connect with. Such changes have a direct impact on the social capital of the underlying community. Hence, we can interpret IANs as capturing the potential for social capital, and ESNs —overlayed on IANs— as measuring actual social capital. Moreover, depending on the kinds of implicit connections that may exist among the same individuals, one can determine what form, bonding or bridging, of social capital is being affected and how.

Table 2 summarizes the relationship between potential and actual social capital based on the connections of a hybrid network. The presence of both implicit and explicit connections between individuals indicates actual bonding social capital as like individuals (IAN links) are linked to one another (ESN links). When only implicit connections exist among individuals, one observes only potential for bonding social capital. For example, in Figure 1, Amy and Bob have linked only implicitly, indicating that there is a potential bond that would be realized if they were to become friends. The absence of implicit connections when explicit connections exist is an indicator of actual bridging capital as diverse individuals (no IAN links) are linked to one another (ESN links). Finally, the absence of either type of connections highlights the potential for bridging social capital, that would be realized when ESN links are established.[1]

		IAN Link: Yes	IAN Link: No
ESN Link	**Yes**	Actual Bonding	Actual Bridging
	No	Potential Bonding	Potential Bridging

Table 2: Potential vs. Actual Social Capital

Table 2 makes it clear that there is no *actual* bonding nor bridging social capital without explicit links. The amount of similarity implicit among individuals determines the amount of bridging and/or bonding that occurs within the network as explicit links are made or removed. Both implicit and explicit connections are therefore necessary to calculate the network's social capital.

Bonding vs. Bridging Social Capital

Social capital is measured from a hybrid network, using both implicit and explicit connections. In general, all connections, or edges, have an associated strength or weight. For explicit edges, the strength, s_{ij}^{ESN}, of the connection between nodes i and j could be as simple as 1 or 0, to reflect the presence or absence of a link between the two nodes, but may actually range over [0,1] to capture degrees of connectivity (e.g., best friend vs. casual friend vs. acquaintance). For implicit edges, the strength, s_{ij}^{IAN}, of the connection between nodes i and j typically ranges over [0,1] and is a measure of the similarity between the nodes it connects, based on their attribute-value sets. In principle, any similarity metric can be used. In practice one generally chooses suitable metrics for the individual attributes (e.g., standard equality for numerical attributes, and adequate string metrics, such as soundex or jaro-winkler, for strings), and then computes an aggregate similarity score through some combination technique, such as Jaccard's index.

Potential bonding social capital between two nodes i and j is simply s_{ij}^{IAN}. Actual bonding social capital between i and j can then be defined as the product of the strength of the implicit edge (i.e., potential bonding social capital) by the strength of the explicit edge. That is,

$$bonding(i,j) = s_{ij}^{IAN} s_{ij}^{ESN}$$

[1]Note here that if IAN links were established first, this situation would of course turn into one of potential bonding social capital, rather than bridging social capital.

Hence, as expected, if either the implicit strength or the explicit strength is 0, that is, if either i and j have nothing in common or they do not know about each other, then there is no bonding social capital. On the other hand, if both implicit and explicit strengths are 1, then bonding is also maximum at 1. Any other configuration reflects the amount of bonding social capital between i and j.

Bonding social capital for an entire social network is the sum, over all edges, of the actual bonding social capital divided by the sum, over all edges, of the potential bonding social capital, as follows.

$$bonding = \frac{\sum_{i,j} bonding(i,j)}{\sum_{i,j} s_{ij}^{IAN}}$$

Conversely, potential bridging social capital between two nodes i and j is simply $1 - s_{ij}^{IAN}$. The more dissimilar the two nodes are the larger the potential for bridging. Then, actual bridging social capital between i and j can be defined as the product of the reciprocal of the strength of the implicit edge (i.e., potential bridging social capital) by the strength of the explicit edge. That is,

$$bridging(i,j) = (1 - s_{ij}^{IAN}) s_{ij}^{ESN}$$

If both implicit and explicit strengths are 0, then there is clearly no bridging social capital. However, potential bridging is maximum at 1, since the individuals have nothing in common. Similarly, if both implicit and explicit strengths are 1, then there is still no bridging social capital, as the individuals are homogeneous. Bridging social capital is maximum at 1 only when explicit strength is 1 but implicit strength is 0. Any other configuration reflects the amount of bridging social capital between i and j.

