doi: 10.36689/uhk/hed/2020-01-062

The Impact of Anonymization on the Geosocial Network Metrics Used in Socio-economic Analysis

Jana MEDKOVÁ

University of Hradec Králové, Hradec Králové, Czech Republic; jana.medkova@uhk.cz

Abstract: The expansion of mobile devices equipped with GPS (Global Positioning System) locators corresponds to the development of the highly customized location-based services including geosocial networks. The usage of customized location-based services positively effects many aspects of users' daily routines from travelling to choosing the best restaurant. On the other hand, providing customized services relates to collecting and storing large amount of users' information and gives rise to many privacy-preserving issues. In this paper, we discuss the privacy concerns connected with publishing geosocial network datasets and the impact of the anonymization on the utility of the geosocial network dataset. Considering the importance of the geosocial network for the socio-economic analysis, we put arguments for the importance of geosocial network anonymization before exploiting the dataset. We apply the clustering anonymization methods according to geographical coordinates and the values of location entropy on the real-world data to prevent the location privacy leakage. Afterwards, we compare the network metrics in the original and anonymized real-world datasets and measure the impact of the anonymization on the metric values.

Keywords: location privacy; geosocial network; anonymization; socio-economic analysis, location entropy

JEL Classification: C88; M31; O35

1. Introduction

Privacy is a concept of preventing the sensitive data and information from an unapproved access. Due to the steady growth of the number of Internet users privacy preserving methods have become a deeply investigated field of study. The rapid development of mobile devices like smartphones or tablets enables the progress of various internet services. Enjoying the benefits offered by the service providers is usually connected with providing the service providers with personal information. Since the quality data have become a highly valued commodity, there has been a demand for sharing data among different subjects. However, publishing collected datasets give rise to many privacy issues like the problem of identity, attribute or inferential disclosure, as noted by Fung et al. (2010).

Location privacy is a special type of information privacy which considers the right of the individual to decide when, how and with what share the location information about themselves. The main aim of the location privacy is to control of location information about the user, as described by Duckham and Kulik (2006). Nowadays, a plenty of internet services, called location-based services, requires the user's location information to provide the user with the service. Gaining the weather forecast with the weather applications, finding the nearest restaurant or using the navigation application requires the user to share her or his location information with the service provider. The terms of services for the applications should contain the information how and for how long the location information will be stored, whether it will be shared with other subjects. However, the users are not able to check whether the providers keep the storage time limits and sometimes they are not even aware that their location information is stored at all, as pointed out by Keßler and McKenzie (2018).

Online social networks (SNs) are Internet services enabling users and organizations to communicate and share various information content with each other. Links between entities represents by the relationships between SN users. Geosocial networks (GSNs) are social networks enhanced with a location information. When user shares her or his current location with the GSN provider, then the location information can be distributed to other users and may lead to highly customized social applications such as real-time discovering friends in the neighborhood, recommending services in the current location or highlighting nearby points of interests, as noted by Gambs et al. (2011). On the other hand, GSNs can be perceived as the special branch of location-based services, since sharing among friends is an additional feature to providing information based on user's current location.

GSNs has become a very important source of information with unlimited access and a simple communication tool for tourism participants. Instead of the traditional information centers, the information available on GSNs are up-to-date and come from all tourism participants including tourists themselves. Information from tourists that had visited a location and consume services in the neighborhood, have significant impact on the future visitors at the locations. Tourists can make quick and competent choices while deciding which service to buy. Furthermore, recommendations about travel-related services influence indirectly the future improvement of the services themselves.

