Addressing Privacy Concerns of Telecommunication Data with Differential Privacy

Aitor Navarro
Bilkent University
06800 Ankara, Turkey
alvar.navarro@ug.bilkent.edu.tr

Selim Eren Bekçe
Bilkent University
06800 Ankara, Turkey
eren.bekce@bilkent.edu.tr

Yunus Emre Tortamış
Bilkent University
06800 Ankara, Turkey
emre.tortamis@bilkent.edu.tr

ABSTRACT
Big data initiatives are being carried out by providers and public institutions with an increasing frequency. Hence there is a growing need of comprehensive study and analysis on that big data. In this project, first we evaluate the areas that are most susceptible to benefiting from a big data analysis and the reasons behind it. We then go through a case study by Korean Government and assess the risks and the conflict arising from implementing such solutions via the subjects involved. This assessment is evaluated with the results of a small poll between diverse groups of friends. At the end we implemented differential privacy with interactive database technique in order to solve corresponding privacy concerns in this theme. As a result, Laplacian noise is added to raw data in order to protect the individuals' data while not losing much in utility.

Keywords
Data privacy, big data, public-private partnership, mobile data, transportation, data anonymization, differential privacy.

1. INTRODUCTION
In recent days we find many news about nations going after the benefits of gathering large amounts of data for accomplishing traditional tasks. Few days ago we could read in the newspapers that the Spanish Institute of Statistics (INE) plans to use the data from the citizens' mobile phones for their next census in 2021. That same article included a statement (without further explanation) claiming they were only interested in aggregated data and therefore your privacy was not endangered. Many are the areas that benefited from Big Data, the cases we have observed mainly revolve around better transportation, better dimensioning and response to population movements and more efficient placement of facilities or PR materials.

Some initiatives are more controversial than others. Despite what we could think in first place, PR material location (that include data items as the age) is less intrusive in terms on privacy than others like the creation of a night bus network. The reason behind it was the usage of sensitive information as the billing address. Our main motivation is to discover this possible threats to privacy in apparently harmless projects, and for it we will have a close look at the Seoul night bus network development [1].

2. SEOUL NIGHT BUS PROJECT
Albeit the city of Seoul counted already with affordable public transportation, it did not run after midnight due to financial issues and the desire of not harming the taxi business.

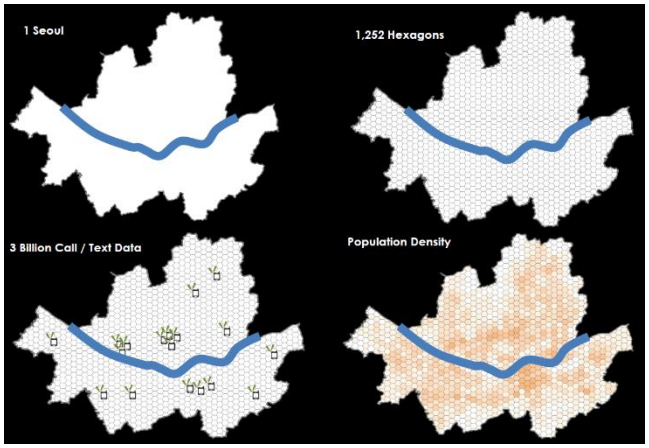

Figure 1: Hexagonal division of Seoul and population density

As the population of the city grew more and more, late night public transportation became more vital for citizens. Although the absence of it was a commonly commented problem among citizens, official discussion on the late night bus system was started by a college student who tweeted to the Mayor. The metropolitan government realized there could be few bus lines after midnight; in fact, initially as little as 10 lines were considered. The question was obvious: how to set the optimal bus route that will satisfy a majority of citizens the answer resulted to lie on the Big Data. As could be seen in Fig.1, first the city was divided into 1,252 hexagonal areas with a diameter of 1 kilometer, which was chosen heuristically and was thought as an acceptable distance that can be walked even at night. Then several sources allowed the gathering of a large amount of data. One of this sources was call data, thanks to the signature of a MOU[1] with the service provider Korea Telecom, 3 billion calls and texts at total were analyzed to map the hexagonal division of the city.

