Acta Technica *62* No. 1B/2017, 437–446 

# Human resource data location privacy protection method based on prefix characteristics

Yulong Qi[1,2], Enyi Zhou[1]

**Abstract.** With the arrival of the big data era, a large amount of human resource data location information is implicitly collected. They cross reference with the spatial and temporal data the user initiatively published, which has caused new human resource privacy disclosure problem in the big data era. The existing location privacy protection mechanism cannot effectively protect the privacy of the users due to the fact that it does not take into consideration that the implicitly collected spatial-temporal data can cross reference with the human resource location data that is initiatively distributed by the user. The privacy protection problem in the implicitly collected temporal and spatial data is defined and studied for the first time, and the privacy protection framework that takes the characteristics of the prefix into consideration is put forward. In particular, a nested loop algorithm which takes the prefix filter into account is proposed to discover the records in the implicitly collected temporal and spatial data that may disclose the privacy of the human resource data. And the dummy filling method based on the frequent moving prefix is put forward to eliminate these records. In addition, a more efficient reverse aprior algorithm and graph based dummy filling algorithm are put forward respectively. Finally, the proposed algorithm is fully tested in a number of real data sets. The experimental results show that these algorithms have relatively high protective effect and performance.

**Key words.** Implicit privacy, temporal and spatial data, privacy protection, human resource data.

## 1. Introduction

In the era of big data, with the development of the location technology, the location-based services are becoming increasingly popular, and the user's temporal and spatial data is distributed through all types of services. While the users take the initiative to publish their own temporal and spatial behaviors through the sign-in and other mobile social network services, a large amount of temporal and spatial data

[1]School of Management, Xi'an University of Architecture and Technology, Xi'an, Shaanxi, 710055, China
[2]Corresponding author

that records people's behaviors when they use mobile phones to make phone calls, send and receive short messages is implicitly collected by the mobile communication operators at the same time [1–2]. As the temporal and spatial data between the mobile phones and mobile communication operators is collected by the mobile phone base station automatically, these implicitly collected data is characterized by large data volume and containing the human behaviors, which has played a key role in the social issues such as the human resource distribution, as well as the important social applications such as the human resource referral [3–6], human resource level [7] and so on. However, these implicitly collected temporal and spatial data will exposes the sensitive privacy information of the user such as the personal identity, purpose of action, health status, interests and hobbies and many other aspects through the cross reference to the temporal and spatial data initially published by the user [8–9]. In recent years, with the enhancement of the human resource privacy concept and the soundness of the laws and regulations in the publishing and using of the data, it is necessary to first eliminate the records which may expose the privacy of the human resource data before the implicitly collected spatial and temporal data is used for scientific research and data mining.

In order to ensure that the human resource sensitive information is not compromised, a large amount of work for the protection of the temporal and spatial data privacy is committed to anonymize the temporal and spatial data that may expose the human resource sensitive information. For example, the k anonymity on the location and trajectory data and other methods can generalize the location records of the user at the specific time range and spatial region, so that the attacker cannot recognize the specific user in a certain time range and spatial region [10–12]. However, these methods do not take into account that the attacker can refer to the temporal and spatial data that is initially published by the user and find the records that can reveal the human resource data privacy from the implicitly collected temporal and spatial data. Therefore, the temporal and spatial data sets which are protected by these methods can still disclose the data of the human resource data privacy.

In order to meet the aforementioned challenges, in this paper, the implicit privacy which is independent of the data initially published by the user is defined on the implicitly collected data set, so as to ensure that no matter how much data initially published by the users is collected by the attacker, the spatial and temporal data sets after the privacy protection will not reveal the additional information.

In this paper, the first section mainly introduces the related technology of the privacy protection in the publishing of the temporal and spatial data. The second section introduces the implicit privacy problem and its prefix feature protection framework. The third section introduces the corresponding discovery and elimination algorithm. And the fourth section is the demonstration of the experimental results.