Except the possibility of sharing information, GSNs provide connection between common users and registered businesses, called venues. Hence, GSNs play an important role in geomarketing strategies. Palos et al. (2018) presented that the geomarketing strategies, usually included in mobile marketing, were based on analyzing the behavior of consumers according to their location. The location information are later used to the business promotion. For instance, users are provided with different advertisements according to their location. Thus, GSN users usually benefit from checking-in at the venue location. They gain coupons, discounts and special offers for revealing their location to the provider, who transmitted it to the venues in a real-time, as mentioned in Palos et al. (2018). Indisputably, GSN contributes in higher profitability of participating businesses. On the other hand, applying geomarketing entails the companies to register their own locations in the network, loading a various content such as up-to-date photographs and videos and implement modern trends and technologies. Furthermore, providing location-based services relates to collecting and storing large amount of users' information, which has to be protected from the unauthorized access. Since data indicating customers' movements, shopping habits or behavior are very valuable for both academic and marketing researches, providers can also decide to share their datasets with another subject.

GSNs have a great potential to be used in various socio-economic analysis. Zhou et al. (2017) presented how geosocial network data could be used to quantify the impact of cultural investment on the urban regeneration process and predict attached socio-economic deprivation changes. They exploited 4 million transition records for 3 years in London from the popular GSN Foursquare and used the network metrics, average clustering coefficient and the centrality, to estimate the likelihood of local growth in response to cultural investment. Then, the findings were used in supervised learning models to deduce socio-economic deprivation changes in London. They proved that the geosocial network data become a powerful tool in social-economic analysis.

Zhou et al. (2017) exploited user mobility records and venue information shared from Foursquare (2020). As stated in the privacy policy of Foursquare, available at Foursquare Labs (2020), the company share only aggregated or anonymous data to other third parties. The anonymization was probably used before the dataset was shared, since the pure aggregation of visits may leak sensitive information about a single user from the data in the used representation of GSN.

Zhou et al. (2017) represented GSN dataset as a spatial network of locations where two locations were connected iff a GSN user passed from one location to the another one in a certain time period. It is a directed graph where nodes represent the locations and edges represents the users' transitions between the locations. The weight of the edge corresponds to the number of transitions made by all users between the two locations. Edges with the very small weights indicate that there are only a few users who passes between the corresponding pairs of locations. The weight equaling to 1 implies the transition of the only user. When this condition meets the background knowledge of the adversary about the visited location of the target user, then the connected location is revealed. Moreover, the user's trace can be compiled with the certain probability depending on the weight of the following edges.

Anonymization enables providers to publish their dataset or share it with the other subjects while preserving privacy of individuals being involved in the dataset. The aim of the anonymization is to modify the original data in the way to prevent the attacker from attaching the records in the dataset with the individual who is related to them. Obviously, the modification affects the data utility and the range of the modification should be as small as possible. However, the anonymized data still has informative value for further data mining and analysis.

Narayanan and Shmatikov (2009) proved that the simplest anonymization method, removing the identifiers of individuals, is proved not to be sufficient for preserving the privacy. The combination of other record in the dataset may identify the target individual in the dataset even without his or her identifier. Methods for the anonymization of relational datasets differs from the anonymization methods addressing the problem in the SNs and the GSNs. While anonymizing relational datasets corresponds to modifying records, anonymizing SNs requires modifying the corresponding graph structure and node attributes, if included. Additionally, the presence of location information attribute in GSNs demand an even more specific approach for anonymizing GSN datasets.

When the further analysis of the anonymized GSN data requires preserving the location traces of users, then the anonymization approach should include the location privacy protection mechanism protecting the users' location privacy, which were studied by Shokri et al. (2011). The dataset is vulnerable even without the temporal information, since the background knowledge about the visited locations may lead to the successful re-identification attack and leaking the user's identity from the data, as proved by Masoumzadeh and Joshi (2011).

