Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset

SEPTEMBER 15, 2014 BY ATOCKAR 57 COMMENTS

In my previous post, Differential Privacy: The Basics, I provided an introduction to differential privacy by exploring its definition and discussing its relevance in the broader context of public data release. In this post, I shall demonstrate how easily privacy can be breached and then counter this by showing how differential privacy can protect against this attack. I will also present a few other examples of differentially private queries.

The Data

There has been a lot of online comment recently about a dataset released by the New York City Taxi and Limousine Commission. It contains details about every taxi ride (yellow cabs) in New York in 2013, including the pickup and drop off times, locations, fare and tip amounts, as well as anonymized (hashed) versions of the taxi's license and medallion numbers. It was obtained via a FOIL (Freedom of Information Law) request earlier this year and has been making waves in the hacker community ever since.

The release of this data in this unalloyed format raises several privacy concerns. The most well-documented of these deals with the hash function used to "anonymize" the license and medallion numbers. A bit of lateral thinking from one civic hacker and the data was completely de-anonymized. This data can now be used to calculate, for example, any driver's annual income. More disquieting, though, in my opinion, is the privacy risk to passengers. With only a small amount of auxiliary knowledge, using this dataset an attacker could identify where an individual went, how much they paid, weekly habits, etc. I will demonstrate how easy this is to do in the following section.

Violating Privacy

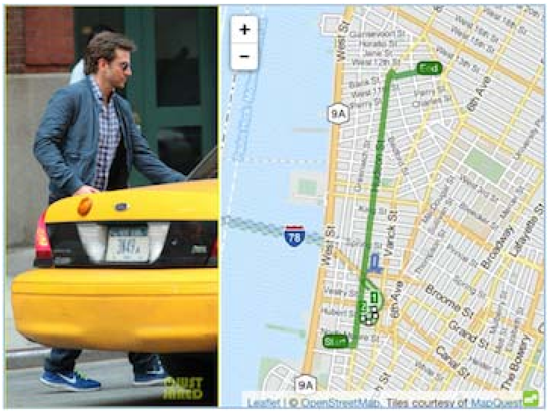
Let's consider some of the different ways in which this dataset can be exploited. If I knew an acquaintance or colleague had been in New York last year, I could combine known information about their whereabouts to try and track their movements for my own personal advantage. Maybe they filed a false expense report? How much did they tip? Did they go somewhere naughty? This can be extended to people I don't know – a savvy paparazzo could track celebrities in this way, for example.

There are other ways to go about this too. Simply focusing the search on an embarrassing night spot, for example, opens the door to all kinds of information about its customers, such as name, address, marital status, etc. Don't believe me? Keep reading…

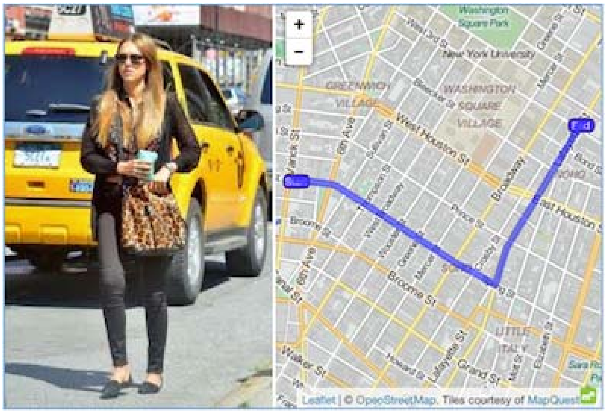
Stalking celebrities

First things first. How might I track a person? Well, to zone in on a particular trip, I can use any combination of known characteristics that appear in the dataset, such as the pickup or drop-off coordinates or datetime, the medallion or license number, or even the fare amount from a receipt. Being the avid fanboy that I am (*note: sarcasm*), I thought it might be interesting to find out something new about some of the celebrities who had been seen in New York in 2013. In particular, where did they go to / come from, and how much did they tip?

In order to do this, I spent some of the most riveting hours of my professional career searching through images of "celebrities in taxis in Manhattan in 2013" to find enough information to identify the correct record in the database. I had some success – combining the below photos of Bradley Cooper and Jessica Alba with some information from celebrity gossip blogs allowed me to find their trips, which are shown in the accompanying maps.



