Anonymization of geosocial network data by the (k, l)-degree method with location entropy edge selection

Jana Medková[a]

[a]*Faculty of Informatics and Management, University of Hradec Kralove, Rokitanskeho 62, Hradec Kralove 50003, Czech Republic*

Abstract

Geosocial networks (GSNs) have become an important branch of location-based services since sharing information among friends is the additional feature to provide information based on user's current location. The growing popularity of location-based services contribute to the development of highly customized and flexible utilities. However, providing the customized services relates to collecting and storing a large amount of users' information. In this paper, we focus on the privacy preserving concern in publishing GSN datasets. We introduce a new (k, l)-degree anonymization method to prevent the re-identification attack in the published GSN dataset. The presented method anonymizes user's social relationships as well as location-based information in GSN. We propose the new (k, l)-degree anonymization algorithm which modifies the network structure with a sequence of edge editing operations. Furthermore, a location entropy metric is used to measure the importance of the visited locations in the edge selection procedure of the algorithm. New edges are added preferably among the users who visited the same places with significant importance to them. This may contribute to making a real social tie between them in the future. Moreover, we explore the usability of the algorithm by running experiments on real-world datasets.

Keywords:
Geosocial network, Privacy, Anonymization, Affiliation network, Location entropy

Email address: jana.medkova@uhk.cz (Jana Medková)

1. Introduction

The world-wide spreading of the smart devices users equipped with GPS locators causes the development of various location based services (LBS). Users of mobile devices are able to query the location providers and benefit from corresponding LBS data during travelling or everyday activities. The LBS data include navigation information, restaurant recommendation or live traffic information. In return, the LBS customers provide the location providers with a real-time information access to their current location. The special branch of LBS are geosocial networks (GSNs) where sharing information among friends is the additional feature to providing information based on user's current location.

The geosocial networks (GSNs) are social networks (SNs) extended with a location attribute. The integrated location information enabling users to share their visited location and recommendations of locations with their friends. Users can check-in their current location, share their location with friends, recommend services at their location or highlight nearby points of interest.

GSNs play an important role in tourism. GSNs as well as other mobile information and communication technologies are widely used by tourism participants. In many aspects, GSNs have taken over the role of traditional information centers. Being a global information source with unlimited access, they influence behaviour of tourists and travellers. Therefore, tourists can make quick and effective decisions while buying products of tourist services. Another advantage of GSNs in comparison with traditional tourists centers is that the offered information come from the tourists themselves. Sharing wishes and complaints about travel-related services influences indirectly the future development of the services.

Academic and industry research has recently gained information from publishing SN datasets. Including the location information into GSNs makes the datasets even more valuable. The dataset can be used in social, tourist or marketing studies to analyze individual's behaviour based not only on their personal attributes, but also on their mobility (Cho et al., 2011; Fire et al., 2012).

However, publishing GSN datasets may threaten the users' privacy and cause the identity disclosure, which means the leaking of the individual's identity from published records. An attacker, equipped with background knowledge, can try to re-identify the target individual combining the records

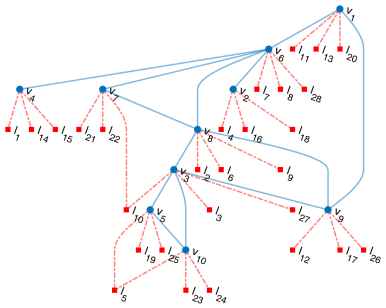
in the published dataset. The knowledge about the visited locations is as valuable information as any knowledge about graph structure or target's relationships. Analyzing the set of user's check-ins may reveal his/her other personal information such as home or workplace locations (Golle and Partridge, 2009; Pontes et al., 2012).

Anonymization enables providers to publish datasets while preserving privacy of individuals. The provider applies an anonymization method on the datasets and publishes only its anonymized version. The issue of preventing the identity disclosure in SNs has been widely studied (Casas-Roma et al., 2017; Liu and Terzi, 2008; Hay et al., 2010; Chakraborty and Tripathy, 2016; Medková, 2018). Nevertheless, collecting location information requires a special approach while anonymizing the GSN dataset (Li et al., 2016). Previous research studies (Li et al., 2016; Masoumzadeh and Joshi, 2013), dealing with the identity disclosure problem in GSNs, proposed different anonymization concepts based on k-anonymity.

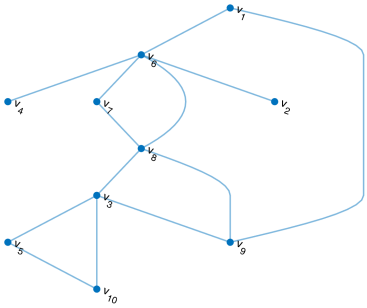
In this paper, we introduce a new concept of (k, l)-degree anonymity for the GSN dataset. The proposed (k, l)-degree anonymization method is based on k-degree anonymization methods for SNs (Casas-Roma et al., 2017; Liu and Terzi, 2008). The proposed algorithm is established on recently presented properties of GSN (Gao et al., 2012; Cho et al., 2011; Scellato et al., 2011). Moreover, the location entropy is considered as the measure of the popularity of the visited locations.

In our proposed framework, GSN is represented by a graph $\mathcal{G}$ with two kinds of nodes; location and user nodes, and two kind of edges; user-user links and user-location links (see *Figure 1a*). User nodes and user-user links form a subgraph G_V and represents social relationships in GSN (see *Figure 1b*). The user nodes, location nodes and user-location links represent information about user's checked-in locations. They are considered to form an affiliation network H_L (see *Figure 1c*). Generally, in the affiliation networks, users are linked to groups of interest, and the groups are linked to their members (Zheleva et al., 2009). In this paper, the groups of interest are locations, at which at least one user has checked in.

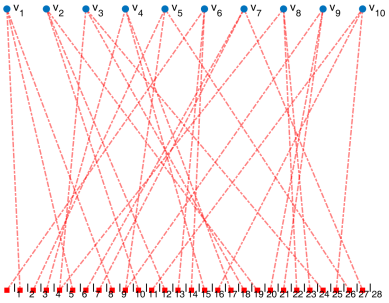
The subgraphs G_V and H_L are anonymized separately using different edge editing algorithms. The algorithm for anonymizing G_V uses the location information from H_L, but it does not change the edge set of H_L. Similarly, the algorithm for anonymizing H_L uses the relationship information from G_V and does not modify the structure of G_V, as described in detail in *Section 4*. By applying the anonymization algorithms, we obtain k-degree anonymous



a) geosocial network $\mathcal{G}$



b) social network G_V



c) affiliation network H_L

Figure 1: Geosocial network graph $\mathcal{G} = (V, E_V, L, E_L)$ and its social and affiliation sub-networks. The vertex set V, marked with blue dots, represents the set of user nodes. The vertex set L, marked with red squares, represents the location nodes. The edge set E_V, marked with blue full lines, represents the social relationships in the network, while the edge set E_L, marked with red dashed lines, represents users' check-ins at locations.

social network G_V^* and l-degree anonymous affiliation network H_L^*. Thus, the result of the whole method is (k, l)-degree anonymous geosocial network $\mathcal{G}^*$.

The presence of location information in GSN enable us to improve the previously presented algorithms for k-degree anonymization in SN (Liu and Terzi, 2008; Casas-Roma et al., 2017) and introduce the new (k, l)-degree

method for GSN anonymization. We propose the algorithm where the location entropy metric is used for quantifying the importance of edges. Moreover, the concept of l-degree anonymous affiliation network as well as the whole representation of GSN as a combination of social and affiliation network is newly introduced in this paper.

The rest of the paper is arranged as follows. The previous research studies on anonymization methods of the GSN data and k-degree anonymization methods of the SN data are summarized in *Section 2*. *Section 3* introduces the (k, l)-anonymization method. The proposed algorithm is described in detail in *Section 4*. The experimental results are presented in *Section 5*. Finally, the paper is concluded in *Section 6*.

2. Related work

The methods, developed for anonymizing undirected and unlabeled graphs representing SNs, usually anonymize the graph structure with modifying the set of edges (Casas-Roma et al., 2017; Liu and Terzi, 2008), aggregating the nodes into clusters (Hay et al., 2010) or adding artificial noise nodes (Chakraborty and Tripathy, 2016). We propose a novel privacy preserving approach for the GSN dataset, which is based on the edge editing method for SN anonymization, well-known as a k-degree anonymization.

The k-degree anonymity was first introduced in (Liu and Terzi, 2008). Liu and Terzi proposed a systematic framework for graph anonymization to prevent the re-identification of individuals by an attacker with structural background knowledge. They decomposed the problem into two parts: finding a k-anonymized degree sequence and constructing a new anonymized graph. At the first step, they proposed an algorithm to extract the degree sequence d from the original graph G and find the k-anonymized degree sequence d^*. In the later step, d^* was used to construct a new anonymized graph G^*. Due to the k-anonymity property of d^*, the resultant graph G^* was k-degree anonymous. In our experiments, we used the proposed *Greedy* algorithm to discover an anonymized degree sequence for constructing the subgraph G_V^*. We add new heuristics to the *Greedy* algorithm to reduce the total count of edge edits in G_V, as it is described in detail in *Subsection 4.1.*

The following research (Hartung et al., 2014; Lu et al., 2012) improved the k-degree anonymity approach in terms of speed by applying different kinds of heuristics. Casas-Roma et al. (2017) introduced a k-degree anonymization

algorithm based on univariate micro-aggregation to anonymize large networks. Their algorithm was experimentally shown to improve the k-degree anonymization of SN in terms of the information loss and data utility. They preserved the two-step approach and split the task into the problem of degree sequence anonymization and graph modification. In the graph modification algorithm, the neighbourhood centrality measure was used for quantifying the edge relevance in the network and selecting auxiliary edges in the edge editing operations. Our proposed social network modification algorithm is based on their graph construction algorithm. Instead of using the neighbourhood centrality, we consider the set of the visited locations and their location entropy.

