Redes sociales basadas en ubicación en Buenos Aires (2009-2015)

Leonardo Salvador Rocco[1], Marcelo A. Soria[2]

[1] Maestría en Explotación de Datos y Descubrimiento de Conocimiento. Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires
2 Facultad de Agronomía. INBA-CONICET. Universidad de Buenos Aires
soria@agro.uba.ar

Resumen. El tipo de redes sociales que se basan en la ubicación de sus usuarios recibe el nombre de redes sociales basadas en localización (LBSN) y son un medio oportuno para el análisis del comportamiento espacio temporal de las personas. Utilizando datos registrados en Foursquare, aplicación para dispositivos móviles que provee a sus usuarios búsquedas personalizadas y localizadas, se construyó la red social de usuarios con actividades en lugares de la Ciudad Autónoma de Buenos Aires entre 2009 y 2015. En este trabajo se describen en primer lugar aspectos metodológicos para la recolección y procesamiento de datos de redes sociales utilizando información pública, y en segundo lugar se estudia las características estructurales de la red social que componen estos usuarios. Entre los hallazgos más importantes se encuentra una estructura fuertemente comunitaria, de mundo pequeño y con un grado neutral de afinidad selectiva, que no se enmarca en una distribución de grados de ley de potencias.

Palabras Clave. LBSN, Foursquare, redes sociales, geolocalización, Buenos Aires, análisis de redes sociales

Received January 2024; Accepted March 2024; Published May 2024

Location-based social networks in Buenos Aires (2009-2015)

Abstract. The type of social networks that are based on the location of their users is called location-based social networks (LBSN). These networks are an adequate means for spatio-temporal users' behavior analysis. Using data from Foursquare, an application for mobile devices that provides its users with personalized and localized searches, the social network for users was built based on activities within Buenos Aires Federal District between 2009 and 2015. In this paper in the first place the methodological aspects for the collection and processing of social network data using public information is described. Secondly the structural characteristics of this social network are analyzed. Among the most relevant findings, we could see that the network has characteristics of a strong communitary, small-world and neutral degree of assortativity structure. It was also found that this network does not fit the power-law degree.

Keywords. LBSN, Foursquare, social networks, geolocation, Buenos Aires, social network analysis

1 Introduction

The integration of online social networks with smart portable devices has revolutionized the way individuals communicate and share information. Location-based social networks (LBSN) rely on the location of their users and serve as a conducive medium for researching sociability at the intersection of online and offline contexts, as well as discovering mobility and spatio-temporal patterns. LBSNs provide a rich, detailed collection of data on individual behaviors, including with whom, when, and where people interact, which is crucial for understanding activity patterns in both time and space, and even the sentiments or preferences of individuals at specific moments and locations.

During the last decade Foursquare was a widely used platform designed to recommend venues (places) to a user according to the search, tastes, habits, needs, geographical location, ranking of places, and the behavior of his/her friend users. Foursquare users could register and share their geographical location by marking their presence (check-in) at a specific site (venue) physically located near where the user was. In addition to checking in at venues, users could read and post comments about their experiences, as well as establish links on the platform with other users (Fig. 1).

Figure 1. Example of places recommended by the Foursquare application based on a search and user location. In blue are the recommended places and their locations.

The aim of this study is to describe the topological characteristics of social networks formed by users of the Foursquare platform with geolocated activities within the Autonomous City of Buenos Aires (CABA). The time frame for the analysis is from December 2009 to March 2015, as access to data became much more restrictive thereafter.

Some key indicators about the use of mobile devices and the internet can be identified to better understand the extent of this application designed primarily for mobile devices. Locally, taking into account data from the World Bank [1], the proportion of households that owned mobile telephony in Argentina and used the Internet from the same was 48% for the year 2010. Although there is no data regarding the penetration of social networks in Argentina for the period analyzed in this work, Foursquare is recognized as the most widespread LBSN worldwide during the period considered in this work [2][3]. Internationally, by 2011 more than 12% of mobile phone users in the United States had used their phone for LBSN such as Foursquare or Gowalla [4].

2 Data extraction and methods

2.1 Data extraction

Data collection was carried out through the platform's API. The area of the City of Buenos Aires was exhaustively covered, collecting all the places registered on the platform, which required using official geographical information for the location of the 12,373 geographical blocks of the city. For this purpose, the “Manzanero” database [5] was used (Fig. 2).

