Does fair anonymisation exist? (ongoing research)

Dr. Zoltán Alexin, PhD.
University of Szeged,
Department of Software Engineering
Árpád tér 2., Szeged, H-6720
e-mail: alexin@inf.u-szeged.hu
http://www.inf.u-szeged.hu/~alexin

Alexin, Z.: Does Fair anonymisation exist?, BILETA 2013, April 10-12, 2013. Liverpool, UK

Acknowledgement

This work was partially supported by the European Union and the European Social Fund through project FuturICT.hu (grant no.: TÁMOP-4.2.2.C-11/1/KONV-2012-0013).

Preliminaries

- Hungarian health government created a registry from health insurance accounting data in 2004
- Registry was created by the Decree 76th of 2004 of the Health Minister
- In 2006 I challenged the Decree before the Constitutional Court, but the case No. 937/B/2006 was dropped in 2007
- Paul Ohm: Broken promises of privacy
- I decided to collect evidences to support my view and obtained (date of birth, living location ZIP) distribution data from the National Population Registry

Content

- What anonymisation is?
- Benefits of anonymisation and its potential dangers for privacy
- Threats of mindless anonymisation
- Historical health insurance accounting databases
- Example: the Hungarian IMD (Itemised Medical Database)
- Estimating re-identification risk using Hungarian Population Registry data
- Conclusion

What anonymisation is?

- Anonymity is derived from the Greek word *ἀνωνυμία* (anonymia), meaning „without a name” or nameless.
- The goal is staying hidden, not to reveal one’s identity.
- In IT technology: remove all information that could lead to an individual being identified.
- Generally, removal of personal names are not enough.
- People can be identified not only by their names, e. g. workplace, job and boss; or exact place of living (number, street, city); or school, date of birth, and form-master.

Benefits and dangers

- Benefits:
  - Data can be processed without doing harm for data subjects
  - The strict Personal Data Protection Act need not be applied
  - Ethical questions do not arise
  - Data sets can be shared, even can be sold
- Dangers:
  - People and companies (workplaces) are publishing more and more information about themselves on the Internet
  - We cannot guess the future (what sorts of information is being published)
  - Companies harvesting public information from the web (names, photos, dates of birth, place of living, schools etc.)
  - Industrialized re-identification of anonymous scientific databases

People	Job	Date of Birth	Health data
10784343	teacher	3rd May 1965	HIV+
13453453	accountant	2nd June 1946	Cancer
53353534	driver	17th August 1964	Syphilis+

Find us on Facebook

	Hobby	Date of Birth	Job
David Simon	Hiking	17th August 1964	driver
John Smith	Sailing	3rd May 1965	teacher
Jackie Chan	Diving	2nd June 1946	accountant

Threats of careless anonymisation

- The effect of anonymisation cannot be reversed
- The publicised information cannot be revoked
- Careless anonymisation can cause harm, that cannot be repaired (compensated)
- It is a future risk (the possibility of reidentification continuously threatens data subjects)
- Cannot be considered as a fair data processing
- Elizabeth France (1998): anonymisation is a kind of processing and therefore shall be done upon authorization of a law

Historical health insurance accounting databases

- Health insurance companies are collecting accounting data
- In each case a patient appears in the health system, an accounting data record is created
- Persons are identified by a unique health ID number
- Data are retained for several years for financial inspections and to avoid crimes (fraud)
- The collected data can be used for secondary purposes (research, planning, quality control, retrospective studies)
- The databases are anonymised with a minimum effort – so the letters of the law is satisfied. But?

