The Power of Indirect Ties in Friend-to-Friend Storage Systems

Xiang Zuo Jeremy Blackburn
University of South Florida
Tampa, Florida USA
{xiangzuo, jhblackb}@mail.usf.edu

Nicolas Kourtellis
Yahoo Labs
Barcelona, Spain
kourtell@yahoo-inc.com

John Skvoretz
University of South Florida
Tampa, Florida USA
jskvoretz@usf.edu

Adriana Iamnitchi
University of South Florida
Tampa, Florida USA
anda@cse.usf.edu

***Abstract*—Friend-to-Friend (F2F) storage systems were shown to suffer from two significant limitations. First, users with a small set of friends are penalized by lack of available storage for their needs, while users with many friends get overloaded with resource requests. Second, friends are typically in close geographical proximity to each other, and thus their online times are synchronized, leading to low data availability when they are offline. This paper addresses these concerns by expanding the set of storage resources while still using a measure of social incentives. It proposes an indirect tie measurement to compute the social strength between possibly distant nodes in a social network. Using datasets from co-authorship networks and a video gaming community, we show that the social strength-based mechanism more than doubles the set of storage candidates motivated by social incentives, invites socially low connected users to contribute more resources and improves data availability by up to 6.5 times. We also show that our method complemented with simple load balancing strategies significantly improves peer engagement and workload distribution in F2F storage systems.**

I. Introduction

Cloud storage services such as Dropbox, Amazon S3, and CrashPlan are attractive for their relatively low cost and high quality-of-service guarantees. The problem, however, with all centrally managed services is the single point of failure. Furthermore, these services are subjected to uncomfortable scrutiny by the government as recent events showed. For these reasons, the alternative of Friend-to-Friend (F2F) storage systems [1] where users leverage social incentives to get access to the available storage resources of their friends, becomes more appealing. F2F storage systems allow users to share resources with their neighbors, according to their existing social relationships, to solve some enduring problems in decentralized systems, such as privacy disclosure and incentives for cooperation. For example, Friendstore [2] demonstrated that the social fabric provided incentives to supply reliable backup and resulted in a system with less churn which, in turn, reduced the cost of data maintenance.

The standing challenge with F2F storage systems is that they typically provide poor quality of service. The main reasons that prevent F2F storage systems to achieve at least the QoS of peer-to-peer systems are the small size of the friendsets of most users [2] and the synchronized online availability patterns of friends [3]. Immediate consequences are low data availability [4] and low storage utilization [5]. One solution to mitigate this poor QoS is to use cloud storage as a fall-back alternative to ensure better data availability [6].

We propose a solution that addresses the above problems of F2F storage systems by extending the group of storage candidates to socially incentivized n-hop-away users ($n \geq 2$). This approach is motivated by the observation that, in real life, favors are often made to friends of friends, depending on the strength of the relationships involved. In fact, some of a user's indirect relations can be more socially "close" than some of his or her directly connected friends. At the same time, these indirect relations contribute in ways the direct friends cannot: with new information and extra resources or, specific to storage sharing, by being online at different times and thus potentially increasing service availability.

These observations have led us to propose a method that recruits socially distant nodes as data placement candidates based on the social strength between the user and those nodes' owners [7] (Section II). The computation of this metric is fully localized and adaptive to the social neighborhood of each user. We show experimentally using three real networks in Section III and IV that this approach (1) provides a larger candidate set to most users; (2) can significantly improve data availability; and (3) makes better use of existing resources in the system by better engaging the poor connected users.

II. Social Strength Definition

We define a metric that quantifies the strength of a social connection between indirectly connected nodes in a social graph. The definition of such a metric is supported by the following observations from sociology.

O1: The strength of a direct social relationship is related to the amount of interactions [8]: the more frequently persons interact with one another, the more likely they will form strong relationships. Moreover, interactions among OSN users were shown to represent more meaningful relations than just declared relationships [9]. Therefore, in the quantification of an *indirect* social tie, we rely on a numerical representation of the strength of a *direct* social tie that can be expressed as number of interactions, number of shared interests, or other recordable outcomes, depending on the semantics of the relationship.