In this paper, we demonstrate how the anonymization effects the network metrics, average clustering coefficient and the ratio of the indegree and outdegree centrality, which was shown to have a meaningful value for socio-economic analysis by Zhou et al. (2017). We compare the relative difference of the metric values in the original and anonymized data. We use the same representation of GSN dataset as Zhou et al. (2017) and exploit the data from the real-world dataset Gowalla that was collected by Cho et al. (2011). We apply the hierarchical clustering method according to the geographical coordinates of locations to cluster the location from GSN into regions. Afterwards, the locations in one region are additionally clustered into subclusters using the hierarchical clustering according to the location entropy values. We examine the impact on both methods on the metric values and demonstrate that the usage of location entropy in the clustering method highly improve the relative difference between the original and anonymized values of the examined metrics and thus preserve the data utility in the anonymized data.

2. Methodology

In this section, we formalize the problem addressed in the paper, introduce the dataset and the network metrics investigated in our experiments and describe in detail the methods used for the anonymization.

The anonymization changes the structure of the graph representing the real data. Hence, it influences the network metrics as well. Clustering anonymization method is often used when the location-based data are anonymized, as in the research performed by Masoumzadeh and Joshi (2011). The locations can be clustered into large regions according to their geographical coordinates.

At first, we introduce our approach. The detailed descriptions are added in the following subsections. Using the exact locations from the sample of the Gowalla dataset, we made the graph G, where nodes represented the exact locations. For every location in the data sample, we computed the entropy location, the metric which would be later used in the anonymization method. Then, we applied the hierarchical clustering method on the same set of locations and obtained the clustered regions. Afterwards, we made the graph G_H, where nodes represented the regions, instead of locations. Then, we applied the additional entropy-based clustering on the regions. Hence, some regions are split into several subregions. After that, we made the graph G_E, where node represented the subregions. We measured the values of network metrics, average clustering coefficient and the ratio of indegree and outdegree centrality, in all three graphs and compute the relative difference between the metric values of G_H and G and the relative difference between the metric values of G_E and G. Our research goal is to answer the following research questions:

1. How does the clustering anonymization methods according to the geographical coordinates effect the values of the examined network metrics? What is the relative difference between the values of network metrics in G_H and G?

2. How does the additional entropy-based clustering effect the metrics measurement? Does the entropy-based clustering reduce the relative difference between the values of network metrics in the original and anonymized graph?

2.1. Data representation

Gowalla was a geosocial network, where users shared their locations by checking-in. Cho et al. (2011) collected a total of 6 442 890 check-ins of 196 591 users over the period of February 2009 – October 2010. We examined the sample of Gowalla dataset related to 411 user which contained over 50 458 locations and 123 548 user transitions. The transition is defined as the successive pair of check-ins created by users. Formally, the Gowalla dataset was represented as a directed graph $G=(V,E)$, where the set of nodes $V=\{v_1, \dots, v_n\}$ represented the locations and the set of edges E was composed of pairs of locations that had at least one transition between each other. The weight of the edge $e(v_i,v_j)$ equaled to the number of transitions from the location v_i to the location v_j.

After the application of the hierarchical clustering, the anonymized data sample was represented as a directed graph $G_H=(V_H,E_H)$, where V_H represented the clustered regions and E_H represented the transition between the regions, instead of locations. Similarly, the weight of an edge equaled to the number of transitions between regions.

Similarly, we composed the graph $G_E=(V_E,E_E)$ representing the data sample after the application of the entropy-based clustering method.

2.2. Network metrics

This research focused on the same network metrics as was addressed by Zhou et al. (2017), the average clustering coefficient of ACC and the ratio of indegree and outdegree centrality IOR_i. Indegree centrality IC_i represented the number of in-flow transitions that the node v_i received. It was computed as the sum of the weights of the incoming edges. Similarly, the outdegree centrality OC_i represented the number of out-flow transitions that the node v_i received. The ratio IOR_i was defined by Zhou et al. (2017) as follows:

$$IOR_i = \frac{IC_i}{OC_i}.$$

For the purpose of the comparison, we also defined the average ratio of the indegree and outdegree centrality $AIOR$ as the average of IOR_i for the entire graph:

