

Személyes adatokat tartalmazó adatállományok entrópiája

Dr. Alexin Zoltán, PhD.
egyetemi adjunktus
SZTE, TTIK, Szoftverfejlesztés Tanszék
e-mail: alexin@inf.u-szeged.hu
<http://www.inf.u-szeged.hu/~alexin>

Tartalom

- Motiváció
- Az új technológiák gondatlan alkalmazásainak következményei
- Az entrópia kiszámítása
- Az entrópia számítás alkalmazása a népesség nyilvántartási adatokra
- Az entrópia számítás alkalmazása az ÁEEK járóbeteg ellátási adatokra

Köszönetnyilvánítás

A szerző köszönetét fejezi ki:

- A Közigazgatási és Elektronikus Közszolgáltatások Központi Hivatalának (KEKKH) az állami népszámlálási adatokból ingyenesen biztosított kutatási adatokért.
- Az Állami Egészségügyi Ellátó Központnak (ÁEEK) a járóbeteg ellátási adatokat tartalmazó adatállományból ingyenesen biztosított kutatási adatokért.

A kutatás előzményei

- A tételes egészségügyi adattár 1998-tól tartalmaz adatokat minden egyes magyar állampolgárról, nem meghatározott ideig (mindhalálig).
- A Nemzeti Egészségügyi Alapkezelő Központ a TAJ azonosítót helyettesíti egy másik azonosítóval, kapcsolati kóddal (pszeudo-TAJ). A TAJ és a pszeudo-TAJ összerendelést tartalmazó adatállomány a NEAK-nál található, és végtelen ideig tárolják.
- Minden egyes rekord (házi orvos, vény, járó- és fekvőbeteg ellátás) tartalmazza a születési dátumot, a lakóhely irányítószámát, és a nemet, továbbá a megjelenés dátumát, orvos pecsétkódokat, intézmények kódja, felírt orvosságok.
- Alkotmányjogi panaszt nyújtottam be 2006-ban ellene száma: 937/B/2006. A határozat szerint, ha a NEAK **személyes adatokat továbbít**, akkor a törvényellenes, ezért rendes bírósághoz forduljak.
- A jog az adatállományt személyazonosításra alkalmatlannak tartja, pedig nem az (az egyes személyazonosításra alkalmatlan ágazati (egészségügyi, szakmai) adatok körének meghatározására, gyűjtésére, feldolgozására vonatkozó részletes szabályokról szóló 76/2004. (VIII. 19.) ESzCsM Rendelet)
- Végül rendes bírósághoz fordultam, és a Kúria végül ellenem ítélt.

A data broker reckoning - Proto X Jon Keegan a Twitteren: "Yes! X" +

https://twitter.com/jonkeegan/status/152224406839410689/photo/1

Placer - Planned Parenthood - X +

https://analytics.placer.ai/#/admin/insights/venues/9ab7ea... | keegan@themarkup.org

Placer.ai

Property

My-Zone Explore Property Chains Advanced Reports Labs Academy Marketplace

Property Reports

Overview

Property Info

Visits

Variance

Visitor Journey

Audience

Trade Area

Area Analysis

Ranking

Loyalty

Void Analysis

Other Reports

Starred Reports

Planned Parenthood

Venues: Planned Parenthood / [redacted]

By: Home Location Metric: Visits Min. Visits: 1 Resolution: Nationwide Visualization: Gradient

of Visits

High

Low

Map Hybrid

Open GIS

Help & Feedback

Keyboard shortcuts Map data ©2022 200 m Terms of Use Report a map error

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

3 3 12

<https://twitter.com/jonkeegan/status/152224406839410689/photo/1>

Fájl Szerkesztés Nézet Előzmények Könyvjelzők Eszközök Súgó

Google, DeepMind face lawsuit X +

← → ↻ https://www.cnbc.com/2021/10/01/google-deepmind-face-lawsuit-over-data-deal-with

Legtöbbször látogatott Bevezetés GoToMeeting Hub FF Free Online QR Code ... Más könyvjelzők