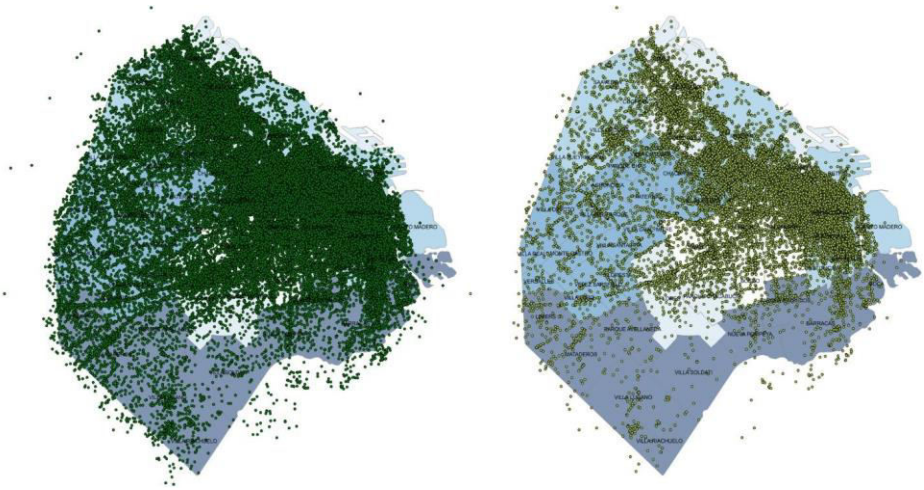

Figure. 2. Places in CABA initially collected (left) and distribution of places considered in the analysis (right).

The colored dots on Figure 2 on the right are places that have users' comments and are included in this work. As can be seen, both distributions spatially resemble the communes of CABA -administrative subdivisions of the city-, and were colored with a color scale associated to population density: lighter tone represents communes with higher density. It is observed that the analyzed places are distributed mainly in areas with high population density.

2.2 Software processes and components

This work used open-source software entirely for data collection, processing, and analysis. The combined use of QGIS and PostGIS allowed the manipulation of the Manzanero database, converting it to latitude-longitude coordinates and then transforming it into a PostgreSQL database for the necessary spatial calculations for interaction with the Foursquare API. Talend Open Studio was used for data integration from the API and process orchestration. Figure 3 shows an overview of high-level processes, data and software supports used.

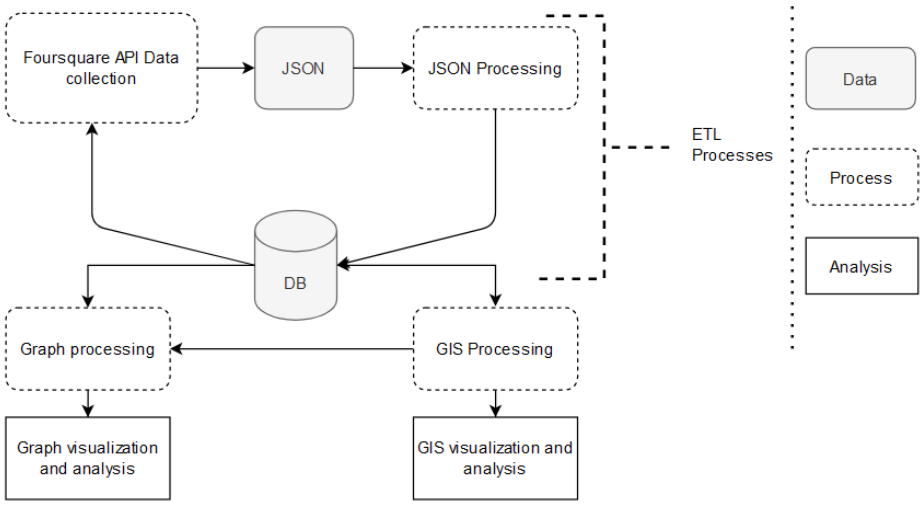

Fig. 3. High-level diagram of processes, information, and analysis performed.

Regarding the software used for social network analysis, Python was chosen for processing, SQL for querying, and Gephi (https://gephi.org/) for the network exploration and visualization. For graph and social network analysis within Python, NetworkX was mainly used. In every case, the relevant software was linked to the database developed in PostgreSQL to facilitate dynamic data processing.

2.3 Data Processing

Once all the venues in the city existing on Foursquare were identified, the collection of comments left on them, the people who had left those comments, and their links with other users was carried out. Of the total number of people who left comments in CABA, outliers were univariately excluded based on their score transformed to a Z scale, discarding 0.5% of users furthest from the average number of comments. Bots and advertising agents on Foursquare, as well as other social networks, have been detected and analyzed in different studies [6], and are widely extended [7].

After pruning users with extreme values, 31,385 remained, with whom the graph of relationships between venue commentators in CABA was built. The data cleaning process also involved identifying the unique users commenting on the places, deduplicating ties between users. Also, the main network components were identified, so 5,234 disconnected small components were discarded.

Although it was not possible to attribute a geographical origin to each of the users considered in the graph, it was possible to use the textual content of the comments they entered for CABA venues to be classified according to their language using a classification model. For practical purposes and to facilitate analysis, it was decided to divide the users into two groups: Spanish speakers (78.2%) and non-Spanish speakers (21.8%).

