A Computational Framework for Social Capital in Online Communities

Matthew S. Smith

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Christophe Giraud-Carrier, Chair
Dan A. Ventura
Dan P. Dewey
Charles D. Knutson
David W. Embley

Department of Computer Science

Brigham Young University

June 2011

ABSTRACT

A Computational Framework for Social Capital in Online Communities

Matthew S. Smith
Department of Computer Science, BYU
Doctor of Philosophy

Social capital is the value of the relationships we create and maintain within our social networks to gain access to and mobilize needed resources (e.g., jobs, moral support). Quantifying, and subsequently leveraging, social capital are challenging problems in the social sciences. Most work so far has focused on analyses from static surveys of limited numbers of participants. The explosion of online social media means that it is now possible to collect rich data about people's connections and interactions, in a completely ubiquitous, non-intrusive manner. Such dynamic social data opens the door to the more accurate measuring and tracking of social capital. Similarly, online data is replete with additional personal data, such as topics discussed in blogs or hobbies listed in personal profiles, that is difficult to obtain through standard surveys. Such information can be used to discover similarities, or implicit affinities, among individuals, which in turn leads to finer measures of social capital, including the often useful distinction between bonding and bridging social capital. In this work, we exploit these opportunities and propose a computational framework for quantifying and leveraging social capital in online communities. In addition to being dynamic and formalizing the notion of implicit affinities, our framework significantly extends current social network analysis research by modeling access and mobilization of resources, the essence of social capital. The main contributions of our framework include 1) hybrid networks that provide a way for potential and realized social capital to be distinguished; 2) the decoupling of bonding and bridging social capital, a formulation previously overlooked which coincides with empirical evidence; 3) the unification of multiple views on social capital, in particular, the seamless integration of resources.

We demonstrate the broad applicability of our framework through a number of representative, real-world case studies to test relevant social science hypotheses. Assuming that the extraction of implicit affinities may be useful for community building, we built a large social network of blogs from an active, tech-oriented segment of the Blogosphere, using cross-references among blogs. We then used topic modeling techniques to extract an implicit affinity network based on the content of the blogs, and showed that potential sub-communities could be formed through increased bonding. A widespread assumption in sociology is that bonding is more likely than bridging in social networks. In other words, people are more likely to seek out others who are like them than attempt to link to those they share little or nothing with. We wanted to test that hypothesis, particularly in the context of online communities. Using Twitter, we created an experiment where hand-crafted accounts would tweet at regular intervals and use varied following strategies, including following only those with maximum affinity, following only those with no affinity, following random users, etc. Using the number of follow-backs as a surrogate for social capital, we showed that the assumed physical social

behavior is also prevalent online, $p < 0.01$. There is much interest in computational social science to compare physical and cyber behaviors, test existing hypotheses on a large scale and design novel experiments. The advent of social media is also impacting public health, with growing evidence that some global health issues (e.g., H1N1 outbreak) may be discovered and tracked more efficiently by monitoring the content of social exchanges (e.g., blogs, tweets). In collaboration with colleagues from Health Sciences, we wanted to test whether broadly applicable health topics were discussed on Twitter, and to design and guide the process of discovering such themes. We gathered a large number of tweets over several regions of the United States over a one-month period, and analyzed their content using topic modeling techniques. We found that while clearly not a mainstream topic, health concerns were non-negligible on Twitter. By further focusing on tobacco, we discovered several subtopics related to tobacco (e.g., tobacco use promotion, addiction recovery), which indicate that analysis of the Twitter social network may help researchers better understand how Twitter promotes both positive and negative health behaviors. Finally, in collaboration with colleagues from Linguistics, we wanted to quantify the effect of social capital on second language acquisition in study abroad. Using questionnaire data collected from about 200 study abroad participants, we found that students participating in bridging relationships had significantly higher levels of language improvement than their counterparts, $F(1, 201) = 12.53$, $p < .0001$.

Keywords: social capital, affinity networks, online communities, social resources

Contents

Part I

Introduction

Research in building, discovering and analyzing online communities is increasingly important as the Internet becomes the largest collection of ideas, personalities, and cultures in history. These communities represent groups of individuals connected by some 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.

Not surprisingly, the need for studying and understanding the social phenomena underlying such communities has recently given rise to a new field of work, known as Computational Social Science [Lazer et al., 2009], which is materializing at the crossroads of computer science and the social sciences. The key motivation for this new field is, of course, the unprecedented size and dynamic nature of online communities. While social scientists have been successful in analyzing small and static groups, the power of more advanced computational tools is required to visualize and make sense of the huge social networks generated by online communities. Furthermore, as was noted in the context of Twitter, "[social media] data may help answer sociological questions that are otherwise hard to approach, because polling enough is too expensive and time consuming" [Savage, 2011].

Indeed, there is tremendous synergy at the confluence of these two disciplines. The social graph that is now becoming available online is more comprehensive and pertinent than those generated from manual surveys. Computer scientists, especially those in the area of

knowledge discovery and data mining, have been developing and cultivating the techniques and algorithms necessary to learn from these massive data sets. For years, social scientists have been developing theory around social behavior and human interactions, but the data has been limited and rarely dynamic. Now, the tremendous amount of social data being recorded at an unprecedented rate, offers social scientists new possibilities for testing social theories and accelerating research focused on the core issues that challenge societies across the world. Research in this area offers both computer science and the social sciences increased opportunities to contribute to the public good.

The research presented in this dissertation belongs to the area of computational social science. Our specific contribution is in the design of a computational framework for quantifying and reasoning about social capital. Social capital is unlike other forms of capital in that it is not possessed by individuals, but resides in the relationships that individuals have with one another [FAST, 2006]. Social capital fosters reciprocity, coordination, communication, and collaboration. While the notion of social capital has been around for at least a century, the last two decades have witnessed a surge of theory and research in this area. Sociologists appear to have been most aggressive in studying the topic [Lin, 2001], while political scientists have greatly contributed to its popularity [Putnam, 2000]. The interest in social capital has quickly expanded into other areas including business, computer science, economics, organizational studies, psychology, and healthcare. For example, in a study about CEO compensation, Belliveau and colleagues show that social capital plays a significant role in the level of compensation offered to CEOs [Belliveau et al., 1996]. In another study on social capital in the workplace, Erickson concludes that "good networks help people to get good jobs" [Erickson, 2004].

The analysis of social capital encompasses both the individual and public good. Generally speaking, social capital is higher when members of a community are connected and working together. While there has been a significant amount of work in the social sciences to define and measure social capital, much of it has been of a qualitative nature,

and limited to relatively small static communities. We know of no attempt at unifying the various perspectives and theories, or creating a uniform framework to compute social capital over large dynamic communities, and reason about it, as we do here.

The dissertation is organized in two parts, the first describing our proposed framework, and the second reporting on a number of case studies where we use our framework to approach relevant questions in the social sciences.

Part II consists of three papers. In the first paper (published in *Proceedings of the 17th Annual Workshop on Information Technologies and Systems*, reprinted with permission), we explain that social networks are typically constructed around an explicit and well-defined relationship among individuals. Thus analyses of such networks generally ignore other possibly interesting connections that reside in the amount of similarity among individuals. We therefore present another class of social networks, known as Implicit Affinity Networks (IANs), where links are implicit in the patterns of natural affinities among individuals. These networks exhibit rich dynamics and uncover interesting patterns of community evolution.

The second paper (published in *Information Technology and Management*, reprinted with permission) shows how implicit affinity networks in combination with explicit, well-defined relationships are important for determining social capital within an online community. We present an initial mathematical framework for social capital that 1) distinguishes between potential and actual social capital, 2) decouples bonding and bridging social capital, and 3) can be used for community tracking. These first two papers, are supplemented by practical applications of our framework within two Web communities, one focused on people's interests and one focused on topics written in blogs.

The third paper extends our framework for quantifying and reasoning about social capital to accommodate social resources. Social capital is not grounded only in the relationships that exist among individuals but also in the resources that are available to individuals or the group through these relationships. By thus extending our framework to resource-aware social networks, we bridge the gap between social networks that could leverage social capital

but have no explicit resources to be mobilized (e.g., Facebook), and resource-sharing social networks that currently give no thought to leveraging the notion of social capital (e.g., Freecycle). In addition, we show how experimentation within such an environment confirms access to social resources through social interaction, generosity, and reciprocity.

Part III consists of four papers, each containing a case study within a particular domain to which our social capital framework is applied. The first paper (published in *Proceedings of the AAAI Spring Symposium on Social Information Processing*, reprinted with permission) targets the Blogosphere. In this paper, the affinities, or inherent similarities, and the explicit relationships among bloggers are exploited and discussed in the context of the framework. Affinities, in this case, are derived using the topics discussed by authors of each blog. In particular, the framework identifies the difference between potential and actual bonding/bridging that is occurring within this highly dynamic network. In addition, potential sub-communities that would result through increased bonding are highlighted.

The second paper (published in *Proceedings of the International Symposium on Social Intelligence and Networking*, reprinted with permission) takes the widely-held view of social scientists that bonding interactions are more likely than bridging interactions in social networks, and tests it within the context of the large online Twitter community. Using our social capital framework, we confirm that this assumption seems to carry over to the online world, as users who request to follow others having similar profile descriptions (i.e., attempting to bond) are more likely to be followed back than others. From a practical standpoint, this result also informs how a new user might interact on Twitter to maintain a high follow-back ratio.

The third paper (published in *Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction*, reprinted with permission) examines how public health topics can be discovered on Twitter. Although this paper does not address social capital explicitly, it supplements our research by delivering a viable technique for creating an implicit affinity network (IAN), which may benefit future social capital studies.

Additionally, the methods used in this paper equip public health researchers and practitioners to better understand public health problems through large social datasets.

The fourth paper (published in the *33rd Annual Conference of the Cognitive Science Society*, reprinted with permission) leverages our framework to quantify and reason about social capital in an off-line social sciences study. We consider the role of social capital in language acquisition during study abroad. Using data collected from over 200 participants in Japanese study abroad programs, we show that students who leverage social capital through bridging relationships achieve higher levels of language improvement. Furthermore, an analysis of the topics participants discuss with locals suggests that there are significant differences between students who have a tendency to build close-knit networks and students who cast a broader net.

Finally, we conclude the dissertation with a summary and thoughts about some promising areas of future work.

Part II

Social Capital Framework

M. Smith, C. Giraud-Carrier, and B. Judkins. Implicit Affinity Networks. In *Proceedings of the 17th Annual Workshop on Information Technologies and Systems*, pages 8–13, 2007.

M. Smith, C. Giraud-Carrier, and N. Purser. Implicit Affinity Networks and Social Capital. *Information Technology and Management*, 10(2–3):123–134, 2009. *(The original publication is available at www.springerlink.com)*

M. Smith, C. Giraud-Carrier, and S. Stephens. Measuring and Reasoning About Social Capital: A Computational Framework. *(Submitted)*

Implicit Affinity Networks

M. Smith[1], C. Giraud-Carrier[1] and B. Judkins[2]
[1]Dept. of Computer Science, Brigham Young University, USA, {cgc@cs.,smitty@}byu.edu
[2]Amazon.com, USA, {brockjudkins@gmail.com}

Abstract
Social networks are typically constructed around an explicit and well-defined relationship among individuals. In this paper, we describe another class of social networks, known as Implicit Affinity Networks (IANs), where links are implicit in the patterns of natural affinities among individuals. Preliminary results with two Web communities, one focused on people's interests and one focused on people's blogs, exhibit rich dynamics and show interesting patterns of community evolution.

1. Introduction

Online communities, also referred to as neo-tribes [3], have sprung up like mushrooms all over the Internet. 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. The resulting social networks are relationship-centered, and their analysis typically assumes that the network is static, or evolves sufficiently slowly to make the study of snapshots relevant and meaningful. Much work has been done to capture, understand, and model the structure of such social networks (e.g., see [13,15]).

Although useful from a practical (computational) standpoint, the assumption of a static network tends to limit the kinds of analyses that may be performed. Recently, some researchers have begun to study the actual 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 [4,6,7,11,14]). We propose to go further and allow the nature of the underlying relationship to vary by focusing on implicit affinities. We take an individual-centered rather than a relationship-centered view of social networks. We consider individuals as social actors characterized by a wide range of attributes and we let relationships among them emerge naturally as a result of commonalities across attributes. Unlike traditional social networks where links represent *explicit* relationships, the links in our approach are based strictly on affinities, or inherent similarities, among the social actors, which create *implicit*, and multi-faceted, relationships. We call the resulting networks *Implicit Affinity Networks* (IANs). Because individuals are complex entities whose attitudes and behaviors change over time, IANs are intrinsically dynamic, and evolve naturally with such factors as their participants' age, occupation, interests, and life circumstances.

In this paper, we describe how IANs can be generated from information about individuals, to visualize and analyze affinities among groups of these individuals. We then report on the early evolutionary stages of a Web community based on implicit affinities as well as the richer dynamics of a blog-inspired community.

2. Community Generation: IANs

We represent individuals by collections of attributes and associated value sets. Each attribute captures some information about individuals, such as occupations, hobbies, research interests, birth place, etc. In our context, an individual may be characterized by any number of attributes and each attribute may have any number of its possible values. Whenever two individuals share an attribute whose value sets overlap, we say that there is an *affinity* between them. A group of individuals together with their affinities can be represented as a graph or network, known as an Implicit Affinity Network (IAN), where each node corresponds to an individual and each edge to an affinity. Any time an individual X adds a value, say v, to one of its attributes, say A, new edges are automatically added between X's node and all existing nodes whose individuals have value v for A. Since we are interested in tracking evolution, our networks are

actually time graphs, as defined in [5], where every node and every edge in the network is time-stamped with the time at which it was added.

In principle, any similarity function defined over pairs of individuals may be employed to build an IAN. Our focus, here, is on the analysis of the network rather than the specific underlying similarity function. Hence, we propose a relatively simple function, as follows. Let Γ be a set of attributes, and for each attribute $A \in \Gamma$, let V_A denote the arbitrary value-set of A. For any individual X, let $Attr(X) \subseteq \Gamma$ denote the set of attributes of X, and $V_A(X) \subseteq V_A$ denote the set of values of attribute A for individual X. Then, the affinity score between X and Y is given by:

$$AffScore(X,Y) = \frac{\sum_{A \in Attr(X) \cap Attr(Y)} AffScore_A(X,Y)}{\left|Attr(X) \cap Attr(Y)\right|}, \text{ where } AffScore_A(X,Y) = \frac{\left|V_A(X) \cap V_A(Y)\right|}{\left|V_A(X) \cup V_A(Y)\right|} \times \alpha_A$$

The term α_A is an optional weighting factor for the Jaccard's index $AffScore_A(X,Y)$. This weight may be used to reflect the relative importance of A in a community. In the most general case, α_A is a composite of individual user preferences and a mined community preference. The former is elicited from individuals, e.g., using a kind of 5-star rating. The latter is the ratio of the number of individuals that have at least one value for attribute A to the total number of individuals in the community. It acts as a global, learned, community weight that evolves with changes in the behavior of individuals and favors frequently used attributes.

3. Social Capital for Community Tracking

Several measures have been proposed to capture the structure and evolution of social networks, including nodal degree, diameter and density (e.g., see [15]). Here, we propose a measure, based loosely on the notion of social capital, which originates in political science and sociology (e.g., see [8]). The notion of social capital seems relevant, and rather intuitive, in the context of implicit affinity networks. 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. It has been suggested that "social capital can be viewed as based on social similarity, the shared affiliations or activities that indicate *how* one knows someone" [1]. In this sense, social capital is not limited to explicit relationships but also implicit ones that result from similarities that may exist in individuals' attitudes and behaviors.

Two main components of social capital have been defined: bonding social capital and bridging social capital [9,10]. 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 heterogeneous groups of people. Associations and clubs typically create bonding social capital; neighborhoods and choirs tend to create 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 [2].

Because IANs capture implicit affinities, they can only be used to compute the *potential* for social capital rather than social capital itself. Social capital really accrues when individuals are aware of it, that is, when they establish explicit and intentional relationships with each other. It is still informative to understand and track what social capital may be available to individuals and communities.

Here, we define bonding and bridging potentials simply, as reciprocal of each other, by the following formulas, where N denotes the number of nodes and E the set of edges in the network:

$$BondingPotential = \frac{2}{N(N-1)} \sum_{\{X,Y\} \in E} AffScore(X,Y) \text{ and } BridgingPotential = 1 - BondingPotential$$

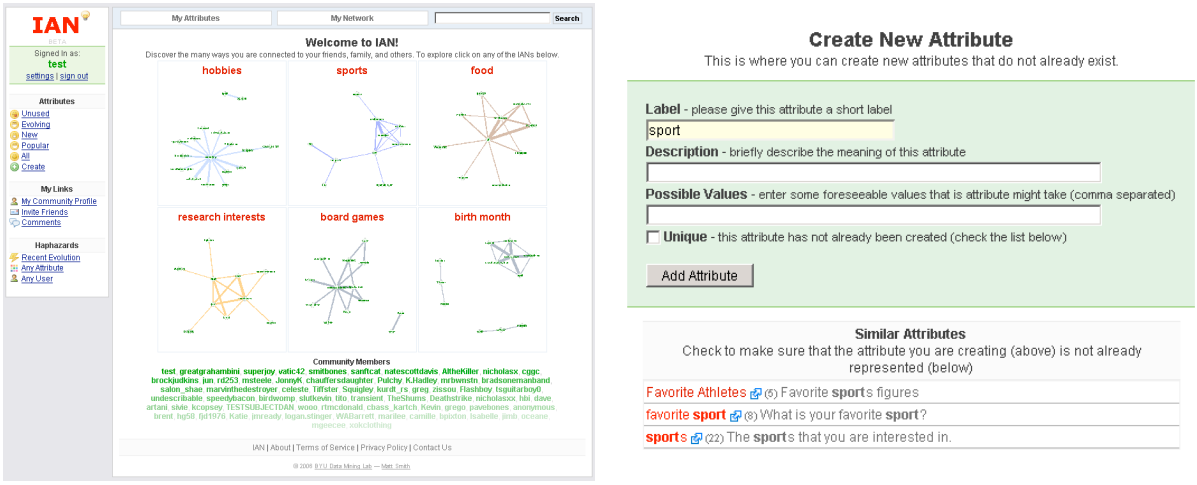
Whereas *BondingPotential* is indeed a measure of homogeneity, *BridgingPotential* is a measure of diversity within the network, suggesting how individuals may further connect. Every time an individual

adds a new attribute or a new value to an existing attribute, we therefore say that this individual is attempting to bridge out by seeking new connections with new people.

Notice that *BondingPotential* is essentially a weighted version of the density measure Δ defined in [15], where the edges' weights are given by overall affinity scores. One of the unique features of IANs is that these weights are not fixed, but naturally adapted as the profiles of individuals change. Specifically, for an edge $\{X,Y\}$, if X adds a value to one of its attributes, say A_k, and that value is not shared by Y, then the weight of $\{X,Y\}$ decreases.

4. Experiments

As mentioned earlier, most existing online communities are based on a single type of explicit relation among individuals. Even when additional data is available about individuals beyond the relation itself, such data typically lacks the time element necessary to analyze the evolution of implicit affinities required by IANs. Hence, for our first experiment, we implemented a Web application that allows individuals to create and edit their profile in the form of dynamic attribute-value sets, where each attribute captures some characteristic or personal dimension of interest that individuals wish to be represented by and share (see Figure 1). Any and all changes to attribute and attribute-values are time-stamped.



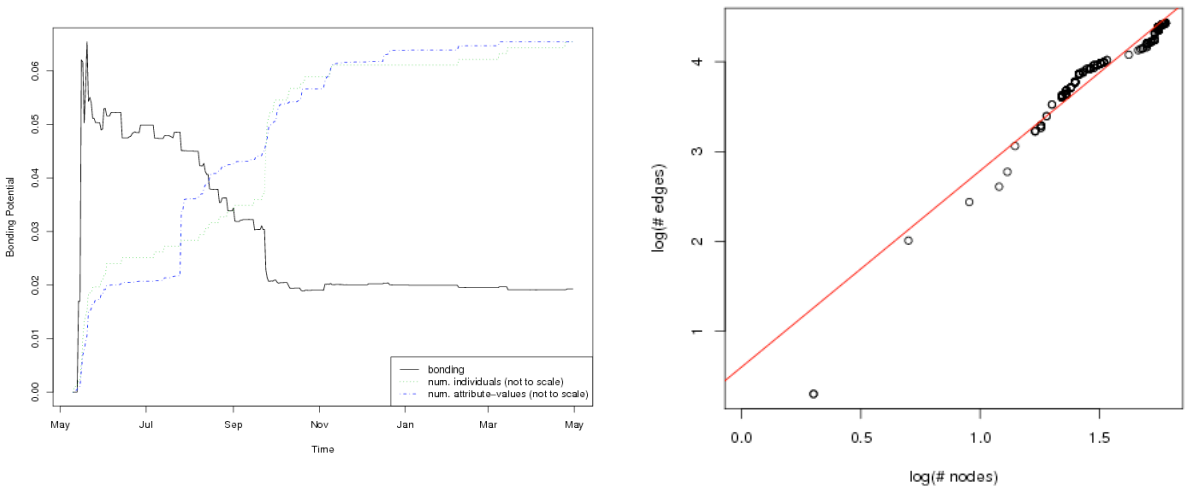
(a) User's Home Page (b) New Attribute Creation

Figure 1: Interests-based Community

In its current form, the IAN community is a general community that enables sub-communities to emerge over a variety of topics. As of a year from its inception, there were 72 individuals signed up. On average, a user 1) was active within the community for 52 days, 2) visited every 10 days, 3) added 2 attribute-values to their profile per visit, and 4) had 95 attribute-values across 21 attributes.

Recent work on social networks has examined the evolution of the average degree of nodes (i.e., $2|E|/N$) over time [6,7]. We wish to do the same with bonding potential. Figure 2(a) shows the global evolution of bonding potential in the IAN community during our experiment. The number of individuals and the number of attribute-values are also shown to facilitate interpretation. The overall trend in the evolution of bonding potential is decreasing, as might be expected of a still fairly new and rather varied community. At a lower level, the graph may be split into four time periods:

- May. This is the ``birth" of the community. As one might expect, bonding potential rises as a small number of people join in and begin sharing values on a small set of attributes.
- June-July. This is a period of relative stability, where the number of attribute-values remains relatively constant and only a few new individuals join the community. Note that arrival of new individuals generally results in a short-time drop in bonding potential, followed by an increase, as bonding replaces bridging.



(a) Bonding Potential (b) #Edges vs. #Nodes

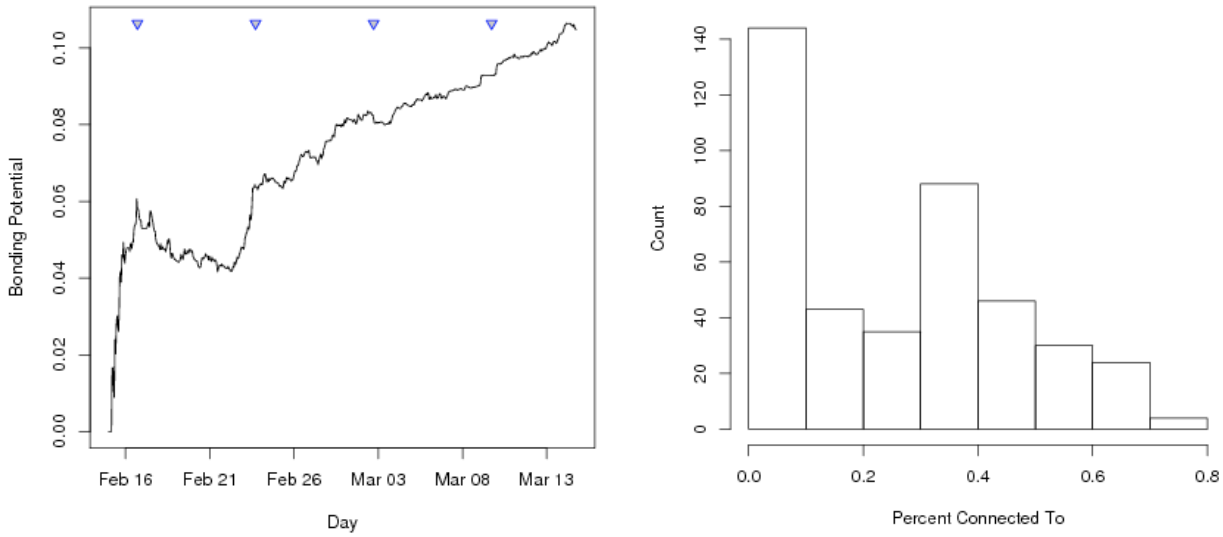
Figure 2: Interests-based Community Evolution

- August-October. This is a period of high activity, partly due to our extending the availability of IAN. During this period a significant number of individuals join IAN and create a significant number of new attribute-values, faster than current members can exploit, thus leading to a decrease in bonding potential. New members are ``casting their lines out,'' attempting to bridge out by offering new possibilities for affinities with current and new members.
- November-May. As the number of individuals and the number of attribute-values begin to stabilize again, bonding potential plateaus out. The addition of new individuals and new attributes, which causes small troughs on the bonding potential curve, seems to be compensated by the capitalization of individuals on existing attribute-values, i.e., bonding with others rather than bridging out.

Recent studies of dynamic social networks have highlighted characteristics or laws that seem to have broad applicability. In particular, it seems that social networks exhibit densification (i.e., the relation of the number of edges to the number of nodes follows a power law, $E(t)=N(t)^a$ for $1<a<2$) and shrinking diameters (i.e., the 90th percentile of the shortest path lengths between all pairs of nodes decreases over time) [7]. We wish to see whether IANs obey similar laws. We restrict our attention to densification, realizing that our network is still relatively small at this stage. Figure 2(b) plots the log of the number of edges versus the log of the number of nodes, when edges are aggregated across the values of each attribute (i.e., at most one edge per attribute between any 2 nodes). As can be seen, it appears that densification in IANs also follow a power law. Given the way new links arise in IANs, i.e., with probability proportional to the richness of existing individuals' profiles rather than the richness of their connections, we might expect that the exponent be larger than 2. Indeed, the slope of the regression line is 2.18, with $R^2=0.92$.

In our second experiment, we generate and analyze an implicit affinity network based on blogs. Rather than modeling blog communities based on explicit hyper-linked cross-references as in [5], we model them implicitly, based on blog content. We mine blogs from the public reading list of an influential technology journalist, Robert Scoble [12]. From his list we extracted 19,337 individual blog entries authored by 2,041 bloggers, over a period of a month (from 15 February 2007 to 15 March 2007). To build an IAN from this space of blogs, we represent each blogger as an individual, with a set of attributes and associated values that we mine from the individual's blog entries. Clearly, the more sophisticated the text mining technique used, the richer the description of individuals. For the sake of simplicity, we focus, here, on a single attribute, *Company*, which holds the names of the companies (from a pre-compiled list of 1,914 company names) that may appear in a blog entry. For each blogger, we add the company name value *X* to the attribute *Company* whenever *X* occurs in the body of one of that blogger's entries.

Figure 3(a) shows the global evolution of bonding potential in the IAN community during our experiment. The tick marks at the top of the graph correspond to weekends. The community appears to be bonding overall, as might be expected with a single attribute. Interestingly, however, there appears to be a kind of weekly cyclical activity. It is most visible in the first week, but seems to hold, although to a lesser degree, in the following weeks. The beginning of the week is marked by a decrease in bonding potential, followed by a period of overall increasing bonding, suggesting that bloggers may be bridging out early on and bonding later. This could be due to more prolific or active bloggers that are on the look-out for new information (especially about companies, here) that they add to their blogs early on in the week, followed by a group of less active bloggers (or weekend bloggers) that later catch up with current conversations.



(a) Bonding Potential (b) Connectedness

Figure 3: Blog-based Community Evolution

The number of active bloggers, not shown here, does indeed seem to peak towards the end of the week. Figure 3(b) shows a histogram of the number of bloggers connected to a given fraction of other bloggers in the community. The heavy tail, and the significant number of bloggers connected to between 30% and 40% of the others, suggest that many bloggers are implicitly connected to a relatively high percentage of the community. This is in contrast to the micro-communities found in [5].

One potential benefit of IANs in the blogsphere is that unlike hyper-linked cross-references, IAN links 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. 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). Knowing where bloggers fit in the IAN today provides information on what they need to do to get where they would like to be in the future.