The location entropy was introduced together with other location-based measures for analyzing the social context of a geographic region in (Cranshaw et al., 2010). They showed that there existed a positive relationship between the entropy of the locations, which the user visited, and the number of social ties, that the user had in the network. As presented in (Scellato et al., 2011), the location entropy was exploited to define features for the link predictions in GSN. They experimentally showed that sharing locations with a low entropy value was an important indicator in establishing new relationships in the network. In our framework, the location entropy is used to measure the popularity of locations and their importance for visitors. the lower entropy indicates greater importance for visitors. Additionally, when two users visit the same location with the low entropy, they are more likely to make a social tie in the network in the future (Cho et al., 2011). Thus, when it is necessary to make a new link in the network, users visiting the same location with the low entropy are preferred. It increases the probability that the anonymized network resembles the future development of the network.

The concept of connecting the social network and its user's attributes using the affiliation network is presented in (Zheleva et al., 2009). They stated that each SN co-existed with a two-mode affiliation network, in which users were linked to the groups of interest, and the groups were linked to their members. Moreover, they presented a generative model for social and affiliation networks. They studied such groups that users had voluntarily chosen to be part of them, for example, being in a book-reading club. They did not focus on natural groups identified according to age or sex. In this paper, users are formed into the groups according to the checked-in locations in GSN. Users are in the group z_i iff they visited the location z_i. As far as we know, this is the first study where GSN is represented with the combination

of the undirected graph and the affiliation network.

Many approaches have been recently presented to handle different privacy preserving problems in GSN. Possible types of attacks on GSNs as well as the properties that should be satisfied in the privacy preserving GSN model were introduced in (Carbunar et al., 2013). The privacy preserving issues, related to gaining rewards for checking in at locations, were discussed in (Carbunar et al., 2014). The problem of providing a social recommendation while keeping the users' social relations private, was addressed in (Liu and Hengartner, 2013; Zhang et al., 2014). Rahman et al. (2013) introduced a concept of location centric profiles, aggregating statistics built from the profiles of users. Their framework guaranteed strong privacy to users and correctness assurances to GSN providers.

The problem of preserving location privacy in GSNs was adressed in (Alrayes and Abdelmoty, 2016; Xue et al., 2017; Ma et al., 2018; Siddula et al., 2018; Kotzanikolaou et al., 2016).

The theoretical framework for evaluating location privacy was proposed in (Shokri et al., 2011). Shokri et al. formally define the terms of tracking and localization attack on anonymous traces. Furthermore, they provided a tool for evaluating the effectiveness of various location-privacy mechanisms. The model is further developed in (Shokri et al., 2012) into the framework that includes the adversarial knowledge into a privacy-preserving process.

The principles and existing approaches to preserve the location privacy was summarized in (Wernke et al., 2014). The compact data structure based on Bloom filters was designed to store the location information and preserve the location privacy in (Calderoni et al., 2015). The research in (Alrayes and Abdelmoty, 2016) focused on location content awarness in relation to privacy on GSNs. Ma et al. in (Ma et al., 2018) proposed the algorithm to preserve the user's location privacy by replacing his/her location with circle overlapping regions.

Unlike the previous studies focusing on spatio-temporal dimension of GSNs, we consider different model of geosocial network which does not deal with the timeline of visiting the locations. Hence, the adversary is not able to compile the actual trace of the target user and perform the localization attack. We focus on the spatial and social dimension of GSN data and the identity disclosure problem in GSN.

The identity disclosure problem in the GSN datasets was addressed in (Li et al., 2016; Masoumzadeh and Joshi, 2013). Both research studies focused on the re-identification attack performed by an adversary with the background

knowledge about user locations. Masoumzadeh and Joshi (2013) introduced a top location representation and presented the concept of $\mathcal{L}_k$-anonymity and $\mathcal{L}_k^2$-anonymity based on k-anonymity. Moreover, they proposed a clustering algorithms for anonymizing a GSN dataset. Clusters with at least k nodes were created and an anonymized value was produced for each cluster. The GSN datasets were represented as a 4-tuple graph, where each user was assigned to one location value in the location domain. In our representation, each user is usually assigned to multiple location values corresponding to the number of the visited locations. Instead of clustering approach where the locations are clustered into regions, we propose the edge editing algorithm where the original location value is preserved.

In (Li et al., 2016), the GSN dataset was represented as a hypergraph. Li et al. presented relaxations of $\mathcal{L}_k$-anonymity and $\mathcal{L}_k^2$-anonymity called (k, m)-anonymity and (k, m, l)-anonymity respectively. They proposed a two-step anonymization algorithm to achieve the (k, l, m)-anonymity. In the first step, locations were generalized and the location links were anonymized. In the second step, only the social network links were modified.

Although we also anonymize G_V and H_L separately, the choice, which subgraph is anonymized as the first one, is made during the run of the algorithm. The first to process is the subgraph, which requires more structural changes. The decision is made not only by the structure of $\mathcal{G}$, but also by the required level of anonymity given by parameters k and l.

3. Problem definition

In this study, the geosocial network is represented as a combination of a social network describing the social relationships within the network and an affiliation network describing the location information. For reference, the summary of the notation, used throughout this paper, is included in *Table 1.*

Definition 3.1 (GSN dataset)**.** *The GSN dataset is described by graph* $\mathcal{G} = (V, E_V, L, E_L)$*, where* $V = \{v_1, \ldots, v_n\}$ *is a set of vertices representing the users connected within the GSN,* $E_V \subseteq V \times V$ *is a set of edges representing the relationships among users,* $L = \{z_1, \ldots, z_m\}$ *is a set of vertices representing the visited locations and* $E_L \subseteq V \times L$ *is a set of edges representing the visits of users in locations.*

Let $\Theta_i = \{z \in L;\ (v_i, z) \in E_L\}$ denote the set of locations visited by the user $v_i \in V$. In case that the edge set E_V is omitted, the bipartite

graph $H_L = (V, L; E_L)$ is an affiliation network of users and locations. Every edge from E_L connects a vertex from V to one from L. There is an edge between the nodes $v \in V$ and $z \in L$ iff the user v visited the location z. On the other hand, the graph $G_V = (V, E_V)$, describing the social interactions among users, is a typical social network. Thus, $(v_i, v_j) \in E_V$ iff the user v_i is in the relationship with the user v_j within the network $\mathcal{G}$. An example of a simple GSN and its social and affiliation network is shown in *Figure 1.* The terms "check-in at location" and "visit a location" are used interchangeably as well as the terms "node" and "vertex".

The anonymization of $\mathcal{G}$ is done in two steps: anonymizing G_V and anonymizing H_L. We used two parameters k and l for identifying the level of anonymization in each subnetwork. The edge editing anonymization methods change only the edge sets and leave the node sets untouched. Since G_V and H_L share only the node set V, the anonymization of H_L has no impact on the structure of G_V and vice versa. The gained k-degree anonymized social network G_V^* and l-degree anonymized affiliation network H_L^* results in (k, l)-degree anonymous GSN network $\mathcal{G}^* = (V, E_V^*, L, E_L^*)$.

The aim of the k-degree anonymization is to prevent an attacker from re-identifying his/her target individual in the published dataset using the background knowledge about the graph structure. Let assume that the attacker knows the degree of his/her target node. When the graph structure is changed in the way that all nodes have at least $k-1$ other nodes sharing the same degree, then the probability of re-identifying attack equals to $\frac{1}{k}$. We provide the definitions of k-degree anonymous vector and graph, proposed in (Liu and Terzi, 2008).

