TALKING ABOUT SMART HEALTHY AGE-FRIENDLY ENVIRONMENTS

International Conference LEARNING AND IMPLEMENTING SOCIAL INNOVATION

Zoltán Alexin: The entropy of personal data

Are visit-date pairs identifying?

Organization Committee

Escola Superior de Enfermagem de Coimbra

Funded by the European Union

Co-funded by the European Union

Co-funded by Programme for Employment and Social Innovation (EaSI) of the European Union

Content

- Motivation
- Privacy risks when applying the new technologies
- Application of the entropy formula on datasets
- Entropy values of the Population Registry data
- Entropy values of the visit pairs
- Conclusions

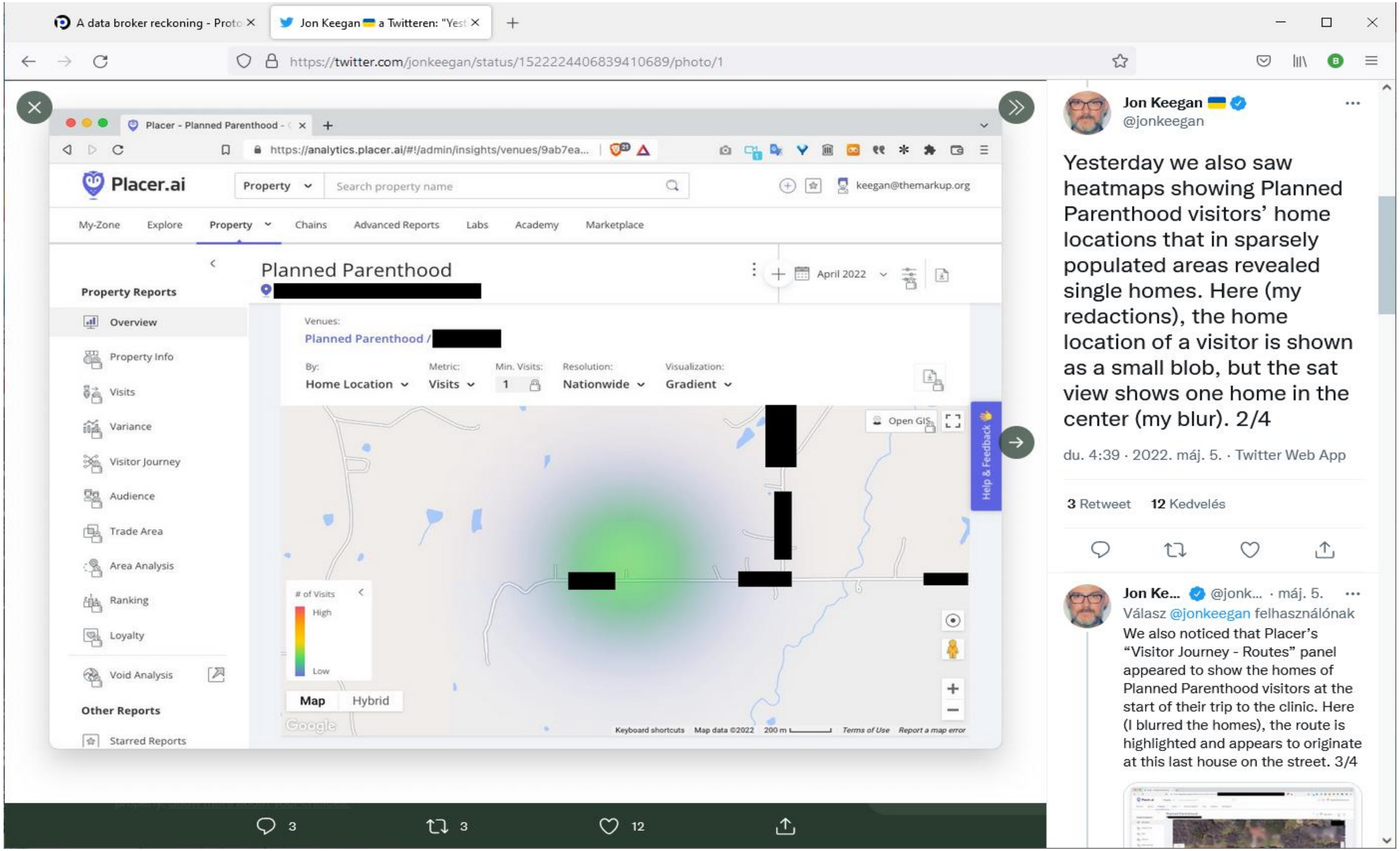
Acknowledgements

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 story behind

- Itemized Medical Data Warehouse, IMDW, that stores accumulated health insurance accounting data of all Hungarian citizens beginning from 1998 forever.
- National Health Insurance Fund (NEAK) replaces Social Security Identifiers by a pseudonym. The interchange table is maintained by and kept endlessly by the insurance fund.