2.4 Graph Construction

Given the characteristics of user relationships enabled on Foursquare, an undirected and unweighted graph was used to represent the social interactions. Figure 4 shows a schema depicting the users' roles and their relationship to the venues, showing which users (nodes delimited by a continuous circle) and relationships are considered in the graph. The same person can have comments in different places, and a friend of a commenter can be: 1) a commenter at the same venue, 2) a commenter at a different venue, or 3) a direct first-degree friend of a commenting person even though they have not commented at any venue in CABA. The third case has been omitted from this work.

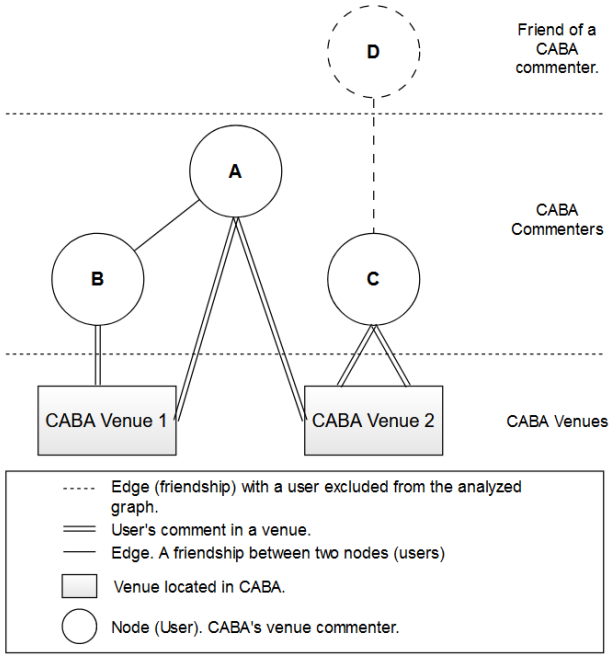

Figure. 4. Graph Scheme in relation to user's relationships, comments and venues in CABA.

3 Results and Discussion

3.1 General Characteristics of the Graph

Real social networks are not random, but they are also not highly regular; they have unique characteristics where order and structure coexist with disorder to different extents. In the following sections, some of the properties of the graph that will allow the identification of the general structure, communities, and characteristics

of this social network are analyzed. Although there are numerous centrality metrics, this work concentrates on the most used centrality indices in the field [8]. It had been observed that many graph metrics are correlated with each other and partially describe the same properties, so only some of them are sufficient for practical purposes of social network analysis [9].

Table 1 lists the main topographical features of the graph of interactions among Foursquare users in Buenos Aires. While the heatmap on Fig. 5 shows the correlation between pairs of centrality measures. The high r values are expected, have been previously reported, and are due to the high levels of association that occur in data organized in networks.

Table 1. Graph Characteristics. NetworkX was used for metric calculations.

Metric	Value
Nodes	31385
Edges	229224
Connected components	1
Mean centrality degree	14.6
Density	0.000465
Mean clustering coefficient	0.1304
Transitivity	0.0134

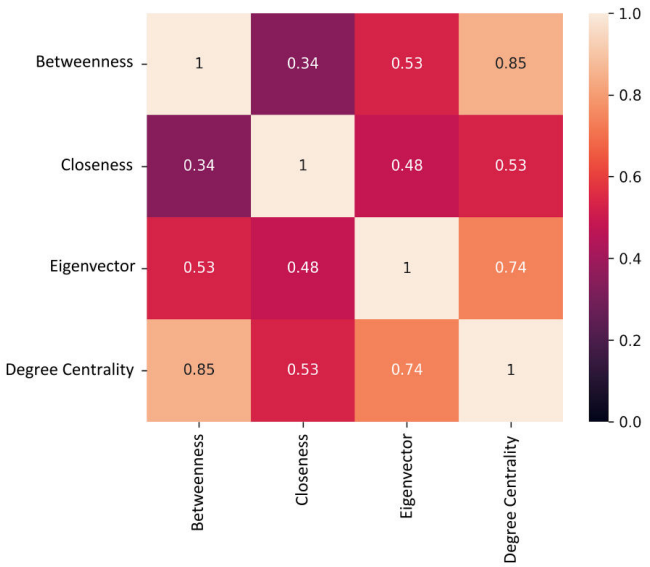

Fig. 5. Pearson's r correlation values between centrality measures represented in a heatmap.

The average centrality degree is somewhat lower than in other studies [10][11] on the same platform, but the differences may be due to the way the analyzed graph was constructed, restricted only to users with comments in CABA. Regarding other network structure measures, as can be observed in Figure 6, it was found that the clustering coefficient decreases as the degree of the nodes increases, indicating that nodes with higher degrees may be using the platform less oriented towards relating to real friendships, and instead establishing indiscriminate relationships.