MARKETS BUSINESS INVESTING TECH POLITICS CNBC TV INVESTING CLUB PRO

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

WATCH LIVE

KEY POINTS

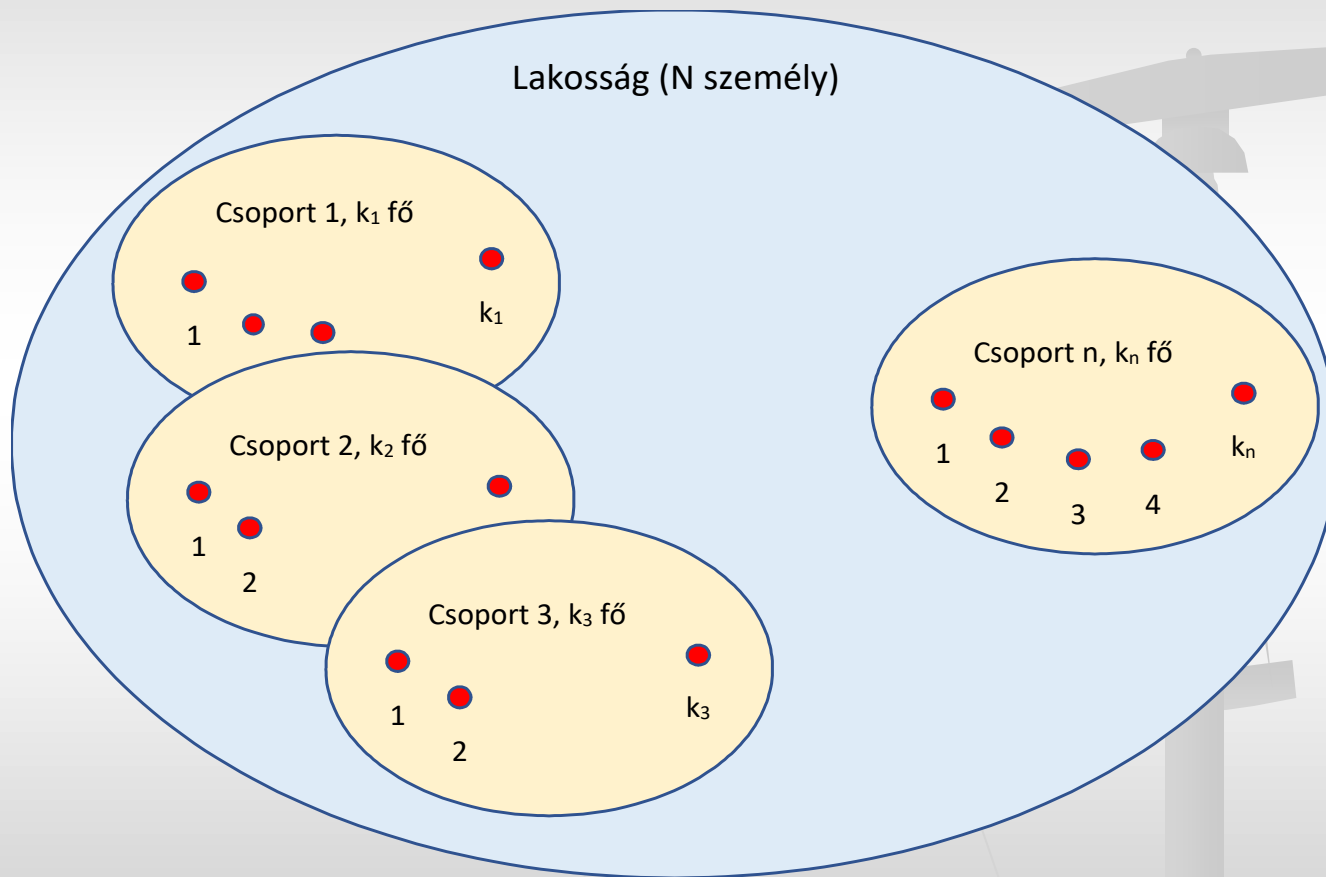
- DeepMind found itself in the spotlight in 2016 when the New Scientist reported that its collaboration with the U.K.'s National Health Service went beyond what was publicly announced.
- British law firm Mishcon de Reya told CNBC Friday it had filed a claim with the High Court on behalf of Andrew Prismall 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

<https://www.cnbc.com/2021/10/01/google-deepmind-face-lawsuit-over-data-deal-with-britains-nhs.html>

