An improved privacy Apply algorithm in PINQ

Qianying Li

State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, 100876, China

liqian_ying@sina.com

Keywords: Differential privacy, PINQ, Vehicle information.

Abstract. It can save a lot of manpower and resources through the statistical traffic flow to manage the high speed road. But the weak protection of the vehicle information may make trouble. In this paper, an improved privacy algorithm named Apply based on the common implementation framework of differential privacy PINQ[1] is proposed to provide a better security protection. We compare the new Apply with the old Apply and another popular privacy preserving method K-Anonymization[2], and the experiments results demonstrate that the new Apply works better in security protection and data availability.

Introduction

Bad protection of the vehicle information may bring unnecessary trouble to the owners. For example, when counting the total number of the vehicles in an exact high speed road, the time and the port will be published. If the information is released nakedly, the routes of the cars may be revealed.

Privacy preserving model based on group[3] such as K-Anonymization and differential privacy model are two popular privacy preserving methods. PINQ[1]is a differentially-private analysis platform. Differential privacy has become tremendously popular and been widely used in various environments such as network trace analysis[4]and recommendation systems[5]. But functional and meaningful systems using PINQ are less[6]. The Apply algorithm is the key function in PINQ[1]. It is used to check whether the left budget is enough for the next query. But the existing Apply does not limit the times for query, that is to say it is possible to get the right information after analyzing a large number of results.

The rest of this paper is organized as follows: In section II, we give a brief background for differential privacy and K-Anonymization. And we describe PINQ, the common implementation framework of differential privacy. In section III, we describe the new Apply in PINQ; In section IV, we give the proper value for budget by contrasting some typical values, and we analyze the experiments results to illustrate the better method; Finally, we summarize this paper in section V.

Background

Differential privacy. A randomized computation M provides ε-differential privacy if for all D and D ' and any subset SM of the outputs of the computation:

$\Pr[\mathrm{M(D)} \in S_M] \le \exp(\varepsilon) \times \Pr[M(D') \in S_M]$ (1)

ε represents budget[6], which reflects the privacy preserving level. When ε equals 0, the privacy preserving will be the best.

K-Anonymization. Two popular techniques in K-Anonymization are generalization and hiding[2]. Generalization is a simple thought, which means the original data can be replaced by an unclear and more general value. After that, if there still are unqualified tuples, hiding will be used to reach the standard.

PINQ. PINQ is a common implementation framework of differential privacy. It provides an opaque PINQueryable object supporting various SQL-like operations. Different operations supply different functions. The Apply algorithm which is in charge of the budget is one of the key functions in PINQ.

The improvement for Apply in PINQ

After every query, the budget should be added. If the updated budget exceeds the threshold, the next query should be stopped to prevent the privacy leakage. But the existing Apply does not check whether the refreshed budget could satisfy the following queries, so the new Apply is improved as follow.

The whole budget is b; every query will cost ε.

Input: ε.

Output: 1 represents the database could be queried and 0 represents the database could not be queried.

1. get query conditions;
2. Contrast to the query conditions in the query log file dig.txt, if the conditions are the same, do3, or4;
3. contrast the two query times and check whether there is overlap, if the answer is yes, then update the budget ε*2=>ε and4, otherwise the budget is still ε and 4;
4. compare ε with b, if ε<b then return 0, else do this query and b-ε=>b;
5. update the query log file dig.txt.

Data analysis

Values of budgets for improved Apply. It is known that the smaller the privacy budget ε is, the bigger the noise is, and the better encryption effect is, the lower data availability is[6].

Different budgets have different effects on data, and the comparison is as follow.

