Jonathan Dyer
CS 1699: Privacy in the Electronic Society
Homework 3

# 1. Google Maps navigation

## EVENT:

I used Google Maps to navigate to my jazz piano teacher's house, roughly 30 minutes away. It was turn-by-turn on my smartphone, so I had constant updates about traffic patterns as I went. This is a weekly drive, typically around the same time every week (but I'm not comfortable enough with the route yet that I can do it without assistance).

## DISCLOSURES:

Quite a bit of physical information was given up in the course of this interaction, as well as perhaps some behavioral information. Let's proceed from more obvious disclosures to perhaps less apparent ones.

Firstly, I obviously gave away my real-time physical location to Google, in order to receive directions as well as traffic updates as I went. Additionally, because this is a repeated route for me (weekly), I gave up the information about my behavior (*drives to this point every week around this time*). It is worth noting here that already, even in a dataset of many users of this service, just the two points given up here (my home and my piano teacher's house) are enough to **uniquely identify** me. No one else spends the bulk of their time at my house and also drives to my piano teacher's once a week.

Now notice that all of this information is not just associated with my physical device but with my actual identity (name, birthday, education, gender, etc.) as stored by Google *somewhere*, since of course I have a Google account (providing email and other services) which I was logged into at the time of use. In fact, by examining the other data Google has, it will be discovered that I keep a weekly reminder of this piano lesson in Google Calendar, under a title that includes my piano teacher's name. This is significant, because a quick check of the address to which I drove will verify that in fact that person still lives there, and has something to do with pianos; thus I not only gave up my own data, but my piano teacher's as well!

Now we may wish to believe that no further data is disclosed, but it has to be considered that everything I just described is stored somewhere, potentially a third-party's servers which are licensed by Google for storage of just this kind of data. Who is this other entity? I have no way of finding out, at least not easily. Thus my location data is, hopefully in an encrypted form, stored in this other location.

Finally, if my location information is being shared by Google with anyone else (which it is, for advertising purposes), then they also have it, along with their data storage solutions, etc., bringing an untold number of other entities into the picture. I may as well have broadcast my real-time location to the whole internet.

## THREAT MODEL:

In order for this activity of navigating through Google Maps to be "worth it" for the average user, several assumptions must be made. Firstly, in order for this information sharing to be **safe** (truly the matter of physical location privacy is one of safety), we must assume that Google and any associated parties that are given access to location info are not malicious in intent.

If it is shared in some aggregated or anonymized form with other parties, we have to suppose that the data has been obfuscated effectively--I don't know of an easy way to discover how. Did they rely on *k-anonymity* or some other measurement? I'm not an Apple user, so I can't count on any kind of differential privacy, can I? What about federated learning? These are big assumptions, since I have no way of knowing if my tracking data is de-anonymizable from whatever set is provided to ad companies and the like.

As if those assumptions weren't enough, we must also suppose that Google and friends (i.e. anyone with this data) are competent and secure enough not to allow unauthorized access to the info. This means that their servers are physically and electronically secure from hackers or physical compromise. That in turn requires a set of assumptions about the supply chain of all these parties being trustworthy and reliable as well.

Finally, if we don't want to feel **spied upon**, we also assume that Google and any third-party associates are not using such location information to actively take advantage of users or bring about any negative effects for them.

If there is truly reasonable doubt about *any* of the above assumptions, then an average user might be well-advised not to feel that their privacy is upheld in such a usage, and may question whether continued use of the Google Maps service is indeed "worth it".

## EVALUATE THREAT MODEL:

It would be a little too optimistic to suppose that Google and others are not leveraging this information for *some* kind of financial gain. However, Google in particular is a large enough company, serving enough people, that I (and certainly the average user) believe that they are competent enough to keep my information safe from **other unauthorized parties**. This addresses the direct security of my information. What it doesn't consider, however, is the security policies of **authorized third parties** who have access to my data. I do *not* believe that without exception they have all the same security in place as Google, although I imagine the average user doesn't consider this much.

Additionally, the average user is likely not totally aware of who precisely has their info, and almost certainly has not considered de-anonymization as a potential threat to their privacy. I do believe that Google makes a reasonable effort to protect my anonymity (due to the reputation they have to protect), but I think the average user might be more wary about using the service if they knew all of the security assumptions that are implied.

## IMPROVEMENT IDEAS:

One idea I have for decreasing the necessary trust in the system is de-linking the location data with your account, and instead associating it with the physical device being used to navigate. This would have the advantage of not risking personally identifiable information on a potentially risky site or company. This would allow us to relax the assumptions surrounding the security of each party involved in acquiring, storing, and processing the location information. However this does reduce the utility of the service if one relies on the suggestions, saved places, etc.