O2: The strength of an indirect tie decreases with the length of the shortest path between the two individuals. This has been quantitatively observed by Friedkin [10], who concluded that people's awareness of others' performance decreases beyond 2 hops. The three degrees of influence theory, proposed by Christakis et al. [11], states that social influence does not end with people who are directly connected but also continues to 2- and 3-hop relationships, albeit with diminishing returns. This theory has held true in a variety of social networks examined [12], [13]. In accordance with these observations, the social strength metric we propose focuses on 2- and 3-hop relationships with a decreasing value as a function of distance.

O3: Multiple types of social interactions result into a stronger (direct) relationship than only one type of interaction [14]. Furthermore, sociology studies [10] observed that the relationship strength of indirectly connected individuals greatly depends on the number of different direct or indirect paths connecting them. Therefore, we consider the strength of multiple shortest paths in our definition of the strength of an indirect social tie.

O4: Typically, social ties between individuals are asymmetrically reciprocal [15]. We preserve this asymmetry in quantifying indirect ties, such that Alice and Charlie, indirectly connected via Bob, are entitled to have different views about their indirect tie.

Therefore, expanding the definition from [16] to quantify the social strength of an indirect social tie between users i and m, we consider relationships at n (n = 2 or n = 3) social hops, where n is the shortest path length between i and m. We assume a weighted interaction graph model that connects users with edges weighted based on the intensity of their direct social interactions. Assuming that $\mathcal{P}^n_{i,m}$ is the set of different shortest paths of length n joining two indirectly connected users i and m and $\mathcal{N}(p)$ is the set of nodes on the shortest path $p, p \in \mathcal{P}^n_{i,m}$, we define the social strength between i and m from i's perspective over an n-hop shortest path as:

$$SS_n(i,m) = 1 - \prod_{p \in \mathcal{P}^n_{i,m}} \left(1 - \frac{\min\limits_{j,\ldots,k \in \mathcal{N}(p)} [NW(i,j), \ldots, NW(k,m)]}{n}\right) \tag{1}$$

This definition uses the normalized direct social weight $NW(i,j)$ between two directly connected users i and j, defined as follows:

$$NW(i,j) = \frac{\sum_{\forall \lambda \in \Lambda_{i,j}} \omega(i,j,\lambda)}{\sum_{\forall k \in N_i} \sum_{\forall \lambda \in \Lambda_{i,k}} \omega(i,k,\lambda)} \tag{2}$$

Equation 2 calculates the strength of a direct relationship by considering all types of interactions $\lambda \in \Lambda$ between the users i and j such as phone calls, interactions in online games, and number of co-authored papers (observation O3). These interactions are normalized to the total amount of interactions of type λ that i has with other individuals. This approach ensures the asymmetry of social weight (observation O4) in two ways: first, it captures the cases where $\omega(i,j,\lambda) \neq \omega(j,i,\lambda)$ (such as in a phone call graph). Second, by normalizing to the number of interactions within one's own social circle (e.g., node i's neighborhood N_i), even in undirected social graphs, the relative weight of the mutual tie will be different from the perspective of each user.

III. Expanding Resource Peer Sets in F2F Systems

We propose a simple, local algorithm (Section III-A) that leverages social strength SS_n for expanding the resource sets in F2F systems while still using a measure of social incentives. We empirically evaluate the success of this algorithm with real datasets and discuss the results in Section III-B.

A. Friendset Expansion Algorithm

Some socially aware systems have explored indirect ties among users in the design of their systems. Some previous work directly involves all of a user's friends-of-friends (or even longer distant relationships) [17]. However, not every 2- or 3-hop distant friend has enough incentives, for example, for storing data, hosting computation tasks, or routing messages. Therefore, instead of directly involving all of a user's indirect ties within some radius, we use the quantitative power of the social strength metric, SS_n, to select socially "close" distant nodes, that is, indirect connections with comparable social strength with the user's direct (1-hop) friends.