Bridging social capital for an entire social network is the sum, over all edges, of the actual bridging social capital divided by the sum, over all edges, of the potential bridging social capital, as follows.

$$bridging = \frac{\sum_{i,j} bridging(i,j)}{\sum_{i,j} 1 - s_{ij}^{IAN}}$$

Blog Experiment

The Blogosphere refers to the growing, worldwide social network of people who write web logs, or blogs. This large, heterogeneous network is made up of a number of communities, often organized around some common topic of interest. The social capital existing within such communities is somewhat nebulous and largely unknown, and thus underexploited. We focus here on one technology-oriented community and show how social capital can be used to influence its behavior.

We started by creating a large database of blog entries using the unofficial Google Reader API (Kennedy 2007). The database included 13,000,000 entries from over 38,000 blogs from the period of July 1st, 2006 to July 1st, 2007. We determined which blogs to retrieve entries from by following the links (i.e., HTML A/anchor tags) in the blog entries, beginning with the influential technology journalist Robert Scobles blog (http://scobleizer.com). We began with Scoble because of the large amount and wide variety of content available on his blog. We anticipated that, within only a few degrees of separation, or levels, away from Scoble we would find a rich social network.

Topic	Most Likely Topic Components (10 of 20 listed for each topic)
1	de la, regionsdash details, la ciudad, de mayo, de abril, de junio, nelson blogcast, de las, de los, distrito federal
2	elliott back, google news, news articles original, commentsoffice depot featured gadget, platinum system packs, hard drive, geek chic, nvidia geforce, santa rosa, mobile pc
3	technorati tags, open source, social media, san francisco, windows vista, web site, search engine, years ago, social networking, york times
4	pdd nos, autism spectrum disorder, autistic children, autistic child, autistic persons, developmental disabilities, ancient greek, michael phelps, autistic son, unstrange minds
5	fourth quarter, stock symbol, related articlesread, etfs type, call transcripts, research stocks, related stocks, net income, cash flow, seeking alpha
6	lindsay lohan, san francisco, wesmirch permalink, bay area, paris hilton, bed jumping, ice cream, mark pritchard, ed jew, san jose
7	windows vista, visual studio, net ajax, scott hanselman, download advertisment, windows xp, sql server, windows server, pure evil, web service
8	feed preferences powered, unified communications, siemens networks, acme packet, mobile convergence, vosky exchange, internet telephony, sip trunking, siemens ag, oliver rist
9	john mccain, rudy giuliani, white house, mitt romney, homeland security, hillary clinton, fred thompson, al qaeda, real id, barack obama
10	roxanne darling, ukulele experiment, wines tasted, beach walks, sports racer intros, download quicktimedownload ipoddownload, gary vaynerchuk, shozurobert scoble, discollection hair, joanne colanstory

Table 3: N-gram Results of LDA (used for IAN links)

To retrieve a level of blog entries to store in the database, a three step process was followed:

1. Using the pyrfeed Google Reader interface (Google 2007) entries were retrieved for all blogs on a level.
2. All links were extracted from the blog entry content.
3. We determined whether or not the URL in the link was to another feed by parsing the HTTP headers for a content-type that implied it was a feed. If content-type in the HTTP headers was 'text/html' then we parsed the HTML header to check if it contained a link HTML tag that specified a feed. If we could not find a feed for the url using either of these two methods we assumed that the link was to some other type of content besides a feed and did not consider it in our analysis.

Following this pattern, we retrieved all entries for feeds located within two levels of Scoble. We have since retrieved a third level of feed content resulting in a database of almost 20,000,000 entries from over 150,000 blogs for the same time period. We focus here only on the first two levels.

We constructed the IAN as follows. Two blogs were implicitly linked to each other if they shared common attributes. In this study, attributes are defined as the main topics of discussion found in blogs. We used Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan 2003) to model prevalent topics in blog entries throughout the 12 months of the

