International Workshop on Spatial and Spatiotemporal Data Mining, SSTDM-13
Proc. 2013 IEEE 13th International Conference on Data Mining Workshops,
Dallas, TX, December, 7, 2013, pp 1036-1044.

Analyzing the Proximity and Interactions of Friends in Communities in Gowalla

Tommy Nguyen
Rensselaer Polytechnic Institute
110 8th St, Troy, NY 12180
Tel/Fax: (518) 276-{2094,2529}
Email Address: nguyet11@rpi.edu

Mingming Chen
Rensselaer Polytechnic Institute
110 8th St, Troy, NY 12180
Tel/Fax: (518) 276-{2094,2529}
Email Address: chenm8@rpi.edu

Boleslaw K. Szymanski
Rensselaer Polytechnic Institute
110 8th St, Troy, NY 12180
Tel/Fax: (518) 276-{2716,2529}
Email Address: szymab@rpi.edu

***Abstract*—We collected friendship information and location data from a social media website called Gowalla to analyze the relationship between geographical space and friendship. First, we analyzed how geographic proximity shapes the structure of the social network by limiting joined activities among distant users. Second, we incorporated information about geographic locations that users visited into three selected community detection algorithms (Clique Percolation Method, Inference Algorithm, and GANXiS) to detect friendship communities where members are on average separated by one friendship link and also likely to be close to each other geographically. Third, we proposed a technique to generate covers of fixed sizes by using a combination of social and geographic information for the purpose of comparing them to communities detected by the selected algorithms. Finally, we used community quality measurements based on friendship link connectivity and geographic locations visited by users to examine detected communities.**

I. Introduction

Contrary to the belief in the death of distance barrier to forming social ties [1], we find that the creation of friendship between two people in Gowalla is more likely to occur when they are geographically closer, and the likelihood of users being friends rapidly decreases as the geographic distance between them increases. Such geographic effects may help in designing spatially-aware community detection algorithms where on average every two people in a community are separated by a few hops and also likely to be within spatial proximity.

A common approach in community detection is to divide a network into multiple partitions by maximizing the number of edges within each partition and minimizing the number of edges between them. The often used quality measurement for the partitions is modularity that compares the difference between the fraction of edges inside and fraction of edges across a partition and such expected difference if edges in the network were randomly distributed [2]. Greedy approaches like hierarchical clustering [3] and spectral approaches such as minimum cuts [4] divide a network into disjoint partitions by combining or separating clusters of nodes so that modularity is maximized at every step. As studied by authors in [5][6], a problem with this modularity maximization approach is that it inclines to merge two separated communities together, increasing the value of modularity, but creating the merger that does not reflect the ground truth.

Another approach to community detection is to divide a network into multiple partitions so that the majority of members within each partition shares a common attribute [7]. A proposed attribute is based on friendship similarity defined as the density of common friends between pairs of nodes [7]. A problem with this proposed attribute is that it allows for a community consisting of people who have a lot of friends in common but are not friends of each other. However, this imperfect definition works well in practice because people who have a lot of friends in common are likely to be friends themselves. Since community detection is an active area of research, our goal in this paper is not to provide another technique that detect communities (many have been proposed) but to incorporate the spatial information of nodes into existing algorithms for analyzing Gowalla and propose a null model (generating covers) to benchmark the detected communities.

In this paper, we combine these two approaches in community detection by incorporating the location information of users and geographic distances between them into three selected algorithms taken from the literature. First, we want to minimize the number of edges between communities and maximize the number of edges within them. Second, we want members inside a community to be within spatial proximity by giving geographically correlated friends more weight than distant friends during the detection process. This combined approach applies a natural interpretation of a friendship community where members are well connected and also likely to be geographically close.

Our goal is to extract information about friendship in communities and measure their physical interactions in a large-scale social network called Gowalla. Applications that we foresee might benefit from such spatial effects include recommendation systems and link prediction by designing systems based on the knowledge of users' geographical locations, their social connections, and the structure of their friendship communities. For instance, recommendation systems could be enriched by incorporating geographical information of users, their friends and location-based ratings to increase the quality of the recommended item [8]. An example of this is to personalize the ranking of information on the web for mobile phones where users get information relevant to their current location, friends and/or followers, and the structure of their

communities [9]. Link prediction could be enriched by using pairs of users that are geographically close and belong to the same community to predict how likely they will become friends or connected in the future [10].

The rest of this paper is organized as follows. In section II, we described the data analysis and provided details of the information that we collected from Gowalla for analyzing the relationship between space and friendship. In section III, we incorporated geographic information of users into three selected community detection algorithms consisting of a modified version of Clique Percolation Method (CPM) [11], Inference Algorithm (IA) [2][3], and GANXiS [12] to detect disjoint communities of friends in Gowalla. In section IV, we designed an experiment in which we generated different types of covers by using a combination of social and geographic information. In section V, we used quality metrics based on the link connectivity, geographic proximity, and physical interactions among members to examine detected communities as a function of their sizes and used covers as a baseline for examining communities. Before concluding in section VII with a summary of the results, we presented a literature review of community detection and location-based social networks focusing on the spatial aspects of friendship and their applications in section VI.

II. Preliminaries

We collected data from a location-based social networking provider called Gowalla that allowed people to use their internet-enabled and sensing-capable mobile phones to record and share their current location with their friends. Additional details of the collected dataset are given in [13]. As of now, Gowalla is no longer operated by itself since it has been integrated into Facebook. Foursquare is another location-based social media that provides the same functionality as Gowalla and is still active. However, Foursquare has a stricter privacy policy that limits the API from providing checkins, even though the information is available on web browsers. As the result, the data collected from Gowalla allows us to measure physical interactions of friends and their hidden communities.