## 2. Algorithm and analysis

This section describes the algorithms of discovery and elimination of $(\varepsilon, k)$ privacy disclosure, respectively.

### *2.1. Discovery of* $(\varepsilon, k)$ *privacy disclosure*

Firstly, we introduce the nested loop algorithm of the discovery of $(\varepsilon, k)$ privacy disclosure that takes the prefix filter into consideration, and then point out its defect in lack of efficiency, and put forward a more efficient reverse aprior algorithm.

According to the privacy protection parameters $\varepsilon$ and $k$, the basic idea of the temporal and spatial data of the discovery of the $(\varepsilon, k)$ privacy disclosure is the set composed of all $k$ temporal and spatial points by enumeration, and check whether each combination is associated with a unique user. To this end, the prefix filter based nest loop (referred to as PF–NL for short) is used to implement $k$ nested loop for the temporal and spatial points by enumerating the k-bit non-repeating numerical numbers, and conducts pruning in the enumeration using the prefix filter method. Firstly, we introduce the important properties of the $(\varepsilon, k)$ privacy.

**Property 1**. Denote the set of all the temporal and spatial points that expose the $(\varepsilon, k)$ privacy as $U_k$, and denote the set of all the temporal and spatial points with the size of $k$ that can uniquely associate with moving prefix as $u_k$, then $U_k = u_1 \cup \ldots \cup u_k$.

We skim the trivial proof of Property 1, which shows that, we can start from the set of temporal and spatial points with the size 1, until we enumerate all the sets of temporal and spatial points with the size not exceeding $k$. To this end, we number $n$ temporal and spatial points from 1, and each temporal and spatial set with the size k can be regarded as a numerical number with $k$ bits and scale $n$, and each of the same set of the temporal and spatial points has the unique representation method after sorted according to the sequence.

Thus, the set of temporal and spatial points with the size $k$ $(k > 1)$ has prefixes with the length $\{1, \ldots, k-1\}$. For example, we can denote the six effective temporal and spatial points $p_{\mathrm{A},\mathrm{T}_1}$, $Sp_{\mathrm{A},(\mathrm{T}_1+\varepsilon_1/2)}$, $Sp_{\mathrm{B},\mathrm{T}_1}$, $Sp_{\mathrm{C},\mathrm{T}_2}$, $Sp_{\mathrm{D},\mathrm{T}_2}$ and $Sp_{\mathrm{merge}}$ in Table 1 with the number 1 to 6, respectively. Considering the ($\varepsilon = 0$, $k = 3$) privacy, for the temporal and spatial point set $\{1, 2, 3\}$ with the size 3, it has the prefix $\{1, 2\}$. It is known that $|Sp_1 \cap Sp_2| = 0$, therefore, the set of temporal and spatial points with this prefix will definitely not expose the (0,3) privacy. In the prefix filter method, we avoid enumerating the temporal and spatial point set containing such prefix.

Combined with the aforementioned basic ideas and prefix filter optimization method, we put forward the basic algorithm PF–NL that discovers the implicit privacy violation. In the algorithm PF–NL, we enumerate the temporal and spatial point set *num* with the size $k$ in turn, let the maximum value that each corresponds to be denoted by *bound* array (Line 1). When the size of each moving prefix set included in a certain set of temporal and spatial points after intersection is 1, we add it into the set $R$ of all the temporal and spatial points that violate the privacy requirements (Line 3 and Line 4). Thus, the process of generating a new set of new temporal and spatial points is the process of adding an element in the array (Line 8∼Line 13). It is worth noting that, when we are checking whether a certain temporal and spatial point has exposed the $(\varepsilon, k)$ privacy, the prefix that it may exist is calculated (Line 6), and the prefix is filtered out when a new set of temporal and spatial points is generated (Line 8). When the new temporal and spatial points cannot be generated, the algorithm PF–NL ends (Line 10).