5. Conclusion

We have shown how to generate a novel class of individual-centered social networks, known as implicit affinity networks. Rather than being built around an explicit relationship, these networks capture dynamic, multi-faceted relationships implicit in the shared characteristics or attributes of individuals. We have discussed the use of the notion of social capital to measure the evolving potential of a community, and have used it to report on experiments with two Web communities, one built around interests and the other around blog content.

In addition to extending the use of IANs to other areas, such as online health communities, to detect trends (e.g., in symptoms, experiences and feelings) that may otherwise remain undetected by physicians,

we are working on the idea of overlaying the IAN of a community with the explicit social network (ESN) of the same community. This has already led to the development of measures of bonding and bridging social capital that are not reciprocal. The resulting hybrid network should provide a clean formalism to track the actual social capital of a community more accurately, and thus serve to effectively model important problems in the political and social sciences.

References

[1] Belliveau, M.A., O'Reilly, C.A. III and Wade, J.B. (1996). Social Capital at the Top: Effects of Social Similarity and Status on CEO Compensation, *Academy of Management Journal*, **39**(6):1568-1593.

[2] Burt, R. (2008). Network Duality of Social Capital, in Bartkus, V. and J.H. Davis (Eds.), *Reaching Out, Reaching In: Multidisciplinary Perspectives on Social Capital*, Edward Elgar Publishing.

[3] Johnson, G.J. and Ambrose, P.J. (2006). Neo-Tribes: The Power and Potential of Online Communities in Health Care, *Communications of the ACM*, **49**(1):107-113.

[4] Katz, J.S. (2005). Scale Independent Bibliometric Indicators, *Measurement: Interdisciplinary Research and Perspectives*, **3**:24-28.

[5] Kumar, R., Novak, J., Raghavan, P. and Tomkins, A. (2003). On the Bursty Evolution of Blogspace, in *Proceedings of the 12th International Conference on World Wide Web*, 568-576.

[6] Kumar, R., Novak, J. and Tomkins, A. (2006). Structure and Evolution of Online Social Networks, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 611-617.

[7] Leskovec, J., Kleinberg, J. and Faloutsos, C. (2005). 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*, 177-187.

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

[9] Putnam, R.D. (2000). *Bowling Alone: The Collapse and Revival of American Community*, NY: Simon & Schuster.

[10] Putnam, R.D., Feldstein, L.M. and Cohen, D.J. (2003). *Better Together: Restoring the American Community*, NY: Simon & Schuster.

[11] Redner, S. (2005). Citation Statistics from 110 Years of *Physical Review*, *Physics Today*, **58**:49-54.

[12] Scoble, R. (2007). *Scobleizer's Blogs*, online at: http://www.bloglines.com/public/scobleizer.

[13] Scott, J. (2000). *Social Network Analysis: A Handbook*, SAGE Publications.

[14] Tantipathananadh, C., Berger-Wolf, T. and Kempe, D. (2007). A Framework for Community Identification, in *Dynamic Social Networks, in Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 717-726.

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

Implicit Affinity Networks and Social Capital[1]

M. Smith[a], C. Giraud-Carrier[a] and N. Purser[b]
[a]Dept. of Computer Science, Brigham Young University, USA, {cgc@cs.,smitty@}byu.edu
[b]Omniture.com, USA, {natepurser@gmail.com}

Abstract
Social networks are typically constructed based on explicit and well-defined relationships among individuals. In this paper, we describe another class of social networks, known as Implicit Affinity Networks (IANs), where links are implicit in the patterns of natural affinities among individuals. An effective mathematical formulation of social capital based on implicit and explicit connections is given. Results with two Web communities, one focused on people's interests and one focused on people's blogs, exhibit rich dynamics and show interesting patterns of community evolution.

1. Introduction

Online communities, also referred to as neo-tribes [10], have sprung up all over the Internet. 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. The resulting social networks are relationship-centered, and their analysis typically assumes that the network is static, or evolves in a sufficiently slow manner to make the study of snapshots relevant and meaningful. Much work has been done to capture, understand, and model the structure of such social networks (e.g., see [24,29]).

Although useful from a practical (computational) standpoint, the assumption of a static network tends to limit the kinds of analyses that may be performed. Recently, some researchers have begun to study the actual 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 [11,13,14,22,28]). It is possible to go even further by focusing on implicit affinities thus allowing the nature of the underlying relationship to vary over time. In this context, individuals are viewed as social actors characterized by a wide range of attributes, and relationships among them emerge naturally as a result of commonalities across attributes. Unlike traditional, relationship-centric social networks where links represent *explicit* relationships, the links in this individual-centric approach are based strictly on affinities, or inherent similarities, among the social actors, which create *implicit*, and multi-faceted, relationships (i.e., the sharing of characteristics induces some level of similarity or strength of affinity among actors). Because individuals are complex entities whose attitudes and behaviors change over time, these networks are intrinsically dynamic, and evolve naturally through time (e.g., with such factors as their participants' age, occupation, interests, and life circumstances).

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." [2] (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 describe how IANs can be generated from information about individuals, to visualize and analyze affinities among groups of these individuals. We then show how to build hybrid social networks from IANs and ESNs to derive an effective mathematical formulation of social capital. Finally, we report on the early evolutionary stages of a Web community based on implicit affinities as well as 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.

2. Implicit Affinity Networks

We represent individuals by collections of attributes and associated (discrete) value sets. Each attribute captures some information about individuals, such as occupations, hobbies, research interests, birthplace, etc. In our context, an individual may be characterized by any number of attributes and each attribute may have any number of its possible values (e.g., John=<hobbies:{hiking, reading}, languages:{English, French}, hair:{brown}>,

1 This is an extended version of our WITS paper [26], with material from [27], as well as new material.

Becky=<hair:{brown}, eyes:{blue}, hobbies:{scrapbooking, skydiving, reading}>).

Using the attribute information, we can compute a degree of similarity between individuals, which we call an affinity score. For a given attribute, the affinity score corresponds to the amount of overlap among that attribute's values between the two individuals, and thus ranges from 0 (no overlap) to 1 (all values shared). When there is more than one attribute, the overall affinity score between two individuals is simply the sum of their attribute-level affinity scores normalized by the number of attributes they have in common. Formally, let □ be a set of attributes, and for each attribute A□□, let V_A denote the arbitrary value-set of A. For any individual i, let $Attr(i)$□□ denote the set of attributes of i, and $V_A(i)$□V_A denote the set of values of attribute A for individual i. Then, we define the overall affinity score between individual i and individual j by:

$$s_{ij}^{IAN} = \frac{\sum_{A \in Attr(i) \cap Attr(j)} AffScore_A(i,j)}{|Attr(i) \cap Attr(j)|}, \text{ where } AffScore_A(i,j) = \frac{|V_A(i) \cap V_A(j)|}{|V_A(i) \cup V_A(j)|} \times \alpha_A$$

When i and j share no attributes (i.e., $Attr(i) \cap Attr(j) = \emptyset$), their affinity score is 0. Note that because our focus here is on the analysis of the network rather than the specific underlying similarity function, and because we assume that attributes have discrete values, we have chosen a relatively simple similarity function, based on Jaccard's index. In principle, any similarity function defined over pairs of individuals may be employed to build an IAN. 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.

The term □$_A$ in $AffScore_A(i,j)$, is an optional weighting factor. This weight may be used to reflect the relative importance of A in a community. In the most general case, □$_A$ is a composite of individual user preferences and a mined community preference. The former is elicited from individuals, e.g., using a kind of 5-star rating, where 5 stars may correspond to □$_A$=1 and 1 star to □$_A$=0.2. The latter is the ratio of the number of individuals that have at least one value for attribute A to the total number of individuals in the community. Hence, it acts as a global, learned, community weight that evolves with changes in the behavior of individuals and that favors frequently used attributes.

The set of affinity scores over a group of individuals may be naturally represented in matrix form, and indeed, most of the computations discussed in the remainder of the paper may be performed in that context. Note that the use of affinity scores and the corresponding matrix is essentially a transformation of what may be viewed as 2-way 2-mode data, in the spirit of [4], where one mode is the set of individuals and the other is the set of attribute-value pairs, into 2-way 1-mode (individuals) data. Although the former could be pursued, it makes little sense here as 1) in most cases, the 2-mode matrix will be very sparse, and 2) we will be using affinity matrices in conjunction with relationship matrices (or networks), which are inherently 1-mode. Another significant difference from [4] is that our matrices are dynamic.

Here, however, instead of the matrix form, we choose to use the corresponding graph or network representation, which we call an implicit affinity network (IAN). When edge thickness is used to express the relative affinity score (the thicker the stronger the affinity, with missing edges corresponding to 0 scores), the graph representation provides a compelling mechanism to visualize the community, especially as it evolves over time. Additionally, the graph representation is most useful when overlaying IANs with explicit social networks as shown in section 3. In this context, not only is the representation rather natural, it is also generally more compact than the corresponding matrix, which may be rather sparse.

Any time an individual i adds a value to one of its attributes all affinity scores between i and the other individuals in the community are immediately updated. Since we are interested in tracking evolution, our networks are actually time graphs, as defined in [12], where every change in the network is time-stamped with the time at which it was made.

3. Social Capital for Community Tracking

Several explicitly quantitative measures have been proposed to capture information about the structure and evolution of communities in social networks, including nodal degree, diameter and density (e.g., see [29]). Here, instead, we focus on the more qualitative, yet rich, idea of social capital, used by sociologists, political scientists, economists and others (e.g., see [16]), and seek to establish a quantitative context for it. As it turns out, such an endeavor is not trivial, and depends in part on the perspective one adopts.

When considering social capital, the focus 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 other's) views of social capital using a 2x2 table, which considers both type of actor and type of focus, as shown in Table 1 [5].

	Type of Focus	
Type of Actor	**Internal**	**External**
Individual		One's 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 [5])

There are further variations on these views of course. For example, Hobbes suggested that having a few powerful friends is more important than having many powerless friends [9], 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 [15]. We do not pursue these here.

The social capital measures we present in this paper belong to the group-internal category, where we focus on measuring a network's overall social capital based on links among individuals within the selected network. This view of social capital has been championed by Putnam, who further divides it into two main components: bonding social capital and bridging social capital [19,20].

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. Associations and clubs typically create bonding social capital; neighborhoods and choirs tend to create 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 [6]. Individuals may seek to bond to enlarge their support group, to focus their attention, or to galvanize their efforts; or they may seek to bridge to reach out to others (e.g., philanthropic activities), broaden their horizons, or capitalize on mutually-beneficial collaboration (e.g., cross-disciplinary research). In principle, there is no dichotomy between bonding and bridging. Either, both or neither may be accrued at any one time.

3.1. Actual vs. Potential Social Capital

Because individuals are complex entities whose attitudes and behaviors are many, small changes to one individual's profile may have a number of (unexpected) effects on the overall structure of the IAN. While it is clear that some changes may be sudden, such as becoming a father, and others are more gradual, such as becoming an avid chess player, no distinction is made here; at some point in time, the individual exposes his/her new self to the rest of the network.

Every time an individual's profile changes, typically by adding a new attribute or a new value to an existing attribute, the corresponding update creates an opportunity for existing implicit connections to be strengthened or 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. Changes to an ESN are more purposeful and localized. An individual chooses precisely which other individuals to connect with. In that sense, IANs capture only the *potential* for social capital, rather than social capital itself. Social capital only actually accrues when individuals become aware of it, that is, when they establish explicit and intentional relationships with each other, as part of an ESN.

Hence, we define a *hybrid social network* as the combination of an implicit affinity network and an explicit social network defined over the same set of individuals. Hybrid networks can be visualized by overlaying ESNs onto corresponding IANs. In social network analysis terminology, a hybrid network is a multigraph having both implicit and explicit relations amongst its actors. Table 2 together with Figure 1 provides a simple example of a hybrid social network. The dashed lines are implicit links while the solid lines are explicit links.

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

Table 2: Sample Individuals and Attributes.

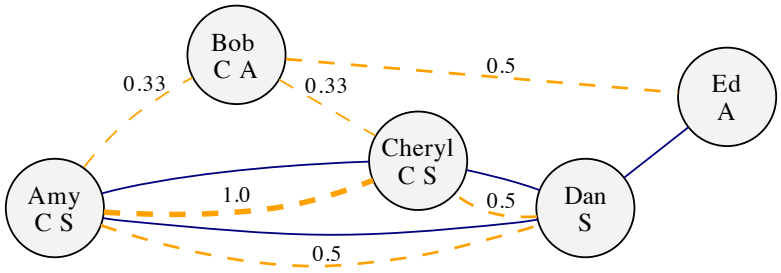


Figure 1: Sample Hybrid Network

The values on the implicit links correspond to their strength, i.e., the affinity score. Here, all explicit links are assumed to have the same weight or strength. This, of course, need not be the case and individual weights may also be placed on explicit connections, as discussed below.

We contend that hybrid social networks provide basic components that contribute to measuring *actual* social capital. Depending on the kinds of connections that may exist among the same individuals, one can also determine what form of social capital, bonding or bridging, is being affected and how, as summarized in Table 3.

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

Table 3: Potential vs. Actual Social Capital in Hybrid Networks

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. 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. 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.

Note that the notions of bonding and bridging discussed here are different and somewhat orthogonal to the idea of near and far (or strong and weak) ties introduced in [8]. In this latter context, it is mostly the frequency of interactions among actors that determines the strength of their connection, or bond. According to the definitions of bonding and bridging we use, however, individuals that interact a lot (a thing that "happens" at the ESN level, since it is voluntary and hence explicit) do cause an increase in *actual* social capital (see Table 3), but the kind is determined by their similarity (a thing that "happens" at the IAN level). Consider again the example of clubs and choirs, as mentioned above. In both cases, there is varying level of interaction among members, but even with high levels of interaction, clubs create bonding since members tend to be similar (at least in terms of the club's focus), while neighborhoods (typically) create bridging since actors have no reason a priori to share affinities. Similarly, a group of elderly cancer patients and young healthy people who interact a lot are bridging (barring any other affinity among them). The nature and frequency of their interaction (e.g., lunch together everyday) would be represented in the ESN strengths, while the attributes (e.g., age, cancer) would be represented in the IAN strengths (i.e., affinity scores). Hence, (actual) social capital may be viewed as both structure-based, as suggested in [8], and affinity-based, as advocated in [19,20] and pursued here.

There is neither *actual* bonding nor *actual* 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 only as explicit links are made or removed. Both implicit and explicit connections are necessary to calculate the network's social capital. It is clear that we may not be able to identify all of the attributes of an individual that may form a bond. That potential may be there but unknown to us as no implicit links are found; and so a link (in the hybrid network) that we label as bridging may actually be bonding. There is no way to avoid that. Our hope is that using the available data, and observing the dynamics of the network, will provide enough information to improve network understanding, and through time reveal the true nature of the embedded social capital.

3.2. Bonding and Bridging Social Capital

Note that although the notions of bonding and bridging have been discussed and used in various studies, they have not yet been really operationalized. We try to do so here, by providing a quantitative context for has been treated mostly qualitatively so far. We first define potential social capital, and then derive a formulation for actual social capital.

3.2.1 Potential Social Capital

We define bonding and bridging potentials for a network as reciprocal of each other by the following formulas, where N denotes the number of nodes and E the set of edges in the network:

$$BondingPotential = \frac{2}{N(N-1)} \sum_{\{i,j\} \in E} s_{ij}^{IAN}$$

$$BridgingPotential = 1 - BondingPotential$$

We note that *BondingPotential* is essentially a weighted version of the density measure defined in [29], where the edges' weights are given by the affinity scores. One of the unique features of IANs is that these weights are not fixed, but naturally adapted as the profiles of individuals change. Specifically, for an edge $\{i,j\}$, if i adds a value to one of its attributes, say A_k, and that value is not shared by j, then the weight of $\{i,j\}$ decreases. Of course, this in turn also decreases the bonding potential between i and j. This may seem counterintuitive as j may not be aware of the value added by i. If j were to also have that value but has simply not provided it yet, then one could argue that the computation would be inconsistent with the true state of the network. Stated otherwise, this suggests that we may need to treat "unknown" values differently from the way we treat "missing" values.

Rather than making this distinction, we propose two simpler alternatives. In the first one, whenever a new attribute (or a value) is added, the new attribute (or value) is broadcast to the network so that every individual may update his/her profile (i.e., decide whether or not it is applicable to them), and affinity scores remain unchanged until all individuals have had a chance to react to the change. In the second one, the new attribute (or value) is not broadcast, affinity scores are updated immediately (causing a temporary decrease in bonding), and the system waits the natural process of time for things to adapt, in hope that over time individuals will add the appropriate attributes if they are applicable to them. Thus affinity scores, and subsequently potential social capital, are continually changing through time as individuals create or update their profile. In our current implementation, we use the second approach as it requires no additional computation across the network and seems to be more natural. In real life, affinities are discovered in the process of time (e.g., through interaction); there may be more potential bonding available (or it may not change because of one individual's change), but until individuals make their values known, this cannot be detected.

Although we define how much potential bonding and bridging exist within a network, we cannot predict how much of that will be "actualized." As stated above, actual bonding and bridging do not occur merely because people have or do not have affinities, but when explicit links are established. In that sense, it seems reasonable to consider potential bonding and bridging as reciprocal, as per the above definition. Consider a simple 2-individual network. Let us say that there are a total of A affinities the two may share (and no more); for each one, they either share it or they do not; each one they share gives an opportunity to bond (potentially). Let us say they share k of the A affinities, then potential bonding in our model is k/A. The remaining A-k affinities may be viewed as offering an opportunity for bridging; hence potential bridging is $(A-k)/A$, or 1 minus potential bonding. Given that there is a (theoretically) finite number of affinities, or amount of potential social capital, every time one kind of potential capital increases, the other one must decrease.

Finally, recall that the measures presented here belong to the group-internal category, hence the "normalization" by the total number of possible links, in the above equations. Indeed, we divide the sum of implicit strengths within

the network by the amount of implicit strengths possible (similar to how density is computed), thus factoring out the size of the network and making comparisons across networks meaningful. This is particularly important in our context where networks evolve dynamically over time. Given one such dynamic network, we may track changes in its underlying potential social capital by considering the network's instantiations at each "time-step" as individual networks and using the above formulas for each instantiation. For example, if at time t_1, the network consists of two individuals, *a* and *b*, that are perfectly similar to each other, the resulting bonding potential is 1. If at time t_2, two more individuals *c* and *d* join the network, where *c* and *d* are perfectly similar, but neither has any affinity with *a* or *b*, then the bonding potential naturally drops to 2/6=1/3, since only two of the possible six implicit connections among the four individuals are present. While for each subgroup, {*a*, *b*} and {*c*, *d*}, the bonding potential is the same (namely 1), at the network level potential bonding has decreased. It is clear that such results would seem counter-intuitive if one were working under, for example, an individual-external perspective of social capital. Given our group-internal view, the normalization is warranted, even if, on the surface, it seems to give more weight to the number of individuals

3.2.2. Actual Social Capital

We now turn to the computation of actual social capital, which as stated above requires both implicit and explicit links. In general, all connections, or edges, have an associated strength or weight. 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 (see section 2). 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). The notion of frequency of interaction from [8] would also offer a viable weighing mechanism for ESN links. For example, we might generate an ESN based on email activity. Strong ESN scores (i.e., near 1) would be assigned among individuals that regularly exchanged emails, while weaker scores (i.e., near 0) would be assigned among those that were only sending an occasional email to each other.

Actual bonding social capital between two nodes *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,

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

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.

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

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.

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

One important aspect of the above formulation is that, although the two kinds of *potential* social capital are reciprocal as explained, *actual* bonding social capital and *actual* bridging social capital are not. 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:

> *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.* (Putnam, personal communication)

Our formulation is not merely a mathematical trick, but is rooted in what we understand to be the nature of actual vs. potential bonding and bridging social capital. Cast in our hybrid network framework, we would consider the "friend" relationship as explicit and the "race" or "skin color" attribute as implicit. Thus, in a hybrid network, the IAN part would consist of the connections among white individuals and among non-white individuals, while the ESN part would consist of the connections among friends within and across these groups. The kind of white individuals referred to by Putnam would have many explicit connections with individuals who differ from them (non-white) as well as individuals like them. In our framework, such individuals are characterized by *both* high bonding social capital and high bridging social capital, thus accurately modeling the underlying empirical finding.

Consider yet another example that illustrates the difference between potential and actual, and bonding and bridging, social capital. Often actual bridging will arise when there is mutual benefit or a mutual cause. This is exactly the information that could be input to our model. Let us say that two individuals *X* and *Y* in a community have nothing in common aside from both being in favor of a local municipal initiative. How strongly explicitly connected *X* and *Y* are (e.g., how frequently they meet with one another) determines the amount (and which type) of *actual* social capital exists between them. If *X* and *Y* share no explicit link (e.g., they do not know each other), then there can be neither actual bonding nor bridging. If they do, then there is *actual* bridging. Of course, there could be *actual* bonding too; maybe *X* and *Y* both play tennis and enjoy Mexican food. This is similar to the situation of two researchers in two different disciplines, with maybe little in common, deciding to work together on cross-disciplinary issues. There is clear bridging even if there may be little or no bonding.

4. Experiments

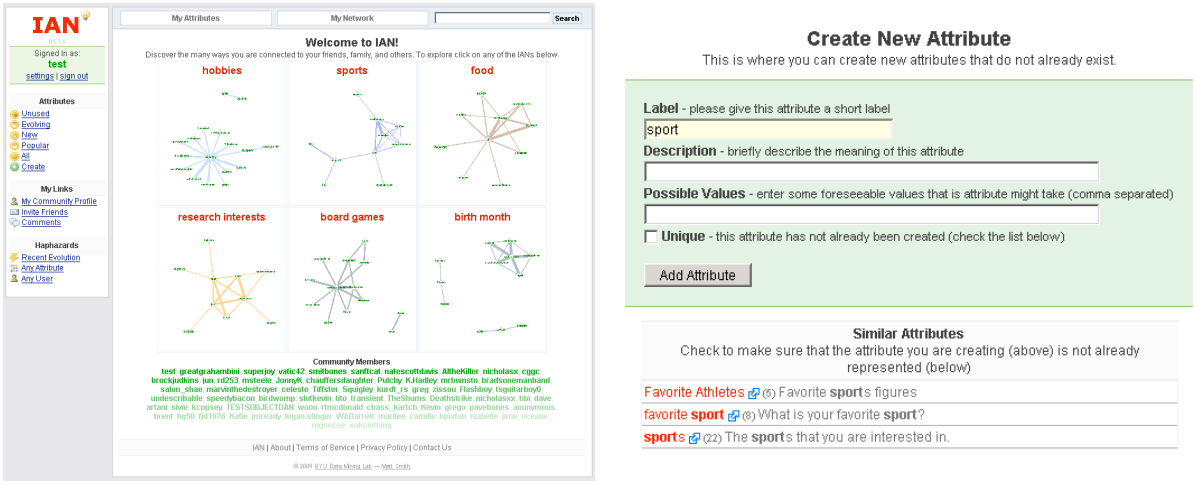
In this section, we report on two experiments with implicit affinity networks and community tracking using our definitions of social capital.

We note, at the outset, that there is no gold standard for social capital. Given the several definitions proposed for social capital, validating any approach is difficult and typically done against its underlying assumptions rather than some accepted ground truth (e.g., see [18,25,15]). We follow a similar pattern here.

4.1. Implicit Affinities and Potential Social Capital

As mentioned earlier, many existing online communities are based on a single type of explicit relation among individuals. Even when additional data is available about individuals beyond the relation itself, such data typically lacks the time element necessary to analyze the evolution of implicit affinities required by IANs. Hence, for our first experiment, we implemented a Web application that allows individuals to create and edit their profile in the form of dynamic attribute-value sets, where each attribute captures some characteristic or personal dimension of interest that individuals wish to be represented by and share (see Figure 2). Any and all changes to attribute and attribute-values

are time-stamped.



(a) User's Home Page (b) New Attribute Creation

Figure 2: Interests-based Community

In its current form, the IAN community is a general community that enables sub-communities to emerge over a variety of topics. Each member of the community is described as a tuple of attribute-value pairs. For added flexibility, the Web application allows attribute-level views (i.e., one attribute at a time) as well as aggregated views (several attributes combined) of IANs. Affinity scores and social capital are computed on the aggregated IAN as discussed above.

The experiment was started gradually by inviting a user or two at a time, and allowing these users to invite others as desired. We did not collect any demographics about these individuals, other than what they provided in their profile. It is very likely that the word-of-mouth approach means that many of the participants know each other. Although we only focused on the implicit affinity network here, and thus this has no impact on *actual* social capital, it may still introduce a bias in favor of bonding in *potential* social capital.

As of a year from its inception, there were 72 individuals signed up. On average, a user:

1. was active within the community for 52 days,
2. visited the site every 10 days,
3. added 2 attribute-values to their profile per visit, and
4. had 95 attribute-values across 21 attributes.

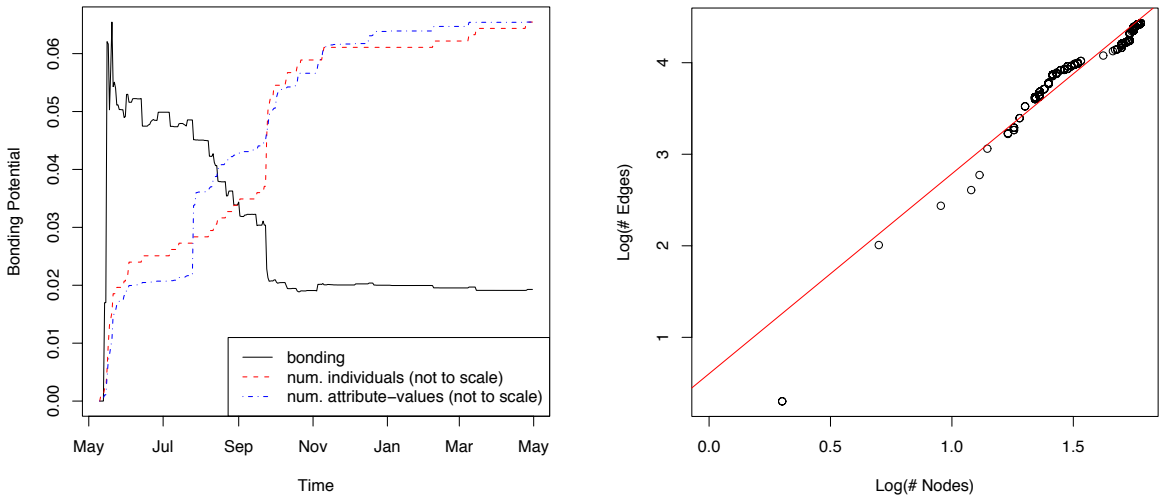
Recent work on social networks has examined the evolution of the average degree of nodes (i.e., $2|E|/N$) over time [13,14]. We wish to do the same with bonding potential. Figure 3(a) shows the global evolution of bonding potential in the IAN community during our experiment. The number of individuals and the number of attribute-values are also shown to facilitate interpretation.

The overall trend in the evolution of bonding potential is decreasing, as might be expected of a still fairly new and rather varied community. Recall that the community was started by one or two individuals who created a few attributes and then invited others to join. Hence, the number of attributes (and values) is rather small at the beginning, as seen on Figure 3(a). As new individuals join the community, they may use some of these attributes' values, but they are most likely to add new values and define their own attributes, that are more relevant to them. This is also apparent in the closely related increasing trends for the number of individuals and number of attribute-values displayed in Figure 3(a). As a result, in addition to the natural effect of the number of individuals on the value of potential social capital (see section 3.2.1) the tendency is to bridging (i.e., adding values to profiles that differentiate people) rather than bonding (i.e., capitalizing on what values are already there).

At a lower level, the graph may be split into four time periods, which provides a more detailed view of the above phenomenon:

- May. This is the "birth" of the community. As one might expect, bonding potential rises as a small number of people join in and begin sharing values on a small set of attributes.
- June-July. This is a period of relative stability, where the number of attribute-values remains relatively constant and only a few new individuals join the community. Note that arrival of new individuals generally results in a short-time drop in bonding potential, followed by an increase, as bonding replaces bridging.

- August-October. This is a period of high activity, partly due to our extending the availability of IAN. During this period a significant number of individuals join IAN and create a significant number of new attribute-values, faster than current members can exploit, thus leading to a decrease in bonding potential. New members are "casting their lines out," attempting to bridge out by offering new possibilities for affinities with current and new members.



(a) Bonding Potential (b) #Edges vs. #Nodes

Figure 3: Interests-based Community Evolution

- November-May. As the number of individuals and the number of attribute-values begin to stabilize again, bonding potential plateaus out. The addition of new individuals and new attributes, which causes small troughs on the bonding potential curve, seems to be compensated by the capitalization of individuals on existing attribute-values, i.e., bonding with others rather than bridging out.