Itemized Medical Database

- The Hungarian National Health Insurance Fund (OEP) maintains an accounting database
- Retention time is 15 years
- Patients are identified by TAJ (national Social Security Identifier, 9 digits-long)
- OEP quarterly sends updates to the Itemized Medical Database (TEA), which is maintained by the Quality- and Organisational Development Institute (GYEMSZI) of the health government
- The OEP replaces TAJ numbers with a pszeudo-TAJ numbers that are also unique identifiers so as the care events belonging to the same people can be joined together. Otherwise the registry contains detailed data equivalent to the accounting database
- In- and outpatient care events and all prescription data are included
- Retention time is unspecified (lifelong)

Hungarian Population Registry

- 10 million people who have a citizenship or have residential address in Hungary
- Personal names, date of birth, mother's name, ZIP code, city (village), street, number, national personal identifier, document ids issued, a photo taken from the person
- Permanent and temporary residential address
- Maintained by the Central Office for Administrative and Electronic Public Services

Statistical Database for Risk Estimation

- It contains ZIP code, gender and date of birth from all citizens in Hungary (10 004 090 people).
- P-twins (pseudo twins): people, who are living in the same ZIP code district, have the same gender, and was born on the same day. If we have no more information, then they cannot be distinguished.
- The biggest clone contains 11 P-Twins (1 clone, 1975), 12 clones contain 10 P-Twins, etc.

…
1011;1989.01.23.;N;2
1011;1989.02.01.;N;1
1011;1989.03.11.;N;1
(8 million lines)
…

Re-identification with DoB and Living Location Data

- IMD contains date of birth, gender and living location's ZIP
- Whenever one gets to know these data then he can identify people with big probability
- Since everyone's data can be found in the database, the uncertainty of the identification is small
- Other data is available, like medical institution, medical professional's name and body height, weight (in the case of specialists' visits), all screening results, ICD-10 codes
- Since IMD is anonymous, therefore its use is not transparent, no ethical approval, no data protection supervision of the uses

Zip Codes

Alexin, Z.: Does Fair anonymisation exist?, BILETA 2013, April 10-12, 2013. Liverpool, UK

Villages and cities

Inhabitants	Number of ZIP districts	Population	P(1)	P(1&2)
n < 1000	1339	725628	98,218%	99,973%
1000 ≤ n < 5000	1296	2800312	94,798%	99,811%
5000 ≤ n < 20 000	402	3883348	82,026%	97,839%
20 000 ≤ n	73	2594802	49,838%	80,315%
Sum:	3110	10004090		

One is uniquely identifiable if he/she has no P-Twins (singleton).
P(1) = probability of identification. (The number of uniquely identifiable persons / all persons.)