Definition 3.2 (k-anonymous vector)**.** *A vector of integers $u = (u_1, \ldots, u_n)$ is* k-anonymous*, if every distinct value u_i, $i = 1, \ldots, n$, appears in u at least k times.*

Definition 3.3 (k-degree anonymous graph)**.** *A graph $G = (V, E)$ is* k-degree anonymous *if the degree sequence d_G is k-anonymous. It means, that for every vertex $v \in V$, there exist at least $k-1$ other vertices $v_1, \ldots v_{k-1}$ with the same degree, $deg(v) = deg(v_1) = \cdots = deg(v_{k-1})$.*

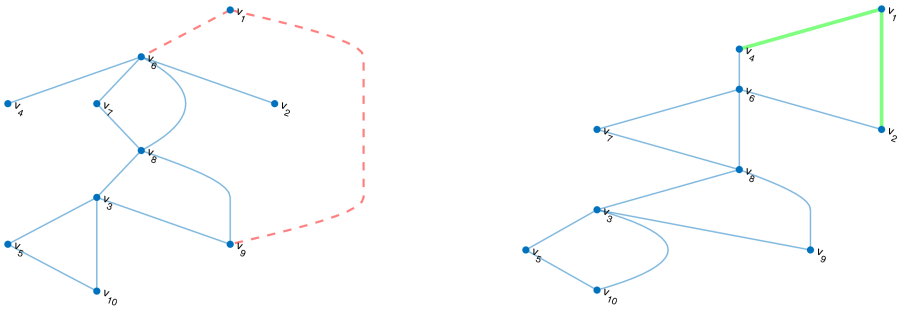
Figure 2 shows the small social network G_V with the degree sequence $d_V = (2, 1, 4, 1, 2, 5, 2, 4, 3, 2)$ and its 3-degree anonymized version G_V^* with the degree sequence $d_V^* = (2, 2, 4, 2, 2, 4, 2, 4, 2, 2)$. The 3-degree anonymized

Notation	Description
$\mathcal{G} = (V, E_V, L, E_L)$	geosocial network graph
V	set of user nodes
$n = \|V\|$	number of user vertices
L	set of location nodes
E_V	user-user links
E_L	user-location links
$G_V = (V, E_V)$	social network; subgraph of $\mathcal{G}$
$H_L = (V, L; E_L)$	affiliation network; subgraph of $\mathcal{G}$
$v \in V$	user vertex/ user node/user
$z \in L$	location vertex/ location node/location
(v_1, v_2)	an edge connecting v_1 and v_2
$\mathcal{G}^*$	anonymized $\mathcal{G}$
Θ_i	set of locations visited by the user v_i
k	parameter of k-degree anonymization of G_V
l	parameter of l-degree anonymization of H_L
d_V	degree sequence of G_V
d_L	degree sequence of location nodes in H_L
$deg(z)$	degree of a node z
$C(z)$	a total number of check-ins that all users have at z
$q_i(z)$	a fraction of check-ins the user v_i has at location z
$\mathcal{E}(z)$	location entropy of the location z
δ_V	a vector indicating needed changes in d_V
δ_V^+	set of vertices from V that have to increase their degree
δ_V^-	set of vertices from V that have to decrease their degree
δ_L^+	set of vertices from L that have to increase their degree
$\sigma(d_V)$	sum of all elements in d_V

Table 1: Summary of the notation.

graph was obtained with four changes in the edge set. Two edges (v_1, v_6), (v_1, v_9) have been removed and two edges (v_1, v_2), (v_1, v_4) have been added.

We introduce a new concept of the l-degree anonymous affiliation network. The aim of the anonymization of H_L is to prevent an attacker to re-identify the target user using the background knowledge about the locations visited by the target user. Let assume that the attacker knows a location z_T, visited by the target individual. If the location z_T were not visited by other users,



a) social network G_V b) 3-degree anonymized social network G_V^*

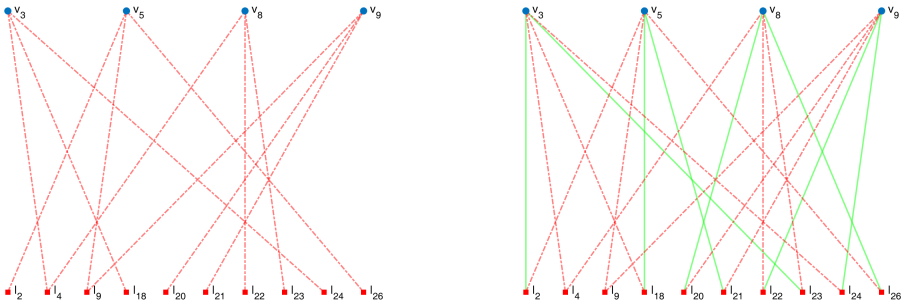
Figure 2: Social network $G_V = (V, E)$ and its 3-degree anonymized version G_V^*. The dashed red edges have been removed from E and the green bold edges have been added to E to achieve the 3-degree anonymity.

then the attacker can surely connect the individual with its node within the network. The number of users who visited the location equals to its degree in H_L. In case that the degree of every location is at least l, which means that every location from L is connected with at least l users from V, then the probability of re-identification attack is equal to $\frac{1}{l}$. The idea is formalized in the following definition.

Definition 3.4 (l-degree anonymous affiliation network)**.** *Let $H = (A, B; E)$ be an affiliation network where A is a set of objects and B is a set of affiliations. Then H is* l-degree anonymous *if $\forall b \in B : deg(b) \geq l$.*

Figure 3 shows a subgraph $\widehat{H_L}$ of the bipartite graph H_L representing a small affiliation network. The degree sequence of location nodes is $d_L = (1, 2, 2, 1, 1, 1, 1, 1, 1, 1)$. The degree of all nodes has to be equal at least 2 to create a 2-degree anonymous $\widehat{H_L}^*$.

Let assume that the attacker has both structural and location information about the target user. The attacker is considered to know the degree of the target node in G_V and one of the locations visited by the target user in H_L. We propose the following (k, l)-degree anonymous model of GSN to prevent GSN from the re-identification attack.



a) an affiliation network $\widehat{H_L}$ b) 2-degree anonymized affiliation network $\widehat{H_L}^*$

Figure 3: Affiliation network $\widehat{H_L}$ and its 2-degree anonymized version $\widehat{H_L}^*$. The bold green edges have been added to the edge set to achieve the 2-degree anonymity.

Definition 3.5 ((k,l)-degree anonymous GSN)**.** *A GSN $\mathcal{G} = (V, E_V, L, E_L)$ is* (k,l)-degree anonymous *if the graph $G_V = (V, E)$ is k-degree anonymous and the affiliation network $H_L = (V, L; E_L)$ is l-degree anonymous. It means, that for every user $v \in V$ there exist at least $k-1$ other users $v_1, \ldots v_{k-1} \in V$ with the same number of relationships, $deg(v) = deg(v_1) = \cdots = deg(v_{k-1})$, and for every location $z \in L$ there exist at least l users which visited it, $deg(z) \geq l$.*

Finally, we provide the definition of the location-based feature called the location entropy, introduced in (Cranshaw et al., 2010). The location entropy is used for weighing locations in the social network modification algorithm (see *Subsection 4.2*).

Definition 3.6 (Location entropy)**.** *Let $z \in L$ be a location and $C(z)$ be a total number of check-ins that all users have at z. Let $v_i \in V$, $i = 1, \ldots, n$ be a user and $c_i(z)$ be the user's number of check-ins at the location z. Then $q_i(z) = \frac{c_i(z)}{C(z)}$ is a fraction of check-ins the user v_i has at location z and $\{q_1(z), \ldots, q_n(z)\}$ is a discrete probability distribution that the location z is visited by a certain user. The* location entropy *of z is defined as follows:*

$$\mathcal{E}(z) = -\sum_{i=1,\ldots,n} q_i(z) \log q_i(z)$$

4. (k, l)-degree anonymization algorithm

The aim of the proposed algorithm is to anonymize $\mathcal{G} = (V, E_V, L, E_L)$. We solve separately the k-degree anonymization of $G_V = (V, E)$ and l-degree anonymization of $H_L = (V, L; E_L)$. The two tasks are not completely independent on each other, since information stored in H_L is used in the anonymization method of G_V and vice versa.

Let d_V be a degree sequence of G_V. At first, we construct the k-anonymous degree sequences d_V^* using one of the previously presented degree sequence anonymization algorithms (Liu and Terzi, 2008; Casas-Roma et al., 2017). Let set $\delta_V = d_V - d_V^*$ to be a vector indicating which vertices from G_V must change their degree to fulfil the k-degree anonymity. For example, if $\delta_V(1) = 2$ and $\delta_V(2) = -3$, then $deg_V(v_1)$ need to be decreased by 2 and $deg_V(v_2)$ need to be increased by 3. The number of necessary modifications in the graph G_V equals to $c_V = \sum_{i=1}^{n} |\delta_V(i)|$.

Since the definition of l-degree anonymity requires nodes to have their degree equalling or being greater than l, the number of necessary modifications of the edge set E_L equals to $c_L = \sum_{z \in \delta_L^+} l - deg(z)$, where $\delta_L^+ = \{z \in L;\ deg(z) < l\}$.

If $c_V \leq c_L$, then G_V is modified before H_L with a *Social network modification algorithm with the location edge selection*. The algorithm exploits the location information which is stored in H_L. Then H_L is modified with an *Affiliation network modification algorithm*. The algorithm searches for information about the social ties in already anonymized G_V^*. If $c_V > c_L$, H_L is anonymized as the first one and G_V as the latter one. Thus, both parameters k, l have an impact on the anonymization of both G_V and H_L.

4.1. Heuristics in degree anonymization algorithm

The social network modification algorithm is independent on the choice of the degree sequence anonymization algorithm used for anonymizing d_V. In our implementation, the *Greedy* algorithm, introduced in (Liu and Terzi, 2008), was applied. We used the version of *Greedy* algorithm that considers a simultaneous edge addition and removal during the successive graph modification.

Assuming the degree sequence d_V is sorted in descending order, the *Greedy* algorithm first forms a group consisting of the first k vertices with the highest degrees. Then the degree of all vertices from the group is set to the same value. The new value is the median value of the original degrees. Then

the algorithm checks whether it is better to join the $(k+1)$-th vertex into the previously formed group or start a new group at position $(k+1)$. The detailed description of the *Greedy* algorithm is omitted with reference to (Liu and Terzi, 2008).