TABLE I: Characteristics of social networks used in experiments. APL: average path length, CC: clustering coefficient, A: assortativity, D: diameter, EW: range of edge weights.

Networks	Nodes	Edges	APL	density	CC	A	D	EW
CA-I	348	595	6.1	0.0098	0.28	0.173	14	[1–52]
CA-II	1,127	6,690	3.4	0.0100	0.33	0.211	11	[1–127]
TF2	2,406	9,720	4.2	0.0034	0.21	0.028	12	[1–21,767]

The expansion algorithm follows two steps:

- For each user i, find the weakest direct social contact j such that $NW(i,j) = \min\limits_{k \in N_i} [NW(i,k)]$. Let this minimum normalized weight be referred to as θ_i.
- For each m of i's n-hop friends, if $SS_n(i,m) \geq \theta_i$, the user m is inserted in the candidate peer set of i. Intuitively, this ensures that the social strength between i and m, located at distance n in the social graph, is at least as strong as i's weakest direct tie.

We note that the algorithm expands each candidate set using a user-specific, thus local, threshold. Such local thresholds are needed in the distributed setting of a F2F system.

B. Expansion Results

Datasets: To evaluate *how much* the candidate set can be expanded, we implemented SS_n presented in Eq. 1, and measured the size of the candidate set selected based on the expansion algorithm on three online networks (Table I): an interactions network among friends in the online game Team Fortress 2 (TF2) [18], and two co-authorship networks from ArnetMiner [19]. The TF2 interacting friends network is composed of edges between players who had at least one in-game interaction while playing together and supplies each player's online/offline status that can be used for the data availability experiments presented later, in Section IV. In the co-authorships networks, nodes represent authors and are labeled with the author's affiliation, and edges exist if the two researchers co-authored at least one paper together and are weighted with the number of papers co-authored. We map each author's affiliation information to a timezone, which can be further used in simulating users' online/offline behavior in Section IV. From this dataset we extracted two largest connected components as our networks: *Co-authorship I (CA-I)* and *Co-authorship II (CA-II)*.

Results: For 2-hop expansion, 63.6% users in CA-II and 36.6% of players in TF2 expanded their candidate sets. Even in the sparser CA-I, 34.2% users augmented their friend sets. When considering the expansion brought in by 3-hop distant nodes m who satisfy the requirement that $SS_3(i,m) \geq \theta(i)$ the expansion is still taking place in all three networks: even in the sparser network CA-I, 10.6% users augment their friendsets and about 1% users have expanded their candidates with more than five friends. The denser network CA-II has more than 50% of users expanding their candidate sets, and TF2 has

27.2% (with the number of expanded 3-hop friends being 1,032). As expected, 3-hop augmentation is not as strong as 2 hops' since as the social distance increases, the social strength weakens. Thus, using social strength for recruiting peers indirectly connected in the social graph augments users' peer-sets and potentially solves problems caused by the limited number of friends in F2F systems.

IV. Improving Data Availability in F2F Systems

Expanding the candidate set is a necessary but insufficient solution for improving the performance of F2F systems. In particular, as shown in the context of F2F storage systems, F2F service availability depends on user online activity patterns [3], [4]. In addition, resource utilization was shown to be low and workload unevenly distributed.

In this section we show that a larger resource candidate set complemented with simple mechanisms for resource selection significantly improves both service availability and resource utilization in F2F storage systems. We stress that we do not propose a complete mechanism that improves the performance of F2F systems. Instead, we focus on exploring the potential of using social strength (via the expansion algorithm) in F2F solutions. Thus, Section IV-A shows that data availability significantly improves compared to "traditional" 1-hop F2F when using the expanded peer set algorithm with a greedy data placement algorithm [3]. Also, Section IV-B shows that a simple peer selection strategy can lead to a significant increase in peer engagement and resource utilization.

A. Online Storage Availability

In this section we simulate users' online presence and data placement to estimate file availability in F2F storage systems with service candidate sets augmented by social strength.