Recent studies of dynamic social networks have highlighted characteristics or laws that seem to have broad applicability. In particular, it seems that social networks exhibit densification (i.e., the relation of the number of edges to the number of nodes follows a power law, $E(t)=N(t)^a$ for $1<a<2$) and shrinking diameters (i.e., the 90th percentile of the shortest path lengths between all pairs of nodes decreases over time) [14]. We wish to see whether IANs obey similar laws. We restrict our attention to densification, realizing that our network is still relatively small at this stage. Figure 3(b) plots the log of the number of edges versus the log of the number of nodes, when edges are aggregated across the values of each attribute (i.e., at most one edge per attribute between any 2 nodes). As can be seen, it appears that densification in IANs also follow a power law. Given the way new links arise in IANs, i.e., with probability proportional to the richness of existing individuals' profiles rather than the richness of their connections, we might expect that the exponent be larger than 2. Indeed, the slope of the regression line is 2.18, with $R^2=0.92$.

4.2. Social Capital in the Blogosphere

In our second experiment, we generate and analyze an implicit affinity network within the Blogosphere. 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 under-exploited. 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 [7]. 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 Scoble's blog [23]. 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	real estate, pet food, real estate marketing, technorati tags, articles tagged, ny times, wheat gluten, menu foods, north carolina, science blogging conference
2	en el, los usuarios, de forma, se trata, de los, comp rtelo, nos permite, im genes, de momento, una serie
3	united states, white house, president bush, bush administration, york times, middle east, years ago, health care, washington post, national security
4	posts filed, search blogs linking, sam rubenstein, slam online, lang whitaker, xbox live, visit thebbps forums, san antonio, peoples champ, golden state
5	supreme court, death penalty, district court, law school, ninth circuit, justice scalia, law review, lethal injection, justice kennedy, oral argument
6	related articles, technorati tags, windows vista, tablet pc, search engine, open source, social media, web site, search engines, related posts
7	personal finance, real estate, credit card, interest rates, paul krugman, federal reserve, monetary policy, social security, united states, credit cards
8	fourth quarter, stock symbol, earnings call transcripts, related articlesread, etfs type, related stocks, cash flow, seeking alpha, conference call, email alerts
9	paris hilton, fashiontribes fashion, american idol, los angeles, britney spears, related posts, high school, lindsay lohan, years ago, hot product
10	time warner, general electric, general motors, competitive strategy, linking blogs, apple computer, ford motor, consumer experience, mart stores, sirius satellite radio

Table 4: 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 [21] entries were retrieved for all blogs on a level.
2. All hypertext links were extracted from the blog entry content.
3. We determined whether or not the URL in the link was to another blog by parsing the HTTP headers for a content-type that implied it was a blog. 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 blog. 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 blog and did not consider it in our analysis.

Using this pattern, we retrieved all entries for blogs located within two levels of Scoble. We then 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 224 blogs, within the first two levels, that had at least one substantial explicit relationship to another blog.

Next, we constructed the IAN as follows. We applied Latent Dirichlet Allocation (LDA) [3] to model prevalent topics in the blog entries throughout the 12 months of the experiment. To determine the *n*-grams within each topic, we chose to input all the entries used by the 224 blogs identified in the previous paragraph. The ten topics, shown in Table 2, were generated using MALLET's implementation of LDA [17]. Based on this list, we determined whether a blog was a member of a topic group by checking if its entries contained at least half of the *n*-grams from that particular topic. Each blog was then characterized by a single attribute, *topics addressed*, whose values were all of the topic groups it belonged to. For example, if a blog *X* contained at least ten of the bi-grams in topic 3 (e.g., white house, president bush, …, health care) and at least half of the bi-grams from topic 9 (e.g., american idol, lindsay lohan, high school, …, britney spears) it would be a member of topic groups 3 and 9 from Table 4 and thus would be represented as *X*=<topics addressed:{3, 9}>. For simplicity, we restricted our attention to affinities of score 1, i.e., where implicitly linked blogs addressed exactly the same set of topics.

Finally, we created the resulting hybrid network consisting of 224 nodes, representing blogs, and 2,664 links, 580 of which were explicit and the other 2,084 were implicit. Note that, to keep computations tractable (especially LDA), we have ignored the time element in both the IAN and ESN networks here, and assumed that topics and cross-references were not changing over time. We will revisit this issue in future work.

Figure 4, drawn using the "Organic" layout in the Cytoscape 2.6.0 software package, shows a graph of this network. In the graph, each node represents a blog while each edge represents reciprocal links (resulting in 1,332 links: 290 explicit and 1,042 implicit). The darker, solid blue lines between blogs represent explicit links and the

lighter, dashed orange lines represent implicit links. Additionally, the ESN and IAN components of the hybrid network are shown by themselves to aid in the analysis (Note: 60 isolates that are not implicitly connected to other blogs are not shown in the IAN network). Six regions labeled in the graph, to facilitate discussion in following paragraphs, are centered near the most significant affinity cliques described in Table 5.

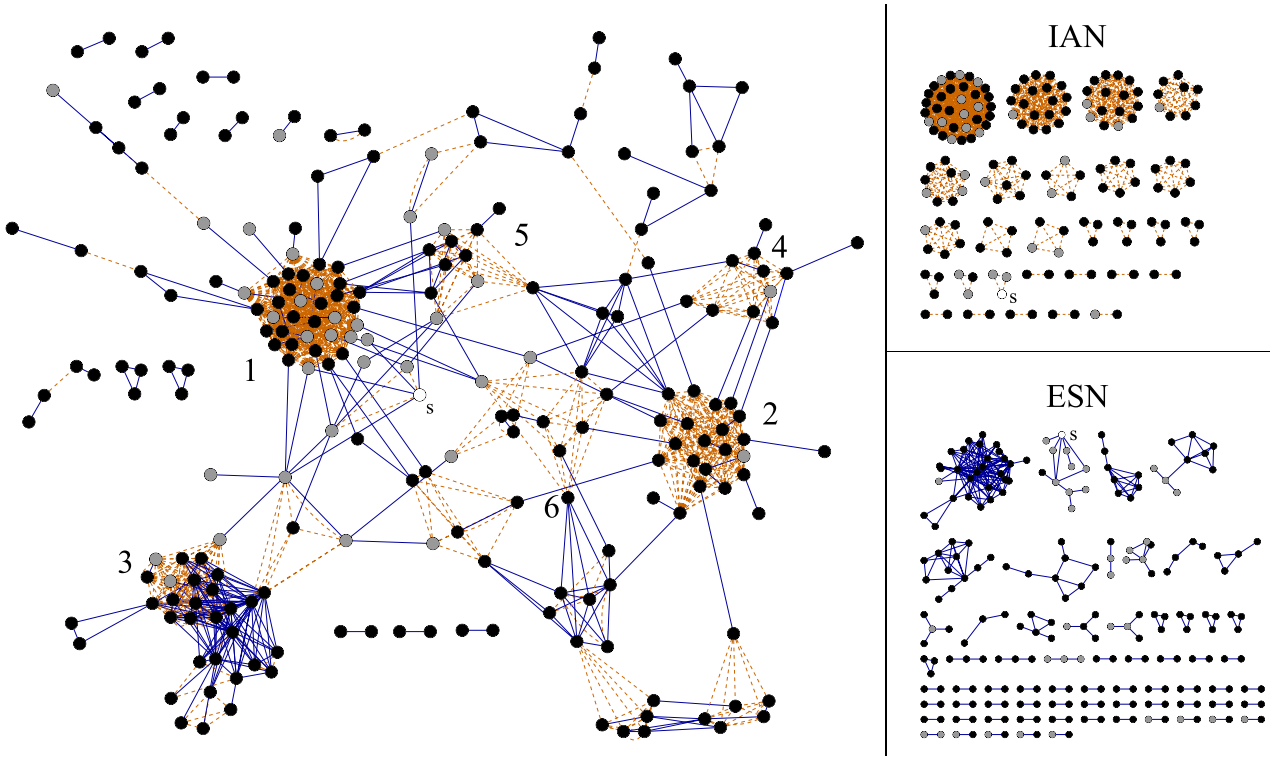


Figure 4: Hybrid Network for Blog Experiment (regions discussed are labeled 1-6, Scoble is labeled with an `S') The corresponding explicit social network (ESN) and the implicit affinity network (IAN) are shown separately to the right. Nodes colored grey are one degree of separation from Scoble, while nodes colored black are two degrees away.

Clique	Topics Addressed	Size
1	6	34
2	3	19
3	3,6,7,8,10	16
4	3,7	9
5	3,6	9
6	3,6,7,9	7

Table 5: Topic Sets Addressed by Significant IAN Cliques

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. Part of this may be due to the fact that we are only considering two levels of separation out. It is probably true to some degree, but the effect may diminish as we go another degree of separation out. The following are worthy of note:

- Region 1, centered on the most prominent affinity clique, tends to focus solely on Topic 6. Although there are nine explicit links among this group, which indicate some amount of actual bonding has occurred, the majority remains explicitly unconnected. This suggests that there is potential for bonding in this region.
- Region 2 is connected implicitly by Topic #3 making a clique of 19 blogs. Three cross-references are present among the group to each other, while 19 cross-references are to other blogs outside of the clique. Thus, approximately 74% (14 of 19) of the IAN clique is bridging while 26% (5 of 19) is participating solely on bonding. One blog (i.e., http://drsanity.blogspot.com) is both bonding and bridging. As an additional observation, through manual introspection of these blogs, a number of

pictures were observed on blog entries, often more prevalent than the corresponding text.

- Region 3 includes the affinity clique that addresses five topics and the largest ESN component in the graph. It is somewhat of an anomaly, as all of the blogs strictly in the ESN component of the region are sub-domains of the single blog, bloggingstocks.com. This occurs because the site automatically includes links to the relevant sub-domains (i.e., one for each ticker symbol) for every post. On the other hand, the IAN component of this graph is connected to other sites that cover similar topics including ResourceShelf.com and SeekingAlpha.com. Thus, actual bridging and potential bonding are observed.
- Similarly, near Region 6, there is a dense number of explicit links that connect the blogs together. However, the linking of these blogs is not automated, but agreed upon. In their words, "Each Sunday, the editors of every site—from LAist to Londonist—choose their most interesting article, a list that is compiled into the network-wide feature Elsewhere In The Ist-a-Verse." (http://torontoist.com/2008/11/elsewhere_in_the_istaverse_23.php). Interestingly, despite the fact that this group is well connected explicitly, the blogs in this network are members of five different IAN cliques, as they address different topic sets. This is evidence of actual bridging.
- Region 4 is a clique of political blogs, which offer opportunities for actual bonding. This is a set of blogs that address both Topic #3 and #7, which deal with both politics and money.
- Throughout the graph there are several implicitly connected regions with few explicit links among them (e.g., regions 1, 2, 4, and 5). This presents a significant amount of potential bonding that could occur to create new sub-communities. For instance, region 5 includes blogs with content about search and social media. They do not link to each other explicitly although they do have a strong tendency to address the same topics. Capitalizing on such links (through explicit connections) could add value to members of these communities who would suddenly have access to new resources (in the form of complementary blog contents) that they may have ignored up to this point.

Such hybrid network analyses are particularly interesting to bloggers within these networks, as they describe where their blog resides within some greater community. So far, our model (and corresponding analysis) is descriptive rather than predictive. No claim is made as to what specific action is recommended. However, bloggers may evaluate where they are and where they hope to be within the community being analyzed.

5. Conclusions and Future Work

We have shown how to generate a novel class of individual-centered social networks, known as implicit affinity networks. Rather than being built around an explicit relationship, these networks capture dynamic, multi-faceted relationships implicit in the shared characteristics or attributes of individuals. 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.

We have used our measure of social capital to report on experiments with two Web communities, one built around interests and the other around blog content. 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 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. 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 between entities, particularly 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 individuals. 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 (or other online social actor) 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

[1] Adler, P.S. and Kwon, S-W. (2002). Social Capital: Prospects For a New Concept. *The Academy of*

Management Review, **27**:17-40.
[2] Belliveau, M.A., O'Reilly, C.A. III and Wade, J.B. (1996). Social Capital at the Top: Effects of Social Similarity and Status on CEO Compensation, *Academy of Management Journal*, **39**(6):1568-1593.
[3] Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research* **3**:993–1022.
[4] Borgatti, S. P., and Everett, M. G. (1997). Network Analysis of 2-mode Data. *Social Networks*, **19**(3):243-269.
[5] Borgatti, S. P., and Everett, M. G. (1998). Network Measures of Social Capital. *Connections*, **21**(2):27-36.
[6] Burt, R. (2008). Network Duality of Social Capital, in Bartkus, V. and J.H. Davis (Eds.), *Reaching Out, Reaching In: Multidisciplinary Perspectives on Social Capital*, Edward Elgar Publishing.
[7] GoogleReaderAPI (2008). Available online at http://code.google.com/p/pyrfeed/wiki/GoogleReaderAPI.
[8] Granovetter, M. (1973). The Strength Of Weak Ties. *American Journal of Sociology*, **78**:1360-1380.
[9] Hobbes, T. Leviathan. Collier, New York, 1962.
[10] Johnson, G.J. and Ambrose, P.J. (2006). Neo-Tribes: The Power and Potential of Online Communities in Health Care, *Communications of the ACM*, **49**(1):107-113.
[11] Katz, J.S. (2005). Scale Independent Bibliometric Indicators, *Measurement: Interdisciplinary Research and Perspectives*, **3**:24-28.
[12] Kumar, R., Novak, J., Raghavan, P. and Tomkins, A. (2003). On the Bursty Evolution of Blogspace, in *Proceedings of the 12th International Conference on World Wide Web*, 568-576.
[13] Kumar, R., Novak, J. and Tomkins, A. (2006). Structure and Evolution of Online Social Networks, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 611-617.
[14] Leskovec, J., Kleinberg, J. and Faloutsos, C. (2005). 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*, 177-187.
[15] Licamele, L. and Getoor, L. (2006) Social capital in friendship-event networks. In Proc. of the Sixth IEEE Int. Conf. on Data Mining (ICDM'06). Pages 959-964.
[16] Lin, N. (2001). *Social Capital: A Theory of Social Structure and Action*. NY: Cambridge University Press.
[17] McCallum, A.K. (2002). MALLET: A Machine Learning for Language Toolkit. Available online at http://mallet.cs.umass.edu.
[18] Narayan, D. and Cassidy, M.F. (2001). A Dimensional Approach to Measuring Social Capital: Development and Validation of a Social Capital Inventory. *Current Sociology*, **49**(2):59-102.
[19] Putnam, R.D. (2000). *Bowling Alone: The Collapse and Revival of American Community*, NY: Simon & Schuster.
[20] Putnam, R.D., Feldstein, L.M. and Cohen, D.J. (2003). *Better Together: Restoring the American Community*, NY: Simon & Schuster.
[21] Pyrfeed (2008). Available online at http://code.google.com/p/pyrfeed/.
[22] Redner, S. (2005). Citation Statistics from 110 Years of *Physical Review*, *Physics Today*, **58**:49-54.
[23] Scoble, R. (2008). *Scobleizer's Blogs*, online at: http://www.bloglines.com/public/scobleizer.
[24] Scott, J. (2000). *Social Network Analysis: A Handbook*, SAGE Publications.
[25] Silva, M.J.D., Harpham, T., Tuan, T., Bartolini, R., Penny, M.E. and Huttly, S.R. (2006). Psychometric and Cognitive Validation of a Social Capital Measurement Tool in Peru and Vietnam. Social Science Medicine, **62**(4):941-953.
[26] Smith, M., Giraud-Carrier, C. and Judkins, B. (2007). Implicit Affinity Networks. In *Proceedings of 17th Annual Workshop on Information Technologies and Systems*, 1–6.
[27] Smith, M., Purser, N. and Giraud-Carrier, C. (2008). Social Capital in the Blogosphere: A Case Study. In *Papers from the AAAI Spring Symposium on Social Information Processing*, K. Lerman et al. (Eds.), Technical Report SS-08-06, AAAI Press, 93-97.
[28] Tantipathananadh, C., Berger-Wolf, T. and Kempe, D. (2007). A Framework for Community Identification, in *Dynamic Social Networks, in Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 717-726.
[29] Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*, NY: Cambridge University Press.

Measuring and Reasoning About Social Capital: A Computational Framework

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

Abstract

Social capital is grounded in the relationships that exist among individuals and the resources that are available to them, or the group as a whole, through these relationships. We propose a general framework for quantifying and reasoning about social capital in resource-aware social networks. In doing so, we bridge the gap between social networks that could leverage social capital but have no explicit resources to be mobilized (e.g., Facebook), and resource-sharing social networks that currently give no thought to leveraging the notion of social capital (e.g., Freecycle). We report on a number of experiments within such an environment that confirms access to social resources through social interaction, generosity, and reciprocity.

Keywords: Social Capital; Social Resources; Social Network Analysis; Social Behavior Modeling

1 Introduction

Whereas most forms of capital are mainly a function of what an individual possesses, social capital is an inherently social concept, grounded in 1) the *relationships* that exist among individuals, and 2) the *resources* that are available to individuals or the group because of these relationships (Coleman, 1988; Lin, 2001; Putnam, 2000). In our work on building a framework for quantifying and reasoning about social capital, our main focus so far has been on relationships, their nature and role. In particular, we have introduced the notion of implicit affinities (Smith et al., 2007), and showed how implicit affinities interact with explicit connections in the creation of social capital (Smith et al., 2008, 2009; Smith and Giraud-Carrier, 2010).

Here, we wish to add the elements necessary to provide a uniform treatment of resources. While the use of social capital in social network analysis tends to be restricted to a single resource, which is the focus of the study (e.g., CEO compensation (Belliveau et al., 1996), conference authorship (Licamele and Getoor, 2006), citations (Redner, 2005)), we continue with our goal of designing a framework that is relevant to the social sciences and thus offer a general account of resources. This is consistent with the fact that many see social capital as an attempt at actually quantifying the value of social relationships in achieving some individual or group benefit based on the resources present in the underlying network (Borgatti et al., 1998; Portes, 1998; Adler and Kwon, 2002). The following are popular definitions of social capital, that make very clear the critical role played by resources (emphasis added).

- "The sum of the *resources*, actual or virtual, that accrue to an individual or a group by virtue of possessing a durable network of more or less institutionalized relationships of mutual acquaintance and recognition." (Bourdieu and Wacquant, 1992).
- "The process by which social actors create and mobilize their network connections within and between organizations to gain access to other social actors' *resources*." (Knoke, 1999)
- "The number of people who can be expected to provide support and the *resources* those people have at their disposal." (Boxman et al., 1991)
- "The sum of the actual and potential *resources* embedded within, available through, and derived from the network of relationships possessed by an individual or social unit. Social capital thus comprises both the network and the assets that may be mobilized through that network." (Nahapiet and Ghoshal, 1998)

Most social networks today focus exclusively on facilitating interaction among participants, and providing limited information about participants in the form of simple profiles, generally including hobbies, likes, dislikes, interests, and in the more professional ones (e.g., LinkedIn), a possible range of skills or expertise. Resources per se are seldom, if ever, accounted for in any exploitable way in these highly interactive social networks.[1] On the

[1]It is true that, for example, inherent to a friendship network (e.g., Facebook) is access

other hand, traditional ideas are taking on more social forms that seem to be gaining traction. For example, the concept of recycling is being re-spun into "freecycling," which uses technology to enable people to freely give items to others within a community, and the idea of "micro-lending," which leverages technology to drive entrepreneurship by extending tiny loans to individuals in poverty. Other examples include people extending their sharing of personal things such as cars (e.g., zipcar), clothes (e.g., ThredUP), and spare bedrooms (e.g., airbnb) beyond the traditional family and close friends circle to a broader social community of strangers. In such cases, resources flow rather freely through the network of participants, with little concern for the nature and strength of the underlying relationships among them.

From the perspective of social capital, its creation and use, there is therefore a disconnect between social networks that could leverage social capital but have no explicit resources to be mobilized, and resource-rich social networks that currently give no thought to leveraging the notion of social capital. We attempt to bridge this gap, by providing the mechanisms necessary to treat resources as first-class citizens in highly dynamic social networks. For this extension of our framework, we consider social capital as "assets in networks" or more specifically "access to and use of resources embedded in social networks" (Lin, 1999). Previously, the access to social resources had not been measured, but was hypothesized to occur whenever sufficiently strong relationships among individuals existed. This paper more fully incorporates social resources into the social capital framework and shows how resources can be effectively mobilized through purposeful social interactions.

The remainder of the paper is organized as follows. Section 2 briefly reviews other work related to ours, as well as popular social sites and applications. Section 3 describes the nature, role, evolution and dependencies among relationships, resources and interactions within our proposed framework. Section 4 shows how social capital must be contextualized, and provides a computational definition of social capital. Section 5 reports on experiments that exercise the proposed framework and test two hypotheses about the effects of social interaction, generosity and reciprocity on social capital. Finally, section 6 concludes the paper.

---

to intangibles such as a sense of belonging, which indeed may be viewed as resources. Here, we take a more general view of resources, as discussed above.

2 Related Work

Our work belongs to the general area of computational social science, an emerging field that "leverages the capacity to collect and analyze...vast, emerging data sets on how people interact [thus offering] qualitatively new perspectives on collective human behavior." (Lazer et al., 2009). Specifically, we are interested in constructing a framework wherein social capital can be measured and reasoned about to test and/or discover hypotheses, as well as understand and influence behavior, in rich online social contexts.

Table 1: Relevant Features of Popular Applications. This table shows the features currently provided by popular social networking and resource exchange services (as of March 2011), and our framework. The features have been grouped under relationships, affinities, and resources. The following services are represented: Facebook (FB), Twitter (TW), LinkedIn (LI), FreeCycle (FC), Swap.com (SW), and our Social Capital Framework (SCF).

Feature	**FB**	**TW**	**LI**	**FC**	**SW**	**SCF**
relationships						
simple (friend or not)	✓	✓	✓			✓
complex (strength)						✓
asymmetric		✓				✓
dynamic						✓
affinities						
simple (like or not)	✓		✓		✓	✓
complex (how much)						✓
dynamic						✓
resources						
desired by ego				✓	✓	✓
desired by others				✓		✓
available by ego				✓	✓	✓
...willing to give				✓	✓	✓
...to specific others						✓
available by others				✓	✓	✓
...willing to give				✓	✓	✓
...to ego						✓
dynamic rel. (on exchange)						✓
...update on giving						✓
...update on receiving						✓

There are two major conceptualization of social capital, one in which

social capital is viewed as a property of individuals, and the other in which it is viewed as a property of the group. The first has been advocated by Bourdieu (1986), who considers social capital as the "aggregate of the actual or potential resources" of members of a group, while the second has been pursued mainly by political scientists, such as Putnam (2000); Putnam and Feldstein (2003). Our framework follows in Bourdieu's tradition.

Stone and Hughes (2002) argued that "creating a single index of social capital made no statistical (or substantive) sense." Yet they conceded that composite measures involving norms of trust and reciprocity, as well as network characteristics could form "useful summary measures of social capital and can be used in ways similar to that which a single, overall index of social capital might be used." Our single measure of social capital is in fact a composite measure in that sense, since its computation depends on norms, such as reciprocity, as well as network characteristics, such as density and size.[2]

While our attempt at designing a general framework that harmonizes several operationalizations of social capital may be somewhat unique, much research has been done to improve understanding of the source, value and use of social capital. We review some of the most relevant work here.

Coleman (1988), in his work on the relationship between social and human capital, discusses the important ideas of obligations, expectations and trust in social networks, where what someone may expect of others depends both on what one has done for them and whether one can safely count on their reciprocating. We capture these ideas through directed, weighted connections.

As theorized by Lin (1999), 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, and 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. Lin (2008) further characterizes access and mobilization as theoretical approaches that describe how social capital is expected to produce returns. 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

[2]While these are not used explicitly, they are implicit in the number of resources available and the number of individuals from whom resources may be obtained.

capital largely 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." Our framework accounts naturally for both access and mobilization perspectives.

Most computational social science studies so far have been done in the context of static networks. Recently, however, some researchers have begun to study the actual 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 (Katz, 2005; Kumar et al., 2006; Leskovec et al., 2005; Redner, 2005; Tantipathananandh et al., 2007)). Other studies have focused on analyzing explicit group formation and evolution (Backstrom et al., 2006; Leskovec et al., 2008; Zheleva et al., 2009). Similarly, our formalism takes into account the inherently dynamic nature of social networks, which, according to Coleman (1988) is essential to the formation of social capital. In particular, the notion of implicit affinities is used in our framework to further allow the nature of underlying relationships and groupings to vary over time.

Whereas social network analysis has mostly focused on the value of relationships, possibly in the context of one resource, social exchange theory introduces a general definition of resources used to explain social behavior. Although views vary, the theory posits that people value social relationships as the difference between the rewards and costs associated with them. In this theory, resources are defined as "an ability, possession, or other attribute of an actor giving him the capacity to reward (or punish) another specified actor" (Emerson, 1976). For example, whenever the relationship costs rise above the rewards, one may consider ending the relationship. Yet, before ending any unrewarding relationship, the alternatives and costs associated with switching to another relationship are considered. Blau described social exchange as "actions that are contingent on rewarding reactions from others" (Blau, 1964). As people connected by a social relationship benefit by it and remain satisfied, they will more likely remain in it. Recently, Schaefer (2011) extended the traditional and limited standard exchange resource (i.e., non-duplicable and non-transferable) to the much more general and realistic information-type resource exchange (i.e., duplicable and transferable). He was then able to show that the location of advantageous positions within a network differs by resource characteristics. Like social exchange theory, our framework treats resources generically. As Schaefer, it also emphasizes the

importance of resource characteristics and how they may impact the value of social capital.

From a pragmatic standpoint, there are certainly some popular social sites and applications that hit upon specific aspects of our framework. Table 1 compares a number of these services against our framework, by the features they currently offer. Since most of these sites are updated regularly, the features offered are likely to increase as the unchecked concepts become more mainstream and technology facilitates their inception.

3 Relationship Modeling

In our framework, the underlying social network is a hybrid network, or multigraph, consisting of explicit connections and implicit affinities among actors. An implicit affinity connects individuals together based on loosely defined affinities, or inherent similarities, such as shared interests, hobbies, political views, preferences, etc. We call these implicit because individuals may not be aware of the similarities in attitudes and behaviors that exist among them. On the other hand, individuals are aware of all explicit connections among them. These explicit ties are generally modeled using various types of information (Wasserman and Faust, 1994). Some of the more prevalent types that may be considered in a resource-sharing network include:

- **Behavioral interaction**: A connection is present between individuals when they engage in some form of communication, interaction, or information exchange (e.g., sending emails, visiting).
- **Evaluation of one person by another**: A connection is present between individuals when at least one of them places a value judgement on the other (e.g., considering as a friend, liking, showing respect).
- **Formal relationships**: A connection is present between individuals when there exists a well-defined, formal relationship between them (e.g., manager-employee, co-worker, student-teacher, colleague).
- **Biological relationships**: A connection is present between individuals when they are related by kinship or descent (e.g., parent-child, sibling, cousin).

These relationships may vary over time. The rate of change often depends on the type of relationship. The order in which the relationship types are listed

above is from most dynamic to most static. Indeed, biological relationships are clearly static by nature, while formal relationships may change in the natural process of time (e.g., retirement, graduation, job change). Value judgments have a tendency to evolve even more rapidly as they are affected by external events (e.g., gift giving, offense taking), and relationships based on behavioral interaction are probably the most dynamic, as they can vary almost continuously.

In practice, relationships may be elicited in a number of ways. In the social sciences, researchers typically design and administer carefully crafted questionnaires to a sample of their population of interest. Questions may include personal information, that can be used to find affinities among respondents, as well as lists of relationships (e.g., friends, colleagues), that can be used to build a network of explicit connections among respondents. With the proliferation of online data and the emergence of social media applications, much of this type of data can now be harvested automatically and on a much larger scale. Indeed, most social media applications, including blogs and social networks, allow people to create individual profiles, exchange messages, post comments, and create and maintain links among themselves (e.g., becoming friends on Facebook, following on Twitter, hyperlinking in the blogosphere).