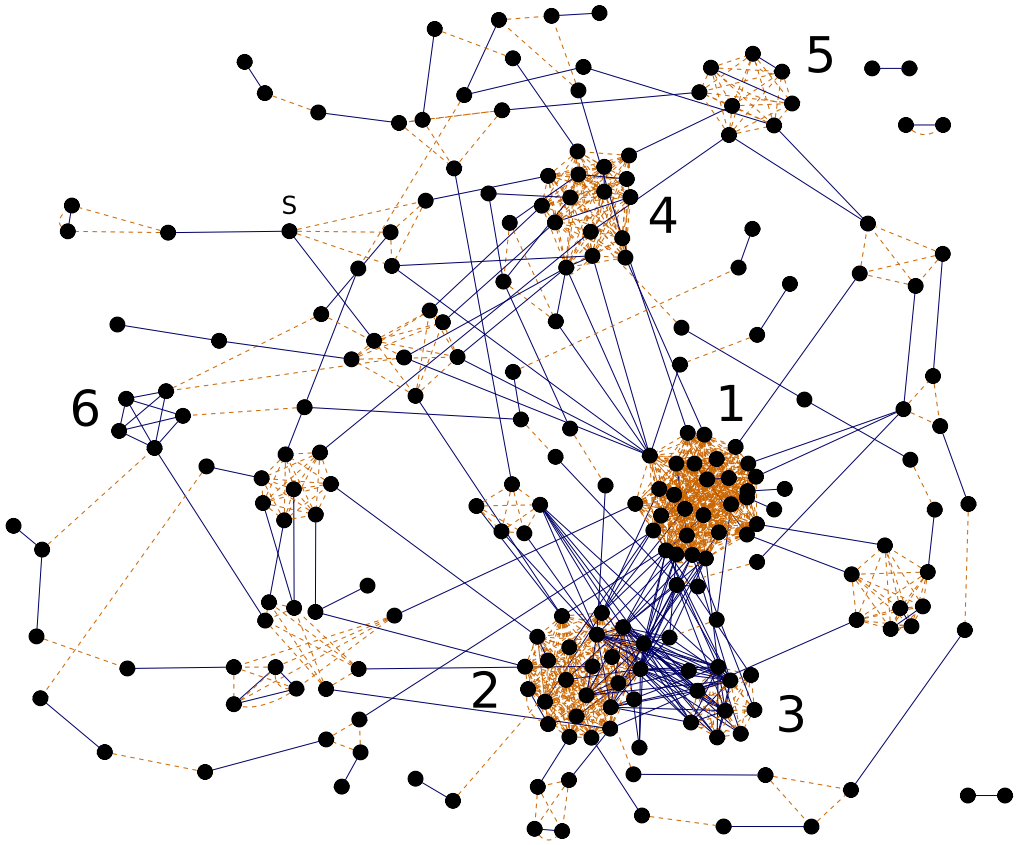

Figure 2: Hybrid Network for Blog Experiment (significant clusters are labeled 1-6, Scoble is labeled with an 'S')

experiment. To determine the n-grams within each topic, we chose to input all the entries from the first level of blogs away from Scoble. The ten topics, shown in Table 3, were generated using MALLET's (McCallum 2002) implementation of LDA. Based on this list, we determined whether a blog was a member of a topic by checking if its entries contained any of the n-grams from that particular topic. Finally, we defined two blogs to be implicitly linked if they shared the exact same set of topics. In other words, only implicit links of strength 1 are considered here. Interestingly, by manual inspection of the blogs matching this criteria, none were found to be exact replicas of other blogs, often considered spam blogs or "splogs". Future work will extend the analysis to weaker implicit links (i.e., where two blogs share only a subset of topics).

Similarly, we constructed the ESN as follows. Two blogs were considered explicitly linked to each other if they had reciprocal cross-references (i.e., hyperlinks to one another). To keep computations tractable, explicit connections between blogs were restricted to blogs that reciprocally cross-referenced each other at least 30 times during the year. Using this threshold allowed us to narrow the set of blogs to the 224 blogs, within the first two levels, that had at least one substantial explicit relationship to another blog.

Finally, we created the resulting hybrid network consisting of 224 nodes, representing blogs, and 2358 links, 494 of which are explicit and the other 1,864 are implicit. Figure 2 shows a graph of this network. In the graph, each node represents a blog while each edge represents two reciprocal links (resulting in 1179 links). The darker, solid blue lines between blogs represent explicit links and the lighter, dashed orange lines represent implicit links. Significant clusters of blogs, or sub-communities are numbered.

The network is largely connected by either implicit or explicit links, which is interesting because it suggests that most blogs are part of some larger social community. The following are worthy of note:

- Towards the bottom of the graph there are two clusters, labeled 2 and 3, that seem to be tightly linked explicitly, but have few implicit links. This is evidence that there is actual bridging taking place between the two sub-communities. In other words, blogs in each group cover similar topics, but differ across the two groups; yet they cross-reference each other.
- Throughout the graph there are several implicitly connected clusters with few explicit links among them (e.g., clusters 1,4, and 5). This presents a significant amount of potential bonding that could occur to create a new sub-community. For instance, cluster 5 includes blogs with content about the entertainment industry and pop culture. They do not link to each other explicitly although they do have a strong tendency to talk about the same top-

ics. Capitalizing on such links (through explicit connections) would add value to members of these communities who would suddenly have access to new resources (in the form of complementary blog contents) that they insofar ignored.

