

Estimation the level of anonymity of databases



Dr. Zoltán Alexin, PhD.
University of Szeged,
Faculty of Science and Informatics,
Department of Software Engineering
e-mail: alexin@inf.u-szeged.hu
<http://www.inf.u-szeged.hu/~alexin>

About me

- Mathematician, PhD., information retrieval, text processing
- Since 2004, I began studying privacy issues
- Member of a Regional medical research ethics committee, 2009
- Member of the Association on Fair Data Processing, 2009
- Blogger ([Facebook](#), www.tisztesslegesadatkezeles.org)
- Has cases before Civil Courts, Hungarian Constitutional Court, European Commission on fundamental questions of medical data processing
- Achievements: excluding unsubsidized care events from the National Health Insurance Fund database, obligation of ethics approval of medical research projects without intervention, restricting the retention time of medical data at the National Health Insurance Fund
- Data protection officer at the Clinical Center of University of Szeged since 2015.
- Research on Hungarian demographic data (dataset was obtained from the Population Registry of citizens)
- COST IC1206, project WG4 leader, MC member, COST CA19121 („GoodBrother”) MC member, CA19136 („NET4Age”) MC member.



Re- identification attacks

Quasi-identifiers

- Quasi identifiers: data items that don't identify an individual on their own but can become identifying in combination with other quasi-identifiers.
- Quasi identifiers are **not direct identifiers**. Instead, they are partial identifiers such as a zip code, date of birth. There are many people who share a zip code, and many people who share a date of birth but only few share both.
- In other words: such type of data, that an adversary can acquire together with formal identifiers like name, mother's name etc. and can use this information to re-identify a de-identified dataset.
- The question is: what could be a quasi-identifier? Date of birth, zip, job, gender, qualifications, schools, workplace, disease, medical intervention, childbirth.
- A record then could be $r(q_1, q_2, q_3, q_4, q_5, \dots, q_n, d_1, d_2, \dots, d_m)$.



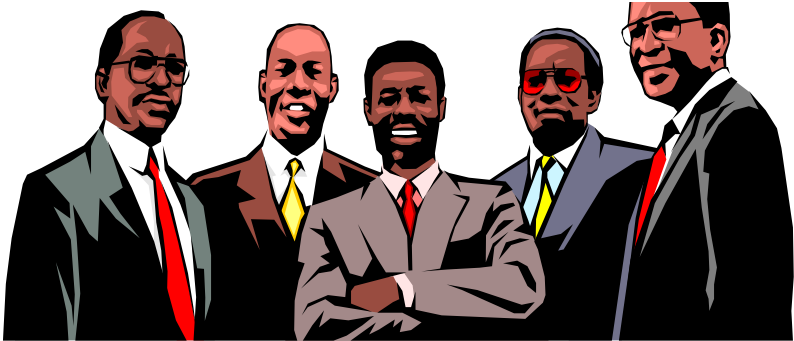
pseudonymous (de-identified) data

Code	Profession	Birthdate	ZIP	Health data
10784343	teacher	1965. 05. 03.	64100	HIV+
13453453	accountant	1946. 06. 02.	65200	Cancer
53353534	driver	1964. 08. 17.	66300	Syphilis+



Name	Hobby	Birthdate	ZIP	Profession
George Simon	hiking	1964. 08. 17.	66300	driver
Stephan King	sailing	1965. 05. 03.	64100	teacher
Luis Brown	diving	1946. 06. 02.	65200	accountant

Additional data for re-identification



What data have been stolen by the hackers?

- FBI and Dept. of Homeland Security employees
<https://www.trendmicro.com/vinfo/us/security/news/cyber-attacks/hackers-leak-information-of-30-000-fbi-and-dhs-employees>
- Psychiatry clinics in Finland
<https://www.theguardian.com/world/2020/oct/26/tens-of-thousands-psychotherapy-records-hacked-in-finland>
- 23andme.com 6.9 million customers' data (DNA profile))
<https://time.com/6342551/23andme-hack-health-data-profiles-compromised/>
- eKRÉTA (nationwide Hungarian public school system)
<https://nedolgozzingyen.hu/2024/04/23/a-kretabol-a-dark-web-re-kerulhettek-tanulok-szulok-iskolai-alkalmazottak-adatai/>
- <https://techcrunch.com/2024/08/12/2024-in-data-breaches-1-billion-stolen-records-and-rising/>



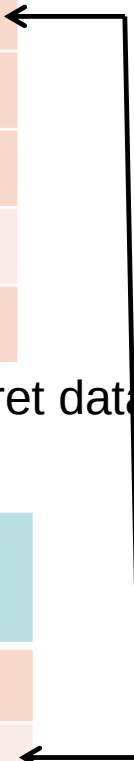
Code	Profession	Birthdate	ZIP	Health data
10784343	teacher	1965. 05. 03.	64100	HIV+
25343567	teacher	1965. 10. 16.	64100	Diabetes
43242444	teacher	1965. 12. 28.	64100	Glaucoma
13453453	accountant	1946. 06. 02.	65200	Lung Cancer
53353534	driver	1964. 08. 17.	66300	Syphilis+

quasi-identifiers



secret data

Name	Hobby	Birthdate	ZIP	Profession
George Simon	hiking	1964. 08. 17.	66300	driver
Stephan King	sailing	1965. 05. 03.	64100	teacher
Luis Brown	diving	1946. 06. 02.	65200	accountant



Code	Profession	Year of birth	ZIP	Health data
10784343	teacher	1965.	64100	HIV+
25343567	teacher	1965.	64100	Diabetes
43242444	teacher	1965.	64100	Glaucoma
13453453	accountant	1946.	65200	Lung Cancer
53353534	driver	1964.	66300	Syphilis+

quasi-identifiers

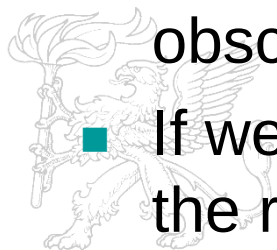


secret data

Name	Hobby	Birthdate	ZIP	Profession
George Simon	hiking	1964. 08. 17.	66300	driver
Stephan King	sailing	1965. 05. 03.	64100	teacher
Solomon Aire	sport	1965. 10. 16.	64100	teacher
Thomas Cook	hunting	1965. 12. 28.	64100	teacher
Luis Brown	diving	1946. 06. 02.	65200	accountant

What can protect individuals?

- It can be presumed that hackers do always HAVE the additional data (quasi-identifiers plus the corresponding identifying information) to re-identify any data record. C.f. GDPR (26) preamble: *Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person.*
- Individuals can be protected only by the embedded obscurity (mathematics, statistics).
- If we can measure the obscurity, then we can estimate the resilience of a dataset against re-identification attacks.





The k-anonymity



k-anonymity

- A dataset is called k -anonymous if for all records in the dataset at least another $k-1$ records can be found that share the same combination of quasi-identifiers.

Code	Profession	Year of birth	ZIP	Health data
10784343	teacher	1965.	64100	HIV+
25343567	teacher	1965.	64100	Diabetes
43242444	teacher	1965.	64100	Glaucoma
13453453	accountant	1946.	65200	Lung Cancer
23130445	accountant	1946.	65200	COVID-19
56432242	accountant	1946.	65200	Stomach cancer
53353534	driver	1964.	66300	Syphilis+
56327578	driver	1964.	66300	Bone fracture
28453576	driver	1964.	66300	Diabetes

- The above dataset is 3-anonymous

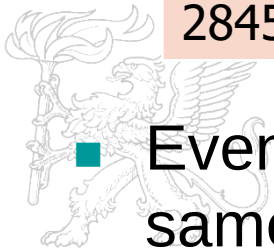
k-anonymity

- It means that an adversary cannot select one single person by a re-identification attack, but only k or more.
- What is the acceptable value of k ? Medical doctors usually say: 3, mathematicians say: 100 or more.
- Very hard to create such a dataset without losing essential information. It is almost impossible.