Hence, it is possible to extract rich networks of relationships from the web. For example, Matsuo et al. (2007) use Google searches to set both explicit connections and implicit links (or affiliations). A set of names is provided to the system, and queries are issued for each pair (x_i, x_j) of names independently, and then together. Let n_i (resp., n_j) be the number of documents returned by a query with x_i (resp., x_j), and n_{ij} be the number of documents returned by a query with x_i and x_j. An explicit link is set between x_i and x_j if $n_{ij}/min(n_i, n_j)$ exceeds a threshold. Similarly, an implicit link is set between x_i and x_j if the similarity between the keywords found in documents associated with x_i and x_j exceeds a threshold. For their part, Adamic and Adar (2003) use links among home pages of students at MIT and Stanford to set explicit connections among the students. They assume that if a user's home page links to another's or is linked to by another's then the two users must know each other and can thus be labeled as friends.

In general, the context of the study, the availability of historical data, and the objectives of the analysis dictate how and what type of information can be used, or should be collected, to initialize relationships among network participants. Our framework is independent of how the network is

constructed. However, unlike most other approaches, it does assume that the network is dynamic, in that new participants may join, existing participants may become disaffected, and all relationships, both explicit and implicit, may change over time.

3.1 Relationship Strength

More so even than the type of relationship itself, what matters most in the context of resource sharing is the value associated to the relationship, or what we might call its strength, which adds one more dimension of change. Indeed, even static relationships may have variable strength, which can in some cases change the very nature of these relationships. It is possible, for example, for a grievously offended parent to disown a child, thus essentially nullifying the otherwise immutable biological relationship between them. While things need not be this dramatic, it is clear that all relationships, whatever their type, have varying degrees of strength. In general, such strength, or value placed on the relationship, is directly correlated to the amount of interaction between the related individuals. Hence, as alluded to above, any relationship, including a biological one, if it is not nurtured, risks losing (some of) its value in granting access to resources. Conversely, relationships that are cultivated tend to foster resource sharing. We formalize this idea shortly.

We call explicit social network (ESN) that part of the hybrid social network involving only explicit connections, and implicit affinity network (IAN) that part of the hybrid social network involving only implicit affinities. The IAN is an undirected graph, reflecting the fact that the notion of similarity among individuals is clearly symmetric (i.e., if i has some level of similarity with j then j has the same level of similarity with i, and vice-versa). The ESN, on the other hand, is a directed graph, since i and j may value their relationship to each other differently. For example, i may think of j as a great friend, while j considers i only as an acquaintance.[3] That distinction is important because what i may obtain from j does not depend so much on how i views j as it does on how j views i. For example, a startup company courting an angel investor for funding will not receive funding based on the value it places on its relationship to the investor, but rather based on the

[3]Of course, some explicit connections, such as the biological ones, are clearly symmetric. Since this is not true of all explicit connections, we choose the more general setting of a directed graph for the ESN. Symmetry, when applicable, is easily handled by having two edges of identical strength between individuals.

value that the investor places on its relationship to the company. This view is consistent with the prevailing idea that "to possess social capital, a person must be related to others, and it is those others, not himself, who are the actual source of his or her advantage." (Portes, 1998).

Both types of relationships are essential to the definition of social capital, especially the distinction between potential and actual social capital in a network, as demonstrated in (Smith et al., 2008, 2009; Smith and Giraud-Carrier, 2010). The strength of the explicit connection between two nodes i and j is denoted by s_{ij}^{ESN}, and the strength of the implicit affinity between i and j is denoted by s_{ij}^{IAN}. As per the above, $\forall(i,j)\ s_{ij}^{IAN} = s_{ji}^{IAN}$, but in general, given any pair (i,j) of nodes, $s_{ij}^{ESN} \neq s_{ji}^{ESN}$.

The value of s_{ij}^{IAN} depends only on the similarity between i and j. In our framework, we measure similarity based on the value of attributes or characteristics that participants expose to the network, either deliberately in the form of user profiles or more tacitly in the form of comments, blogs and other digital footprints from which sentiment, attitude, and behavior may be inferred. Given these attributes, we use the Jaccard Index to define the amount of affinity between i and j. Let A_i (resp., A_j) be the set of attributes exposed by i (resp., j). Then,

$$s_{ij}^{IAN} = \frac{A_i \cap A_j}{A_i \cup A_j}$$

While s_{ij}^{IAN} is likely to be relatively more static than s_{ij}^{ESN}, it may still change over time through such things as life events (e.g., from working to retired, from single to married), education (e.g., acquiring a new skill), and interactions with others (e.g., picking up a new hobby, changing one's mind). The initialization and evolution of s_{ij}^{ESN} are discussed below since they are dependent upon resources and interactions between i and j.

3.2 Resources

We generally define a *social resource* as a specific asset, material or symbolic, available through social connections within a network. Social resources are introduced to a network whenever individuals who possess them decide to expose them so they may become available to others. Resources come in a variety of forms, which have an impact on how they are handled within a social network. The following is a (non-exhaustive) list of typical resource characteristics. Note that these characteristics are not necessarily mutually exclusive. For each, we provide simple examples.

- **Duplicability**: A resource is duplicable if it can be reproduced through a ubiquitous process. For example, digital media is duplicable, while analog media is not. Note that information and knowledge may also be viewed as duplicable in the sense that a piece of information/knowledge may be given to another without loss of the "original" by the owner. An item of clothing or protected data, on the other hand are not duplicaple.

- **Transferability**: A resource is transferrable if ownership can be moved from one individual to another without restriction. In particular, if one individual receives a resource from another then it can freely pass it on to other individuals.

- **Exhaustibility**: A resource is exhaustible if it exists in finite quantity and can thus be completely used up. Examples of exhaustible resources include food and fossil fuels; examples of non-exhaustible resources include knowledge and wind energy.

- **Returnability**: A resource is returnable if it must be given back to the lender sometime later. A car is generally considered a returnable resource as it is expected to be brought back to its owner after use. On the other hand, a sandwich is clearly not a returnable resource as it is expected to be eaten by the receiver and thus could no longer be returned.

- **Quantifiability**: A resource is quantifiable if it can be enumerated, i.e., one can decide how many instances of the resource are available. For example, donuts are quantifiable, while support is not.

- **Durability**: The durability of a resource is a measure of how long it is capable of withstanding wear and tear and decay. A leather jacket might last many years and live through many exchanges; a jug of milk, on the other hand, may last just a few days and is likely to average only a single exchange.

The characteristics of resources affect both their access and mobilization. For example, exhaustible resources can only be mobilized until they are used up, but access to duplicable or shareable resources is almost unrestricted. There are also some dependencies among resource characteristics. For example, non-returnable resources may be viewed as exhaustible from the perspective of the giver, most non-duplicable resources are likely to be

returnable, and the exhaustibility of some resources may be defined in terms of their durability (i.e., used up over time) or quantifiability (i.e., used up over available amount). External factors may also result in specific characteristics for certain resources. For example, classified or secret information, whether so designed by a government agency or simply agreed upon among friends, tends to have strict rules of shareability.

Resource characteristics determine, in part, the value that one places on them. For example, scarce resources (i.e., small quantity) tend to be more highly valued, while abundant resources, especially non-exhaustible ones, may carry less value. Note that, in addition, such value may not be the same for all individuals, and it may vary over time, or based on external circumstances. An individual may attach much value to their material possessions while another might not. An individual may have a naturally altruistic attitude while another may find it more difficult to part with or share resources with others. An individual may value a resource highly today and much less tomorrow, because that resource ages (e.g., it may be easier to let someone borrow your old beat up truck than your brand new sports car), or more of the resource has become available (e.g., it is generally easier to give away money when there is a surplus of it than when there is only enough to meet one's own needs), or attitude towards the resource changes (e.g., as children get older they find it easier to pass their once most valued toys on to their younger siblings), or the resource becomes obsolete (e.g., upgrading a home appliance, moving from a detached house with a garden in the country to an apartment in the city). We denote by v_i^r the value given by individual i to resource r.

In addition to the somewhat absolute value individuals may assign to the resources they own, there is an interesting link between resources and relationships, which also affects whether, or how, resources flow from possessors to requestors. Indeed, it is generally the case that the flow of a resource r from an individual i who possesses r to an individual j who requests r depends not only on how i values r but also on its relationship to j. For example, if r is i's personal vehicle, i will likely allow access to a family member or a close friend, but not to a stranger. On the other hand, if r is one of i's screwdrivers, i will likely let almost anyone borrow it. In other words, j's access to a resource r owned by i may be captured by a function, $access_j(i, r)$, of v_i^r and s_{ij}^{ESN}, where $access_j(i, r)$ is true if j can access r from i, and false otherwise.

In our model, r is obtainable from i by j (provided j is requesting it) if j has a sufficient social relationship with i, as gauged by the resource holder i. In other words,

$$access_j(i, r) = \begin{cases} True & \text{if } s_{ij}^{ESN} \geq v_i^r \\ False & \text{otherwise} \end{cases}$$

This is illustrated in Figure 1, for a small network of three individuals and three resources. Individual j has access to all of the resources of individual i whose value to i is less than or equal to the strength of the explicit link from i to j (i.e., how i values its relationship to j).

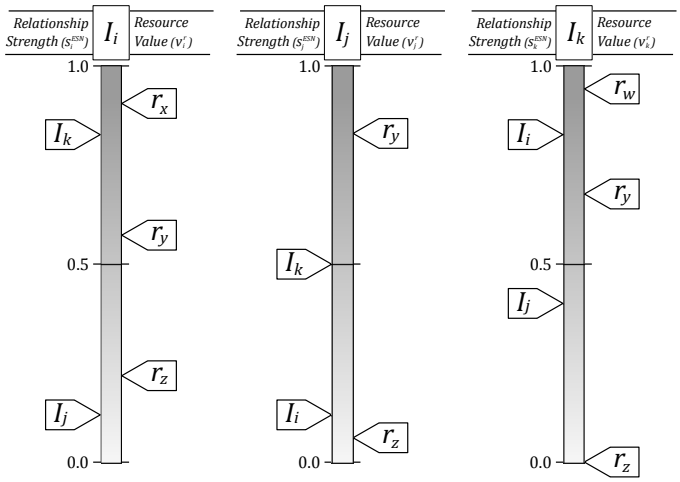


Figure 1: Resource Availability Scale for a Small Network. The network has three individuals, labelled I_i, I_j, and I_k. I_i has three resources, r_x, r_y, and r_z, two of which are available to I_k. I_j has two resources, r_y and r_z, one of which is available to both I_k and I_i. Finally, I_k has three resources, two of which are available to I_i and one (i.e., r_z) to I_j.

From a resource owner's perspective, this mechanism means that individuals introducing resources to the network can decide, in a general way, what other individuals will have access to it. For example, if v_i^r is set to a very small value (0 in the limit), i essentially intends to share r with everyone, while if v_i^r is set to a value close to 1 (1 in the limit), i is prepared to share r only with individuals it values most highly. Hence, v_i^r provides a spectrum of values for i to assign to r, to capture i's willingness to make r available to the network. In other words, resource owners can indirectly determine which of their resources are available to whom.

Note that the model assumes a fixed value for each resource by resource holders (i.e., v_i^r is only a function of r and i, not j). The only variable in $access_j(i, r)$ is the relationship strength. This, of course, makes certain patterns of access not feasible. For example, in Figure 1, it is not possible for I_i to let I_k have access to r_y and r_z, but not r_x, while at the same time allowing I_j access to r_x and r_y, but not r_z. To handle such scenarios, we would need to define resource value based on both the resources and the individuals. We contend that not only would such flexibility add to the complexity of the model, it is not strictly necessary. Since the strength of the relationship from the owner to the requestor is used to determine access, it is reasonable to assume that if I_i is willing to give I_k access to r_z because it values its relationship to I_k sufficiently highly, then I_i would also make r_z available to I_j since it values its relationship to I_j even higher than that to I_k. In general, it seems natural to assume that all individuals whose relationships are valued at the same or higher level should have access to the same resources from an individual, which is precisely what our model captures.

The foregoing discussion applies to all resources that members of a social network are willing to share. It is clear that just because someone owns a resource, they need not be willing to share it. Only so called *social resources* are available for sharing. A resource possessed by an individual becomes social once that individual declares it as such. All other resources remain essentially invisible to the social network.[4] For simplicity, we will ignore this distinction, and assume that all uses of the term resources is a reference to social resources.

Recall that social capital is a measure that describes an individual's ability to access and mobilize resources (Lin, 2008). In our framework, an individual j's social capital is created and increased by fostering relationships with individuals who possess resources j needs, and who are inclined to strengthen their own relationship to j. Once the value of any of these relationships with an individual i (i.e., s_{ij}^{ESN}) exceeds the value placed by i on the sought-after resource (i.e., v_i^r), $access_j(i, r)$ is True and j may access the resource from i. It is also possible for j's social capital to increase without any action on j's part. Indeed, if an individual i declares a new resource r as social and there already exists an explicit relationship between i and j such that the

[4] Alternatively, we could let all possessed resources be visible, but allow owners unwilling to share a resource to set its value to $+\infty$.

value of that relationship exceeds the value i places on r, then j may access r immediately. Hence, our framework naturally handles both aspects of social capital, namely access (a relationship must have enough strength to make access to resources possible) and mobilization (a relationship may have to be created and/or strengthened through purposive action to get to the needed resource).

We formalize these ideas in a computational definition of social capital shortly. First, however, we briefly discuss the notion of interaction, which makes resource mobilization possible.

3.3 Interactions

As pointed out above, interactions among social agents are an essential ingredient in the evolution of the strengths of relationships. These interactions may take on various forms, involving different levels of engagement (e.g., simple greeting when passing each other, emailing, attending a meeting, going out to dinner) and the possible exchange of resources (e.g., information, goods, services). For example, one person might ask a friend to loan her some money, or for help with a problem. If resources are understood in the broadest sense, all interactions involve at the very least the expenditure of some time. Yet, people often do not consider intangible resources such as time or friendly support as resources. Hence, we make a distinction in our framework between interactions in which resources are exchanged and interactions in which resources are not exchanged.

In terms of modeling relationships, the two main characteristics of interactions that matter are:

1. whether the interaction involves the exchange of resources, and
2. how the interaction is perceived (positive, negative, or neutral).

Both of these impact the strength of relationships. Whether it involves resources or not, the effect of an interaction on a relationship depends largely on how the interaction is perceived by the people participating in it. For those who view the interaction as a good thing, it is a positive interaction, which will likely strengthen their relationship with the others involved in the interaction; for those to whom the interaction is a bad thing, it is a negative interaction, which will likely weaken their relationship. Some interactions may even be viewed as neutral, having no measurable effect on the relationships of those involved. The possible exchange of resources in an interaction may further impact the perception of the participants, depending on the

value of the resource to the giver, the nature of the resource, the potential for reciprocity from the receiver, and the general disposition of the giver. For example, some individuals are naturally more giving than others, while some are rather attached to their resources; some individuals are naturally more philanthropic than others, while some always consider "what's in it for them." Similarly, some people are more naturally grateful than others (thus feeling a sense of reciprocity, or a desire to strengthen their relationship to the giver), while others may suffer from some sense of entitlement (thus feeling detached from the giver). These various attitudes in turn affect resource-sharing incentives either positively or negatively.[5]

In addition, it is likely that the frequency of interaction also affects the strength of relationships. In general, strength increases with interactions and decreases in the absence thereof. In our framework, every interaction may produce a change in the corresponding relationship's strength, and the prolonged lack of interaction is accounted for via a decaying mechanism, as discussed below. Note that one can also envisage situations where strength may decrease with an increasing number of interactions. For example, a benevolent individual may grow tired of an endless stream of interactions from a demanding friend, and thus wish to essentially reduce the strength of its association to said friend. While, we do not pursue this idea here, it can be accommodated in our framework simply by adapting the strength-updating function accordingly.

We shall assume that the participants in a social network purposely and consistently try to create and leverage their social capital, either to acquire the resources they need now or to enable their future access to new resources. We shall also assume that all individuals know what resources are possessed by what other individuals in the social network.[6] At any time, an individual can either do nothing or interact with another individual in the social network. The choice of the individual i with which j will interact is determined by a function of the form:

$$i = sel_j(R_i^p, R_j^s, s_{ij}^{ESN}, s_{ij}^{IAN})$$

[5]Note here again that all of this requires the ESN to be a directed graph.

[6]This is a reasonable assumption as it is always possible to find out what resources others are willing to share before asking them for these resources, and asking without first finding out is simply inefficient.

where

R_i^p : Set of resources currently possessed by i
R_j^s : Set of resources currently sought by j
s_{ij}^{ESN} : Strength of the explicit link from i to j
s_{ij}^{IAN} : Strength of the implicit affinity between i and j

As described above, some of these selection functions might require that j knows the strengths that other individuals assign to their explicit links (i.e., s_{ij}^{ESN}). While the exact values may not be known in reality, as j interacts with i and observes what i does (or does not do) in return, j gains some level of appreciation for the value that i may be placing on their relationship. So as to not unduly add complexity to our framework, we simply assume that j knows the value of s_{ij}^{ESN} for all i.

The exact definition of sel_j depends on j's goal. For example, if j is most concerned about getting all of the resources it needs as fast as it can, its selection function could return:

$$i = argmin_k(min_{r \in R_j^s \cap R_k^p} (v_k^r - s_{kj}^{ESN}))$$

In other words, j would always try to locate the individual i that possesses one of the resources it needs that is easiest to acquire, i.e., either $s_{ij}^{ESN} \geq v_i^r$[7] so that the resources can be obtained immediately, or the difference between v_i^r and s_{ij}^{ESN} is smallest, thus increasing its chances that an interaction with i might cause s_{ij}^{ESN} to exceed v_i^r, subsequently making r available to j.[8] Alternatively, j could be seeking to obtain its resources in some order of priority (imposed on R_j^s), so that its function would return:

$$i = argmin_k(v_k^{r_h} - s_{kj}^{ESN})$$

where r_h is the resource with highest priority in R_j^s. On the other hand, if j wishes to extend its reach and increase future access to resources (i.e., its social capital), its selection function could return:

$$i = argmax_k(\mid R_k^p \mid + s_{kj}^{IAN})$$

[7]Equivalently, $access_j(i, r) = True$.

[8]This, of course, requires the further assumption that j knows the value that i places on the resources it holds.

In other words, j would find the individual that has the most resources available and is most similar to itself, thus increasing the chances of i reciprocating. Indeed, those with whom affinities are shared are more likely to comply with requests (e.g., (Smith and Giraud-Carrier, 2010; Burger et al., 2001)). Any other selection function making use of the information available to j can thus be designed to match j's specific attitude and goals. In addition, as j's goal may change over time, so does sel_j.

As stated above, every interaction has an impact on the strengths of the explicit links of the individuals engaged in it. As before, we will denote by j the individual under consideration. For every individual in the social network, we define the following functions.

- $\Delta s_{ji}^{ESN}(event)$ determines how s_{ji}^{ESN} changes based on *event*, where *event* is one of the following:
  - j receives resource r from i: We assume that when j receives a resource from i, j is likely to feel some sense of reciprocity towards i, which we would capture by increasing the strength of the relationship from j to i, thus increasing i's chances of access to j's resources. The magnitude of the change in strength is likely to depend in part on the value that j assigns to r.
  - j gives resource r to i: We assume that when j gives a resource to i, j may wish to alter the strength of its relationship to i. In this case, the change may be a reduction as j may feel "used" by i. Again, the magnitude of the change in strength is likely to depend in part on the value that j assigns to r, as well as j's predisposition (e.g., altruistic vs. egotistic).
  - j engages in a (resourceless) interaction with i: We assume that every time j interacts with i, this may affect the strength of its relationship to i. This is particularly true of the first interaction j initiates with i. In that case, j can decide what value it places on its relationship to i, and set s_{ji}^{ESN} accordingly.
  - j receives a (resourceless) interaction from i: We assume that every time i interacts with j, this may cause j to change the strength of its relationship with i. For example, some level of reciprocity may be envisaged where j increases its strength to i as a response to i's reaching out to it. One special case of course is the setting of the original value of s_{ji}^{ESN}, following i's first interaction

with j (when j is not connected to i already). It would seem reasonable for the magnitude of the change here to depend on such things as s_{ij}^{IAN}, i.e., the amount of similarity between i and j (since bonding relationships are easier than bridging ones (Smith and Giraud-Carrier, 2010)), and the resources, R_i^p, that i possesses, i.e., what j can hope to gain from reciprocating.

- $decay_j(i)$ determines how s_{ji}^{ESN} decays over time in the absence of positive interactions from i. This is a function of the frequency of interaction from j to i, and from i to j. For example, if j reaches out to i but i does not react to j over some period of time, then j may wish to diminish the strength of its relationship to i. In general, it seems reasonable to assume that the strength of a relationship that is not maintained by some kind of positive interactions will likely decline over time. Figure 2 shows some examples of decay functions that may be used. Decay functions can be specified by the user or derived empirically (e.g., see (Burt, 2000)).

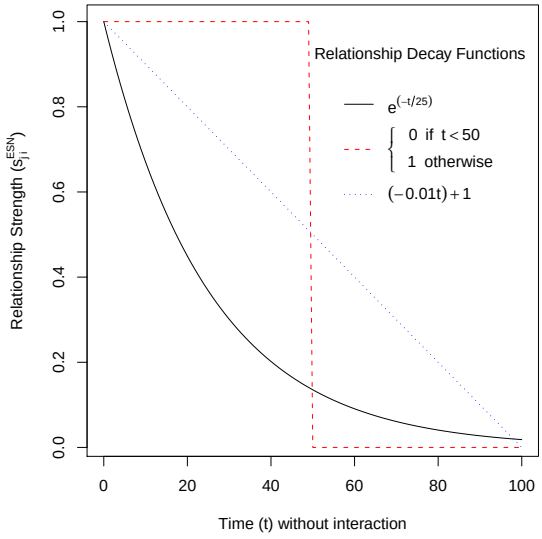


Figure 2: Decay Function Examples. Relationships tend to decay over time without significant interaction among individuals. In these example functions, the relationship strength nears zero in less than 100 time increments. In practice, decay functions can be specified by the user or derived empirically.

It should be clear that, except for $decay_j$, all functions work in pairs, in a kind of dual relationship. Indeed, if j receives resource r from i, then i gives resource r to j, and vice-versa; and if j engages in a (resourceless) interaction with i, then i receives a (resourceless) interaction from j, and vice-versa. Hence, for every $\Delta s_{ji}^{ESN}(event)$, there is a corresponding $\Delta s_{ij}^{ESN}(dual\,event)$.

The foregoing set of functions allows us to distinguish clearly among individuals making claims (i.e., using their social capital to access/mobilize resources), individuals agreeing to these demands (i.e., sharing their resources as sources of social capital) and the actual resources, a distinction that Portes (1998) claims is essential to a systematic treatment of social capital. Furthermore, it also accounts for both the structural aspect of social capital (i.e., the underlying social networks) and its cultural aspect (i.e., reciprocity, obligations, trust) (van Deth, 2008).

4 Social Capital Computation

While there may generally be many resources in a social network, it should be clear that not all of them are relevant in all situations. For instance, an individual k may only be interested in leveraging its social network to obtain a specific resource (e.g., a tool, a job, a piece of information). If no offering of the resource is currently available in k's network, then k can be viewed as having no social capital in that context. However, in a different context where k may be seeking to acquire other resources, it may actually have a significant amount of social capital if its network consists of individuals from whom k may access said resources. In other words, the value of social capital must be *contextualized* to the resources of interest.

Note that this contextualization in our framework addresses an important criticism of the view that social capital is a direct consequence of access and mobilization of resources. Indeed, Portes (1998) claims that "equating social capital with the resources acquired through it can easily lead to tautological statements." He gives as an example the case of an individual who gets a large tuition from his kin and another who does not. His question is then: should the first be said to have more social capital than the second, when in fact the second may not have gotten the tuition because it is not available in its network, even though those in the network may be highly committed to helping that individual? If so, Portes argues, this is like saying that the successful succeed (hence, the tautology). We argue that Portes' criticism fails to recognize the need for contextualization. Instead of a tautology,

the statement is conditional: *given a specific context*, the successful succeed. Hence, where the context is obtaining a tuition, the first individual does indeed have more social capital than the second since it can obtain the sought after resource while the other cannot. In another context, however, i.e., one more appropriate to the help the committed members of the second individual's network are able to offer, the second individual may have more social capital than the first.

To enable contextualization of social capital, we resort to a very simple mechanism of a *context vector*. The size of the context vector is the total number of (unique) resources possessed by members of the social network, and each entry in the vector is either No if the corresponding resource should be ignored, or Yes if the resource should be included as part of the context. In this way, a job seeker could focus only on resources related to a particular industry, a learner could focus only on resources corresponding to a specific area of expertise, and a community owner could focus only on resources known to have a positive impact on revenue. Of course, in a given analysis, several individuals, possibly all of them, will share the same context vector. In fact, as demonstrated in our discussion of Portes' criticism, this is critical to guarantee consistency and comparability in terms of social capital when studying a group of individuals.

We denote by C_j the context vector for individual j and by $C_j(r)$ the value of the entry for resource r in C_j.We can now turn to a computational definition of social capital in our framework. We first define the following function, which computes the number of resources that j may obtain from i within a given context C_j:

$$\#res_j(i) = \mid \{r \in R_i^p : access_j(i, r) = True \wedge C_j(r) = Yes\} \mid$$

Now, as social capital is a measure of an individual's access to social resources, the social capital for individual j can be defined as:

$$sc(j) = \sum_i \#res_j(i)$$

That is, $sc(j)$ is the total number of contextualized resources that are currently available to j. Note that resources may be counted more than once if the same resource is available to j from more than one individual in the network. We would certainly like to think that j's ability to obtain a resource in more than one place is indeed an indication of higher social capital.

In our earlier work, especially in several of our case studies (e.g., (Smith and Giraud-Carrier, 2010)), we sometimes found it useful to distinguish between bonding social capital and bridging social capital, to allow finer-grained analyses of social networks (Putnam, 2000; Putnam and 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. Bonding social capital increases through closure, as individuals strengthen existing links among themselves, while bridging social capital increases through brokerage, as individuals establish new links across structural holes (Burt, 2009). Network variety, or bridging social capital, has been argued to be valuable to both employers and employees in the hiring process (Erickson, 2004), while network similarity, or bonding social capital tends to be most beneficial for support and affirmation. The distinction between the two forms of social capital carries naturally in our extended framework:

$$b(j) = \sum_i \#res_j(i) s_{ij}^{IAN}$$

$$br(j) = \sum_i \#res_j(i)(1 - s_{ij}^{IAN})$$

so that $b(j)$ is the number of resources that j acquires through individuals who are similar to j (i.e., bonding) and $br(j)$ is the number of resources that j acquires through individuals who have little in common with j (i.e., bridging). As expected, we still have

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

Finally, when considering a community as a whole, we take the approach advocated by Bourdieu (1986), and define the social capital of the community as the aggregate of the social capital of its members. That is,

$$sc = \sum_j sc(j)$$

5 Experimental Results

General social resources are seldom available to social capital studies. Even at the time of this writing, machine-readable social resource data is

hard to come by.[9] Hence, our experiments rely on synthetic resource data. To conduct these experiments, we use NetLogo, a multi-agent programmable modeling environment, which allows us to simulate the necessary components of our social capital framework (Wilensky, 2011). The experimental testbed was designed to generate networks, distribute resources, simulate exchange, and calculate social capital over time. NetLogo conveniently provides a visual representation of the simulation and monitors on the variables of interest including social capital metrics.

The simulation allows us to model networks where the explicit relationships are determined using familiar techniques (specialized for directed graphs) including the scale-free model (Barabási and Albert, 1999), the small-world model (Watts and Strogatz, 1998), a random model (Erdos and Renyi, 1959), and even no pre-existing explicit network. Social resources are then initially distributed using a chosen mechanism, such as disparately, equally, randomly, or preferentially.

We run several experiments within this simulation environment to observe how the network behaves under different conditions. Three experiments, in particular, are presented. The first experiment has a small number of nodes and is included to provide intuition on the dynamics of a resource sharing simulation, the second compares networks of altruistic individuals versus networks of selfish individuals, and the third tests the effects of resourceless interactions on the network. These experiments are described in detail below.

5.1 Preliminaries

Before describing our experiments, some things must be understood about the experimental testbed. For practical reasons, we consider time as a sequence of discrete time steps, known as rounds. A round is a period of time during which all individuals take a turn at either doing nothing or interacting with another individual in the social network. The set of possible interactions is as described in Section 3.3, with the strength updating functions (here all constant for simplicity) given in Table 2.

We initialize the networks, such that 1) each node possesses a single resource, and 2) there are no explicit connections among individuals. We then begin the simulation and allow individuals to interact with one another and

[9]This is explained in part by the fact that social resources can be complex and challenging to describe unambiguously and uniformly, since a single resource can be given different descriptions by different individuals, thus requiring entity resolution.

