Self-organized Population Segmentation for Geosocial Network Neighborhood

Low Shen Loong, Syarulnaziah Anawar, Zakiah Ayop, Mohd Rizuan Baharon, Erman Hamid
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia

***Abstract*—Geosocial network neighborhood application allows user to share information and communicate with other people within a virtual neighborhood or community. A large and crowded neighbourhood will degrade social quality within the community. Therefore, optimal population segmentation is an essential part in a geosocial network neighborhood, to specify access rights and privileges to resources, and increase social connectivity. In this paper, we propose an extension of the density-based clustering method to allow self-organized segmentation for neighbourhood boundaries in a geosocial network. The objective of this paper is two-fold: First, to improve the distance calculation in population segmentation in a geosocial network neighbourhood. Second, to implement self-organized population segmentation algorithms using threshold value and Dunbar number. The effectiveness of the proposed algorithms is evaluated via experimental scenarios using GPS data. The proposed algorithms show improvement in segmenting large group size of cluster into smaller group size of cluster to maintain the stability of social relationship in the neighbourhood.**

Keywords—Segmentation; geosocial network; virtual neighbourhood; density-based clustering; dunbar's number

I. Introduction

Since Web 2.0 technology has gained its popularity, the use of online social networks (OSN) like Facebook and Instagram has increased to the point of becoming pervasive. With the introduction of social networking, people have more interaction in a neighbourhood level. Social networks tend to take over some of the functions of neighbourhood communities [1]. These virtual communities allow better quality of social interaction among neighbours in a social network. Geosocial network neighbourhood is one of the branches in social networking that allow user to share information and communicate with other people within a virtual neighbourhood or community. Typically, geosocial networking application uses location awareness to track geolocation information that consists of current user location coordinates; longitude and latitude. Location-aware features in users' mobile device will assist GPS self-check-in function to match users' house address and current location.

One of the important research areas in geosocial network is defining neighbourhood boundaries. Previously, several studies have proposed techniques for optimal definition of neighborhood boundaries in geographic information system (GIS) such as using collaborative tagging system [2], classification and regression trees [3], and clustering [4]. In the context of geosocial network neighbourhood, optimal neighbourhood boundaries can be achieved through population segmentation. Population segmentation is a key component of neighbourhood management strategy. Population segmentation is the process of dividing population into segments based on various characteristics. Segmentation is an essential part in a geosocial networking application for safer virtual neighborhood environment. Segment is utilized to determine residents in a neighbourhood and to differentiate whether a resident is staying within its own living areas. This is particularly important in order to specify access rights or privileges to applications' resources. In addition, social connection between each resident plays a vital role in the quality of geosocial network neighbourhood. A large and crowded neighbourhood will degrade the social quality within the community. Therefore, population segmentation will increase social connectivity and allows user to share information and communicate with other people effectively.

Population segmentation for defining neighbourhood boundaries have been first applied using clustering in geosocial network by [5]. Their work applies cluster technique using DBSCAN algorithms based on user-defined parameters, which is not suitable for population segmentation in a geosocial network neighbourhood. The study has been extended in [6], where the proposed solution considered user check-in time during the clustering process. In [7], EBSCAN algorithms is used to improve the shortcomings in [5] and [6], by removing the user-defined parameters thus improving system performance. However, several problems pertaining to defining neighborhood boundaries remains: (1) low accuracy for distance calculation, (2) inability for self-organized population segmentation, (3) less social connectivity within population. To address these challenges, we propose to design self-organized population segmentation for geosocial network neighbourhood. The improved population segmentation technique is important to provide better representation and accuracy in determining the residents of geosocial neighbourhood for better social connectivity within a neighbourhood segment.

The objective of this paper is two-fold: First, to improve the distance calculation in population segmentation in geosocial networking. Second, to implement self-organized population segmentation using threshold value and Dunbar number. The rest of this paper is organized as follows: Section 2 explain population segmentation in geosocial network neighborhood and review some related work. In Section 3, the basic concepts of DBSCAN and Dunbar Numbers used in the proposed work are explained. In Section 4, the design of self-organized population segmentation is presented. Finally, we present the proof-of-concept implementation of our proposed algorithms in Section 5. This paper is summarized in the last section.