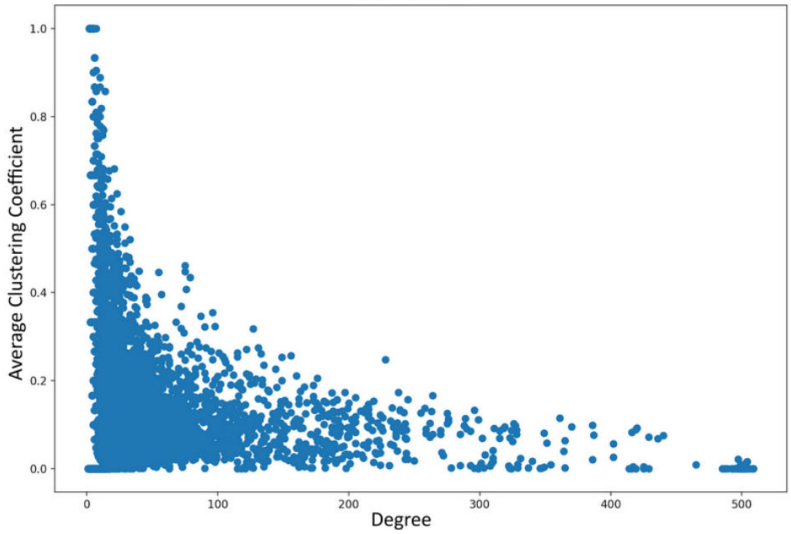

Figure 6. Average Clustering Coefficient (y-axis) according to degree (x-axis).

3.2 Resilience

Resilience measures the impact on connectivity and cohesion of the network when one or more nodes are removed. Most social networks are robust to the removal of nodes, but not so much to the removal of high-degree nodes. Resilience is an important property as it measures the redundancy of paths between different nodes and the interconnectivity of the network, since when an edge is not available, connectivity still exists in the network. Due to the computational complexity required to measure the number of nodes or edges needed to turn the connected component into a disconnected component, the calculation of density was used instead, which measures the fraction of possible edges that exist. The greater the number of edges that exist in a network, the more redundant paths between nodes. In this case, the density calculation was 0.000465 (out of a maximum value of 1, where all nodes are interconnected) indicating a very sparse network.

This density value is reasonable considering that the graph contains only people who have left comments in CABA and the links between them, but not the totality of relationships that the nodes could have outside the city.

3.3 Assortativity

Real social networks often present characteristics of assortativity among their nodes based on their attributes (age, gender, ethnicity, etc.) as well as structural or intrinsic attributes of the nodes (degree of closeness, etc.). This assortativity, also called homophily in sociology, measures the tendency of people to relate to similar people.

Initially, to understand degree assortativity, the relationship between the degree of the nodes and the average degree of their neighbors was analyzed (Figure 7).

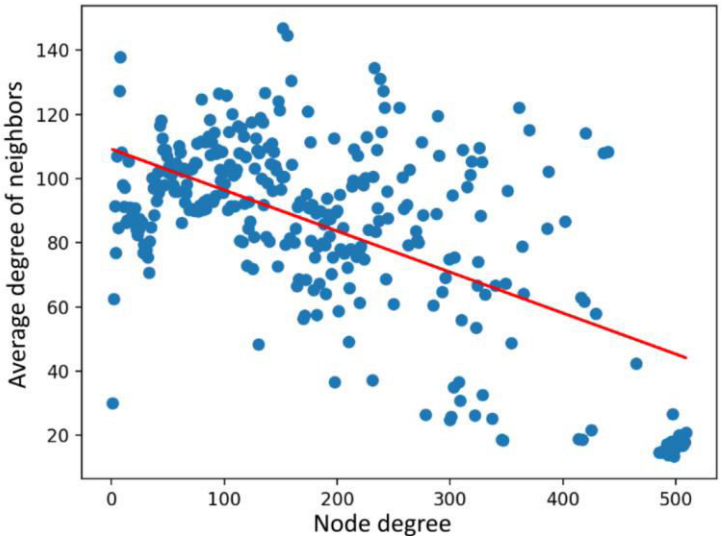

Figure 7. Association between the node degree (x-axis) and the average degree of neighbors (y-axis). Red line: Simple linear regression fit.

At first glance, there is a negative assortativity trend with a Pearson's r value of 0.59, where higher-degree nodes have neighbors with a lower average degree. A similar pattern emerges when analyzing the degrees of connected node pairs. In Fig. 8, it is observed that as the degree of a node increases, it tends to relate more to nodes of a lower degree. This pattern is confirmed with the Newman coefficient, which was -0.11, indicating that the nodes in the network are at an intermediate or neutral level of assortativity with a tendency towards inverse assortativity. This value is consistent with some studies [12] that propose that real networks have a low degree of assortativity and 58% of them are classified in a "neutral" degree (between -0.19 and

+0.19). This value is also consistent with the degree found in other online social networks like YouTube [13].