Table 2: Strength Updating Functions. This table shows how the strength of explicit links are updated as individual j interacts with individual i.

Event	Δs_{ji}^{ESN}	Δs_{ij}^{ESN}
j gives resource r to i	-0.2	+0.4
j receives resource r from i	+0.4	-0.2
j engages in a (resourceless) interaction with i	-	+0.1
j receives a (resourceless) interaction from i	+0.1	-

exchange resources. For simplicity, we ignore implicit affinities and resource-driven reciprocity. Hence, all interactions, initial ones as well as subsequent ones, among pairs of individuals, update the relationship strengths (i.e., s_{ij}^{ESN} and s_{ji}^{ESN}) via only the constant strength updating functions specified in Table 2.

We make the following additional assumptions about resources:

- If j has access to r from i, the resource r must flow from i to j. In other words, i cannot refuse j a resource it holds for which j qualifies.
- Resources have intrinsic value and they are valued the same by each individual in the network (i.e., all individuals have the same resource availability scale). The idea is that people within our simulated social networks are not vastly different in the way they value resources (e.g., the richest and poorest individuals value resources the same).
- Resources are unique (an assumption also made by Schaefer (2011)), constant (i.e., neither created nor destroyed beyond initialization), transferrable, non-duplicable, and non-exhaustible.
- Individuals desire to acquire all resources not in their possession.

Also, our simulation has several options that can be set to affect the behavior of the network. For the purposes of these experiments, the following options are relevant:

1. **Altruism**: When this option is turned on, individuals in the network will sometimes simply give a resource to another individual without being asked, regardless of current relationship strength. When the option is off, individuals in the network will only give resources when someone with a sufficiently strong relationship asks them for one.

2. **Interactions**: When this option is on, individuals will randomly interact with other individuals in the network, strengthening relationships without exchanging resources. This models positive interactions without resource exchange, as discussed in Section 3.3. When the option is off, relationships can only be strengthened by resource exchange.
3. **Relationship decay rate**: This controls the relationship decay function $decay_j(i)$ by setting the number of rounds before the relationship from j to i begins to weaken. After this number of rounds, relationships decay linearly.

For the experiments described here, the relationship decay rate was set equal to the number of individuals within the network, thereby never decaying without at least offering an opportunity for individuals to interact. The values of the other options are specified in each experiment.

Finally, the selection function, sel_j, chooses an individual i to interact with by randomly considering one of the following interactions:

1. *Obtain a resource* from a neighbor willing to give it
2. *Give a resource* to a non-neighbor that is in need of it
3. Engage in a *resourceless interaction*

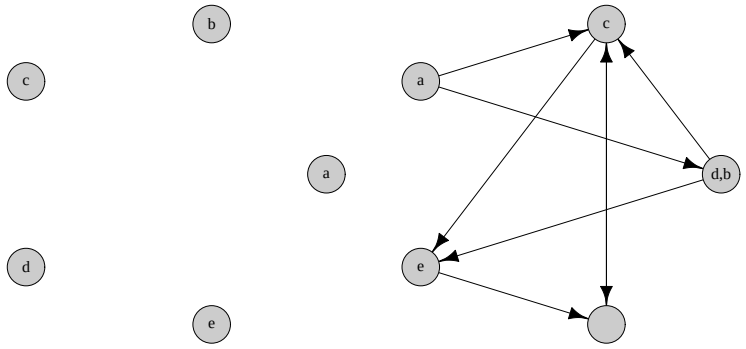
The random selection of interaction type is intended to allow individuals to participate in each of the interaction types at nearly the same frequency. This style of selection provides an initial baseline that future studies could be compared against. Whenever one of the above actions cannot be completed by the individual (e.g., "give a resource" was chosen, but the individual has none to give), then nothing is done and the simulation moves to the next individual.

5.2 Basic Experiment

First, we report on a basic experiment, with five individuals, that confirms the simulated dynamics of a resource-sharing network. This small experiment is included to show that our simulated network behaves according to the functions and framework discussed above. It should also provide the reader with some intuition that will help in understanding the two larger experiments that follow and how other context-sensitive experiments might be set up.

The initial network consists of five unconnected nodes, each possessing a resource as shown in Figure 3(a). For the duration of the experiment, both

the altruism and interactions settings are turned on. Using the strength updating functions described above, receiving a needed resource creates the most social capital towards the giver (+0.4), while an individual benefits only slightly after making a resourceless interaction (+0.1).



(a) Initial Network (b) Final Network

Figure 3: **Basic Experiment - Network**. The network consists of five individuals, starting with no relationships among them. Each individual begins with a single resource (i.e., a, b, c, d, e) as labeled on each node. After 500 rounds, the experiment concludes with a social capital value of 10. At that point, resources are no longer distributed uniformly (e.g., one node has no resources, and another has two (i.e., "d, b")). Over 1200 exchanges occurred during the experiment, averaging 2.4 exchanges per round.

For each round, every individual (in random order) interacts with one of the other individuals as determined by the selection function (i.e., sel_j). Following each round, the network social capital is computed. The evolution of social capital within the network during the simulation is shown in Figure 4.

The social capital increases as individuals have positive interactions, including receiving resources, one with another. The social capital decreases from givers towards receivers and over time as individuals neglect interacting with one another (i.e., *decay*). Social capital reaches the maximum of 20 (i.e., $\#resources * (\#individuals - 1)$) after 294 rounds and again at 384 rounds, while averaging 9.9 during the 500 rounds of the experiment. Figure 3(b) shows the resulting network.

This example shows how social capital increases and oscillates over time as individuals continue to interact and exchange resources with one another. The factors at play show continual changes even for a very small and relatively simple experiment. As additional real-world situations are added to

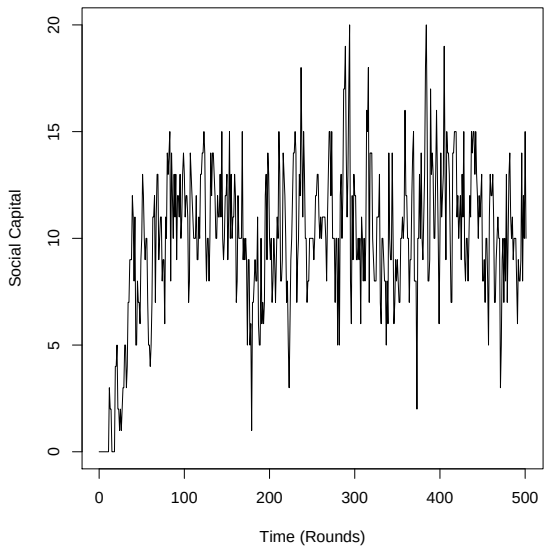


Figure 4: **Small Experiment - Social Capital**. The value of the network's social capital oscillates throughout the experiment averaging 9.9 with a standard deviation of 3.6.

the simulation, social capital can change even more dramatically and in accordance with the behavior of the individuals and community. For instance, if we simulated individuals introducing new or destroying old resources, then social capital would also increase or decrease accordingly. Furthermore, we could very easily scatter varied strength updating functions and selection functions, which could significantly change the behavior of individuals, and social capital would then reflect this. This experiment, used simple functions to offer some basic intuition into how a resource-sharing network can be simulated.

5.3 Altruism vs. Selfishness

Building upon the result of Theorem **??**, we wish to analyze more closely the effect that the presence or absence of altruism in a network has on the overall social capital of a network. We hypothesize that networks with altruistic individuals in them would have higher social capital. The following experiment was designed to test that hypothesis.

We start out with a network of size 25, and the simulation is run for 1000 rounds. Figure 5 depicts the evolution of social capital over time, with different settings of the Altruism option.

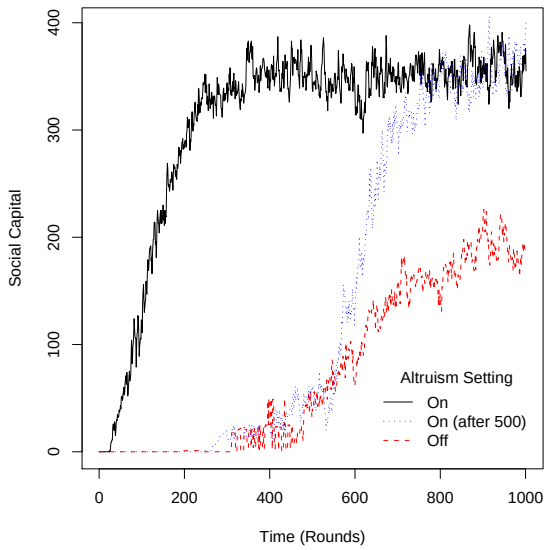


Figure 5: **Altruism vs. Selfishness Social Capital**: This graph plots the change in social capital over time of an altruistic network, a selfish network, and one that switches halfway through the simulation. Notice how after altruism is turned on, social capital surges upwards and rises to the level achieved by the network that had altruism on since the beginning of the simulation.

When the Altruism option is turned on, i.e., the network consists entirely of altruistic individuals, the value of social capital increases rapidly at first. After a while, this increase begins to level off, until it reaches a reasonably stable level (just above 300). The network's social capital stays around that level (with only relatively slight variations) until the simulation ends.

When the Altruism option is turned off, the network behaves noticeably differently. Social capital remains at 0 for nearly 300 rounds before it finally begins to rise. By contrast, it takes under 30 rounds for the altruistic network's social capital to start increasing. The social capital of our non-altruistic network then begins to increase, at a much lower rate than the altruistic network, until it reaches a maximum of 226 (averaging just about 76). The altruistic network, on average, had over four times the amount of social capital during the experiment.

Finally, we consider what happens when the Altruism option is first off, but then turned on about half way through the simulation. The network behaves nearly the same as the non-altruistic network for the first half of the

experiment. Then, as the altruism option is turned on (after 500 rounds), the network's social capital shoots up to quickly reach the same high level found in the altruistic network.

To check the robustness of these observations, we repeated the same experiment with networks of sizes from 15 and 50 individuals. In all cases, the difference in social capital between the altruistic behavior and the non-altruistic behavior was pronounced. The differences were largest for the larger networks. For networks of size 50, the altruistic network averaged 6 times as much social capital as the non-altruistic one. For networks of size 15, the difference was less, but still significant, with the altruistic network averaging 3 times as much social capital as the non-altruistic one.

The results of this experiment confirms our hypothesis that altruism does indeed lead to increased social capital.

5.4 Effect of Resourceless Interactions

Our operationalization of social capital is grounded in resources and their ability to be accessed and mobilized by members of a network. As such, one may wonder about the interplay between resourceless interactions and resource sharing, and how it affects social capital. In particular, we wish to isolate the effect of resourceless interactions. Our hypothesis is that social capital decreases without these interactions, as it becomes more difficult to maintain strong relationships. The following experiment was designed to test that hypothesis.

The network is initialized with 20 individuals, each having a single resource and no relationships. The simulation is run for 1000 rounds. Figure 6 depicts the evolution of social capital over time in networks of varying sizes, with the Interactions option turned off half way through the experiment.

The simulation begins with resource exchange (including altruism) and resourceless interactions on. As before, the network quickly increases in social capital, as relationships are formed among individuals in the network. After the network's social capital reaches a fairly steady state, we turn off the resourceless interactions. The network's social capital decreases sharply, maintaing itself at about half of the value it had previously reached.

When the experiment is repeated with networks of 10 and 30 individuals, similar results are obtained, except that the drop in social capital without resourceless interactions scales up or down as the size of the network increases or decreases. As might be expected, social capital is correlated with the size of the network. However, this is not always the case. Aside from simply

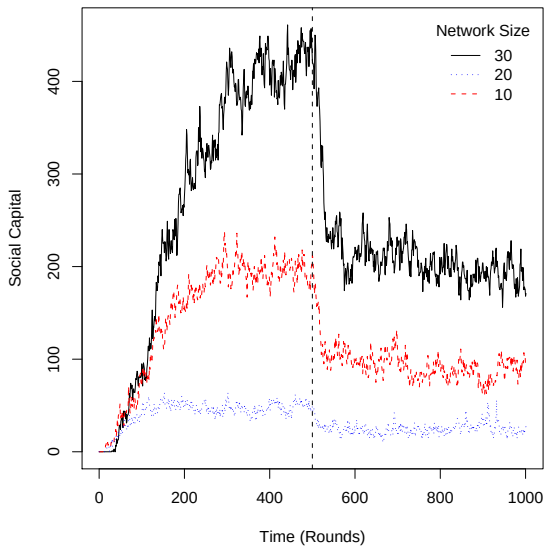


Figure 6: **Interactions Experiment, Social Capital**: This graph plots the change in social capital over time of networks of different sizes when resourceless interactions are turned off halfway through the experiment. The vertical line indicates when resourceless interactions are turned off.

having more individuals in a network, social capital increases as resources are produced and then shared with the network or simply as people become more willing to share what they already have. On the flip side, social capital decreases as individuals become less willing to share (see the previous experiment) or consume and destroy resources faster than they are created.

To further test the impact the interplay between the two kinds of interactions, we conduct the same experiment with resource exchange deactivated, i.e., the only possible interactions in the network are resourceless. In this case, the increase in social capital is very gradual and never reaches double-digit values, even when the simulation is allowed to continue running up to 5000 rounds. Of course, when resourceless interactions are deactivated, the network's social capital soon goes to zero.

The results of this experiment confirms our hypothesis that interactions without exchange (i.e., resourceless interactions) play an important role in strengthening relationships and therefore in increasing social capital in a network. They also suggest that these resourceless interactions are insufficient on their own to create a cohesive network.

6 Conclusion

Social resources are an important part of social networks and critical to any analysis of social capital. In this paper, we proposed an extension to our previous social capital framework that accounts for resource-aware social networks that create and maintain social capital through purposeful interactions. We showed how information about relationships and resources positively impacts social capital, and created a simulation in which interesting aspects of resource-sharing social networks were highlighted that can be of interest in real social networks.

The design of our computational framework for reasoning about social capital and its instantiation in a simulation environment of a social network with resource flow are indeed valuable contributions. For example, the findings on altruism would likely have been more difficult to discover in a real-world study than in the simulation. Clearly, real-world studies are still required to validate fully the results of our simulations. There are also a number of aspects of the framework that ought to be exercised and refined. In particular,

- We showed how a fully-connected group with maximum relationship strengths had maximum social capital. This is representative of the situation of several kinds of historical and current real-world communal experiments. It would be interesting to verify the impact of defectors and free-riders would be on such social networks. More generally, it would be interesting to compare communal social networks to more individualistic ones (e.g., where entrepreneurs emerge). Portes (1998) gathers some interesting thoughts about this in his work.

- We have ignored the situation where one individual does not (maybe cannot) seek a resource directly from an individual but rather asks one of his connections to help him or her get at the resource he or she needs. In other words, how would our framework need to be extended to model transitivity of relationships?

- We have assumed that contextualization was a binary function, where resources were either considered or excluded. These values could instead range over [0,1], indicating the relative value of resources in a particular context. As an example, weights could be chosen for an importance vector, such that r_x would be twice as valuable as r_y and four

times as valuable as r_z. In this case, an individual would have access to more social capital by having access to resource r_x than r_z or r_y.

- We have considered only reciprocal relationships, where directly interacting individuals update the strength of their relationships as a result of their interaction. It would be interesting to consider one-to-many update functions. For example, there may be a difference between my giving to one individual and my giving to many. Either may impact the strength to the one I give to, but if I give to many people, I may become less giving to every one over time, even those I never gave anything to before (feeling "abused" as many people try to get to me, e.g., philanthropists being hit up for money by everyone they meet); in this case, maybe giving to one causes a decrease of strength not just to the one but to all I have connections to.

References

Adamic, L., Adar, E., 2003. Friends and neighbors on the web. Social Networks 25 (3), 211–230.

Adler, P. S., Kwon, S.-W., January 2002. Social capital: Prospects for a new concept. The Academy of Management Review 27 (1), 17–40.

Backstrom, L., Huttenlocher, D., Kleinberg, J., Lan, X., 2006. Group formation in large social networks: membership, growth, and evolution. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 44–54.

Barabási, A.-L., Albert, R., October 15 1999. Emergence of scaling in random networks. Science Magazine 286 (5439), 509–512.

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

Blau, P., 1964. Exchange and power in social life. New York: Wiley.

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

Bourdieu, P., 1986. The forms of capital. In: Richardson, J. (Ed.), Handbook of Theory and Research for the Sociology of Education. NY: Greenwood Press, pp. 241–258.

Bourdieu, P., Wacquant, L., 1992. An invitation to reflexive sociology. University of Chicago Press.

Boxman, E., De Graaf, P., Flap, H., 1991. The impact of social and human capital on the income attainment of dutch managers. Social Networks 13 (1), 51–73.

Burger, J. M., Soroka, S., Gonzago, K., Murphy, E., Somervell, E., 2001. The effect of fleeting attraction on compliance to requests. Personality and Social Psychology Bulletin 27 (12), 1578–1586.

Burt, R., 2000. Decay functions. Social Networks 22 (1), 1–28.

Burt, R., 2009. Network duality of social capital. In: Bartkus, V., Davis, J. H. (Eds.), Social Capital: Reaching In, Reaching Out. Edward Elgar Publishing, Cheltenham, Ch. 2, pp. 39–65.

Coleman, J. S., 1988. Social capital in the creation of human capital. American Journal of Sociology 94, S95–S120.

Emerson, R. M., 1976. Social exchange theory. Annual Review of Sociology 2, 335–362.

Erdos, P., Renyi, A., 1959. On random graphs I. Publicationes Mathematicae 6, 290–297.

Erickson, B. H., 2004. Good networks and good jobs: The value of social capital to employers and employees. In: Lin, N., Cook, K. S., Burt, R. S. (Eds.), Social Capital: Theory and Research. Aldine Transaction, Ch. 6, pp. 127–158.

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

Knoke, D., 1999. Organizational networks and corporate social capital. In: Leenders, R., Gabbay, S. (Eds.), Corporate Social Capital and Liability. Kluwer Academic Publishers, pp. 17–42.

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

Lazer, D., Pentland, A., Adamic, A., Aral, S., Barabási, A.-L., Brewer, D., Chrsitakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., Van Alstyne, M., 2009. Computational social science. Science Magazine 323 (5915), 721–723.

Leskovec, J., Backstrom, L., Kumar, R., Tomkins, A., 2008. Microscopic evolution of social networks. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 462–470.

Leskovec, J., Kleinberg, J., Faloutsos, C., 2005. 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. pp. 177–187.

Licamele, L., Getoor, L., 2006. Social capital in friendship-event networks. In: Proceedings of the IEEE International Conference on Data Mining. pp. 959–964.

Lin, N., 1999. Building a network theory of social capital. Connections 22 (1), 28–51.

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

Lin, N., 2008. A network theory of social capital. In: Castiglione, D., van Deth, J. W., Wolleb, G. (Eds.), The Handbook of Social Capital. Oxford University Press, pp. 50–69.

Matsuo, Y., Mori, J., Hamasaki, M., Nishimura, T., Takeda, H., Hasida, K., Ishizuka, M., 2007. POLYPHONET: An advanced social network extraction system from the web. Journal of Web Semantics: Science, Services and Agents on the World Wide Web 5 (4), 223–278.

Nahapiet, J., Ghoshal, S., 1998. Social capital, intellectual capital, and the organizational advantage. Academy of management review 23 (2), 242–266.

Portes, A., 1998. Social capital: Its origins and applications in modern sociology. Annual Review of Sociology 24, 1–24.

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

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

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

Schaefer, D. R., 2011. Resource characteristics in social exchange networks: Implications for positional advantage. Social Networks In Press, Corrected Proof, doi: 10.1016/j.socnet.2010.12.002.

Smith, M., Giraud-Carrier, C., Judkins, B., 2007. Implicit affinity networks. In: Proceedings of the 17th Annual Workshop on Information Technologies and Systems. pp. 1–6.

Smith, M., Giraud-Carrier, C., Purser, N., 2009. Implicit affinity networks and social capital. Information Technology and Management 10 (2–3), 123–134.

Smith, M., Purser, N., Giraud-Carrier, C., 2008. Social capital in the blogosphere: A case study. In: Papers from the AAAI Spring Symposium on Social Information Processing. pp. 93–97.

Smith, M. S., Giraud-Carrier, C., 2010. Bonding vs. bridging social capital: A case study in twitter. In: Proceedings of the 2nd International Symposium on Social Intelligence and Networking. pp. 385–392.

Stone, W., Hughes, J., 2002. Social capital: Empirical meaning and measurement validity. Research Paper No. 27. Australian Institute of Family Studies.

Tantipathananandh, C., Berger-Wolf, T., Kempe, D., 2007. A framework for community identification in dynamic social networks. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 717–726.

van Deth, J., 2008. Measuring social capital. In: Castiglione, D., van Deth, J., Wolleb, G. (Eds.), The Handbook of Social Capital. Oxford University Press, pp. 150–176.

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

Watts, D. J., Strogatz, S., 1998. Collective dynamics of 'small-world' networks. Nature 393, 440–442.

Wilensky, U., 2011. NetLogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL. (online at: http://ccl.northwestern.edu/netlogo/).

Zheleva, E., Sharara, H., Getoor, L., 2009. Co-evolution of social and affiliation networks. In: Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 1007–1015.

Part III

Case Studies

M. Smith, N. Purser, and C. Giraud-Carrier. Social Capital in the Blogosphere: A Case Study. In *Papers from the AAAI Spring Symposium on Social Information Processing*, pages 93–97, 2008.

M. Smith and C. Giraud-Carrier. Bonding vs. Bridging Social Capital: A Case Study in Twitter. In *Proceedings of the International Symposium on Social Intelligence and Networking (SIN-10)*. The Second IEEE International Conference On Social Computing (SocialCom-10), Minneapolis MN (USA), Aug. 20–22, 2010.

K. Prier, M. Smith, C. Giraud-Carrier, C. Hanson. Identifying Health-Related Topics on Twitter: An Exploration of Tobacco-Related Tweets as a Test Topic. In *Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction (SBP 2011)*, College Park, MD (USA), Mar. 28–31, 2011. *(The original publication is available at www.springerlink.com)*

M. Smith, C. Giraud-Carrier, D. Dewey, S. Ring, D. Gore. Social Capital and Language Acquisition during Study Abroad. *33rd Annual Conference of the Cognitive Science Society*, Boston, MA: Cognitive Science Society, 2011.

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).

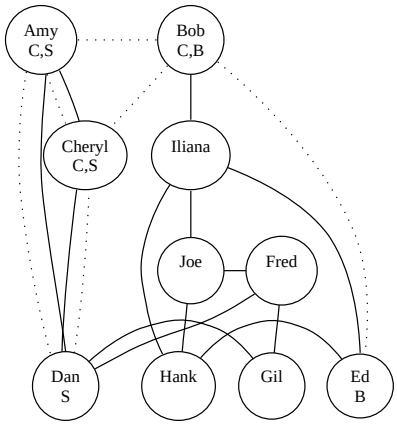


Figure 1: Sample Hybrid Network

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

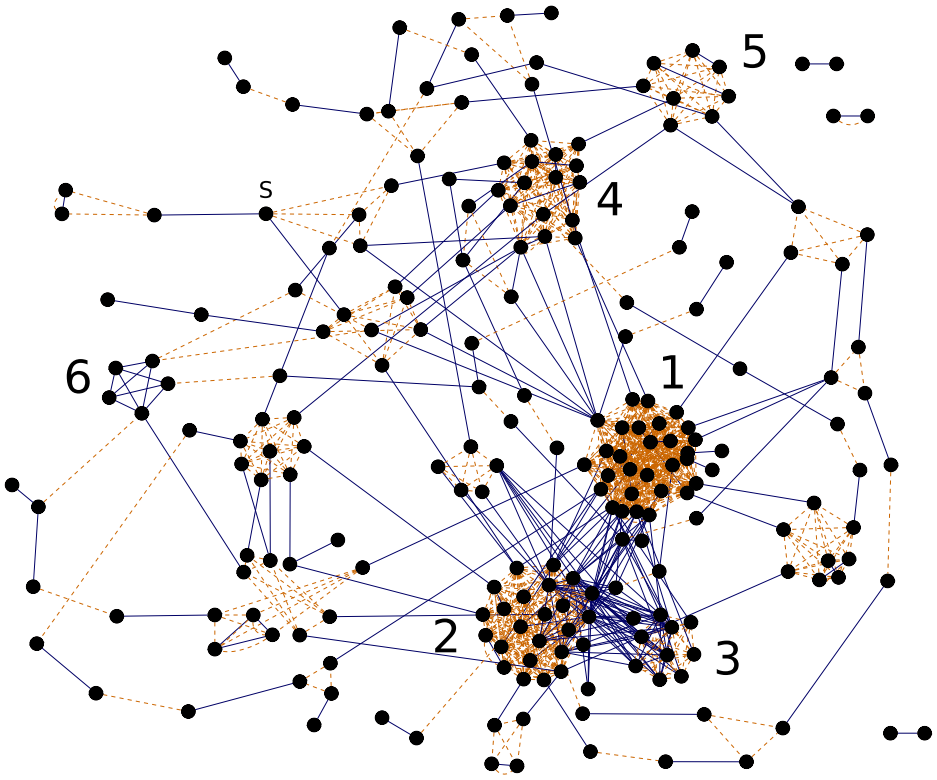


Figure 2: Hybrid Network for Blog Experiment (significant clusters are labeled 1-6, Scoble is labeled with an 'S')

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.

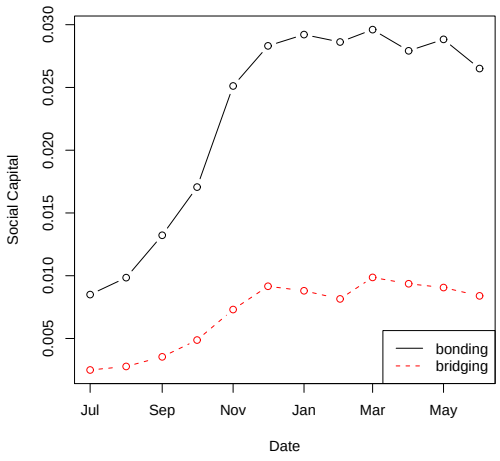


Figure 3: Social Capital by Month

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.

Bonding vs. Bridging Social Capital: A Case Study in Twitter

Matthew S. Smith
Department of Computer Science
Brigham Young University
Contact: http://m.smithworx.com

Christophe Giraud-Carrier
Department of Computer Science
Brigham Young University
Email: cgc@cs.byu.edu

***Abstract*—Online communities are connecting large numbers of individuals and generating rich social network data, opening the way for empirical studies of social behavior. In this paper, we consider the widely-held view of social scientists that bonding interactions are more likely than bridging interactions in social networks, and test it within the context of the large online Twitter community. We find that indeed users who request to follow others having similar profile descriptions (i.e., attempting to bond) increase the number of Twitter users who reciprocate their follow requests. From a practical standpoint, this result also informs how a new user might interact on Twitter to maintain a high follow-back ratio.**

I. Introduction

Online communities are groups of individuals connected by some generally 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. Technology, of course, has been the great enabler for the creation and evolution of such communities. Many of the most visited websites on the Internet, such as YouTube, Facebook, Wikipedia, and Blogger, allow users to connect and maintain ties via a social network. Furthermore, various organizations and initiatives are advocating the creation of standards that support this trend. For example, the Friend of a Friend project describes itself as "a simple technology that makes it easier to share and use information about people and their activities," while the OpenSocial initiative observes that "the web is more interesting when you can build apps that easily interact with your friends and colleagues."

The emergence of global and easily accessible online communities is revolutionizing the way in which individuals, and now even businesses, interact with each other. In turn, the science of building, discovering, understanding and leveraging such communities, or social networks, becomes increasingly important, as the Internet continues to grow into the largest collection of ideas, attitudes, personalities, and cultures in human history.

At the heart of social network analysis is the notion of social capital, aggressively pursued and popularized in the past couple of decades by sociologists and political scientists, such as Coleman [1], Lin [2], and Putnam [3]. Unlike other forms of capital that are centered around the individual, social capital is a property that emerges from the relationships that exist among individuals. 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 based on the resources present in the underlying network. 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 [4], [5]. An interesting study of the role of social capital in creating group-level benefits is Paxton's work on the mutually reinforcing effects of social capital and democracy [6]. In this paper, we restrict our attention to a consideration of the relationship of social capital to individual-level benefits or goods.