ℓ-diversity

Code	Profession	Year of birth	ZIP	Health data
10784343	teacher	1965.	64100	Diabetes
25343567	teacher	1965.	64100	Diabetes
43242444	teacher	1965.	64100	Diabetes
13453453	accountant	1946.	65200	Lung Cancer
23130445	accountant	1946.	65200	COVID-19
56432242	accountant	1946.	65200	Stomach cancer
53353534	driver	1964.	66300	Syphilis+
56327578	driver	1964.	66300	Bone fracture
28453576	driver	1964.	66300	Diabetes

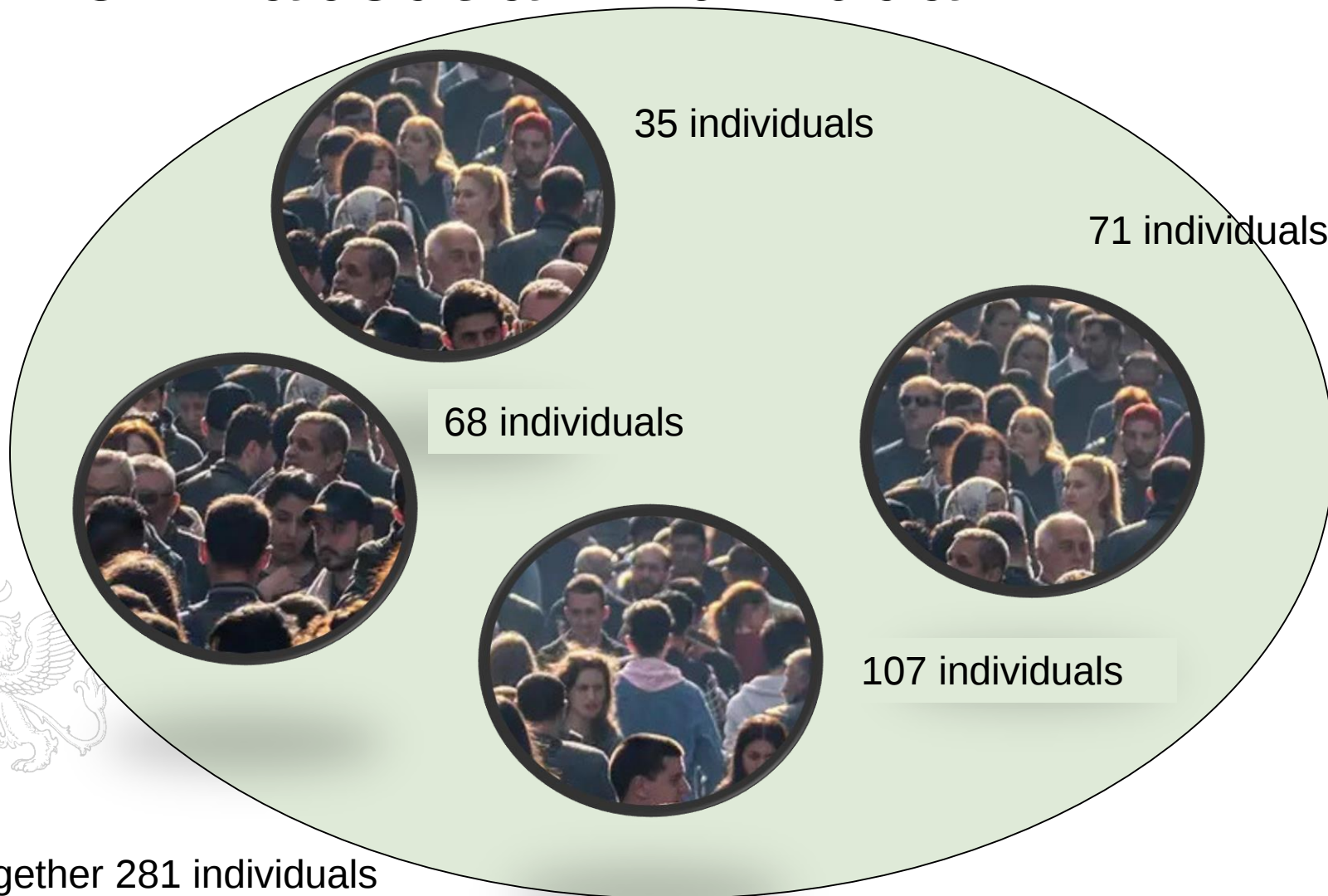


- Even if a dataset is k -anonymous, if the secret data is the same, then we can get to know the secret data without identification.



The entropy of set systems

What amount of information is known about an individual?



How many bits are known about an individual?

68 individuals



To distinguish N different individuals (population) from each other we need $\log_2(N)$ bits.

But, if $k = 68$ individuals are indistinguishable then we need $\log_2(k)$ less bits. We know less information about the individuals in the group.

10111000111010100010101011

$$I = \log_2(N) - \log_2(k) = \log_2(N/k) = -\log_2(k/N)$$
$$I = \log_2(281) - \log_2(68) = \log_2(281/68) = -\log_2(68/281) = 2.0469$$

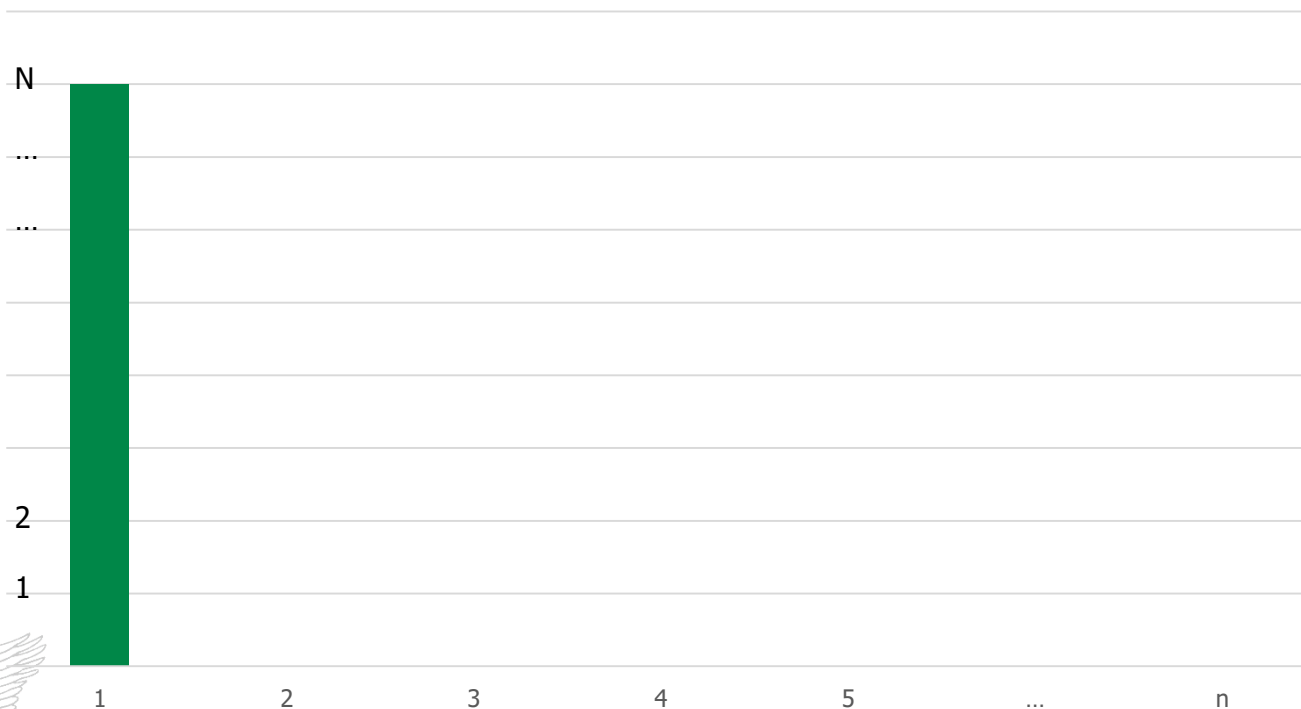
In case of multiple groups, we can sum up the information content of groups.