1) Online Presence Behavior: To estimate peer availability, we augment each network with online presence empirically deduced from real traces. For CA-I and CA-II, we fit a distribution to online presence information extracted from Skype traces presented in [3]. The distribution was applied to each author by shifting it to match the timezone of her affiliation. For the TF2 network, we use one month of recorded playing times (Figure 4 in [20]).

To determine whether the social strength selection mechanism improves the availability of storage resources, we measure the percentage of a node's selected candidates available throughout the day, by binning online presence into 1-hour time slots. We also map each user's affiliation to a timezone, then match the timezone to an hour of a day. If a user is online at some point during a time slot, we mark him as available for that time slot. For CA-I and CA-II networks, in each time slot, we randomly select users to be online but keep the same percentage of online users from the Skype trace. We repeat this random sampling process for multiple iterations to obtain stable results.

2) Data Placement: Replicating data across all friends allows a user to get maximum achievable data coverage but results in high costs for storing and transferring data to multiple copies, in particular for users with a large number of friends. Instead, we adopt the greedy heuristic data placement algorithm proposed in [4] to backup files with a subset of friends who can cover the maximum online time.

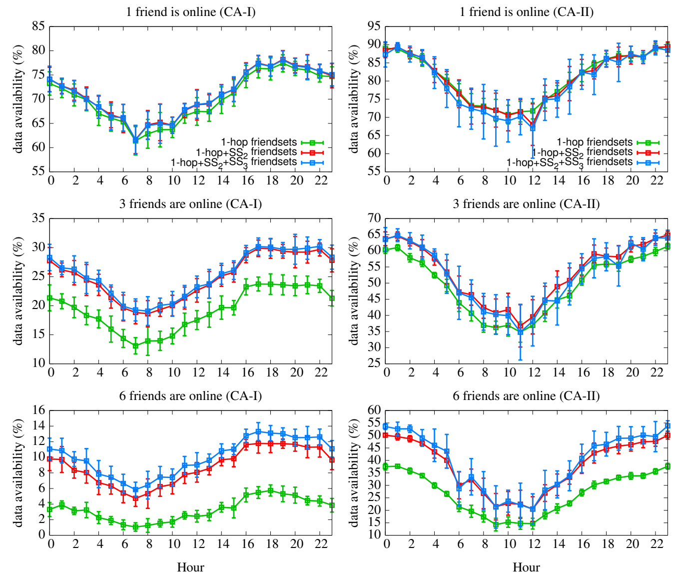

Fig. 1: Average percentage of data availability per hour for CA-I and CA-II.

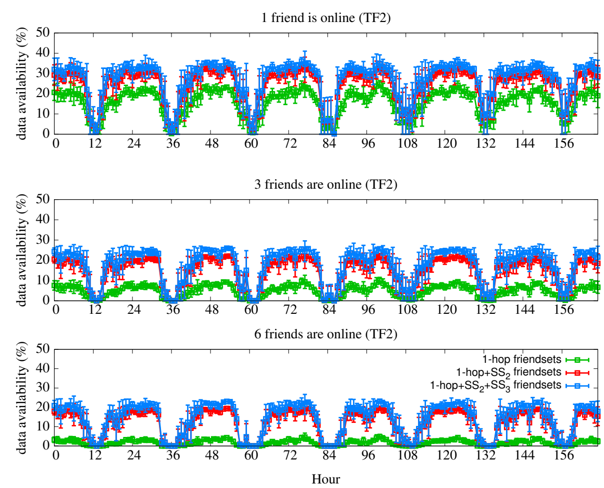

Fig. 2: Average percentage of data availability per hour of the week for TF2.