Bradley Cooper (Click to Explore)

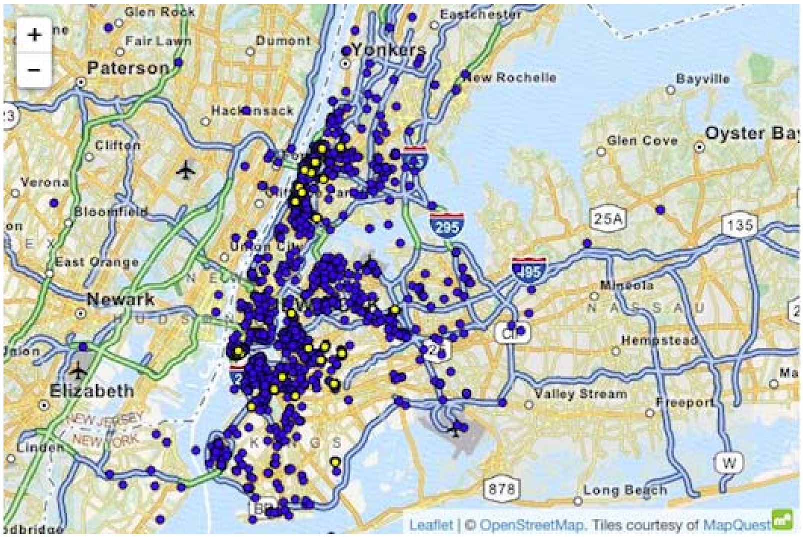


Jessica Alba (Click to Explore)

In Brad Cooper's case, we now know that his cab took him to Greenwich Village, possibly to have dinner at Melibea, and that he paid $10.50, with *no recorded tip*. Ironically, he got in the cab to escape the photographers! We also know that Jessica Alba got into her taxi outside her hotel, the Trump SoHo, and somewhat surprisingly also did not add a tip to her $9 fare. Now while this information is relatively benign, particularly a year down the line, I have revealed information that was *not previously in the public domain*. Considering the speculative drivel that usually accompanies these photos (trust me, I know!), a celebrity journalist would be thrilled to learn this additional information.

A few innocent nights at the gentlemen's club

But OK, perhaps you're not convinced. After all, this dataset is (thankfully) not real-time. How about we leave the poor celebrities alone and consider something a little more provocative. Larry Flynt's Hustler Club is in a fairly isolated location in Hell's Kitchen, and no doubt experiences significant cab traffic in the early hours of the morning. I ran a query to pull out all pickups that occurred outside the club after midnight and before 6am, and mapped the drop-off coordinates to see if I could pinpoint individuals who frequented the establishment. The map below shows my results – the yellow points correspond to drop-offs that are closely clustered, implying a frequent customer.



Click to Explore

The potential consequences of this analysis cannot be overstated. Go ahead, zoom in. You will see that the GPS coordinates are terrifyingly precise. Using this freely-obtainable, easily-created map, one can find out where many of Hustler's customers live, as there are only a handful of locations possible for each point. Add a little local knowledge, and, well, it's not rocket science. *"I was working late at the office"* no longer cuts it: Big Brother is watching.

Even without suspicions or knowledge of the neighborhood, I was able to pinpoint certain individuals with high probability. Somewhat shockingly, just googling an address reveals all kinds of information about its inhabitants. Take the following example:

> *Examining one of the clusters in the map above revealed that only one of the 5 likely drop-off addresses was inhabited; a search for that address revealed its resident's name. In addition, by examining other drop-offs at this address, I found that this gentleman also frequented such establishments as "Rick's Cabaret" and "Flashdancers". Using websites like Spokeo and Facebook, I was also able to find out his property value, ethnicity, relationship status, court records and even a profile picture!*

Of course, I will not publish any of this information here, but this was by no means a unique situation. While the online availability of all this potentially private information is part of a wider discussion, it’s fair to say that this guy has a right to keep his nighttime activities a secret.

To reiterate: the findings in this section were *not hard* to uncover. Equipped with this dataset, and just a little auxiliary information about you, it would be quite trivial for someone to follow your movements, collecting data on your whereabouts and habits, while you remain blissfully unaware. A stalker could find out where you live and work. Your partner may spy on you. A thief could work out when you’re away from home, based on your habits. There are obvious mitigating factors here, such as the population density of Manhattan and the time delay of the data release, but the point still stands.