Given a set of users U, let $u \in U$ be a particular user, L_u be a set of its shared locations known as checkins, and F_u be a set of its friends. A shared location $l \in L_u$ of the user u is a tuple of three elements denoted as l^1, l^2, and l^3 corresponding to the latitude, longitude, and timestamp of the location l, respectively. The friendship network denoted as $F = (U, E_U)$ is an undirected and non-weighted graph where an edge represents reciprocal friendship; that is, $e = (u, u') \in E_U$ means $u' \in F_u$ and $u \in F_{u'}$. The geographic distance $d(u, u')$ between two users u and u' is estimated by averaging the locations in L_u and $L_{u'}$ and using the haversine formula to calculate arch distances.

The level of physical interaction between user u and u' denoted as $I(u, u')$ is calculated from their shared locations as follows. Two locations $l \in L_u$ and $l' \in L_{u'}$ are equivalent if they are within geographic proximity $d(l, l') < d_\epsilon$ and occurred within a time interval $|l^3 - l'^3| < t_\epsilon$. Have such two equivalent locations l_u and $l_{u'}$ means we infer u and u' have gone to the place l together since the purpose of Gowalla was to help friends meet at different places.

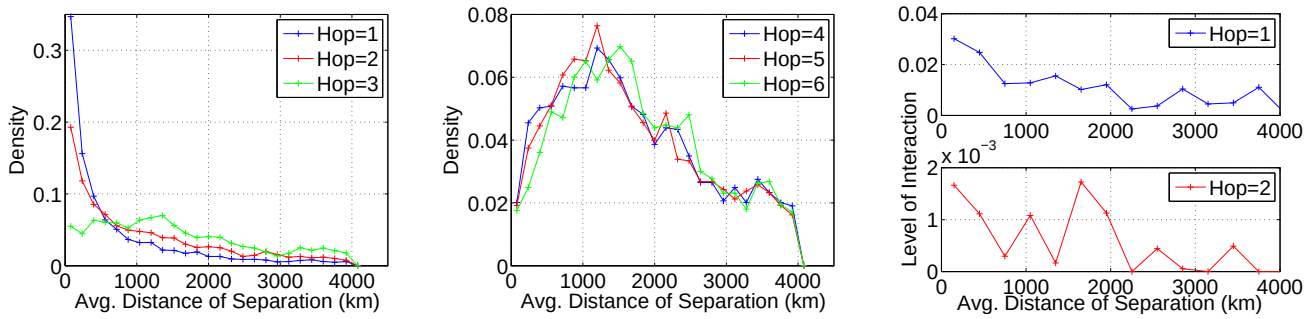
The maximum pair-wise equivalence between L_u and $L_{u'}$ is defined as the longest sequence of equivalent location pairs $((l_1, l'_1), \ldots, (l_k, l'_k))$, such that for each $1 \leq i \leq k$, $l_i \in L_u, l'_i \in L_{u'}$ and l_i is equivalent to l'_i. The level of physical interaction $I(u, u')$ is defined as the length k of the maximum pairwise equivalence divided by the size of the smallest locations set (i.e., k / min$(|L_u|, |L_{u'}|)$). Finding the maximum pairwise equivalence can be reduced to a network flow problem where polynomial running time algorithms such as Ford-Fulkerson can be used to calculate the maximum number of matches.

A. Data Analysis

In Fig. 1(a,b), we plotted the density of friends (hop=1), friends-of-friends (hop=2), and pairs of users up to six degrees of separation as a function of the average geographic distance between two users in km. For each level $1 \leq k \leq 6$ of indirection (measured in the number of hops), we randomly selected 5,000 non-cyclic paths of length k and created from the ends of these paths 5,000 pairs from the Gowalla dataset, each pair with k indirection of friendship. We analyzed pairs that were within 4,000 km distance from each other. In Fig. 1(a), the density of direct friends (4,317 total) reaches the highest value of 0.35 (in other words, 1511 pairs) at the lowest geographic separation in the range from 0 to 160 km (each point at distance x represent users with distances from x-160km to x+160 km) and continues to decrease as the distance between them increases. At the second level of indirection, the density of friends-of-friends (3,464 total) achieves the highest value 0.19 in the range from 0 to 160 km and continues to decrease as the geographic distance between them increases. Geographic proximity has an effect where friends (hop=1) and friends-of-friends (hop=2) are more likely but not necessary required to be within proximity. For instance, 61% of friends are within 480 km and 47% of friends-of-friends are within 640 km.

Another way of looking at the results is that people who are separated by three or more hops are unlikely to be within geographic proximity. In Fig. 1(b), we plotted pairs of users who are separated by four, five, and six hops. We noticed that they are not likely to be within geographic proximity. The density of those pairs reaches the highest value 0.07 at the 160 km range centered at 1,200 km and continues to decrease regardless of their degrees of separation.

In Fig. 1(c), we plotted the average amount of physical interactions $I(u, u')$ of friends (hop=1) and friends-of-friends (hop=2) as a function of their geographic distance in km. The larger the geographic distance between friends, the less likely they physically interact by going to the same places together. The highest peak (0.027) is at the lowest geographic separation from 0 to 266 km and continue to gradually decrease (with some small fluctuations) as the distance between them increases. For friends-of-friends, the physical interactions reflect



(a) (b) (c) t_ϵ=30mins, d_ϵ=1km

Fig. 1: Measuring Density of Pairs (a,b) and Level of Interaction (c)

the probability that they happened to be together.