[1] Memorandum of understanding (MOU) is a formal agreement between two or more parties. Companies and organizations can use MOUs to establish official partnerships.

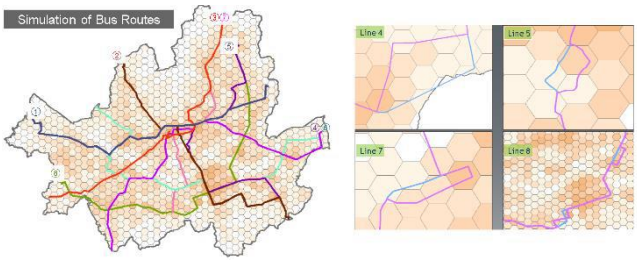

Figure 2: Simulation of the bus routes and local adjustments

Drawing optimal bus lines is more or less obvious by just inspecting the population density map, but the direction of the bus is another problem. That problem was resolved by analysis call location of individuals after midnight (departure point; nightlife area) and their billing address (arrival point; residence area). Naturally, detecting optimal departure and arrival points for every possible user required exhaustive simulations on big data. Further, the results were refined and crosschecked with other data sources like the Taxi DTG[2] and the Seoul transportation card. The data of the former was especially useful since individuals who live very far from midnight destinations tend to take cab. At the end, these bus routes are validated through the analysis of the accumulated Big Data and appropriate adjustments were made as in Figure 2. The result of the after-implementation analysis was promising. Table 1 shows 42% of Seoul citizens live near bus stops of post-midnight lines.

Table 1: Population coverage with post-midnight lines.

	Number of Census Output Area	Population
Seoul	16,471	9,512,346
Within 500m radius	6,901	3,980,391
Ratio	41.9%	41.8%

The problem is that this big data comes out of the collaboration of different parties. Quantitative analyses were conducted by metropolitan government of Seoul on Transportation Card System, the nighttime taxi database, and the Big Data from Korea Telecom. In summary in order to get a late-night transportation service, Seoul citizens had to give up some of their personal data such as the location of their personal calls, the time they took the bus, where they called taxi, their billing address, effectively the home address. Hence there is a privacy conflict that needs to be addressed.

3. CONFLICT

In this particular case study, the usage of the data from Korea Telecom is of vital importance. Despite the usage of other sources of data like the transportation card, call data is the enabler of this project. And the Seoul Owl Bus is not one-of-a-kind but rather a commonplace phenomenon these days, at the moment the city of Seoul has numerous ongoing initiatives that directly rely on the usage of 3rd party's data. At first sight, it might seem that this approach has only benefits, but it actually raises some concerns as well.

Despite the novelty of the problem, some institutions have already put the spotlight on this issue. It is the case of ENISA, which stands for the European Union Agency for Network and Information Security with this 80 pages-long document published just few months ago.

Unfortunately, its role is mainly limited to recommend the implementation of some privacy enhancing tools. This is due to the fact that the current applicable law falls under the directive 95/46/EC which dating from two decades ago fails to provide adequate solutions to such a recent concern.

Talking now about Korea itself, it is worth to mention that its Data Privacy Law traditionally was limited to the fulfilment of the OECD's privacy Guidelines and its derivative, the APEC Privacy Framework. This changed not so long ago, in 2011, when the Personal Information Protection Act was promulgated. The Act was set to replace the existing Public Agency Data Protection Act in whole and in relation to the private sector it replaces in part the Act on Promotion of Information and Communications Network Utilization and Information Protection (ICN Act). Aiming to agglutinate and formalize Privacy law, and set a high standard of privacy, which provoked privacy experts around the globe to name it "The strongest law in Asia".

It is rather surprising however, that such a recent law does not address Big Data specifically (not even in its amendments passed on April 26th 2016).