$$E(D) = \sum_{group} -\frac{\#group}{N} \log_2 \frac{\#group}{N}$$

Altogether 281 individuals

The minimal entropy

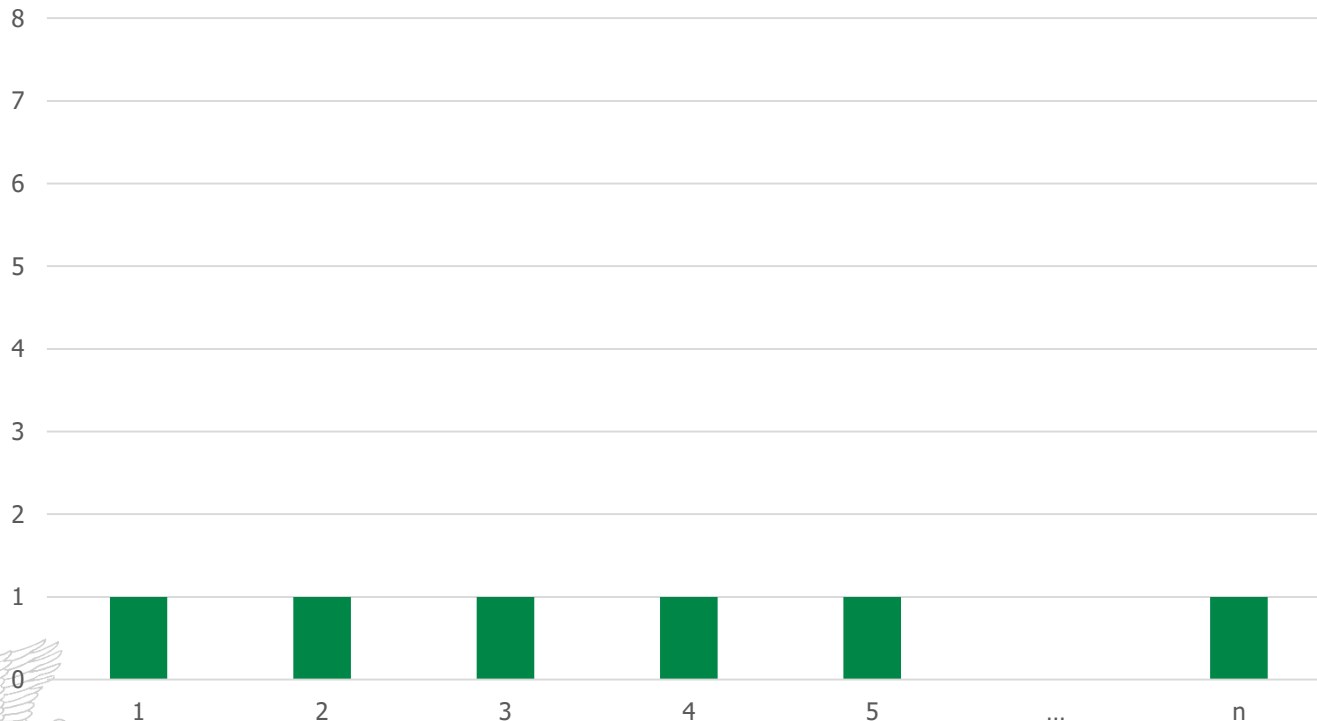
All population is undistinguishable.



$$E(D) = \sum_{i=1,}^{N=N} -\frac{N}{N} \log_2 \frac{N}{N} = -\log_2 1 = 0$$

The maximal entropy

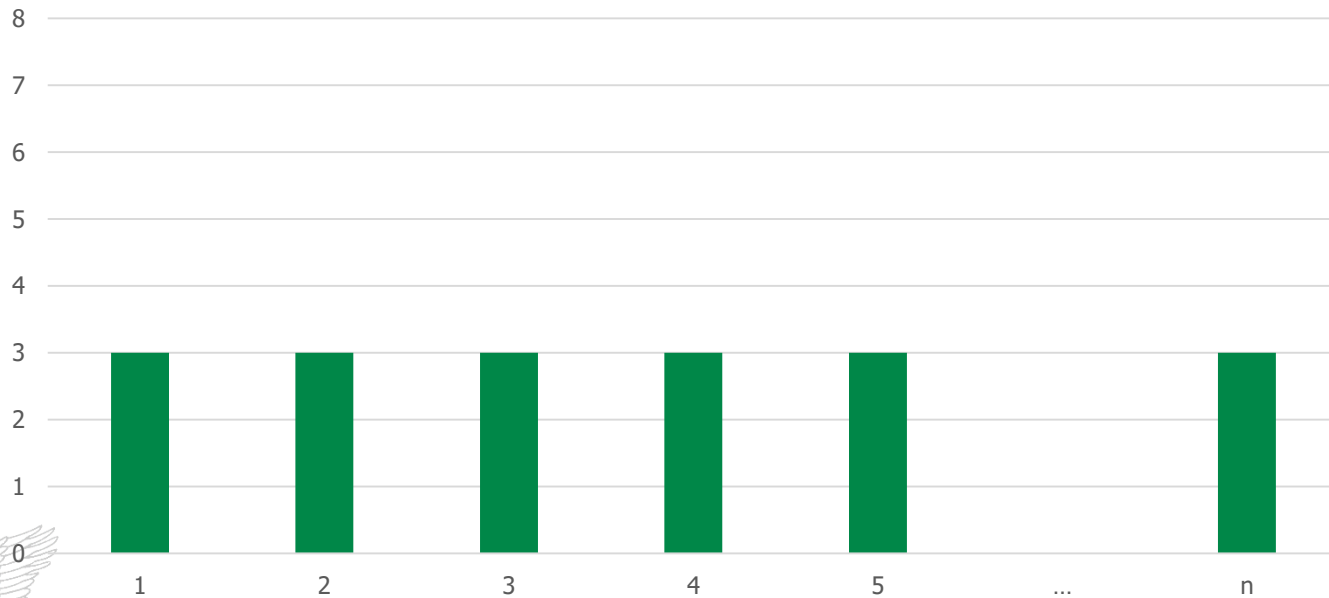
All individuals are singletons.



$$E(D) = \sum_{i=1,2,\dots,n}^{1+1+\dots+1=N} -\frac{1}{N} \log_2 \frac{1}{N} = -N \frac{1}{N} \log_2 \frac{1}{N} = -\log_2 \frac{1}{N} = \log_2 N$$

Each set contains three undistinguishable individuals

All groups contain three undistinguishable person.



$$E(D) = \sum_{i=1,2,\dots,N/3}^{3+3+\dots+3=N} -\frac{3}{N} \log_2 \frac{3}{N} = -\frac{N}{3} \frac{3}{N} \log_2 \frac{3}{N} = -\log_2 \frac{3}{N} = \log_2 N/3$$

The k -anonymity and the entropy

The entropy is closely connected to k -anonymity. If a dataset is k -anonymous then the following formula on entropy holds: $E(D) < -\log_2 (k/N)$. In general case we can obtain an approximate $\hat{k}$ value by reordering the above formula.