Algorithm PF–NL uses the prefix filter to try to filter out the temporal and spatial point set with empty intersection in the enumeration, which has accelerated the searching process of the temporal and spatial point set that violates the $(\varepsilon, k)$ privacy. But as many prefixes can be used in the filter, we cannot record each one, it has still conducted a lot of unnecessary work. Next, we will introduce the reverse aprior algorithm, making use of the breadth-first search idea, to avoid containing the relatively small temporal and spatial point set that does not violate the $(\varepsilon, k)$ privacy in the search for a larger set of temporal and spatial points.

Algorithm 2 shows the reverse aprior (referred to as RA algorithm for short), in which we use $u_i, z_i$ and $c_1$ to represent the temporal and spatial point set that violates the $(\varepsilon, k)$ privacy with the size i, cannot violate the $(\varepsilon, k)$ privacy, and may violate the $(\varepsilon, k)$ privacy in the larger set of temporal and spatial points.

In the RA algorithm, we first check all the temporal and spatial points, and add the temporal and spatial points that violate the $(\varepsilon, k)$ privacy into $u_1$; as the temporal and spatial points must contain the prefix, they are all added into $c_1$ (Line 1).

Next, we consider the set composed of more temporal and spatial points. The relatively large temporal and spatial point set consist of the temporal and spatial point set with relatively smaller probability to violate the $(\varepsilon, k)$ privacy (Line 5). As the set with the size $i+1$ has multiple combination methods, different combinations may lead to different sizes of the temporal and spatial point sets to be tested. In order to reduce the time overhead, we choose two sets of temporal and spatial points with the smallest Cartesian product to generate it (Line 4). In the same time, it is used to check whether the temporal and spatial point set that violates the $(\varepsilon, k)$ privacy should not include smaller temporal and spatial point set that cannot violate the $(\varepsilon, k)$p rivacy (Line 6). In the end, violate the moving prefix sets contained in each temporal and spatial point in the temporal and spatial point set to solve the intersection, and return its size. The temporal and spatial point set with the size 1 is in violation of the $(\varepsilon, k)$ privacy, and added into $u_{i+1}$ 1 (Line 6), while the set of temporal and spatial points with the intersection size of 0 will not violate the $(\varepsilon, k)$ privacy, therefore is added into $z_{i+1}$, and removed from $c_{1+1}$ (Line 8 and Line 9).

Example 1: For Table 1, we first check whether the set of temporal and spatial points which are composed of six effective temporal and spatial points is in violation of the $(\varepsilon, k)$ privacy. Firstly, as $Sp_1$ and$Sp_2$ contains a set each, $u_1 = \{1, 2\}$, $c_1 = \{3, 4, 5, 6\}$. Next, according to the temporal and spatial point set to be tested with the size 2 which is generated by the Cartesian product of $c_1$ and $c_1$, that is {{3,4}, {3,5}, {3,6}, {4,5}, {4,6}..., {5,6}}. After checking, the four pairs of temporal and spatial point set that violates the $(\varepsilon, k)$ privacy is $u_2$ = {{3,4}, {3,5}, {4,6}, {5,6}}, while the temporal and spatial point set $\{3, 6\}$ and $\{4, 5\}$ does not contain the same moving prefix. Therefore, the set of temporal and spatial points containing them cannot violate the $(\varepsilon, k)$ privacy. Hence they are removed from $c_2$. Finally, $c_2 = \emptyset$, $z_2$ = {{3,6}, {4,5}} is obtained. Therefore, algorithm RA ends. It is worth noting that, in general, the larger the set of temporal and spatial points is, the more temporal and spatial point sets we need to check. However, in Example 1, when the size of the set of empty points increases, there is no increase in the set of

temporal and spatial points to be examined, which proves the effectiveness of the RA algorithm intuitively.