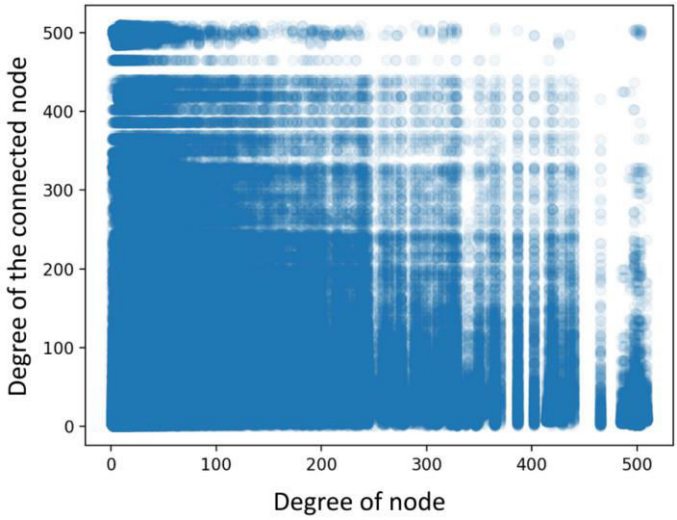

Figure 8. Degrees of nodes (both) connected by edges.

To get a deeper insight into assortativity we analyzed two attributes that were significant to the demographic aspect of the social network: declared gender and user language classified based on their comments. For the case of user language, we observe in Table 2 that 76.9% of the nodes relate to nodes of the same language, which the affinity coefficient confirms with a positive value of 0.11. While 73% of the edges connect Spanish-Spanish speakers, only 3.9% do so among non-Spanish speakers. Considering the total number of edges established by Spanish-speaking people, Spanish speakers relate to each other 86% of the time. Whereas considering the total relationships established by non-Spanish speakers, 74% of the time they do so with Spanish speakers.

Table 2. Assortativity Matrix by Language Classification

Edges	Non-hispanic	Hispanic	Subtotal
Non-hispanic	3.9%	11.5%	15.4%
Hispanic	11.5%	73.0%	84.6%
Subtotal	15.4%	84.6%	100%

Regarding the fraction of nodes that relate to the same gender, i.e., the proportion of edges that connect women users (10%) plus the edges that connect male users (45.9%), they add up to just over 55% of the total edges. Therefore, it can be said that

the social network is not homophilic from a gender perspective, as confirmed by calculating the assortativity coefficient for gender, which is just -0.0078.

3.4 Fit to a Power Law

Networks can be classified based on their degree distribution. In the original premise of Barabási and Albert [14], they stated that a large number of real networks exhibit a power-law distribution, with a majority of nodes having a low degree and a few having a high degree [15]. This degree distribution represents one of the three main properties commonly found in social networks, along with short distances and a high degree of clustering (described in the section on small-world networks). The power law is not only the least intuitive of these properties but also the most studied and debated since its description at the end of the last century [16]. It should be noted that, while it was commonly understood that social networks usually exhibit this property, this assumption is currently under question [17].

Despite an initial analysis shows that the degree distribution in the network resembles a power law, the goodness of fit of alternative distributions, such as log-normal or exponential, was also tested. For this purpose, the power law distributions were graphically compared against log-normal and exponential distributions, as seen in Figure 9.

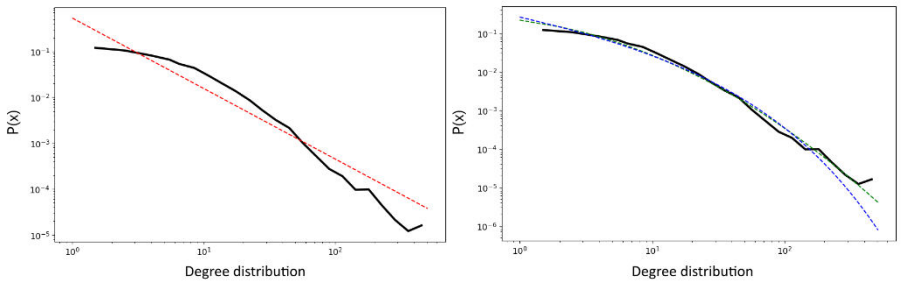

Fig. 9. Log-log plots for probability density functions P(x) fits to degree distribution. Left panel: Power law (red line) fit to the degree distribution (continuous line). Right panel: Log-linear (green line) and narrow exponential (blue) fits to the degree distribution.

Although difficulties in distinguishing log-normal distributions from power laws are common [18], a comparative evaluation using the likelihood ratio was carried out to identify which of the two distributions fits better. The log-normal (also called log-linear) distribution fit better. The likelihood ratio value between the power law and the log-linear distribution was -48.408. This number is negative as it indicates that it is more likely for the distribution to fit a log-linear model. The p-value significance was < 0.001, leading to the rejection of the power law as the best theoretical distribution fitting the degree distribution of the analyzed graph.

