The Importance of Being Placefriends: Discovering Location-focused Online Communities

Chloë Brown Vincenzo Nicosia
Salvatore Scellato Anastasios Noulas Cecilia Mascolo
Computer Laboratory
University of Cambridge
Firstname.Lastname@cl.cam.ac.uk

ABSTRACT

Discovering groups of online friends who go to the same physical places has numerous potential applications including privacy management, friend recommendation, and contact grouping as in Google+ circles. Until recently, little information was available about places visited by users of online social networking services, so community detection on the social graph could not take this into account. With the rise of services such as Foursquare, Gowalla, and Facebook Places, where users *check in* to named venues and share their location with their friends, we now have the right data to make this possible. In this work, we propose a way to extract place-focused communities from the social graph by annotating its edges with check-in information. Using traces from two online social networks with location sharing, we show that we can extract groups of friends who meet face-to-face, with many possible benefits for online social services.

Categories and Subject Descriptors

G.2.2 [**Graph theory**]: Graph algorithms, graph labeling; H.3.5 [**Online information services**]: Web-based services

Keywords

Online social networks, location-based services, community detection

1. INTRODUCTION

Physical places have always been important to social communities, with people meeting and forming friendships in locations where shared activities take place [7]. Although people may communicate online regardless of their location, OSN users do connect with friends they meet in person [4]. It is therefore reasonable to suppose that OSNs contain *place-focused* communities: groups of online friends who go to the same places. Location is becoming increasingly integrated into online social networks (OSNs), from incidental features such as Facebook Places, to explicitly location-based services such as Foursquare and Gowalla. As a result, data is available about the places users visit, enabling a level of analysis that has until recently been impossible. Communities in OSNs have been well-studied [1, 14], but without location data community detection on the social graphs can consider only the network topology [8, 17], and may fail to isolate groups of users who visit common sets of places.

Being able to find place-focused communities has many potential applications. Research examining online and offline social networks has found that shared locations between users are a positive predictor of social ties [3, 5, 6, 9, 18]. Studies by Pan et al. [16] and by Kostakos and Venkatanathan [11] compare users' Facebook networks and Bluetooth contacts of mobile devices, observing that the fused online and offline network is denser than either of the separate networks. More ties exist between the same set of users than in either network alone, which confirms that location gives additional information about social links and could help friend recommendation for online social services.

A further example application is in privacy; community detection has been proposed as a way to sort a user's OSN contacts into groups to aid privacy management. Jones and O'Neill [10] study how users create sets of their Facebook friends for selective privacy settings, and find that geographic location is a widely used criterion. They demonstrate that a network clustering algorithm can approximate these groups, with 33.8 to 76.1% of contacts receiving the correct settings. The identification of location-based communities could therefore help to improve OSN privacy controls. Automatic grouping of friends based on factors such as location could also more generally help users to manage the simultaneous existence of multiple logical groups of contacts [12], like the *circles* seen in Google+.

In this work, we propose a means to reveal place-focused communities in OSNs with location sharing, by annotating the social graph with information about the places users visit. We experiment using two large-scale network traces, and show that *community detection using only the social graph topology may fail to reveal groups of friends who visit the same places*. We find that *we can extract groups of users connected not only by social ties, but also by common places.* The potential implications of our findings are manifold: as the offline and the online realms converge, better services and applications can be designed, focusing on users who visit the same physical places in their daily lives.

2. PLACE-FOCUSED COMMUNITIES

We now describe in detail our proposed means of finding place-focused communities by annotating the social graph with information about the places where users go.

WOSN'12, August 17, 2012, Helsinki, Finland.

2.1 Finding place-focused communities

Community structure is an important feature of many networked systems, so much research has focused on devising methods for dividing large graphs into meaningful communities of nodes [8]. However, these methods generally use only the network topology, not application-specific information such as user location. This makes it difficult to use these methods to find place-focused communities. Furthermore, many algorithms insist that every node is assigned to a community, which may pose a problem when specifically seeking place-focused communities: some users may not belong to such groups, and should therefore be excluded.

