Dissertation Proposal:
Social Capital in Online Communities

Matthew Smith

August 3, 2009

Abstract

Online communities are connecting hordes of individuals and generating rich social network data. The social capital that resides within these networks is largely unknown. We propose to create a general framework for measuring and leveraging social capital based upon explicit social networks, implicit affinities, and the mobilization of social resources. The resulting quantitative models are used to characterize social capital in several online communities.

1 Introduction

The science of building and discovering communities is increasingly important as the Internet becomes the largest collection of ideas, personalities, and cultures in history. The continual emergence of new online communities requires better techniques for understanding these phenomena. Online communities, also referred to as neo-tribes [20], have proliferated the Internet. In particular, *the blogosphere*, the growing community of people that read and write Weblogs, has been more than doubling each year [38]. These communities represent groups of individuals connected by some well-defined, explicit relation, such as a shared medical condition in a health community, a trusted contact link in a business network, or an established friend or family relationship in a photo-sharing community. Online communities continue to rise in popularity by bringing people together to socialize, work together, and communicate.

The amount of data generated by these communities far exceeds everything collected previously. In the past, the available social network data has been limited and very static. For instance, it has been prohibitively expensive for researchers to survey individuals requesting each to name their friends, allowing a simple social network graph to be created for analysis. Due to the increased ability to connect on the Internet, social network data is available, not only for static snapshots, but dynamically over time. The social graph that is now becoming available online is more comprehensive and pertinent than those generated from manual surveys.

Social capital is a fundamental idea that originates in political science and sociology (e.g., see [27]). "Unlike other forms of capital, social capital is not possessed by individuals, but resides in the relationships individuals have with one another." [16]. Social capital fosters reciprocity, coordination, communication,

and collaboration. It has been used to explain, for example, how certain individuals obtain more success through using their connections with other people. In an interesting study about CEO compensation, Belliveau and colleagues show that social capital plays a significant role in the level of compensation offered to CEOs [3]. In another study on social capital in the workplace, Erickson concludes that "good networks help people to get good jobs" [13].

Social capital within a community is grounded on:

1. relationships (e.g., see [11])
2. individuals' attributes (e.g., see [33, 13])
3. available social resources (e.g., see [27])

To exploit (1) and (2) above, 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, and focus on their complementary natures in the context of social capital. While there is no consensual definition of social capital, most definitions focus on the value of social relations in achieving some individual or group benefit. Indeed, "social capital can be viewed as based on social similarity, the shared affiliations or activities that indicate *how* one knows someone." [3] (emphasis added). In this sense, social capital is naturally interested in implicit connections. On the other hand, social capital can really only accrue when individuals are aware of it, that is when they establish explicit connections among themselves.

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 [39]. For example, the characterizations of Table 3 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. We call a network that has both implicit and explicit links a *hybrid network*.

In regards to (3), Lin suggests that accessing social resources within a network should consider the position of ego in hierarchical structures, the nature of the tie between ego and the other actors, and the location of the ties in the networks [27].

Knowing how much social capital exists within these communities allows us to more effectively answer important questions, such as:

- Who should a community newcomer attempt to connect with?
- What influence does an individual have upon online friends in terms of mobilizing them to act? (e.g., click a link, respond to a question)
- Who should one connect with in order to gain access to additional resources?
- How heterogenous is an individual's network and what bonding/bridging opportunities exist?
- Do the attributes gleaned from an individual's data stream seem accurately describe how the individual hopes to be perceived?
- What social resources were mobilized within the community during the past month?
- Which individuals tend to mobilize the most social resources?

We propose to formalize the notion of social capital by enhancing previous metrics by incorporating the mobilization of social resources through purposive actions. This includes evaluating nodes based not only on their relationships and attributes, but on their social resources. The result is a quantitative model for characterizing social networks and providing social analytics that aid in decision making.

2 Related Work

We have organized the most relevant work related to this research in the sub-sections below.

2.1 Social Network Analysis

For decades, researchers have performed social network analysis. A plethora of structural properties and measures have been invented for social network analysis [45]. Interestingly, most properties and measures have been designed for static social networks. Some, however, such as nodal degree, diameter, and density, can easily be adapted to capture aspects of network evolution over time (e.g., see [45, 37]). Recently, some researchers have begun to study the dynamics of social network formation and evolution, leading to the discovery of several interesting patterns such as degree power laws and shrinking diameters (e.g., see [22, 24, 25, 35, 41]). Dynamic social network analysis techniques are increasingly important as the pertinent data becomes available.

