Analysis of Fusing Online and Co-presence Social Networks

Susan Juan Pan, Daniel J. Boston, and Cristian Borcea
Computer Science Department, New Jersey Institute of Technology
jp238@njit.edu, djb38@njit.edu, borcea@cs.njit.edu

***Abstract*—This paper explores how online social networks and co-presence social networks complement each other to form global, fused social relations. We collected Bluetooth-based co-presence data from mobile phones and Facebook social data from a shared set of 104 students. For improved analysis accuracy, we created weighted social graphs based on meeting frequency and duration for co-presence data, and based on wall writing and photo tagging for Facebook data. By analyzing the overall structural properties, we show the two networks represent two different levels of social engagement which complement each other. By fusing them together, the average path length and network diameter is shortened, and consequently the social connectivity increases significantly. By quantifying the contribution of each social network to the fused network in terms of node degree, edge weight, and community overlap, we discovered that the co-presence network improves social connectivity, while the online network brings greater cohesiveness to social communities.**

***Keywords*-Social computing, social network analysis, co-presence traces, smart phones.**

I. Introduction

Many pervasive computing applications, such as Foursquare, Brightkite, Loopt, and Google's Latitude, employ social information to personalize or customize their results. Commonly, this social networking information is either self-declared or collected from online interactions. The ubiquity of smart phones can improve this situation by collecting user co-presence information, which allows us to identify social ties grounded on real world interactions. A combination of online and offline social interactions can lead to a more complete and accurate picture of people's social ties.

Online social networks provide a relatively stable social graph. Users slowly add new relationships after an initial bootstrap time and rarely delete relationships from their profile [1]. However, due to the dynamics of each individual's behavior, such declared online friendships may not carry complete information on social ties (i.e, two friends who rarely interact online may see each other frequently in real life). Meanwhile, inferring social information from mobile devices alone is limited by the difficulty to differentiate simple co-presence from social interaction, especially in densely populated environments. Therefore, it makes sense to use both types of social information: online and co-presence.

Some systems, such as Prometheus [2], maintain these two separately and use them according to the application's needs. For example, online social information can be used to improve the results of search engines (e.g., Microsoft's Bing uses Facebook data), while co-presence information can be used to improve forwarding in delay-tolerant networking [3]. Other applications, on the other hand, may need both types of information. For instance, Quercia et al. [4] maintains the two networks separately, but utilizes both to help balance youngsters' social connections. Furthermore, by fusing the two types of information, new social relations can be discovered (e.g., friend-of-friend relations) and leveraged in applications such as friend recommendation systems.

Our focus is to explore the fusion of online and co-presence social networks. Specifically, the main questions addressed in this paper are:

- Do these two representations of an individual's social network just reinforce each other or do they capture different types of social ties?
- If they are different, does it make sense to fuse them?
- How can we quantify the benefits of this fusion?
- Can we measure the contribution of each source network to the fused network?

To answer these questions, we collected smart phone co-presence data using Bluetooth and online social data from Facebook for a shared set of 104 users. For improved analysis accuracy, we created weighted social graphs based on meeting frequency and duration for co-presence data, and based on wall writing and photo tagging for Facebook data. We then compared and analyzed the three graphs (online, co-presence, and fused) in terms of global structural properties and individual node/community similarities.

We discovered that the fused network is more connected and has larger sized communities. Its average path length is shortened, and its average node degree is increased by 80% compared to the online network. The results also show that the co-presence network contributes more strongly to the fused network in terms of node degree, weighted degree, and edge weight. In contrast, the online network contributes stronger community structures to the fusion.

The paper is organized as follows. Section II presents related work. Sections III and IV describe our methods of data collection and weighted graphs construction. The results and their analysis are presented in Section V. The paper concludes in Section VI.

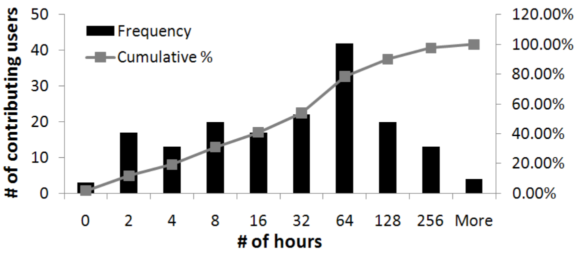

Figure 1. Histogram and cumulative distribution of total data hours collected by study participants

II. Related Work

To the best of our knowledge only one paper has tackled the fusion of on-line and co-presence networks [5]. The networks considered in this work are very sparse, and therefore have limited social information. Furthermore, the co-presence social ties are based on only one meeting; thus, they could be misleading. Finally, this work uses unweighted graphs. In contrast, our work contains highly accurate weighted social graphs with good network densities. Therefore, our results capture meaningful social similarities and differences between the online, co-presence, and fused networks.