We propose to address these difficulties by using information about places users visit to annotate the social graph, assigning to every edge a weight derived from location information. The aim is to give higher weights to edges denoting friendships for which physical places are important. Edges that do not have high enough weights to be important to a place-focused community can then be removed, and users who do not belong to place-focused communities and are therefore left with no ties can be excluded.

2.1.1 Notation

We assume a social network consisting of a set of N users $V = \{u_1, \ldots, u_N\}$ and the set E of ties between them, containing e_{ij} when u_i and u_j are friends. We represent the social network as an unweighted, undirected graph $G(V, E)$.

We also assume place data in the form of *check-ins* as used by online location-based social services such as Foursquare, Gowalla, and Facebook Places. Users *check in* by indicating that they are at a named location, and notify their friends. We represent the set of L places where users have checked in by $M = \{m_1, \ldots, m_L\}$, and we write as c_{ij} the number of check-ins that user u_i has made to place m_j. U_j denotes the set of users who have checked in to place m_j, and M_i represents the set of places where user u_i has checked in.

2.2 Method

Given the social network $G(V, E)$, the set of places M, and the associated user check-ins, we define an *annotation function* $f : E \to \mathbb{R}_{\geq 0}$. f takes an edge e_{ij} from the social graph, and assigns to it a weight $f(e_{ij})$ derived from the check-in information about u_i and u_j. Detection of place-focused communities is then performed as follows:

1. Assign to each edge e_{ij} in E a weight $f(e_{ij})$ based on the check-ins of u_i and u_j.
2. Remove from the graph all edges with weight lower than a threshold $t > 0$.
3. Remove from the graph all nodes with no incident edges.
4. Apply a standard community detection algorithm.

2.2.1 Annotation functions

We experimented with several different definitions of f:

- `binary`: This results in the subgraph of the unweighted graph G with only those edges e_{ij} where u_i and u_j have checked in to at least one of the same places. We call users who have a place in common *placefriends*:

$$f_{binary}(e_{ij}) = \begin{cases} 1 & \text{if } |M_i \cap M_j| > 0 \\ 0 & \text{otherwise} \end{cases}$$

  Note that `binary` forces use of the threshold $t = 1$, as use of any higher value would remove all edges in the thresholding step, where all edges with weight lower than t are eliminated.

- `places`: The number of places where u_i and u_j have both checked in:

$$f_{places}(e_{ij}) = |M_i \cap M_j|$$

- `checkins`: The sum, over all of the places that u_i and u_j share, of the lower of the numbers of check-ins that u_i and u_j have made to each place:

$$f_{checkins}(e_{ij}) = \sum_{m_p \in (M_i \cap M_j)} \min(c_{ip}, c_{jp})$$

  This aims to capture the extent to which users visit the same places, without giving undue weight to edges where one user visits a place many times and the other very few times.

- `ratio`: We compute the ratio of total check-ins at a place to the number of users who have checked in there. We use the maximum value over the places that u_i and u_j share:

$$f_{ratio}(e_{ij}) = \max_{m_p \in (M_i \cap M_j)} \left(\frac{C_p}{|U_p|} \right)$$

  where C_p is the total number of check-ins users in V made to place m_p. Places where many people go infrequently, such as an airport, will give low values and are not likely to indicate important place-based friendship ties. Places with high values are visited by a few users many times, for example, somebody's house, and may be more important. Thus, we weight edges where users share such places more highly.

2.2.2 Thresholding

The thresholding step aims to remove users who do not belong to sufficiently place-focused communities. After the annotation step, we eliminate edges with weights lower than a threshold value t. We then remove from the graph any users left with no ties. Without thresholding, these users will be assigned to communities by algorithms that insist that every node must be placed in a community, and decrease the place-focus of resulting communities. Choice of t is discussed in section 3.3.3.