Table 1. Comparison of different groups of budget

b	ε_1	ε_2	ε_3	Q_1	Q_2	Q_3	CQ_1	CQ_2	CQ_3
1.0	0.01	0.05	0.1	100	20	10	6	4	3
1.0	0.04	0.1	0.2	25	10	5	4	3	2
2.0	0.01	0.05	0.1	200	40	20	7	5	4
2.0	0.04	0.1	0.2	50	20	10	5	4	3

In Table 1, b is the total budget, ε_1 is the budget which is required to query the current road information, ε_2 is the budget which is required to query all roads information of the one specific freeway, ε_3 is the budget which is required to query all roads information of the whole freeway, Q_1 is the time of the query for the current road information, Q_2 is the time of the query for all roads of one specific freeway information, Q_3 is the time of the query for all roads of the whole freeway information,CQ_1 is the time of the query for the current road information with time cover, CQ_2 is the time of the query for all roads of one specific freeway information with time cover, and CQ_3 is the time of the query for all roads of the whole freeway information with time cover.

Analysis of the four groups typical data is as follow: The budget combinations of the first group and the third group may cause a greater probability of privacy leakage because of the high Q1, but the Q1 in second group is too low to influence the practicability of the system, considering the system practicability and the degree of privacy protection, the fourth group data wins. And also we choose the fourth group data for the system.

Privacy protection analysis of the original and the improved. We operate the data of the fifth grade cars which enter at Beijing and exit at Bazhou in freeway named G45 Daguang from 12:00 pm to 12:01 pm on 6-2-2015.

Unimproved algorithm can query infinitely, so we might as well choose 300 results to analyze. Using the results, we can get a diagram Fig.1. According to Fig.1, we can forecast the true value may be 44. Compared with the exact true value 44, we can know that this situation might cause privacy disclosure.

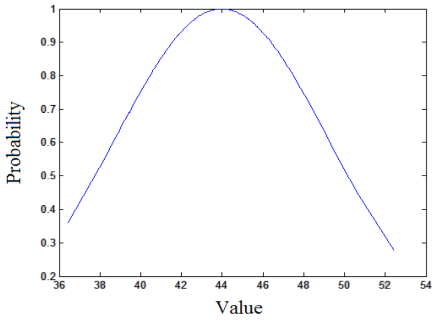

Fig. 1 Probability of the original values

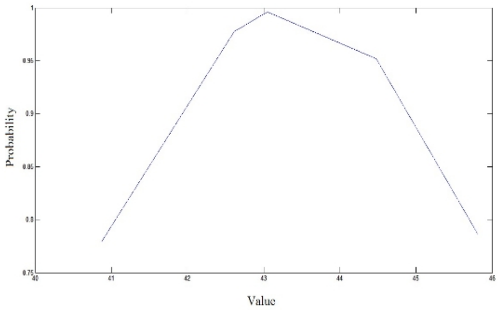

Fig. 2Probability of the improved values

While improved algorithm can query only five times, we draw a diagram Fig.2. According to the Fig.2, we can forecast the true value may be 43. Compared with the exact true value 44, we can know that in this situation privacy is better protected.

Comparison analysis between differential privacy and K-Anonymization. K-Anonymization is another popular privacy protection method, and we will compare these two methods in the performance in system.

Operate Data with K-Anonymization. We choose the algorithm named entropy which is based on entropy classification to operate the data[7].

We select 10 records in database arbitrarily into Table 2.

Table 2. Ten Arbitrary Records

	ID	type	time stamp for Entering high speed	time stamp for leaving high speed	Port
1	JingG88282	3	1433270362	1433274975	Chifeng-Tongliao
2	HeiP89249	3	1433278286	1433282142	Chengde-Beijing
3	YunY69217	2	1433278802	1433283615	Chifeng-Tongliao
4	XinQ79246	1	1433263917	1433268368	Daqing-Songyuan
5	XiangN94494	3	1433229400	1433235908	Chengde-Beijing
6	ShanW20858	5	1433282127	1433285977	Daqing-Songyuan
7	EJ96079	4	1433270375	1433276002	Chifeng-Tongliao
8	JingG88282	3	1433280352	1433285252	Chengde-Beijing
9	JiH20384	4	1433247664	1433252066	Daqing-Songyuan
10	ES86154	5	1433225653	1433232435	Daqing-Songyuan

Private information is freeway port.