Centrality measures have historically been used to determine the relative importance of a particular node within a network graph. For instance, Google's PageRank algorithm utilizes a form of centrality to provide ranked search results [32]. Common centrality measures include degree, betweenness, closeness, and eigenvector centrality [45]. *Degree centrality* consists simply of the in-degree or out-degree of a particular node [17]. Often, high in-degree centrality represents popularity, while high out-degree centrality represents gregariousness. *Betweenness centrality* takes a different slant by calculating the shortest paths between every node within the network and assigning high values to nodes included in more shortest paths, thus signifying

which nodes are most "central" [18]. *Closeness centrality* is the mean shortest path geodesic distance between the node and all other nodes reachable from it [2]. Thus, the node with the lowest value is the closest to the most other nodes. *Eigenvector centrality* is the principal eigenvector of the adjacency matrix defining the network [7]. The eigenvector provides a score for each node within the network such that a high scoring node is one that is adjacent to nodes that are themselves high scoring. In addition, the notion of individual centrality has been extended for application on groups [15, 14]. All of these centrality measures can provide a measure of individual (or group) importance that is based solely on the connections in the network.

2.2 Social Capital

The notion of social capital has been around for at least a century, however the surge of theory and research has been during the last two decades. Sociologists appear to have been most aggressive in studying the topic [27], while political scientists have made it popular [33]. The interest in social capital has since expanded to other areas including business, computer science, economics, organizational studies, and health.

Two main components of social capital have been defined: bonding social capital and bridging social capital [33, 34]. Bonding social capital refers to the value assigned to social networks among homogeneous groups of people. Bridging social capital refers to the value assigned to social networks among socially heterogeneous groups of people. As described in [30], the "conceptual distinction [between bonding and bridging social capital] should be seen as a continuum rather than a dichotomy because in practice many groups serve both bridging and bonding functions, but networks can be classified as falling closer to one end of this spectrum or the other." Associations and clubs typically create more bonding social capital; neighborhoods and choirs tend to create more bridging social capital. Whereas bonding social capital increases through closure, as individuals strengthen existing links among themselves, bridging social capital increases through brokerage, as individuals establish new links across structural holes [10]. Erickson argues that network variety (i.e., bridging capital) is a form of social capital valuable to both employers and employees in the hiring process [13]. In order to create either bonding or bridging social capital, individuals must interact.

In general, bonding interactions are more likely to occur than bridging interactions [27]. Interacting homogeneously (i.e., bonding) "should be the expected pervasive pattern of interactions observed", because it requires the least effort [27]. On the other hand, interacting heterogeneously (i.e., bridging) demands effort due to resource differentials and the lack of shared sentiments and is therefore relatively less likely to occur [27].

As theorized by Lin, *personal* and *social resources* can be characterized for individual actors. These resources are defined as either material goods (e.g. land, houses, car, and money) or symbolic goods (e.g., education, memberships in clubs, reputation, or fame). Personal resources (i.e., human capital) are in the possession of the individual, while social resources (i.e., social capital) are accessible through social connections [27]. Resources gained through bridging interactions are perceived to be of greater worth as they are more likely to be dissimilar than the resources already available.

Lin characterizes *access* and *mobilization* as theoretical approaches that describe how social capital is expected to produce returns [28]. Access estimates the amount of social capital (known to be) available to an individual. This approach is based on the assumption that the amount of accessible social capital largely

	Type of Focus	
Type of Actor	**Internal**	**External**
Individual		Ones relationships with others Me ↔ Them
Group	Structure of the relationships within the group Us ↔ Us	Structure of the relationships of the group with outsiders Us ↔ Them

Table 1: Forms/Views of Social Capital (adapted from [8])

determines the returns, without regard to the particular actions taken to use the social capital. Alternatively, the theoretical approach of mobilization reflects "a selection of one or more specific ties and their resources from the pool for a particular action at hand" [28]. For example, using a specific contact having certain resources (e.g., a highly trafficked blog, or domain-knowledge) to boost sales on an e-commerce site could be indicative of mobilized social capital.

The focus of social capital may be on the relations one specific individual maintains with other individuals, on the structure of the relations within a group of individuals, or on a combination of these [1]. Borgatti and Everett attempt to summarize these (and others) views of social capital using a 2x2 table, which considers both type of actor and type of focus, as shown in Table 1 [8].

There are further variations on these views. For example, Hobbes suggested that having a few powerful friends is more important than having many powerless friends [19], an idea taken up in a recent individual-external study, where social capital for an event was defined as the number of organizers with whom the actor is friends [26].