We like to mention that it is possible the locations of some users are irrelevant to their distant friends. This may be a source of potential bias where the geographic proximity of friends may be enlarged by a friendship selection process in Gowalla in which users subjectively add friends who are within their geographic proximity. However, we noticed that 38% of friends are geographically separated by more than 520 km. Also, the Gowalla data and other social media indicate that distant friends are selected, perhaps for the purpose of keeping in contact [14].

III. Community Detection Algorithms

We selected three community detection algorithms based on their popularity (CPM subsection III-A), promising experimental results (IA subsection III-B), and ability to scale to millions of nodes and edges (GANXiS TTL subsection III-C) for the purpose of capturing and measuring the interactions of users inside a community. Below we summarize the selected algorithms and describe how we incorporated geographic information of users into the process of detecting friendship communities in Gowalla (since level of interactions is correlated with distance as seen in Fig. 1(c)).

A. Clique Percolation Method

The CPM algorithm was proposed in [11] to detect overlapping communities by combining cliques or fully connected subgraphs. Given an undirected graph $F = (U, E_U)$, let H_m denotes the set of all cliques in F of the size m. The clique-graph $G = (H_m, E)$ consists of cliques in H_m represented as nodes, and edges between pairs of cliques if they have $m-1$ overlapping members. Each connected component of the graph G is a community consisting of many fully connected subgraphs of F.

A problem of the CPM algorithm is its lack of scalability because the number of cliques explodes as m increases for large networks. Unfortunately, the problem of finding the clique with the largest size in a given graph is NP-hard [15] preventing the algorithm from using cliques with the near largest size.

We modified CPM to incorporate geographic information of nodes and made the algorithm scalable as follows. Instead of finding cliques of large sizes, we find triangles ($m = 3$) since they can be efficiently identified in parallel using map-reduce. To limit the number of triangles, we select a subset of disjoint triangles from all possible triangles by using geographic distances between pairs of nodes as follows. The average geographic distance of a triangle t is defined as $(1/3)\sum d(u, u')$ for $u \neq u' \in t$. We take a triangle one at a time from a sorted list of triangles until all possible disjoint triangles have been taken. If a user is not part of any disjoint triangle, we assign it to a triangle that maximizes the number of edges between this user and the triangle and use geographic distances to break ties by assigning a user to the geographically closest triangle.

The clique-graph G' is defined as $G' = (T, E_T)$ where T is the set of modified triangles and E_T is the set of edges between triangles that are assigned as follows. For each triangle, we create a single clique edge from this triangle to the one that maximizes the number of friendship edges between them, and use geographic distances to break ties if necessary. Like in the original CPM algorithm, each connected component of G' is a community consisting of geographically correlated and well connected subgraphs of F.

B. Modularity Maximization

Modularity maximization is a popular technique used to find communities proposed in [2][3]. Given a graph $F = (U, E_U)$ and a set P containing disjoint partitions or subsets of U, the modularity Q of the partitions in P is defined as:

$$Q = \sum_{p_i \in P} e_{ii} - a_i^2 \quad (1)$$

where e_{ij} is the fraction of edges between nodes in the partitions p_i and p_j, and $a_i = \sum_j e_{ij}$ is the fraction of edges leaving the partition p_i [2]. A positive value of Q correlates with the difference between densities of edges inside and edges leaving the partitions compared to a null model.

To maximize modularity, a greedy approach based on hierarchical clustering were proposed in [3][16]. Initially, every

node in U belongs to its own community. Then the pair of communities with the highest increase in modularity is merged together. The process of merging repeats $n-1$ times where $n = |U|$. The clusters with the highest overall value of modularity at each iteration are taken as a set of communities.

For weighted networks, Newman proposed a simple technique to map weights of integer values to multigraphs [17]. For every edge of the weight w_{ij}, there will be $w_{ij} - 1$ additional unweighed edges added between node i and j, and the weight w_{ij} is set to 1. The definition of modularity remains the same, since the fraction of edges e_{ij} between partition p_i and p_j can simply incorporate multiple edges between nodes. We incorporated geographic information about users into the Inference Algorithm by assigning weights to edges based on spontaneousness and typical means of travel: walking up to 1.6km, biking/using public transportation up to 25km, short car/train ride up to 100km, long car/train ride up to 500km, and plane flight above 500km. Friends who are within walking distance (1.6 km) get the highest weight of 2^4. Friends who are within biking distance (25 km) get the second highest weight of 2^3. Friends who are within driving distance get a weight of 2^2, and so on.

C. GANXiS

GANXiS was proposed in [12] based on a probabilistic propagation process that spread labels between speakers and listeners. Given a graph $F = (U, E_U)$, each node $u_i \in U$ initially carries a unique label i in its pocket $p_i = \{i\}$. When a node u is randomly selected to speak, it requests all members of its neighborhood, nodes that are adjacent to u to randomly send a label in their pocket to u. The probability of a label being chosen by u' in its pocket $p_{u'}$ is proportional to number of times the label was added; the more times a label was added, the more likely it will be chosen. The probability of a speaker u_i choosing a label from a listener u_j is based on the weight w_{ij}/w_i where w_i is the sum of all weighted edges coming out of u_i. For unweighted networks, $w_{ij} = 1$.

The algorithm repeats until the maximum number of iterations is completed where in each iteration everyone gets to speak exactly once in a random order. At the end, labels that have a probability of being chosen to send to a speaker less than a threshold r are deleted. Finally, the labels that a node carries determine the communities that to which it belongs. For instance, nodes that carry a label i will belong to the community c_i. Time to live (TTL) has been recently proposed to limit the number of labels that nodes propagate. TTL defines the number of times a label can be sent (so it reaches limited number of nodes within TTL hop distance).