Applying Differential Privacy

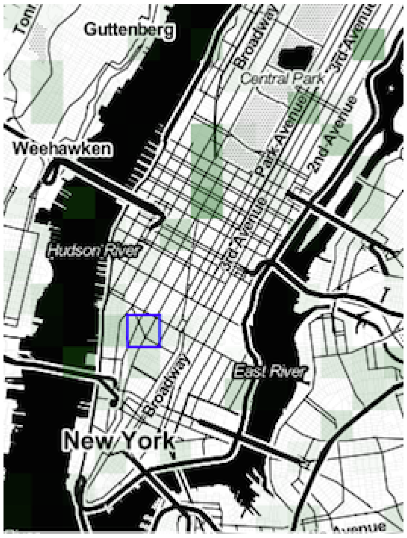
So, we’re at a point now where we can agree this data should not have been released in its current form. But this data has been collected, and there is a lot of value in it – ask any urban planner. It would be a shame if it was withheld entirely.

Enter differential privacy. Sure, the data could be anonymized, but that obviously didn’t work out too well. More care could be taken, for example hashes could be salted as suggested in this forum. But that still doesn’t really protect against either of the attacks above. If you read my first post, you will know that only differential privacy guarantees the protection of all individuals in the data.

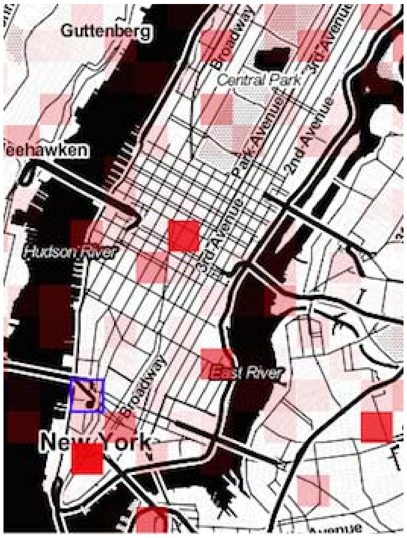
So how should we go about applying differential privacy? Remember that differential privacy works by adding noise to the query. Our three maps above are *point* queries: they filter the data rather than aggregating it. This means we have to be extra careful. We cannot simply run the query and add noise to the result (as we might do with an aggregation query) since this will release private information! To understand why, suppose we framed our question like this: *“How many taxis picked up a passenger outside The Greenwich Hotel at 19:35 on July 8, 2013?”* By running our original query, even after adding noise to the output coordinates, we have indirectly answered this question accurately. This constitutes a privacy breach, as information about specific individuals can be learned in this way.

Instead, we have to turn our query into an aggregation. If we place a grid over our map, and count the number of output coordinates that fall into each cell, we end up with a set of counts that are generally independent of each other. Differential privacy is now straightforward to apply, as we are working with aggregate quantities. The privatized versions of these three queries are displayed below. Please refer to the appendix for a more technical discussion of how this was done.

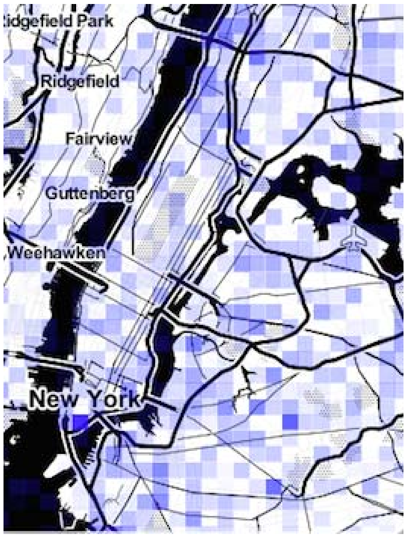
We can appreciate from these privatized maps the importance of ε, the privacy parameter, which I introduced in Differential Privacy: The Basics. When ε is low we can no longer accurately track either celebrity, nor learn how much they spent. In fact, it takes unreasonably high levels of ε to reliably locate them. We could opt for a finer grid, indeed we could all but replicate the point query with a fine enough grid, but differential privacy’s Laplace mechanism is robust enough to effectively ensure that the noise obfuscates any actual difference in the data. The privatization of the Hustler query is more interesting – since it is less targeted, the difference caused by the participation of one individual is less pronounced. As a result, there is more information in the privatized output – for example, the square over Wall Street still stands out, which echoes the actual evidence shown above.