- All data items contain the date of birth, ZIP code of resident address and gender, dates, physicians, institutes, medicines
- I filed the controller of the IMDW before the Constitutional Court in 2006 without success, case no: 937/B/2006.
- The law declares that the dataset is anonymous (Decree 76 of 2004 on collection and processing of medical sector data not suitable for personal identification)
- I turned to the civil court later and asked them to declare that the data is personal (not anonymous). Finally the Supreme Court declined the case.



Planned Parenthood

[redacted]

Venues:

Planned Parenthood / [redacted]

By: Home Location

Metric: Visits

Min. Visits: 1

Resolution: Nationwide

Visualization: Gradient

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

Jon Keegan 🇺🇦
@jonkeegan

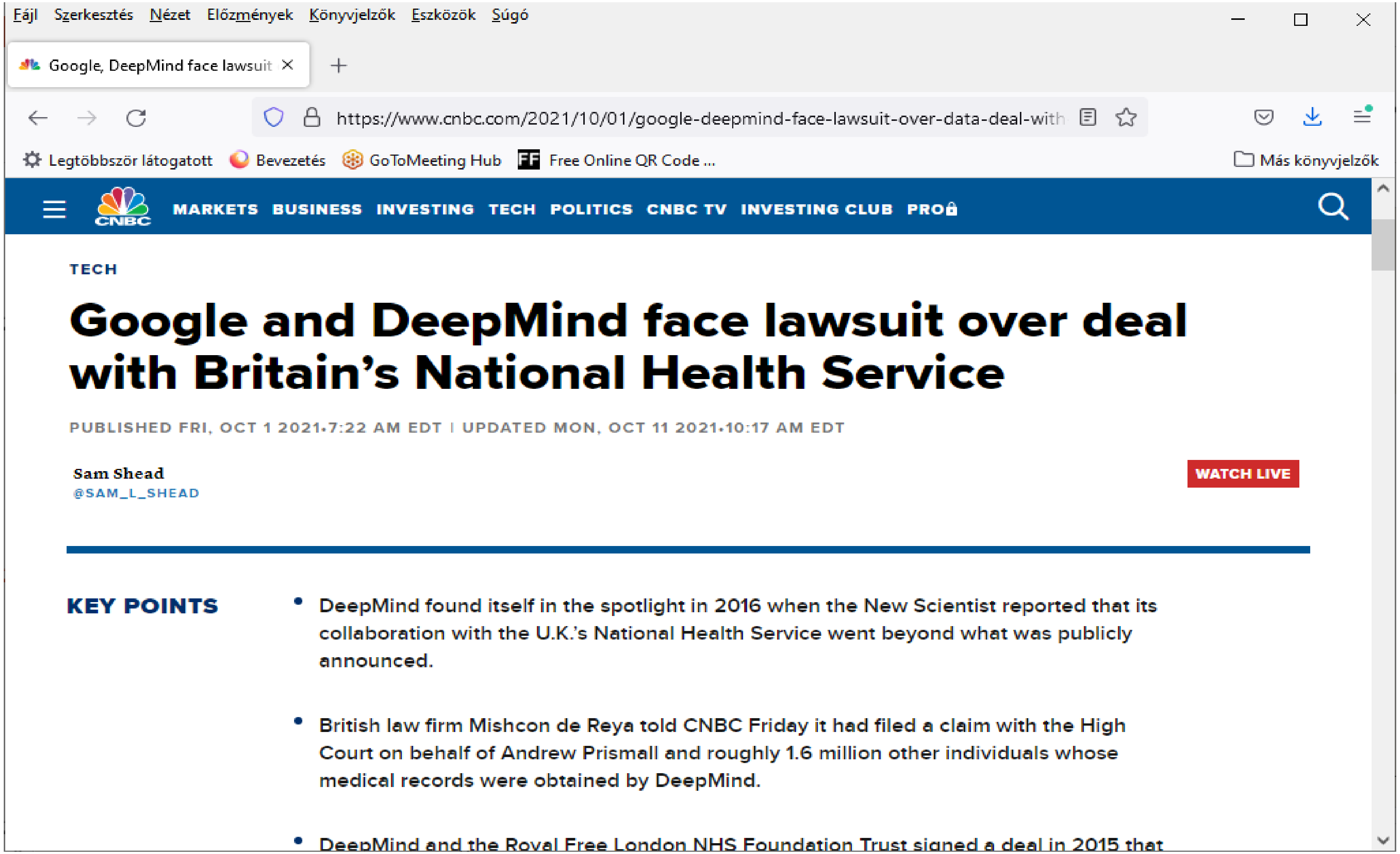
Yesterday we also saw heatmaps showing Planned Parenthood visitors' home locations that in sparsely populated areas revealed single homes. Here (my redactions), the home location of a visitor is shown as a small blob, but the sat view shows one home in the center (my blur). 2/4