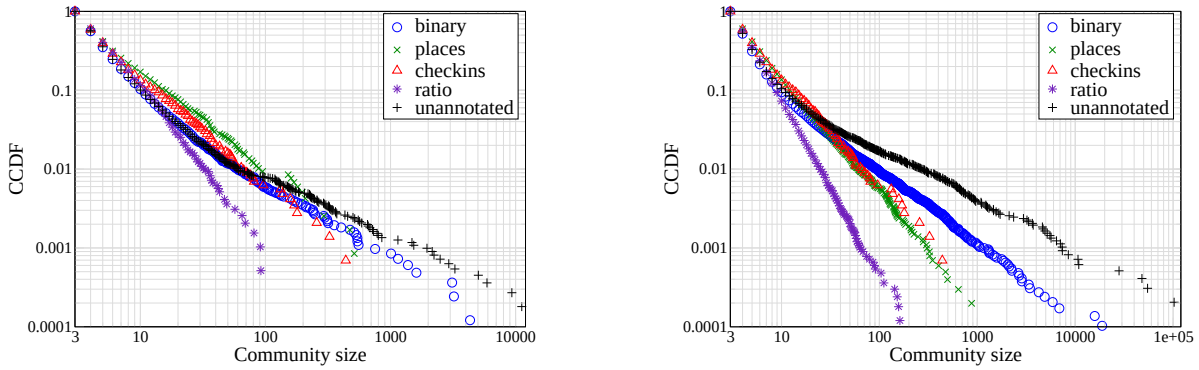
3. RESULTS AND EVALUATION

We now evaluate the effectiveness of our approach in finding groups of online friends who also go to the same physical places. All our definitions of the annotation function f, described in section 2.2.1, remove edges e_{ij} from the social graph where the users u_i and u_j have no places in common. Therefore, we take as given that the communities yielded by these graphs will contain users who share common places, as we aim to find. To check that this sharing of places is potentially meaningful, we specifically consider whether or not their members have been colocated, so that friends visit the same places *together*. Note that colocation is not the same as simply having visited the same places, which is the only information we use to weight edges before community detection.

In this section, we define two measures of community colocation. We then describe the datasets and present our results. We confirm that we can successfully extract place-focused communities from the annotated graphs that are not revealed by community detection on the unannotated graph. In particular, we find that not only do these users visit common sets of places, they also visit them *at*

Dataset	N	K	N_{GC}	$\langle k \rangle$	$\langle c \rangle$	L	C
Gowalla	165,051	765,872	157,622	9.28	0.23	1,541,951	13,561,773
Twitter	663,198	11,959,895	657,553	36.1	0.21	4,274,022	34,037,471

Table 1: Properties of the datasets: number of nodes N and edges K in the social network, number of nodes in the giant connected component N_{GC}, mean node degree $\langle k \rangle$, mean clustering coefficient $\langle c \rangle$, total number of places L and total number of check-ins C.



(a) Gowalla

(b) Twitter

Figure 1: CCDFs of community sizes using each of the annotation functions and the unannotated graph

the same time as other group members, suggesting that the groups do represent place-focused communities of friends who also meet in the offline world.

3.1 Colocation measures

We verify the place-focus of communities by examining whether their members have been colocated. We argue that if users in a group have declared ties in the social network, then their being in the same places at the same time is meaningful, and can be considered confirmation that the grouping represents a place-focused community.

We consider users to have been colocated if they have checked in to the same place within one hour of each other. We define two measures involving community members' colocation: *colocation density*, and *colocation fraction*. Let $\mathbf{M}$ be the *colocation matrix* such that the ij^{th} entry $\mathbf{M}_{ij}$ is 1 if users u_i and u_j have been colocated and 0 otherwise. Let $\mathbf{A}$ be the adjacency matrix of the unannotated social graph G so that $\mathbf{A}_{ij}$ is 1 if the users u_i and u_j have a tie e_{ij} in E and 0 otherwise.