Another type of research has focused on using co-presence data to predict friendship. The Reality Mining project [6] has developed behavioral characteristics of friendship for their user set by analyzing co-presence data in conjunction with location and phone logs. Cranshaw et al. [7] improved the prediction results by adding a location entropy feature. Our work is complementary to these projects as it focuses on analyzing the benefits and characteristics of fusing online and co-presence social networks.

Yet other studies focused on revealing the structure and role similarity and dissimilarity between co-presence and self-reported (or online) networks. Mtibaa et al. [8] showed that subjects generally spend more time with their friends, therefore concluding that the two graphs are similar. However, their experiment was performed at a conference over the course of a single day. As such, these results cannot be broadened to more than a contained event, where it is expected that subjects spend more time with their friends. Unlike this work, our results (on a larger user set: 104 vs. 27) demonstrate that the two networks, online and co-presence, are different.

III. Data Collection

We collected one month of Bluetooth co-presence data and Facebook friend lists for a set of 104 students at our university. The study took place on our medium size urban campus; 73% of our subjects were undergraduates, and 29% were women. All participants were volunteers and received monetary compensation. As well, they were representative of the various colleges and departments at NJIT.

The subjects installed a Facebook application to participate in a followup survey, and gave us permission to collect and analyze their friend lists, comments, and photo tags.

Similar to the Reality Mining traces [6], mobile phones were distributed to students, and a program quietly recorded the Bluetooth addresses of nearby devices and periodically transmitted them to a server. These periodic transmissions form a trace of interactions over time. This method of collecting co-presence data is non-intrusive and does not miss meetings as long as the phones are on.

It is possible, however, that some recorded meetings are just chance encounters without social significance (e.g., two students sitting at nearby tables in the cafeteria). To reduce the impact of such scenarios, we consider certain thresholds for meeting time and meeting frequency between two people (as shown in Section IV). Furthermore, ground-truth evidence from a different paper suggests that social groups can be detected with high accuracy using this method [9].

Given that our sample size (104 volunteers) was small compared to the university population (9000 students) and that many students are commuters, our trace data is relatively sparse. For example, about half the subjects collected less than 49 hours of data for the entire month (see Figure 1), and only 24% of the scans detected other Bluetooth devices in proximity. The typical user provided a few hours of data per day, especially during the week days.

IV. Social Graph Representation

In a social graph, $G = \langle V, E \rangle$, the vertices V are users and the edges E are social ties between users. For the online Facebook data, there is an edge between any pair of friends. For the Bluetooth co-presence data, there is an edge between two users who spent a certain amount of time and met with a certain frequency. The weight of Facebook ties indicate the number of interactions and the weight of Bluetooth ties exhibit the frequency and duration of co-presence.

A. Thresholds Selection for Co-presence Social Graph

In order to have accurate co-presence social data, we need to determine when to add an edge between two users. Very short and infrequent co-presence does not indicate the presence of a social tie. On the contrary, online friendship declarations need both users' direct involvement, strongly indicating the presence of a social tie. Thus, we keep the online social graph unchanged, but vary the co-presence graph to find thresholds that result in the least noise. Specifically, we need to select the right total meeting duration and meeting frequency between two users who are considered socially connected. If the threshold is too loose, then the co-presence graph may have too many social tie edges that are not important; if the threshold is too tight, then some important edges are lost. Hence, a good threshold is crucial to ensure the elimination of noisy data without the loss of true social ties.

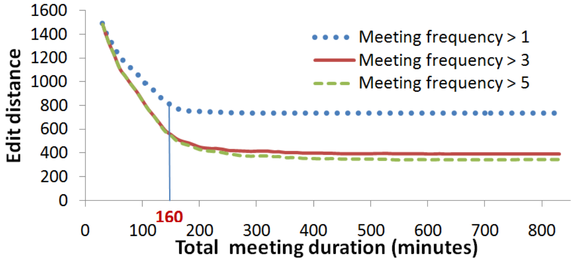

Figure 2. Edit distance for different meeting frequencies as function of total meeting duration