3 Project Description

In this section, we provide an overview of the proposed work, along with areas where experiments will be conducted.

3.1 Preliminary Work

We have begun formalizing the notion of social capital by building a mathematical model that reflects some of the main requirements (e.g., bonding and bridging) utilized in previous attempts (e.g., see [33]). There are several key features to our model, which we detail in the following sections.

1. The distinction between potential and realized social capital is made.

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

Table 2: Potential vs. Realized Social Capital in Hybrid Networks.

2. Bonding and bridging social capital are not reciprocal.

3. The model can be readily applied to available community data.

3.1.1 Potential vs. Realized Social Capital

Because individuals are complex entities whose attitudes and behaviors are prone to change over time, IANs are intrinsically dynamic, evolving with such things as their participants' age, occupation, interests, and life's circumstances (e.g., marriage, retirement). The network continually and automatically shifts as new participants create and current ones update their profile. Indeed, 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 realized social capital. Moreover, depending on the kinds of implicit connections that may exist among the same individuals, one can determine what form of social capital is being affected and how.

Table 2 summarizes the relationship between potential and realized social capital based on the connections of a hybrid network.

The presence of both implicit and explicit connections between individuals indicates realized bonding social capital as like individuals (IAN links) are connected to one another explicitly (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 realized 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]

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

Individual	Attributes
Amy	Health: {Cancer}, Habit: {Smoke}
Bob	Health: {Cancer, Alopecia}
Cheryl	Health: {Cancer}, Habit: {Smoke}
Dan	Habit: {Smoke}
Ed	Health: {Alopecia}

Table 3: Sample Individuals and Attributes. The data contains three distinct attribute-values (i.e., *cancer*, *alopecia*, and *smoke*) for two attributes: *health* and *habit*.

Table 2 makes it clear that there is no *realized* 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. Thus, *potential* bonding or bridging occurs among individuals when no explicit links are present among them. Both implicit and explicit connections are therefore necessary to calculate the network's social capital.

3.1.2 Bonding and Bridging Social Capital

Recall that a hybrid social network consists of an implicit affinity network (IAN) and an explicit social network (ESN) defined over the same set of individuals. Hybrid networks can thus be visualized by overlaying ESNs onto corresponding IANs. In social network analysis terminology, a hybrid network is a multigraph having an explicit and implicit relation among actors (e.g., see Figure 1).

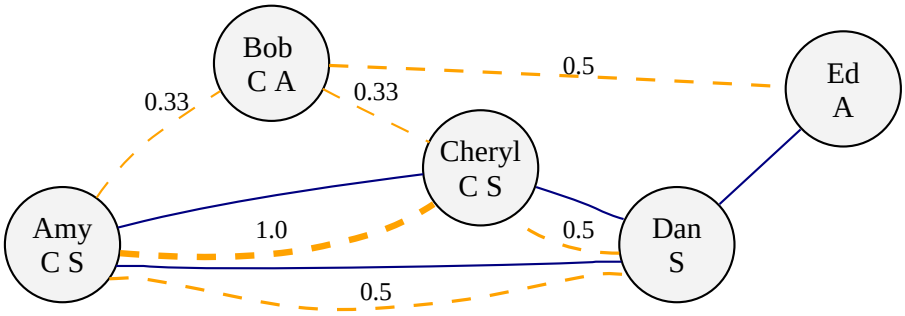

Figure 1: **Sample Hybrid Network**. Each node is labeled with the individual's name and the first letter of each attribute they possess (see Table 3). Additionally, implicit (dashed) edges are labeled with a corresponding strength, while explicit (solid) links are assumed to be of strength 1.

In [40], we showed how to derive an effective mathematical formulation of social capital by exploiting

capital, rather than bridging social capital.

the complementarity of IANs and ESNs. We formalized aspects of social capital to show precisely when the community was either bonding or bridging for the particular context. These analyses highlighted the effects that individual changes had on the community; the occurrence of an individual bridging out or showing their attributes in new areas was of particular interest.

We discussed the computation of realized social capital, which as stated above requires both implicit and explicit links. For implicit edges, the strength, s_{ij}^{IAN} , of the connection between nodes i and j ranges over [0,1] and is a measure of the similarity between the nodes it connects. 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 also range over [0,1] to capture degrees of connectivity (e.g., best friend vs. casual friend vs. acquaintance).