$$\hat{k} = \frac{N}{2^{E(D)}}$$

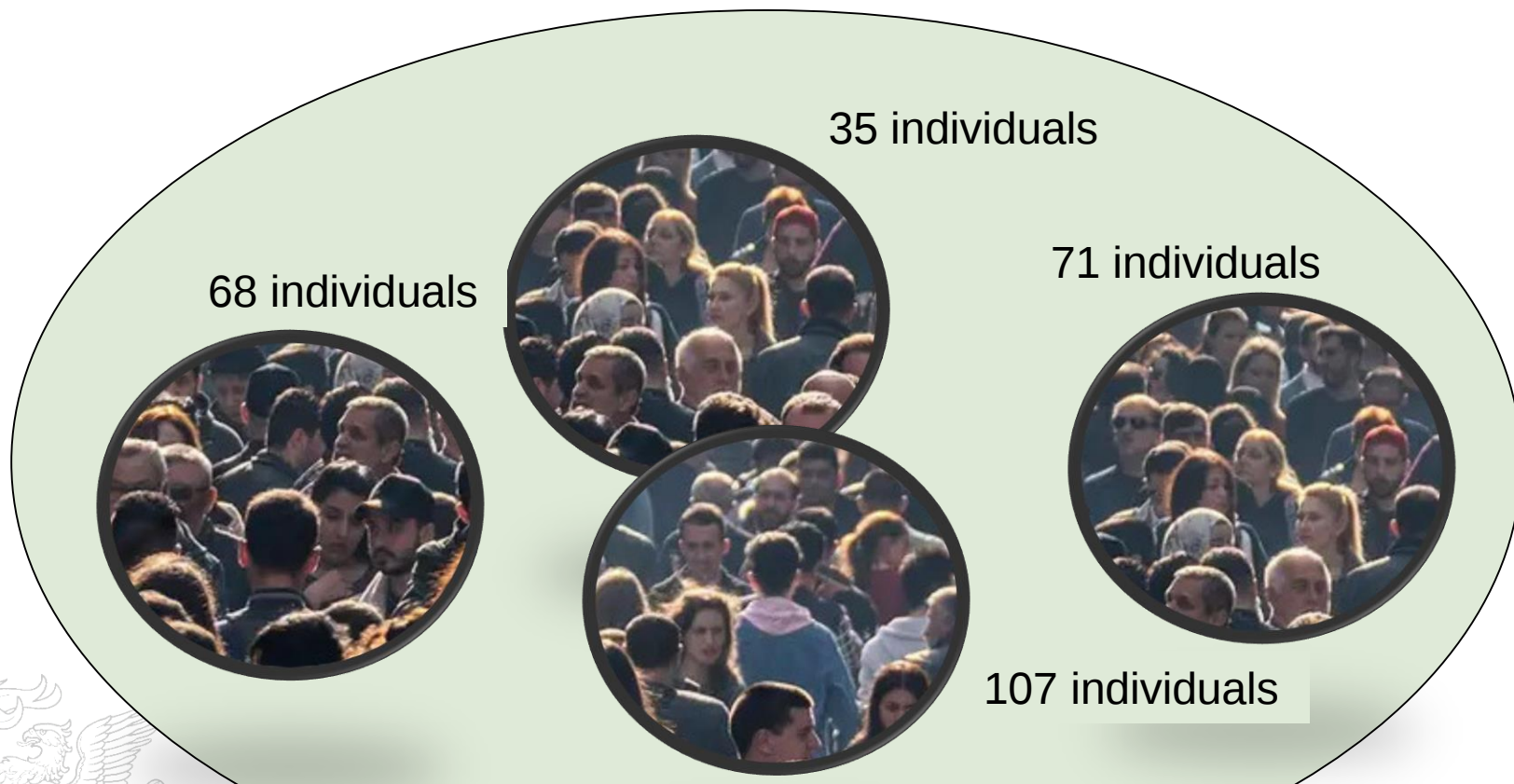
If the entropy is large enough, that means it is greater than $\log_2 N - 1$, then supposing all sets contain one or two undistinguishable individuals we can get an estimation on the number of sets with one person (singletons). This can be greater, if the sets contain three, four, ... undistinguishable individuals.

$$n_{\text{singletons}} \geq (E(D) - (\log_2(N) - 1)) * N$$



- Alexin, Z.: [Entropy based approach to personal data](#), In Proceedings of the International Conference on Privacy-friendly and Trustworthy Technology for Society — COST Action CA19121, pp. 18-31. DOI: [10.5281/zenodo.6813377](https://doi.org/10.5281/zenodo.6813377), 28th June 2022. Zagreb, Croatia (2022.)

How to compute the entropy?



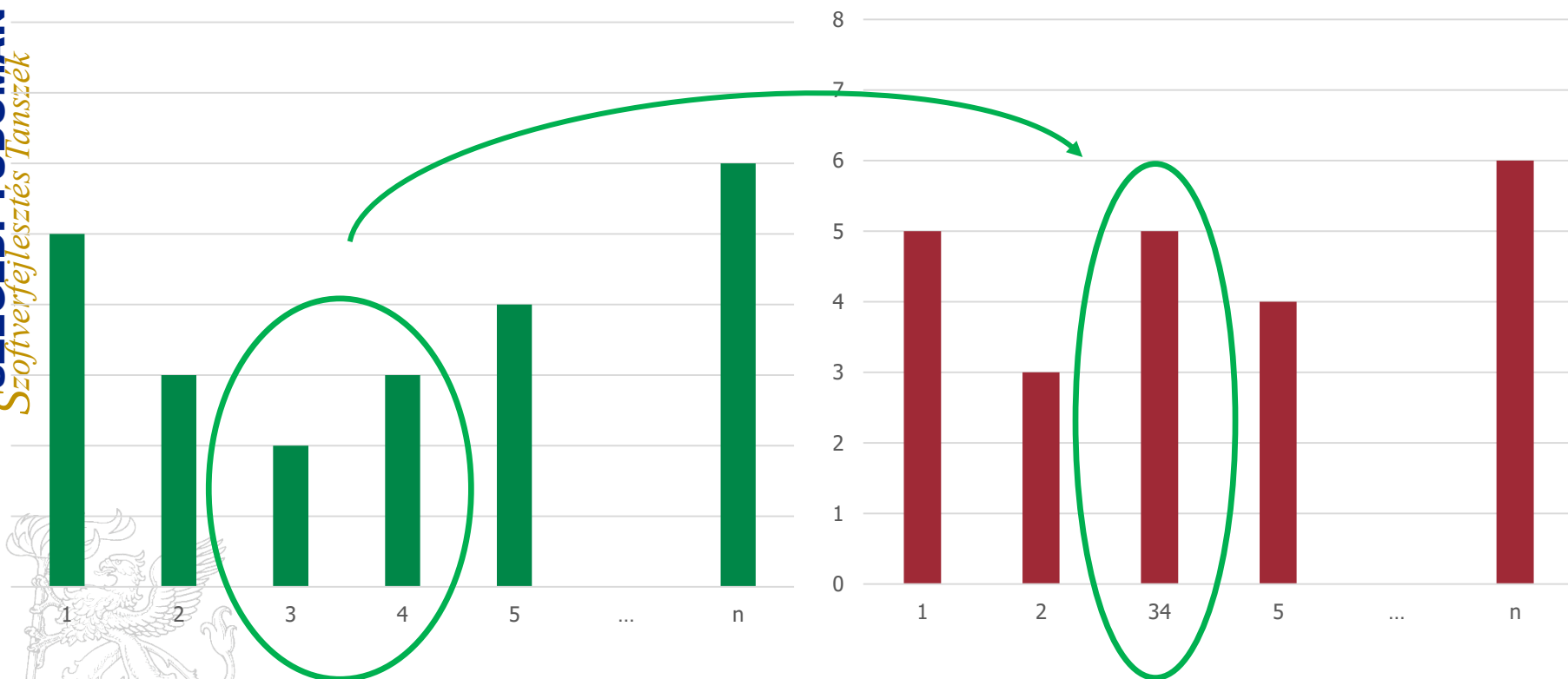
$$N = 35 + 68 + 107 + 71 = 281 \quad \log_2(281) = 8,1344$$

$$E(I) = \frac{68}{281} \log_2\left(\frac{281}{68}\right) + \frac{35}{281} \log_2\left(\frac{281}{35}\right) + \frac{107}{281} \log_2\left(\frac{281}{107}\right) + \frac{71}{281} \log_2\left(\frac{281}{71}\right) = 1,90153$$

Entropy based approach

Undistinguishable individuals (D_1)

Undistinguishable individuals (D_2)



$$E(D_1) > E(D_2)$$

Acknowledgement

The author would like to express his thanks to the:

- Central Office for Administrative and Electronic Public Services (in Hungarian: Közigazgatási és Elektronikus Közszolgáltatások Központi Hivatala, KEKKH) for the provision of a research dataset extracted from the national population registry.
- National Healthcare Service Centre (in Hungarian: Állami Egészségügyi Ellátó Központ, ÁEEK) for the out-patient research dataset extracted from the IMDW (Itemized Medical Data Warehouse).