3) Data Availability: Some methods store files in a distributed fashion such as erasure codes that require k storage sites to be available for retrieving a file [21]. We vary the number of friends necessary for a node's storage needs to be met under such storage schemes. We measure the fraction of nodes who have enough candidates online to meet their needs when selected by either the pure F2F approach or the social strength mechanism. We compare three scenarios: 1) storage candidates selected only from direct social contacts; 2) storage candidates selected from the SS_n-based expanded candidate set, with $n = 2$ and 3) $n = 3$. Figures 1 and 2 plot the average fraction of users whose storage needs are met with the requirement that at least $k \in \{1, 3, 6\}$ candidates are online at a given time for the co-authorship networks and TF2, respectively. Error bars represent the 95% confidence interval on the average.

Using the expanded candidate set results in higher service availability. In particular, when 6 friends are needed to cooperate on completing a storage task, about 4 times higher data availability can be reached in CA-I, 1.6 times higher in CA-II and 6.5 times higher in TF2. Further, the social strength mechanism does not degrade as quickly as the 1-hop selection when increasing the number of friends that are required to be online simultaneously. We also see that for sparse networks such as CA-I, social strength over larger distance n improves data availability, especially when larger number of friends are required to be online simultaneously.

Finally, CA-II shows higher levels of availability than CA-I under the same conditions. This is likely because CA-II has more users with larger expanded candidate sets under the social strength mechanism than CA-I. Moreover, we note that CA-I shows better performance than TF2 under the same requirements. In the scenario that requires at least one friend online, 73% of users in CA-I have candidates available at midnight, compared to only 20% of TF2 users. One explanation could be the limited number of concurrent players the gaming server supports (at most 32 simultaneous players). Another explanation is that CA-I users are spread out over multiple timezones, while most of the TF2 users are geographically close to the server to minimize latency, and thus are time synchronized in their gaming patterns.

To conclude, using three datasets, we show that the social strength-based algorithm increases data availability by up to 6.5 times compared to the pure F2F approach.

B. *Peer Engagement & Workload Distribution*

In F2F systems the load placed on a peer is directly related to the peer's degree [20], i.e., nodes with high degree are the ones most often chosen while resources from low degree nodes are almost never requested. As a result, F2F systems suffer from low resource utilization and imbalanced workload distribution.

Expanding the candidate set shifts some of the load to low-degree peers, as they will be included in the candidate set of more peers, but does the same to high-degree peers, that will be even more loaded. Although the candidate sets of high-degree users expand more than those of low-degree users, a modified selection strategy could improve the utilization rate. One straightforward candidate selection strategy is giving high priority to low-degree users at the initial phase. Then, if all of a peer's friends have already contributed, friends with the lowest load will be chosen. We name this candidate selection scheme *Low Degree First then Low Workloads First* (LDLW).

Workload distribution and resource utilization in distributed systems depend on many factors, such as user requests, resource distribution, network characteristics, etc. To isolate the effect of social strength-expanded candidate set from other factors, we assume the following experiments are in a homogeneous environment: all peers issue the same number of resource requests, bandwidth and storage are unlimited, and files are equal in size. In the following we refer to *peer engagement* as the fraction of contributing peers in the system. We measure peer engagement in simple F2F systems and in expanded F2F systems in the same networks as above, when we vary the number of fragments ("friends" in F2F systems) needed for replication with erasure code [22] and under two load allocation schemes: random and LDLW.

1) Peer Engagement: Figure 3 plots peer engagement across all three networks in our experiments. The general trend is that the traditional F2F system has lower peer engagement

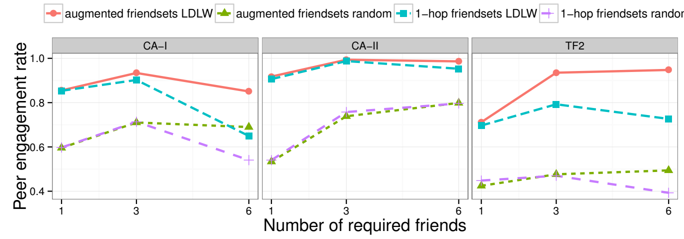

Fig. 3: Peer engagement of general F2F systems vs. F2F augmented by social strength. The requests are implemented by the random and LDLW schemes, respectively.

than the augmented F2F under the random load allocation, especially when the number of required peers for storage availability increases (e.g., CA-I and TF2). The augmented F2F solution implemented with the LDLW scheme strikingly improves peer engagement, especially when the number of backup friends is 6: CA-I's utilization rate jumps from 71.7% in 1 hop to 93.4%, CA-II achieves the biggest value at 99.3%, and TF2 shows the largest increment that raises from traditional F2F's 39.2% to augmented F2F's 94.8%.