### 2.2. *Elimination of the* $(\varepsilon, k)$ *privacy violation temporal and spatial data*

In this section, we introduce the method of dummy filling to eliminate the temporal and spatial data that violates the $(\varepsilon, k)$ privacy. In particular, we introduce the frequent moving prefix based dummy filling method that does not need to look for the temporal and spatial point set that violates the $(\varepsilon, k)$ privacy and the graph based privacy disclosure elimination method that can realize relatively high data effectiveness. Regardless of the set of the temporal and spatial points that have been found to violate the $(\varepsilon, k)$ privacy, a simple method of privacy protection is to perform dummy filling with the same user id to each user for each temporal and spatial point where it exists, for example, to protect the (0,2) privacy disclosure caused in Table 1. After such dummy filling, the following can be obtained $Sp_{\mathrm{A},\mathrm{T}_1} = \{u_1, u_5\}$, $Sp_{\mathrm{A},(\mathrm{T}_1+\varepsilon_1/2)} = \{u_2, u_6\}$, $Sp_{\mathrm{B},\mathrm{T}_1} = \{u_3, u_4, u_7, u_8\}$, $Sp_{\mathrm{C},\mathrm{T}_2} = \{u_1, u_3, u_5, u_7\}$, $Sp_{\mathrm{D},\mathrm{T}_2} = \{u_2, u_4, u_6, u_8\}$, $Sp_{\mathrm{merge}} = \{u_1, u_5, u_2, u_6\}$. At this point, the temporal and spatial data for user $u_5 \sim u_7$ has added 8 entries of dummy, and no longer in violation of (0, 2) privacy. However, this method requires filling 100% (or more) dummy. As an improvement, we add two moving prefix that occur the most frequently for each temporal and spatial point.

Algorithm 3 (frequent moving object, referred to as the FMO algorithm for short) shows the process of the dummy fixing based on the frequent moving object. Given the privacy parameter $(\varepsilon, k)$ with uniqueness, we first find the most frequently occurred two moving prefix, and in this process there are a lot of fast algorithms; secondly, we add these moving prefix to each set of the temporal and spatial points

As the set of the temporal and spatial points of unique privacy disclosure found in Section 3.1 is not taken into consideration, the algorithm FMO performs dummy filling indiscriminately to the temporal and spatial points, resulting in the filling of a lot of data that is not necessary.

Example 2: For the following four temporal and spatial points: $\{u_1, u_6\}, \{u_2, u_3\}, \{u_3, u_4\}, \{u_7, u_8\}$, according to the FMO algorithm, perform dummy filling and obtain $\{u_1, u_6, u_2, u_3\}, \{u_2, u_3\}, \{u_2, u_3, u_4\}$, $\{u_7, u_8, u_2, u_3\}$, and 62.5 % of the data has been filled. However, the first temporal and spatial point clearly does not need to fill any dummy.

Taking the set of temporal and spatial point set that violates the $(\varepsilon, k)$ privacy found in Section 3.1, algorithm 4 shows the graph based dummy filling (referred to as G-DF for short) process. In algorithm G-DF, we regard the human resource location data as a graph, in which each temporal and spatial point represents a node in the graph. If two temporal and spatial points exist in a set of temporal and spatial points that disclose the unique privacy, we add an edge to them in the graph (the second line). We only run algorithm 3 for each connected component in the graph, that is, looking for the most frequent two moving prefixes in each connected component, and then adding them to each node in the connected component (Line 3∼Line 5).

Example 3: It is noted that in Example 2, only the set of temporal and spatial points $\{\{u_2, u_3\}, \{u_3, u_4\}\}$ with the size 2 has violated the $(\varepsilon, k)$ privacy, and Fig. 1 is the case that Example 2 is converted into graph. In this way, according to algorithm G–DF, the data in Example 2 is converted into $\{u_1, u_6\}, \{u_2, u_3\}, \{u_2, u_3, u_4\}, \{u_7, u_8\}$, which has increased 62.5% more data compared with the FMO algorithm, while the G–DF algorithm has only added 12.5 % of the data.

