

Disclaimer:

- The following positions are purely personal
- The details do NOT reflect in any way the products or solutions from employer

Big Data everywhere – The Tricks and Pitfalls of Pseudonymisation and Anonymisation

Philipp Schaumann

<http://sicherheitskultur.at/>

philippschaumann@mailbox.org

Philipp Schaumann, sicherheitskultur.at

CONFERENCE 2016
PRAGUE
APRIL 12-14

1

Smart anything = Big Data everywhere

Technology

Tomorrow's cities: How changing the world

By Jane Wakefield
Technology reporter



Home

Guardian sustainable business Technology and Innovation

A truly smart city is more than sensors, big data and an all-seeing internet

Investing billions in big data and smart technology isn't the only answer to building more sustainable urban areas. We need to focus on the big levers

How Smart City Barcelona Brought the Internet of Things to Life

Cities Get Smart with Big Data

by Ellis Booker | September 25, 2014 5:30 am | 0 Comments

Comments

Forget moving sidewalks and robot police. The biggest technology change in our cities will involve data, and lots of it.

Big Data is incompatible with data protection

Data protection:

- Data can only be collected for a well-defined purpose
- Data can only be collected on well-defined legal basis or with consent

Big Data

- As much Data as possible is collected without knowing what it can be used for
- Consent of the person is assumed or data is considered “anonymous”

Anonymity (legal definition)

Anonymity = no name (no identifying data field)

. . . that it can no longer be used to identify a (natural) person by using “**all the means likely reasonably to be used**” by either the controller or a third party - the processing must be irreversible

Also, „events“ that are relating to 1 person shall not be linkable (no „tracking“)