II. Population Segmentation in Geosocial Network Neighbourhood

A. Geosocial Networking

Geosocial network is a relatively new research field emerging as an integration of social network and location-based services. In recent years, some of the most prominent geosocial network applications are Waze, Foursquare and Facebook Places with millions of active users. Geosocial network applications use several techniques to provide its service, such as geocoding, geotagging, and geolocation (see Figure 1). Geocoding can be used to generate route in a traffic report application like Waze. On the other hand, geotagging is the process of adding geographical identification metadata.

Geolocation is a technique to estimate or identify a person or place geographical location within a set of geographic coordinates. With the use of mobile phone GPS to track user location, a geosocial neighborhood application will track user location through the GPS sensor with accuracy of 10 meter. Geolocation application such as Foursquare, Brightkite and others encourage user to provide details recent visited places, hometown or neighbourhood using GPS system. One example of geosocial network neighbourhood application that used geolocation technique is NextDoor.

Population segmentation is the process of dividing the population (users) in a geosocial network application into smaller group. In the process of dividing and segmenting, those who shared common characteristics such as common interest, common needs, similar lifestyles or even similar demographics profiles will be divided into segment. Population segmentation is important in understanding the distinctive needs of different parts of the population. Understanding the characteristics of population needs is important to identify services to be offered in neighborhood geosocial network. Tailoring services to specific segments is the best way of ensuring the most effective use of resources. The starting point for population segmentation strategy is identifying target populations.

Population segmentation in a geosocial network neighbourhood can be characterized physically and logically. The former is based on the distance and geolocation data, while the latter is based on the user behavior and activities within the neighbourhood. In physical segmentation, users must be divided into clusters to be more effectively targeted. One way to implement physical segmentation in geosocial network is through clustering techniques. Clustering algorithms have three basic categories that are hierarchical, partitioning and density-based [8]. Large number of data in huge databases can be deals by all these algorithms. Partitioning algorithm construct k clusters given n data object. Each cluster cannot have same common data object and can have as many group and object, where $k \leq n$. Hierarchical algorithms create hierarchical decomposition presented in a dendrogram, where a tree splits set of data object into smaller subset until one data object represents a subset. Hierarchical algorithms can be classified into two types that is divisive and agglomerative. On the other hand, density-based algorithms are designed to discover arbitrary shape of cluster that has higher density than remainder data object. Low-density region of data object considers outlier or noise.