du. 4:39 · 2022. máj. 5. · Twitter Web App

3 Retweet 12 Kedvelés

Jon Ke... @jonk... · máj. 5.
Válasz @jonkeegan felhasználónak

We also noticed that Placer's "Visitor Journey - Routes" panel appeared to show the homes of Planned Parenthood visitors at the start of their trip to the clinic. Here (I blurred the homes), the route is highlighted and appears to originate at this last house on the street. 3/4


Fájl Szerkesztés Nézet Előzmények Könyvjelzők Eszközök Súgó
Google, DeepMind face lawsuit
https://www.cnbc.com/2021/10/01/google-deepmind-face-lawsuit-over-data-deal-with
Legtöbbször látogatott
Bevezetés
GoToMeeting Hub
Free Online QR Code ...
Más könyvjelzők
CNBC
MARKETS
BUSINESS
INVESTING
TECH
POLITICS
CNBC TV
INVESTING CLUB
PRO
WATCH LIVE

TECH

Google and DeepMind face lawsuit over deal with Britain’s National Health Service

PUBLISHED FRI, OCT 1 2021•7:22 AM EDT | UPDATED MON, OCT 11 2021•10:17 AM EDT

Sam Shead
@SAM_L_SHEAD

KEY POINTS

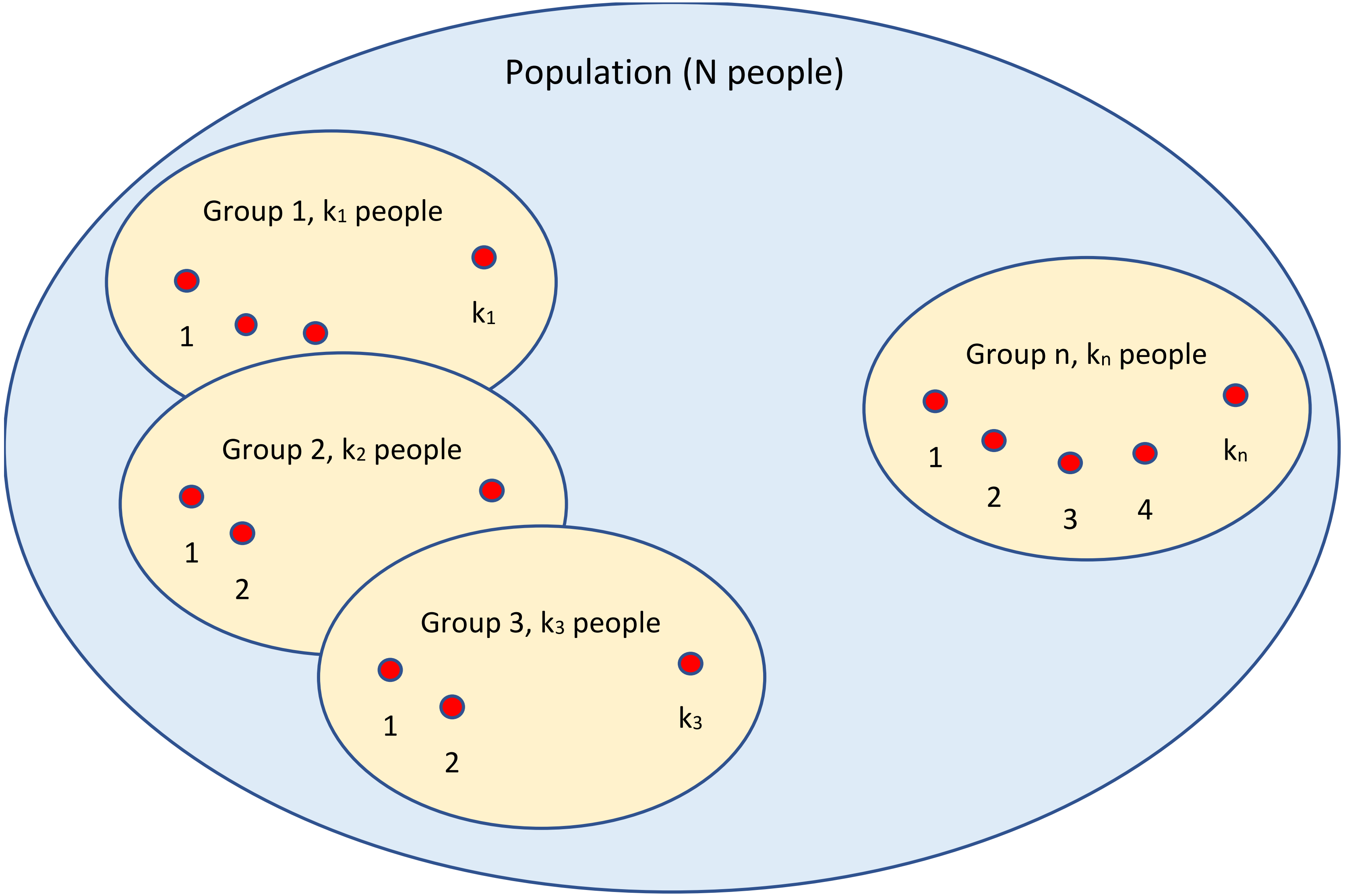
- DeepMind found itself in the spotlight in 2016 when the New Scientist reported that its collaboration with the U.K.’s National Health Service went beyond what was publicly announced.
- British law firm Mishcon de Reya told CNBC Friday it had filed a claim with the High Court on behalf of Andrew Prismall and roughly 1.6 million other individuals whose medical records were obtained by DeepMind.
- DeepMind and the Royal Free London NHS Foundation Trust signed a deal in 2015 that

Quasi-identifiers