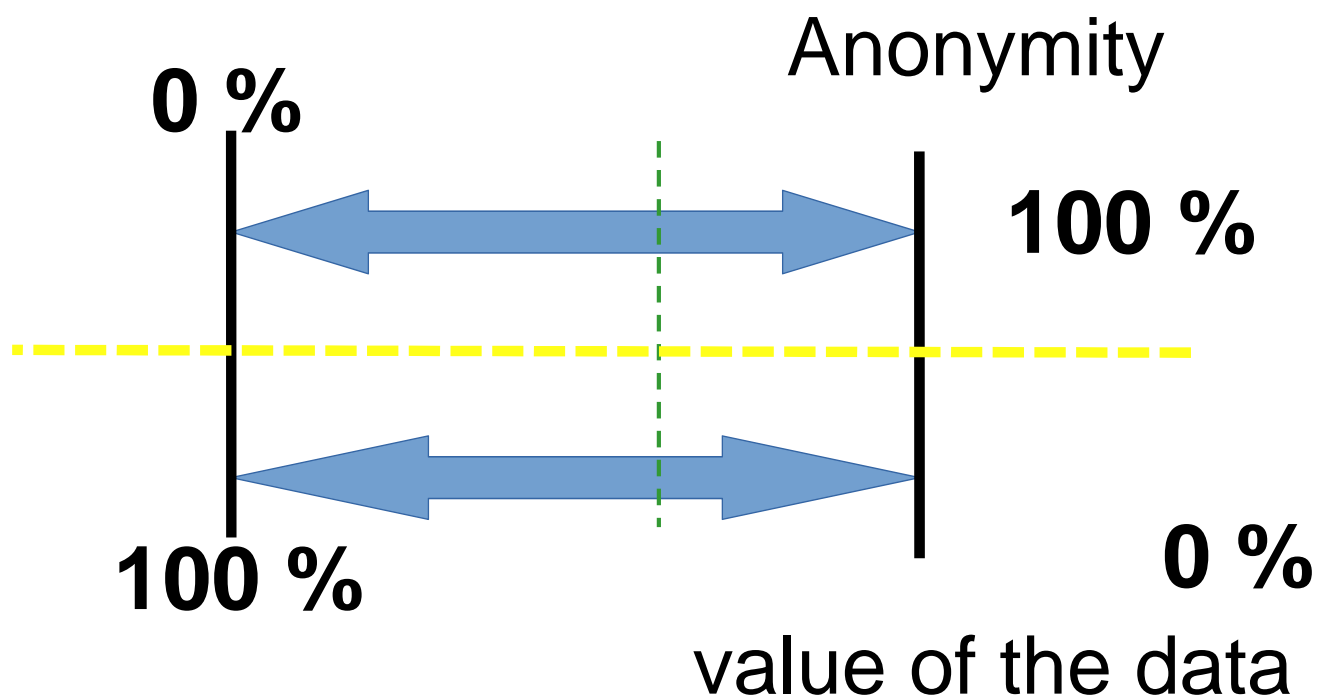
Degree of Anonymization

100 % Anonymity means that

- no entity in the (known) universe can,
- even with lots of compute power and
- unlimited access to all other data sources,
- now and in the foreseeable future re-identify the person

Complete anonymity is very hard to do and the data loses most of its value – it is equal to synthetic data

Anonymity versus Value



Pseudonymity

Pseudonymity = name and other identifying data fields have been replaced by values that allows for the **original data controller** to match the data with other data – e.g. it allows matching of events concerning the same person, etc.

The **recipient of the data** cannot (by legal means) identify the person.

Anonymous data are no longer protected by D.P. law

Pseudonymous data is still protected by D.P. law

Multiple events per person

- If there are multiple events per person (either in the same record or in multiple records) then pseudonymization and even in-complete anonymization will still allow „tracking“ of the events of 1 person = violation of strict anonymity

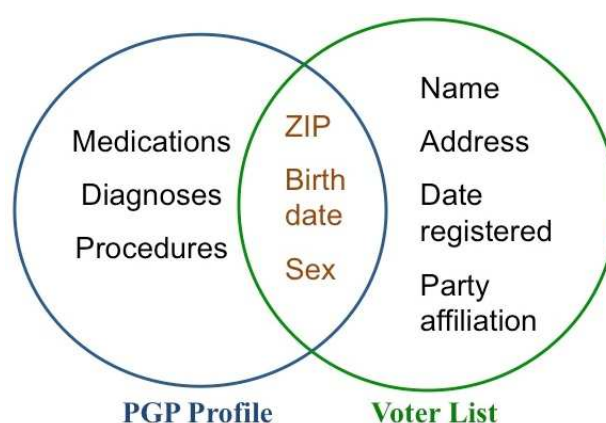
Identifier vs. Quasi Identifier

Identifiers need to be removed (name, PP#, Soc. Sec. No, email-address, account #, phone #,

Quasi-Identifiers (QI)/ key attributes need to be considered in combination with each other + possible other sources for this data (Zip Code, Gender, Age, Address, education, profession, memberships, participation in groups or events,)

De-Anonymization (1) Unicity = Uniqueness

- De-Anonymization is possible if the combination of several data items generate uniqueness.
- E.g.
sex + birthdate + 5-digit ZIP-code identify 87% of US-population, sex + birthdate + city identify 53 % (all identifiable against voter register)



Source:
Personal Genome Project
(PGP) De-Anonymisation

AOL search data (2006)

AOL apologizes for exposing search data

A spokesman for the ISP-turned-portal says the release of keyword search information from about 658,000 anonymous AOL users was a "screw up" that was based on good intentions.

The Netflix case (2007)

Projecting two databases onto the same subspace

Each record consists of anonymized personal data plus movie ratings with the date of the rating. (movie name, date, rating value)

Each record can be represented by a point in a multidimensional space, where each attribute is a coordinate.

They are so distant that after partitioning the space into regions, each region contains only one record. In the Netflix case, records were sufficiently unique with just 8 movie ratings given within 14 days of distance.

The very same selection of just 8 rated movies constituted a fingerprint of the expressed ratings, not shared between two data subjects within the database.

The researchers matched the supposedly anonymous Netflix data set with another public database with movie ratings (the IMDB), thus finding users who had expressed ratings for the same movies within the same time intervals.