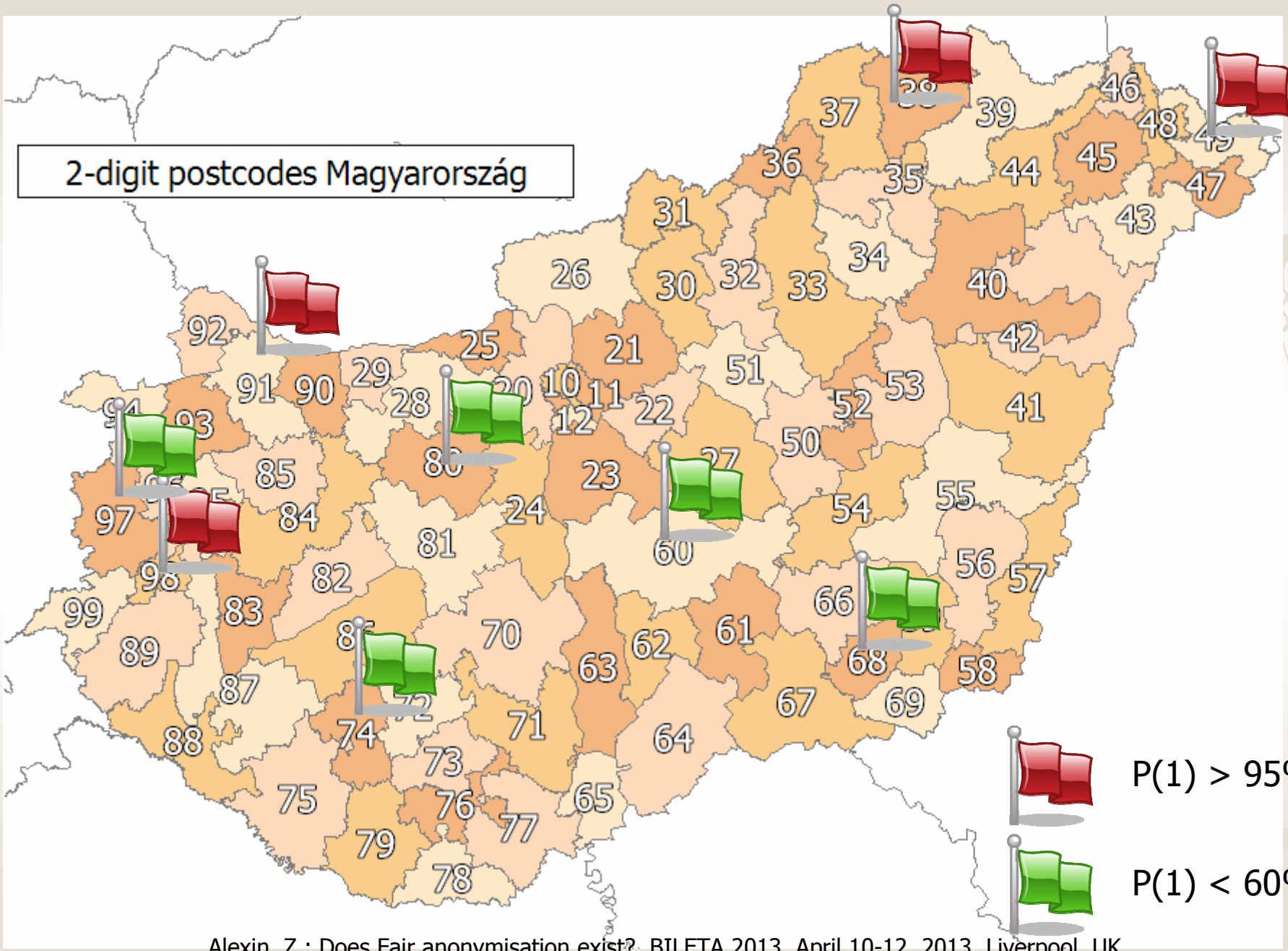
If I can always choose the correct one from two given people, then all single or 2-twin people can be uniquely identified.

P(1&2) = number of singles + 2-twins / all persons.

Distribution of P-Twins

ZIP	Region	1	2	3	4	P(1)	P(1&2)
1xxx	Budapest	1311381	147157	16027	1815	78,902%	96,610%
2xxx	Middle	1364579	154293	27018	5560	76,460%	93,751%
3xxx	N-East	978924	69693	9255	1555	**84,835%**	**96,915%**
4xxx	East	897630	80463	13854	3622	79,942%	94,274%
5xxx	M-East	589923	63741	11594	2604	76,957%	93,588%
6xxx	S-East	690776	83418	18849	5607	**72,585%**	**90,116%**
7xxx	S-West	686907	53665	8645	1795	82,811%	95,750%
8xxx	M-West	780474	75089	19937	6205	75,736%	90,309%
9xxx	West	545256	51508	11789	3142	77,632%	92,300%
Sum:		**7845850**	**779027**	**136968**	**31905**	78,4264%	94,001%

Alexin, Z.: Does Fair anonymisation exist?, BILETA 2013, April 10-12, 2013. Liverpool, UK


2-digit postcodes Magyarország
P(1) > 95%
P(1) < 60%

Years 1..50

365 days of a year

50 x 365 = 18250 boxes

The distribution is quite linear and random across the years and days, therefore if we drop randomly 1000 balls into 18250 boxes they will fall in different box with a high probability.

Conclusion

- Sometimes people are obliged to publicise his date of birth and living location (academics, politicians)
- The transparent pocket (glass pocket law) of Hungary
- MPs are obliged to publicize their biography and assets
- Many times the living location of a minister is said in a TV programme
- Some districts of Budapest behave like a village

- More attention should be paid when creating such a database!

Thank you for the Attention!