The entropy of ZIP code (Hungary)

ZIP code	City	Population	Bits ($\log_2(N/k)$)	Entropy ($-(k/N) \cdot \log_2(k/N)$)
1011	Budapest I.	3286	11.5719	0.003800
1012	Budapest I.	4446	11.1357	0.004948
1013	Budapest I.	3404	11.5210	0.003920
...				
9982	Apátistvánfalva	589	14.0519	0.000827
9983	Szakonyfalu	769	13.6672	0.001050
9985	Felsőszölnök	589	14.0519	0.000827
Altogether:		10,004,090		10.303428

- The result of the computation is that a ZIP code contains 10.3 bits information in Hungary. If we choose a random citizen and get to know the ZIP code of their resident address, then we get 10.3 bits on average. This corresponds to 7916-anonymity. Not risky.
- To uniquely identify everybody in Hungary, where the population is 10 004 090 we need $\log_2(N) = 23.254$ bits.

Birthdate and ZIP code

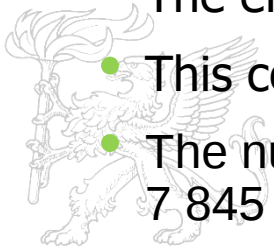
Birthdate x ZIP code	Population	Bits ($\log_2(N/k)$)	Entropy ($-(k/N) \cdot \log_2(k/N)$)
(1894.12.31., 3744)	1	23.254	2.324458e-6
...			
(1975.08.04., 9400)	4	21.254	8.498159e-6
(1975.08.04., 9407)	1	23.254	2.324458e-6
(1975.08.04., 9473)	1	23.254	2.324458e-6
(1975.08.04., 9523)	1	23.254	2.324458e-6
(1975.08.04., 9600)	1	23.254	2.324458e-6
(1975.08.04., 9700)	6	20.669	1.239640e-5
...			
Altogether:	10,004,090		22.79385

- The entropy is 22.7985 bits. This corresponds to 1.37-anonymity. Such a dataset is very risky from the point of view of re-identification.
- The forecasted number of singletons were 54% of the population, but in fact it was greater: 6 635 838 individuals (66,33%).

Birthdate + ZIP + sex

Birthdate x ZIP x sex	Population	Bits	Entropy
(1894.12.31., 3744, M)	1	23.254	2.324458e-6
...			
(1954.04.14., 6000, M)	1	23.254	2.324458e-6
(1954.04.14., 6041, M)	1	23.254	2.324458e-6
(1954.04.14., 6066, M)	2	22.254	4.448998e-6
(1954.04.14., 6070, F)	1	23.254	2.324458e-6
(1954.04.14., 6097, F)	1	23.254	2.324458e-6
(1954.04.14., 6097, M)	1	23.254	2.324458e-6
...			
Sum:	10,004,090		22.992721

- The entropy is 22.992721 bits.
- This corresponds to 1.19-anonymity. Very risky.
- The number of singletons (uniquely identifiable individuals) were: 7 845 850 individuals (78,43% of the population).



Conclusion

- Entropy is a single, floating-point number, that well characterise the information content of a pseudonymised database.
 - Perhaps, the use of $\hat{k}$ value gives a more understandable result.
 - Dates convey lots of information, they shall be avoided, use only the year (c.f. HIPAA law, privacy rule).
 - Before releasing a dataset for example, for research purposes (like AI), custodians shall be ascertained that the dataset contains small amount of information about data subjects. C.f. HIPAA law, expert rule.
-
- What is the entropy of visit sequences?
 $S = \{ \text{person}, (\text{date}_1, \text{event}_1), (\text{date}_2, \text{event}_2), (\text{date}_3, \text{event}_3), \dots \}$
 - Can somebody be identified by two visits?

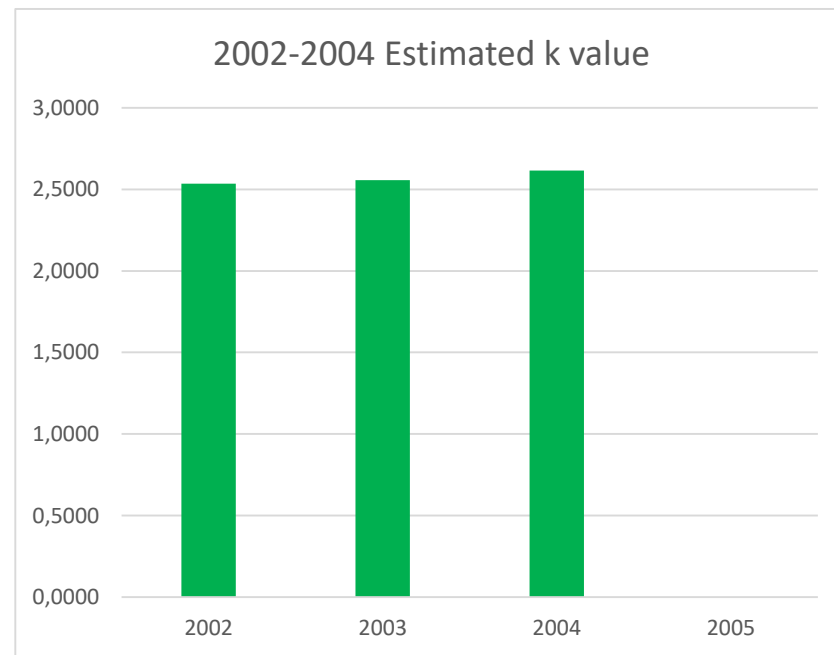




2002-2004 distribution of visit pairs

Maximal entropy	34,250899
Computed entropy	32,890720
Number of visit pairs	20 443 157 017
Estimated k value	2,5671694

1	9 901 739 166	16,58958386
2	1 485 377 009	4,831946533
3	509 305 701	2,441444976
4	237 195 716	1,496789378
5	131 251 931	1,024973559
6	81 479 517	0,757258345
7	54 807 826	0,590097998
8	38 983 655	0,476746038
9	28 844 276	0,394683351
10	22 025 099	0,33322331
11	17 225 208	0,285390545
12	13 759 479	0,247680322
13	11 186 074	0,217315558
14	9 226 291	0,192354534
15	7 700 918	0,171458391
16	6 508 938	0,154106316
17	5 539 587	0,138950194
18	4 746 775	0,1257231
19	4 095 906	0,114214143
20	3 551 655	0,103993115
...		



Thank you for your attention!