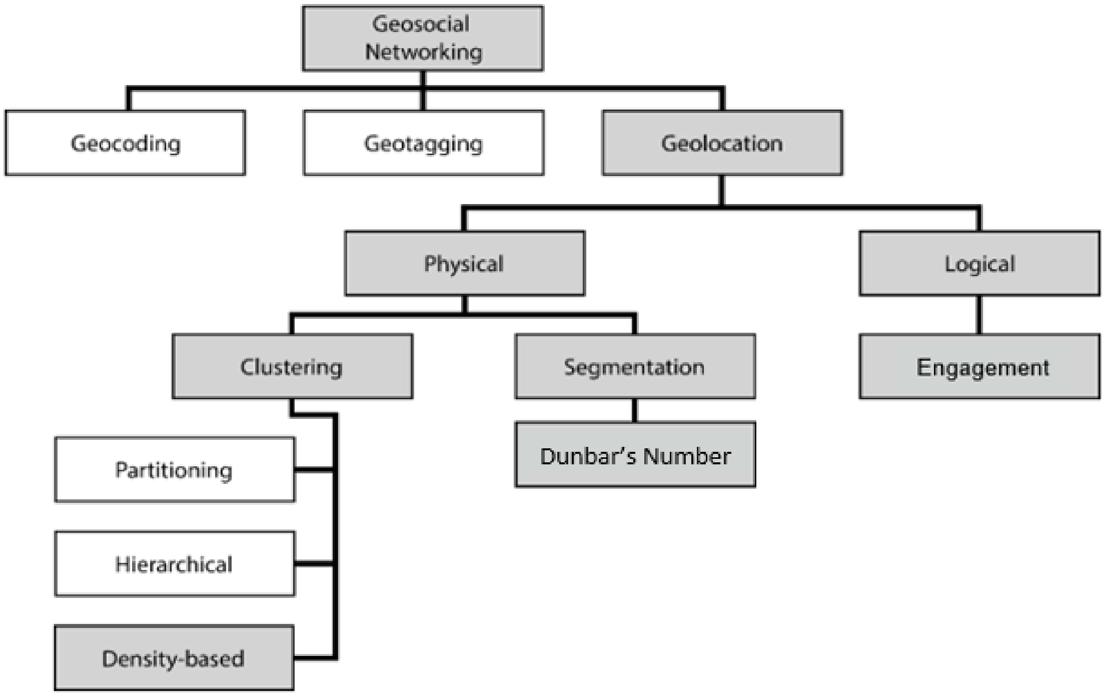

Fig. 1. Classification of Geosocial Networking.

B. Density-Based Clustering

The density-based notion is a common approach for clustering. Density-based clustering algorithms are based on the idea that objects which form a dense region should be grouped together into one cluster. The algorithms use a fixed threshold value to determine dense regions. They search for regions of high density in a feature space that are separated by regions of lower density.

In this paper, we extend the work in [5] that used DBSCAN algorithms [9] because it has the ability in discovering clusters with arbitrary shape such as linear, concave, and oval. Furthermore, in contrast to some clustering algorithms, it does not require pre-determination of the number of clusters. This algorithm is a data clustering algorithm that given a set of points in some space, it groups together points that are closely packed together, marking as outlier points that lie alone in low-density regions. DBSCAN has been proven in its ability of processing very large databases [10],[11].

In the context of geosocial networking, the DBSCAN algorithms have been first applied for clustering in geosocial network by [5]. However, as the DBSCAN algorithms only has cluster technique that is based on user-defined parameters, the algorithms is less appropriate for population segmentation in geosocial network.

III. Basic Concept used in the Proposed Work

In this section, two basic concepts used in the proposed algorithms, namely DBSCAN algorithms and Dunbar Numbers are explained.

A. Understanding DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) is designed to discover arbitrary-shaped clusters in any database, D, and at the same time can distinguish noise points. The main idea of DBSCAN is developing a cluster from each point with two parameters, Eps and MinPts. Eps (Epsilon) accepts a radius value based on a user defined distance measure and a value MinPts (Minimum Points) for the number of minimal points that should occur within Eps radius.

DBSCAN use distance function such as Euclidean Distance, Manhattan Distance, Haversine Formula for points x_i and x_j to determine as neighbourhood denoted by $\mathrm{dist}(x_i, x_j)$. Eps is the maximum radius between points for them to be considered as in the same neighborhood. The Eps-neighbourhood is denoted as $\{x_j \in D \mid \mathrm{dist}(x_i, x_j) \leq \mathrm{Eps}\}$.

Number of objects within Eps-neighbourhood can be differentiate using three types of object that is core object, border object and noise object. These three objects will rely on the second parameter that is MinPts. MinPts specifies the minimum amount of points in a neighbourhood. A neighbourhood that contains more than, or minimum amount of points (MinPts) is define as core object. The core object will denote as x_{core} derived in density $x_{core} \geq$ MinPts. Border object, x_{border}, is an object where the density is reachable from another core object, but it is not a core object. Border object belongs to a neighbourhood of core object, and the density that is less than MinPts is define as density $x_{border} <$ MinPts. Lastly, noise object is a point that fall within the neighbourhood radius, Eps, but less than the minimum amount of MinPts where no core object exists in it. Therefore, noise object does not belong to any clusters.