DNA Study (2013)

Apr 25, 2013 @ 03:47 PM 18,872 views

Harvard Professor Re-Identifies Anonymous Volunteers In DNA Study

... has re-identified the names of more than 40% of a sample of anonymous participants in a high-profile DNA study, highlighting the dangers that ever greater amounts of personal data available in the Internet era could unravel personal secrets.



Harvard Professor Latanya Sweeney

Montjoye Mobility Study 2013

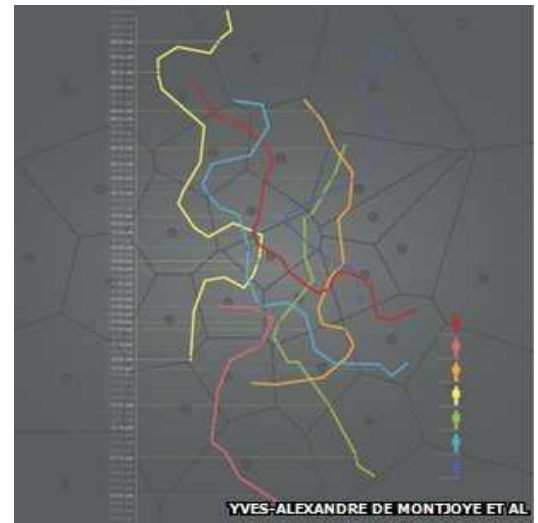
... fifteen months ... for one and a half million individuals:

... in a dataset where the **location of an individual is specified hourly**, and with a **spatial resolution** equal to that given by the **carrier's antennas**, four spatio-temporal points are enough to **uniquely identify 95% of the individuals**

De-Anonymization (2)

Unicity = Uniqueness

- Day-long location tracking (mobility pattern):
Week-day daylight location + main night time location is unique for 90+ % of the people
- In a village, a bank might have only a single male/female customer in a certain age range – how many customers of an Autstian bank are living in a mid-size town in Norway?
Probably just 1



NY Taxi Rides (2014)

Dataset released by the New York City Taxi and a Limousine Commission - every taxi ride 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.

NY Taxi Rides (2014) - “Chasing” Celebrities

searching through images of “celebrities in taxis in Manhattan in 2013” to find enough information to identify the correct record in the database.



Jessica Alba (Click to Explore)

Philipp Schaumann, sicherheitskultur.at

Slide 17

NY Taxi Rides (2014) “Chasing” others –

A few innocent nights at the gentlemen’s club

. . . ran a query to pull out all pickups that occurred **outside the club after midnight and before 6am**, and mapped the drop-off coordinates

. . . was able to pinpoint certain individuals with high probability.

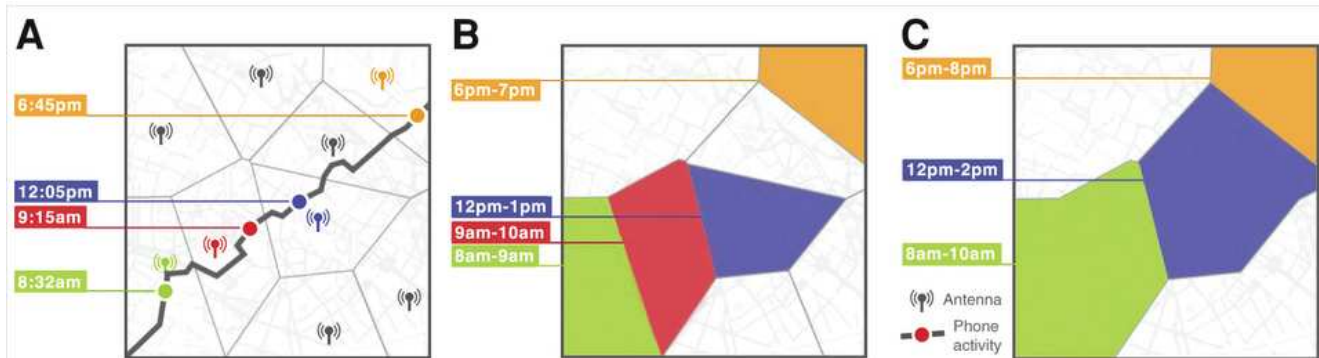
Examining one of the clusters in the map above revealed that only one of the 5 likely drop-off addresses was inhabited; a search for that address revealed its resident’s name.