3.1.1 Colocation density

The *colocation density* of a community C is given by:

$$\frac{\sum_{u_i,u_j \in C} \mathbf{M}_{ij}}{|C|(|C|-1)}$$

That is, the number of pairs who have been colocated, as a fraction of the possible number of pairs. This indicates whether a group is a place-focused community in the offline world, having higher values when higher proportions of community members meet face-to-face.

3.1.2 Colocation fraction

Colocation density is useful for determining whether a community represents a group who tend to meet each other, but its value will be affected by the size of the community: it is less likely that a group of 20 people will all have been pairwise colocated than it is that a group of 5 would have been. We thus define *colocation fraction*, which controls for the size of the community by considering the *social* links that are present, and which of these pairs of friends are also colocated. The *colocation fraction* of a community C is defined:

$$\frac{\sum_{u_i,u_j \in C} \mathbf{M}_{ij} \cdot \mathbf{A}_{ij}}{\sum_{u_i,u_j \in C} \mathbf{A}_{ij}}$$

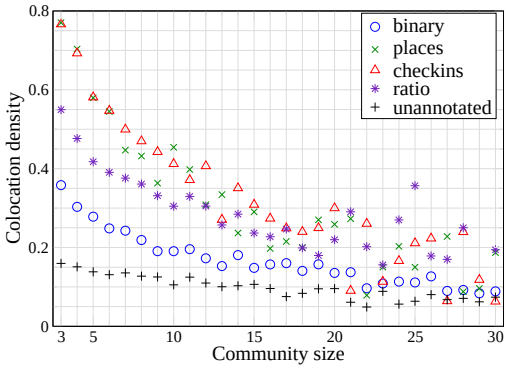
That is, the proportion of the social ties between community members where the users concerned have been colocated. This measures the extent to which online social ties between community members reflect offline meetings.

3.2 Dataset description

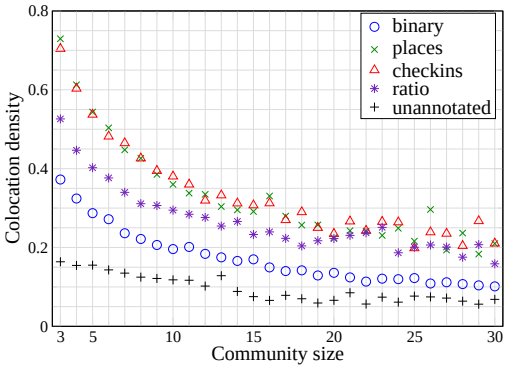
We apply our method to datasets from two OSNs with user check-in information, Gowalla and Twitter. Properties of the two datasets are shown in Table 1.

Gowalla is a location-based social network created in 2009, and discontinued when the company was acquired by Facebook in December 2011. Users declare friendship ties to form a social network, and use their mobile phones to check in at their location and notify their friends. We consider a complete snapshot of the service downloaded in August 2010.

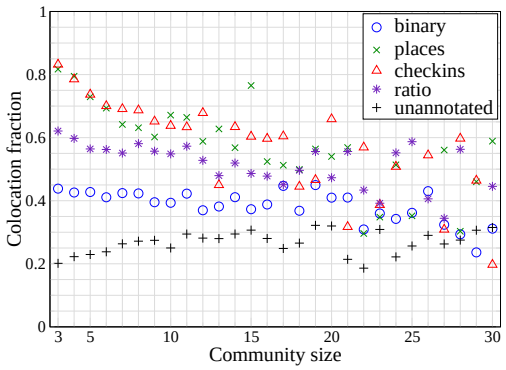
Twitter is one of the most popular online social networking services, with over 300 million registered users at the end of 2011. We obtain check-in information for Twitter users by considering those who have shared Foursquare check-ins publicly through Twitter. Foursquare is the most popular online location-based social network, with over 15 million users in January 2012. The dataset consists of Foursquare check-ins pushed to Twitter between May and November 2010, the users who shared these check-ins, and the social links between them on Twitter. We take two users u_i and u_j to have an edge e_{ij} between them in the social graph if each follows