$$AIOR = \frac{1}{n} * \sum_{i=1}^{n} IOR_i$$

, where n was the number of nodes in the graph. The local clustering coefficient of the node v_i, denoted by CC_i, described the connectivity of the nodes in its neighborhood of the node v_i. It was defined by Zhou et al. (2017) as follows:

$$CC_i = \frac{Li}{ki*(ki-1)}$$

, where L_i was the number of edges between the k_i neighbors of the node v_i. The average clustering coefficient was the mean of the CC_i over all nodes of the graph:

$$ACC = \frac{1}{n} * \sum_{i=1}^{n} CC_i.$$

Since we measured the metrics in three graphs, the metric computed in the graph G were denoted by $ACC(G)$ and $AIOR(G)$.

2.3. Clustering anonymization methods

We used the hierarchical clustering method with the average linkage, which was described in detail by James et al. (2013), to cluster the locations according to their geographical coefficients. Locations, which were geographically closed enough to each other, were grouped together and created a geographic region. The level of the anonymization depended on the height of the cut of the corresponding dendrogram. For instance, if the height of the cut equaled to 10 kilometers, then the distance between all pairs of locations in one region was less than 10 kilometers. Hence, the locations within 10 kilometers were indistinguishable in the anonymized data. However, clustering based only on the geographic closeness may cause a significant information loss, since locations that are geographically close to each other might have very different location entropy.

Scellato et al. (2011) described the location entropy as a metric for measuring the popularity of locations in GSN. It expresses the possibility whether users who visited the particular location will have a social tie with each other in the future. Users who visited the location with low entropy are more likely to become friends in the GSN than user who visited the same locations with higher entropy. Furthermore, low entropy locations are usually places with significant importance for their visitors, for instance home places or work offices, as stated by Scellato et al. (2011). On the other hand, high entropy locations are likely to be public places, such as coffee shops or railway stations. We omit the formal definition of the location entropy metric with the reference to Cranshaw et al. (2010).

After the locations were clustered into regions according to their geographical coordinates, then all locations belonging to the same region were clustered into subclusters according to their location entropy. To do the entropy clustering we used the hierarchical clustering method with the complete linkage. The crucial parameter in the method was again the height of the cut of the corresponding dendrogram, which became the input parameter of the implemented algorithm. Hence, the final subclusters consisted of geographically closed locations with similar entropy values.

2.5. Relative difference measurement

The relative difference between the values m_1 and m_2, where m_1 is the controlled value, was defined by Kušnerová et al. (2013) as follows:

$$RD(m_1, m_2) = \frac{|m_1 - m_2|}{m_1} * 100\ (\%).$$

3. Results

All experiments were performed on a PC running Windows 10 operating system with 16 GB RAM and 3.2 GHz processor. We compiled the procedures described in the previous section into an algorithm which was implemented in Matlab 9.6.0.1214997 (R2019a).

Since the experiments were executed on a single data sample and the various height of the cut in the hierarchical clustering method did not influence the runtime, the runtime of one run of the algorithm did not vary for the different values of parameters and was about 6 minutes for all parameter values.

Table 1. Number of clusters and subclusters corresponding to several C_H values. The number of nodes $|V_H|$ in the graph G_H corresponds to the number of clusters. Similarly, $|V_E|$ corresponds to the number of subclusters. The entropy cut-off C_E=0.5.

C_H (km)	$\|V_H\|$	$\|V_E\|$
0.6	9 475	12 512
3.3	2 837	4 751
10	1 381	3 113
44.5	530	2 000

The input parameters evaluating during experiments were the height of the cut-off point for the clustering according to geographical coordinates C_H and the cut-off point for the clustering according to location entropy C_E. During each evaluation of the algorithm one parameter was fixed and another one took its value from its domain. The values of C_H varied from *0.6* to *55.6* kilometers, while the values of C_E varied from *0.25* to *1.75* (nats). The cut-off parameter values C_H and C_E had the impact on the number of clusters and subclusters. The number of clusters and subclusters belonging to some of the C_H and C_E values are summarized in Table 1.