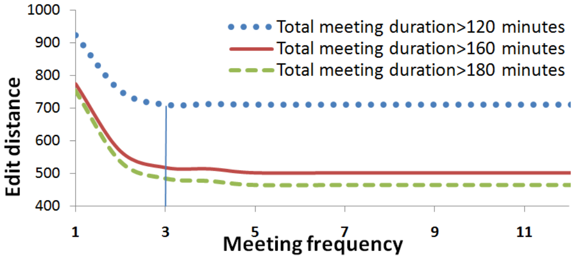

Figure 3. Edit distance for different total meeting durations as function of meeting frequency

To achieve this goal, we compute the Edit distance between online and co-presence networks. The Edit distance is the number of edit operations (add or delete) needed to change one graph into another. We use the adjacency matrices of the two networks to compute the Edit distance:

$$Online_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are Facebook friends} \\ 0 & \text{otherwise} \end{cases}$$

$$Co-presence_{ij} = \begin{cases} 1 & \text{if } T_{ij} \geq \alpha,\ F_{ij} \geq \beta \\ 0 & \text{otherwise} \end{cases}$$

T_{ij} is the total time users i and j spent together; F_{ij} is the total number of meetings in the encounter history. α and β are thresholds for meeting duration and meeting frequency that we vary during the analysis (α within [30min, 1800min] and β within [1, 10]).

We pick 160 minutes and 3 meetings as thresholds based on the results in Figures 2 and 3 that show the Edit distance remaining stable beyond these values.

B. Weight Computation

The weight of the online Facebook graph is the number of interactions between each pair of users. We consider a friendship request as an interaction. Thus, the minimum weight is 1. In our analysis, the interactions include both wall writing and photo tagging.

The weight of the Bluetooth co-presence social graph is obtained from two perspectives: the total meeting duration D and the meeting frequency F. We use both to capture different types of social interactions. Some friends prefer to meet longer but infrequently, while some prefer to see each other more but for shorter times.

Table I
MEAN, STANDARD DEVIATION AND MAXIMUM IN THE DATA SETS

	Max	Mean	Standard Dev.
Meeting Duration	220hr 2min	1hr 16min	7hr 34min
Meeting Frequency	51	2.2	3.7
Online Interaction	40	2	4

In order to make the co-presence and online social graphs comparable, we adjust the weights for the total meeting time and meeting frequency to be within the same range with the weights for online interactions. Hence, for each pair of users who meet for a total of D seconds and F times, the weight of the meeting duration is $Weight_d = \lceil (D \times 40)/MAX_d \rceil$ and the weight of the meeting frequency is $Weight_f = \lceil (F \times 40)/MAX_f \rceil$. Table I shows the maximum, mean, and standard deviation of meeting duration, frequency, and number of online interactions. We can see that the typical interactions are within reasonable expectations; we use the maximum for proper normalization.

We merge $Weight_d$ and $Weight_f$ to form the final co-presence weight: $Weight_{co-presence} = \lceil 0.5 \times Weight_f + 0.5 \times Weight_d \rceil$. While other ways to aggregate the two source networks are possible, we consider a simple average as the final weight of an edge in the fused network: $Weight_{fused} = \lceil 0.5 \times Weight_{co-presence} + 0.5 \times Weight_{online} \rceil$.

V. RESULTS AND ANALYSIS

Conceptually, the co-presence network and online network represent different levels of engagement in social relationships. The online social network services focus on building and reflecting virtual social relationships among people. It is explicitly self-declared, long-term, and allows for social interaction across space and time. The co-presence social network represents the dynamics of the individual behavior during a certain period. It is implicitly derived from the phone traces and represents face-to-face communication. The fused network is the combination of both. In this section, we assess the similarity and difference between the three networks in terms of both global and local parameters. We used JUNG [10] and igraph [11] to compute the results.

A. Structural Comparison of Global Network Parameters

Degree Distribution. Figure 4 shows that the degree distribution of the online social network follows closely a power-law distribution, which is typical for social networks [12]. However, the co-presence degree does not resemble a power-law distribution. This is due to the noise introduced by people's mobility, which can result in meetings with familiar strangers that do not translate into social

