Front.Comput.Sci.
DOI

SUPPLEMENTARY

A privacy – enhancing scheme against contextual knowledge – based attacks in location – based services

Online Resource 2 (Performance Evaluation)

1 Simulation Setup

In our simulation, about 20,000 LBS users are deployed in a $10km \times 8km$ area in the downtown of Shanghai. The area is divided into 8,000 cells, with the size of each cell being $100m \times 100m$. Fig.1 depicts the local map combined with the density of AP deployment, and we will also follow that deployment in our simulation. Query probabilities are computed as the users' density in each cell, and the query preferences of users are randomly assigned under normal distribution. Two taxi trajectory datasets from ShanghaiOpen Data Apps (http://soda.datashanghai.gov.cn) in Oct. 2013 (hereafter termed *Taxi-2013*) and Apr. 2015 (hereafter termed *Taxi-2015*), which involves more than 20,000 trajectories, are used to describe the mobility patterns of users.

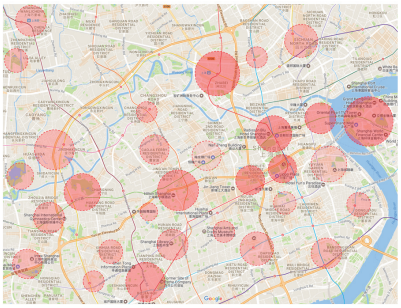


Fig. 1 Density of AP Distribution in Shanghai

There are some parameters used in our simulation. Privacy profile k is set from 2 to 15. # of query types is $m = 5$ and # of sets is $ns = 100$. The various query probability threshold $\beta = 0.0015$, and the query preference threshold $\theta = 0.2$. In addition, the distance preference μ (in the index of distance) is set randomly from 1 to 4.

We compared our proposals with *Random* [1] as the baseline scheme, which randomly chooses dummy locations to protect privacy. *DLS* (*enhanced-DLS*) [2], which is one of state-of-the-art methods, is also selected as a comparison.

2 Evaluation Results

All the following results (except for cache hit ratio and system utility) are achieved from several snapshots after the simulation time $t = 200\ min$.

2.1 k vs. privacy metrics

Fig.2(a)-2(b) and Fig.3(a)-3(b) show the relationship between k and the entropy of a k-anonymity set under different trajectory datasets.

Gross query probability (*i.e.* statistics of query probabilities don't consider query types) is used in Fig.2(a) and Fig.3(a). We can see that all schemes except for *Random* perform well because most schemes take some side information (query probabilities) into account.

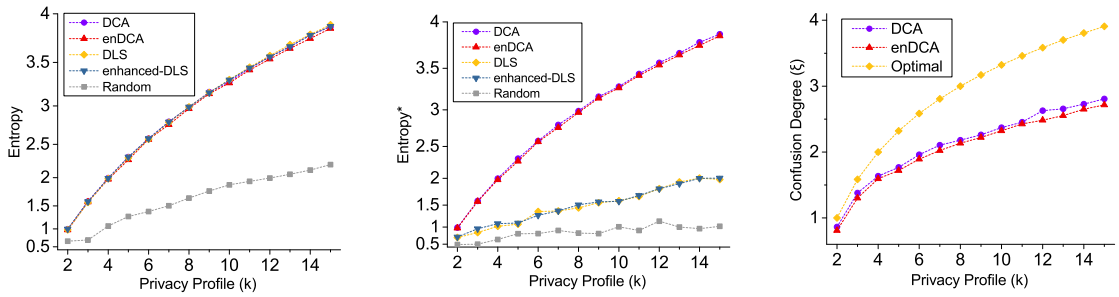
On the contrary, *various query probability* is used in Fig.2(b) and Fig.3(b), which highlights the advantages of our schemes, where richer side information is involved.

As to our proposed privacy metric (confusion degree ξ), we can see in Fig.2(c) and Fig.3(c) that *DCA* edges out *enDCA*, as *enDCA* sacrifices a little confusion degree to decrease the exposure of query preferences and bandwidth overhead (see Fig.6(a)-6(b)). Our schemes have high but not theoretically optimal results because finding $k-1$ nearby users who have approximately the same query preferences as the target user's (*i.e.* preference of the user who actually issues the query) is literally tough when k is large. However, our schemes lead others here (since they don't consider query preferences at all).