DBSCAN clustering have other important definition to define the relationship between objects, namely density reachable, density connected and cluster. Density reachable happen when two objects x_1 and x_n are in a chain of objects $x_1, x_2, \ldots, x_n$ such as x_{i+1} and x_1 are direct density reachable and x_n as core object is density reachable to x_1, followed by the requirement of Eps and MinPts. Meanwhile, two objects x_1 and x_2 are considered as density connected when x_1 and x_2 are density reachable to x_i with respect to MinPts and Eps. Lastly, a set of objects that density reachable to a core object will form cluster object. The relationship between the objects is illustrated in Figure 2.

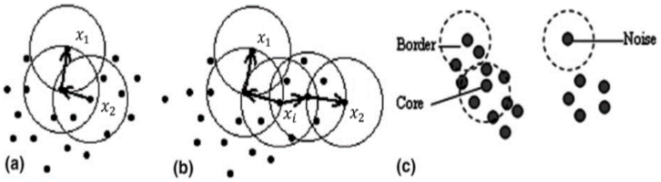

Fig. 2. Terms and Concept of DBSCAN a) Density-Reachable b) Density Connected c) Noise, Core Object and Border Object.

DBSCAN algorithms start from random points in a database, and retrieve all density-reachable of points within Eps radius. Points that are more than MinPts parameter will become core object and a new cluster will be formed. The chain of objects for the cluster will recursively detect all density-reachable objects. A seed list in DBSCAN algorithms is used to store all core objects in the chain and update newly discovered core objects into the list. dbscan and ExpandCluster are two major functions in DBSCAN algorithms as presented in [11].

B. Dunbar's Number

The size of a neighbourhood segment has an impact on social interaction. With large amount of information in each neighbourhood, each user must make their own choices about the best way to handle and use the information given the priority of personal preferences, interests, and needs. In 1992, Dunbar measured the correlation between neocortical volume and typical social group size in a community. The limit imposed by neocortical processing capacity appears to define the number of individuals with whom it is possible to maintain stable interpersonal relationships in a group. The results indicate that humans' social network size is limited to between 100 and 200 individuals, i.e. Dunbar's number [12].

IV. Design of Self Organized Population Segmentation

This study will extend previous effort by Shi et al. in [5], [6] for population segmentation in a geosocial network neighbourhood.

A. Improving Distance Accuracy between Residents using Haversine Formula

This study proposed to improve the accuracy of distance calculation between residents in [5] and [6] using Haversine formula [13] due to its suitability in calculating the great-circle distance between two points on a sphere given their longitudes and latitudes. Current algorithms use Euclidean distance to calculate distance of the two points. However, Euclidean distance is only applicable on cartesian plane but cannot be applied on sphere shape. Since the Earth is nearly spherical, the great-circle distance formulas give the distance between points on the surface of the Earth with correctness within 0.5%. Thus, Haversine formula is chosen to replace Euclidean distance in the improved algorithms.

In the Haversine formula [13], d is the distance between two points, r is the radius of sphere 6371km, φ_1, φ_2 is latitude of point 1 and latitude of point 2, in radians, and λ_1, λ_2 is longitude of point 1 and longitude of point 2, in radians.

B. Defining threshold Value to Self-Organize Segment Density

Next, the improved algorithms define a threshold value for both Eps and MinPts parameters to self-organize segment density, which are previously user defined. In [5], Eps parameter accepts radius value and MinPts accept number of minimal points in Eps radius. Therefore, both parameters should have a threshold value to achieve self-organize segment density without any user defined parameters. Thus, this study proposed that Eps and MinPts parameter is predefined with a threshold value.

To define the threshold value of Eps, the minimum distance between neighbours needs to be defined and understood. In real life society, a neighbour means a person living nearby or next door to the person referred to. Person who stay in front could also be considered as a neighbour. Thus, the threshold distance between both neighbours are the radius of house area and the width of roads. In Malaysia, there are few road design which the shortest width among others types with only 2.75 meter width length. The width length of road is the standard length that are set under Malaysian Public Works Department (JKR) which is responsible for construction and maintenance of public infrastructure in Malaysia.