Kvázi-azonosítók

- Kvázi-azonosítók: olyan adatelemek, amelyek magukban nem elegendőek egy személy azonosításához, más kvázi-azonosítókkal kombinálva azonban már azonosíthatnak személyeket.
- A kvázi azonosítók **nem közvetlen személyazonosítók**. Ehelyett olyan adatok, amelyek például a földrajzi helyre utalnak, irányítószám, születési dátum. Sok ember lakóhelyének lehet ugyanaz az irányítószáma, és sok embernek lehet azonos a születési dátuma, azonban nagyon kevés esetben azonos mindkettő.
- Olyan típusú adatokról van szó, amelyeket egy ellenfél (adversary) megszerezhet a természetes személyazonosítókkal együtt mint például a név, anyja neve stb. és ezen a módon feltörheti az álnevesített adatállományt.
- Egy álnevesített adatállomány: $r(q_1, q_2, q_3, q_4, q_5, \dots, q_n, d_1, d_2, \dots, d_m)$.
- Egy ellenfélnek birtokában lehet: $a(q_2, q_3, q_4, q_5, \text{név})$.
- Kérdés: mik lehetnek kvázi-azonosítók? Születési dátum, irányítószám, nem, munkakör, végzettség, iskolák, munkahely, betegség, műtét, gyermekszületés.

Személyes csoportosítása kvázi-azonosítóik alapján



Mi az entrópia?

0110011101100110011

1010

Csak a csoportot ismerjük,
ahol a keresett személy
található.

$$\log_2(N) - \log_2(k) = \log_2(N/k)$$

k megkülönböztethetetlen
személy, $\log_2(k)$ bit

- Az entrópia egy súlyozott összeg (várható érték, átlag) azon információ mennyiségének átlaga, amelyet az egyes személyekről tudunk.

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

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

Az entrópia tulajdonságai

Az entrópia kapcsolatban áll a k -anonimitással, és teljesül rá, hogy ha egy adatállomány k -anonim, akkor az entrópia: $E(D) < -\log_2(k/N)$. Általános esetben azonban az egyes csoportok elemszáma változó lehet, de ekkor is képezhető a képletből, átrendezéssel egy hozzávetőleges $\hat{k}$ érték.

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

Ha pedig az entrópia elég nagy, azaz $\log_2 N - 1$ -nél nagyobb, akkor feltételezve, hogy a csoportok csak egy és kételeműek, akkor az egyelemű csoportok számára van egy becslés. Ez a szám lehet nagyobb is, ha vannak három, négy, ... elemű csoportok is.

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

Az irányítószám entrópiája

Irányítószám	Település	Lakosság	Bitek ($\log_2(N/k)$)	Entrópia ($-(k/N) \cdot \log_2(k/N)$)
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
Összesen:		10,004,090		10.303428

- A számítás eredménye szerint az entrópia az irányítószám átlagos információtartalma 10,3 bit. Ez azt jelenti, hogy egy véletlenül kiválasztott személy esetén, ha tudjuk a lakóhelye irányítószámát, akkor az átlagosan 10,3 bit információ. Ez 7916-anonimitásnak felel meg.
- A teljes lakosság egyértelmű azonosításához Magyarországon (a lakosság 10 004 090) $\log_2(N) = 23.254$ bitre van szükségünk.

A születési dátum + irányítószám

Születési dátum x irányítószám	Lakosság	Bitek ($\log_2(N/k)$)	Entrópia ($-(k/N) \cdot \log_2(k/N)$)
(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
...			
Sum:	10,004,090		22.79385