The output of the algorithm was the metric values $ACC(G_H)$, $AIOR(G_H)$, $ACC(G_E)$, $AIOR(G_E)$. The values of the metric computed from the original graph $ACC(G)$ and $AIOR(G)$ did not vary and was computed during the first run of the algorithm. At first, we fixed $C_E=0.5$ and examined the dependency of $RD(ACC(G), ACC(G_H))$, $RD(ACC(G), ACC(G_E))$, $RD(AIOR(G), AIOR(G_H))$, $RD(AIOR(G), AIOR(G_E))$ on the values of C_H, which is shown on Figure 1. Then, we fixed $C_H=5.6$ km and focused on the dependence of the relative differences on the values of C_E, which is shown on Figure 2.

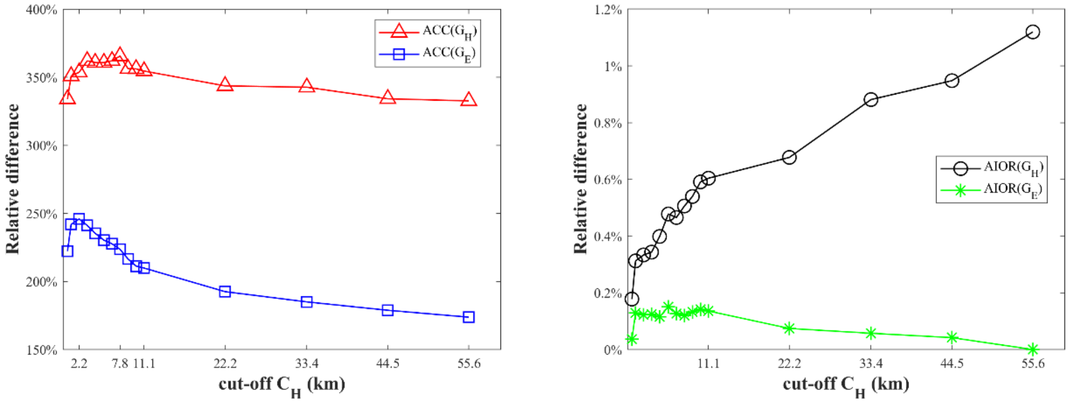

Figure 1. Dependence of the relative differences on the values of C_H with $C_E=0.5$.

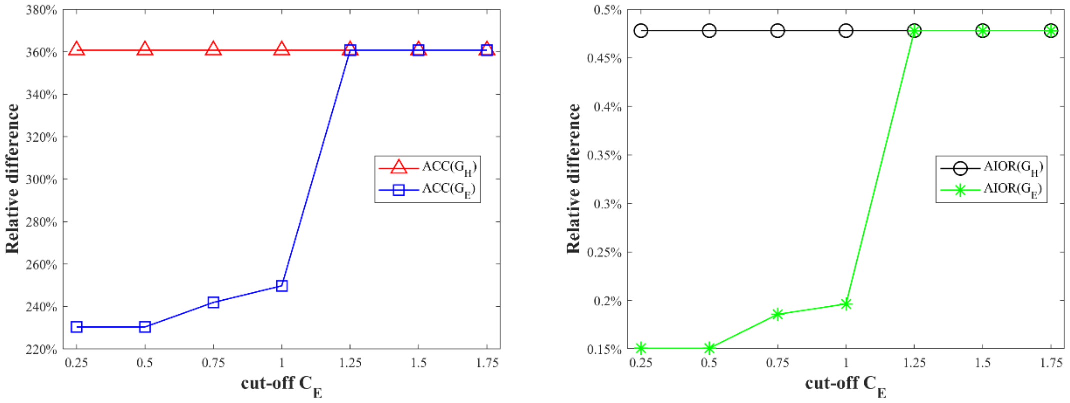

Figure 2. Dependence of the relative differences on the values of C_E with $C_H=5.5$ km.