Vizitpárok entrópiája

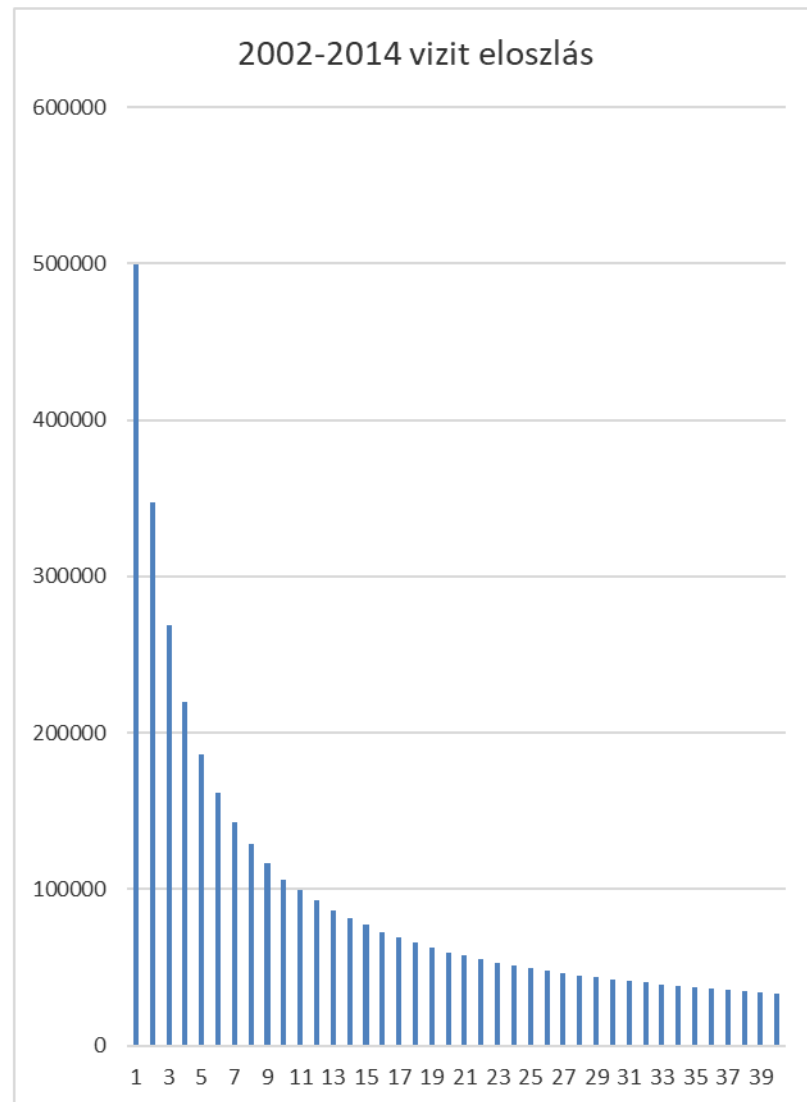
Mi az egyéni élettörténet?

- Egy személy fontosabb életeseményeinek listája
- Dátumokat és az események minimális jellemzőit tartalmazza (amelyeket talán maga az érintett nyilvánosságra hozott, vagy más módon tudható róla)
- Kettőnél több eseményt tartalmaz



Vizitek a 2002-2014 években

Vizitek száma			709 645 121
Entrópia			20,7758268
Estimated k			395,2702822
	1	499735	0,020705377
	2	347748	0,027836224
	3	268784	0,031608398
	4	219512	0,033905299
	5	185847	0,03546031
	6	161860	0,036700233
	7	143215	0,037570594
	8	128803	0,03833718
	9	116836	0,038870423
	10	106175	0,039021012
	11	99516	0,040018987
	12	92739	0,040487189
	13	86275	0,04062145
	14	81879	0,04134446
	15	77481	0,041755245
	16	72929	0,041769172
	17	69138	0,041927934
	18	65653	0,042019204
	19	62550	0,04212666
	20	59635	0,042152935

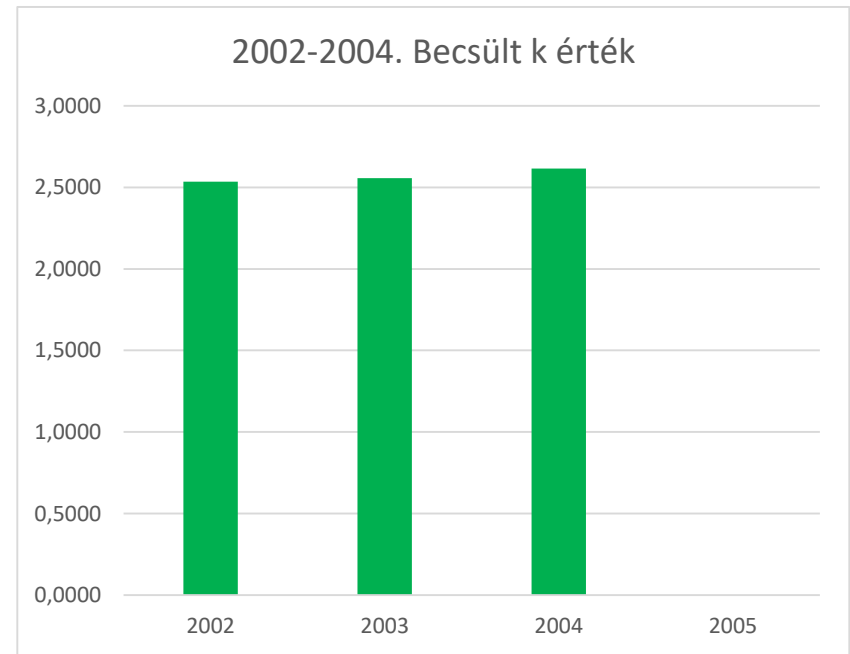




2002-2004 évek adatai párokra

Maximális entrópia	34,250899
Számított entrópia	32,890720
Vizitpárok száma	20 443 157 017
Becsült k érték	2,5671694

1	9 901 739 166	16,58958386
2	1 485 377 009	4,831946533
3	509 305 701	2,441444976
4	237 195 716	1,496789378
5	131 251 931	1,024973559
6	81 479 517	0,757258345
7	54 807 826	0,590097998
8	38 983 655	0,476746038
9	28 844 276	0,394683351
10	22 025 099	0,33322331
11	17 225 208	0,285390545
12	13 759 479	0,247680322
13	11 186 074	0,217315558
14	9 226 291	0,192354534
15	7 700 918	0,171458391
16	6 508 938	0,154106316
17	5 539 587	0,138950194
18	4 746 775	0,1257231
19	4 095 906	0,114214143
20	3 551 655	0,103993115





Első eredmények a tripletekre

		2002. Január 1-13	
Population		2 008 486 075	
Entrópia		30,7335334	
Estimated k		1,1250023	
log2(N)		30,9034613	
	1	1 838 400 257	28,28644516
	2	33 385 925	0,994136558
	3	8 814 094	0,385986253
	4	3 871 100	0,222830898
	5	2 140 234	0,152281786
	6	1 311 645	0,110960642
	7	849 430	0,083176942
	8	588 576	0,065415669
	9	448 080	0,055684522
	10	373 843	0,051337987
	11	261 470	0,039300096
	12	181 287	0,029589383
	13	148 130	0,026081628
	14	118 008	0,022288331
	15	97 844	0,01972719
	16	86 566	0,018552681
	17	90 917	0,020635698
	18	89 102	0,021347526
	19	74 512	0,018788771
	20	64 113	0,016970213

		2002. január 1-8. Budapest	
Population		1 540 600 325	
Entrópia		30,2599110	
Estimated k		1,1982546	
log2(N)		30,5208455	
	1	1364351062	27,02916992
	2	31294465	1,199323473
	3	8080114	0,455287259
	4	3713181	0,2749657
	5	2033292	0,18608536
	6	1263415	0,137457898
	7	848838	0,106886806
	8	621567	0,088827967
	9	453368	0,072439482
	10	368935	0,065134561
	11	291025	0,056231994
	12	220741	0,046313274
	13	182690	0,041345999
	14	152308	0,036973572
	15	126706	0,032832796
	16	107144	0,029511088
	17	100133	0,029207132
	18	90767	0,02794514
	19	77753	0,025193533
	20	70466	0,023966409

Legal cases

■ Before the Constitutional Court

- Az OEP adatmegőrzési idejének megállapításáért (Ab. 1034/E/2005)
- Nem támogatott vények adatainak összegyűjtése ellen (Ab. 29/2009. III. 21.)
- Az OEP korlátlan adatigénylésének lehetősége ellen (Ab. 53/B/2009)
- Az eü. adatok összekapcsolhatóságának korlátozása ellen (Ab. 728/B/2004)
- A bírói közreműködés nélkül bekövetkezett sérelemre hivatkozva indítani Ab. eljárást másodlagos adatvédelmi törvény ellen (Ab. 3110/2013. VI. 4.)
- A szükségtelen kényszer üzemorvosi alkalmassági vizsgálat ellen, Ab. IV/2718/2012
- A beavatkozás nélkül történő kényszer orvosi kutatás ellen (EJEB), 129/B/2008
- Az OEP adatmegőrzési idejének csökkentéséért, a vények adattartalmának csökkentéséért, Ab. IV/2702-14/12
- Az kórházak, szakrendelők adatmegőrzési idejének csökkentéséért, Ab. 67/2011.
- Az orvosi titoktartás állami elismeréséért (Ab. IV/2689/2012)
- A jogorvoslat nélkül történő adatkezelés és az EESZT ellen (Ab. 753/2018) folyamatban
- A foglalkozás-egészségügyi orvos nem követelheti a TAJ-t, (Ab. 3096/2021.)