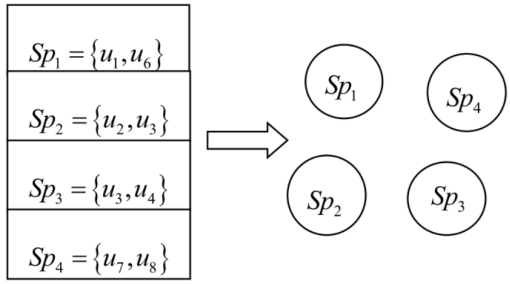


Fig. 1. Convert the temporal and spatial points into graph

## 3. Experiments

We use two real data sets to compare the performance and effectiveness of the two algorithms that discover the temporal and spatial data violating the $(\varepsilon, k)$ implicit privacy and two algorithms that eliminate such violation in the prefix characteristic framework. In particular, we will answer the following questions through the experiment:

(1) How the privacy protection parameter $(\varepsilon, k)$ affect the $(\varepsilon, k)$ privacy in the temporal and spatial data.

(2) The influence of the privacy protection parameter $(\varepsilon, k)$ on the performance of the privacy violation discovery algorithm.

(3) The effect and performance of the privacy protection parameter $(\varepsilon, k)$ on the privacy protection algorithms.

### *3.1. Experimental environment and data set description*

We use Java 1.7 to implement Algorithm 1 $\sim$ Algorithm 4 in this paper. The experimental environment is a Linux server, Intel Xeon E5645 2.4 GHz processor, 128 G RAM, and 1T SATA hard disk.

In addition to the PF–NL method and RA method applied to discover the temporal and spatial set that violates the $(\varepsilon, k)$ privacy in this paper, as well as the FMO and G–DF methods in this paper that are used to eliminate the temporal and spatial point set that violates the $(\varepsilon, k)$ privacy. And other comparative methods include the following:

YCWA[3]: This method is the latest method that adopts the trajectory

anonymization technology to protect the privacy of the temporal and spatial data. It divides the trajectories into dwell point, and protects the privacy information by anonymizing these dwell points. This method mainly focuses on the performance of the trajectory privacy protection.

This method focuses on the data availability of the trajectory anonymity technology. While anonymizing the trajectory, it minimizes the distance between the changed temporal and spatial points and the original temporal and spatial points at the same time.

We use two public data sets, GeoSocial [5] and GeoLife [6] as the implicitly collected temporal and spatial data sets to conduct experiment. Their data size and the number of users are shown in Table 2.

Table 1. Data set

| Data set | Number of record entries | Number of moving prefix |
|---|---|---|
| GeoSocial | 4 M | 18 K |
| GeoLife | 20 M | 17 K |

### *3.2. Comparison of the privacy protection effect*

Figure 2 shows the effects of the spatial and temporal data privacy protection for each method on the GeoSocial and GeoLife data sets under relatively stringent $(\varepsilon, k)$ privacy condition ($\varepsilon_1 = 10\,\text{min},\ \varepsilon_2 = 1\,\text{km},\ k = 10$). Our method RA * G–DF uses the RA method that has the best performance in the discovery phase, and the G–DF method that has the best performance in the elimination phase. We can see that, there are still a large number of temporal and spatial sets that violate the $(\varepsilon, k)$ privacy in the trajectory after the YCWA and SQL–ANON treatment. This is because these methods do not take the privacy disclosure caused by the cross-reference of the implicitly collected temporal and spatial data and the user initially published temporal and spatial data into consideration.