The assumption of a power law fit in social networks is currently being questioned, and specific LBSN studies have also characterized that the degree distribution partially fits the log-linear distribution [19], while others describe it as a power law [20].

3.5 Small World

To determine if this social network has a small-world structure, with short distances and a high degree of clustering, an experiment was conducted using seven rewiring probabilities (p), using a Watts-Strogatz-based graph generation algorithm. The Watts-Strogatz algorithm starts from a ring-shaped network (which ensures a high clustering coefficient value) and randomly rewires some edges based on a given probability p. This rewiring is done between distant nodes to ensure a decreasing average shortest path length, reaching its highest randomness in node wiring when p=1.

During the experiment, the topological characteristics of the number of nodes (31385) and their average degree (14) as seen in the analyzed graph were maintained. The experiment results are shown in Table 3.

Table 3. Experiment results for randomized graphs using Watts-Strogatz

p	Average Clustering	Average shortest path length
0.000001	0.6923	1066.26
0.00001	0.6922	955.34
0.0001	0.6921	210.59
0.001	0.6902	40.03
0.01	0.6721	11.48
0.1	0.5078	5.81
1	0.0004	4.23

It can be observed that as the rewiring probability p increases, the clustering coefficient decreases much more slowly than the graph's average shortest path length, with the clustering value dropping only in graphs with very high p values. Figure 10 clearly shows the behavior of both measures using a logarithmic scale for p.

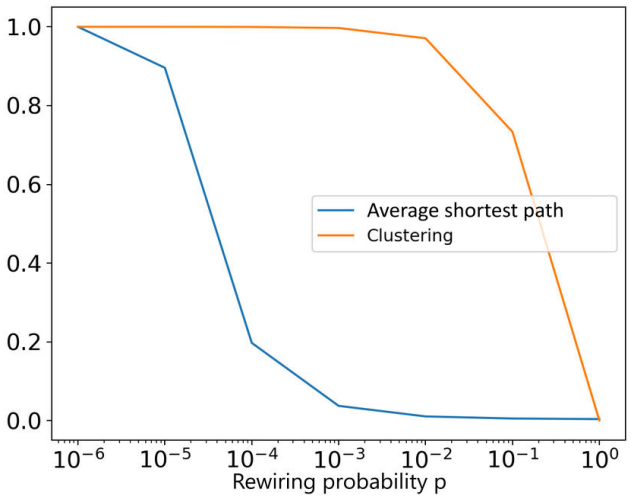

Fig. 10. Clustering coefficient and average shortest path (y-axis) according to log(p) (x-axis) for the experiment with randomized graphs using Watts-Strogatz.

Considering that the analyzed graph has an average clustering coefficient of 0.148 and an average shortest path length value of 4.23, we can conclude that it is not at the extreme where rewiring is completely random, thus classifying this network as a small-world spatial structure.

Similar values have been found for social networks like Facebook [21], where the average distance is 4.7, and other online social networks [22] or LBSNs like Foursquare, where the average shortest path length distance is less than 6 and the average clustering coefficient is between 0.18 and 0.26 [23]. These properties are also in line with the small-world nature of LBSNs [24] studied.

3.6 Community Structure

Most social networks have a community structure [25], composed of groups, also called clusters, meaning that there are groups of nodes densely connected to each other and loosely connected to other groups. For the identification and partitioning of communities, the Louvain heuristic was used, based on optimizing the graph's modularity. Although the heuristic used is not deterministic, it provided similar results in different executions for segmenting the graph into communities. The modularity resulting from the partition of the graph was 0.52. This value indicates a high community structure, where edges tend to be incident among nodes of the same community. Out of a total of 37 identified communities, they have an average of 848 nodes, with a maximum of 4229 and a minimum of 3 nodes. The Kernel density diagram of the community size distribution can be observed in Figure 11.

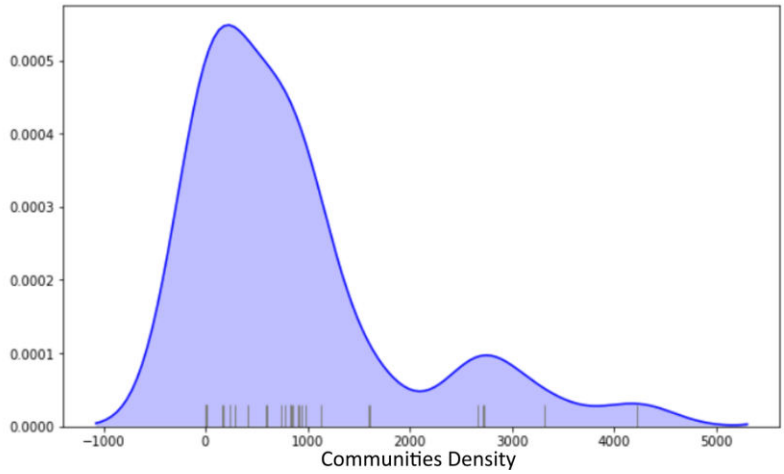

Fig. 11. Kernel density estimation of community size

As an example, Figure 12 shows the graph of one of the communities, of fewer than 200 nodes, characterized by having 55% of users classified as non-Spanish speakers (in pink), deviating from the average language distribution found in other communities.