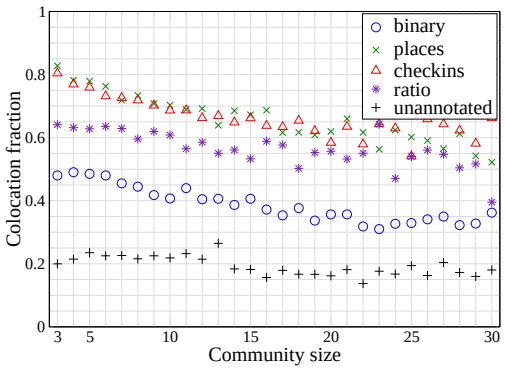
(a) Colocation density: Gowalla



(b) Colocation density: Twitter



(c) Colocation fraction: Gowalla



(d) Colocation fraction: Twitter

Figure 2: Colocation-based measures for community place-focus

the other on Twitter. This aims to exclude links where many users follow brands or celebrities but the tie is not bidirectional, and does not represent friendship.

3.3 Results of community detection

We used the Louvain algorithm [2] to perform community detection on the annotated graphs and on the original social graphs. As mentioned in section 2.2.1, `binary` forces choice of the threshold $t = 1$: only the edges e_{ij} where u_i and u_j are not placefriends are removed. This results in removing 85% of edges and 31% of users from the Twitter graph, and 65% of edges and 46% of users from the Gowalla graph.

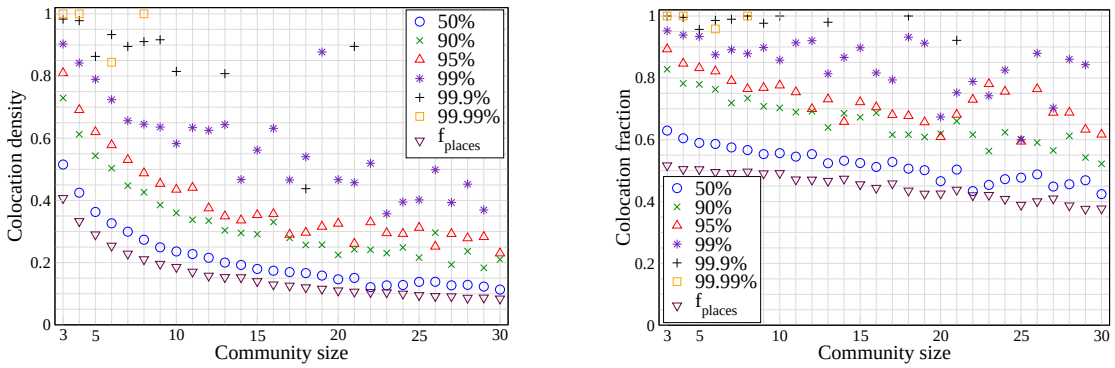
For the other three annotation functions we present results for t selected such that the 10% highest weighted edges are retained in the thresholding step. The choice of t is discussed further in section 3.3.3.

3.3.1 Community size distribution

Figure 1 shows the Complementary Cumulative Distribution Function (CCDF) of community sizes. For the unannotated graph, there are a few large communities with thousands of nodes returned in both networks. *Many of the smaller communities detected in the annotated graphs, confirmed to be more place-focused in the following section, are contained within these huge communities.* Since place information is not exposed to the detection algorithm by the unannotated graph, spatially important edges are lost amongst other social links and subsumed into large communities with little spatial meaning.

Some improvement is obtained by using `binary` to remove social edges that are not between placefriends. Figure 1 shows that fewer very large communities result, but some still persist. This is likely due to the fact that `binary` forces $t = 1$ and users who happen to have a social tie to a placefriend, but who are not part of a place-focused community, remain in the graph. The other three annotation functions yield community size distributions where the largest communities have fewer than 1000 members. The problem of many place-focused communities being hidden inside large communities that may be significant to the topology of the OSN social graph, but which are not place-focused, is reduced.