### *3.3. Performance of the algorithm that discovers the* $(\varepsilon, k)$ *privacy*

We compare and verify the running efficiency of the two proposed $(\varepsilon, k)$ implicit privacy discovery algorithms with the real data sets GeoLife and GeoSocial. Figures 3–5 show the run time of the two algorithms in the GeoSocial data set to find all the $(\varepsilon, k)$ privacy violation in different $(\varepsilon, k)$. It can be seen from the comparison of Fig. 3 and Fig. 4, the RA algorithm is 1∼2 orders of magnitude faster than the PF–NL algorithm under the same privacy parameter with the same privacy parameter $(\varepsilon, k)$. Particularly, PF–NL algorithm cannot even calculate the situation when $k > 3$.

For the RA algorithm, we have adjusted the GeoLife data set size and the number of the moving prefix and tested its performance. Figure 3 (c) shows that, the size of GeoLife data set has the greatest impact on the algorithm, as it has changed the number of the temporal and spatial points in the data set; however, the number of

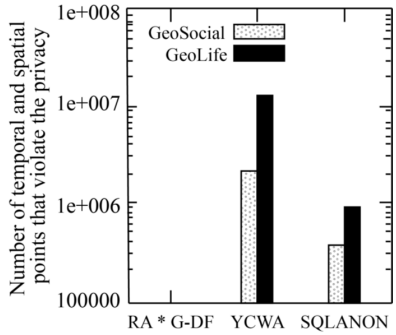


Fig. 2. Comparison of privacy protection effect

prefix has little effect on the algorithm.

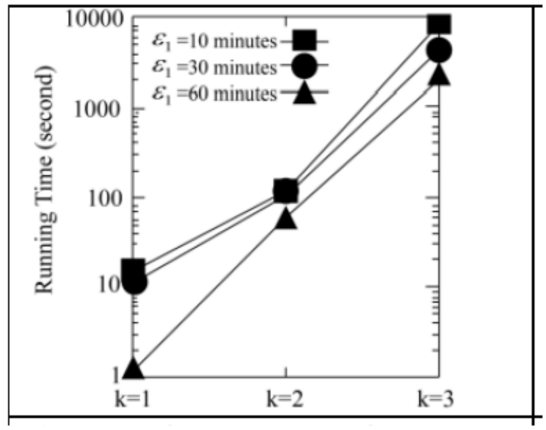


Fig. 3. Performance of PF–NL algorithm on GeoSocial data set ($\varepsilon_2 = 1\,\mathrm{km}$)

## 4. Conclusion

In this paper, the definition of the $(\varepsilon, k)$ implicit privacy in the temporal and spatial data is proposed for the first time in view of the situation of the cross-reference of the data set initially published by the user and the temporal and spatial data set implicitly collected. And the prefix characteristic privacy protection framework is put forward. In particular, two highly efficient algorithms are proposed in this paper to discover the temporal and spatial point set that violates the $(\varepsilon, k)$ privacy. In addition, the dummy filling anonymous protective method is put forward in this paper. In order to improve the effectiveness of data, this paper further proposes the graph based dummy filling method. The full experiment on the real data set shows that, the proposed algorithm is highly efficient. In the future work, we will further improve the performance of the proposed method in the paper.

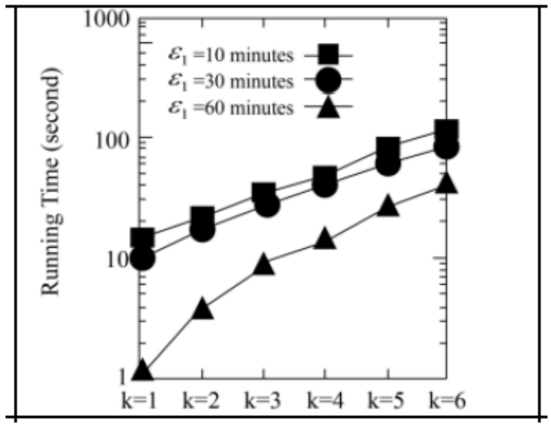


Fig. 4. Performance of RA algorithm on GeoSocial data set ($\varepsilon_2 = 1\,\mathrm{km}$)