4. Discussion

Since the relative difference values on the left-side graphs are significantly higher than the relative difference values on the right-side graphs on Figures 1 and 2, the used anonymization method is proved to have a larger impact on the values of the clustering coefficient than on the values of ratio of indegree and outdegree centrality.

In the first research question we focused on the impact of the clustering based on the geographical coordinates, which corresponding to the performance of the relative difference $RD(ACC(G), ACC(G_H))$ and $RD(AIOR(G), AIOR(G_H))$ on Figure 1. Table 1 illustrates that the higher cut-off C_H, meaning the higher anonymization level, corresponds to the smaller amount of clusters. The smaller amount of clusters causes naturally the larger utility loss in the data, which is also proved on the right-side graph on Figure 1, where the difference between $AIOR(G)$ and $AIOR(G_H)$ grows steadily with the increasing C_H. However, $RD(AIOR(G), AIOR(G_H))$ is under 1.2% for all parameter values, hence we can deduce that the geographical-based clustering anonymization preserved the ratio of the indegree and outdegree centrality.

On the other hand, $RD(ACC(G), ACC(G_H))$ values, which were between 202% and 366%, proved that the neighborhood of the clusters in G_H did not resemble the neighborhood of the locations in G.

The explanation is that most of the locations in the neighborhood of the location v_i in G were also near the location v_i geographically, thus they were clustered in the same cluster as v_i in G_H. Hence, the socio-economic researchers exploiting GSN datasets and examining the neighborhood of nodes should be careful about interpreting the clustering coefficient correctly and specify whether the clustering coefficient corresponds to the locations or some larger regions.

Our second research question addresses the entropy-based clustering method. The additional entropy-based clustering split the clusters into subclusters, thus it increases the amount of nodes in the anonymized graph G_E, as shown on Table 1. The comparisons of $RD(ACC(G), ACC(G_H))$ and $RD(ACC(G), ACC(G_E))$ on the left-side graph and $RD(AIOR(G), AIOR(G_H))$ and $RD(AIOR(G), AIOR(G_E))$ on the right-side graph on Figure 1 prove that the clustering according to the location entropy values reduces the impact of the anonymization on the examined metrics. On the right-side graph on Figure 1 there is visible a different trend in $RD(AIOR(G), AIOR(G_H))$ and $RD(AIOR(G), AIOR(G_E))$. While $RD(AIOR(G), AIOR(G_H))$ increases with the growing C_H, $RD(AIOR(G), AIOR(G_E))$ decreased. It indicates that using the location entropy values in the anonymization positively effects the preserving of data utility.

Figure 2 shows the dependence of the relative differences on the cut-off value C_E. If C_E is greater than 1, then $RD(AIOR(G), AIOR(G_H))$ equals $RD(AIOR(G), AIOR(G_E))$ and $RD(ACC(G), ACC(G_H))$ equals $RD(ACC(G), ACC(G_E))$. Hence, nearly no subclusters were made, if $C_E>1$. The value $C_E=0.5$ is proved to be the proper cut-off value for the hierarchical clustering according to the location entropy.

5. Conclusions

GSNs are valuable source of information for socio-economic researches. Since the privacy of users has to be preserved in the exploited GSN dataset, the dataset is usually anonymized before the further analysis. In this paper, we examined the impact of the hierarchical clustering anonymization method on the values of the network metric, average clustering coefficient and the ratio between the indegree and outdegree centrality, which was used in the socio-economic analysis performed by Zhou et al. (2017).

We applied the hierarchical clustering method according to the geographical coordinates on the data sample of the real-world dataset Gowalla. Moreover, we focused on the impact of the additional clustering according to the location entropy values on the clustered data. The geographical-based clustering anonymization preserved well the ratio of the indegree and outdegree centrality, while it had a huge impact on the values of the clustering coefficient. Applying the entropy-based clustering improved the metric values significantly and we recommend to use the values of location entropy in the anonymization of location-based data.