As mentioned earlier, social capital is comprised of the two types of social capital. Therefore, the social capital for an individual i is the sum of the bonding capital and bridging capital:

$$sc(i) = b(i) + br(i)$$

We define the *potential bonding social capital* of an individual i, where N is the set of individuals in the network, as the sum of the individual's implicit similarity strength to every other individual. That is,

$$pb(i) = \sum_{j \in N, j \neq i} s_{ij}^{IAN}$$

Likewise, we define the *potential bridging* social capital of an individual i as the sum of the individual's implicit dissimilarity strength to every other individual. That is,

$$pbr(i) = \sum_{j \in N, j \neq i} (1 - s_{ij}^{IAN})$$

For a network, we define the *potential bonding (pb)* as the sum of each individual's potential bonding score divided by two. The division by two ensures that the potential social capital shared between pairs of individuals is counted only once rather than twice.

$$pb = \frac{\sum_{i \in N} pb(i)}{2}$$

Similarly, for a network, we define *potential bridging (pbr)* social capital as:

$$pbr = \frac{\sum_{i \in N} pbr(i)}{2}$$

Normalized formulations of potential bonding and bridging that factor out the number of possible connections among individuals are defined as:

$$npb = \frac{2}{N(N-1)} pb \qquad npbr = \frac{2}{N(N-1)} pbr$$

Intuitively, these normalized equations represent the potential bonding/bridging opportunities without regard to the network size. Thus, offering a metric for comparing network diversity. For example, a relatively small niche network (e.g., Match.com) might be relatively homogeneous compared to a large general purpose network (e.g., MySpace.com). The range for each of these functions is [0,1]. Furthermore,

$$npb + npbr = 1$$

The bonding social capital realized by an individual i, when (explicitly) connecting with individual j is the product of the strength of the implicit edge by the strength of the explicit edge:

$$s_{ij}^{IAN} s_{ij}^{ESN}$$

Hence, as expected, if either the implicit strength or the explicit strength is 0, that is, if either i and j share nothing in common or they are not explicitly connected, then there is no bonding social capital. On the other hand, if both implicit and explicit strengths are 1, then bonding is maximized at 1.

We define the *bonding* social capital for an individual by summing the realized bonding for all j in N (except when $i = j$). That is,

$$b(i) = \sum_{j \in N, j \neq i} s_{ij}^{IAN} s_{ij}^{ESN}$$

Likewise, *bridging* social capital for an individual is defined, where the implicit affinity strength (i.e., s_{ij}^{IAN}) is replaced by the implicit dissimilarity strength (i.e., $1 - s_{ij}^{IAN}$). That is,

$$br(i) = \sum_{j \in N, j \neq i} (1 - s_{ij}^{IAN}) s_{ij}^{ESN}$$

Network bonding social capital is the sum of the bonding social capital for all individuals divided by two, as follows.

$$b = \frac{\sum_{i \in N} b(i)}{2}$$

Network bridging social capital is the sum of the bridging social capital for all individuals divided by two, as follows.

$$br = \frac{\sum_{i \in N} br(i)}{2}$$

A normalized version of network bonding capital is calculated by dividing the bonding (b) by the potential bonding (pbr) currently available in the social network. That is,

$$nb = b/pb$$

A normalized version of network bridging capital is calculated by dividing the bridging (br) by the potential bridging (pbr) currently available in the social network. That is,

$$nbr = br/pbr$$

From the above formulation, we can see that bonding social capital and bridging social capital are not reciprocal of each other. Instead, their values are completely decoupled, allowing each to vary independently of the other. The motivation for such a decoupling is found in the following puzzle, posed by Putnam (Personal Communication).

> Too often, without really thinking about it, we assume that bridging social capital and bonding social capital are inversely correlated in a kind of zero-sum relationship —if I have lots of bonding ties, I must have few bridging ties, and vice versa. As an empirical matter, that assumption is often false. In the US, for example, whites who have more non-white friends also have more white friends. (This generalization is based on our extensive analysis of the 2000 Social Capital Community Benchmark Survey.) In other words, high bonding might well be compatible with high bridging, and low bonding with low bridging. Of course, one can artificially create a zero-sum relationship between bridging and bonding by asking what proportion of (say) friendships are bridging or bonding, or on relative trust of in-groups and out-groups, but the result is a mathematical trick, not an empirical finding.

Our formulation is not merely a mathematical trick, but is rooted in what we understand to be the nature of realized vs. potential bonding and bridging social capital.

In [40], we 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.

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 patient communities having implicit connections to connect explicitly, thus forming support groups. The political domain could use hybrid graphs to determine where political candidates should concentrate grass roots efforts online. Furthermore, the expanding blogosphere creates numerous social capital applications across many unique domains.