Another threshold value for MinPts is defined by the minimum number of individuals to form a group. A group is a number of people that are located, gathered or classed together. The group size of people can vary from two persons to thousands of people. A German sociologist, Georg Simmel study the connections between group size and group actions, as well as the effect of the group size on social life. For MinPts parameter, a minimum value need to be assigned in the improved algorithms to group neighbours into a cluster. According to Simmel's studies of group size, dyad, or a group of two people, is the simplest group form that may exist between individuals [14]. Thus, 2 is selected as threshold value for MinPts parameter.

C. Improving Social Connectivity through Re-Segmentation using Dunbar Number

In order to improve social connectivity, the improved algorithms implements cluster re-segmentation using Dunbar's number. Current cluster technique did not cluster based on the concern of social connectivity between neighbours. Therefore, this study proposed a method to determine group size of a cluster and re-segmentation of a cluster using Dunbar's number as shown in Figure 3.

The improved algorithms use Dunbar's Number as a threshold number for a cluster. Based on [15], a community should have a mean group size of 150 peoples to maintain a stable social relationship between each other. Thus, the improved algorithms should determine cluster group size that is larger than 150 peoples and save the cluster id into an array for later population segmentation.

Algorithm 1: Cluster_size (cluster_data)

```
For $C_i$ in clustered data set
    If $C_i$ size $C > 150$
        add $C_i$ to list
    End
End
Return list
End // cluster_size
```

Fig. 3. Determination of Cluster Group Size.

One major problem for this improvement is the algorithms will discover cluster in arbitrary shape which makes it difficult to equally segment the cluster. To solve this problem, we propose to use inverse Haversine formula to form a rectangle border to segment equally as seen in Equation 1 below:

$$\varphi_2 = a\sin(\sin\varphi_1 \cdot \cos\delta + \cos\varphi_1 \cdot \sin\delta \cdot \cos\theta)$$

$$\lambda_2 = \lambda_1 + a\tan 2(\sin\theta \cdot \sin\delta \cdot \cos\varphi_1, \cos\delta - \sin\varphi_1 \cdot \sin\varphi_2\,) \quad (1)$$

Where,

φ is latitude,

λ is longitude,

θ is the bearing (clockwise from north),

δ is the angular distance d / R; d being the distance, R the earth's radius.

The inverse Haversine formula is implemented in the improved segmentation algorithms as shown in Figure 4. In order to form a rectangle border for equal segmentation, the algorithms will first determine the distance of latitude and longitude by finding the cluster farthest north and south latitude, and east and west longitude points. Then, the number of slice is determined by dividing the total points for oversize segment with 150. The improved algorithms will then re-segment cluster group that have more than 150 people by creating a new cluster using the inverse Haversine formula.

Algorithm 2: Segmentation (over_size_cluster)

```
If $DLat > DLng$
    Distance segmented area, $DSeg = DLat / S$
    New Cluster List, $NList = inverse\_haver\sin e(NLat, SLat, DSeg)$
Else
    Distance segmented area, $DSeg = DLng / S$
    New Cluster List,
        $NList = inverse\_haver\sin e(WLng, ELng, DSeg)$
End
Return $NList$
End // segmentation
```

Fig. 4. Determination of Rectangle Border for Arbitrary Shape Cluster.

Where,

$NLat =$ Most North of Latitude point

$SLat =$ Most South of Latitude point

$WLng =$ Most West of Longitude point

$ELng =$ Most East of Longitude point

Distance of Latitude, $DLat = NLat - SLat$

Distance of Longitude, $DLng = WLng - ELng$

Number of slice, $S =$ Total point for over-size / 150

V. Implementation

This section discusses the implementation of the proposed population segmentation algorithms.

A. Experimental Setup

The experiment is setup in Taman Bukit Melaka, Malaysia. Global Positioning System (GPS) coordinates are plotted based on Taman Bukit Melaka housing area as shown in Google Maps. All GPS coordinates are represented as the local residents. Two different scenarios of resident density are simulated in this experiment; low and high densities. All GPS coordinates data are recorded in JavaScript Object Notation (JSON) format for the evaluation of improved algorithms.