Since the degree distribution in SN is a power-law distribution as it is presented in (Mislove et al., 2007), there are a few nodes with a very high degree. Where the high-degree nodes are merged into one group with other nodes, the highest degrees are very distant from the median. It causes removing a large amount of edges from G_V in the graph modification algorithm, which could not be always possible. Hence, after computing d_V^* with the *Greedy* algorithm, we slightly increase the anonymized degree value in a few first groups. It increases the probability that d_V^* is feasible and the original graph can be modified with the edge editing operations to meet d_V^*.

4.2. Social network modification algorithm with the location entropy selection

The graph modification algorithm with the location entropy selection is based on the graph modification algorithm proposed in (Casas-Roma et al., 2017). Let $\sigma(d_V)$ be the sum of each element in d_V, $\sigma(d_V) = \sum_{i=1}^{n} d_V(i)$, and let $\sigma(d_V^*)$ be the sum of each element in d_V^*, $\sigma(d_V^*) = \sum_{i=1}^{n} d_V^*(i)$. We create a set of vertices which have to decrease their degree $\delta_V^- = \{v_i \in V;\ \delta_V(i) > 0\}$ and the set of vertices which have to increase their degree $\delta_V^+ = \{v_i \in V;\ \delta_V(i) < 0\}$.

In case that $\sigma(d_V^*) < \sigma(d_V)$, the original graph is modified with sequence of the edge removal operations. For every vertex $v_i \in \delta_V^-$, we select the most suitable vertex $v_j \in \delta_V^-$ in such a way that there exists an edge $(v_i, v_j) \in E_V$ and remove the edge from E_V. Therefore, the degree of both vertices has decreased and $\delta_V(i)$ and $\delta_V(j)$ are increased by 1. The procedure is repeated unless $\sigma(d_V^*) = \sigma(d_V)$.

In case that $\sigma(d_V^*) > \sigma(d_V)$, the original graph is modified with the sequence of the edge addition operations. For every vertex $v_i \in \delta_V^+$, we select the most suitable vertex $v_j \in \delta_V^+$ such that $(v_i, v_j) \notin E_V$ and add the edge into E_V.

When $\sigma(d_V^*) = \sigma(d_V)$, the anonymized graph with d_V^* is obtained with the sequence of the edge switching operations. For every vertex $v_i \in \delta_V^-$, we select the most suitable neighbour $v_m \in V$, $(v_i, v_m) \in E_V$, and the most suitable $v_j \in \delta_V^+$ such that $(v_m, v_j) \notin E_V$. Then (v_i, v_m) is removed from E_V and (v_m, v_j) is added to E_V. Therefore, $deg(v_i)$ is decreased, $deg(v_j)$ is

increased and $deg(v_m)$ does not change. The procedure is repeated with all $v_i \in \delta_V^-$, unless $d_V = d_V^*$ and $\delta_V = (0, \ldots, 0)$.

In every graph modification step, it is necessary to select the most suitable pair of vertices to remove the existing edge or add a new one. It is possible to select the pairs of vertices at random or used the neighborhood centrality properties of the network, it is as proposed in (Casas-Roma et al., 2017). However, we decided to exploit the location information kept in the GSN. Before every graph modification step, we investigate locations connected with participating users v_i, v_j.

As it is presented in (Scellato et al., 2011), the location entropy is an indicator of whether a certain location is likely to result in social ties among its visitors. The locations with the low entropy might result in more social links among their visitors than the locations with the higher entropy. Hence, by linking the pairs of users who visited the same location with the low entropy, we make connections that are likely to be really added to the network in the future. When $deg(v_i)$ has to be increased during the edge addition step for $v_i \in \delta_V^+$, the suitable vertex $v_j \in \delta^+$ is selected with the *Minimal entropy selection algorithm for edge addition* (see *Algorithm 1*).

Algorithm 1 Minimal entropy selection for the edge addition

Input: $v_i \in \delta_V^+$
Output: $v_x \in \delta_V^+$
1: Set $W_i = \{v_j \in \delta_V^+ : (v_i, v_j) \notin E_V\}$.
2: Find $v_x = \arg\min_{v_j \in W_i}(\min_{z \in \Theta_i \cap \Theta_j} \mathcal{E}(z))$.
3: **return** v_x.

On the other hand, the locations with the low entropy are likely to be places with significant importance for their visitors, such as a home place or work office (Scellato et al., 2011). The locations with the high entropy are likely to be public places. The information that two users visited a location with a high entropy (e.g. railway station, shop), is less significant than the information, that they visited low-entropy location (e.g. home or work location). Hence, two users, who visited the same low entropy location, are likely to have a strong social tie with each other. Removing the link between them may cause a larger information loss then deleting the link among users, who did not visit the same location or visited the same location with the high entropy. Therefore, the links among users, who visit the same place with the low entropy, are kept untouched in the edge removal procedure. Suitable

vertices for the edge removal are selected with the *Maximal entropy selection algorithm for the edge removal*, described in detail in *Algorithm 2*.

During the edge switch operation, both approaches are combined. At first, the minimal entropy selection algorithm for the edge removal is used to select the suitable neighbour, then the minimal entropy algorithm for the edge addition is used to select a suitable vertex for the edge addition.

Algorithm 2 Maximal entropy selection for the edge removal

Input: $v_i \in \delta_V^-$
Output: $v_x \in \delta_V^-$
1: Set $W_i = \{v_j \in \delta_V^- : \ (v_i, v_j) \in E_V\}$.
2: Set $\overline{W_i} = \{v_j \in W_i : \ \Theta_i \cap \Theta_j = \emptyset\}$.
3: **if** $\overline{W_i} \neq \emptyset$ **then**
4: **return** random $v_x \in \overline{W_i}$.
5: **else**
6: Find $v_x = \arg\max_{v_j \in W_i}(\max_{z \in \Theta_i \cap \Theta_j} \mathcal{E}(z))$.
7: **return** v_x.
8: **end if**

4.2.1. Complexity

Let assume $\sigma(d_V) < \sigma(d_V^*)$ without any loss of generality. Then the algorithm requires adding new edges unless $\sigma(d_V) = \sigma(d_V^*)$. Since a new edge is created between two elements of δ_V^+ every edge addition step, every additional step increases $\sigma(d_V)$ with 2. Hence we need $c = (|\sigma(d_V) - \sigma(d_V^*))/2$ simple edge additions to fulfil the condition $\sigma(d_V) = \sigma(d_V^*)$.

Each addition step requires one run of *Algorithm 1*, where the most complex step is the computation of $\min_{v_j \in W_i}(\min_{z \in \Theta_i \cap \Theta_j} \mathcal{E}(z))$. As it is described in *Section 5*, all values of $\mathcal{E}(z)$ are computed in the preprocessing step and stored in H_L. Moreover, before the individual vertices from δ_V^+ are processed, we also compute $\min_{z \in \Theta_i \cap \Theta_j} \mathcal{E}(z)$ for every $v_i, v_j \in \delta_V^+$ and store them in matrix M of size $|\delta_V^+| \times |\delta_V^+|$. Therefore, the whole edge addition means to find a minimum in M for c times, which implies the complexity $\mathcal{O}(c * |\delta_V^+|^2)$.

The computation of M requires to find the intersection of all visited locations in H_L for every pair of elements of δ_V^+ and find the minimum in the intersection. Finding neighbours in H_L is in $\mathcal{O}(\hat{d}_L)$, where $\hat{d}_L$ is the maximum degree of the user node in H_L, the intersection of the neighbours sets is in $\mathcal{O}(\hat{d}_L \log \hat{d}_L)$ and finding the minimum $\mathcal{O}(\hat{d}_L)$. Overall, the computation of

M is in $\mathcal{O}(|\delta_V^+| * \hat{d}_L \log \hat{d}_L)$. Moreover, the upper estimate for $|\delta_V^+|$ is n and the upper estimate for c is $(\hat{d}_V * n)/2$, where $\hat{d}_V$ is the maximal degree in G_V. Overall, the whole edge addition method is in $\mathcal{O}(\hat{d}_V * n^3)$.

Computing the complexity of the edge removal process is similar and gives the same complexity $\mathcal{O}(\hat{d}_V * n^3)$. In the edge switching method, the following operations are done for every $v_i \in \delta_V^+$: finding a set of neighbours N_V in G_V ($\mathcal{O}(\hat{d}_V)$), finding $\min_{z \in \Theta_i \cap \Theta_j} \mathcal{E}(z)$ for every $v_m \in N_V$ ($\mathcal{O}(\hat{d}_V)$) and sorting N in the descend order according to the minimal entropy values ($\mathcal{O}(\hat{d}_V \log \hat{d}_V)$). Moreover, for every $v_m \in N_V$ we find $C_m = \{v_j \in \delta_V^-;\ (v_m, v_j) \notin E_V\}$ ($\mathcal{O}(|\delta_V^-| \log |\delta_V^-|)$) and compute $\min_{z \in \Theta_m \cap \Theta_j} \mathcal{E}(z)$ for every $v_j \in C_m$ ($\mathcal{O}(|\delta_V^-|)$). Estimating both $|\delta_V^+|$ and $|\delta_V^-|$ with n, the whole edge switching method is in $\mathcal{O}(\hat{d}_V * n^3 \log n)$.