3.1.3 Data Availability

Social network data is currently being generated at an unprecedented level. Popular social media websites such as, Facebook, MySpace, Blogger, and Twitter are all building social graphs (ESN) and collecting user attribute data (IAN). It is reported that millions of users are contributing data to these sites everyday [36].

Much of this data is available for public consumption. For instance, the data on Twitter including Followers and Updates is open for anyone to view and consume. As another example, there is a huge

amount of data available in the public blogosphere including a plethora of rich data and explicit connections among blogs.

3.2 Proposed Work

Up to this point, we have described social capital without discussing the role that specific social resources have within social networks. Recall that Lin characterizes how social capital is expected to produce returns through accessible social resources that can be mobilized [27, 28]. The bonding and bridging measures focused on in our preliminary work provide an intuitive sense of the homogeneity and connectedness of a community over time. However, these metrics alone fail to account for how social capital is expected to produce returns. In order to be able to give an accounting of how social capital is being used within a community, specific resources available through social connections must be considered.

The next stages of our research will address the following:

1. **Evaluate nodes based on their relationships, attributes, and social resources.** An important area of research is improving our understanding about how much social capital each individual has access to within a given community. Flow models [21] may be used to evaluate nodes to include social resources. Flow models incorporate a measure of prestige based on explicit links, however, they do not consider how similar nodes are (i.e., implicit affinities based on individuals' attributes).

2. **Identify a set of measurable social resources accessible within online communities.** These resources might include referring visitors, guest authorship, wiki contributions, blog comments, exposed sponsor information, job information, and exposure to ideas or products. These social resources will be chosen within the context of a particular domain so that the results of this research can be directly applied to existing online social networks. Part of the challenge will be to make sure that many of these social resources are measurable within the community of interest. It is possible that some simple measures, such as unique visitors to a site, which are already being collected, may serve as a good starting point and could easily be used with existing data.

3. **Formalize the notion of accessible and mobilized social resources.** Our current social capital models will be extended to include social resources identified (in the previous point). These modifications will provide a measure based upon the social capital accessible to an individual over time and a mechanism for dynamically tracking social resources as they are mobilized. These additions will provide a more accurate assessment of the social capital available to each individual and within a given online community.

4. **Run experiments to validate our formal models of social capital.** To validate the models above, experiments will need to be conducted that compare the estimated social capital to known values of social capital within publicly available community data sets. Experimental areas are detailed in the next section (3.3) and validation is discussed in Section 4.

3.3 Experimental Areas

In this section, we describe the areas where online communities will be studied.

3.3.1 Twitter

Twitter is an open community that was estimated to have 4-5 million users in November of 2008 [31] and was ranked as the third largest social network behind MySpace and Facebook in February 2009 [44]. This relatively new community allows users to contribute short free-form status updates about themselves and follow the updates of others. The status updates, called *tweets*, are a rich source of data that can be used to build implicit affinity networks, while the following and followers information can be used for explicit social network links. Furthermore, rich status update information among individuals including web links and re-tweets that might be used to quantify mobilized social resources.

3.3.2 Blogosphere

Experiments within the blogosphere can be conducted to increase our understanding of this important phenomenon. For these experiments, we extract an explicit network and generate an implicit affinity network based on blog links and entries. Rather than modeling blog communities based solely on explicit hyper-linked cross-references as in [23], we model them with an implicit overlay, based on blog content. We have performed preliminary experiments in this domain, which demonstrate promise [40].

Here, a *blog* refers to a single online journal, a *blog entry* refers to an entry in such a journal, and a *blogger* refers to an author of a blog. To build an IAN from the space of blogs, we represent each blogger as an individual (a single blog may have multiple authors), with attributes and associated values that we mine from the individual's blog entries.

Determining blogger's attributes is a significant sub-task that allows for various feature extraction techniques to be used. In previous studies we have used probabilistic Dirichlet processes [43, 5, 6, 4] to discover attributes that represent the underlying concepts that bloggers tend to write about, rather than simply the terms they choose to use. In subsequent studies, we will continue to use this approach, yet alternative techniques will be considered as they arise.

As the entire blogosphere is difficult, if not impossible, to capture and study, our experiments will focus on a sample of the blogosphere. For example, a community C could be sampled from a publicly defined set of blogs, such as a Leaderboard onTechMeme. Another possibility, would be to randomly select blogs from one or more blog aggregators (e.g., Technorati, Google Reader, Bloglines). Yet another possibility, would be to choose a blog to start from and then spider all of the explicit links within the blog recursively up to some finite level. Alternatively, sampling from the blogosphere could be discussed with an expert in the Statistics department. A handful of these methods could even be compared to determine which should be used for larger studies.