There are two reasons behind this phenomenon. First, users' expanded friend sets supply more opportunities for users to donate their resources. In particular, when more friends are needed, some low-degree users in the traditional F2F scenario are punished by lack of resources due to few directly connected friends. The second is related to specifics of our test networks. CA-II has more users with expanded friend sets and TF2 has a larger friends expansion rate. Therefore, after friend sets expansion, these two networks involve more peers supplying backup service and ultimately lead to higher peer engagement.

The LDLW scheme achieves a better utilization rate than the randomized scheme. Additionally, if the size of a user's directly connected friends is large enough to satisfy their service demands, the traditional and the augmented F2F systems demonstrate similar resource utilization rates. However, when more peers are simultaneously needed, the augmented F2F implemented with LDLW scheme shows higher peer engagement.

To conclude, by enlarging peers' friend sets, the overall peer engagement increases compared to traditional F2F systems and leads to increased resource utilization. In addition, the simple LDLW candidate selection scheme improves peer engagement even more.

2) Workload Distribution: Another way to look at the benefits of our approach is to understand how a user's position in the network affects the workload of her associated peer. This is important because many social networks display a heavy-tailed degree distribution. That is, a few nodes have a high degree while many have a low degree. We address this by placing users into 5 bins representing the percentile rank of their degree centrality.

Figure 4 plots average workload (number of storage requests) for peers in each bin for the aforementioned candidate selection methods. As we increase the number of required friends (the columns in the figure) the workload increases across all groups of users for the random selection mechanism. With the random selection schemes, we see a skewed load distribution due to the skewed degree distribution. Our

Fig. 4: Average workload as a function of node-degree rank in three scenarios.

proposed LDLW selection scheme, however, is effective at evenly distributing loads among all degree bins. For all three networks, given the same number of required friends, the load across bins is comparable, with the largest improvement occurring in the CA-II network: the LDLW solution doubles the average workload of the lowest degree bin compared to the 1-hop random selection scheme and has a 1.7 times improvement when compared to random selection from social strength augmented candidate sets. Increasing the average contribution of low-degree users thus helps balance load distribution by shifting workloads away from high-degree users.

V. Summary and Discussions

In this paper, we propose a metric that quantifies the strength of indirect ties and use it to extend the candidate sets for users in a F2F system while preserving the social incentives that are the very motivation of F2F systems. Specifically, by considering both the intensity of interactions and the number of connected paths, we extend the candidate set of a user. This expanded candidate set includes not only her direct contacts in the social network, but also peers located n-hops away with whom she has a social strength higher than that of her weakest direct contact. This approach allows users access to many more resources, involves the lonely users to contribute resources to the system, and significantly improves the data availability by expanding the geographical coverage of the candidate set.

This work is the first step in understanding and further applying the indirect tie metric in solving problems in friends-only systems. The cost of using the social strength metric is computational complexity, especially in large, dense networks. Building a practical framework that enables fast computation of social strength on dense networks is left for future work.

Acknowledgment

This research was supported by the National Science Foundation under Grant No. CNS 0952420. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] J. Li and F. Dabek, “F2F: reliable storage in open networks,” in *Proceedings of the 4th International Workshop on Peer-to-Peer Systems*, ser. (IPTPS), 2006.

[2] D. N. Tran, F. Chiang, and J. Li, “Friendstore: cooperative online backup using trusted nodes,” in *Proceedings of the 1st ACM Workshop on Social Network Systems*, 2008, pp. 37–42.