In all cases, more than 90% of communities have fewer than 30 members. We are most interested in these small communities, as they may feasibly represent real communities of OSN users who are friends and visit the same physical places. Our choice of this upper bound for the possible size of a place-focused community is supported by the finding of Onnela et al. [15] that groups of 30 people or below tend to be geographically tight, but become more spread out beyond this size. In the following section we consider these communities of size 30 or below.



(a) Colocation density

(b) Colocation fraction

Figure 3: Effect of varying the threshold parameter t to exclude different proportions of the links from the Twitter graph after application of the annotation function `places`.

3.3.2 *Colocation-based measures*

We now examine the colocation-based measures. We present results for the different annotation functions and for the unannotated graph. For these measures, `binary` is a second baseline over that of the unannotated graph: by comparing the other annotation functions with thresholding to `binary`, we can judge how much of any observed effect is due to the chance of users who are *placefriends* and friends happening to be colocated, and how much is due to the use of the other annotation functions.

Figure 2 shows the mean colocation density and mean colocation fraction for communities of a given size. The communities detected having annotated the graphs with `places`, `checkins` and `ratio` have consistently higher values for both measures than those found after application of `binary`, and in the unannotated graph. We therefore see that we can use our annotating and thresholding mechanism to remove spatially unimportant edges from the graph and reveal communities of users who not only go to the same places – these place-focused communities we are seeking – but also who tend to go to the same places together.

The functions `places` and `checkins` give higher values of the colocation-based measures than `ratio`. This could be because the thresholding step removes users who have checked in to fewer places, and it is possible that if users check in to enough of the same places they will eventually be colocated by chance. However, these users have declared online friendship, so we know that we are not merely grouping so-called *familiar strangers* [13] who meet regularly but are not friends. We therefore argue that colocation of these users is likely to be meaningful.

We also note that for `places` and `checkins`, thresholding necessarily discards users who have checked in to fewer than t places and with fewer than t check-ins, respectively. It may be that to detect place-focused communities it is necessary to ignore users for whom we do not have enough data. As location continues to become more important to online social networking, data sparsity may be less of a problem. For now, the verifiably place-focused communities we can detect include those users who actively make use of location features, and these are the users for whom applications of the detection of these communities could be most useful.

3.3.3 *Effect of choice of threshold value*

The threshold value t acts as a tuning parameter, determining how aggressive we are in excluding users from consideration on the basis that they do not have social ties that are strongly place-focused enough. To illustrate the effect of changing t, we present results for `places` and the Twitter network. We saw similar results for the other functions and for Gowalla; while these are not shown due to space limitations, the following gives an idea of the rôle of the parameter t and considerations in choosing its value.

The lowest possible threshold value for `places`, $t = 1$, corresponds in this network to removing 50% of links from the graph, hence this is the lowest proportion for which values are shown.

Figure 3 shows how communities found using higher t are increasingly place-focused. Pruning less spatially important links from the graph will increase place-focus, but users may be removed due to not having highly-weighted incident links. When choosing t the needs of the application must be considered: it may be that any communities found need to be highly place-focused even if most users will be excluded from consideration, in which case t should be high. Conversely, if moderate place-focus is acceptable and we want to assign more users to communities, lower t will be appropriate. For example, if using these communities for friend recommendation, it might be better to have lower place-focus but more users available for consideration. For applications such as privacy management, higher place-focus may be more important and a more restricted group of users would be an acceptable cost.

4. CONCLUSIONS

As location data becomes more available on online platforms, information about the places where users go can be used to distinguish between social ties that connect friends who meet face-to-face, and those ties that are maintained mostly or entirely online. We have presented a way to find place-based communities in online social networks, by annotating the social graph with weights derived from information about physical places users visit. We have applied our technique to two real networks, and seen that while community detection using the network topology alone can fail to group friends who visit the same places, we can find place-focused groups where users are often colocated. Our work has many poten-

tial applications, from friend recommendation algorithms to privacy management and automatic contact organization.

5. ACKNOWLEDGEMENTS

This work was funded in part through EPSRC Project MOLTEN (EP/I017321/1). Chloë Brown is a recipient of the Google Europe Fellowship in Mobile Computing, and this research is supported in part by this Google Fellowship.