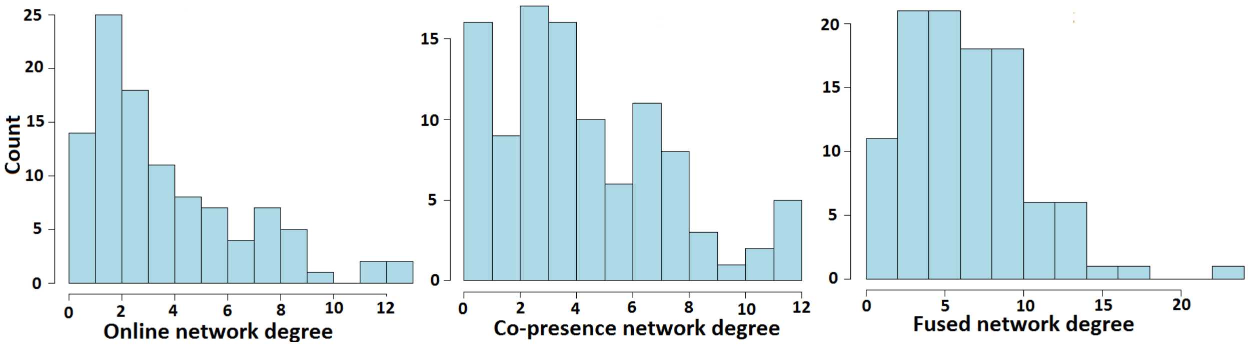

Figure 4. Comparison of the degree distribution among online, co-presence, and fused networks

Table II
GLOBAL STRUCTURAL MEASUREMENT

	Online	Co-presence	Fused	Weighted
Number of edges	165	196	310	N
Size of the largest connected component	63	84	98	N
Diameter of the largest connected component	7	8	7	N
Average weighted degree	9.54	13.73	11.63	Y
Average degree	3.17	3.77	5.96	N
Average weighted betweenness	49.1	90.13	94.83	Y
Average edge weight	3.02	3.64	1.95	Y
Average weighted path length between all reachable nodes	12.30	21.98	8.77	Y
Average weighted cluster coefficient	0.156	0.122	0.157	Y

relationships. As such, the fused network also does not closely follow a power-law distribution.

The rest of the global parameters are presented in Table II and discussed in the following.

Number of Edges and Average Degree. There are 51 shared edges between the online and co-presence networks, which is less than a third of the number of edges in each of the two networks. Therefore, the fused network has a much higher number of edges than each of the source networks. As such, the average unweighted degree in the fused network is significantly larger than those of the contributing networks.

Simple calculation shows the co-presence network contributes 27% more edges to the fused network structure than does the online network. This is due to the fact that the co-presence network captures a wider network of people without respect to the nature of their relationship. In contrast, the online network contributes less as it requires stronger user involvement (i.e., explicit request and confirmation) to establish a relationship.

Diameter and Average Path Length. Intuitively, the diameter measures the longest shortest path in the network. The weighted shortest path is the path with the greatest capacity of carrying information [13], [14]. The diameter and the average weighted shortest path are reduced in the fused network when compared to their values in the source networks. Therefore, people can become closer and more involved in each other's lives if the fused network is leveraged in social applications.

Average Weighted Degree and Average Edge Weight. The weighted node degree [15] is the sum of the weights of the edges attached to it. It measures the extent to which the user is involved in social activity. The average edge weight measures the extent that a pair of users interacts across the whole network. In our results, both the average weighted degree and average edge weight in the fused network are smaller than in the co-presence network but greater than in the online network. This demonstrates that people generally interact more in real life than online. However, online and face-to-face interactions are different types of social communication which complement each other: the person who is highly socially active online is not necessarily highly socially active in real life, leading to smaller values of the two parameters in the fused network. This is further confirmed by the extremely low weighted degree Spearman correlation [16] of 0.0588 and edge weight Spearman correlation of -0.0207 between the co-presence and online networks.

Betweenness Centrality. The betweenness centrality [17] counts the number of times a node occurs on the shortest path of other pairs of nodes. The weighted betweenness centrality [14] is the classic version measured on the weighted shortest paths. The average weighted betweenness score improves in the fused network because this network exhibit two types of social activity, thus, the social involvement of the node is increased.

The average weighted betweenness is higher in the co-presence network than in the online network. The explanation is that the average path length is longer in the co-presence network; therefore, a node has a greater chance of occurring on the shortest path between pairs of nodes. However, we note that in the fused network, the average weighted shortest path is only 8.77, but the average weighted betweenness score is still high. This is explained by the fused

Table III
Distance of node degree, weighted node degree and edge weight

	$Dist_{online,co-presence}$	$Dist_{online,fused}$	$Dist_{co-presence,fused}$
Weighted node degree	0.558	0.306	0.256
Node degree	0.399	0.305	0.225
Edge weight	0.560	0.324	0.295

network's larger connected component of size 98, indicating that more nodes can reach each other, which increases the chance of individual nodes to be on other nodes shortest paths.