4.3. Affiliation network modification algorithm with the neighbourhood selection

The aim of the algorithm is to achieve l-degree anonymized bipartite graph H_L^* by modifying the original graph H_L. According to *Definition 3.4*, it is necessary to increase the degree of locations from the set δ_L^+. The straightforward solution is to select $l - deg(z)$ vertices $v \in V$ for every $z \in \delta_L^+$, such that $(v, z) \notin E_L$ and add those edges to E_L. The crucial part is the selection of the most suitable vertices $v \in V$, as in the social network modification algorithm.

As it is presented in (Cho et al., 2011), there is a correlation between the relationship in GSN and the mobility of users. Moreover, the social ties were used for predicting the user's location. We observe the situation from the terms of locations. Having a location $z \in L$, we look for a user $v_i \in V$, who is likely to visit z in the future. Let $W_z = \{v \in V :\ (v, z) \in E_L\}$. Then the most suitable vertex for the edge addition $v_i \in V$, $(v_i, z) \notin E_L$, is searched among the vertices, that are connected with vertices from W_z in G_V. More precisely, the v_i is a member of the set $\overline{W_z} = \{v \in V :\ \exists w \in W_z :\ (v, w) \in E_V\ \&\ (v, z) \notin E_L\}$. The addition procedure is called the *Neighbourhood addition algorithm* and is described in detail in *Algorithm 3*.

The *Neighborhood addition algorithm* finishes on the input z when $deg(z) = l$ or all vertices from $\overline{W_z}$ are linked with z. Hence, after the algorithm is run on all $z \in \delta_L^+$, there can still be locations with $deg(z) < l$. Before adding the rest of the links randomly, we reduce the number of newly added edges with the sequence of the edge switching operations. The distribution of the

Algorithm 3 Neighbour addition

Input: $z \in \delta_L^+$, E_L, l
Output: modified E_L

Set $W_z = \{v \in V : (v, z) \in E_L\}$.
while $deg(z) < l$ & $W_z \neq \emptyset$ **do**
    Set $v_T \in W_z$ to be the user that has done the greatest number of check-ins at location z among all users from W_z
    $W_z = W_z \setminus \{v_T\}$
    Set $N_{T,z} = \{v \in V;\ (v_T, v) \in E_V$ & $(v, z) \notin E_L\}$.
    while $deg(z) < l$ & $N_{T,z} \neq \emptyset$ **do**
        randomly select $w \in N_{T,z}$
        $N_{T,z} = N_{T,z} \setminus \{w\}$
        $E_L = E_L \cup \{(w, z)\}$
    end while
end while

number of check-ins in GSN is power-law (Gao et al., 2012). A few locations have many check-ins while most of the locations have few check-ins. Locations with a very high number of check-ins correspond to the locations with a very high degree in H_L. Moreover, we experimentally found that the high degree locations have the high entropy. Thus they are likely to be locations with insignificant importance for visitors. Thus, for every location z_h with the high degree, we randomly select a part of its edges and switch the location node in them for some location node from δ_L^+. The *High degree switching algorithm* is described in detail in *Algorithm 4*.

Algorithm 4 High degree switching

Input: z_h location with high degree, δ_L^+, E_L, p_z
Output: modified E_L

randomly select a subset $E_{z_h} \subset \{(v, z_h) \in E_L;\ v \in V\}$ such that $|E_{z_h}| < p_z$
$\forall (v, z_h) \in E_{z_h}$: find $z \in \delta_L^+$ such that $(v, z) \notin E_L$
remove (v, z_h) from E_L
add (v, z) into E_L

4.3.1. Complexity

For every $z \in \delta_L^+$ we find a set of neighbours W_z, which takes $\mathcal{O}(m)$ steps in the worst case, where $m = |L|$. All $v_T \in W_z$ are sorted in the descend order according to the number of their check-ins $q_T(z)$. The sorting is in $\mathcal{O}(\hat{d}_L \log \hat{d}_L)$, where $\hat{d}_L$ is the maximal degree of the location node in H_L. Then for every $v_T \in W_z$, the following operations are done: the set $N_{T,z}$ is set ($\mathcal{O}(n \log n)$) and at most l edges are added. Since the upper estimate for $\hat{d}_L$ is n and usually $l < n$, the complexity of the neighbour addition is $\mathcal{O}(|\delta_L^+| * m) + \mathcal{O}(|\delta_L^+| * n^2 \log n)$. Moreover, m is larger than n and $|\delta_L^+|$ can be estimated with m in the worst case. Hence the complexity equals $\mathcal{O}(m^2 * n \log n)$.

The high degree switching method is clearly less demanding than the addition, since it deals only with few highly degree locations and it goes only once through the set δ_L^+.

5. Experimental results

In this section, the implementation of the algorithm and results of the accomplished experiments with two real datasets are presented. All experiments were performed on a PC running Windows 10 operating system with 16 GB RAM and 3,2 GHz processor. The programs were written in Matlab 9.6.0.1214997 (R2019a).

5.1. Tested networks and data preprocessing

We tested the algorithm on the datasets of real geosocial networks Gowalla and Brightkite, where users shared their locations by checking-in. The datasets were collected by Cho et al. in (Cho et al., 2011). Each dataset is composed of two text documents, one containing the checked-in information and the second the social ties in the network. Since the whole Gowalla dataset contains nearly 200,000 nodes and over 6 million check-ins, our experiments were run only on samples of both networks. The algorithm was tested on several samples with a number of nodes between 1,101 and 10,101 and number of top locations between 2,154 and 19,317. *Table 2* summarizes the main characteristics of the tested networks.

Before the actual run of the algorithm, the data were preprocessed to form two graphs G_V and H_L. The graph G_V contained only the social relationships. On the other hand, the graph H_L involved the user-location links, the number of check-ins $q_i(z)$ that user v_i had done at location z, the total

Gowalla						
$\|V\|$	1,101	2,101	3,101	4,101	5,101	10,101
$\|E_V\|$	10,240	22,253	31,303	46,577	63,329	142,272
$\|L\|$	2,522	4,772	7,035	8,993	10,884	18,142
$\|E_L\|$	2,886	5,371	7,956	10,364	12,785	22,622
Brightkite						
$\|V\|$	1,101	2,101	3,101	4,101	5,101	10,101
$\|E_V\|$	8,028	16,344	25,234	30,085	35,924	71,904
$\|L\|$	2,154	4,195	6,383	8,550	10,198	19,317
$\|E_L\|$	2,778	5,435	8,112	10,765	13,059	25,267

Table 2: Sample characteristics of the Gowalla and Brightkite datasets used in the experiments

number of check-ins at location z $C(z)$ and the value of location entropy $\mathcal{E}(z)$.The location entropy was computed using the relation in *Definition 3.6*.

5.2. Top location model

In real datasets, a large amount of locations is visited only once and is meaningless for further data analysis. Thus, it is necessary to extract the most representative locations for each user from the dataset. The (k, l)-degree anonymization model, as well as the algorithm, were proposed with no regards to the location model. The top location model, based on the frequency of a user visit, was used in our implementation. For every user the three locations with the highest $C(z)$ were extracted from the data, processed and stored in H_L. The top location models were also used in recently presented researches dealing with the k-anonymity of GSN (Li et al., 2016; Masoumzadeh and Joshi, 2013). Although considering only the top locations caused an information loss, the data are still valuable for a statistical analysis and further research. Moreover, the top location model is not too large for performing the (k, l)-degree anonymization within a reasonable time.

5.3. Measures

In order to evaluate our results, we use several structure metrics, presented in (Casas-Roma et al., 2017) and a metric for the information loss, introduced in (Li et al., 2016). Since the graphs G_V and H_L were anonymized separately, we separately evaluate the results of the anonymization process

for G_V and H_L. The information loss measure and the average degree are computed for both networks. The average distance, transitivity and the largest eigenvalue of the adjacency matrix are computed only for the social network since measuring them in the affiliation network is not valid.

Let $G = (V, E)$ be a graph. The *information loss* $z_{eb}(G, G^*)$ is defined as the normalized cardinality of the symmetric difference between the original edge set E and the modified edge set E^*:

$$z_{eb}(G, G^*) = \frac{(E \setminus E^*) \cup (E^* \setminus E)}{|E|}$$

The *average degree* $\langle deg \rangle(G)$ is the average of degrees of all nodes from V in the graph G:

$$\langle deg \rangle(G, V) = \frac{\sum_{v \in V}^{n} deg(v)}{n}$$

The *average distance* $\langle dist \rangle(G)$ is an evaluation of connectivity. It is defined as the average of the path lengths between each pair of vertices in G:

$$\langle dist \rangle(G) = \frac{\sum_{i,j=1}^{n} dist(v_i, v_j)}{\binom{n}{2}}$$

where $dist(v_i, v_j)$ is the length of the shortest path from v_i to v_j, meaning the number of edges along the path. The *transitivity* $C(G)$ is a probability of revealing the existence of tightly connected communities in the network. It measures the presence of local loops near the vertex, as it is defined in (Girvan and Newman, 2002):

$$C(G) = \frac{3 * (\text{number of triangles on the graph})}{(\text{number of connected triples of vertices})}$$

The *largest eigenvalue* $\lambda(G)$ *of the adjacency matrix* of G is a spectral measure which encodes the information about the cycles of the networks and their diameter.