Another possibility inspired by our discussion in class is to use **mix zones** or some other temporal/physical decrease in granularity to provide the service. Traffic data is relatively localized, so perhaps reporting travel details in smaller chunks, say every hundred yards or so (under different pseudonyms) would allow for the same functionality while protecting privacy more fully.

I think that following this idea while still consistently providing personalized services is simply not as valuable for Google. They would be giving up precious marketing information about habits and locations of their users, which is simply not as lucrative in areas such as advertising. The fact of the matter is that not enough of their users care/are concerned about it for such a change to be cost-effective to Google.

## 2. Signing into My Pitt to print a document

### EVENT:

The other day I used my smartphone to log in to My Pitt in order to access my university-provided Box account. I downloaded a document from there and then used the email client on my phone (connected to my Pitt email account) to send it to the Pitt "MobilePrint" server to be added to my print queue.

### DISCLOSURES:

A surprising amount of information is given up in this seemingly menial interaction with my university account. Certainly basic information was recorded by Pitt's servers such as my IP address, time of login, and user credentials. But the portal to Box means that the company was also made aware of my access to the account, as well as to the document in question. Further, due to Pitt's mandatory multi-factor authentication, all parties involved could infer that I was with my smartphone at the time of access.

Upon downloading the file, a couple of anti-malware apps on my phone scanned it to ensure that it was safe, potentially uploading parts of the file for analysis. Then I used my university email to send the file, which meant giving up the document directly both to Pitt's servers (wherever they keep documents ready to be printed) as well as to Microsoft (who hosts the Outlook email service upon which PittMail is based).

### THREAT MODEL:

In order for this interaction to be "worth" the disclosures mentioned above for the average user, they must assume a large number of potentially unjustified facts. I will mention as many as possible.

First, there is an entire set of assumptions about the device that I was using that I implicitly accepted in order to feel **comfortable** that my data was private. This includes that the entire supply chain of the physical hardware was secure and uncompromised, that the software development process was safe, and that nothing untoward had happened to my device since the time I received it (via malicious applications or viruses, for example).

Additionally, it is supposed that all parties involved (Pitt, Box, Microsoft, Samsung (device manufacturer), etc.) are not malicious in intent nor compromised in their data security. If the document were private or sensitive, then it could have implications in any number of areas. Additionally, if my physical proximity to my cell phone is a matter of concern, that data could even be a safety concern in extreme circumstances.

Now consider the matter of data storage at every point in the process. I assumed that Box (or some hacker) had not accessed and modified my file, and that Pitt's email servers would keep it in some secure way until I released it at a printer somewhere on campus. After that, I suppose that they securely store or delete that file so as to provide a sort of *forward secrecy* in case of some security breach. Lastly consider that I assumed my phone was secure enough to keep the file locally without harm (as I still have not deleted it).

Finally, look at the data in motion. I assumed that the wifi access point I was using was secure from malicious actors, or at least that my traffic was encrypted well enough to avoid compromising its confidentiality or integrity. Similarly for the email server, I supposed that my email could not be compromised or intercepted (or if so that it would not be decipherable). In essence I expected complete security (and hence quite a bit of privacy) in the overall interaction.

## EVALUATE THREAT MODEL:

There are a *lot* of assumptions above, all of which must be accepted either implicitly or explicitly (mostly the former) in order for the average user to really feel **comfortable** that no unwanted entity will have access to their data. I believe them, with varying degrees of confidence--for instance, I know essentially nothing about the supply chain of my phone or the encryption status of Outlook emails.

I think the average user understands most of the assumptions above on some level, and accepts them (thinking about each one explicitly every time is exhausting and impossible) typically without question. If brought to their attention more explicitly, these assumptions might be cause for some reflection and caution about the way in which users view and use this system (here, specifically Box and Outlook), and what kind of data they toss around in those environments.

## IMPROVEMENT IDEAS:

One way to reduce information disclosure that comes immediately to mind is to bring data storage in-house to Pitt-owned servers. That is, rather than relying on Box for university data storage, Pitt could open its own data center and simply care for the data of its constituents there. This would cut out one of the third parties that have access to your data as well as eliminating all of the assumptions about that third party that were necessary to make for the above interaction. Rather, we could rely on the assumptions that Pitt's security is sufficient (which is seems to be on the surface at least, two-factor authentication and all).

However, this solution is obviously *extremely* expensive on Pitt's end (we outsource things like data storage for a reason), far too costly to implement. Additionally, if the My Pitt website is anything to go off of, this would likely decrease the usability of the service as well… It simply makes the most financial sense to use a dedicated third party (Box) to provide this service instead.