B. Results and Discussions

Current work in progress is to identify low and high density residential areas. The improved algorithms consist of segmentation technique to control group size of a cluster based on pre-defined parameters and Dunbar number with average 150 persons per group. In this section, the comparison of current algorithms [5], and the improved algorithms are shown in Figure 5, Figure 6 and Figure 7. Both algorithms are implemented in Javascript programming language.

The first experimental scenario is to compare the population segmentation for a low-density neighbourhood. The result in Figure 5 shows that both current and improved algorithms produce the same segments. This indicate that both algorithms work well in a low-density neighbourhood.

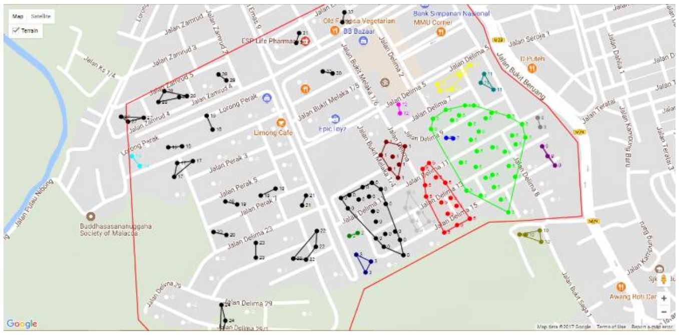

Fig. 5. Segmentation Algorithms for Low-Density Area using Current [5] and Improved Algorithms.

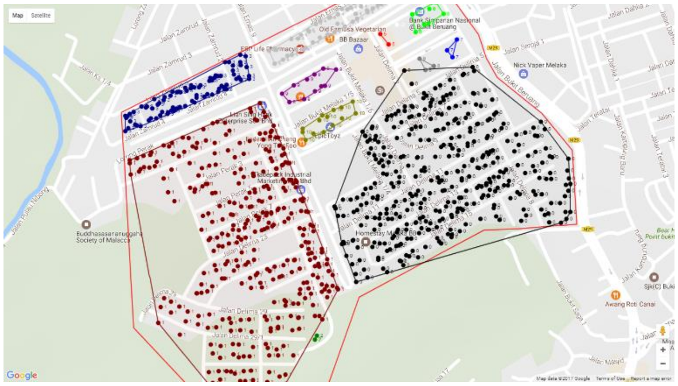

Fig. 6. Segmentation for High-Density Area using Current Algorithms [5].

The second experimental scenario is to compare the population segmentation for a high-density neighbourhood. Figure 6 shows the segmentation result for current algorithms [5] while Figure 7 shows the segmentation result for the proposed algorithms. The result in Figure 6 shows that current algorithms only cluster nearby GPS coordinates into few large segment. From the results, it can be seen that segmentation of the cluster group is not equally distributed and some of the segments are too crowded.

On the other hand, the result in Figure 6 shows that the proposed algorithms do not only cluster GPS coordinates, but also segment large group size of cluster into smaller group size of cluster with average of 150 people to maintain the stability of social relationship.

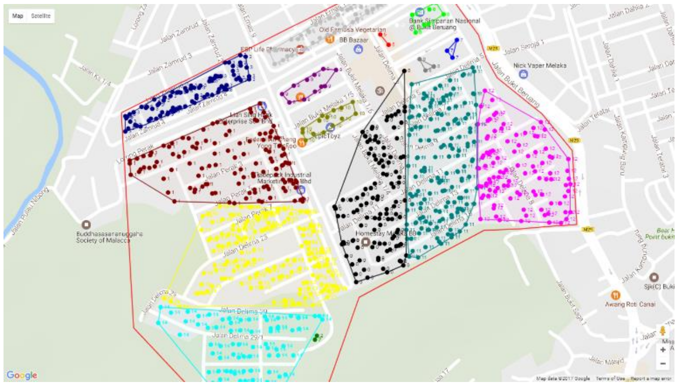

Fig. 7. Segmentation for High-Density Area using Improved Algorithms.