5.4. Results evaluation

Although the results for G_V and H_L are evaluated separately, the degree sequence anonymization algorithm, social network modification algorithm and affiliation network modification algorithm were processed as one run of the (k, l)-degree anonymization algorithm with variables k, l, n where $n =$

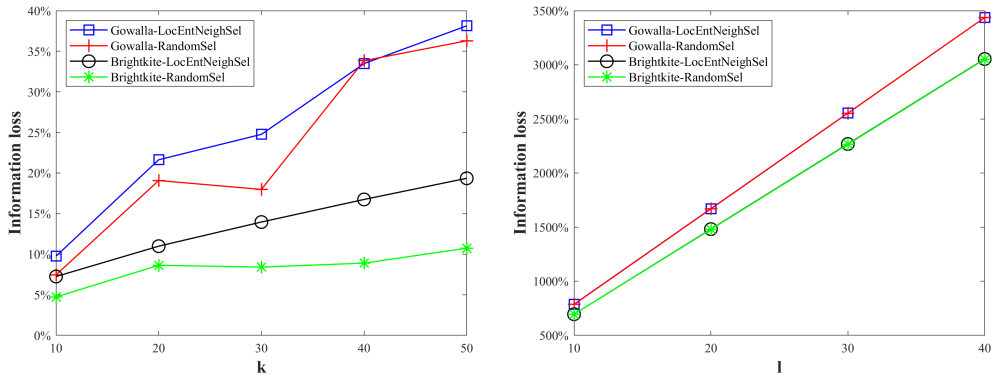
$|V|$. The domains of k and l are $\mathcal{D}(k) = \{10, 20, 30, 40, 50\}$ and $\mathcal{D}(l) = \{10, 20, 30, 40\}$ respectively. The domain of n is shown in *Table 2*. Every run of the algorithm, two parameters were fixed and the remaining one took values from its domain.

We compare the output of the (k, l)-degree anonymization algorithm with location and neighbourhood edge selection, as it is described in *Section 4*, with the original data and the output of (k, l)-degree anonymization algorithm with the random edge selection. In this modified version of the (k, l)-degree anonymization algorithm, the auxiliary vertices needed for the edge edition operations are selected randomly omitting the location entropy values or neighbourhood of vertices. More precisely, in the social network modification algorithm, the minimal entropy selection algorithms are omitted and the auxiliary vertices for the edge addition, edge removal and edge switch are selected randomly. In the affiliation network modification both neighbour addition and high degree switching are omitted and new edges are added randomly only according to the difference between $deg(z)$ and l. In the rest of the paper, the (k, l)-degree anonymization algorithm with the location and neighbourhood edge selection is denoted **LocEntNeighSel**, while the (k, l)-degree algorithm with random selection is denoted **RandomSel**.

Since the greedy algorithm, the maximal entropy selection and the neighbourhood addition include the step where a vertex from a set is randomly selected, the algorithm was run 20 times for every parameter setting. The presented results are the average metric values of the 20 runs of the algorithm. No metric varies considerably with the 20 runs of the algorithm. The most varying values was the values of the transitivity. The relative standard deviation of its values in the 20 runs was 2.7% in the worst case. The relative standard deviation of the values of the other metrics in the 20 runs was under 1%.

The information loss is naturally increasing with the increasing parameters k and l, as it is shown in *Figure 4*. *Figure 4a* shows the increase of $z_{eb}(G_V, G_V^*)$ for growing k and fixed $l = 10$, $n = 3101$. The increase of $z_{eb}(G_V, G_V^*)$ is slower with the Brightkite data, **RandomSel** gets better results for both datasets. The $z_{eb}(G_V, G_V^*)$ is quite good for all values of k, even the worst case for Gowalla with $k = 50$ equals to 38%.

When $k = 20$ and $n = 3101$, then $z_{eb}(H_L, H_L^*)$ grows linearly in l for both algorithms and both datasets, as it is shown in *Figure 4b*. Unfortunately, the information loss is huge, even in the best case it equals 695% with $l = 10$. There is no difference between the results of **RandomSel** and



a) $z_{eb}(G_V, G_V^*)$ with $n = 3101$, $l = 10$

b) $z_{eb}(H_L, H_L^*)$ with $n = 3101$, $k = 20$

c) $z_{eb}(G_V, G_V^*)$ with $k = 20$, $l = 10$

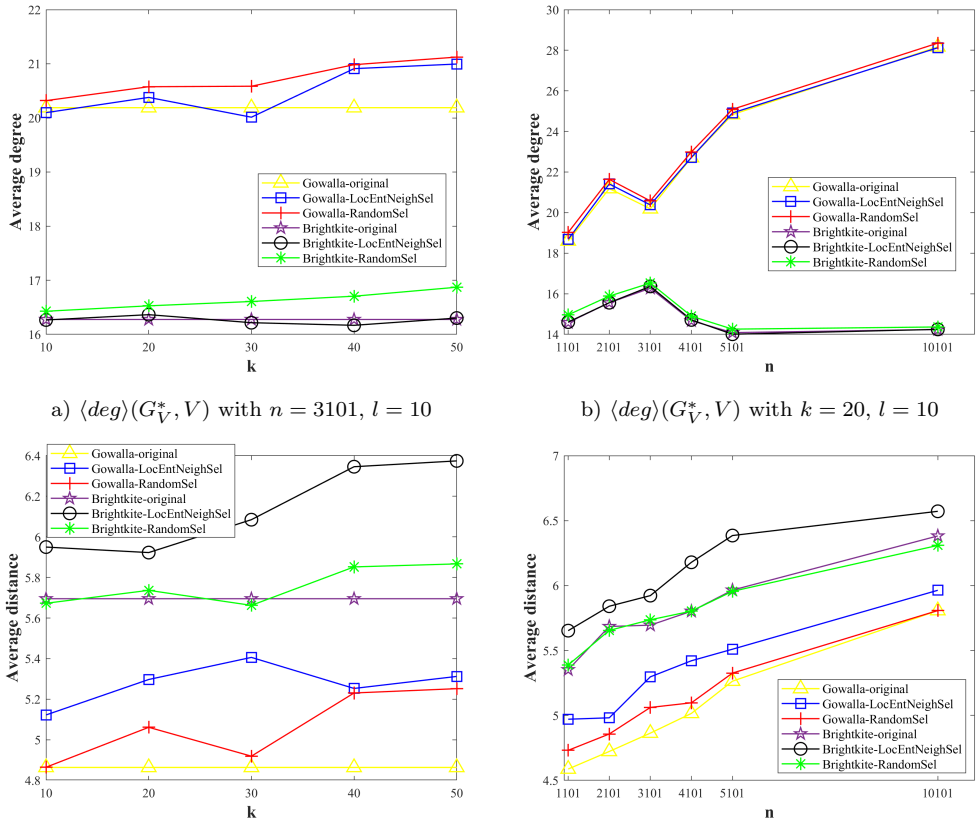
d) $z_{eb}(H_L, H_L^*)$ with $k = 20$, $l = 10$

Figure 4: Information loss measurement.

LocEntNeighSel.

Fixing $k = 20$ and $l = 10$, we measured the dependency of $z_{eb}(G_V, G_V^*)$ and $z_{eb}(H_L, H_L^*)$ on the size of the network n, as it is shown in *Figure 4c)d)*. Both $z_{eb}(G_V, G_V^*)$ and $z_{eb}(H_L, H_L^*)$ decrease for Gowalla in growing n. For Brightkite, the best results are also with $n = 10,101$ for both graphs. The explanation is that in larger networks, every user node is more likely to meet other $k - 1$ vertices with the same degree and every location is more likely to be visited by l users. Hence, the number of the necessary edge edits decreases with growing n and fixed k and l for both G_V and H_L. It indicates the algorithms' usability in large networks. **RandomSel** shows better results for both datasets. The reason for getting a smaller information loss with **RandomSel** is that the edge edits equally distributed in the graph

positively affect the data utility.



a) $\langle deg \rangle(G_V^*, V)$ with $n = 3101$, $l = 10$

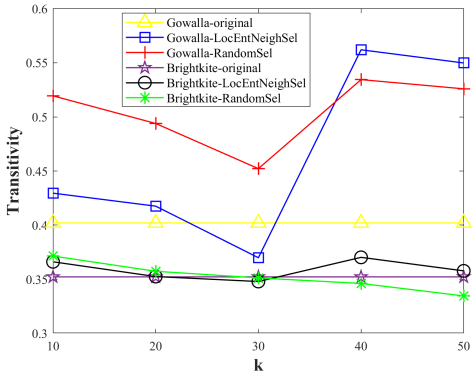
b) $\langle deg \rangle(G_V^*, V)$ with $k = 20$, $l = 10$

c) $\langle dist \rangle(G_V^*)$ with $n = 3101$, $l = 10$

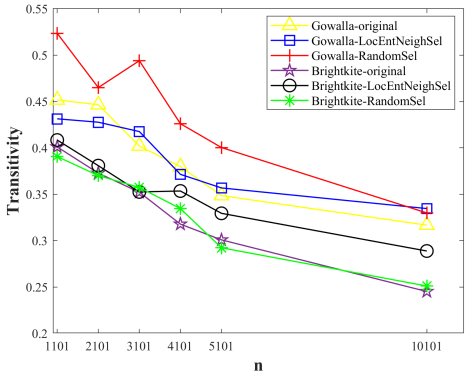
d) $\langle dist \rangle(G_V^*)$ with $k = 20$, $l = 10$

Figure 5: Average degree and average distance measurement for G_V^*.