The advantage of GANXiS is that it scales linearly with the number of edges, but the disadvantage is that the relationship between convergence and the number of iterations is yet unknown. GANXiS is capable of discovering overlapping communities, but we selected its running parameters in such a way that the results included only disjoint communities to make them compatible with the results of other algorithms. We incorporated geographic information of users into GANXiS by assigning weights equal to the reciprocal of geographic distances between friends. This is an extension of the interpretation of speaker-listener propagation algorithm where a listener is more likely to be able to hear a speaker if they are within spatial proximity (close to each other).

TABLE I: Six Techniques for Generating Covers

Algorithm	Abbreviation	Spatial Info.?	Social Info.?
Completely Random	CR	no	no
Random Walk	RW	no	yes
Closest Friend First	CFF	yes	yes
Farthest Friend First	FFF	yes	yes
Closest to All	CTA	yes	yes
Farthest to All	FTA	yes	yes

IV. A Null Model

We proposed to integrate spatial and friendship information of nodes into a process of generating covers. The purpose of the covers is to serve as a baseline for analyzing the performance of various community detection algorithms under a quality measurement. In subsection IV-A, we described how we generated six covers by using a combination of spatial and friendship information in traversing the network. In subsection IV-B, we selected a few quality measurements for examining covers and detected communities. In subsection IV-C, we examined the covers using the selected quality measurements.

A. Generating Covers

Given a graph $F = (U, E_U)$, a cover $C \subset U$ of size k is a subgraph of F with k nodes selected in a specific way. A completely random cover CR is one where each user $u \in U$ has the same probability of being added during the selection. In a random walk cover RW, we first randomly add a seed into the cover, then randomly select a friend of the most recently added user, and continue selecting friends until the cover reaches the size k. The closest-friend-first cover CFF is similar to RW but instead of adding a random friend, we add the spatially closest friend not in the cover of the last added user. If all of that user's friends have already been added into the cover, we go back one step to the previously last added user and branch out from there. We call this the roll-back mechanism. The farthest-friend-first cover FFF is similar to CFF except that we take the spatially farthest friend instead of taking the closest one. The closest-to-all cover CTA is similar to CFF but instead of adding the spatially closest friend to the last added user, we add the spatially closest friend with respect to all members already in the cover. Finally, the farthest-to-all cover FTA is one where we take the spatially farthest friend with respect to all members already in the cover. Cover generation algorithms such as CTA and FTA are described in Fig. 2 without the roll back mechanism for simplicity. We listed the covers and their details in Table I.

B. Measuring Covers

We use three types of quality measurements based on the link connectivity and location of members to measure covers and communities.

```
procedure CoverGeneration(k)
    F = (U, E_U)
    seed = rand(1, |U|), cover = [seed]
    while len(cover) < k do
        distances = [ ], m = len(cover)
        for u in F_seed do
            // Calc. haversine distance from u to cover[i].
            d_u = (1/m) Σ_{i=1}^{m} d(cover[i], u)
            distances.append((u, d_u))
        end for
        // sort d_u from least to greatest or vice-versa
        distances = sort(distances, key = x: x[1])
        for u, d_u in distances do
            if u ∉ cover then
                cover.append(u)
                seed = u
            end if
        end for
    end while
    return cover
end procedure
```

Fig. 2: Generating CTA & FTA Covers

The first type of measurements is based on the intra-edge count IEC defined as the number of edges whose both ends are inside the cover. The contraction $CONT$ of a cover is computed by dividing intra-edge count by the size of the cover. The intra-density IND of a cover is calculated by dividing intra-edge count by the intra-edge count of a completely connected cover of the same size. For these three measures (IEC, $CONT$, IND), higher the value, better formed is the community.

The second type of measurements is based on the boundary-edge count BEC defined as the number of edges whose one end is inside the cover while the other is outside. This metric is useful for taking into account the effect of adding high degree users into covers of large sizes since such users are likely to increase both the intra- and boundary-edge counts. The expansion EXP of a cover is computed by dividing the boundary-edge count by the size of the cover. The conductance $COND$ of a cover is defined as $COND(C) = \frac{BEC(C)}{2IEC(C)+BEC(C)}$. For these three measures (BEC, EXP, $COND$), lower the value, better formed is the community.

The third type of measurements is based on pair-similarity that measures a given metric such as friendship similarity among pairs of nodes. This is applicable to the definition of a community of which members have a lot of commonality [7]. We replace friendship similarity ratio with three additional measurements based on the geographic proximity and location of nodes. The first one is the geographic diameter of a cover GDI defined as the geographic distance between the two farthest nodes. The second one is the average geographic distance AGD among pairs of nodes. Here, lower the measure (GDI and AGD), better formed is the community. The third one is the sum of the levels of physical interactions SLI among pairs of nodes for which higher the measure, better formed is the community.