VI. Conclusion and Future Work

In summary, this paper propose an improved population segmentation algorithms to provide better representation and accuracy in determining the residents of geosocial neighbourhood. In this paper, we show how the current population segmentation in geosocial network neighbourhood using density-based clustering method can be extended to improve the quality of social community. Three improvements are implemented in the proposed algorithms namely improving distance accuracy between residents using haversine formula, eliminating user-defined parameter by defining threshold value to self-organize segment density, and improving social connectivity through re-segmentation using Dunbar number. The improved algorithms have achieved a great result particularly on the segmentation for high-density neighbourhood and able to segment a crowded area into smaller group size of cluster to maintain the stability of social relationship.

In geosocial network neighbourhood, identification of demographic sub-group will help to better understand the needs and requirements for the people who are related to the segment. Therefore, future extension of our work may include and put an emphasis on integration of demographic information that may help to reveal patterns or differences between groups of people who may be similar in age, gender, race, religion or socioeconomic status. The addition of demographic variables will allow the most effective use of resources in a geosocial network neighbourhood application.

Acknowledgment

A high appreciation to Center for Advanced Computing Technology (C-ACT), Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM) for supporting the work done in this paper.

References

[1] K. Hampton, and B. Wellman, "Neighboring in Netville: How the Internet supports community and social capital in a wired suburb," in City & Community, vol. 2(4), pp. 277-311, 2003.

[2] F. Wilske, "Approximation of neighborhood boundaries using collaborative tagging systems," in GI-Days, vol. 32, pp. 179-187, 2008.

[3] R. Füss, and J.A. Koller, "The role of spatial and temporal structure for residential rent predictions," in International Journal of Forecasting, vol. 32(4), pp. 1352-1368, 2016.

[4] Y. van Gennip, B. Hunter, R. Ahn, P. Elliott, K. Luh, M. Halvorson, S. Reid, M. Valasik, J. Wo, G.E. Tita, and A.L. Bertozzi, "Community detection using spectral clustering on sparse geosocial data," in SIAM Journal on Applied Mathematics, vol. 73(1), pp. 67-83, 2013.

[5] J. Shi, N. Mamoulis, D. Wu, and D.W. Cheung, "Density-based place clustering in geo-social networks," in Proceedings of the 2014 ACM SIGMOD international conference on Management of data, pp. 99-110, June 2014,

[6] D. Wu, J. Shi, and N. Mamoulis, "Density-Based Place Clustering Using Geo-Social Network Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 30(5), pp. 838-851, 2018.

[7] S. Yokoyama, Á. Bogárdi-Mészöly, and H. Ishikawa, "EBSCAN: An entanglement-based algorithm for discovering dense regions in large geo-social data streams with noise," in Proceedings of the 8th ACM SIGSPATIAL International Workshop on Location-Based Social Networks, p. 7, November 2015.

[8] K. Mumtaz, and K. Duraiswamy, "A novel density based improved k-means clustering algorithm–Dbkmeans," in International Journal on computer science and Engineering, vol. 2(2), pp. 213-218, 2010.

[9] M. Ester, H.P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Kdd, vol. 96, no. 34, pp. 226-231. 1996.

[10] A. Zhou, S. Zhou, J. Cao, Y. Fan, and Y. Hu, "Approaches for scaling DBSCAN algorithm to large spatial databases," in Journal of computer science and technology, vol. 15, no. 6, pp. 509-526, 2000.

[11] M. Ester, H.P. Kriegel, J. Sander, and X. Xu, "Clustering for mining in large spatial databases," in KI, vol. 12, no. 1, pp. 18-24, 1998.

[12] R.I. Dunbar, "The social brain hypothesis," in Evolutionary Anthropology: Issues, News, and Reviews, vol. 6, no. 5, pp. 178-190, 1998.

[13] R.W. Sinnott, "Virtues of the Haversine," in Sky Telesc., vol. 68, p.159, 1984.

[14] R.L. Moreland, "Are dyads really groups?," in Small Group Research, vol. 41, no. 2, pp. 251-267, 2010.

[15] R. Dunbar, How many friends does one person need? Dunbar's number and other evolutionary quirks. London: Faber & Faber, 2010.