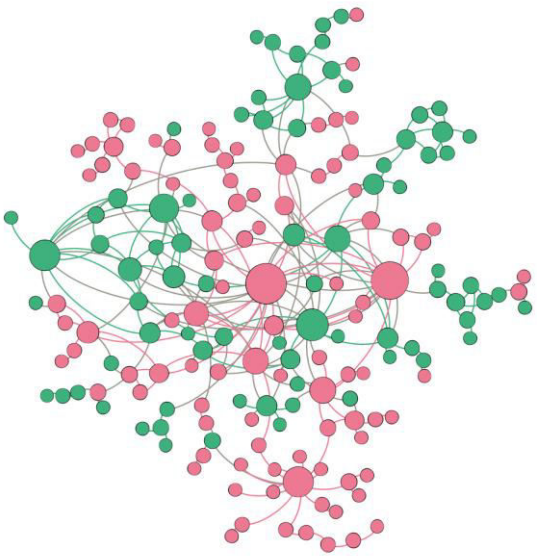

Fig. 12. Graph of selected community. Spanish speakers in green. Non-Spanish speakers in pink. Size by degree centrality.

This community gives us an indication that people seem to relate to each other according to their language, but this does not seem to be largely extendable to the entire graph. This was addressed in more detail in the section on assortativity. However, when plotted on the city map in Figure 13, a spatial indication was found that non-Spanish speakers frequent more concentrated sectors of the city than Spanish speakers, a hypothesis that will be addressed in more depth in another work.

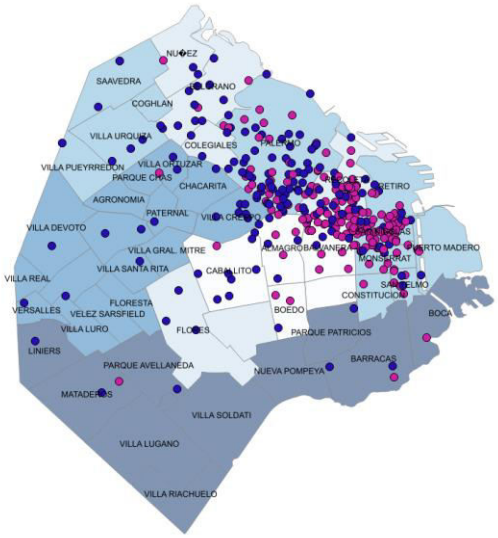

Fig. 13. Geographic centroid of comments from each member of the blue and pink community. The composition of Spanish speakers for the first is 55%, while for the second, it is 15%.

Finally, while the identification of communities is useful for discovering underlying properties in the social network that may be exclusive to that community or then extend to the entire network, it is also possible to identify how each of these communities relates to each other and understand the centrality of each of them for different interests such as message diffusion or user segmentation.

4 Conclusions and Future Work

This work uses social network analysis (SNA) to characterize the structural aspects of a geolocation-based social network in the Autonomous City of Buenos Aires. Additionally, the methodology used for data collection, processing, and analysis using open-source software and publicly available data sources in the state and online

platforms is described. Regarding the characteristics of the analyzed social network, the following structural aspects and properties were found:

- It is concluded that the power law is not the best theoretical distribution fitting the degree distribution of the analyzed graph, finding the log-linear distribution more suitable, in line with the most recent academic literature on the subject.

- It is verified that the network presents a low degree of negative assortativity, within the parameters presented by other similar social networks.

- Similarly, the small-world structure is verified, coinciding with other studies.

- Lastly, a high community structure is found, and the structural characteristics of different communities present are described.

From an applied perspective, this study shows the usefulness of the graph approach to analyze tourist behaviors in CABA. For instance, it helps identify areas of the city gaining tourist interest, segmented by their region of origin. Also, the degree of connectivity found suggests that through proper targeting of advertising and marketing actions, diffusion levels comparable to those obtained with more expensive campaigns seeking high saturation can be achieved. In a future work, the spatial component of the social network will be analyzed, focusing on the links between users in relation to spatial distance and the structure of their social network. Although there is agreement in the literature that the probability of a link between two users decreases as the geographic distance between them increases [26][27], the characteristics and importance of such a relationship are still being debated according to different types of networks, platforms, and countries, making it important to analyze what characteristics it acquires in the local context.