Finally, we note another important potential benefit of IANs in the blogosphere. Explicit links, captured by hyper-linked cross-references, are "already known" to the bloggers, while affinities are implicit and therefore may not be known to bloggers. In particular, bloggers may not realize how or where they fit within

a particular community based on blog entry content. Our implicit links (i.e., affinities) are derived from text in a blogger's blog entries. Thus, an IAN might be used to inform bloggers as to where they reside in the implicit network. For example, are they blogging about things that few others in the community are (i.e., bridging) or are they blogging about the same things that many others are (i.e., bonding opportunity)? The notion of social capital is used to understand the state of the community.

3.3.3 Medical Communities

Online medical communities are also becoming increasingly common. In general, they are designed to enable patients to discuss symptoms and treatments and to get support. Daily Strength [12], for example, is a community that offers support groups on many different medical conditions, including those that are less common. For some of the more common conditions, independent communities have been created, for example, there is a separate community for breast cancer [9], lung cancer [29], testicular cancer [42], and bladder cancer [9]. In addition, some of these communities incorporate doctors and other experts that provide advice and treatment options. Although, medical support groups have existed for some time, only recently have they become available online, thus offering many unexploited affinities among individuals.

We propose to conduct experiments within the medical domain to show the direct applicability to success of these communities. Individuals within these communities often share the challenges they face in hopes that they can find others in their same situation. However, sometimes these individuals are not able to find the desired support group and remain isolated even though others with related challenges exist within the community. The experiments we plan to conduct will measure the social capital within these communities.

3.3.4 Language Acquisition

We also wish to include an experimental area, namely the area of Language Acquisition, that is *not* an online community. We do this to compare the validity and significance of the proposed modeling with traditional social scientific evaluation.

The Linguistics department at Brigham Young University (BYU) is very interested in acquiring new languages and how to do it more effectively. Some researchers in this department have been studying the effects of social networks on language acquisition. Past studies have involved testing subjects on language proficiency and various surveys that seek to understand behavior, including social interactions, while abroad in a foreign country. BYU maintains a number of study abroad programs which provide significant opportunities for research in the social sciences.

We have currently been collaborating with linguistics researchers to design upcoming studies. We have been using the insights we have gained through our preliminary experimentation in online communities and the social capital literature. Lastly, they are interested in the additional analysis that our proposed model could provide.

4 Validation

Our proposed quantitative model of social capital in online communities will be validated using an ensemble of techniques as briefly described below:

- **Twitter #1**: We will generate j ($5 <= j <= 25$) ego networks for an assortment of individuals where the bonding/bridging social capital ranking is known among the j groups, or at least agreed upon by some number of individuals k ($5 <= k <= 10$). Next, the individual social capital will be modeled. The results will then be compared using the non-parametric Mann-Whitney U/Wilcoxon rank-sum test. Lastly, meaningful qualitative examples will be provided.

- **Twitter #2**: To begin this experiment, a set A of new Twitter accounts will be created and assigned names that will initially vary only by a random two-digit number appended to the username (e.g., $john12$, $john54$, $john65$), where $5 <= |A| <= 50$. A selection of Twitter users, U, will be sampled from the Twitter public timeline until the the number of individuals is greater than or equal to $|A| * f_A$, such that the number of individuals in U is evenly divisible by the number of accounts in A (i.e., $|U|mod|A| = 0$), where f_A is the number of individuals that each account will be allowed to follow during the entire study. Furthermore, the status updates for each of the users in U will be retrieved and used to describe individuals by the content they have published through Twitter. Next, each of the accounts ($a \in A$) will publish s (where $10 <= s <= 100$) status updates focused on a niche topic (e.g., web development, shopping at walmart), so that an implicit affinity network can be built among all individuals in $U \cup A$. Each account will be assigned a strategy for determining which individual to follow next. These planned strategies are listed below.