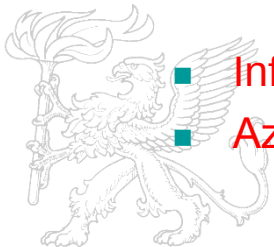
Legal cases

■ Before the civil courts:

- Az eü. adatokról az érintett kaphasson utólagos tájékoztatást, az adatvédelmi szabályzatok nyilvánosságáért
- Az engedélyezett orvosi kutatások nyilvánosságáért
- Az OEP adatmegőrzési idejének visszamenőleges növelése ellen
- Az eü. adatok törlésének lehetőségéért, a kötelező előzetes tájékoztatásért
- A TEA személyes adatokat tartalmaz-e?

■ Ombudsman on Fundamental Rights

- A foglalkozás-egészségügyi beutalón nem lehet TAJ azonosító



- Infotv.-ből hiányoznak a jogalapok (2012.)
- Az EESzT kötelező adatkezelése miatt

Kézikönyvek

- ▶ Information Commissioner's Office: Anonymisation: managing data protection risk, Code of practice,
<https://ico.org.uk/media/1061/anonymisation-code.pdf>
- ▶ Simson L. Garfinkel: De-Identification of Personal Information, NISTIR 8053, U.S. Department of Commerce, National Institute of Standards and Technology,
<https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf>
- ▶ Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques, WP216, Adopted on 10 April 2014,
https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf



Anonimitást növelő transzformációk

- A kvázi-azonosítók értékkészlete nagy, ezért minden rekord különböző lehet.
- A cél az, hogy a rekordok megkülönböztethetetlenek legyenek a kvázi-azonosítók alapján. Azaz, több rekordban is azonosak legyenek a kvázi-azonosítók.
 - ▶ Csökkenteni kell kvázi-azonosítók dimenziójának a számát – a szükségtelen oszlopokat el kell távolítani.
 - ▶ Csökkenteni kell a kvázi-azonosítók értékkészletét: alsó levágás, felső levágás, csoportosítás (kvantálás), torzítás módszerével.
 - ▶ Ha vannak teljesen egyedi rekordok, akkor azt a sort törölni kell.
 - ▶ Úgy nevezik ezeket, hogy: generalization és suppression.
 - ▶ Születési dátum helyett születési év: általánosítás (generalization).
 - ▶ Cella vagy rekord törlése: suppression (elfedés) egy módja.



Példa

Azonosító	Kor	Nem	Vizit napja	Diagnózis		
1001011	8	F	1992/05/03	B3730		
1001012	Azonosító	Kor	Nem	Vizit napja	Diagnózis	
1001013						
1001014	1001011	0-9	F	1992/05/03	B3730	
1001015	1001012	10-19	Azonosító	Kor	Nem	Vizit napja
1001016	1001013	10-19	1001011	Kisebb20	F	1992/05/03
	1001014	20-29	1001012	Kisebb20	F	1992/05/03
	1001015	20-29	1001013	Kisebb20	F	1992/05/03
	1001016	20-29	1001014	20-29	M	1991/10/08
			1001015	20-29	M	1991/10/08
			1001016	20-29	M	1991/10/08

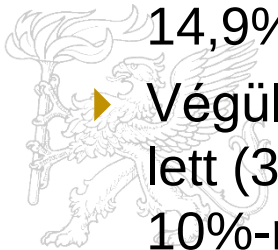
Csoportképzés

Alsó levágás



Kanadai kísérlet valódi egészségügyi adatokkal

- ▶ Egy gyermekkórházból származó vény adatbázist vizsgáltak meg
- ▶ *Samarati* és *El Emam Khaled* végezte el a „μ_argus” nevű szoftverrel.
- ▶ Ez automatikusan alkalmaz általánosítási és törlési műveleteket az adatállományokon.
- ▶ 94 100 rekordot dolgoztak fel, ebben 10 364 vizit szerepelt és 6970 különböző gyermek (páciens).
- ▶ Az általánosítási lépések után még el kellett távolítaniuk a rekordok 14,9%-át.
- ▶ Végül az újra-azonosítási kockázat a rekordok 95%-a esetében 33% lett (3-anonimitás). A rekordok 80%-a esetében pedig lecsökkent 10%-ra (10-anonimitás).