TABLE II: Measurements for Cover C of the size k

Measurement	Definition
IEC [18]	$\lvert\{(v_i, v_j) \in E \mid v_i \in C \ \wedge \ v_j \in C\}\rvert$
BEC [19]	$\lvert\{(v_i, v_j) \in E \mid v_i \in C \ \vee \ v_j \in C\}\rvert$ - IEC
CONT	IEC/k
EXP [20]	BEC/k
IND [18]	$IEC/(0.5k(k-1))$
COND [20][19]	$BEC/(2IEC + BEC)$
GDI	$\max d(u, u') \ \forall u, u' \in C$
AGD	$\sum_{u \neq u' \in C} d(u, u')/(0.5k(k-1))$
SLI	$\sum_{u \neq u' \in C} I(u, u')$

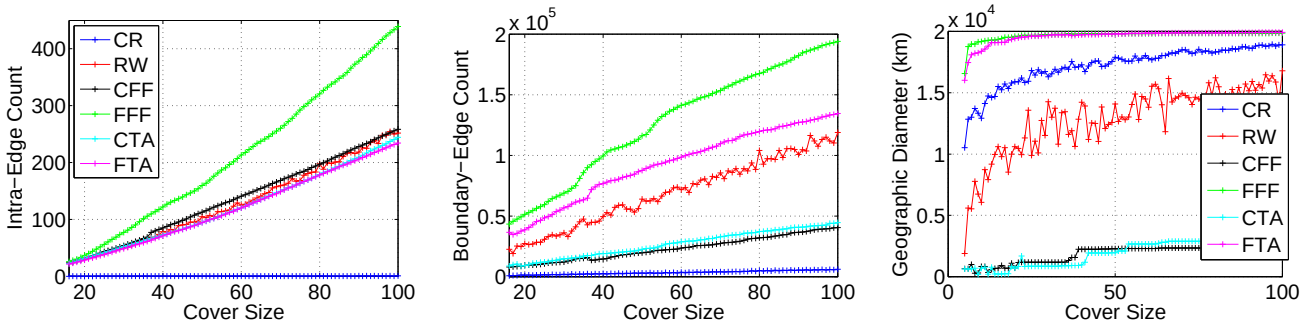
C. Measuring Covers

For each technique, we generated covers of fixed sizes from 5 to 100 with an increment of 1. For each cover size, we generated 100 covers and calculated the average intra-edge count, boundary-edge count, geographic distance, and geographic diameter. We then derived the remaining measurements.

In Fig. 3(a), we noticed that FFF outgrows the other techniques in terms of intra-edge count as the cover size increases. In Fig. 3(b), we noticed that FFF and FTA outgrow the other techniques in terms of boundary-edge count by a great margin suggesting that they strategically add users with very large degrees. While RW is decent at generating covers with high intra-edge counts as seen in Fig. 3(a), it is also biased since users with high degrees are more likely to be added, which increases the intra-edge count as the cover continues to grow. However, FFF and FTA are even more biased than RW and FFF outgrows the other five techniques because the radius of the farthest friend would cover everyone including common friends in between. On the other hand, we noticed that CFF and CTA are most effective out of the six techniques at increasing the intra-edge count while minimizing the boundary-edge count at the same time.

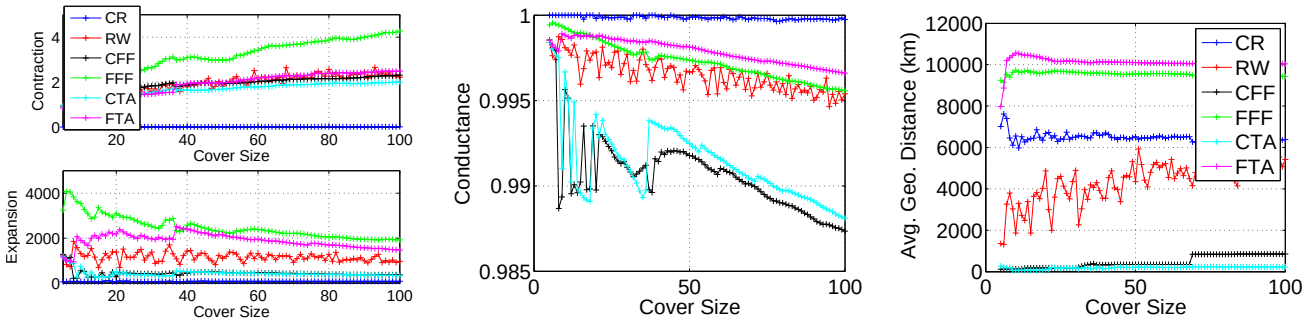
In Fig. 3(c), we measure the geographic diameter of a cover as a function of its size. As expected from how covers are generated, FFF and FTA are most effective at maximizing the geographic diameter while CFF and CTA are most effective at minimizing this measurement. The geographic diameter of FFF and FTA reaches the limit within 20 iterations, while the diameter for CTA and CFF slowly continues to grow. A similar trend is seen in Fig. 4(c) which shows the average geographic distance in contrast to the growth rate of intra- and boundary-edge counts seen in Fig. 4(a).

Last but not least, conductance is a measurement used to determine the quality of a community by considering both the intra- and boundary-edge counts. As seen in Fig. 4(b), CFF is the most effective out of the six covers at minimizing conductance since it preserves some geographic structure of the social network by traversing the edges based on who is the geographically closest friend, and adding friends who are likely to be friends with the members already in the cover. CTA is not as effective as CFF because geographic distances get diluted as the size of the cover increases. FFF and FTA



(a) (b) (c)

Fig. 3: Three Measurements of Covers (Intra-edge Count, Boundary-Edge Count, and Geographic Diameter)



(a) (b) (c)

Fig. 4: Measurements of Covers

are worse than RW at minimizing conductance. We later use the physical interactions of users to compare and contrast the results generated by the CFF cover to results detected by the community detection algorithms.

V. Examining Detected Communities

We first examined the results by looking at the total number of communities detected and the number of members in each one. The modified CPM algorithm[1] with geographic information detected 2.6K communities whose average size was 60 with the size of the largest one being 69K. IA without geographic information detected 1.2K communities with the average size of 134 and the size of the largest one being 52K. IA with geographic information detected 349 communities with the average size of 442 and the size of the largest one being 45K. GANXiS without geographic information detected 7.2K communities with the average size of 21 and the size of the largest one being 3K. Finally, GANXiS with geographic information detected 16K communities with the average size of 10 and the size of the largest one being 455. Additional information relating to community sizes is listed in Table III.