Cooper, Privatized
(Click to Interact)

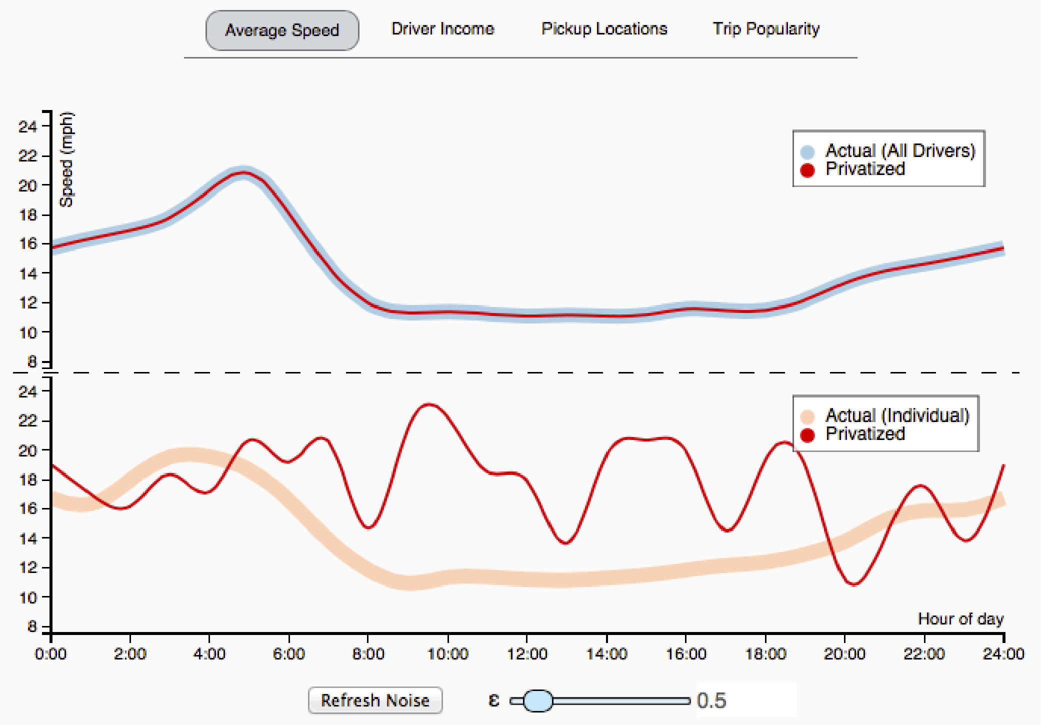


Alba, Privatized
(Click to Interact)



Hustler Customers, Privatized
(Click to Interact)

What about other queries? After all, did we just turn the data into nonsense? Clearly, our ability to learn what we sought from these queries was severely compromised. However, since they were so targeted, one could argue it is a good thing that the results are hard to interpret. The visualization below shows the results of other queries, and demonstrates that useful insights may still be extracted, particularly when aggregating over all records.



Click to Interact

Concluding Remarks

Now that you're an expert in differential privacy, I urge you to consider: what should have been released by the NYC Taxi and Limousine Commission? How should the public be able to access it? Is this in line with Freedom of Information Law? If not, how should the law be changed to accommodate private data release?

It is only by considering these questions that we can hope to motivate a change in practices. With data privacy and security mishaps cropping up in the news almost daily, the topic has never before received so much attention. This post is yet another reminder of the fallibility of our systems and practices. As we know, differential privacy, with its strong guarantees – borne out of its sturdy fundamentals – offers a solution to many of these concerns.

As data scientists, it is in our best interests to encourage free data flow in our society, and lead the charge to ensure participants in these datasets are protected. The science is clearly there, and will continue to be developed in the coming years. The onus is now on industry and governments to pay attention and embrace differential privacy to protect their stakeholders and citizens.

Appendices

SQL queries used in this analysis

Tracking Bradley Cooper and Jessica Alba:

```
SELECT D.dropoff_latitude, D.dropoff_longitude, F.total_amount, F.tip_amount
FROM tripData AS D, tripFare AS F
WHERE D.hack_license = F.hack_license AND D.pickup_datetime = F.pickup_datetime
  AND pickup_datetime > "2013-07-08 19:33:00" AND pickup_datetime < "2013-07-08 19:37:00"
  AND pickup_latitude > 40.719 AND pickup_latitude < 40.7204
  AND pickup_longitude > -74.0106 AND pickup_longitude < -74.01;
```

```
SELECT D.dropoff_latitude, D.dropoff_longitude, F.total_amount, F.tip_amount
FROM tripData AS D, tripFare AS F
WHERE D.hack_license = F.hack_license AND D.pickup_datetime = F.pickup_datetime
  AND dropoff_datetime > "2013-09-07 12:19:00" AND dropoff_datetime < "2013-09-07 12:25:00"
  AND dropoff_latitude > 40.727 AND dropoff_latitude < 40.728
  AND dropoff_longitude > -73.994 AND dropoff_longitude < -73.993;
```