There is still an active discussion in the social sciences of exactly what social capital is, what forms it may take, or what it may entail. It is clear though that in order to create and leverage social capital, individuals must interact. For the purposes of our study, we consider one of the three forms of social capital identified by Coleman, namely information. Informational social capital arises from relations that provide information that, in turn, facilitates action [1]. Furthermore, we adopt Putnam's high-level dichotomy of social capital into bonding social capital and bridging social capital, where bonding social capital refers to the value assigned to social networks among homogeneous groups of people and bridging social capital refers to the value assigned to social networks among heterogeneous groups of people [3], [7].

The study of social capital requires the availability of sufficiently rich social network data. In the physical world, the acquisition of such data remains one of the biggest challenges. To cite just one example, Haynie's recent work on delinquency is based on the National Longitudinal Survey of Adolescent Health, which draws information from kids at 132 schools, yet the network sample includes kids from only 16 schools [8]. Studies have to be rather large to obtain even adequate network data. The cost of compiling such studies is significant as it involves the design and administration of expensive surveys. By contrast, cyberspace has no such limitations, either of size or cost. Indeed, the ease with which connections can be made online means that rich social network data is becoming available, opening the way for authentic, large-scale analyses of social behavior. We show one such analysis here, focused

on bonding and bridging social capital, in the context of the Twitter community.

Twitter is a fast-growing online social network, which went from 2-4 million users at the beginning of 2009 to about 40 million users by the end of that year. This relatively new community allows users to contribute short free-form status updates, called *tweets*, about themselves, and to follow the updates of others. Individuals are using this service to interact with friends, while businesses are beginning to use it to reach out and respond to customers. Twitter status updates can be a rich source of information about individuals, while the *following* and *follower* relationships provide the backbone of the underlying social network.

The principle of homophily, that contact between similar people occurs at a higher rate than dissimilar people, has been examined extensively [9]. Social capital researchers have also suggested that bonding interactions are more likely to occur than bridging interactions. Lin, for example, points out that interacting homogeneously (i.e., bonding) "should be the expected pervasive pattern of interactions observed," because it requires the least effort, while interacting heterogeneously (i.e., bridging) demands effort due to resource differentials and the lack of shared sentiments [2]. Or, as Burt puts it,"closure is the more obvious force. People advantaged by barriers between insiders and outsiders have no incentive to bring in outsiders. People too long in their closed network have difficulty coordinating with people different than themselves." [10]. In this paper, we use Twitter to verify that bonding interactions are indeed the pervasive pattern.

The paper is organized as follows. We first give a brief overview of some of the most relevant related work. We then proceed to describe our framework to quantify and measure both bonding and bridging social capital in online communities. Finally, we outline our experimental design and present the results.

II. Related Works

The interactions and structure of online social networks is dynamic and complex. Social network analysis assumes that the relationships among interacting units are a critical source of information [11], [12]. Within the social sciences, the study of these interactions has given rise to a number of interesting results. We mention only a few here, that are most relevant to our own analysis. Granovetter introduced the idea of the strength of weak ties, where otherwise dissimilar individuals engage in significant social interactions [13]. While it precedes such work, this idea is captured by the notion of bridging social capital as we use it. Haythornthwaite, in her work on the impact of communication media on social interactions, distinguishes among three types of ties, namely latent ties, weak ties and strong ties [14]. Latent ties correspond to technically possible, but not yet activated communication channels (e.g., belonging to the same email network); weak ties exist once individuals begin to use any medium of communication between them; and strong ties eventually arise as individuals expand their use of existing and create new media of communication to maintain their interactions. If sharing a communication medium is regarded as type of affinity among individuals, then Haythornthwaite's latent ties are the same as our implicit affinities, and the weights we assign to explicit connections capture the variable strength of ties among individuals. Coleman, in his work on the relationship between social and human capital, discusses the important ideas of obligations, expectations and trust in social networks, where what someone may expect of others depends both on what one has done for them and whether one can safely count on their reciprocating [1]. We capture these ideas through directed, weighted connections.

Most studies have been done in the context of static networks. Recently, however, some researchers have begun to study the actual 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 [15], [16], [17], [18], [19]). Other studies have focused on analyzing explicit group formation and evolution [20], [21], [22]. Similarly, our formalism takes into account the inherently dynamic nature of social networks, which, according to Coleman is essential to the formation of social capital [1]. In particular, the notion of implicit affinities is used to further allow the nature of underlying relationships and groupings to vary over time.

In practice, social network analysis has been used to understand an assortment of complex group phenomena, such as terrorist networks [23], [24], [25], [26], animal sociality [27], wasp colonies [28], and spread of diseases and behaviors [29]. Studies with an explicit focus on social capital have 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 [30]. In another study on social capital in the workplace, Erickson concludes that "good networks help people to get good jobs" [31]. Social capital has also been used in computer science to analyze the impact of the number of organizers with whom a potential author is friend on that authors publication records [32], and indirectly to distinguish between factual and relational content in social media communities [33]. Our work continues this tradition of using computational methods to explain social behavior.

III. Social Capital Framework

Social capital within a community is grounded in relationships, individuals' attributes, and available social resources. To exploit this information, we find it useful to distinguish between two types of relationships among individuals, as follows.

- An *explicit* connection links one individual to another based on some purposive action (e.g., sending an email, visiting) or a well-defined relationship (e.g., being a friend of, collaborating with). Individuals thus linked are aware of the explicit connections among them.

- An *implicit* affinity connects individuals together based on loosely defined affinities, or inherent similarities, such as similar hobbies or shared interests. Individuals 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 arising from implicit affinities [34]. A network with both implicit affinities and explicit connections is a *hybrid network*. In social network analysis terminology, a hybrid network is a multigraph having an explicit and an implicit relation among actors.

Implicit affinities are weighted by the amount of similarity estimated between individuals. The similarity metric chosen uses relevant attributes derived from the description and behavior of an individual. We make the important assumption that online personas are accurate. In other words, we assume that "you are what you say you are" online.

Social capital is naturally interested in implicit affinities, since it clearly has some relation to shared affiliations or activities among individuals [30]. 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. Hybrid networks thus play a key role in the definition of social capital, and the kinds of connections that exist among individuals determine whether that capital is realized or not. Note that in a strict sense, social capital is only realized, or accrued, once actions are taken and their result evidences the presence of said social capital. Hence, typical studies of social capital are retrospective. Here, however, we wish to use the notion of social capital to predict how one should leverage one's relations. For example, given that X and I are friends, that X is a head hunter and that I am looking for a job, I would want to ask X to help me find a job. Clearly, I may be misguided in the trust I place in X in this context (i.e., there may not be any social capital for me to leverage here), but it seems most reasonable to assume that I am not and to try to leverage what I perceive as social capital. For simplicity, we say that such capital is realized.

1) Implicit affinities only. In this case, the individuals have much in common (e.g., similar occupation or hobbies) but they are unaware of it. If they were to connect explicitly, they would be bonding, but since they have not yet, we say that there is only potential for bonding social capital here.
2) Implicit affinities and explicit connections. In this case, we say that the potential for social capital is now realized as similar individuals connect to one another explicitly.
3) No implicit affinities and no explicit connections. In this case, the individuals have little or nothing in common and they are unaware of each other. If they were to connect explicitly, they would be bridging, but since they have not yet, we say that there is only potential for bridging social capital here.
4) No implicit affinities but explicit connections. In this case, the mostly dissimilar individuals are now connected to one another (e.g., colleagues collaborating across disciplines or members of a church choir). Hence, we say that there is realized bridging social capital.

Both implicit affinities and explicit connections are therefore necessary to predict the network's social capital. Based on this framework, we have derived an effective mathematical formulation of social capital, as follows. An earlier version of this formulation is in [35].

Let s_{ij}^{IAN} be the strength of the implicit affinity between nodes i and j. s_{ij}^{IAN} ranges over [0,1] and is a measure of the similarity between nodes i and j. Similarly, let s_{ij}^{ESN} be the strength of the explicit connection between nodes i and j. s_{ij}^{ESN} may be as simple as 1 or 0, to reflect the presence or absence of a link, but may also range over [0,1] to capture degrees of connectivity (e.g., best friend vs. casual friend vs. acquaintance). Finally, let N be the set of nodes in the network.

We define the *potential* bonding social capital of an individual i as the sum of the individual's implicit affinity 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,

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

While it seems appropriate for implicit affinities to be "undirected," since two people either share or do not share a specific affinity, it is less so for explicit edges. It is clear that the value of some (explicit) relationships is not necessarily reciprocal and may vary among participants. For example, one person may consider another person as their best friend, while that other person may look at the first as only a good friend. Thus, our framework recognizes that the amount of social capital an individual i may realize from a relationship with another individual j is not predicated upon the value that i places in the relationship, but rather upon the value that j places in it. While i may think highly of that connection, for example in the context of obtaining a job reference from j, the reference will only be as strong as j thinks of i, and not the other way.

Accordingly, we define the bonding social capital *realized* by a node i, when (explicitly) connecting with node j, as the product of the strength of the implicit affinity between i and j by the strength of the explicit edge connecting j to i: $s_{ij}^{IAN} s_{ji}^{ESN}$. Now, as expected, if j is unaware of i, even when i may be aware of (and possibly even count on) j, there is no social capital available for i from that relationship. The (realized) bonding social capital of an individual i is the sum of its realized bonding social capital with all other individuals. That is,

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

Likewise, the (realized) bridging social capital of an individual i is the sum of realized bridging social capital with all other individuals. That is,

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

Finally, 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 its bonding capital and bridging capital. That is,

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

IV. Experimental Setup

The following experiment was designed to test the social scientists' hypothesis that bonding is more likely than bridging in social networks. Given our framework, this hypothesis may be recast into the following measurable Twitter hypothesis.

Hypothesis. *Following users with whom the most affinities are shared (i.e., attempting to bond) produces more follow-backs (i.e., bonding) than other following strategies.*

The idea is that individuals who follow-back others who deliberately adopt a following strategy motivated by a desire to bond create bonding social capital. If there is a stronger tendency for individuals to follow-back those of their followers who are like them rather than others, then this, in some sense, establishes that bonding is more pervasive than bridging. Hence, by comparing the relative number of follow-backs and followers (i.e., the bonding social capital accrued) by adequate strategies, we can verify our hypothesis.

For our experiment, we created a set $\mathcal{A}$ of Twitter accounts. Each account was setup to behave approximately the same as all others for everything, except for the individuals it chooses to follow. Specifically, each Twitter account in $\mathcal{A}$ was given a screen name that varied only by the random three-digit number appended to a pre-specifed name (e.g., $jon287$, $jon797$, $jon853$). Each account was also given the same profile information (see Figure 1 for an example), and all accounts were scheduled to tweet at approximately the same times. This rigorous setup allowed us to test the unique following strategies assigned to each account.

The implicit affinity network among Twitter users was derived from the profile description, labeled *Bio* in Figure 1. Alternatively, the implicit affinity network could naturally be derived from tweets made by users, thus creating a more dynamic but also more computationally intensive network. We chose to be conservative by utilizing just the profile description for this study. Collecting status updates (i.e., tweets) requires additional calls to the Twitter API and could require a significant amount of text mining, which could slow down the pace of the experiment and possibly limit the applicability of the results for regular Twitter users. Profile descriptions are relatively easy to obtain, are less dynamic, and can be sufficiently descriptive of a particular user. Implicit affinities were calculated by counting the number of matching unigrams and bigrams within profile descriptions after removing common stop words (e.g., a, by, on, the, with). Thus, users having more profile affinities (with the accounts in $\mathcal{A}$) offer opportunities for bonding. Those with few or no affinities offer opportunities for bridging.

Name Jon
Location USA
Web http://bit.ly/*unique-hash*
Bio I love data mining, social networks, machine learning, business intelligence, pattern recognition, and natural language processing.

Fig. 1. Twitter profile information for each account in $\mathcal{A}$.

A selection of Twitter users, $\mathcal{U}$, was sampled from the Twitter public timeline. Rather than including anyone that recently tweeted something, we decided to restrict our sampling to those that had recently tweeted either "data mining" or "social networks". Sampling from the Twitter public timeline without restriction opens the possibility that all of the accounts sampled in $\mathcal{U}$ could share very little or even nothing with the niche accounts in $\mathcal{A}$. We chose to avoid this possibility, and effectively narrowed the candidate pool to those that could possibly share an interest with the targeted focus of the accounts in $\mathcal{A}$.

Next, each Twitter account in $\mathcal{A}$ was assigned a following strategy as follows.

A *Bonding* - this strategy attempts to bond by following users, in $\mathcal{U}$, with whom the most affinities are shared. In other words, at the time of selection, the user with the largest s_{ij}^{IAN} is followed.

B *Bridging* - this strategy attempts to bridge by following users, in $\mathcal{U}$, with whom the least affinities are shared. In other words, at the time of selection, the user with the smallest s_{ij}^{IAN} is followed.

C *Median affinities* - this strategy follows the users in $\mathcal{U}$ having the median number of affinities shared at the time of selection.

D *Randomly* - this strategy randomly follows a user in $\mathcal{U}$ at the time of selection. Every remaining user in $\mathcal{U}$ has the same probability of being followed.

E *Minimum absolute following/ers difference* - This strategy follows the user in $\mathcal{U}$ having the smallest difference between following and followers (i.e., $|following_{count} - followers_{count}|$) at the time of selection. For example, the absolute following difference is 400 for a user following 100 and followed by 500.

F *Maximum absolute following/ers difference* - This strategy follows the users in $\mathcal{U}$ having the largest absolute difference between following and followers (i.e., $|following_{count} - followers_{count}|$). The absolute following difference is calculated as described in the previous strategy.

G *Median number of followers* - This strategy follows the user in $\mathcal{U}$ having the median number of followers at the time of selection.

rank_f	strategy	following	follow-backs	↓ followers	rejects	churn	follow_{total}	follower_{total}
1	*bonding* (A)	500	158 (32%)	202 (40%)	12	127	512	329
2	*max. following/ers diff.* (F)	500	84 (17%)	172 (34%)	12	324	512	496
3	*random* (D)	500	118 (24%)	154 (31%)	20	103	520	257
4	*median affinities* (C)	500	99 (20%)	123 (25%)	25	93	525	216
5	*bridging* (B)	500	99 (20%)	120 (24%)	25	91	525	211
6	*min. following/ers diff.* (E)	500	87 (17%)	99 (20%)	50	55	550	154
7	*median num. followers* (G)	500	63 (13%)	86 (17%)	31	51	531	137
8	*min. num. followers* (H)	500	33 (07%)	42 (08%)	79	29	579	71
9	*follow nobody* (I)	0	0 (—%)	3 (—%)	0	24	0	27

TABLE I

FOLLOWER STATISTICS: EACH OF THE NINE ACCOUNTS ARE LISTED BY *strategy* AND RANKED BY THE NUMBER OF *followers* OBTAINED DURING THE EXPERIMENT, DENOTED $\textit{rank}_f$. THE *following* COLUMN IS THE NUMBER OF USERS THAT THE ACCOUNT WAS FOLLOWING AT THE END OF THE EXPERIMENT. THE *follow-backs* COLUMN REPORTS THE NUMBER OF USERS FOLLOWED THAT WERE FOLLOWING THE ACCOUNT BACK AT THE END OF THE STUDY, THE PERCENT OF FOLLOWING IS SUPPLIED FOR REFERENCE (I.E., *follow-backs*/*following*). THE *followers* COLUMN IS THE NUMBER OF USERS FOLLOWING THE ACCOUNT AT THE END OF THE EXPERIMENT (INCLUDING THOSE THAT WERE NEVER FOLLOWED BY THE ACCOUNT). THE PERCENT OF FOLLOWING IS ALSO SUPPLIED FOR REFERENCE (I.E., *followers*/*following*). THE *rejects* COLUMN REPORTS THE NUMBER OF USERS THAT COULD NOT BE FOLLOWED ON TWITTER AT THE TIME (E.G., ACCOUNT WAS PROTECTED, USER ATTEMPTING TO FOLLOW WAS BLOCKED, OR USER WAS SUSPENDED). THE *churn* STATISTIC REPORTS THE NUMBER OF USERS THAT FOLLOWED THE ACCOUNT FOR A TIME, BUT WERE NO LONGER FOLLOWING THE ACCOUNT AT THE END OF THE EXPERIMENT. THE $\textit{follow}_{total}$ IS THE TOTAL NUMBER OF USERS THAT WERE FOLLOWED BY THE ACCOUNT, I.E., THE SUM OF *following* AND *rejects*. THE $\textit{follower}_{total}$ IS THE TOTAL NUMBER OF USERS THAT FOLLOWED THE ACCOUNT DURING THE EXPERIMENT, I.E., THE SUM OF *followers* AND *churn*.

H *Minimum number of followers* - This strategy follows the user in $\mathcal{U}$ having the fewest number of followers at the time of selection.

I *Follow nobody* - this strategy chooses not to follow any users. It may naturally be viewed as a control group.

A *following round* consisted of each Twitter account in $\mathcal{A}$ selecting users from $\mathcal{U}$ one at a time according to its assigned strategy. Users were removed from $\mathcal{U}$ as soon as they were selected. Thus, users from the pool could only be followed by a single account in $\mathcal{A}$. Each following round began by randomizing the order in which accounts selected users. On the days that following rounds occurred, accounts selected 50 or less users to follow. Following rounds were planned to occur sporadically until every account in $\mathcal{A}$ was following 500 users (an arbitrary, yet substantial number of individuals). Following rounds occurred on 22 of the 105 days in which the experiment was conducted.

For the duration of the study, each of the Twitter accounts in $\mathcal{A}$ published identical status updates to their respective Twitter stream at approximately the same time. There were 117 status updates made across 19 different days during the experiment. Over 90% (106) of the status updates published included a link that tracked the number of times it was clicked. Each link was shortened (using bit.ly) and associated to specific account in $\mathcal{A}$. After all of the users in $\mathcal{U}$ had been selected by the accounts in $\mathcal{A}$ and all status updates had been published the experiment concluded. The following statistics were analyzed for each account at the conclusion of the experiment:

- number of followers
- number of click-thrus (tweets and profile click-thrus)
- individual bonding capital
- individual bridging capital

V. RESULTS

The final follower statistics for each account after the experiment are shown in Table I. Each account in $\mathcal{A}$ is listed by the assigned strategy and ranked by the number of *followers*. The number of *follow-backs* is the subset of *following* users that reciprocated follow requests made by the account. The *followers* column reports the number of followers that the account had at the end of the experiment. Unlike the statistic reported in *follow-backs*, this statistic includes followers that discovered the account through alternative methods. Although we do not know all of the ways that accounts can get noticed through the numerous Twitter apps, a few alternative methods for being noticed on the Twitter website include being discovered through Twitter search or by traversing the explicit social network (e.g., being discovered through a "friend of a friend").

The *rejects* column reports the number of users that could not be followed on Twitter, at the time the request was made, due to some reason, such as the account was protected, the user attempting to follow was blocked, or the user was suspended. Twitter is a constantly evolving community where users can block other users on a whim and where users are regularly suspended for "strange activity." For instance, attempting to follow a user that has been suspended produces the following error message: "Could not follow user: This account is currently suspended and is being investigated due to strange activity." Rejection errors occurred most often for strategies H and E, perhaps suggesting something about users that fall into these groupings (i.e., users in in group H likely block the accounts in $\mathcal{A}$ and users in group E tend to get suspended more often.)

The *churn* statistic represents the number of users that followed the account for some time during the study, but no longer followed the account at the end of the study. The current guidelines on Twitter's website state "if you decide to

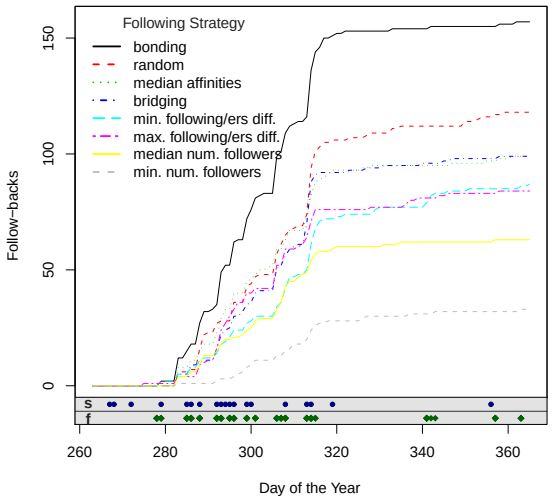


Fig. 2. **Follow-backs Over Time**: Follow-backs obtained by accounts in $\mathcal{A}$ throughout the duration of the study. Days in which following rounds occurred (i.e., accounts in $\mathcal{A}$ followed users in $\mathcal{U}$) are marked in the row labeled **f**. Days that new status updates were posted to the accounts in $\mathcal{A}$ are marked in the row labeled **s**.

strategy	significantly different
(A) *bonding*	B, C, E, G, H
(B) *bridging*	A, H
(C) *median affinities*	A, H
(D) *random*	E, G, H
(E) *min. following/ers diff.*	A, D, F, H
(F) *max. following/ers diff.*	E, G, H
(G) *median num. followers*	A, D, F, H
(H) *min. num. followers*	A, B, C, D, E, F, G

TABLE II
FOLLOWERS-TO-FOLLOWING: PAIRWISE PROPORTION TEST RESULTS. ($\alpha = 0.01$, BONFERRONI CORRECTED p-VALUES)

strategy	significantly different
(A) *bonding*	B, C, E, F, G, H
(B) *bridging*	A, H
(C) *median affinities*	A, G, H
(D) *random*	G, H
(E) *min. following/ers diff.*	A, H
(F) *max. following/ers diff.*	A, H
(G) *median num. followers*	A, D, C
(H) *min. num. followers*	A, B, C, D, E, F

TABLE III
FOLLOWBACKS-TO-FOLLOWING: PAIRWISE PROPORTION TEST RESULTS. ($\alpha = 0.01$, BONFERRONI CORRECTED p-VALUES)

follow someone and then change your mind later, thats fine!" However, they discourage *aggressive follow churn*, which they define as "when an account repeatedly follows and un-follows large numbers of users."[1] Churn was observed most often for strategy F (more than double strategy A, the next highest) — perhaps more users selected by this group are using automated tools to aggressively follow and un-follow.

The second to last column, $follow_{total}$, is the total number of users that were followed by the account, or the sum of *following* and *rejects*. The last column, $follower_{total}$, is the total number of users that followed the account during the experiment, or the sum of *followers* and *churn*.

Figure 2 shows a plot of the number of follow-backs that each account had during the experiment. Following rounds occurred on the days marked in the row labeled **f**. Status updates occurred on the days marked in the row labeled **s**. The follow-backs plotted is the cumulative sum of followers obtained on the day indicated and that remained at the conclusion of the experiment. Users that followed back but were no longer following at the end of the study are not included. Following rounds are accompanied by noticeable increases in follow-backs for all following strategies, with significantly larger such increases for strategy A, a first indication that bonding may indeed be easier.

We formally tested our hypothesis using proportion tests. Strategy I is left out as it does not follow anyone (i.e., *following*=0) and thus proportions would be undefined (division by 0). All other strategies are included in the results, but our focus is on strategies A and B. Users who follow A, especially when reciprocating (i.e., follow-backs), are clearly bonding since they were first picked by A because they were similar to A. Users who follow B, again especially when reciprocating, cannot be bonding, and must be bridging, since B explicitly chose them for their dissimilarity with itself.

A test comparing the *followers*-to-*following* proportions showed that strategies A and B were significantly different having a p-value < 0.001. Upon performing a pairwise proportion test across all of the strategies, we observe that many of the strategies were significantly different, as shown in Table II. Note that the p-values were Bonferroni adjusted and considered significant only if they were less than alpha ($\alpha = 0.01, p < \alpha$). While it appears that A's bonding strategy is not significantly different from D's random following strategy, this is probably due to the fact that we pre-selected the set $\mathcal{U}$ of users to follow based on their affinities with A. Hence, if our hypothesis holds, a random strategy would exhibit a fair amount of bonding.

As an additional check, a pairwise proportion test was performed on the *follow-backs*-to-*following* proportion, as shown in Table III. Again, this test shows that strategies A and B are significantly different. As above, the p-values were Bonferroni adjusted and considered significant only if they were less than alpha ($\alpha = 0.01, p < \alpha$). These results are similar to the above.

Next, we investigate the click statistics. Table IV shows the clicks obtained through each account (and the number of mentions). Each of the nine accounts are listed by *strategy* and ranked by the number of *total clicks* received, denoted $rank_c$. Each account made approximately 117 status updates

[1] *Following Limits and Best Practices* available at: http://help.twitter.com/forums/10711/entries/68916 *(Jan. 06, 2010).*

(i.e., tweets), of which 106 included a clickable tracking link. The $clicks_t$ column shows the number of clicks that came through links posted in status updates for each account. The next column, $clicks_p$, shows the number of times that the profile link (i.e., unique tracking link immediately after *Web* in Figure 1) was clicked for each account. The next column, *total clicks*, is the sum of the previous two columns and is the total number of clicks obtained for each account. Lastly, the number of *mentions*, or any Twitter update that contained *@username* in the body of the status update, is listed for each account. Note that due to a small configuration error, not all of the click data was recorded for strategy G. It is therefore not included in the ranking (only 78 of the 106 links posted were tracked).

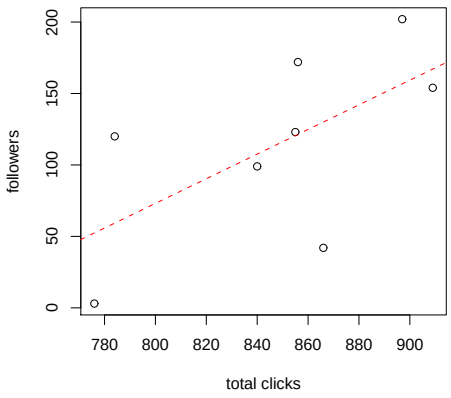


Fig. 3. **Clicks vs. Followers**. The linear model shown by the regression line (dashed) poorly fits the data having an R^2 value of 0.28. There is, however, a positive Pearson correlation of 0.62, yet it is not as high as might be expected.

The click results were somewhat unexpected. First, the number of clicks obtained for strategy I (i.e., followed nobody) is surprisingly high and similar to the number of clicks obtained for the other strategies. We think that this may be due to how rigorously Twitter data was being consumed by automated tools and web crawlers during the study. Secondly, we expected that the number of clicks for each account would be linearly proportional to the number of followers. This was not the case. Figure 3 plots the number of clicks versus the number of followers. The adjusted R^2 value of 0.28 for the linear model confirms that it poorly fits the data. The number of clicks did not appear to be proportional to the number of followers, nor did the number of clicks vary significantly among strategies with a standard deviation of 46. These results suggest that, in terms of obtaining clicks, tweeting is more important than obtaining more followers.

Finally, we consider the social capital accrued by each strategy. Using the formulas defined in our framework, we compute the social capital realized by each strategy at the end of the experiment. Table V shows the proportion of bonding social capital (to total social capital) for each strategy,

$rank_b$	strategy	↓ bonding
1	*bonding* (A)	10%
2	*max. following/ers diff.* (F)	3%
3	*follow nobody* (I)	3%
4	*random* (D)	2%
5	*bridging* (B)	2%
6	*min. following/ers diff.* (E)	2%
7	*median num. followers* (G)	2%
8	*median affinities* (C)	1%
9	*min. num. followers* (H)	1%

TABLE V
SOCIAL CAPITAL RESULTS: EACH OF THE NINE ACCOUNTS ARE LISTED BY *strategy* AND RANKED BY THE PROPORTION OF BONDING SOCIAL CAPITAL THEY ACCRUED (I.E., $b(i)/sc(i)$), DENOTED $rank_b$. STRATEGY A HAS SIGNIFICANTLY MORE BONDING SOCIAL CAPITAL THAN ANY OF THE OTHER STRATEGIES.

in descending order. Again, these results set A's bonding strategy as a clear winner over all strategies, and in particular significantly higher than B's bridging strategy. Note that the seeming rise of strategy I is due to the fact that social capital is accrued based on followers rather than following, as discussed above. Hence, while I did not follow anybody, it did garner three followers as shown in Table I. The 3% proportion of bonding social capital is, however, artificially inflated by I's small number of followers.

In passing, we note that the above results also seem to confirm the intuition that utilizing a random following strategy produces more follow-backs than following nobody at all. We do have to be a little careful here since, as mentioned above, the random strategy here may be confounded by our "bonding-friendly" pre-selection of users.

VI. CONCLUSION

Social media is becoming an important channel for sharing news and information. For many individuals and businesses, the very dynamic Twitter community is a particularly attractive social network to participate in. We have used a novel computational framework for social capital, together with a well-defined experiment, to verify the widely-held view that bonding interactions are more likely than bridging interactions in social networks.