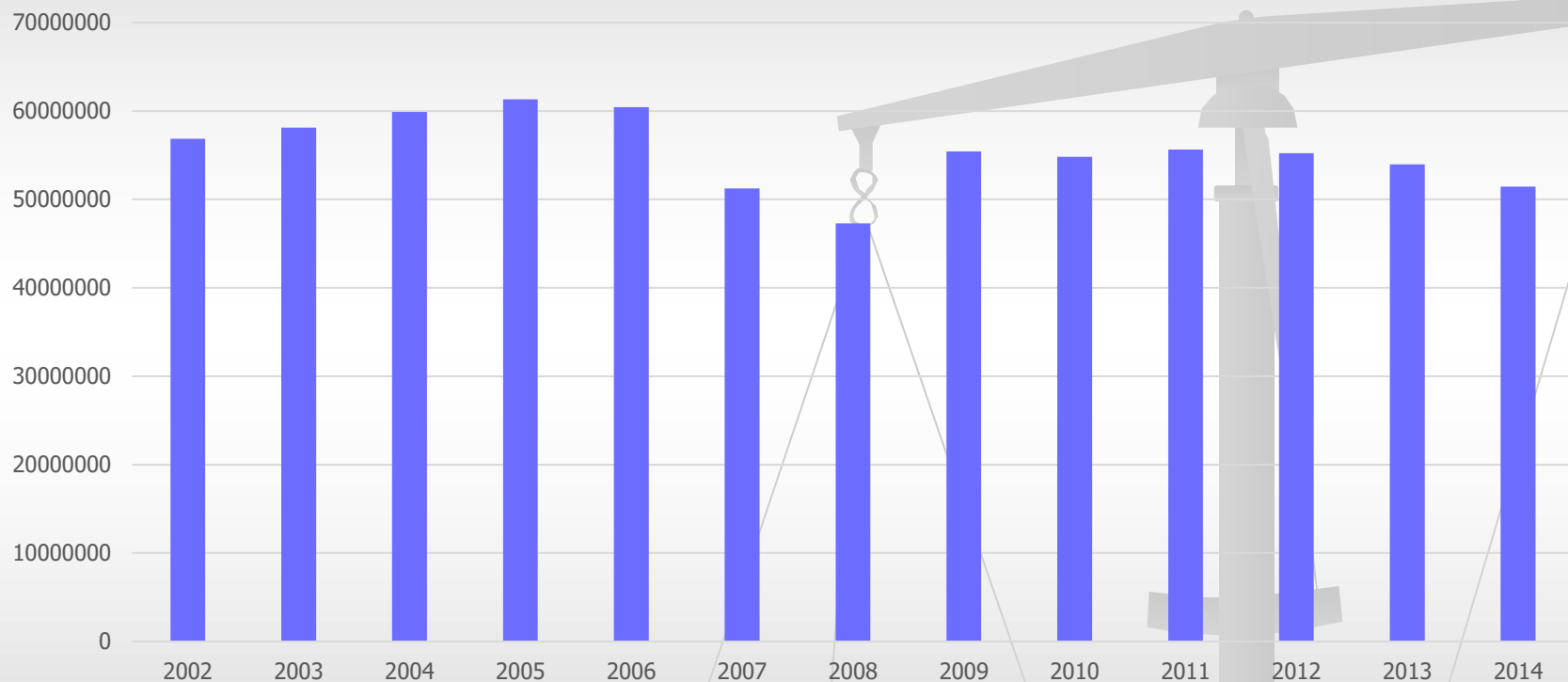
- Az entrópia 22.7985 bit volt. Ez 1.37-anonimitásnak felel meg. Egy ilyen adatállomány jelentős kockázatot jelent az újra-azonosítás szempontjából.
- Az egy elemű csoportok előrejelzett száma: 54% volt, a valóságban azonban több fordult elő: 6 635 838 személy.