Identifying Hustler customers:

```
SELECT dropoff_latitude, dropoff_longitude
FROM tripData
WHERE pickup_latitude > 40.767249 AND pickup_latitude < 40.768
```

```
    AND pickup_longitude > -73.996782 AND pickup_longitude < -73.995538
    AND HOUR(pickup_datetime) < 6
    AND trip_distance > 5;
```

Privacy calculations

I have chosen to use the Laplace mechanism to privatize the above 3 queries, as it is relatively easy to apply and explain. However, in general, the Laplace mechanism is *not appropriate* for geospatial data. For example, it does not consider topographical features – inspections of the privatized maps show positive counts in the middle of the East River! Rather, there are more complex methods for adding noise to spatial data – Graham Cormode's paper, Differentially Private Spatial Decompositions, offers a more thorough mechanism, while this blog covers the topic more generally. However, I will proceed with Laplace, in the hope that, by moving away from discussions of complicated mechanisms, the core tenets of differential privacy may be grasped more easily.

So how should we go about privatizing the above queries using the Laplace mechanism? As mentioned above, these are point queries. We should not apply noise directly to the output, because our sensitivity is essentially undefined. Why? Let's consider the identity query, *Q(D) = D*. This is a point query – there is no aggregation. What happens if we remove a record from the dataset? We have *Q(D') = D'*. Fine. But now look at what happens to our sensitivity. Recall the formula:

$$\Delta f = \underset{D,D'}{max}||Q(D) - Q(D')||_1$$

This formula is only defined when *Q()* returns a vector of fixed length (independent of *D*), such as an average, or a sum. With *Q(D)=D*, this is of course not the case, and so we cannot calculate the sensitivity. Intuitively, this result is true for other point queries. So we need to find another way to privatize point queries.

We know that privatizing aggregation queries is a lot more straightforward. Is there a way to turn our point query into an aggregation query? What if we bucketed the data into similar groups and then counted how many records returned by the query fall into each group? If we define our groups tightly enough, we would fully replicate the point query. We end up with a series of independent counts, which is easily privatized – the change from removing one record, *any* record, is at most 1. So this is our sensitivity. Adding Laplace noise with a sensitivity of 1 to the count in each bucket guarantees differential privacy.

Of course, if we have too many buckets, we will have lots of 0s and 1s. Since adding Laplace noise to this essentially hides this difference (when ε is at a reasonable level of course), we may struggle to learn anything from this data. By allowing each bucket to incorporate a greater range of values, we can obtain more meaningful privatized counts, albeit by sacrificing a degree of accuracy. The optimal way to group the data therefore depends on which features of the output are of interest.

This method is not confined to one-dimensional responses. As can be seen in the privatized maps above, the histogram is 2-dimensional, reflecting the latitude / longitude output from each query. In fact, any mapping of n-dimensional data into independent buckets is sufficient to ensure differential privacy. For the celebrity queries, I did treat the latitude/longitude coordinates and the fare/tip amounts as two separate histogram queries, as these pairs are broadly uncorrelated with each other.

For additional detail on the privacy calculations used throughout this post, the reader is encouraged to examine the code behind the various visualizations. While it is clear that the Laplace mechanism is applied slightly differently each time, it is not hard to imagine how this might be automated so differential privacy could be applied more generally.

FILED UNDER: DATA SCIENCE, GENERAL TAGGED WITH: DATA SCIENCE, DIFFERENTIAL PRIVACY, NYC TAXICAB DATA, PRIVACY

« Differential Privacy: The Basics HLL talk at SFPUG »

Comments

cjbprime says:
September 17, 2014 at 5:45 am

Amazing post. If you said to someone "Hey, I wanted to know where you went after the cab picked you up last year, so I called up the cab company and asked them where they dropped you off and they told me", they would be outraged at (your behavior and) the breach of privacy shown by the cab company. But the city released a dataset that allows exactly this query. What were they thinking?

Another attack: if someone you were with got in a cab in 2013, and they told you where they were going, and you remember the approximate time and location, you can tell whether it was their true destination *regardless of how many other people were being picked up at the time*, because you don't have to identify the exact ride they took, you only have to see whether any rides went to the place they told you they were going.