Our experiments involved analyzing the behavior of a group of Twitter users in reaction to a number of artificial users with pre-defined strategies. The results considered such quantities as ratio of follow-backs to followings as well as accrued social capital. In particular, they show that users who request to follow others having similar profile descriptions (i.e., attempting to bond) increase the number of Twitter users that reciprocate their follow requests, thus generating significantly more bonding social capital. Indirectly, this highlights a strategy that a new user could employ to maintain a high follow-back ratio when interacting with people on Twitter.

ACKNOWLEDGMENT

We wish to thank Dennis L. Eggett for his assistance with the statistical analysis, and Mikaela Dufur for valuable comments that greatly improved the quality of the paper.

$\mathbf{rank}_c$	$\mathbf{rank}_f$	strategy	$\mathbf{clicks}_t$	$\mathbf{clicks}_p$	↓ total clicks	mentions
1	3	*random* (D)	900	9	909	2
2	1	*bonding* (A)	882	15	897	3
3	8	*min. num. followers* (H)	850	16	866	1
4	2	*max. following/ers diff.* (F)	849	7	856	1
5	4	*median affinities* (C)	846	9	855	1
6	6	*min. following/ers diff.* (E)	821	19	840	4
7	5	*bridging* (B)	773	11	784	2
8	9	*follow nobody* (I)	775	1	776	1

TABLE IV

CLICK STATISTICS: EACH OF THE NINE ACCOUNTS ARE LISTED BY *strategy* AND RANKED BY THE NUMBER OF *total clicks* RECEIVED, DENOTED $rank_c$. EACH ACCOUNT MADE APPROXIMATELY 117 STATUS UPDATES (I.E., TWEETS), OF WHICH 106 INCLUDED A CLICKABLE TRACKING LINK.

REFERENCES

[1] J. S. Coleman, "Social capital in the creation of human capital," *American Journal of Sociology*, vol. 94, pp. S95–S120, 1988.

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

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

[4] S. P. Borgatti, C. Jones, and M. G. Everett, "Network measures of social capital," *Connections*, vol. 21, no. 2, pp. 27–36, 2 1998.

[5] P. S. Adler and S.-W. Kwon, "Social Capital: Prospects for a New Concept," *The Academy of Management Review*, vol. 27, no. 1, p. 17, January 2002.

[6] P. Paxton, "Social capital and democracy: An interdependent relationship," *American Sociological Review*, vol. 67, no. 2, pp. 254–277, 2002.

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

[8] D. L. Haynie and D. W. Osgood, "Reconsidering peers and delinquency: How do peers matter?" *Social Forces*, vol. 84, no. 2, pp. 1109–1130, 2005.

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

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

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

[12] J. P. Scott, *Social Network Analysis: A Handbook*. Thousand Oaks, CA: Sage Publications Ltd; 2nd edition, 2000.

[13] M. Granovetter, "The strength of weak ties," *American Journal of Sociology*, vol. LXXVIII, 1973.

[14] C. Haythornthwaite, "Strong, weak, and latent ties and the impact of new media," *The Information Society*, vol. 18, no. 5, pp. 385–401, 2002.

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

[16] 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*, 2006, pp. 611–617.

[17] 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*, 2005, pp. 177–187.

[18] S. Redner, "Citation statistics from 110 years of *Physical Review*," *Physics Today*, vol. 58, pp. 49–54, 2005.

[19] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe, "A framework for community identification in dynamic social networks," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007, pp. 717–726.

[20] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, "Group formation in large social networks: membership, growth, and evolution," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006, pp. 44–54.

[21] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, "Microscopic evolution of social networks," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008, pp. 462–470.

[22] E. Zheleva, H. Sharara, and L. Getoor, "Co-evolution of social and affiliation networks," in *Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2009.

[23] N. Memon, D. L. Hicks, H. L. Larsen, and M. A. Uqaili, "Understanding the structure of terrorist networks," *International Journal of Business Intelligence and Data Mining*, vol. 2, no. 4, pp. 401–425, 2007.

[24] J. Xu and H. Chen, "The topology of dark networks," *Communications of the ACM*, vol. 51, no. 10, pp. 58–65, 2008.

[25] M. A. Shaikh and W. Jiaxin, "Network structure mining: locating and isolating core members in covert terrorist networks," *WSEAS Transactions on Information Science and Applications*, vol. 5, no. 6, pp. 1011–1020, 2008.

[26] S. P. Borgatti, "Identifying sets of key players in a social network," *Computational & Mathematical Organization Theory*, vol. 12, no. 1, pp. 21–34, 2006.

[27] T. Wey, D. T. Blumstein, W. Shen, and F. Jordán, "Social network analysis of animal behaviour: a promising tool for the study of sociality," *Animal Behaviour*, vol. 75, no. 2, pp. 333–344, 2008.

[28] A. Bhadra, F. Jordán, A. Sumana, S. A. Deshpande, and R. Gadagkar, "A comparative social network analysis of wasp colonies and classrooms: Linking network structure to functioning," *Ecological Complexity*, vol. 6, no. 1, pp. 48–55, 2009.

[29] N. A. Christakis and J. H. Fowler, "The collective dynamics of smoking in a large social network," *New England Journal of Medicine*, vol. 358, no. 21, pp. 2249–2258, 2008.

[30] 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*, vol. 39, no. 6, pp. 1568–1593, 1996.

[31] B. H. Erickson, "Good networks and good jobs: The value of social capital to employers and employees," in *Social Capital: Theory and Research*, N. Lin, K. S. Cook, and R. S. Burt, Eds. Aldine Transaction, 2004, ch. 6, pp. 127–158.

[32] L. Licamele and L. Getoor, "Social capital in friendship-event networks," in *Proceedings of the IEEE International Conference on Data Mining*, 2006, pp. 959–964.

[33] V. Barash, M. Smith, L. Getoor, and H. Welser, "Distinguishing knowledge vs social capital in social media with roles and context," in *Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media*, 2009.

[34] M. Smith, C. Giraud-Carrier, and B. Judkins, "Implicit Affinity Networks," in *Proceedings of the 17th Annual Workshop on Information Technologies and Systems*, 2007, pp. 1–6.

[35] M. Smith, C. Giraud-Carrier, and N. Purser, "Implicit affinity networks and social capital," *Information Technology and Management*, vol. 10, no. 2–3, pp. 123–134, 2009.

Identifying Health-Related Topics on Twitter

An Exploration of Tobacco-Related Tweets as a Test Topic

Kyle W. Prier[1], Matthew S. Smith[2], Christophe G. Giraud-Carrier[2], and Carl L. Hanson[1]

[1] Department of Health Science
[2] Department of Computer Science
Brigham Young University, Provo UT 84602, USA
kyle.prier@byu.edu, smitty@byu.edu, cgc@cs.byu.edu, carl_hanson@byu.edu

Abstract. Public health-related topics are difficult to identify in large conversational datasets like Twitter. This study examines how to model and discover public health topics and themes in tweets. Tobacco use is chosen as a test case to demonstrate the effectiveness of topic modeling via LDA across a large, representational dataset from the United States, as well as across a smaller subset that was seeded by tobacco-related queries. Topic modeling across the large dataset uncovers several public health-related topics, although tobacco is not detected by this method. However, topic modeling across the tobacco subset provides valuable insight about tobacco use in the United States. The methods used in this paper provide a possible toolset for public health researchers and practitioners to better understand public health problems through large datasets of conversational data.

Keywords: Data Mining, LDA, Public Health, Social Media, Social Networks, Tobacco Use, Topic Modeling

1 Introduction

Over recent years, social network sites (SNS) like Facebook, Myspace, and Twitter have transformed the way individuals interact and communicate with each other across the world. These platforms are in turn creating new avenues for data aquisition and research. Such web-based applications share several common features that we will use to define social network sites as described by Boyd & Ellision [3]. Although there are slight variations in the actual implementation of these features, each service enables users to (1) create a public profile, (2) define a list of other users with whom they share a connection, and (3) view and discover connections between other users within the system. Since SNS allow users to visualize and make public their social networks, this promotes new connections to be formed among users because of the social network platform [3, 9]. Not only do SNS enable users to communicate with other users with whom they share explicit social connections, but with a wider audience of users with whom

they would not have otherwise shared a social connection. Twitter, in particular, provides a medium whereby users can create and exchange user generated content with a potentially larger audience than either Facebook or Myspace.

Twitter is a social network site that allows a users to relay short messages no longer than 140 characters (known as "tweets") to those who choose to subscribe to that user's profile (known as "followers"). This process is defined as "microblogging," in that users can send short, concise messages that can be read by their followers. Twitter provides tools for users to communicate in real-time with each other and continues to grow in popularity. As of August 2010, a rough estimate of 54.5 million people used Twitter in the United States with 63% of those users under the age of 35 years. Forty-five percent were between the ages of 18 and 34[17]. In addition, the US accounts for 51% of all Twitter users worldwide. An analysis of 2,000 tweets in the US and England revealed that the majority (41%) of tweets were pointless babble. However, the balance of tweets have been designated as conversational (38%), pass-along value (9%), self-promotion (6%), spam (6%) and news (4%)[11]. Despite the extent of pointless babble among tweets, Twitter provides conversational data that are beneficial for researchers.

Because Twitter status updates, or tweets, are publicly available and easily accessed through Twitter's Application Programming Interface (API), Twitter offers a rich environment to observe social interaction and communication among Twitter users[15]. Although very little is known about the use of Twitter for communicating health-related information and experiences, such information can provide researchers and professionals with additional resources and tools for their research. Some studies have specifically used Twitter data to understand "real," or offline, behaviors and trends. Chew and Eysenbach[5] analyzed Tweets in an effort to determine the types and quality of information exchanged during the H1N1 outbreak. The majority of Twitter posts in this study were news related (46%) with only 7 of 400 posts containing misinformation. Scanfeld et al. [14] explored Twitter status updates related to antibiotics in an effort to discover evidence of misunderstanding and misuse of antibiotics. Their research indicated that Twitter offers a platform for the sharing of health information and advice. In an attempt to evaluate health status, Cullota [6] compared the frequency of influenza related Twitter messages with influenza statistics from the Centers for Disease Control and Prevention. Findings revealed a .78 correlation with the CDC statistics suggesting that monitoring Tweets might provide cost effective and quicker health status surveillance. Additionally, through identification and qualitative analysis of public health-related tweets, researchers are enabled to tailor health interventions more effectively to their audiences.

We chose to test our topic modeling effectiveness by focusing on tobacco use in the United States. Although we are interested in devising a method to identify and better understand public health-related tweets in general, a test topic provides a useful indicator through which we can guide and gauge the effectiveness of our methodology. Tobacco use is a relevant public health topic to use as a test case, as it remains one of the major health concerns in the

US. Tobacco use, primarily cigarette use, is one of the leading health indicators of 2010 as determined by the Federal Government [10]. Additionally, tobacco use is considered the most preventable cause of disease and has been attributed to over 14 million deaths in the United States since 1964 [16, 1]. Also, there remain approximately 400,000 smokers and former smokers who die each year from smoking-related diseases, while 38,000 nonsmokers die each year due to second-hand smoke [1, 12].

In this study, we address the problem of how to effectively identify and browse health-related topics within large datasets of conversational data, specifically Twitter. Although recent studies have implemented topic modeling to process Twitter data, these studies have focused on identifying high frequency topics to describe trends among Twitter users. Current topic modeling methods prove difficult to detect lower frequency topics that may be important to investigators. Such methods depend heavily on the frequency distribution of words to generate topic models. Public health-realted topics and discussions use less frequent words and are therefore more difficult to identify using traditonal topic modeling. In this study we want to answer the following questions:

- How can topic modeling be used to most effectively identify relevant public health topics on Twitter?
- Which public health-related topics, specifically tobacco use, are discussed among Twitter users?
- What are common tobacco-related themes?

In the following sections, we discuss our data sampling and analysis of tweets. We used Latent Dirichlet Allocation (LDA) to analyze terms and topics from the entire dataset as well as from a subset of tweets created by querying general, tobacco use-related terms. We highlight interesting topics and connections as well as limitations to both approaches. Finally, we discuss our conclusions regarding our research questions as well as possible areas of future study.

2 Methods and Results

In this section, we introduce our methods of sampling and collecting tweets. Additionally, we discuss our methods to analyze tweets through topic modeling. We use two distinct stages to demontrate topic modeling effectiveness. First, we model a large dataset of raw tweets in order to uncover health-related issues. Secondly, we create a subset of tweets by querying a raw dataset with tobacco-related terms. We run topic modeling on this subset as well and report our findings.

2.1 Data Sampling and Collection

In order to obtain a representative sample of tweets within the United States, we chose a state from each of the nine Federal Census divisions through a random selection process. For our sample, we gathered tweets from the following randomly

selected states: Georgia, Idaho, Indiana, Kansas, Louisiana, Massachusetts, Mississippi, Oregon, and Pennsylvania.

Using the Twitter Search API, recent tweets for each state were gathered in 2 minute intervals over a 14 day period from October 6, 2010 through October 20, 2010. To prepare the dataset for topic modeling, we remove all non-latin characters from the messages, replace all links within the dataset with the term "link," and only include users that publish at least two tweets. This process results in a dataset of 2,231,712 messages from 155,508 users. We refer to this dataset as the "comprehensive" dataset.

2.2 Comprehensive Dataset Analysis

We analyze the comprehensive dataset by performing Latent Dirichlet Allocation (LDA) on it to produce a topic model. LDA is an unsupervised machine learning generative probabilistic model which identifies latent topic information in large collections of data including text corpora. LDA represents each document within the corpora as a probability distribution over topics, while each topic is represented as a probability distribution over a number of words [2, 8]. LDA enables us to browse words that are frequently found together, or that share a common connnection, or topic. LDA helps identify additional terms within topics that may not be directly intuitive, but that are relevant.

We configure LDA to generate 250 topics distributions for single words (unigrams) as well as more extensive structural units (n-grams). We suspected that this topic model would provide less relevant topics relating to public health and specifically tobacco use, since such topics are generally less frequently discussed. The LDA model of the comprehensive dataset generally demonstrates topics relating to pointless babble, various sundry conversational topics, news, and some spam as supported by Kelly [11]. However, the model provides several topics, which contain health-realted terms, as shown in Table 1.

Several themes are identifiable from the LDA output: physical activity, obesity, substance abuse, and healthcare. Through these topics we observe that those relating to obesity and weight loss primarily deal with advertising. Additionally, we observe that the terms relating to healthcare refer to current events and some political discourse. Such an analysis across a wide range of conversational data demonstrates the relative high frequency of these health-related topics. However, as we suspected, this analysis fails to detect lower frequency topics, specifically our test topic, tobacco use. In order to discover tobacco use topics, we create a smaller, more focused dataset that we will refer to as the "tobacco subset."

2.3 Tobacco Subset Analysis

To build our "tobacco subset," we query the comprehensive dataset with a small set of tobacco-related terms: "smoking," "tobacco," "cigarette," "cigar," "hookah," and "hooka." We chose these terms because they are relatively unequivocal, and they specifically indicate tobacco use. We chose the term "hookah"

Topic	Most Likely Topic Components (n-grams)	%
44	gps app, calories burned, felt alright, race report, weights workout, christus schumpert, workout named, chrissie wellington, started cycling, schwinn airdyne, core fitness, vff twitter acct, mc gold team, fordironman ran, fetcheveryone ive, logyourrun iphone, elite athletes, lester greene, big improvement, myrtle avenue	0.01
45	alzheimers disease, breast augmentation, compression garments, sej nuke panel, weekly newsletter, lab result, medical news, prescription medications, diagnostic imaging, accountable care, elder care, vaser lipo, lasting legacy, restless legs syndrome, joblessness remains, true recession, bariatric surgery, older applicants, internships attract, affordable dental	0.01
131	weight loss, diet pills, acai berry, healthy living, fat loss, weight loss diets, belly fat, alternative health, fat burning, pack abs, organic gardening, essential oils, container gardening, hcg diet, walnut creek, fatty acids, anti aging, muscle gain, perez hilton encourages	0.04
Topic	**Most Likely Topic Components (unigrams)**	
18	high, smoke, shit, realwizkhalifa, weed, spitta, currensy, black, bro, roll, yellow, man, hit, wiz, sir, kush, alot, fuck, swag, blunt	13.63

Table 1. Comprehensive Dataset Topics Relevant to Public Health. The last column is the percent of tweets that used any of the n-grams within each topic.

because of an established trend, particularly among adolescents and college students, to engage in hookah usage [7]. The recent emergence of hookah bars provide additional evidence of the popularity of hookah smoking [13].

Querying the comprehensive dataset with these terms, results in a subset of 1,963 tweets. The subset is approximately 0.1% of the comprehensive dataset. After running LDA on this subset, we found that there were insufficient data to determine relevant tobacco-related topics. We thus extended the tobacco subset to include tweets that were collected and preprocessed in the same manner as the comprehensive dataset. The only difference is that we used tweets collected from a 4-week period between October 4, 2010 and November 3, 2010. This resulted in a larger subset of 5,929,462 tweets, of which 4,962 were tobacco-related tweets (approximately 0.3%). Because of the limited size of the subset, we reduced the number of topics returned by LDA to 5, while we retained the output of both unigrams and n-grams.

The topics, displayed in Table 2, contain several interesting themes relating to how Twitter users discuss tobacco-related topics. Topic 1 contains topics related not only to tobacco use, but also terms that relate to substance abuse including marijuana and crack cocaine. Topic 2 contains terms that relate to addiction recovery: quit smoking, stop smoking, quitting smoking, electronic cigarette, smoking addiction, quit smoking cigarettes, link quit smoking, and link holistic remedies. Topic 3 is less cohesive, and contains terms relating to addiction recovery: quit smoking, stop smoking, secondhand smoke, effective steps. Additionally, it contains words relating to tobacco promotion by clubs or

bars: drink specials, free food, ladies night. Topic 4 contains terms related to both promotion by bars or clubs (ladies free, piedmont cir, hookahs great food, smoking room, halloween party, million people) and marijuana use (smoking weed, pot smoking). Topic 5 contains several terms that relate to anti-smoking and addiction recovery themes: stopped smoking, smoking kills, chain smoking, ban smoking, people die, damn cigarette, hate cigarettes. In this case, topic modeling has helped to understand more fully how users are using tobacco-related tweets.

Topic	**Most Likely Topic Components (n-grams)**	%
1	smoking weed, smoking gun, smoking crack, stop smoking, cigarette burns, external cell phones, hooka bar, youre smoking, smoke cigars, smoking kush, hand smoke, im taking, smoking barrels, hookah house, hes smoking, ryder cup, dont understand, talking bout, im ready, twenty years people	0.16
2	quit smoking, stop smoking, cigar guy, smoking cigarettes, hookah bar, usa protect, quitting smoking, started smoking, electronic cigarette, cigars link, smoking addiction, cigar shop, quit smoking cigarettes, chronical green smoke, link quit smoking naturally, smoking pot, youtube video, link quit smoking, link holistic remedies, chronical protect	0.27
3	cigarette smoke, dont smoke, quit smoking, stop smoking, smoking pot, im gonna, hookah tonight, smoking ban, drink specials, free food, ladies night, electronic cigarettes, good times, smoking session, cigarette break, secondhand smoke, everythings real, effective steps, smoking cigs, smoking tonight	0.22
4	smoking weed, cont link, ladies free, piedmont cir, start smoking, hate smoking, hookahs great food, cigarette butts, thingswomenshouldstop-doing smoking, lol rt, sunday spot, cigarettes today, fletcher knebel smoking, pot smoking, film stars, external cell, fetishize holding, smoking room, halloween party, million people	0.25
5	smoke cigarettes, smoking hot, im smoking, smoking section, stopped smoking, chewing tobacco, smoking kills, chain smoking, smoking area, ban smoking, people die, ring ring hookah ring ring, love lafayette, link rt, damn cigarette, healthiest smoking products, theyre smoking, hate cigarettes, world series, hideout apartment	0.06

Table 2. Tobacco Subset Topics. The most likely topic components for each of the five topics generated by LDA. The last column is the percent of tweets that used any of the n-grams within each topic.

3 Discussion and Conclusion

In this study, we directly address the problem of how to effectively identify and browse health-related topics on Twitter. We focus on our test topic, tobacco

use, throughout the study to explore the realistic application and effectiveness of LDA to learn more about health topics and behavior on Twitter. As expected, we determine that implementing LDA over a large dataset of tweets provides very few health topics. The health topics that LDA does produce during this first stage suggeset the popularity of these topics in our dataset. The topic relating to weight loss solutions indicate a high frequency of advertisements in this area. The topic relating to healthcare, Obama, and other current health issues indicate the current trend in political discourse related to health. Additionally, the high frequency of marijuana-related terms indicates a potentially significant behavior risk that can be detected through Twitter. While this method did not detect lower frequency topics, it may still provide public health researchers insight into popular health-related trends on Twitter. This suggests that there is potential research needed to test LDA as an effective method to identify health-related trends on Twitter. The second method we used to identify tobacco-related topics appears to be most promising to identify and understand public health topics. Although this method is less automated and requires us to choose terms related to tobacco use, the results indicated this method to be a valuable tool for public health researchers.

Based on the results of our topic model, Twitter has been identified as an effective tool to better understand health-related topics, such as tobacco. Specifically, this method of Twitter analysis enables public health researchers to better monitor and survey health status in order to solve community health problems[4]. Because LDA generates relevant topics relating to tobacco, we are able to determine themes and also the manner in which tobacco is discussed. In this way, irrelevant conversations can be removed, while tweets related to health status can be isolated. Additionally, by identifying relevant tweets to monitor health status, public health professionals are able to create and implement health interventions more effectively. Researchers can collect almost limitless Twitter data in their areas that will provide practitioners with useful, up-to-date information necessary for understanding relevant public health issues and creating targeted interventions.

Finally, the results from the second method suggest that researchers can better understand how Twitter, a popular SNS, is used to promote both positive and negative health behaviors. For example, Topic 4 contains terms that indicate that establishments like bars, clubs, and restaurants use Twitter as a means to promote business as well as tobacco use. In contrast, Topic 2 contains words that relate to addiction recovery by promoting programs that could help individuals quit smoking.

The use of LDA in our study demonstrates its potential to extract valuable topics from extremely large datasets of conversational data. While the method proves a valuable outlet to automate the process of removing irrelevant information and to hone in on desired data, it still requires careful human intervention to select query terms for the construction of a relevant subset, and subsequent analysis to determine themes. Research is required to further automate this process. In particular, new methods that can identify infrequent, but highly relevant

topics (such as health) among huge datasets will provide value to public health researchers and practitioners, so they can better identify, understand, and help solve health challenges.

Acknowledgements

We would like to thank Matthew Gardner for valuable discussions regarding LDA.

References

1. Armour, B.S., Woolery, T., Malarcher, A., Pechacek, T.F., Husten, C.: Annual Smoking-Attributable Mortality, Years of Potential Life Lost, and Productivity Losses. MMWR Morb Mortal Wkly Rep. 54, 625-628 (2005)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation Journal of Machine Learning Research 3, 993-1022, (2003)
3. Boyd, D.M., Ellison, N.B.: Social Network Sites: Definition, History, and Scholarship. Journal of Computer-Mediated Communication. 13, 210-230 (2008)
4. Centers for Disease Control and Prevention, http://www.cdc.gov/od/ocphp/nphpsp/essentialphservices.htm
5. Chew, C.M., Eysenbach, G.: Pandemics in the age of Twitter: Content analysis of "tweets" during the H1N1 outbreak. Paper presented September 17, 2009, Medicine 2.0, Naastricht, NL.
6. Culotta, A.:Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. Paper presented at the KDD Workshop on Social Media Analytics, 2010.
7. Eissenberg, T., Ward, K.D., Smith-Simone, S., Maziak, W.: Waterpipe Tobacco Smokking on a U.S. College Campus:Prevalence and Correlates. Journal of Adolescent Health. 42, 526-529 (2008)
8. Griffiths, T.L., Steyvers, M.: Finding Scientific Topics. Proceedings of the National Academy of Sciences. 101, 5228-5235 (2004)
9. Haythornthwaite, C.: Social networks and Internet connectivity effects. Information, Communication, & Society. 8, 125-147 (2005)
10. Health People 2010, http://www.healthypeople.gov/lhi/
11. Kelly, R.: Twitter study - August 2009. San Antionio, TX: Pear Analytics (2009)
12. Mokdad, A.H., Marks, J.S., Stroup, D.F., Gerberding, J.L.: Actual Causes of Death in the United States. Journal of the American Medical Association. 291, 1238-1245 (2004)
13. Primack, B.A., Aronson, J.D., Agarwal, A.A.: An Old Custom, a New Threat to Tobacco Control. American Journal of Public Health. 96, 1339 (2006)
14. Scanfield, D., Scanfield, V., Larson, E.: Dissemination of Health Information through Social Networks: Twitter and Antibiotics. American Journal of Infection Control. 38, 182-188 (2010)
15. Twitter API documentation, http://dev.twitter.com/doc
16. U.S. Department of Health and Human Services (USDHHS): The Health Consequences of Smoking: A Report for the Surgeon General. Report, USDHHS, Centers for Disease Control and Prevention (CDC), National Center for Chronic Disease Prevention adn Health Promotion, Office on Smoking and Health (2004)
17. Quantcast, http://www.quantcast.com/twitter.com

Social Capital and Language Acquisition during Study Abroad

M. Smith (m.smithworx.com) and C. Giraud-Carrier (cgc@cs.byu.edu)
Department of Computer Science, Brigham Young University
Provo, UT 84602 USA

D. Dewey (ddewey@byu.edu), S. Ring (ring.spencer84@gmail.com) and D. Gore (d.gore@byu.edu)
Department of Linguistics, Brigham Young University
Provo, UT 84602 USA

Abstract

We study the role of social capital in language acquisition during study abroad. Using data collected from 204 participants in Japanese study abroad programs, we show that students who leverage social capital through bridging relationships feel they achieve higher levels of language improvement. Furthermore, an analysis of the topics participants discuss with locals suggests that there are significant differences between students who have a tendency to build close-knit networks and students who cast a broader net.

Keywords: Social Capital; Second Language Acquisition; Study Abroad.

Introduction

Research in second language acquisition during study abroad has dealt with a number of issues including language use, proficiency development (Badstübner & Ecke, 2009; Mendelson, 2004), and language socialization (Fraser, 2002; Campbell, 1996). Language socialization involves becoming integrated into a community that allows one to practice the second language in meaningful social contexts (Wang, 2010). Language socialization is a complex process affected by a range of variables, including motivation, attitudes, interlocutor attributes, and a range of other variables (Isabellí-Garcia, 2006), and the few studies conducted to date suggest that socialization can affect language acquisition (Mendelson, 2004; Whitworth, 2006).

One particularly interesting measure of social networks, which has been popularized and aggressively pursued in the past couple of decades, is social capital (Coleman, 1988; Lin, 2001; Putnam, 2000). Unlike most other forms of capital that tend to emphasize what people *possess* individually, social capital is an inherently social measure that focuses on the *relationships* that exist among people. Indeed, social capital attempts to quantify the value of such relationships in achieving some individual or group benefit based on the resources present in the underlying network (Borgatti, Jones, & Everett, 1998; Adler & Kwon, 2002).

An analysis of language acquisition during study abroad from the social capital perspective is unprecedented within the existing body of second language acquisition and study abroad research. Although social capital might be more traditionally thought of by some as future employment opportunities or the capacity to secure social favors, an individual's ability to acquire and utilize social capital during study abroad would appear to be consequential in second language acquisition, primarily as a means of exposure to the second language. From this, it is clear that an exploration of second language acquisition and study abroad in the context of social capital merits the critical consideration of those researching second language acquisition and language socialization. We present one such exploration here for a Japanese study abroad program involving over 200 participants.

The paper is organized as follows. We first provide a brief overview of our social capital framework and show how it is specialized to the context of our language acquisition during study abroad analysis. We then present our data and methodology, and show how a number of indicators such as perceived language proficiency and conversation topics vary based on students' social behavior. Finally, we conclude with a discussion of the novel insight into language acquisition provided by the social capital perspective.

Social Capital Framework

Space does not permit us to give a full account of our computational framework for social capital. We give only a brief description of its main components and state the simplifying assumptions we make to apply it here. Further details about the framework are in (Smith, Giraud-Carrier, & Purser, 2009; Smith & Giraud-Carrier, 2010).

Social capital is grounded in relationships, individuals' attributes, and available resources. To exploit this information, we find it useful to distinguish between two types of relationships among individuals, as follows.