Járóbeteg vizit adatbázis

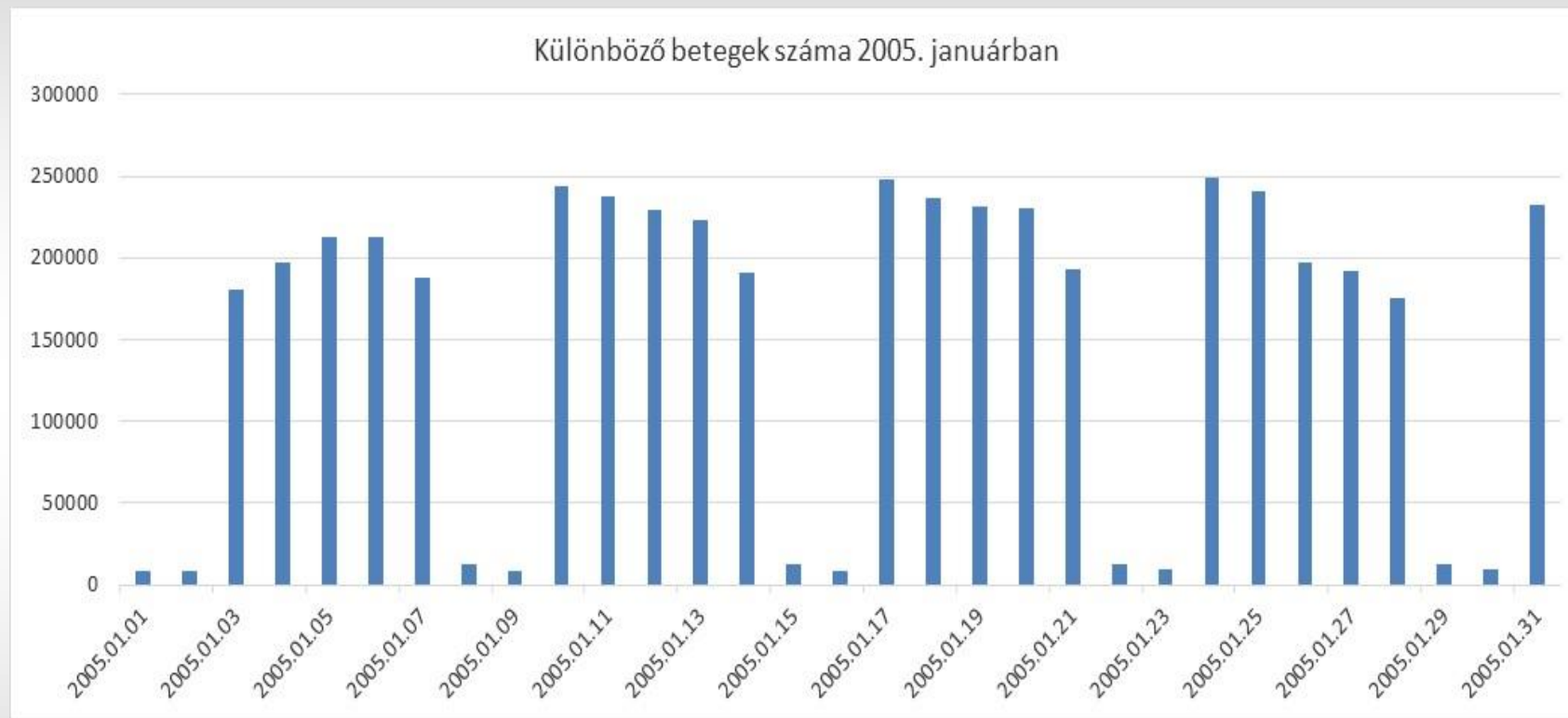
- Az összes NEAK-hoz jelentett járóbeteg ellátás adatát tartalmazza, összesen 721 633 881 vizit adatát. Többségében 2002. január 1. és 2014. december 31-e között.
- Az adatelemek:
 - Kapcsolati kód, amely lehetővé teszi ugyanazon beteg ellátásainak azonosítását;
 - A vizit dátuma;
 - A BNO-kód első betűje kétjegyű számmal kódolva (nem tudom, hogy melyik kód milyen betűt jelent);
 - A rendelőintézet irányítószámának első két számjegye.
 - "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)