The future research can focus on the impact of anonymization on other network metrics as well as the further use of the location entropy in the other GSN anonymization methods.

Acknowledgments: The financial support of the Specific Research Project "Information and Knowledge Management and Cognitive Science in Tourism" of FIM UHK is gratefully acknowledged.

References

Cho Eunjoon, Myers Seth A. and Leskovec Jure. 2011. 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.* ACM Press. pp. 1082-1090.

Cranshaw Justin, Toch Eran, Hong Jason, Kittur Aniket and Sadeh Norman. 2010. Bridging the gap between physical location and online social networks. In: *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*. pp. 119-128. ACM Press. https://doi.org/10.1145/1864349.1864380.

Duckham Matt and Kulik Lars. 2006. Location privacy and location-aware computing. In *Dynamic and Mobile GIS*. CRC press, pp. 63- 80. https:// doi.org/10.1201/9781420008609.ch3

Foursquare. 2020. Foursquare - The Trusted Location Data & Intelligence Company. Available online: https://foursquare.com (accessed on 26th January 2020).

Foursquare Labs, Inc. 2020. Privacy Policy. Available online: https://foursquare.com/legal/privacy (accessed on 26th January 2020).

Fung Benjamin C.M., Wang Ke, Fu Ada Wai-Chee and Philip S. Yu. 2010. *Introduction to privacy-preserving data publishing: Concepts and techniques*. Chapman and Hall/CRC.

Gambs Sébastien, Heen Olivier and Potin Christophe. 2011. A comparative privacy analysis of geosocial networks. *Proceedings of the 4th ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS.*ACM Press, pp. 33-40.

James Gareth, Witten Daniela, Hastie Trevor and Tibshirani Robert. 2013. *An introduction to statistical learning*. New York. Springer.

Keßler Carsten and McKenzie Grant. 2018. A geoprivacy manifesto. *Transactions in GIS*: Vol. 22, pp. 3-19. https://doi.org/10.1111/tgis.12305.

Kušnerová Milena, Valíček Jan, Harničárová Marta, Hryniewicz Tadeusz, Rokosz Krzysztof, Palková Zuzana, Václavík Vojtěch, Řepka Michal and Bendová Miroslava. 2013. A proposal for simplifying the method of evaluation of uncertainties in measurement results. *Measurement Science Review*. Vol. 13, pp. 1-6.

Masoumzadeh Amirreza and Joshi James. 2013. Top Location Anonymization for Geosocial Network Datasets. *Trans. Data Privacy.* Vol. 6, pp. 107-126.

Narayanan Arvind and Shmatikov Vitaly. 2009. De-anonymizing social networks. In: *2009 30th IEEE symposium on security and privacy.* IEEE.

Palos-Sanchez Pedro, Saura Jose Ramon, Reyes-Menendez Ana and Esquivel Ivonne Vásquez. 2018. Users acceptance of location-based marketing apps in tourism sector: An exploratory analysis. *Journal of Spatial and Organizational Dynamics.* Vol. 6, pp. 258-270.

Scellato Salvatore, Noulas Anastasios and Mascolo Cecilia. 2011. Exploiting place features in link prediction on location-based social networks. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM Press, pp. 1046-1054.

Shokri Reza, Theodorakopoulos George, Le Boudec Jean-Yves and Hubaux Jean-Pierre. 2011. Quantifying location privacy. In: 2011 IEEE symposium on security and privacy. IEEE, pp. 247-262.

Zhou Xiao, Hristova Desislava, Noulas Anastasios, Mascolo Cecilia and Sklar Max. 2017. Cultural investment and urban socio-economic development: a geosocial network approach. *Royal Society open science*. The Royal Society Publishing. Vol. 4. http://dx.doi.org/10.1098/rsos.170413.