We used the network community profile (NCP) proposed in [19] to examine detected communities as a function of its size. The authors proposed to take the best partition defined by a quality feature of a given community size because it represents the potential of a partition in a community detection algorithm. We measured the community quality by taking the highest value of intra-density and the lowest value of conductance of the best partition (if any) of a given community size. For intra-density and conductance without geographic information, we consider the number of all possible intra- and boundary-edge counts. For intra-density and conductance with geographic information, we only consider the number of edges that are within geographic proximity (160 km) or roughly 2 hours of driving.

The potential issues resulting from using this approach are discussed below. First, in many situations, taking the average value of a community quality gives a more representative picture and probably is less sensitive in cases containing outliers. Second, the number of communities for a given size might vary from a large number of small communities to very few for large communities. Last but not least, there might be no

TABLE III: Detected Communities and their Sizes

	Community Size				
Algorithms	Avg.	Std.	Smallest	Largest	Total
CPM	60	1,356	6	68,671	2,572
IA	134	1,935	2	52,315	1,151
IA w (w for weighted)	442	2,954	2	45,242	349
GANXiS TTL	21	87	3	3,139	7,236
GANXiS TTL w	10	16	3	455	15,796

[1]We did not run the original CPM algorithm because of the long execution time required to generate the clique graph.

communities of a particular size, and taking the average quality might give a smooth function that is easier to extrapolate at the missing points as seen with the covers. Fig. 5-7 present the results for communities detected by CPM, IA, and GANXiS respectively.

First, intra-density rapidly decreases as the size of the cover increases because adding another member into a large community requires everyone already in it to be connected with this new member, as seen in Fig. 5-7(a). Unlike intra-density, conductance is not correlated with the community size because there are some small and large communities of varying values, as seen in Fig. 5-7(b). Third, GANXiS and IA are a little better than CPM at maximizing intra-edges that are within geographic proximity, as seen in Fig. 5-7(c). IA is the best at minimizing boundary-edges that are within geographic proximity, as seen in Fig. 6(d). Last but not least, GANXiS and IA benefited from incorporating the geographic information of users, as seen in Fig. 6-7(d), where geographically correlated friends are captured in the community detection process.

TABLE IV: Measuring Spatial Conductance

Algorithm	# Spatial Cond. of 0	Total	Ratio
CPM	21	175	0.12
IA	20	78	0.26
IA w (w for weighted)	19	84	0.23
GANXiS TTL	48	126	0.38
GANXiS TTL w	618	5977	0.10

Comparing Fig. 5-7(d) to Fig. 5-7(b), we noticed that some detected communities had a conductance value of 0. This means that every potential node within geographic proximity of a community has already been included in it. For the IA without geographic information, out of the 78 community sizes, 20 of them have geographic conductance of 0, yielding 20/78 $\approx$ 0.26 ratio. For the IA with geographic information, out of the 84 communities, 19 of them have a geographic conductance of 0, yielding 19/84 $\approx$ 0.23 ratio. The remaining values are listed in Table IV.

We examined small-size communities because humans have limited resources and cognitive abilities to keep and maintain social relationships resulting in a bounded number of active friendships known as Dunbar's number [21]. We measured and then plotted in Fig. 8 the NCP level of physical interactions in communities and covers by summing the level of physical interactions among pairs. From the plots, we observed that CPM have small communities where members are statistically more likely than members in covers to physically interact with each other by going to the same places together. In Fig. 8(a), out of 95 communities detected by CPM of the size up to 100, 84 of them have higher amount of physical interaction among members than a proposed null model, CFF. In Fig. 8(b), out of 41 communities detected by IA under the size of 100, 38 of them have higher amount of physical interaction among members than CFF. The remaining values are listed in Table V.

While CPM is the most effective at detecting communities that are intrinsically small (95 total) and where the physical interaction among member is likely to be higher than CFF (88%), IA is the most effective at detecting communities where 93% of them have higher amount of physical interaction than null model, as seen in Table V. GANXiS with geographic information detected more smaller communities (87 vs. 30), but some of the members do not physical interact much compare to a null model (0.80 vs. 0.93).

TABLE V: Measuring Physical Interaction

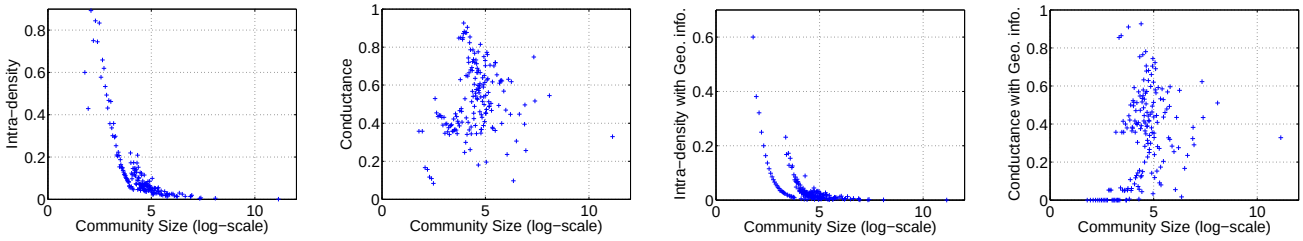
Algorithm	Count	Total	Ratio
CPM	84	95	0.88
IA	38	41	0.93
IA w (w for weighted)	28	30	0.93
GANXiS TTL	60	87	0.69
GANXiS TTL w	70	87	0.80

VI. Related Work

While a lot of effort has been put into detecting communities of generic networks mentioned in surveys such as [18][22][23], little work has included the geographic proximity of friends or nodes into designing algorithms that detect spatially correlated communities. Recent work in location-based social networks [24] has shed insights on spatial-social relationships mentioned in [13][25] and confirmed that friends are geographically correlated because distance matters [26]. Yet, less empirical work has been done on mining the data to capture the geographic effect of friends and their communities.