We can get the entropy, the expectations and the gains.

$$\mathrm{I}(3,3,4) = -\frac{3}{10}\log_2\frac{3}{10} - \frac{3}{10}\log_2\frac{3}{10} - \frac{4}{10}\log_2\frac{4}{10} = 1.571 \tag{2}$$

$$\mathrm{E(type)} = \frac{1}{10}\mathrm{I}(0,0,1) + \frac{1}{10}\mathrm{I}(1,0,0) + \frac{4}{10}\mathrm{I}(1,3,0) + \frac{2}{10}\mathrm{I}(1,0,1) + \frac{2}{10}\mathrm{I}(0,0,2) = 1.47939 \tag{3}$$

$$\mathrm{E(ID)} = \frac{2}{10}\mathrm{I}(1,1,0) + \frac{1}{10}\mathrm{I}(0,1,0) * 2 + \frac{1}{10}\mathrm{I}(1,3,0) + \frac{2}{10}\mathrm{I}(1,0,1) + \frac{2}{10}\mathrm{I}(0,0,2) = 0.398631 \tag{4}$$

$$\mathrm{E(time\ stamp)} = \frac{1}{10}\mathrm{I}(1,0,0) * 3 + \frac{1}{10}\mathrm{I}(0,1,0) * 3 + \frac{1}{10}\mathrm{I}(0,0,1) * 4 = 0.23254 \tag{5}$$

$$\mathrm{Gain(type)} = \mathrm{I}(3,3,4) - \mathrm{E(type)} = 0.09161 \tag{6}$$

$$\mathrm{Gain(ID)} = \mathrm{I}(3,3,4) - \mathrm{E(ID)} = 1.172369 \tag{7}$$

$$\mathrm{Gain(time\ stamp)} = \mathrm{I}(3,3,4) - \mathrm{E(time\ stamp)} = 1.33846 \quad (8)$$

According to the values above, we can get the Fig.3.

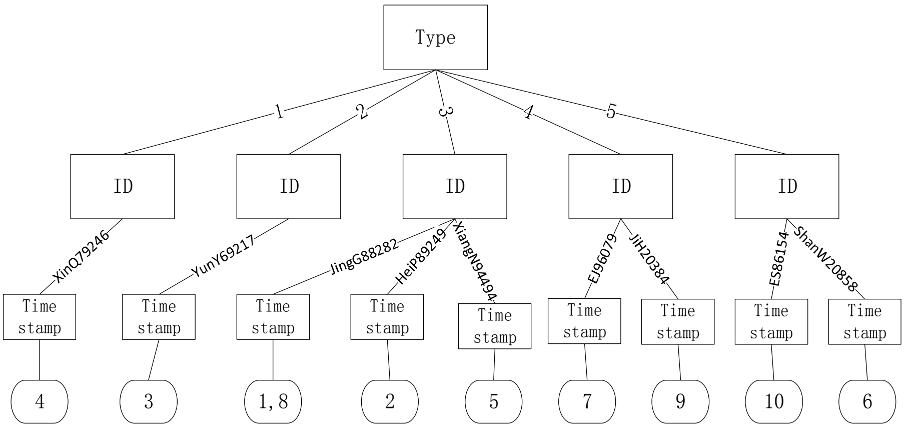

Fig.3Multiway Tree

According to Fig.3, when we set K=2, we can get Table 3.