  - choose those having the highest potential bonding capital
  - choose those having the highest potential bridging capital
  - choose those having bonding/bridging closest to 50%
  - choose randomly

  Some additional baseline strategies to consider:

  - choose those having the fewest followers
  - choose those having the most followers
  - choose those having the median number of followers
  - choose those having the smallest difference between followers and following
  - choose those having the largest difference between followers and following

  Next, each account will take turns following users drawn from U using the strategy they have been assigned. Throughout the study, each of the $|A|$ Twitter accounts will publish identical status updates for their twitter stream at approximately the same time. Furthermore, many of the status updates will

contain links (e.g., bit.ly links) that will be tracked when clicked and matched with the corresponding account in A. After all of the users in U have been selected by the accounts in A and all status updates have been published the study will end. The following statistics will be plotted overtime for each account:

  - number of followers
  - number of click-thrus
  - number of personal website click-thrus (tracked by a bit.ly link)
  - individual bonding capital
  - individual bridging capital

  Lastly, the results will be presented and discussed. The results of this study hope to show whether or not choosing a bonding/bridging strategy produces significantly higher returns (e.g., number of followers, click-thrus, personal website click-thrus).

- **Medical Blogs**: A selection of m ($1 <= m <= 10$) medical blogs will be selected to seed a medical blog network that is focused on a particular ailment (e.g., alopecia, alcoholism, autism, cancer). Next the network of study will be extended to the explicit social network n ($1 <= n <= 5$) degrees of freedom away from the seed blogs. Next, the explicit social network along with a meaningful implicit affinity network will be tracked overtime and analyzed using the proposed social capital modeling. Lastly, meaningful qualitative examples will be identified and investigated.

- **Language Acquisition**: We will apply the proposed modeling on one or two current studies performed by researchers in the BYU Linguistics department. The focus of these studies is centered upon the effects of social networks on language acquisition. The linguistics researchers will use a traditional social scientific approach to analyze the data, while we independently analyze the data using the proposed method. Analyses will then be compared and contrasted to highlight the benefits and limitations of the proposed model. Furthermore, the researchers performing the upcoming studies have agreed to collect some additional individual attribute data that will be used for implicit affinities in our model. The results of this study will be available at the end of the summer (September 2009).

5 Dissertation Schedule

An approximate schedule including relevant milestones is presented in Table 4. Additionally, a list of published and potential papers is presented in Table 5.

6 Conclusion

We have proposed to create a quantitative model for characterizing social capital within social networks. This entails evaluating nodes based on their relationships, attributes, and social resources. The result is a

March-April 2009	Submission of proposal to advisor (first and second drafts)
May-Jun 2009	Submission of proposal to committee members (final draft)
Jun 2009	Schedule Dissertation Proposal
May-Jun 2009	Collect relevant data for Twitter and blog experiments
	Develop social capital modeling and research
	Assist in developing language acquisition experiments
July-September 2009	Perform Twitter and blog experiments
September 2009	Perform analysis on language acquisition study
	Perform analysis on Twitter and blog experiments
October 2009	Prepare dissertation with latest results and findings
December 2009	Submit dissertation to advisor (first and second drafts)
January 2010	Submit dissertation to committee members (final draft)
February 2010	Schedule Dissertation Defense
March 2010	Dissertation Defense

Table 4: Approximate Schedule

mathematical model of social capital that incorporates the mobilization of social resources whenever available.

Submit/(Published)	Description
(2008)	*Social Capital in the Blogosphere: A Case Study*
	In Papers from the AAAI Spring Symposium on Social Information Processing
(2008)	*Social Capital in Online Communities*
	In PIKM 08: Proceeding of the 2nd PhD workshop on Information and knowledge management
(Sep 2009)	*Implicit Affinity Networks and Social Capital*
	Information Technology and Management (Journal)
Nov 2009	*Measuring Social Resources in Online Communities*
	SBP: Social Computing, Behaviour Modeling, and Prediction
	WWW: World Wide Web Conference
Jan 2010	*Social Capital through Social Media* or *Twitter Capital*
	ICWSM: Conference on Weblogs and Social Media
	KDD: Knowledge Discovery and Data Mining
May 2010	*The Latent Value in Social Networks*
	SocialCom: IEEE International Conference on Social Computing
	SNA-KDD: Social Network Mining and Analysis

Table 5: Papers

References

[1] P. S. Adler and S.-W. Kwon. Social Capital: Prospects for a New Concept. *The Academy of Management Review*, 27(1):17, January 2002.

[2] M. A. Beauchamp. An improved index of centrality. *Behavioral Science*, 10(2):161–163, 1965.

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

[4] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation, 2003.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. In *Neural Information Processing Systems 14*, 2001.

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

[7] P. Bonacich and P. Lloyd. Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, 23(3):191 – 201, 2001.

[8] S. P. Borgatti, C. Jones, and M. G. Everett. Network measures of social capital. *Connections*, 21(2):27–36, 2 1998.