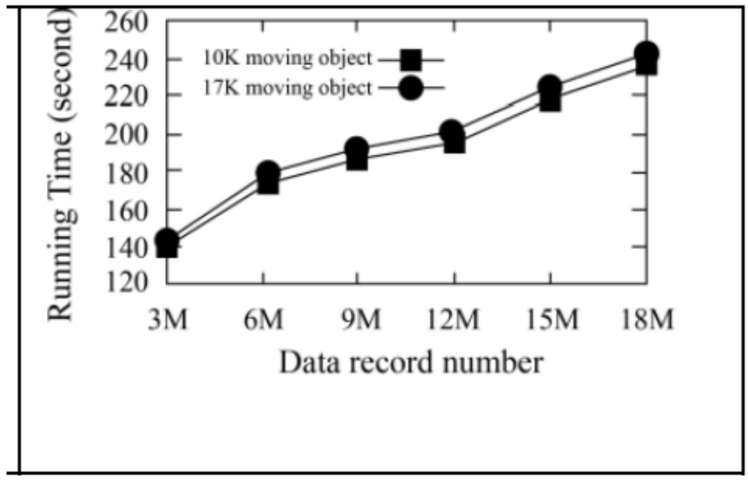


Fig. 5. Performance of RA algorithm on GeoSocial data set ($k = 3$)

## References

[1] A. Pentland, D. Lazer, D. Brewer: *Using reality mining to improve public health and medicine.* Studies in Health Technology & Informatics *149* (2009), 93–102.

[2] A. L. Menach, A. J. Tatem, J. M. Cohen, S. I. Hay, H. Randell, A. P. Patil, D. L. Smith: *Travel risk, malaria importation and malaria transmission in Zanzibar.* Scientific Reports *1* (2011), Srep. No. 93, 1–7.

[3] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, C. O. Buckee: *Quantifying the impact of human mobility on malaria.* Science *338* (2012), No. 6104, 267–270.

[4] J. D. Lilly, D. A. Gray, M. Virick: *Outsourcing the human resource function: Environmental and organizational characteristics that affect HR performance.* Journal of Business Strategies *22* (2005), No. 1, 55–72.

[5] B. L. Tatibekov, J. S. Adams, N. A. Prochaska: *Characteristics of the labor mar-*

*ket and human resources management in the Republic of Kazakhstan.* Advances in Competitiveness Research *12* (2004), No. 1, 44–56.

[6] M. Warner: *Reassessing human resource management 'with Chinese characteristics': An overview.* International Journal of Human Resource Management *19* (2008), No. 5, 771–801.

[7] N. J. Yuan, Y. Zheng, L. Zhang, X. Xie: *T-Finder, a recommender system for finding passengers and vacant taxis.* IEEE Transactions on Knowledge and Data Engineering *25* (2013), No. 10, 2390–2403.

[8] S. B. Wicker: *Cellular telephony and the question of privacy.* Communications of the ACM *54* (2011), No. 7, 88–98.

[9] B. W. Xu, L. Xu, X. F. Meng, G. Yu, Z. D. Lu, Y. X. He, J. Y. Shen: *Web-based information systems and applications, A survey.* Proc. Conference on Web Information System and Applications (WISA), 29–31 October 2004, Wuhan, China.

[10] Y. A. de Montjoye, C. A. Hidalgo, M. Verleysen, V. D. Blondel: *Unique in the crowd, the privacy bounds of human mobility.* Scientific Reports *3*, (2013), Article No. 1376.

[11] I. Ozalp, M. E. Nergiz, M. E. Gursoy, Y. Saygin: *Privacy-Preserving Publishing of Hierarchical Data.* ACM Transactions on Privacy and Security *19* (2016), No. 3, paper 7.

[12] A. E. Cicek, M. E. Nergiz, Y. Saygin: *Ensuring location diversity in privacy-preserving spatio-temporal data publishing.* The VLDB Journal *23* (2014), No. 4, 609 to 625.

Received July 12, 2017