- On the left side of the graph, there is a group of blogs, labeled 6, that are connected explicitly yet there are no implicit links among them. This, again, is evidence of actual bridging.

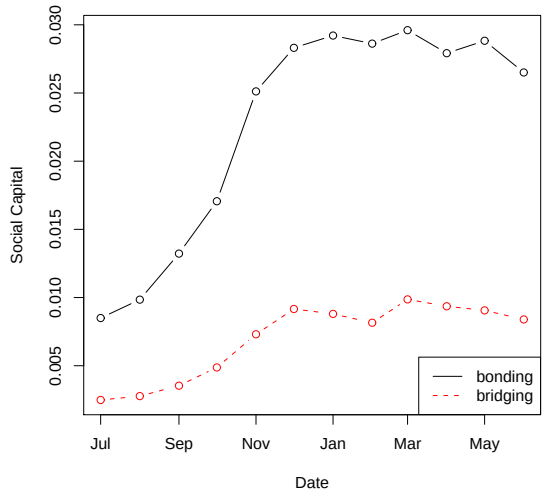

Figure 3: Social Capital by Month

Finally, Figure 3 describes the evolution of bonding and bridging social capital in the network over time. Each one-month interval is calculated using explicit links that occur during the month, while all implicit edges identified throughout the study are used (including implicit edges with strength less than 1). Initially, both types of capital rise, although more bonding occurs than bridging. This community has, and thus capitalizes on, a high level of mutual connectivity. Recall that explicit links, which here cause increasing bonding, appear only when 30 mutual cross-references are established between two individuals. This is more than two references per month on average from both blogs!

Conclusion and Future Work

We have presented a mathematical formulation of social capital based on hybrid networks that combine both implicit and explicit connections among individuals. The framework is such that bonding social capital and bridging social capital are decoupled, so that each may vary independently of the other.

This allowed us to show how a hybrid network within the blogosphere is not only connected explicitly by the blogs they link to, but implicitly by the topics they choose to write about. We showed that these are not necessarily the same groups of blogs, suggesting the emergence of new sub-communities through bonding. Identifying these sub-communities has application in many domains. For example, the medical community could use the hybrid graph to help patients communities having implicit connections to connect explicitly, thus forming support groups. The political domain could use hybrid graphs to determine where political canidates should concentrate grass roots efforts online. The growing Blogosphere creates numerous social capital applications across many different domains.

For future work, we would like to experiment using different metrics for measuring implicit links among blogs. In this study we created topics using LDA over the whole time range. We would like to create topics for smaller periods of time, so that we can accurately represent changes in the implicit network over time. This will be useful for finding trends in social networks and for individual bloggers.

In addition, we would like to extend our study to the data obtained from blogs that were three levels of separation from Robert Scoble's blog. This data will provide more diverse topics. Changing the filtering mechanics that determine which blogs to include in our graphs would also allow us to study a wider variety of blogs. Finally, we would also like to explore the possibilities of suggesting potential connections to a blogger that would allow his/her blog to bridge over into new communities or to further establish itself in sub-communities it implicitly belongs to.

References

Belliveau, M.; O'Reilly, C. I.; and Wade, J. 1996. Social capital at the top: Effects of social similarity and status on CEO compensation. *Academy of Management Journal* 39(6):1568–1593.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3:993–1022.

FAST. 2006. Social capital: Social capital as a theoretical construct. Families and Schools Together, Wisconsin Center for Education Research. Available online at http://fast.wceruw.org/theory/socialcap.htm.

Google. 2007. pyrfeed - google code. Available online at: http://code.google.com/p/pyrfeed/.

Kennedy, N. 2007. Google reader API. Available online at: http://www.niallkennedy.com/blog/archives/2005/12/google_reader_a.html.

McCallum, A. K. 2002. Mallet: A machine learning for language toolkit. Available online at http://mallet.cs.umass.edu.

Putnam, R. D., and Feldstein, L. M. 2003. *Better Together: Restoring the American Community*. Simon & Schuster.

Putnam, R. D. 2000. *Bowling Alone: the Collapse and Revival of American Community*. Simon & Schuster.

Smith, M.; Giraud-Carrier, C.; and Judkins, B. 2007. Implicit Affinity Networks. In *Proceedings of Seventeenth Annual Workshop on Information Technologies and Systems*, 1–6.

Smith, M. 2007. Implicit Affinity Networks. Master's thesis, Brigham Young University.