2.2 k vs. Pearson correlation coefficient

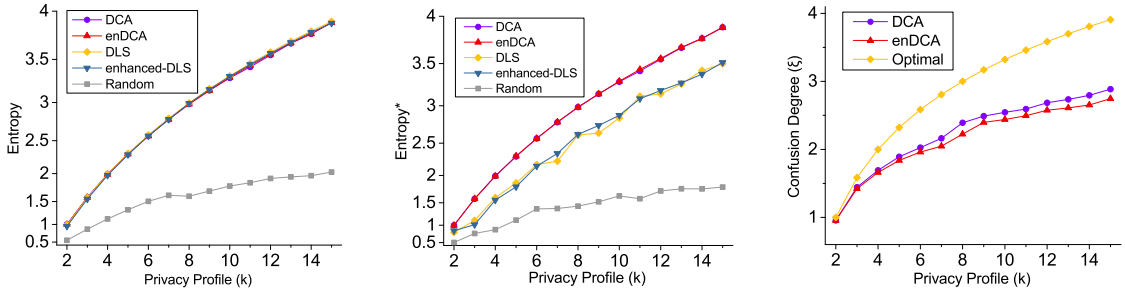
Fig.4(a)-4(b) depict the average correlation coefficient in k-anonymity sets. Our schemes guarantee that $k-1$ other users

E-mail: *cqjin@dase.ecnu.edu.cn*



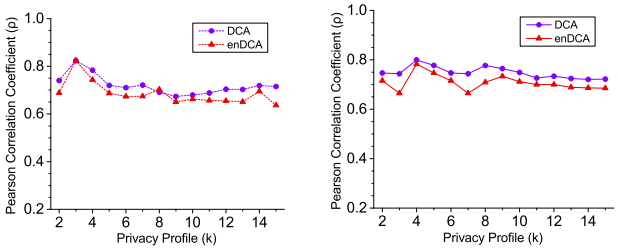
(a) k vs. Entropy (b) k vs. Entropy* (c) k vs. ξ

Fig. 2 Effect of privacy profile k on several privacy metrics (on Taxi-2013 dataset): (a)Entropy (with *gross query probability*); (b)Entropy (with *various query probability*); (c)Confusion degree (with *various query probability*)



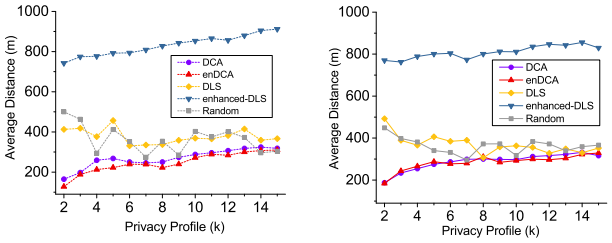
(a) k vs. Entropy (b) k vs. Entropy* (c) k vs. ξ

Fig. 3 Effect of privacy profile k on several privacy metrics (on Taxi-2015 dataset): (a)Entropy (with *gross query probability*); (b)Entropy (with *various query probability*); (c)Confusion degree (with *various query probability*)



(a) on Taxi-2013 dataset (b) on Taxi-2015 dataset

Fig. 4 Effect of privacy profile k on Pearson correlation coefficient



(a) on Taxi-2013 dataset (b) on Taxi-2015 dataset

Fig. 5 Effect of privacy profile k on average distance between the target user and other users in anonymity sets

in the set have relatively similar query preferences with the target user.

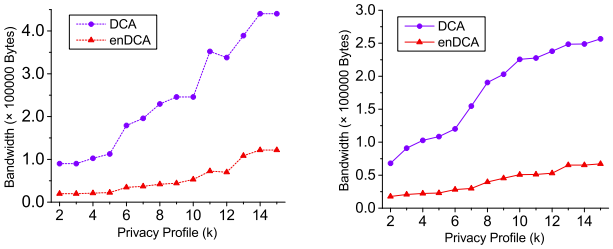
2.3 k vs. average distance

We evaluate the average distance between the target user and other users in Fig.5(a)-5(b). *Enhanced-DLS* has the maximum average distance because it prefers to spread k users as far as possible. In our schemes, average distances increase gradually with k, and the values are around 160-360m, as we set the distance preference μ (in the index of distance) randomly from 1 to 4.

2.4 k vs. bandwidth overhead

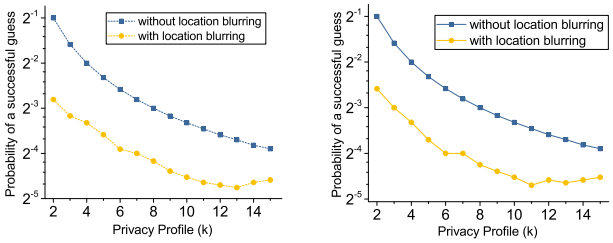
Data such as preference vectors and query probabilities will be transmitted during construction of anonymity sets. In this part, we evaluate the bandwidth overhead of our proposals, with the assumption that the size of a preference vector equals 2048 Bytes, and the size of a cached anonymity set equals 10240 Bytes.

Fig.6(a)-6(b) illustrate that *enDCA* outperforms *DCA*, since it employs *caching* to serve users' requests for anonymity sets.