Average Cluster Coefficient. The local cluster coefficient (also known as transitivity) is a measure of the extent to which nodes in a graph cluster together. It is the fraction of the number of present ties over the total number of possible ties between the node's neighbors. In the Wasserman and Faust weighted version [18], the contribution of each tri-set (visualized as a triangle) of nodes is weighted by a ratio of the average weight of the two adjacent edges of the triangle to the average weight of the node. We then calculate the average cluster coefficients for all 104 nodes to obtain a global point of view.

The average cluster coefficient in the online network is larger than in the co-presence network. It shows, in an online scenario, two people have a higher tendency of declaring friendship if they have friends in common. In the fused network, the average weighted cluster coefficient mainly benefits from the online social network, but it does not increase much.

B. Similarity among Local Network Parameters

Node Degree and Edge Weight. For this analysis, we compute the Euclidean distance of the degree vector (104 nodes) and shared edge vector (51 edges) among the three networks. In order to make the distance comparable among the three networks, the distance is normalized between 0 and 1 as follows: $Dist_{online,co-presence} =$

$$\frac{\sqrt{\sum_i |deg_{online}(i) - deg_{co-presence}(i)|^2}}{\sum_i deg_{online}(i) + \sum_i deg_{co-presence}(i)}$$

The similarity is simply the inverse of the distance. The results in Table III indicate that the co-presence network is more similar to the fused network. It demonstrates the co-presence network contributes more to the fusion network in terms of degree.

Community Overlapping Similarity. For this analysis, we compute the k-clique [19] overlapping clusters on the three networks separately. A k-clique overlapping community is the union of all k-cliques (complete subgraphs of size k) that can be reached from each other through a series of adjacent k-cliques (where adjacency means sharing k-1 nodes). Figure 5 shows the number of 4-clique communities in the three networks: the fused network has relatively larger size communities than the online and co-presence networks.

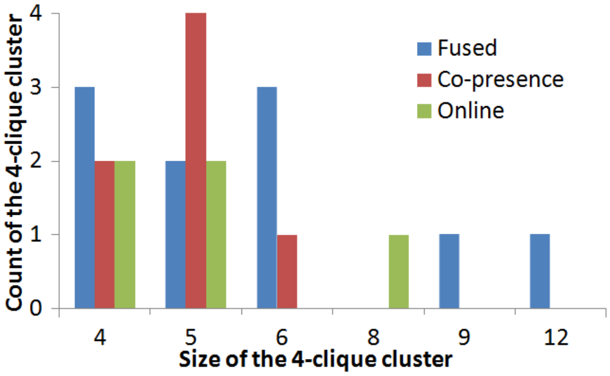

Figure 5. 4-clique cluster size in the online, co-presence, and fused networks

The links of one network are complemented by the links of the other, increasing the likelihood of forming cliques.

Since the k-clique communities are relatively large and very few share all members between any two networks, we decided to transform the one mode network into a two mode [13] network by constructing the community overlapping matrix. We use a $V \times V$ matrix (V is the number of vertices), denoted as H and defined below, for each of the three networks.

$$H_{i,j} = \begin{cases} m & \text{if edge } (i,j) \text{ shares } m \text{ communities} \\ 0 & \text{otherwise} \end{cases}$$

This matrix counts the number of shared communities that any pair of users belongs to. Intuitively, this metric measures the community overlap at the user pair level. We calculate the Edit distance between each pair of matrices to quantify the overall community overlap between the networks. The smaller the value of this metric, the more similar the communities are. Table IV shows that for weaker communities (k=3, k=4), the co-presence network is closer to the fusion network; however, for stronger communities (k=5), the online network contributes more to the fusion network. One possible explanation is that the online network tends to contain stronger social communities than the co-presence network, where recorded meetings do not always have a social meaning.

We notice the community overlapping distance between the co-presence and online network has typically the lowest value. This is due to the significant increase in the number of cliques after fusion, as shown in Table V. After merging the data, the connectivity of the network increases,

Table IV
DISTANCE USING COMMUNITY OVERLAPPING MATRIX

	k=3	k=4	k=5
$Dist_{online,fused}$	2561.0	142.5	26.5
$Dist_{co-presence,fused}$	2289.5	135.5	32.0
$Dist_{online,co-presence}$	894.5	119.0	30.5

Table V
NUMBER OF CLIQUES

online	co-presence	fused
45	51	101

leading to 50% more cliques. Consequently, the community overlapping between the fused network and the two source networks is lower than the overlapping between the two source networks.