Table 3. Anonymous Table

	ID	type	time stamp for Entering high speed	time stamp for leaving high speed	port
3	YunY69217	***	1433278802	1433283615	Chifeng-Tongliao
4	XinQ79246	***	1433263917	1433268368	Daqing-Songyuan
1	JingG88282	3	**********	**********	Chifeng-Tongliao
8	JingG88282	3	**********	**********	Chengde-Beijing
2	**********	3	**********	**********	Chengde-Beijing
5	**********	3	**********	**********	Chengde-Beijing
7	**********	4	1433270375	1433276002	Chifeng-Tongliao
9	**********	4	1433247664	1433252066	Daqing-Songyuan
10	**********	5	1433225653	1433232435	Daqing-Songyuan
6	**********	5	1433282127	1433285977	Daqing-Songyuan

Operate Data with Differential Privacy. We operate the data of the fifth grade cars which enter at Beijing and exit at Bazhou in freeway named G45 Daguang from 12:00 pm to 12:01 pm on 6-2-2015. Query the current road information and we can get 5 groups of results: 45.8141、43.0504、40.8722、44.4754、42.6132 .

Comparison of the K-Anonymization and differential privacy results.

a. Error analysis.

As for K-Anonymization, the total number is added one by one, so there is no error.

As for differential privacy, we can get the error rate from the results in 4.3.2. Five groups of results are 45.8141、43.0504、40.8722、44.4754 and 42.6132, we can forecast the true value may be 43. Compared with the exact true value 44, we can get the error rate is 2.27%.

b. Privacy leak analysis

K-Anonymization cannot resist background knowledge attack[2], that is to say if excepting the published data attacker knows other data tables and connects them together, it is possible to infer the concrete information of a specific car.

The budget in differential privacy is used to restrict the times of query, so that it could reduce the rate of leakage. Differential privacy has a higher degree of privacy protection with an independent background knowledge[6]. The budgets in system may not be the best, but they comfort to the requirements of privacy protection and good usability.

c. Data usability analysis

k is the key parameter in K-Anonymization. The bigger K is, the higher degree of anonymity is, the better data privacy protection is, the lower data availability is[2]. In this comparison, we let the value of K equal 2, and we can see that the table has been hiding a lot of information.

As for differential privacy, the budget is the key parameter. The smaller the budget is, the greater the noise is, the better data privacy protection is, the lower data availability is. In the system, different budgets of different queries are changing dynamically. And we can know that when the whole budget equals 2.0, the system could have a good practicability.

Conclusions

We applied the differential privacy protection method to the transportation information management system, the existing Apply algorithm in PINQ could not satisfy the requirements of the system, so we improved the algorithm in this paper. The improved Apply could adjust the budgets dynamically according to different queries. And we got the budgets which comfort to the requirements of privacy protection and good usability after a large number of experiments results. Besides, we compared two algorithms, and experiments results showed that the improved Apply could provide a better privacy protection.

Acknowledgements

This work is supported by NSFC (Grant Nos. 61300181, 61502044), the Fundamental Research Funds for the Central Universities (Grant No. 2015RC23).

References

[1]. Frank McSherry. Privacy integrated queries[C]// In Proc. ACM SIGMOD International Conference on Management of Data2009.

[2]. Cen T T, Han J M, Wang J Y. Survey of K-anonymity research on privacy preservation[J]. Computer Engineering & Applications, 2008.

[3]. Ninghui Li Tiancheng Li Suresh Venkatasubramanian. (2007). T-closeness: privacy beyond k-anonymity and ℓ-diversity. In Icde, 106 - 115..

[4]. Mcsherry F, Mahajan R. Differentially-private network trace analysis.[J]. AcmSigcomm Computer Communication Review, 2010, 40(4):123-134.

[5]. Machanavajjhala A, Korolova A, Sarma A D. Personalized social recommendations: accurate or private[J]. Proceedings of the Vldb Endowment, 2011, 4(7):440-450.

[6]. Xiong P, Zhu T Q, Wang X F. A Survey on Differential Privacy and Applications[J]. Chinese Journal of Computers, 2014.

[7]. Jian Liu. Research on K-Anonymity for Privacy Preserving, 2010.