- Quasi identifiers: data values that doesn’t identify an individual on its own but can become identifying in combination with other quasi identifiers.
- Quasi identifiers are not direct identifiers. Instead, they are identifiers such as an area code or zip code or 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.
- 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 the de-identified dataset.
- A record then could be $r(q_1, q_2, q_3, q_4, q_5, \ldots, q_n, d_1, d_2, \ldots, d_m)$.
- An adversary can have $a(q_2, q_3, q_4, q_5, \text{name})$.

- The question is: what could be a quasi-identifiers? Date of birth, zip, job, gender, qualifications, schools, workplace, illness, medical operation.

Partitioning people by quasi-identifiers



What is entropy?

$$\underbrace{0110011101100110011}\ \underbrace{\boxed{1010}}$$

We know the group where the target people is.
$\log_2(N) - \log_2(k) = \log_2(N/k)$

k indistinguishable people $\log_2(k)$

- Entropy is a weighed sum (expected value, average) amount of bits we know about a random individual in the database.

$$E(D) = \frac{1}{N}\sum_{people}\log_2\frac{N}{\#group} = \sum_{people} -\frac{1}{N}\log_2\frac{\#group}{N}$$

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

Features of the entropy

- The *k*-anonymity has close connection to the entropy. The following inequality is held: if a dataset D is *k*-anonymous then its entropy $E(\mathrm{D}) < -\log_2(k/N)$, where *N* is the number of individuals (population) in the dataset. This follows from the fact, that in this case all groups have at least *k* members.