(a) on Taxi-2013 dataset (b) on Taxi-2015 dataset

Fig. 6 Effect of privacy profile k on bandwidth overhead

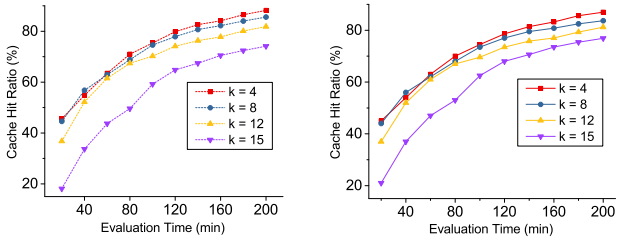


(a) on Taxi-2013 dataset (b) on Taxi-2015 dataset

Fig. 8 Effect of *location blurring* on successful guessing probability

2.5 t vs. cache hit ratio (k = 4, 8, 12, 15)

Fig.7(a)-7(b) depict the relationship among k, cache hit ratio and simulation time. The hit ratio increases gradually with the simulation time t, and smaller k usually results in higher ratio. After a long period of simulation, *enDCA* achieves a satisfying cache hit ratio (over 70%).



(a) on Taxi-2013 dataset (b) on Taxi-2015 dataset

Fig. 7 Effect of simulation time t on cache hit ratio (under different privacy profile values $k = 4, 8, 12, 15$)

2.6 Effect of *location blurring* on the probability of a successful guess

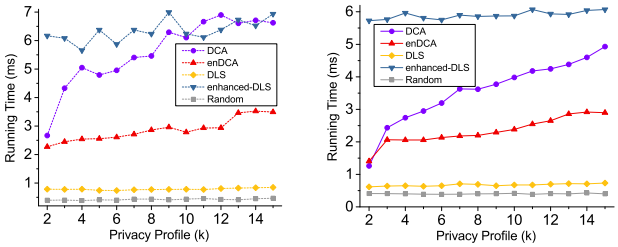
We evaluate the effect of *location blurring* on the probability of successfully guessing the real location in Fig.8(a)-8(b).

Intuitively, the possible real location is in one of the 1-hop areas around k users in the anonymity set. Schemes without *location blurring* (*e.g. DCA, DLS, enhanced-DLS*) have the theoretical k-anonymity (*i.e.* successful guessing probability is $\frac{1}{k}$). The *enDCA*, which is equipped with *location blurring*, owns significantly lower probabilities of successful guesses.

2.7 k vs. running time

Fig.9(a)-9(b) demonstrate the running time of all schemes.

Random runs the fastest because it doesn't take any side information into consideration, while *enhanced-DLS* costs the longest time in most cases because it prefers to find a set with both the maximum privacy level and a bigger cloaking region size. As to efficiency of our schemes when constructing k-anonymity sets, *enDCA* shows better time consumption than *DCA* with the help of *caching*.



(a) on Taxi-2013 dataset (b) on Taxi-2015 dataset

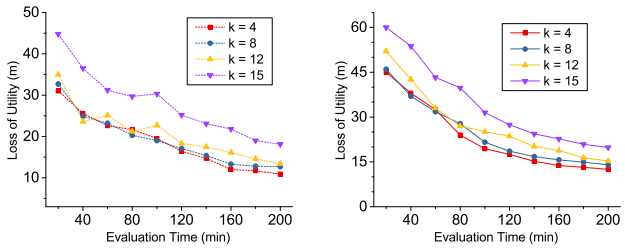
Fig. 9 Effect of privacy profile k on runtime

2.8 Effect of *location blurring* on system utility

As we all know, one of the most frequently used applications in Location-based Services is NN-Query (*e.g.* Finding the nearest gas station). So how accurate the resulting NN answers are affects user experience or system utility very much. We define the *System Utility* as distance difference (accuracy loss): $\left|d_{realNN} - d_{shiftedNN}\right|$, where d_{realNN} denotes the distance between target user's real location and nearest POI based on that location, and $d_{shiftedNN}$ denotes the distance between target user's real location and nearest POI based on the *shifted* location.

When *DCA* Algorithm is run by users, since the k-anonymity sets received by LBS servers includes target user's real location, candidate NN answers must have contained the exact nearest POI. So target user can filter out that POI locally. However, after *Location blurring* is used in *enDCA* Algorithm, there may exist some accuracy loss because target user's shifted location (rather than the real one) is involved in the anonymity set. To survey the influence of *location blurring* over system utility, we conduct some experiment with additional POI data from `dianping.com`.

Fig.10(a)-10(b) illustrate the loss of system utility under different trajectory datasets. With simulation time extended, more anonymity sets will be cached. Consequently, target users will shift their real locations less frequently, and the candidate answers becomes more accurate. Anyway, even in the worst case (few cached sets in the system), accuracy loss is less than 100m (*i.e.* target users need at most an extra distance of 100m to get to the POI which is nearest to the shifted location rather than the real one).



(a) on Taxi-2013 dataset (b) on Taxi-2015 dataset

Fig. 10 Effect of *location blurring* on Loss of Utility (under different privacy profile values $k = 4, 8, 12, 15$)

References

1. H. Kido, Y. Yanagisawa, and T. Satoh. An anonymous communication technique using dummies for location-based services. In *ICPS*, pages 88–97, 2005.
2. B. Niu, Q. Li, X. Zhu, and G. Cao. Achieving k-anonymity in privacy-aware location-based services. In *IEEE INFOCOM*, pages 754–762, 2014.