What it does define, is a strict notification requirement when sharing data with third parties:

> *"Consent for disclosure to third parties is required, and they must be identified (A 17). There are limited exceptions (A 18), but these do not include 'compatible uses' or similar expressions. The consent requirements of the Korean Act are one of its strictest requirements, and an aspect that will be considered onerous by businesses."*

We decided to check whether the public opinion was aligned with the mentioned requirements, and for that purpose we ran a small online poll, whose results has been shown in Figure 3.

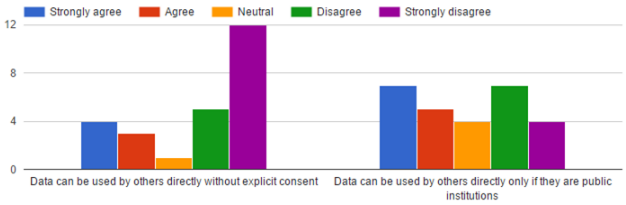

Figure 3: Mini survey results about people's concerns on sharing and disclosure of their data to third parties

It appears evident that people goes a step further, asking not only for notification on the disclosure, but also explicit consent. Unfortunately, this time KT seemed to forgot to do so, no giving notice of it as far as we could find out.

4. SOLUTION

The implications of sharing location data with third parties would be huge, given that this data contains mappings of phone numbers to billing addresses and 3 billion call records with location data. This kind of data is invaluable in the hands of wrong people and

[2] Digital Tacho Gauge keeps speed and distance records

can cause devastating effects on the citizens such as increased thievery and frauds.

The expectations of the solution would need to have following properties:

1. Protecting individual's privacy by default: The privacy of the operator users need to be preserved so that no third party should have raw access to one individual's records. So the raw data must stay local to the operator.
2. The restrictions of the solution should allow Seoul Night Bus Project to be built with only few changes to the methodology.
3. Ideally it should open up an open ecosystem for other new projects to be built, without private contracts.

4.1 Query API

Designing a web based Query API with interactive ε-differential privacy [3] techniques would be the solution for this scheme. It would allow the operator to own the raw data and only share aggregate data with added noise which will increase the privacy.

4.1.1 Ɛ-Differential Privacy

This API would work with interactive ε-differential privacy concept. After the database calculates the query result F(D), it will add some noise to the result with respect to ε-differential privacy then return the query. This way we can guarantee the results would be anonymous or probabilistically indistinguishable from the real case.

4.1.2 Audience

The API would have a wide range of audience like advertisement companies, curious individuals and bus companies as in the real example. The monetary gain of providing such API would also be high, as the service provider can charge by utility and rate. Note that ε value and privacy concept are inversely proportional. In other words, lower ε value means higher privacy (more noise, lower utility) whereas higher ε values mean lower privacy (less noise, higher utility). Therefore, the operator would offer different levels of access to the API with different clearance levels. However, even with the highest utility case, some noise will still be added for people to not get easily de-anonymized.

4.1.3 Query Scheme

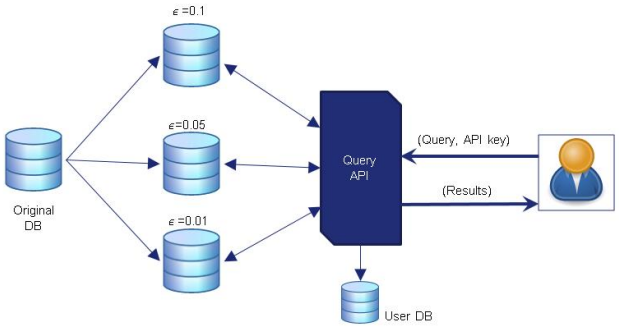

Figure 4: The query scheme

Users of this API will be assigned different levels of clearance respective to their role, which will directly affect ε level applied to the results. In Figure 4, we see multiple virtual databases are created with different ε levels are created and the Query API uses the relevant virtual database to answer the query. Note that a query request consists of the query clauses and the API key, which is used to authenticate the client making the query so that appropriate ε level can be applied by the Query API.