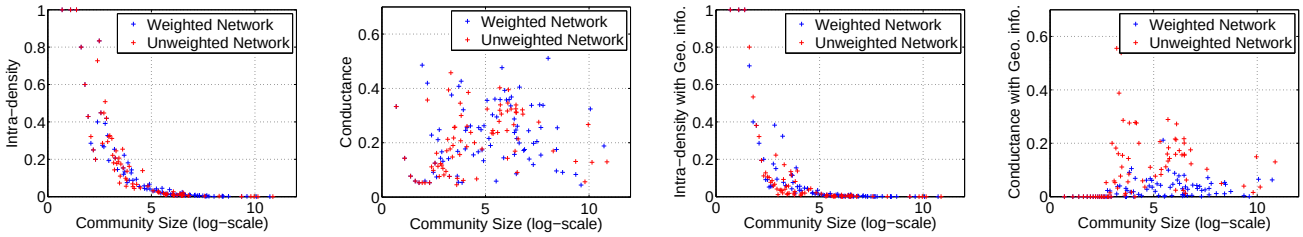
In [27], authors provided network properties (distance strength, social triads, etc.) of online location-based social networks (Gowalla, FourSquare, Brightkite) and they noticed that friendship connections are distributed across a wide range of geographic distances. In [28], authors proposed a new definition of modularity to uncover communities without the effect of geographic distance. However, we argue that in the case of a friendship network, capturing the spatial effect where a community consisting of friends who are within geographic proximity is valuable for analyzing social networks and has applications in link prediction; for instance, using non-friendship pairs in a geographically correlated community to predict whether they will be likely to become friends in the future.

Our work here is similar to the combination of [26] and [19]. In [26], authors defined geographic clustering coefficients and concluded that users with few friends are likely to have friends nearby. In [19], authors proposed network community plot (NCP) to examine detected communities as a function of size and quality features based on link connectivity and network properties of detected communities. Therefore, our work was motivated by a combination of theirs in using community detection algorithms to extract geographically correlated communities of friends. The main differences are that we also proposed and studied the behavior of generating covers as a baseline, and used location and geographic information of users in addition to their link connectivity to benchmark the selected community detection algorithms.



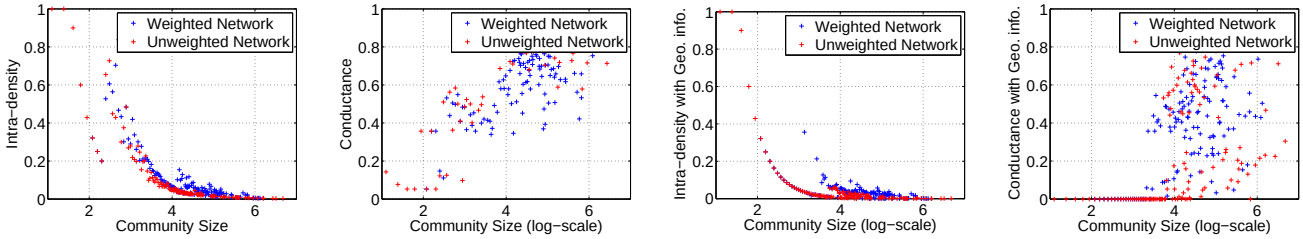
(a) (b) (c) (d)

Fig. 5: Modified CPM



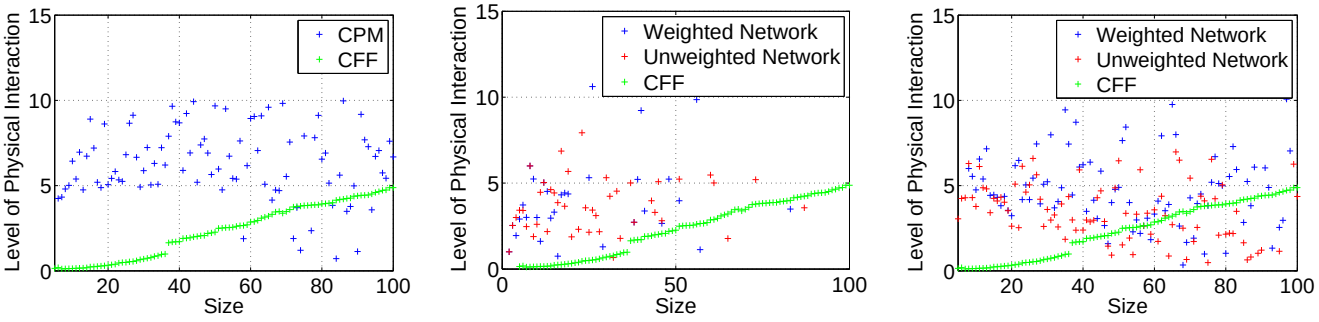
(a) (b) (c) (d)

Fig. 6: Inference Algorithm



(a) (b) (c) (d)

Fig. 7: GANXiS $r = 0.50$



(a) CPM (b) Inference Algorithm (c) GANXiS TTL

Fig. 8: Measuring the Level of Physical Interactions Among Members

VII. Conclusion

First, our data analysis of Gowalla friendship network reveals two degrees of geographically correlated separation where friends and friends-of-friends are more likely to be within geographic proximity. Conversely, pairs of users who are separated by three or more hops of friendship relation are unlikely to be within geographic proximity. Also, friends who are within geographic proximity are more likely to physically interact by going to the same places together than distant friends. Yet, the likelihood of physical interactions among friends-of-friends is minuscule even though they are geographically correlated.