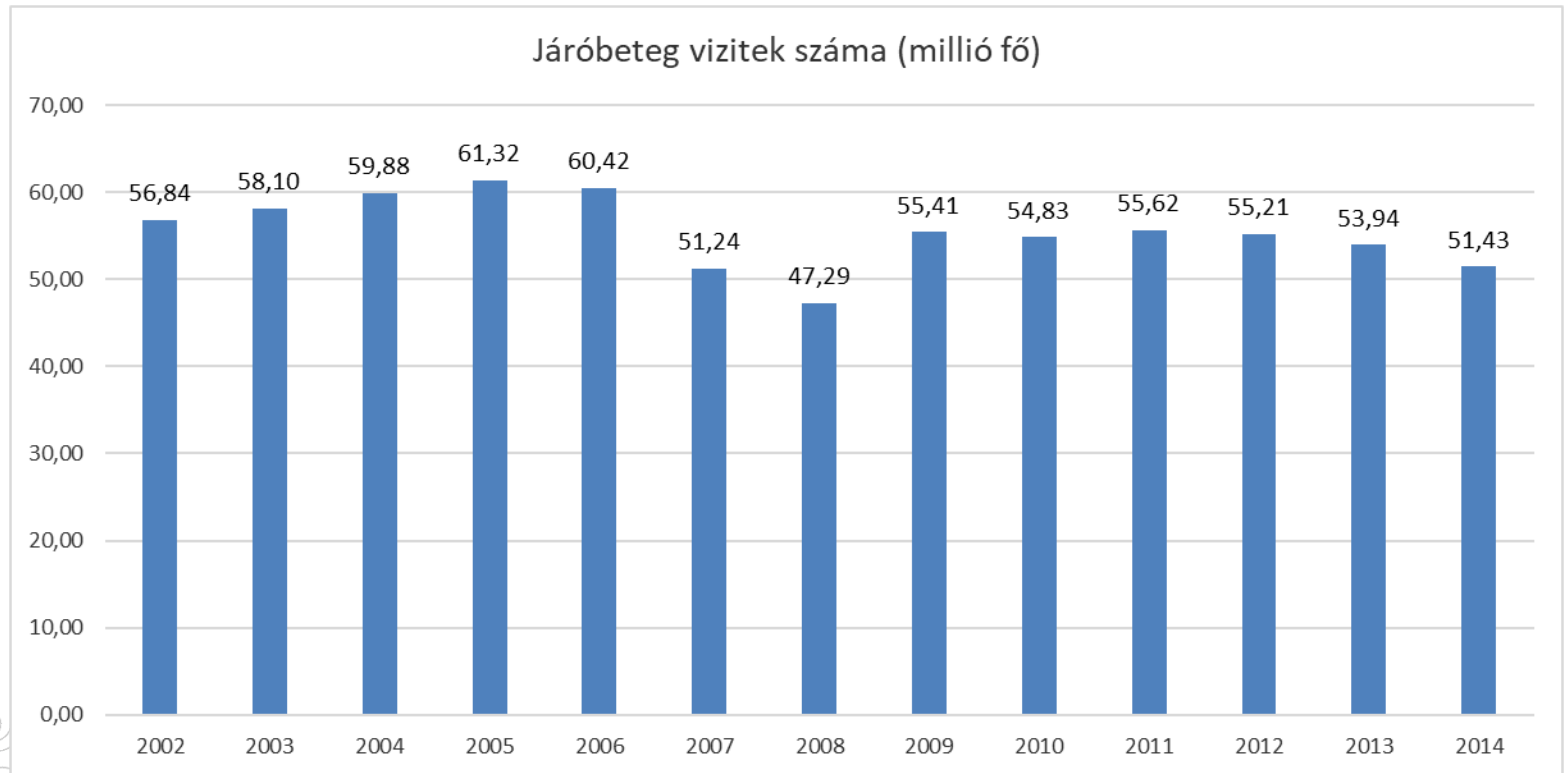


Újra-azonosítási esetek

- Az idősorok két-három-négy esemény dátumának ismerete [és még valamilyen csekély információ] elegendő a személyazonosításhoz.
 - Narayanan, Arvind and Shmatikov Vitaly: Robust De-anonymization of Large Sparse Datasets. IEEE Symposium on Security and Privacy 2008: 111-125
- Sajtóban megjelent balesetekről szóló hírek alapján lehetséges a személyazonosítás
 - Sweeney L. Matching Known Patients to Health Records in Washington State Data. Harvard University. Data Privacy Lab. 1089-1. June 2013.
- Öt-hét laborvizsgálat eredménye egyértelműen azonosíthatja a személyeket (az alapvérvkép hét teszteredménye 98,9%-ban egyértelműen azonosította a betegeket)
 - Atreya, Ravi V, Joshua C Smith, et al.: “Reducing patient re-identification risk for laboratory results within research datasets,” J Am Med Inform Assoc 2013;20:95–101. doi:10.1136/amiajnl-2012-001026.
- A USA HIPAA privacy rule próbája a magyar népességnyilvántartási adatokon
 - Alexin, Z.: Does fair anonymization exist?, International Review of Law, Computers and Technology, Vol. 28 No. 1: pp. 21-44, DOI: 10.1080/13600869.2013.869909, Taylor & Francis Publishing (2014.)



Megjelenések



Összesen: 721 525 095 megjelenés, 12 200 284 különböző személy

Célkitűzés

- Az orvostudományi kutatások egy típusa a betegek életeseményeit vizsgálja.
- Az EESZT adatai esetében különösen vonzó lehet egy ilyen kutatási tevékenység, mert 11,5 millió személy teljes egészségügyi élettörténete megszerezhető belőle.
- A szerző szeretné bebizonyítani, hogy így megszerzett álnevesített adatok sosem lesznek anonim adatok statisztikai szempontból, vagyis hogy néhány életesemény elegendő az érintettek azonosítására.
- Tekintettel arra, hogy éles, névvel ellátott EESZT adatok már kontrollálatlanul kijutottak a védett környezetből így az azonosítási kockázat valós, és közvetlenül fenyegeti a polgárokat.



Az adatok

- Valódi járóbeteg adatállomány, 721 millió eseménnyel (ÁEEK)
- Az eseményt a
 - egy betűkből és számokból álló hash-kód (azonos személy esetén azonos kód)
 - dátum (éééé.hh.nn.),
 - BNO-kód első betűje (számmal kódolva).
 - egészségügyi intézmény irányítószámának első két számjegye,

"P7U4CPB6FG2U4R2"; "2005.04.08."; "31"; "10"

"M302A3UUE2MH780"; "2005.04.13."; "8"; "10"

"0MVIIPDDLI41AD3"; "2005.04.11."; "43"; "59"

"VHBU2C8CNCM3T02"; "2005.04.08."; "39"; "59"

"HSUO9TM7HURF683"; "2005.04.11."; "43"; "59"

"1EE5DM22I0V17P3"; "2005.04.13."; "35"; "86"

"9KSL89QEDIS58P3"; "2005.04.06."; "46"; "86"

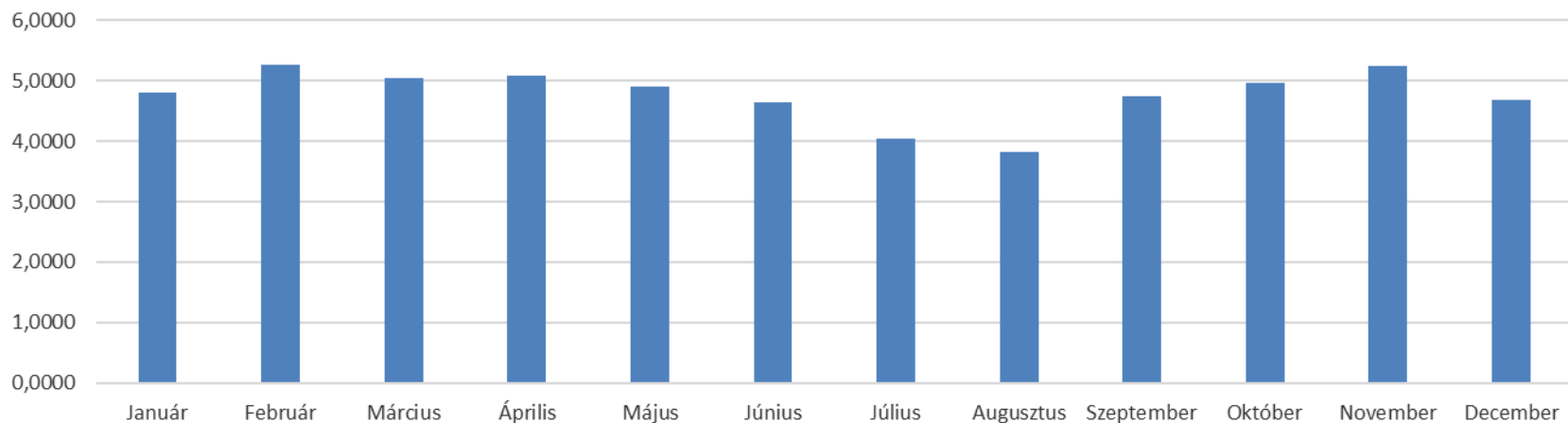
Módszer

- Számítógépes programmal előállítani az összes különböző.
 - vizitet: ~709 645 121 (db) [volt 8 624 389 duplikált vizit].
 - vizitpárt: ~53 718 325 574 (db)
 - vizit-hármast: ~4 742 137 812 244 (db)
- Megvizsgálni azt, hogy az egyes párokhoz / tripletékhez hány személy rendelhető hozzá, milyen arányban fordul elő egyértelmű azonosítás?
- Azokat a betegeket, akik több mint 1000-szer jelentek meg, kizártam.
- Ennyi adat előállításához nincs elegendő memória és tárhely (128 GB RAM, és 2TB SSD), ezért a párokat és a tripletéket növekvő sorrendbe rendeztem. Az első dátum szerint kisebb szeleteket generáltam az összes lehetséges kombinációból, pl. olyan párokat, amelyekben az első dátum 2002-beli (259 GB, 7,142 milliárd sor), 2003-beli (247 GB, 6,787 milliárd sor), 2004-beli (237 GB, 6,512 milliárd sor), 2005-beli (222 GB).



Budapest és a vidék

2002. Becsült k érték



2002. Becsült k érték