- If the entropy of a dataset D is E(D), then we can compute an estimated $\hat{k}$ value from the above formula, which is characteristic to the level of anonymity of the dataset.
  - $\hat{k} = \frac{N}{2^{E(D)}}$

- If the entropy a dataset D is E(D) and it is greater than $\log_2(N) - 1)$, in other words $\hat{k} < 2$ then it guarantees certain number of singletons (uniquely identifiable individuals). If several groups have more members than 2, then the number of singletons will be higher. When $\hat{k} \geq 2$ there still can be singletons, but their existence is not guaranteed.
  - $n_{singletons} \geq (\mathrm{E(D)} - (\log_2(N) - 1)) * N$

Entropy of ZIP codes

ZIP code	Settlement	Population	Bits ($\log_2(N/k)$)	Entropy (-(k/N) *log2(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
Sum:		10,004,090		10.303428

- The result of the computation showed that the entropy of ZIP codes is 10.3 bits. This means that statistically, for a random citizen the expected amount of information in his/her ZIP code is 10.3 bits of average. It corresponds to 7916-anonymity.
- To identify each people in Hungary (the population is 10,004,090) we need $\log_2(N) = 23.254$ bits.

Entropy of birthdate + ZIP codes

Birthdate x ZIP code	Population	Bits ($\log_2(N/k)$)	Entropy (-(k/N) *log2(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
…			
Sum:	10,004,090		22.79385

- The entropy is 22.7985 bits. It corresponds to 1.37-anonymity. This database poses substantial risk for re-identification.
- The predicted ratio of singletons is greater than 54% of the population, in fact it was 6,635,838 individuals.

Out-patient visit database

- It contains all reported out-patient visits to the National Health Insurance Fund, altogether 721,633,881 visits. Mainly between 1st Jan 2002 and 31st Dec 2014.
- The data items are:
  - pseudonym of the patient suitable to join all visits of the same patient
  - date of visit
  - first letter of the ICD-10 code, coded by a decimal number.
  - first two digits of the ZIP code of the medical institution (polyclinic).
  - `"G58FG689DDQ3GE1";"2005.01.27.";"3";"67„`

- Narayanan, A. and Shmatikov, V. (2008): *Robust de-anonymization of large sparse datasets.* In 2008 IEEE Symposium on Security and Privacy (sp 2008) (pp. 111-125). IEEE. (joining the Netflix research database with the IMBD.com)

Entropy of visit pairs

Date	Entropy	Predicted k	Singletons
…			
8th January 2005 (Sat)	19.725	1.247	902,032 (83.45%)
9th January 2005 (Sun)	19.158	1.147	596,462 (88.86%)
10th January 2005 (Mon)	23.254	2.889	12,809,964 (44.32%)
11th January 2005 (Tue)	23.233	2.865	12,561,551 (44.46%)
12th January 2005 (Wed)	23.214	2.750	12,234,144 (45.71%)
January 2005	27.556	2.734	250,588,888 (46,47%)

- The most active patients (who visited clinics 1000 times or more) were excluded (1968 patients (0,016% of all patients, they had made 3,348,081 visits which is 0,46% of the 721 million visits).
- Due, to the size of data, first, only those pairs were generated, where the earlier date fell in January 2005. There were 3,863,139,808 (88.8 GB) of such pairs.
- Considering the visit pairs, the group sizes range from 1 to 3233. The biggest group contained 3233 patients who visited some polyclinics in Budapest, on 24th and 26th January (Monday, Wednesday). There were 539,225,013 distinct visit pairs in the generated dataset, among them there were 250,588,888 singletons.

Alexin, Z. (2023): *What makes the data personal?*, to be submitted to Digital Society (Springer), Special Issue on Privacy-friendly and trustworthy technology for society

Organization Committee

Escola Superior de Enfermagem de Coimbra

FCT Fundação para a Ciência e a Tecnologia

COST EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the European Union

Co-funded by the European Union

Co-funded by Programme for Employment and Social Innovation (EaSI) of the European Union