Second, we showed that covers can serve as a null model for examining community structures. For most quality metrics, small communities are more likely to outperform large ones because it is much easier to find a small group to maximize a particular metric. Therefore, comparing detected communities to covers tell us how much better the algorithm is performing than a proposed null model for a given size of the community.

Finally, we used the results from the covers and compared them to the communities detected by modified CPM, unweighted and weighted IA, and GANXiS. By incorporating spatial information into CPM to make the algorithm scalable, it detected meaningful communities of a large online social network where members are more likely to physically interact than a null model. From the NCP plots, we noticed the importance of small-size communities in large social networks in which it is much harder to find a large community because humans have limited resources to create and maintain relationships. We used the level of physical interactions among members in a community as the final quality measure to compare and validate the performance of the community detection algorithms to the closest-friend-first cover.

VIII. Acknowledgment

Research was sponsored by the Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 and by the the Office of Naval Research Grant No. N00014-09-1-0607. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government.

References

[1] *The death of distance: How the communications revolution is changing our lives*. Harvard Business Review Press, revised edition ed., 2001.

[2] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, Jun. 2004.

[3] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, Dec. 2004.

[4] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[5] S. Fortunato and M. Barthlemy, "Resolution limit in community detection," *Proceedings of the National Academy of Sciences*, vol. 104, no. 1, pp. 36–41, 2007.

[6] M. Chen, T. Nguyen, and B. Szymanski, "A new metric for quality of network community structure," *ASE Human Journal*, vol. 1, no. 4, pp. 226–240, 2013.

[7] M. Goldberg, S. Kelley, M. Magdon-Ismail, K. Mertsalov, and A. Wallace, "Finding overlapping communities in social networks," in *Proceedings of the 4th ASE/IEEE International Conference on Social Computing*, 2010.

[8] J. J. Levandoski, M. Sarwat, A. Eldawy, and M. F. Mokbel, "Lars: A location-aware recommender system," in *Proceedings of the 28th International Conference on Data Engineering (ICDE 2012)*, pp. 1–12, 2012.

[9] T. Nguyen and B. Szymanski, "Social ranking techniques for the web," in *Proceedings IEEE/ACM International Conference Advances in Social Network Analysis and Mining*, 2013.

[10] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *Proceedings of the 12th international conference on Information and Knowledge Management*, pp. 556–559, 2003.

[11] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, pp. 814–818, 2005.

[12] J. Xie and B. K. Szymanski, "Towards linear time cverlapping community detection in social networks," in *Proceedings of the 16th Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD*, pp. 25–36, 2012.

[13] T. Nguyen and B. Szymanski, "Using location-based social networks to validate human mobility and relationships models," in *Proceedings IEEE/ACM International Conference Advances in Social Network Analysis and Mining*, pp. 1247–1253, 2012.

[14] M. Burke, R. Kraut, and C. Marlow, "Social capital on facebook: differentiating uses and users," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pp. 571–580, 2011.

[15] M. Sipser, *Introduction to the Theory of Computation*. PWS Publishing Company, 1997.

[16] B. H. Good, Y.-A. de Montjoye, and A. Clauset, "The performance of modularity maximization in practical contexts," *Physical Review E*, 2010.

[17] M. E. Newman, "Analysis of weighted networks.," *Phys Review E*, vol. 70, 2004.

[18] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, pp. 75–174, 2010.

[19] J. Leskovec, K. J. Lang, and M. Mahoney, "Empirical comparison of algorithms for network community detection," in *Proceedings of the 19th international conference on World Wide Web*, pp. 631–640, 2010.

[20] R. Kannan, S. Vempala, and A. Veta, "On clusterings: Good, bad and spectral," *J. ACM*, vol. 51, pp. 497–515, May 2004.

[21] "Don't believe facebook; you only have 150 friends." accessed on June 7, 2012.

[22] M. E. J. Newman, "Communities, modules and large-scale structure in networks," *Nature Physics*, vol. 8, pp. 25–31, 2011.

[23] J. Xie, S. Kelley, and B. Szymanski, "Overlapping community detection in networks: the state of the art and comparative study," *ACM Computing Surveys*, vol. 45, no. 4, 2013.

[24] E. Cho, S. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'11, pp. 1082–1090, 2011.

[25] C. Brown, V. Nicosia, S. Scellato, A. Noulas, and C. Mascolo, "The importance of being placefriends: discovering location-focused online communities," in *Proceedings of the 2012 ACM workshop on Workshop on online social networks*, WOSN '12, pp. 31–36, 2012.

[26] S. Scellato, C. Mascolo, M. Musolesi, and V. Latora, "Distance matters: geo-social metrics for online social networks," in *Proceedings of the 3rd Conference on Online Social Networks*, WOSN'10, 2010.

[27] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo, "Socio-spatial properties of online location-based social networks," in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[28] P. Expert, T. S. Evans, V. D. Blondel, and R. Lambiotte, "Uncovering space-independent communities in spatial networks," *Proceedings of the National Academy of Sciences*, vol. 108, no. 19, pp. 7663–7668, 2011.

[29] C. Brown, V. Nicosia, S. Scellato, A. Noulas, and C. Mascolo, "Social and place-focused communities in location-based online social networks," *European Physics Journal B*, 2013.

[30] J. McGee, J. A. Caverlee, and Z. Cheng, "A geographic study of tie strength in social media," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pp. 2333–2336, 2011.