4.1.4 Noise Function

Noise is added to the results from the Laplacian Distribution [4].

$$f(x \mid \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

Noise *X* is generated from below function obeys Laplacian Distribution. *U* is a uniform random variable between *(−1/2, 1/2]*. *μ* is the location parameter and *b* is the scale parameter, which is inversely proportional to ε.

$$X = \mu - b\ \mathrm{sgn}(U)\ \ln(1 - 2|U|)$$

4.2 Data Model

There are three types in our data model

4.2.1 API User

The user of the Query API has an internal id, a name, an assigned clearance level and a generated API Key to use in query operations. The API User will send this API Key each time he or she wants to execute a query.

4.2.2 Client

This is the data owner (the telecom operator client) which has an id and a home (billing address).

4.2.3 Data Tuple

The data tuple contains a client id, a timestamp and the data point location.

4.3 Area Model

We have decided to replicate the topology that the Owl Bus initiative used. They mapped the whole city into 1km square hexagons. We labeled this hexagon using X and Y indexes as shown in Figure 5.

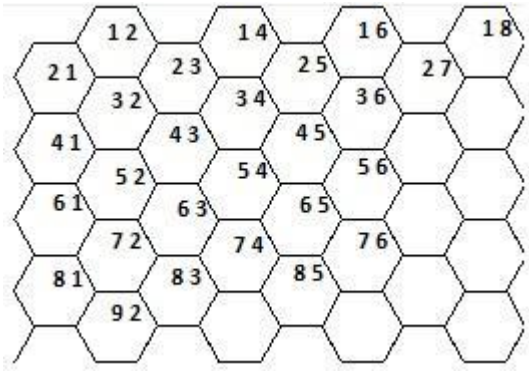

Figure 5: Hexagon Indexes

The reason behind the indexed or "Cell-IDs" is facilitating the application of the Vicinity concept, in order to filter not relevant data for our analysis (i.e.: callers that are already near to their billing address".

Consequently, the X and Y elements of every index, are used for declaring a hexagon as a "neighbor" if it differs as maximum 1 unit to each of the index elements.

4.4 Result Caching

It is important to cache some query results after adding noise, because the adversary may just rerun the query multiple times and

get an average for the whole runs to obtain the real result. Since the API would be public, the results should be cached with respect to the query clauses and clearance level of the client.

4.5 Implementation

The Query API design has been fully implemented as a REST server using Spring Framework and MongoDB as database layer. It serves a query endpoint which accepts a set of query clauses and the API key of the client. The result of query will be aggregated number of unique clients which

4.5.1 Query Clause Types

Each query clause has at least type information and other auxiliary information that is required for respective clause types. This is a conjunctive query design: AND operation will be performed on the results for each clause.

4.5.1.1 Area

As explained, the area model consists of hexagons. The API User can specify multiple hexagons to query on.

```
{"type":"AREA","hexagons":[[2,2],[2,3],[2,4]]}
```

4.5.1.2 Time Clauses

API User can specify the time ranges to perform the query on. Multiple time clauses can be used to further narrow down the results. It supports hour of day, day of week, day of month, month of year and year type time clauses. Some examples include:

```
{"type":"TIME_DAYOFWEEK", "times":[1,7]}
```

Selects Sunday and Saturday. Values from Java Calendar class are used.

```
{"type":"TIME_DAYOFMONTH","times":[18,19,20,21]}
```

Selects data on given days of each month.

```
{"type":"TIME_HOUROFDAY","times":[23,0,1,2,3,4,5]}
```

Selects data on given hours of each day (10 pm to 5 am, inclusive)

```
{"type":"TIME_MONTHOFYEAR","times":[6,7,8]}
```

Selects data on given months of each year (June to August, inclusive)

```
{"type":"TIME_YEAR","times":[2015,2016]}
```

Selects data on given years.