Járóbeteg vizit adatbázis

Különböző járóbetegek száma



Járóbeteg vizit adatbázis



Járóbeteg vizit adatbázis

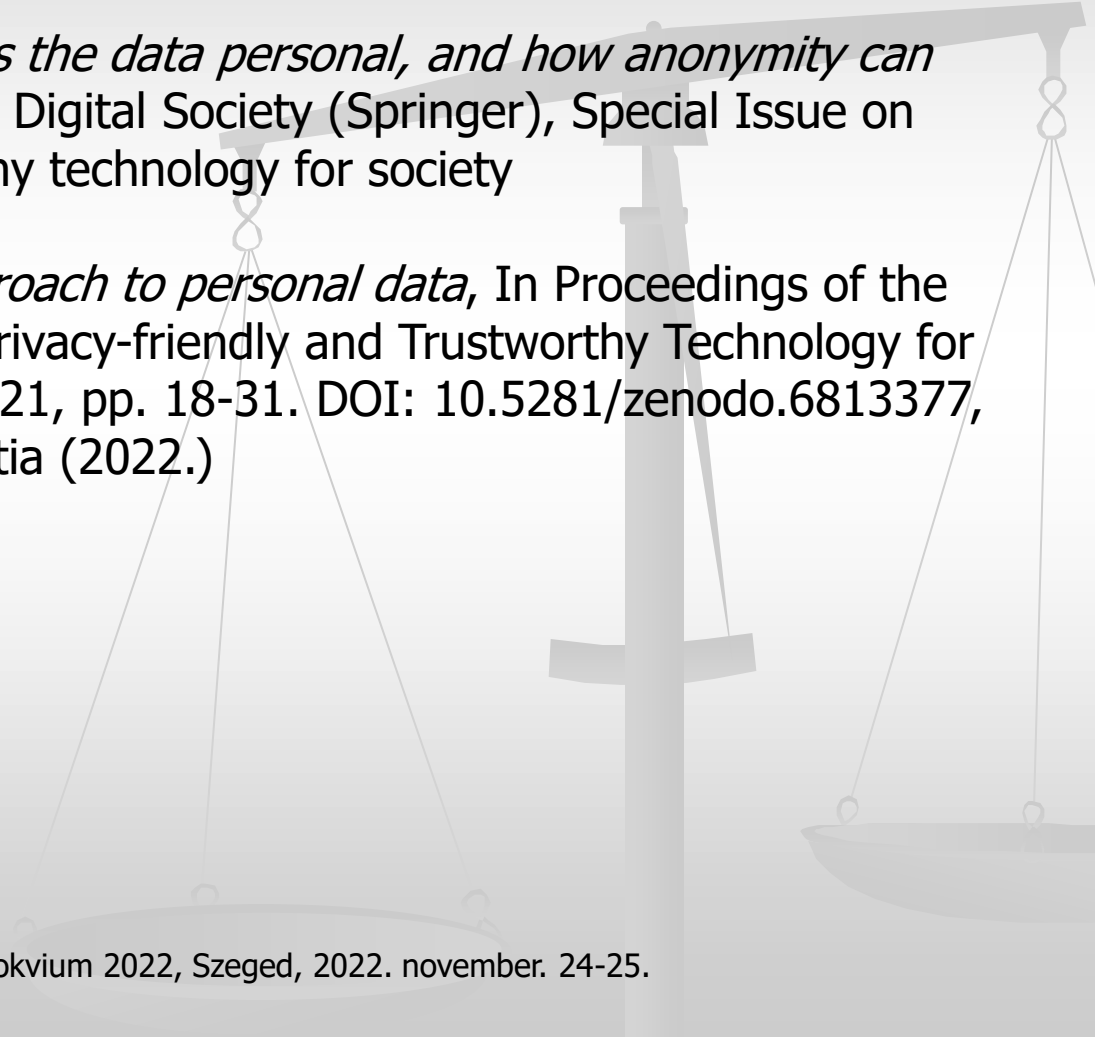
Az orvos-beteg találkozások eloszlása 2005



Vizit-párok entrópiája

Dátum	Entrópia	Becsült k	Egyelemű csoportok
...			
8 th January 2005 (Szo)	19.725	1.247	902,032 (83.45%)
9 th January 2005 (Vas)	19.158	1.147	596,462 (88.86%)
10 th January 2005 (Hét)	23.254	2.889	12,809,964 (44.32%)
11 th January 2005 (Kedd)	23.233	2.865	12,561,551 (44.46%)
12 th January 2005 (Sze)	23.214	2.750	12,234,144 (45.71%)
January 2005	27.556	2.734	250,588,888 (46,47%)

- A legaktívabb betegek (akik ezerszer vagy többször jelentek meg) törlésre kerültek (1968 beteg, az összes beteg 0,016%-a, 3 348 081 vizit került így ki, amely a 721 millió ellátás 0,46%-a).
- A generált párok nagy méretére tekintettel, egy-egy hónapra eső párokat, először a 2005. januárra eső párokat határoztam meg, ahol a korábbi vizit dátuma eset 2005. januárra. Ez 3 863 139 808 (88.8 GB) párt jelentett.
- A vizit párokból álló csoportok mérete 1 és 3233 közé esett. A legnagyobb csoportban 3233 beteg volt, akik budapesti rendelőintézetekben jártak január 24-én és 26-án (hétfő, szerda). Összesen 539 225 013 különböző vizit pár volt, az egyelemű csoportok száma 250 588 888 volt.



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

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, 28th June 2022. Zagreb, Croatia (2022.)

Köszönöm a figyelmet!