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!

Philipp Schaumann, sicherheitskultur.at

Slide 18

Decreasing the resolution of Location Data (time and space)



A shows the full tracking details

B and C show increasing coarseness being applied

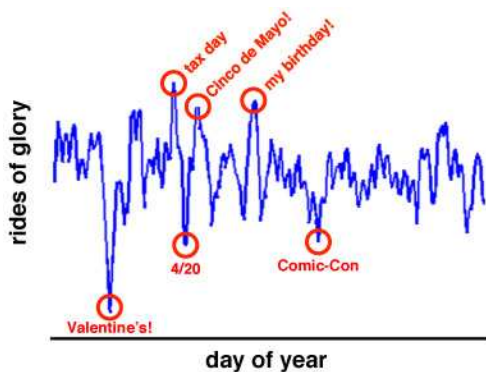
. . . But even coarse datasets provide little anonymity (Montjoye 2013)

Your smartphone trail: Location data is very Talkative

- Where do you spend some of your nights? (full night, only short visits)
- Which pubs, clubs do you visit?
- Which sport clubs do you visit?
- Which churches do you visit?
- Which political meeting do you visit?
- Where do you shop?
- What movies do you watch?

Uber (2015)

- . . . anyone who took a ride between **10pm and 4am on a Friday or Saturday night**, and then took a **second ride** from within 1/10th of a mile of the previous nights' drop-off point **4-6 hours later** (enough for a quick night's sleep).



Philipp Schaumann, sicherheitskultur.at

Slide 21

And the solution for Smart Cities?

1. Collect single data points wherever possible, never full travel data (e.g. for road congestion data)
2. If single data points are not possible, make data very very coarse grained (time + location)

And the solution for Smart Cities?

3. Watch out for uniqueness in the data set! (“quasi identifier”)
4. Hashing without salt allows easy brute force attacks
5. Remove as many data items as possible

And for the Consumer

Be very suspicious if somebody promises you to “anonymize your data”

Leave as little traces as possible

Sources for De-anonymization Examples

2000 L. Sweeney study: <http://dataprivacylab.org/projects/identifiability/paper1.pdf>
 2006 AOL search data: <http://searchsecurity.techtarget.com/news/1208972/AOL-apologizes-for-exposing-search-data>
 2006 AOL search data: http://usatoday30.usatoday.com/tech/columnist/andrewkantor/2006-08-17-aol-data_x.htm
 2007 Netflix: http://archive.wired.com/politics/security/commentary/securitymatters/2007/12/securitymatters_1213
 2007 Netflix: https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf
 2013 DNA study: <http://www.forbes.com/sites/adamtanner/2013/04/25/harvard-professor-re-identifies-anonymous-volunteers-in-dna-study/>
 2013 Montjoye mobility study: <http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html>
 2013 Sweeney Patient records: <http://dataprivacylab.org/projects/wa/1089-1.pdf>
 2014 NY taxi rides: <http://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/> (with tutorial about differential privacy)
 2014 Taxi rides NY: <http://gawker.com/the-public-nyc-taxicab-database-that-accidentally-track-1646724546>
 2015 Uber rides: <https://web.archive.org/web/20140828024924/http://blog.uber.com/ridesofglory>
 2015 credit card data: <http://bits.blogs.nytimes.com/2015/01/29/with-a-few-bits-of-data-researchers-identify-anonymous-people/>
 2015 credit card data: <http://www.sciencemag.org/content/347/6221/468.summary?sid=de616c3b-4b69-416c-b497-1df7cec3d033>

Overview (German language)
http://sicherheitskultur.at/Glaeserner_Mensch.htm#deanom

Where to learn more about Pseudonymization?

Much more technical tutorial from the
 ARTICLE 29 DATA PROTECTION WORKING
 PARTY of the EU

http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

Thanks



Philipp Schaumann
philippschaumann@mailbox.org