[3] G. T. Raúl, M. Sánchez Artigas, and P. García López, “Analysis of Data Availability in F2F Storage Systems: When Correlations Matter,” in *IEEE 12th International Conference on Peer-to-Peer Computing*, 2012, pp. 225–236.

[4] R. Sharma, A. Datta, M. D. Amico, and P. Michiardi, “An empirical study of availability in Friend-to-Friend storage systems,” in *IEEE 11th International Conference on Peer-to-Peer Computing*, 2011, pp. 348–351.

[5] N. Tran, F. Chiang, and J. Li, “Efficient cooperative backup with decentralized trust management,” *ACM Transactions on Storage (TOS)*, vol. 8, no. 8, pp. 1–25, September 2012.

[6] A. Moreno-Martinez, R. Gracia-Tinedo, M. Sanchez-Artigas, and P. Garcia-Lopez, “Friendbox: A cloudified F2F storage application,” in *IEEE 12th International Conference on Peer-to-Peer Computing*, 2012, pp. 75–76.

[7] P. Anderson, N. Koutellis, J. Finnis, and A. Iamnitchi, “On managing social data for enabling socially-aware applications and services,” in *3rd ACM Workshop in Social Network Systems*, 2010.

[8] M. S. Granovetter, “The strength of weak ties,” *American Journal of Sociology*, vol. 78, no. 6, 1973.

[9] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao, “User interactions in social networks and their implications,” in *Proceedings of the 4th ACM European conference on Computer systems*, 2009.

[10] N. E. Friedkin, “Horizons of observability and limits of informal control in organizations,” *Social Forces*, vol. 62, no. 6, pp. 54–77, March 1983.

[11] N. A. Christakis and J. H. Fowler, *Connected: The surprising power of our social networks and how they shape our lives.* Hachette Digital, Inc., 2009.

[12] ——, “The spread of obesity in a large social network over 32 years,” *New England Journal of Medicine*, vol. 357, no. 4, pp. 370–379, 2007.

[13] J. H. Fowler, N. A. Christakis, and D. Roux, “Dynamic spread of happiness in a large social network: longitudinal analysis of the framingham heart study social network,” *BMJ: British Medical Journal*, pp. 23–27, 2009.

[14] L. Pappalardo, G. Rossetti, and D. Pedreschi, “”How well do we know each other?” detecting tie strength in multidimensional social networks,” in *ASONAM*, 2012, pp. 1040–1045.

[15] D. J. Brass, K. D. Butterfield, and B. C. Skaggs, “Relationships and unethical behavior: A social network perspective,” *Academy of Management Review*, vol. 23, no. 1, pp. 14–31, 1998.

[16] N. Kourtellis, “On the design of socially-aware distributed systems,” Ph.D. dissertation, University of South Florida, 2012.

[17] J. A. Pouwelse, P. Garbacki, J. Wang, A. Bakker, J. Yang, A. Iosup, D. H. Epema, M. Reinders, M. R. V. Steen, and H. J. Sips, “Tribler: a social-based peer-to-peer system,” *Concurrency and Computation: Practice and Experience*, vol. 20, no. 2, pp. 127–138, 2008.

[18] J. Blackburn and A. Iamnitchi, “Relationships under the microscope with interaction-backed social networks,” in *1st International Conference on Internet Science*, 2013.

[19] J. Tang, J. Sun, C. Wang, and Z. Yang, “Social influence analysis in large-scale networks,” in *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2009.

[20] X. Zuo, J. Blackburn, N. Kourtellis, and A. Iamnitchi, “The power of indirect social ties,” University of South Florida, http://arxiv.org/abs/1401.4234, Tech. Rep., 2014.

[21] W. K. Lin, D. M. Chiu, and Y. B. Lee, “Erasure code replication revisited,” in *IEEE 4th International Conference on Peer-to-Peer Computing*, 2004, pp. 90–97.

[22] H. Weatherspoon and J. D. Kubiatowicz, “Erasure coding vs. replication: A quantitative comparison,” in *The First International Workshop on Peer-to-Peer Systems (IPTPS)*, 2002, pp. 328–338.