4.5.1.3 'At Home' or 'Not at Home' Clause

It looks at the difference between data location and home location.

```
{"type":"AT_HOME"}
{"type":"NOT_AT_HOME"}
```

If 'At Home' type of query is used, it only selects the data where the hexagon that location and home location are one proximity to each other or are the same. Conversely, if 'Not at Home' type of query is used, it only selects the data where location and home location differs more than one proximity. These two clause types are mutually exclusive so API User cannot use them at the same type.

4.5.2 Test Data

We have generated some random data to conduct experiments. It would have been better if we did

- 3 API Keys with High, Medium and Low clearance levels. This will help us show the difference of utility with different clearance levels for the same query.
- A virtual city with 10 x 10 hexagon grid
- 250 Clients with random home locations on the grid
- 10000 Data tuples in total, each originating from a random Client (created previously), on a random location on the grid. The timestamp is also chosen randomly selected from a predefined month.

4.5.3 Number of Unique Clients Limit

If the query clauses make the number of unique clients too low, then it may be possible to de-anonymize some individuals. Therefore, the API rejects the query if the number of results (after adding appropriate noise) are too low (defined as the 10% of client count).

4.5.4 Experiments

For testing, we have run some sample queries on the generated data and obtained some results. We also measured the utility loss with each clearance level.

```
[
  {"type":"AREA", "hexagons":
[[3,2],[4,3],[2,3],[5,2],[6,3]]},
  {"type":"NOT_AT_HOME"},
  {"type":"TIME_DAYOFWEEK", "times":[2,3,4,5,6]},
  {"type":"TIME_HOUROFDAY",
"times":[3,4,5,6,7,8,9,10,11,12,13]}
]
```

We have run this query multiple times with different clearance levels. Table 2 shows the obtained outputs and also the real result (not given to API User) to compare. It is important to note that the noise scales with the output.

Table 2: Query Results with Multiple Clearance (ε) Levels

Clearance (ε) Level	Output (no. of results)
High	122
Medium	119
Low	130
(Real Result)	123

5. CONCLUSION

In this paper, we explained the privacy concerns emanating from using Big Data through a case study, namely Seoul Night Bus project. We pointed out that the privacy conflict is real, not addressed properly by the agreements or laws and tend to pose serious risk in the near future. Under the scope of this term project, we implemented a Query API with varying utility levels which applies Laplacian noise to provide ε-differential privacy. The results are promising, it has been shown that with this API, it is possible to work on the Night Bus Project and possibly other new projects without risking the privacy of the individuals.

As future work, more efficient implementation with proper geospatial indexes for better scaling could be implemented. Naturally, differential privacy, just like any other privacy protection technique, is prone to malicious attacks; averaging queries, de-anonymization with background information, probabilistic interferences to name a few. Therefore, ideally the API should be self-learning; i.e. it should outsmart the attacker by considering her access patterns and questions. However, this is an open-ended statement and real application of that feature would raise the project to another level.

6. ACKNOWLEDGMENTS

Our deepest gratitude to Ms. Jihyun Kim and the rest of Seoul's Metropolitan and City government. Big thanks to Mr. Earl Burgos for his great predisposition and very special thanks for the NIA[3]'s eGGA[4] team.

7. REFERENCES

[1] Kim G. 2014. What Can We Do with Big Data for Green City? Technical Report. Metropolitan City Government of Seoul.

[2] EU. Privacy by design in big data, an overview of privacy enhancing technologies in the era of big data analytics. Technical report, ENISA (The European Union Agency for Network and Information Security), June 2013.

[3] Dwork, Differential Privacy. ICALP 2006.

[4] Laplace distribution, https://en.wikipedia.org/w/index.php?title=Laplace_distribution&oldid=717924451 (last visited May 8, 2016).

[3] National Information Society Agency

[4] eGoverment Global Academy