[9] breastcancer.org. Breast cancer support and community. Online at: http://www.breastcancer.org/support.html, March 2007.

[10] R. Burt. Network duality of social capital. In V. Bartkus and J. H. Davis, editors, *Reaching In, Reaching Out: Multidisciplinary Perspectives on Social Capital*. Edward Elgar Publishing, 2008.

[11] R. S. Burt. *Brokerage and Closure*. Oxford University Press, 2005.

[12] DailyStrength Inc. Health support groups at dailystrength. Online at: http://dailystrength.org/, March 2007.

[13] B. H. Erickson. Good networks and good jobs: The value of social capital to employers and employees. In N. Lin, K. S. Cook, and R. S. Burt, editors, *Social Capital: Theory and Research*, chapter 6, pages 127–158. Aldine Transaction, 2004.

[14] E. Estrada and J. A. Rodríguez-Velázquez. Subgraph centrality in complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 71(5 Pt 2), May 2005.

[15] M. G. Everett and S. P. Borgatti. The centrality of groups and classes. *Journal of Mathematical Sociology*, 23(3):181–201, 1999.

[16] FAST. 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, 2006.

[17] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 60:35–41, 6 1977.

[18] L. C. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3):215–239, 1979.

[19] T. Hobbes. *Leviathan*. Collier, New York, 1962.

[20] G. Johnson and P. Ambrose. Neo-tribes: The power and potential of online communities in health care. *Communications of the ACM*, 49(1):107–113, 2006.

[21] A. Josang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644, 2007.

[22] J. Katz. Scale independent bibliometric indicators. *Measurement: Interdisciplinary Research and Perspectives*, 3:24–28, 2005.

[23] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 568–576, New York, NY, USA, 2003. ACM Press.

[24] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 611–617, 2006.

[25] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 177–187, 2005.

[26] L. Licamele and L. Getoor. Social capital in friendship-event networks. In *IEEE International Conference on Data Mining (ICDM)*, December 2006.

[27] N. Lin. *Social Capital: A Theory of Social Structure and Action*. NY: Cambridge University Press, 2001.

[28] N. Lin. A network theory of social capital. In D. Castiglione, J. W. van Deth, and G. Wolleb, editors, *Handbook on Social Capital*. Oxford University Press, 2008.

[29] Lung Cancer Support Community. Lung cancer support community. Online at: http://lchelp.org/, March 2007.

[30] P. Norris. The bridging and bonding role of online communities. *Press-Politics*, 7(3), 2002.

[31] J. Owyang. Social Networks Site Usage: Visitors, Members, Page Views, and Engagement by the Numbers in 2008. Online at: http://www.web-strategist.com/blog/2008/11/19/social-networks-site-usage-visitors-members-page-views-and-engagement-by-the-numbers-in-2008/.

[32] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

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

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

[35] S. Redner. Citation statistics from 110 years of *Physical Review*. *Physics Today*, 58:49–54, 2005.

[36] E. Schonfeld. Top Social Media Sites of 2008 (Facebook Still Rising). Online at: http://www.techcrunch.com/2008/12/31/top-social-media-sites-of-2008-facebook-still-rising/, Dec. 2008.

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

[38] D. Sifry. State of the blogosphere. Online at: http://www.sifry.com/alerts/archives/000436.html, August 2006.

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

[40] M. Smith, N. Purser, and C. Giraud-Carrier. Social Capital in the Blogosphere: A Case Study. Number 06 in SS-08. AAAI Technical Report, The AAAI Press, Menlo Park, California, 2008.

[41] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe. A framework for community identification in dynamic social networks. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 717–726, New York, NY, USA, 2007. ACM.

[42] TC-Cancer.com. Testicular cancer support forum. Online at: http://www.tc-cancer.com/forum/, March 2007.

[43] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[44] Twitter.com. Opportunity Knocks. Online at: http://blog.twitter.com/2009/02/opportunity-knocks.html on Twitter Blog, February 14 2009.

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

SIGNATURES

This dissertation proposal by Matthew Smith is accepted in its present form by the Department of Computer Science of Brigham Young University as satisfying the dissertation proposal requirement for the degree of Doctor of Philosophy.

Dr. Christophe Giraud-Carrier
Committee Chairman

Dr. Dan A. Ventura
Committee Member

Dr. Dan Dewey
Committee Member

Dr. Charles D. Knutson
Committee Member

Dr. David W. Embley
Committee Member

Dr. Kent Seamons
Graduate Coordinator