6. REFERENCES

[1] BACKSTROM, L., HUTTENLOCHER, D., KLEINBERG, J., AND LAN, X. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of KDD '06* (New York, NY, USA, 2006), ACM, pp. 44–54.

[2] BLONDEL, V. D., GUILLAUME, J.-L., LAMBIOTTE, R., AND LEFEBVRE, E. Fast unfolding of communities in large networks. *J. Stat. Mech. 2008*, P10008 (2008).

[3] CHANG, J., AND SUN, E. Location3: How users share and respond to location-based data on social networking sites. In *Proceedings of ICWSM '11* (Barcelona, Spain, 2011).

[4] CHUA, V., MADEJ, J., AND WELLMAN, B. *Personal communities: the world according to me*. The Sage Handbook of Social Network Analysis. Sage Publication, London, 2011.

[5] CRANDALL, D. J., BACKSTROM, L., COSLEY, D., SURI, S., HUTTENLOCHER, D., AND KLEINBERG, J. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences 107*, 52 (2010), 22436–22441.

[6] CRANSHAW, J., TOCH, E., HONG, J., KITTUR, A., AND SADEH, N. Bridging the gap between physical location and online social networks. In *Proceedings of UbiComp '10* (New York, NY, USA, 2010), ACM, pp. 119–128.

[7] FELD, S. L. The focused organization of social ties. *American Journal of Sociology 86*, 5 (1981), pp. 1015–1035.

[8] FORTUNATO, S. Community detection in graphs. *Physics Reports 486* (Jan. 2010), 75–174.

[9] HOSSMANN, T., LEGENDRE, F., AND SPYROPOULOS, T. Analysis of three dimensions of human relations: mobility, social and communication interactions. Poster at Applications of Network Theory Conference, Apr. 2011.

[10] JONES, S., AND O'NEILL, E. Feasibility of structural network clustering for group-based privacy control in social networks. In *Proceedings of the Sixth Symposium on Usable Privacy and Security* (New York, NY, USA, 2010), ACM, pp. 9:1–9:13.

[11] KOSTAKOS, V., AND VENKATANATHAN, J. Making friends in life and online: Equivalence, micro-correlation and value in spatial and transpatial social networks. In *IEEE SocialCom* (August 2010), pp. 587 –594.

[12] LAMPINEN, A., TAMMINEN, S., AND OULASVIRTA, A. All my people right here, right now: management of group co-presence on a social networking site. In *Proceedings of GROUP '09* (New York, NY, USA, 2009), ACM, pp. 281–290.

[13] MILGRAM, S. *The individual in a social world: essays and experiments*. Addison-Wesley Pub. Co., Reading, Mass., 1977.

[14] MISLOVE, A., MARCON, M., GUMMADI, K. P., DRUSCHEL, P., AND BHATTACHARJEE, B. Measurement and analysis of online social networks. In *Proceedings of IMC '07* (San Diego, California, USA, 2007), ACM, pp. 29–42.

[15] ONNELA, J.-P., ARBESMAN, S., GONZÁLEZ, M. C., BARABÁSI, A.-L., AND CHRISTAKIS, N. A. Geographic constraints on social network groups. *PLoS ONE 6*, 4 (04 2011), e16939.

[16] PAN, S., BOSTON, D., AND BORCEA, C. Analysis of fusing online and co-presence social networks. In *IEEE PERCOM Workshops* (March 2011), pp. 496 –501.

[17] PAPADOPOULOS, S., KOMPATSIARIS, Y., VAKALI, A., AND SPYRIDONOS, P. Community detection in social media. *Data Mining and Knowledge Discovery* (2011), 1–40.

[18] SCELLATO, S., NOULAS, A., LAMBIOTTE, R., AND MASCOLO, C. Socio-spatial Properties of Online Location-based Social Networks. In *Proceedings of ICWSM '11* (Barcelona, Spain, July 2011).