- An *explicit* connection links one individual to another based on some purposive action (e.g., sending an email, visiting) or a well-defined relationship (e.g., being a friend of, collaborating with). Individuals thus linked are aware of the explicit connections among them.
- An *implicit* affinity connects individuals together based on loosely defined affinities, or inherent similarities, such as similar hobbies or shared interests. Individuals 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 arising from implicit affinities (Smith, Giraud-Carrier, & Judkins, 2007). Social capital is naturally interested in implicit affinities, since it clearly has some relation to shared affiliations or activities among individuals (Belliveau, O'Reilly, & Wade, 1996). On the other

hand, social capital can really only accrue, or be realized, when individuals are aware of it, that is, when they establish explicit connections among themselves. It follows that *hybrid networks*, i.e., networks that include both implicit affinities and explicit connections, play a key role in the definition and analysis of social capital.

Note that in a strict sense, social capital is only realized once actions are taken and their result evidences the presence of said social capital. Hence, typical studies of social capital are retrospective. Within our framework, however, we wish to use the notion of social capital to reason about how one could leverage one's relations. For example, given that X and I are friends, that X is a headhunter and that I am looking for a job, I would probably want to ask X to help me find a job. While evidence of any social capital will truly become apparent only if and when X chooses to help me, it seems most reasonable for me to try to take advantage of my friendship with X. For simplicity here, we equate the presence of an explicit link with the presence of social capital.

We also find it useful to adopt Putnam's high-level dichotomy of social capital into bonding social capital and bridging social capital to provide a general characterization of individuals' (here, learners') behaviors (Putnam, 2000; Putnam & Feldstein, 2003). Bonding refers to the tendency that individuals may have to associate with others who are similar to them, leading to homogeneous groups. Bridging occurs when individuals associate with others who are not like them, leading to heterogeneous groups. The types of links connecting individuals give rise to bonding and/or bridging social capital, as follows.

1. Implicit affinities only. In this case, the individuals have much in common (e.g., similar occupation or hobbies) but they are unaware of it. If they were to connect explicitly, they would be bonding, but since they have not yet, we say that there is only potential for bonding social capital.

2. Implicit affinities and explicit connections. In this case, the potential for bonding social capital is now realized as similar individuals connect to one another explicitly.

3. No implicit affinities and no explicit connections. In this case, the individuals have nothing in common and they are unaware of each other. If they were to connect explicitly, they would be bridging, but since they have not yet, we say that there is only potential for bridging social capital.

4. No implicit affinities but explicit connections. In this case, the dissimilar individuals are now connected to one another (e.g., colleagues collaborating across disciplines or members of a church choir). Hence, we say that there is realized bridging social capital.

The foregoing treats affinities and explicit connections as aggregate binary entities that are either present or absent. In practice, of course, these links may exist with varying degrees of strength. For example, two individuals may have some things in common and others not. Shared attributes, attitudes and behaviors represent opportunities for bonding, while differences among the same represent opportunities for bridging. Thus, there is generally both bonding and bridging social capital between individuals. Furthermore, while it seems appropriate for implicit affinities to be "undirected," since two people either share or do not share a specific affinity, it is not so for explicit edges. Indeed, it is clear that the value of some (explicit) relationships is not necessarily reciprocal and may vary among participants. For example, one person may consider another person as their best friend, while that other person may look at the first as only a good friend. Thus, our framework recognizes that the amount of social capital an individual i may realize from a relationship with another individual j is not predicated upon the value that i places in the relationship, but rather upon the value that j places in it. While i may think highly of that connection, for example in the context of obtaining a job reference from j, the reference will only be as strong as j thinks of i, and not the other way.

We can now turn to a formal account of social capital in hybrid networks. Let $s_{ij}^{IAN} \in [0,1]$ be the strength of the implicit affinity, or measure of similarity, between individuals i and j. It follows that s_{ij}^{IAN} stands for the potential for bonding that exists between i and j, while its reciprocal, $1 - s_{ij}^{IAN}$, stands for the potential for bridging that exists between i and j. Similarly, let s_{ij}^{ESN} be the strength of the explicit connection between individuals i and j. s_{ij}^{ESN} may be as simple as 1 or 0, to reflect the presence or absence of a link, but may also range over [0,1] to capture degrees of connectivity (e.g., best friend vs. casual friend vs. acquaintance). Finally, let *Ind* be the set of individuals in the network.

The bonding social capital *realized* by a node i, when (explicitly) connecting with node j, is naturally given as the product of the strength of the implicit affinity between i and j by the strength of the explicit edge connecting j to i: $s_{ij}^{IAN} s_{ji}^{ESN}$. As expected, if j is unaware of i, even when i may be aware of (and possibly even count on) j, there is no social capital available for i from that relationship. The (realized) bonding social capital of an individual i is then the sum of its realized bonding social capital with all other individuals. That is,

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

Likewise, the (realized) bridging social capital of an individual i is the sum of its realized bridging social capital with all other individuals. That is,

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

Methodology

Participants were 204 former recipients of Bridging Scholarships for study abroad in Japan (101 male and 103 female, average age 21.3 years, $SD = 2.90$). These students had studied Japanese for an average of 2.07 years ($SD = 1.87$) prior

to their departure for Japan. They spent an average of 8.4 months ($SD = 3.70$) in Japan, taking 13.2 hours per week ($SD = 5.27$) of Japanese language courses in 38 language programs across 22 different cities.

To capture learners' perspectives regarding gains in speaking proficiency over study abroad, we had students complete a Then-Now self-assessment (Rohs & Lagone, 1997), based on an oft-used self-assessment instrument designed by Clark (1981). Then-Now measurement is common in educational research as a means of measuring the effectiveness of program interventions and although not as objective as traditional standardized tests of language proficiency, results correlate at moderate degrees with such standardized measures and yield highly reliable results (Dewey, 2002; Lam & Bengo, 2003). Our Then-Now survey presents tasks based on the ACTFL Speaking Proficiency Guidelines (Breinder-Sanders, Lowe, Miles, & Swender, 2000) ranging from Novice-Level to Superior-Level, and asks learners to rate their ability on a 1 (not at all able) to 5 (quite easily) scale. Then-Now reliability estimates were high (Cronbach's Alpha=.97 for Then and .96 for Now).

We measured language use via a web-based version of the Language Contact Profile (LCP), a survey created by Freed et al. (Freed, Dewey, Segalowitz, & Halter, 2004). Social network information was obtained via a thirteen-question version of the Study Abroad Social Interaction Questionnaire (SASIQ) developed by Dewey et al. (2011). The SASIQ consists of items designed to allow the computation of various social network measures, such as size, intensity and dispersion (Scott, 2000). The version of the SASIQ we used also contained items asking learners to identify their friends together with the topics about which they spoke with each one of them. The social network for each participant is thus best represented as a star network (see Figure 1), where the central node is the participant, and the nodes around the periphery represent individuals listed by the participant as friends. The survey allowed for a maximum of 20 friends, but only 56 participants listed 20 friends.

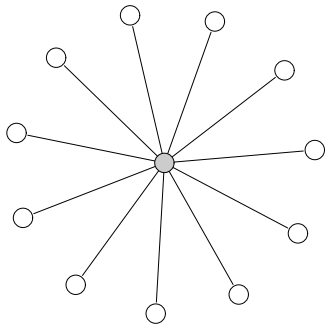


Figure 1: Participants' Social Network

Social Capital for Language Acquisition

Due to the nature of the application, we must specialize our general social capital framework. In particular, we make the following assumptions. (Note: we are only interested in realized social capital, so there is no need to consider implicit links with those not listed as friends.)

1. The only explicit links are between participants and their listed friends. These links are assumed to be undirected and of strength 1, that is,

$$s_{ij}^{ESN} = s_{ji}^{ESN} = \begin{cases} 1 & \text{for each participant } i \text{ and friend } j \\ 0 & \text{otherwise} \end{cases}$$

2. Implicit links are determined only by the topics discussed with friends. Indeed, we have no other information about friends that would allow further affinities to be considered. Each topic of discussion is a possible affinity between individuals.

3. The set T_i of topics discussed by X_i with all of its listed friends is the complete set of possible affinities among them, i.e., we assume that the friends have no other topics of conversation than those pursued with X_i.

4. If T_i is the set of all topics discussed by X_i and $T_{ij} \subset T_i$ is the set of topics that X_i discusses with X_j, then

   (a) both X_i and X_j are interested in the topics in T_{ij}, so that the topics in T_{ij} make up the implicit affinities between X_i and X_j, and

   (b) the ratio $\frac{|T_{ij}|}{|T_i|}$ can be used as a measure of the strength of the affinity between X_i and X_j, that is,

$$s_{ij}^{IAN} = \frac{|\,T_{ij}\,|}{|\,T_i\,|}$$

Bonding and bridging social capitals are then computed as per the general framework's equations. Intuitively, if X_i discusses similar topics with all of his/her friends (i.e., $T_{ij} \simeq T_i$ for all X_j with whom X_i is connected), then X_i has a tendency to bonding, while if X_i discusses different topics with different friends (i.e., $T_{ij} \neq T_{ik}$ for $X_j \neq X_k$), then X_i has a tendency to bridging.

The reader may have noticed that our definition of bridging social capital may be impacted by the number of friends a participant has. Indeed, there is a strong correlation ($r = 0.95$, $p < .0001$) between these two quantities as shown in Figure 2. However, we wish to point out that, in general, bridging social capital is a finer and richer measure as manifested by the vertical dispersion of points on the figure. One extreme case is highlighted by the points labeled a and b. Both of these have 17 friends, and hence would be considered the same under that measure. Yet, b has high bridging, while a has very low bridging.

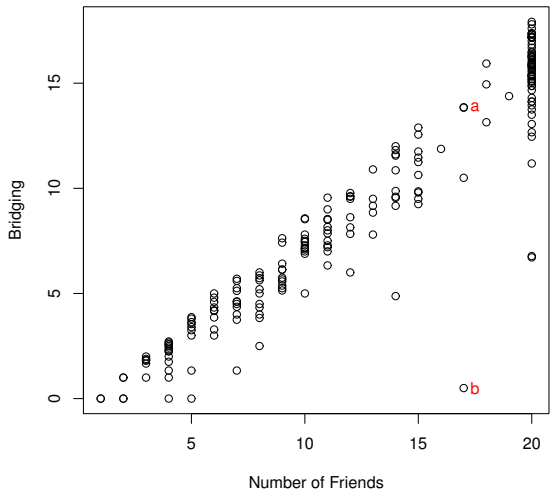


Figure 2: Bridging vs. Number of Friends

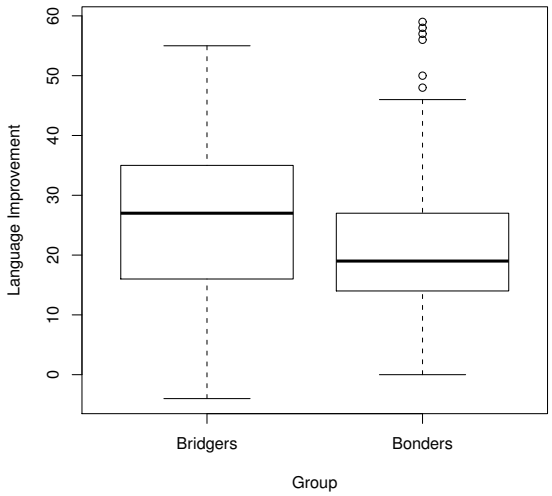


Figure 3: Influence of Social Capital on Language Improvement

Summary of Findings

In this section, we show how learners' self-reported language proficiency and conversation topics vary based on their social behavior.

The aggregate language improvement score is the sum of the differences in the pre- and post- of the 21 self-evaluated language scores. For our participants, the language improvement scores range from -4 to +59.

Rather than carry two scores, one for bonding social capital and one for bridging social capital, we grouped participants into "bonders" and "bridgers" based on their tendency to either behavior. That tendency was computed as the difference between their bonding and bridging social capital values. Individuals with values greater than (or equal to) the mean tendency value were labeled as "bonders", while those with values less than the mean were labeled as "bridgers."

Self-Perceived Gains in Language Proficiency

Figure 3 shows the box-plots comparing the bonders and bridgers groups, with respect to their language improvement. ANCOVA results indicated a significant effect of social capital (bridging vs. bonding) on language improvement (gains from pre- to post-) after controlling for pre-departure proficiency estimate and time in Japan (both found to be predictors of gains in studies cited previously), $F(1,201) = 12.53$, $p < .0001$. Bridgers ($N = 90$, $M = 26.4$, $SD = 12.4$) fared significantly better than bonders ($N = 114$, $M = 21.9$, $SD = 12.1$).

Topic and Group Analysis

Table 1 shows the number of topics used within each of the groups, the number of participants, and the average number of topics that each participant used within each group. On average participants within the bonders group had discussed 7 different topics, while those within the bridgers group discussed 11 topics (4 more).

Table 1: Number of Conversation Topics by Social Capital Group

Group	Topics	Participants	Topics / Participant
Bonders	786	114	6.89
Bridgers	992	90	11.02

Although each topic was discussed by at least one person within each group, some topics were discussed more frequently by participants within each group. Figure 4 shows each topic and which group discussed it most frequently. The upper region of the plot shows the topics that bridgers used more often than bonders, while the lower region shows the topics that bonders used more often than bridgers. The scale represents how many more bridgers/bonders used the given topic. For example, the "life views and ideals" topic (in the upper region of the plot) was discussed by 20 more bridgers than bonders. On the other hand, the "business and economics" topic (in the lower region of the plot) was discussed by five more bonders than bridgers. Topics having the same difference in participants are separated by a slash ('/').

According to this ranking, the most disparate topics were "many topics" and "academics". The two topics used more frequently by bridgers are "many topics" and "random topics", which suggest that students within this group talked about a larger variety of topics than others. On the opposite end, it seems that "academics" is a safe topic for any student

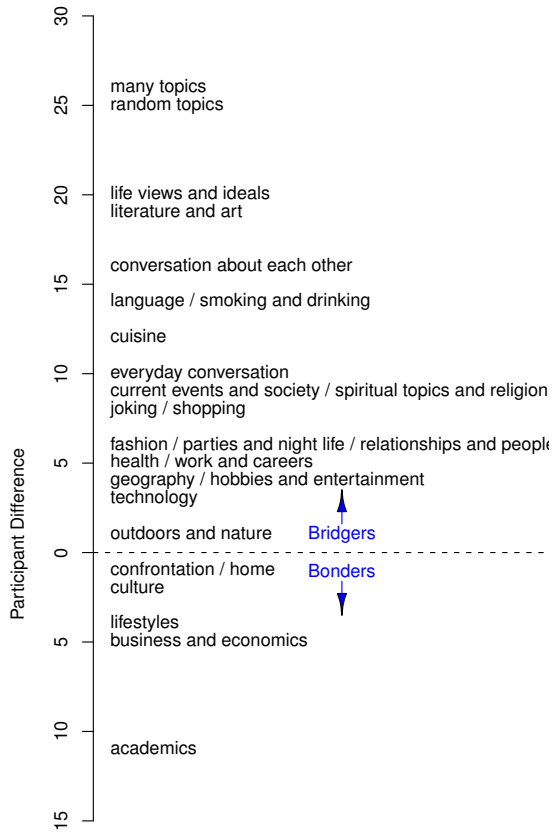


Figure 4: Bridger vs. Bonder Topics

as it is discussed in classes and is something that could possibly even limit who they speak to.

Discussion

The above analysis highlights the potential of social capital as an explanatory variable in study abroad research. While factors such as time abroad, pre-departure proficiency, and grammatical knowledge have received relatively large amounts of attention in study abroad research (Davidson, 2010), social capital and its role in social networking have yet to be explored. Previous studies have shown that developing social networks with native speakers while abroad via volunteer work, part-time employment, club membership, etc. can facilitate language acquisition (Isabellí-Garcia, 2006; Whitworth, 2006). Our research adds to this knowledge by demonstrating that it may not simply be a matter of developing social networks, but also bridging with others by conversing about a range of topics.

The ACTFL Speaking Proficiency Guidelines (Breinder-Sanders et al., 2000) contain a number of descriptions of higher levels of proficiency that indicate learners who wish to become more proficient can benefit from discussing a range of topics. For example, learners at the Superior level are expected to "participate fully and effectively in conversations on a variety of topics in formal and informal settings from both concrete and abstract perspectives," whereas learners one level below (Advanced) "cannot sustain performance at that [abstract] level across a variety of topics," but "are more comfortable discussing a variety of topics concretely" (pp. 14-15). The complexity and variety of topics learners are able to control in the second language decreases with lower level abilities, a pattern in line with the connections between topic range and perceived proficiency level in our research.

This work also has implications for pre-departure preparation. If students are better prepared both linguistically and mentally to engage in a variety of topics while abroad, their chances of making gains while abroad are likely to improve. Those who practice a variety of topics to the extent that they are able to participate in discussions even haltingly prior to going abroad are more likely to be able to bridge while abroad, discussing a variety of topics with a variety of people (DeKeyser, 2007) and taking fuller advantage of a setting where "perhaps the most crucial intervention is to give [students] assignments that force them to interact meaningfully with [locals] and overcome their fear of speaking" (p. 218). Overcoming this fear of speaking may have also been partially responsible for the bridgers' ability to discuss a variety of topics in the second language. In this study we did not measure personality–a weakness we are addressing in follow-up research, where we have collected data investigating the roles of personality, motivation, and affective variables in the creation of social capital during study abroad. Naiman and his colleagues (1996) observed that extroversion and sociability are important in learning one's second language. It is possible that highly extroverted learners are more likely to discuss a variety of (often less familiar) topics than less extroverted learners. We hope to elucidate the role of personality in our ongoing and future work.

Conclusion

In conclusion, this study has explored social capital as it pertains to second language acquisition. Our framework has allowed us to consider participants' language socialization according to their bridging and bonding social capital. By this, we have attempted to assess whether bridging or bonding better predicts improvement in language skills. This method of analysis is valuable from a social capital perspective, as increased language abilities can potentially open doors to new venues for relationship development. Also, the results we have presented confirm the notion that social capital, in terms of bridging and bonding, can provide important insights into language socialization and acquisition.

The information we base our conclusions on is limited to the conversation topics reported by our participants in the survey heretofore discussed. We are currently engaged in additional research investigating the nature of conversations learners have (length and type of discourse, etc.). Additional work might also look at other forms of data or investigate if similar trends explain connections between language socialization

and acquisition that occurs via the Internet. Surely, there are a multitude of potential environments in which the implications of this model can be observed that have yet to be considered.

Acknowledgments

Data collection for this research was funded by a grant from the U.S. Department of Education (International Research and Studies Program, P017A080087).

References

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

Badstübner, T., & Ecke, P. (2009). Student expectations, motivations, target language use, and perceived learning progress in a summer study abroad program in germany. *Die Unterrichtspraxis/ Teaching German*, *42*(1), 41-49.

Belliveau, M., O'Reilly, C. I., & 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.

Borgatti, S. P., Jones, C., & Everett, M. G. (1998, 2). Network measures of social capital. *Connections*, *21*(2), 27-36.

Breinder-Sanders, K., Lowe, P., Miles, J., & Swender, E. (2000). ACTFL proficiency guidelines: Speaking, revised 1999. *Foreign Language Annals*, *33*(1), 13-18.

Campbell, C. (1996). Socializing with the teachers and prior language learning experience: A diary study. In K. Bailey & D. Nunan (Eds.), *Voices from the classroom* (p. 201-223). New York: Cambridge University Press.

Clark, J. (1981). Language. In T. Barrows (Ed.), *College students knowledge and beliefs: A survey of global understanding.* Princeton, NJ: Educational Testing Service.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, *94*, S95–S120.

Davidson, D. (2010). Study abroad: When, how long, and with what results? new data from the Russian front. *Foreign Language Annals*, *43*(1), 6-26.

DeKeyser, R. (2007). Study abroad as foreign language practice. In R. DeKeyser (Ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology* (p. 208-226). Cambridge: Cambridge University Press.

Dewey, D. (2002). *Self-assessment: Possible uses in a standards-based classroom.* Presentation given at the Annual Convention of the American Council on the Teaching of Foreign Languages (ACTFL), Salt Lake City, UT.

Dewey, D., Bown, J., & Eggett, D. (2011). Japanese language proficiency, social networking, and language use during study abroad: Learners' perspectives. *Under Review*.

Fraser, C. (2002). Study abroad: An attempt to measure the gains. *German as a Foreign Language Journal*, *1*, 45-65.

Freed, B., Dewey, D., Segalowitz, N., & Halter, R. (2004). The language contact profile. *Studies in Second Language Acquisition*, *26*, 349-356.

Isabellí-Garcia, C. (2006). Study abroad social networks, motivation, and attitudes: Implications for second language acquisition. In M. Dufon & E. Churchill (Eds.), *Language learners in study abroad contexts* (p. 231-258). Clevedon, UK: Multilingual Matters.

Lam, T., & Bengo, P. (2003). A comparison of three retrospective self-reporting methods of measuring change in instructional practice. *American Journal of Evaluation*, *24*(1), 65-80.

Lin, N. (2001). *Social capital: A theory of social structure and action*. Cambridge, England: Cambridge University Press.

Mendelson, V. (2004). *Spain or bust? assessment and student perceptions of out-of-class contact and oral proficiency in a study abroad context*. Unpublished doctoral dissertation, University of Massachusetts at Amherst, Paper AAI3136758.

Naiman, N., Fröhlich, M., Stern, H., & Todesco, A. (1996). *The good language learner*. Toronto: The Ontario Institute for Studies in Education.

Putnam, R. D. (2000). *Bowling alone: the collapse and revival of american community*. New York, NY, USA: Simon & Schuster.

Putnam, R. D., & Feldstein, L. M. (2003). *Better together: Restoring the american community*. New York, NY, USA: Simon & Schuster.

Rohs, F., & Lagone, C. (1997). Increased accuracy in measuring leadership impacts. *Journal of Leadership Studies*, *4*(1), 150-158.

Scott, J. P. (2000). *Social network analysis: A handbook*. Thousand Oaks, CA: Sage Publications Ltd; 2nd edition.

Smith, M., & Giraud-Carrier, C. (2010). Bonding vs. bridging social capital: A case study in twitter. In *Proceedings of the 2nd international symposium on social intelligence and networking* (p. 385-392).

Smith, M., Giraud-Carrier, C., & Judkins, B. (2007). Implicit Affinity Networks. In *Proceedings of the 17th annual workshop on information technologies and systems* (p. 1-6).

Smith, M., Giraud-Carrier, C., & Purser, N. (2009). Implicit affinity networks and social capital. *Information Technology and Management*, *10*(2–3), 123–134.

Wang, C. (2010). Toward a second language socialization perspective: Issues in study abroad research. *Foreign Language Annals*, *43*(1), 50-63.

Whitworth, K. (2006). *Access to learning during study abroad: The roles of identity and subject positioning*. Unpublished doctoral dissertation, The Pennsylvania State University.

Part IV

Conclusion

Online social networks continue to generate huge amounts of data, thus offering the emerging cross-disciplinary field of Computational Social Science extraordinary opportunities to develop novel visualization and analysis techniques, to test various social theories and to increase our understanding of the core issues that challenge societies. The research presented here falls broadly within this area.

Our main contribution is the general computational framework we present for quantifying and reasoning about social capital, that incorporates affinities, relationships, interactions and social resources, and attempts to unify a number of traditional operationalizations of social capital. We have introduced the notion of implicit affinity networks, and showed how implicit affinities are critical to distinguishing between potential for and actual social capital, as well as to creating both bonding and bridging social capital. Explicit relationships were discussed both formally and practically through examples derived from active online community interaction data. The combined effects of affinities and explicit relationships highlighted potential and actual channels where social capital offers benefit. We have discussed and shown how social resources are important to accessing and mobilizing social capital within a resource-aware community.

Consistent with the aim of computational social science, we validate our framework through social theory confirmation and empirical studies using behavioral data. Well-known

principles from social theory including reciprocity [Gouldner, 1960, Sugden, 1984], homophily [McPhearson et al., 2001], and bonding and bridging [Putnam, 2000, Putnam and Feldstein, 2003] are used and, in some cases, tested with the framework. In particular, we have shown, through case studies, how our framework can be applied in a variety of domains and provide additional insight to problems in these domains. Thus, in addition to the core computational framework, our research also makes the following indirect contributions to the social sciences.

- Within the blogosphere, we measured social capital by combining implicit connections based on shared topics and explicit connections based on cross-referencing, and identified potential sub-communities that would result through increased bonding social capital.
- Using Twitter data, we tested the widely-held view of social scientists that bonding interactions are more likely than bridging interactions (i.e., the principle of homophily) and found that indeed users who request to follow others having similar profile descriptions (i.e., attempting to bond) cause a significantly larger number of Twitter users to reciprocate their follow requests than others. From a practical standpoint, this result also informs how a new user might interact on Twitter to maintain a high follow-back ratio.
- Leveraging the implicit affinity aspects of the framework, we designed a study that examines how to model and discover public health topics and themes in tweets. Some useful, though limited, insight was gained about tobacco-related issues. Importantly however, the methods used provide a possible toolset for public health researchers and practitioners to better understand public health problems through large datasets of conversational, or social media-generated data.
- Using data collected from over 200 participants in Japanese study abroad programs, we showed that students who leveraged social capital through bridging relationships achieved higher levels of language improvement. Furthermore, an analysis of the topics participants discussed with locals suggested that there are significant differences between

students who have a tendency to build close-knit networks and students who cast a broader net.

While we feel that our work is a significant step forward, we do realize that much still remains to be done. Van Deth recently concluded his own review of the literature on social capital stating that "the wide variety of operationalizations should be accepted as an indication of the importance and vitality of the study of social life in complex societies, and empirical research should adapt to this liveliness" [van Deth, 2008]. The study of social capital through computational means is a fascinating area of research that we expect will expand throughout the next several years.

New techniques will need to be developed to help us understand the complexities of individuals and groups interacting within communities. Among other things, continued research in this area may offer promise to educational reform, focused health communities, enhanced collaboration environments, and efficient resource exchange.

Some of the future work, very specific to this research, is mentioned in the conclusion of the individual papers contained in this compilation. A few additional broader avenues of future work are as follows.

- Perform additional resource simulations with multiple types of individuals within the network (e.g., producers and consumers, altruists and free-riders).
- Map practical selection functions (i.e., sel_j) to specific tasks (e.g., finding a job, obtaining support, learning a new skill). For example, if one's task is to *obtain support*, then one will likely wish to select individuals with whom many affinities are shared, in particular, those with whom they share the challenge for which support is sought.
- Utilize prior social exchange theory to seed reasonable hypotheses to then be empirically tested within the framework. Perhaps, suitable individual-level exchange functions could be developed.

- Design a range of relationship strength updating functions (i.e., $\Delta s_{ji}^{ESN}(event)$) that allow external feedback (e.g., holidays, weather, crisis) to affect changes.
- Explore the economics of automatically adjusting resource values based on the supply and demand within the network.
- Identify and test additional well-developed theories of the social sciences to confirm validity within the context of specific online communities (e.g., can bonding social capital on Twitter be leveraged to lose weight?).
- Observe and track socially connected communities that are actively exchanging social resources. Due to the lack of availability at the time of writing, this could not be performed. However, we would expect that this will be something that future researchers will have access to. Moreover, the ideas presented in this dissertation would likely be invaluable towards the creation of such a system.

References

M.A. Belliveau, C.A. III O'Reilly, and J.B. Wade. Social Capital at the Top: Effects of Social Similarity and Status on CEO Compensation. *Academy of Management Journal*, 39(6): 1568–1593, 1996.

Bonnie H. Erickson. Good Networks and Good Jobs: The Value of Social Capital to Employers and Employees. In Nan Lin, Karen S. Cook, and Ronald S. Burt, editors, *Social Capital: Theory and Research*, chapter 6, pages 127–158. Aldine Transaction, 2004.

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.

A.W. Gouldner. The Norm of Reciprocity: A Preliminary Statement. *American Sociological Review*, 25(2):161–178, 1960. ISSN 0003-1224.

David M. Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Alexander Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational social science. *Science*, 323(5915):721–723, 2009.

Nan Lin. *Social Capital: A Theory of Social Structure and Action*. NY: Cambridge University Press, 2001.

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

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

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

Neil Savage. Twitter as Medium and Message. *Communications of the ACM*, 54(3):18–20, 2011.

Robert Sugden. Reciprocity: The Supply of Public Goods Through Voluntary Contributions. *The Economic Journal*, 94(376):772–787, 1984. ISSN 0013-0133.

J.W. van Deth. Measuring Social Capital. In D. Castiglione, J.W. van Deth, and G. Wolleb, editors, *The Handbook of Social Capital*, pages 150–176. Oxford University Press, 2008.