References

1. World Bank. Information, Communication Technologies, & infoDev (Program). Information and communications for development 2012: Maximizing mobile. World Bank Publications. (2012)
2. Preoţiuc-Pietro, D., Trevor, C. Mining user behaviours: a study of check-in patterns in location based social networks. In Proceedings of the 5th annual ACM web science conference. (2013).
3. Zhang, K., Konstantinos, P., and Theodoros, L. Effects of promotions on location-based social media: evidence from foursquare. In International Journal of Electronic Commerce 22.1. (2018).
4. Zickuhr, K. Three-quarters of smartphone owners use location-based services. Pew Internet & American Life Project. (2012)
5. Gobierno de la Ciudad de Buenos Aires.”Manzanas de la Ciudad de Buenos Aires”. http://data.buenosaires.gob.ar/dataset/manzanas, last accessed 12, Jan. 2015.
6. Hussain, A., & Keshavamurthy, B. N. Analyzing Online Location-Based Social Networks for Malicious User Detection. In Recent Findings in Intelligent Computing Techniques. Springer, Singapore. pp. 463-471. (2019)

7. Vasconcelos, M. A., Ricci, S., Almeida, J., Benevenuto, F., & Almeida, V. Tips, dones and todos: uncovering user profiles in foursquare. In Proceedings of the fifth ACM international conference on Web search and data mining, pp. 653-662). (2012)
8. Zhang, J., & Luo, Y. Degree centrality, betweenness centrality, and closeness centrality in social network. In 2017 2nd International Conference on Modelling, Simulation and Applied Mathematics (MSAM2017). Atlantis Press. (2017).
9. Zinoviev, D. Complex network analysis in Python: Recognize-construct-visualize-analyze-interpret. Pragmatic Bookshelf. (2018)
10. Agryzkov, T., Martí, P., Tortosa, L., & Vicent, J. F. Measuring urban activities using Foursquare data and network analysis: a case study of Murcia (Spain). International Journal of Geographical Information Science, 31(1), pp. 100-121. (2017)
11. Ferreira, A. P. G., Silva, T. H., & Loureiro, A. A. F. Beyond sights: Large scale study of tourists' behavior using foursquare data. In 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 1117-1124). (2015)
12. Meghanathan, N. Assortativity Analysis of Real-World Network Graphs based on Centrality Metrics. Computer and Information Science, 9(3), pp 7-25. (2016)
13. Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. Measurement and analysis of online social networks. In Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, pp. 29-42. (2007)
14. Barabási, A. L., & Albert, R. Emergence of scaling in random networks. science, 286(5439), pp. 509-512. (1999)
15. Oliveira, M., & Gama, J. An overview of social network analysis. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. (2012)
16. Muchnik, L., Pei, S., Parra, L. C., Reis, S. D., Andrade Jr, J. S., Havlin, S., & Makse, H. A. Origins of power-law degree distribution in the heterogeneity of human activity in social networks. Scientific reports, 3(1), pp 1-8. (2013)
17. Broido, A. D., & Clauset, A. Scale-free networks are rare. Nature communications, 10(1), pp 1-10. (2019)
18. Malevergne, Y., Pisarenko, V., & Sornette, D. Empirical distributions of stock returns: between the stretched exponential and the power law?. Quantitative Finance, 5(4), pp. 379-401. (2005)
19. Scellato, S., & Mascolo, C. Measuring user activity on an online location-based social network. In 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 918-923). (2011)
20. Wei, W., Zhu, X., & Li, Q. LBSNSim: Analyzing and modeling location-based social networks. In IEEE INFOCOM 2014-IEEE Conference on Computer Communications. pp. 1680-1688). (2014)
21. Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. The anatomy of the facebook social graph. arXiv preprint arXiv:1111.4503. (2011)
22. Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. Measurement and analysis of online social networks. In Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. (2007)
23. Scellato, S., Noulas, A., Lambiotte, R., & Mascolo, C. Socio-spatial properties of online location-based social networks. ICWSM, 11, pp. 329-336. (2011)
24. Leskovec, J., & Horvitz, E. Planetary-scale views on a large instant-messaging network. In Proceedings of the 17th international conference on World Wide Web, pp. 915-924. (2008)
25. Girvan, M., & Newman, M. E. Community structure in social and biological networks. Proceedings of the national academy of sciences, 99(12), 7821-7826. (2002)
26. Levy, M., & Goldenberg, J. The gravitational law of social interaction. Physica A: Statistical Mechanics and its Applications, 393, 418-426. (2014)

27. Backstrom, L., Sun, E., & Marlow, C. Find me if you can: improving geographical prediction with social and spatial proximity. In Proceedings of the 19th international conference on World wide web (pp. 61-70). (2010)