VI. CONCLUSIONS

This paper has analyzed the relation between the online, co-presence, and fused (online + co-presence) social networks for the same set of users. The results demonstrate that the co-presence and online networks represent two different classes of social engagement that complement each other. Therefore, a fused network that incorporates both these networks makes sense for socially-aware pervasive applications that benefit from stronger social connections, but do not care about the specific types of these connections.

Most significant are applications such as friend recommendation and event scheduling systems that can benefit from a more complete understanding of friend-of-friend relationships. Similarly, discussion forums can utilize a fused network to more quickly identify larger and better groups of people to involve in conversations.

Finally, our study of the fused network has found that the online social network contributes to strengthening the community structure and lowering the average path length. The co-presence network, on the other hand, contributes more to increasing the network connectivity and communication strength. We believe that these conclusions are representative of other similar campus environments, but further generalizations can be made only after analysis of similar datasets for different environments.

ACKNOWLEDGMENT

This material is based upon work supported in part by the National Science Foundation under Grant Numbers CNS-0831753. 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 National Science Foundation.

REFERENCES

[1] J. Golbeck, "The dynamics of web-based social networks: Membership, relationships, and change," *First Monday*, vol. 12, no. 11, pp. 1–17, 2007.

[2] N. Kourtellis, J. Finnis, P. Anderson, J. Blackburn, C. Borcea, and A. Iamnitchi, "Prometheus: User-Controlled P2P Social Data Management for Socially-Aware Applications," in *Proceedings of the 11th ACM/IFIP/USENIX International Middleware Conference(Middleware 2010)*, Dec 2010, pp. 212–231.

[3] P. Hui, J. Crowcroft, and E. Yoneki, "Bubble rap: social-based forwarding in delay tolerant networks," in *Proceedings of the 9th ACM international symposium on Mobile ad hoc networking and computing*. ACM, 2008, pp. 241–250.

[4] D. Quercia, J. Ellis, and L. Capra, "Nurturing Social Networks Using Mobile Phones," *IEEE Pervasive Computing*, pp. 12–20, May 2010.

[5] V. Kostakos and J. Venkatanathan, "Making friends in life and online," in *IEEE International Conference on Social Computing*, Aug 2010, pp. 587–594.

[6] N. Eagle, A. Pentland, and D. Lazer, "Inferring Social Network Structure using Mobile Phone Data," *Proceedings of the National Academy of Sciences*, vol. 106, no. 36, pp. 15 274–15 278, 2009.

[7] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh, "Bridging the gap between physical location and online social networks," in *Ubicomp '10: Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, 2010, pp. 119–128.

[8] A. Mtibaa, A. Chaintreau, J. LeBrun, E. Oliver, A. Pietilainen, and C. Diot, "Are you moved by your social network application?" in *Proceedings of the first ACM Workshop on Online Social Networks*, 2008, pp. 67–72.

[9] S. Mardenfeld, D. Boston, S. Pan, Q. Jones, A. Iamntichi, and C. Borcea, "GDC: Group Discovery Using Co-location Traces," in *Proceddings of the 2nd IEEE Symposium on Social Computing Applications (SCA-10)*, Dec 2010, pp. 641–648.

[10] J. OMadadhain, D. Fisher, P. Smyth, S. White, and Y. Boey, "Analysis and visualization of network data using JUNG," *Journal of Statistical Software*, vol. 10, pp. 1–35, 2005.

[11] G. Csárdi and T. Nepusz, "The igraph software package for complex network research," *InterJournal*, vol. Complex Systems, p. 1695, 2006.

[12] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 2007, pp. 29–42.

[13] M. Newman, "Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality," *Physical review E*, vol. 64, no. 1, p. 016132, 2001.

[14] T. Opsahl, F. Agneessens, and J. Skvoretz, "Node centrality in weighted networks: Generalizing degree and shortest paths," *Social Networks*, vol. 32, no. 3, pp. 245–251, 2010.

[15] M. Newman, "Analysis of weighted networks," *Physical Review E*, vol. 70, no. 5, p. 056131, 2004.

[16] J. Maritz, *Distribution-free statistical methods*. Chapman & Hall/CRC, 1995.

[17] L. Freeman, "Centrality in social networks: conceptual clarification," *Social networks*, vol. 1, no. 3, pp. 215–239, 1979.

[18] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 11, pp. 3747–3752, 2004.

[19] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 9, pp. 814–818, June 2005.