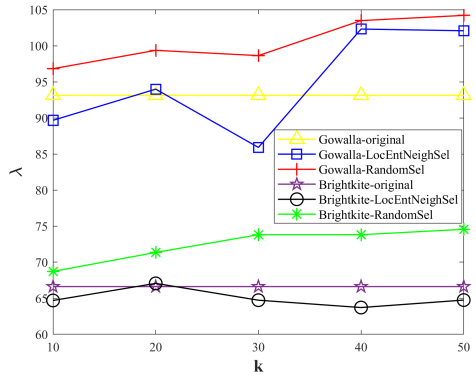
On the other hand, **LocEntNeighSel** kept the structural measures $C(G_V^*)$, $\lambda(G_V^*)$ and $\langle deg \rangle(G_V^*, V)$ closed to the values of the original graph, mainly for the Brightkite dataset, as it is shown in *Figure 5* and *Figure 6*. Only $\langle dist \rangle(G_V^*)$ provided better results with the random edge modifications. The distance between the original and the **LocEntNeighSel** values was not growing with the increasing values of k or n. There is no worsening trend neither in k nor n in any metric for **LocEntNeighSel**. It indicates the usability of **LocEntNeighSel** in large networks. On the other hand, the distance between the original and the **RandomSel** values is visibly enlarging at $\langle deg \rangle(G_V^*, V)$ and $\lambda(G_V^*)$ with the growing k.



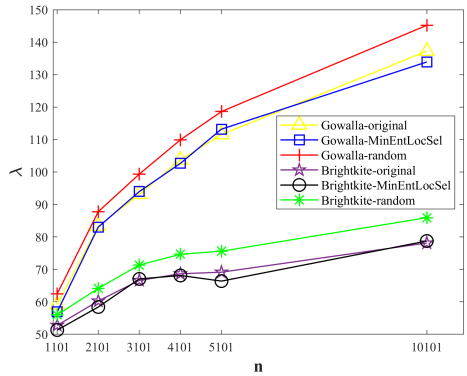
e) $C(G_V^*)$ with $n = 3101$, $l = 10$



f) $C(G_V^*)$ with $k = 20$, $l = 10$



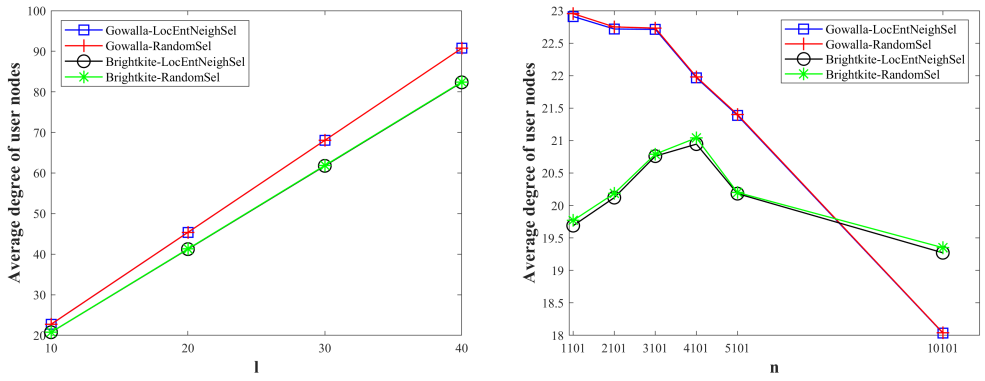
g) $\lambda(G_V^*)$ with $n = 3101$, $l = 10$



h) $\lambda(G_V^*)$ with $k = 20$, $l = 10$

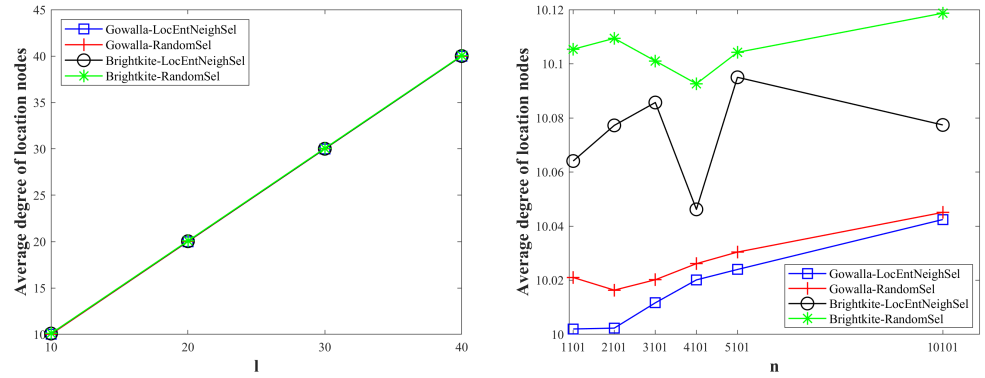
Figure 6: Transitivity and λ measurement for G_V^*.

In the affiliation network, we only measured the average degree of user nodes $\langle deg \rangle(H_L^*, V)$ and the average degree of location nodes $\langle deg \rangle(H_L^*, L)$, as it is shown in *Figure 7*. In all cases, **LocEntNeighSel** provided slightly better results, which is caused by the usage of switching operation. However, the anonymized degree values were still about five times higher than the original values, since the large amount of new edges was added into the network. In future research, this could be solved by clustering locations into larger regions according to their geographical coordinates before the actual anonymization. It would cause a higher information loss in the preprocessing stage, but it would improve the data utility in the anonymization process since the initial degrees of regions would be higher.



a) $\langle deg \rangle(H_L^*, V)$ with $n = 3101$, $k = 20$

b) $\langle deg \rangle(H_L^*, V)$ with $k = 20$, $l = 10$



c) $\langle deg \rangle(H_L^*, L)$ with $n = 3101$, $k = 20$

d) $\langle deg \rangle(H_L^*, L)$ with $k = 20$, $l = 10$

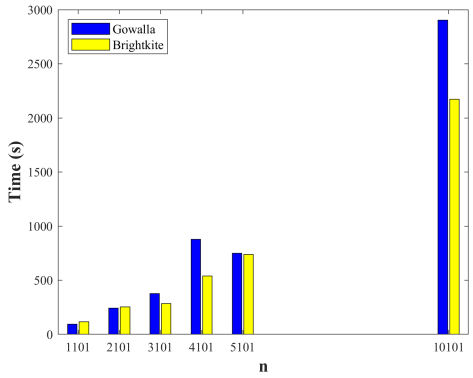
Figure 7: Average degree measurement for H_L.

5.5. Runtime

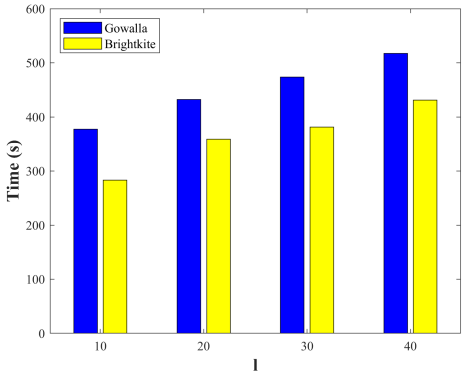
The runtime was measured for the **LocEntNeighSel** algorithm without the preprocessing stage. Within the 20 runs of the algorithm with the same parameter setting, the runtime of one run of the algorithm varied insignificantly. The relative standard deviation of the runtime was up to 2%. Runtime values in *Figure 8* are the average values of the runtimes in the 20 runs of the algorithm with the same parameter setting.

Following from the complexity computations, it grows fast with the growing n, as it is shown in *Figure 8a*. Nevertheless, the slowest computations took around 50 minutes for the sample of the Gowalla data with 10,101 nodes. Hence, the algorithm is usable even for larger networks in real time. The growing l implies only a small linear increase in runtime, as it is shown

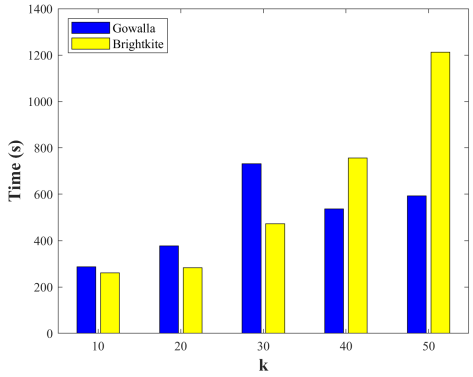
in *Figure 8b*. The networks structure has an impact on the dependence of the runtime on k, as it is shown in *Figure 8c*. While with the Brightkite data the slowest run of the algorithm was with $k = 50$, with the Gowalla data it was with $k = 30$.



a) runtime with fixed $k = 20$, $l = 10$



b) runtime with fixed $n = 3101$, $k = 20$



c) runtime with fixed $n = 3101$, $l = 10$

Figure 8: Runtime measurement.

5.6. Limitations of the experiments

Both the location entropy edge selection and neighbourhood selection in the (k, l)-degree anonymization algorithm were defined to increase the probability that the anonymized $\mathcal{G}^*$ would resembles the future development of the network. Some of the newly added links are likely to really appear in $\mathcal{G}$ in the future. Thus, it reduces the information loss caused by the anonymization. To test the similarity of the anonymized $\mathcal{G}^*$ at time t_1 with

the structure of $\mathcal{G}$ at time t_2, it would be necessary to have several samples of the same networks from different time periods. The requisite data were not available in this research.

Although our experimental results indicate that the algorithm is feasible on larger networks, the maximum tested network sample contains only 10,101 users. It was mainly caused by the limitations of the Matlab program and data management in our implementation.

As it has been mentioned above, the top location model causes the information loss, which was not measured. Future research could involve studying and measuring the effect of the top location model on the data utility, as well as considering other location models.

6. Conclusion

In this study, we addressed the problem of preserving individual's privacy in publishing the geosocial network datasets. The geosocial network was represented as a combination of the social network and the affiliation network connecting users with the checked-in locations. The new (k, l)-degree anonymization method was introduced to prevent the re-identification attack in geosocial networks. Two versions of the proposed algorithm, modifying the edge set of the network, were analysed by running the experiments on real-world datasets Gowalla and Brightkite.

The location entropy edge selection, used in the graph modification method, was shown to improve the preservation of structural properties of the original network in the anonymization process. Since the data utility were preserved better in larger network samples, the algorithm is highly recommended for testing in larger networks.

Future research will involve the analysis of the compatibility of the algorithm with location generalization methods. Clustering locations into regions according to their geographical coordinates will increase the initial vertex degree of the location nodes and decrease the information loss in the anonymization. In addition, it could focus on the application of other location models for extracting representative locations for each user.

Acknowledgements

The financial support of the Specific Research Project "Information and Knowledge Management and Cognitive Science in Tourism" of FIM UHK

is gratefully acknowledged. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Alrayes, F., Abdelmoty, A., 2016. Towards location privacy awareness on geosocial networks, in: 2016 10th International Conference on Next Generation Mobile Applications, Security and Technologies (NGMAST), IEEE. pp. 105–114. doi:10.1109/NGMAST.2016.26.

Calderoni, L., Palmieri, P., Maio, D., 2015. Location privacy without mutual trust: The spatial bloom filter. Comput Commun 68, 4–16. doi:10.1016/j.comcom.2015.06.011.

Carbunar, B., Rahman, M., Pissinou, N., 2013. A survey of privacy vulnerabilities and defenses in geosocial networks. IEEE Commun Mag 51, 114–119. doi:10.1109/mcom.2013.6658662.

Carbunar, B., Sion, R., Potharaju, R., Ehsan, M., 2014. Private badges for geosocial networks. IEEE Trans Mob Comput 13, 2382–2396. doi:10.1109/tmc.2013.95.

Casas-Roma, J., Herrera-Joancomartí, J., Torra, V., 2017. K-Degree anonymity and edge selection: Improving data utility in large networks. Knowl Inf Syst 50, 447–474. doi:10.1007/s10115-016-0947-7.

Chakraborty, S., Tripathy, B.K., 2016. Alpha-anonymization techniques for privacy preservation in social networks. Soc Netw Anal Min 6. doi:10.1007/s13278-016-0337-x.

Cho, E., Myers, S.A., Leskovec, J., 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. doi:10.1145/2020408.2020579.

Cranshaw, J., Toch, E., Hong, J., Kittur, A., Sadeh, N., 2010. Bridging the gap between physical location and online social networks, in: Proceedings of the 12th ACM International Conference on Ubiquitous Computing, ACM Press. pp. 119–128. doi:10.1145/1864349.1864380.

Fire, M., Kagan, D., Puzis, R., Rokach, L., Elovici, Y., 2012. Data mining opportunities in geosocial networks for improving road safety, in: 2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel, IEEE. pp. 1–4. doi:10.1109/EEEI.2012.6377049.

Gao, H., Tang, J., Liu, H., 2012. Exploring social-historical ties on location-based social networks, in: Sixth International AAAI Conference on Weblogs and Social Media, pp. 114–121.

Girvan, M., Newman, M.E.J., 2002. Community structure in social and biological networks. Proceedings of the National Academy of Sciences 99, 7821–7826. doi:10.1073/pnas.122653799.

Golle, P., Partridge, K., 2009. On the Anonymity of Home/Work Location Pairs, in: International Conference on Pervasive Computing. Springer Berlin Heidelberg. volume 5538, pp. 390–397. doi:10.1007/978-3-642-01516-8_26.

Hartung, S., Hoffmann, C., Nichterlein, A., 2014. Improved upper and lower bound heuristics for degree anonymization in social networks, in: International Symposium on Experimental Algorithms, Springer International Publishing. pp. 376–387. doi:10.1007/978-3-319-07959-2_32.

Hay, M., Miklau, G., Jensen, D., Towsley, D., Li, C., 2010. Resisting structural re-identification in anonymized social networks. VLDB J 19, 797–823. doi:10.1007/s00778-010-0210-x.

Kotzanikolaou, P., Patsakis, C., Magkos, E., Korakakis, M., 2016. Lightweight private proximity testing for geospatial social networks. Comput Commun 73, 263–270. doi:10.1016/j.comcom.2015.07.017.

Li, Y., Li, Y., Xu, G., 2016. Protecting private geosocial networks against practical hybrid attacks with heterogeneous information. Neurocomputing 210, 81–90. doi:10.1016/j.neucom.2015.08.132.

Liu, B., Hengartner, U., 2013. Privacy-preserving social recommendations in geosocial networks, in: 2013 Eleventh Annual Conference on Privacy, Security and Trust, IEEE. pp. 69–76. doi:10.1109/pst.2013.6596038.

Liu, K., Terzi, E., 2008. Towards identity anonymization on graphs, in: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, ACM Press. pp. 93–106. doi:10.1145/1376616.1376629.

Lu, X., Song, Y., Bressan, S., 2012. Fast identity anonymization on graphs, in: International Conference on Database and Expert Systems Applications. Springer Berlin Heidelberg. volume 7446, pp. 281–295. doi:10.1007/978-3-642-32600-4_21.

Ma, T., Jia, J., Xue, Y., Tian, Y., Al-Dhelaan, A., Al-Rodhaan, M., 2018. Protection of location privacy for moving kNN queries in social networks. Appl Soft Comput 66, 525–532. doi:10.1016/j.asoc.2017.08.027.

Masoumzadeh, A., Joshi, J., 2013. Top location anonymization for geosocial network datasets. Trans Data Privacy 6, 107–126.

Medková, J., 2018. Composition attack against social network data. Comput Secur 74, 115–129. doi:10.1016/j.cose.2018.01.002.

Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B., 2007. Measurement and analysis of online social networks, in: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, ACM Press. pp. 29–42. doi:10.1145/1298306.1298311.

Pontes, T., Vasconcelos, M., Almeida, J., Kumaraguru, P., Almeida, V., 2012. We know where you live: Privacy characterization of foursquare behavior, in: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, ACM Press. pp. 898–905. doi:10.1145/2370216.2370419.

Rahman, M., Ballesteros, J., Carbunar, B., Rishe, N., Vasilakos, A.V., 2013. Toward preserving privacy and functionality in geosocial networks, in: Proceedings of the 19th Annual International Conference on Mobile Computing & Networking, ACM Press. pp. 207–210. doi:10.1145/2500423.2504577.

Scellato, S., Noulas, A., Mascolo, C., 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. doi:10.1145/2020408.2020575.

Shokri, R., Theodorakopoulos, G., Le Boudec, J.Y., Hubaux, J.P., 2011. Quantifying Location Privacy, in: 2011 IEEE Symposium on Security and Privacy, IEEE. pp. 247–262. doi:10.1109/SP.2011.18.

Shokri, R., Theodorakopoulos, G., Troncoso, C., Hubaux, J.P., Le Boudec, J.Y., 2012. Protecting location privacy: optimal strategy against localization attacks, in: Proceedings of the 2012 ACM conference on Computer and communications security - CCS '12, ACM Press. pp. 617–627. doi:10.1145/2382196.2382261.

Siddula, M., Li, L., Li, Y., 2018. An empirical study on the privacy preservation of online social networks. IEEE Access 6, 19912–19922. doi:10.1109/ACCESS.2018.2822693.

Wernke, M., Skvortsov, P., Drr, F., Rothermel, K., 2014. A classification of location privacy attacks and approaches. Pers Ubiquit Comput 18, 163–175. doi:10.1007/s00779-012-0633-z.

Xue, D., Wu, L.F., Li, H.B., Hong, Z., Zhou, Z.J., 2017. A novel destination prediction attack and corresponding location privacy protection method in geo-social networks. Int J Distrib Sens N 13. doi:10.1177/1550147716685421.

Zhang, J.D., Ghinita, G., Chow, C.Y., 2014. Differentially Private Location Recommendations in Geosocial Networks, in: 2014 IEEE 15th International Conference on Mobile Data Management, IEEE, Brisbane, Australia. pp. 59–68. doi:10.1109/MDM.2014.13.

Zheleva, E., Sharara, H., Getoor, L., 2009. Co-evolution of social and affiliation networks, in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Press. pp. 1007–1016. doi:10.1145/1557019.1557128.