Az élettörténet sosem lehet anonim

Dr. Alexin Zoltán, PhD.
Szegedi Tudományegyetem,
TTIK, Szoftverfejlesztés Tanszék
H-6720 Szeged, Árpád tér 2.
e-mail: alexin@inf.u-szeged.hu
http://www.inf.u-szeged.hu/~alexin

Bemutatkozás

- okl. matematikus 1985. JATE, TTK
- PhD. matematikatudomány 2003.
  - Szabályalapú gépi tanulóalgoritmusok és alkalmazásaik természetes nyelvi problémák megoldására
- EuroSOCAP FP6, QRLT-2002-00771, szakértő, 2007.
  - Európai szabvány a bizalmasság megvalósítására az egészségügyben
- The European Privacy Institute, FP7 terv, tudományos tanácsadó
- Regionális Humán Orvosbiológiai Kutatásetikai Bizottság, tag, 2009.
- Tisztességes Adatkezelésért Egyesület, 2009. november 6.
- Európai Bizottság, új adatvédelmi rendelet konferencia, Brüsszel, 2010.
- Szenzorhálózatok/Future-ICT.hu TÁMOP projektekben alprojekt vezető
- SZTE Szent-Györgyi Albert Klinikai Központ, belső adatvédelmi felelős
- COST IC1206, project WG4 leader, MC member, 2015.
- VICTORIA, H2020 project, Privacy Advisory Board member, 2017.
- MediConSec Project 2021, adatvédelmi és etikai tanácsadó
- COST CA19121 („GoodBrother”), CA19136 („NET4Age”) MC Member, 2021.

Jogi ügyek

- **Alkotmánybíróságon:**
  - 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.)

Jogi ügyek

- **Rendes bíróságon:**
  - 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:**
  - A foglalkozás-egészségügyi beutalón nem lehet TAJ azonosító

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

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

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.

Referenciák

- 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 alapvérkép 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.)

Köszönetnyilvánítás

A szerző köszönetét fejezi ki:

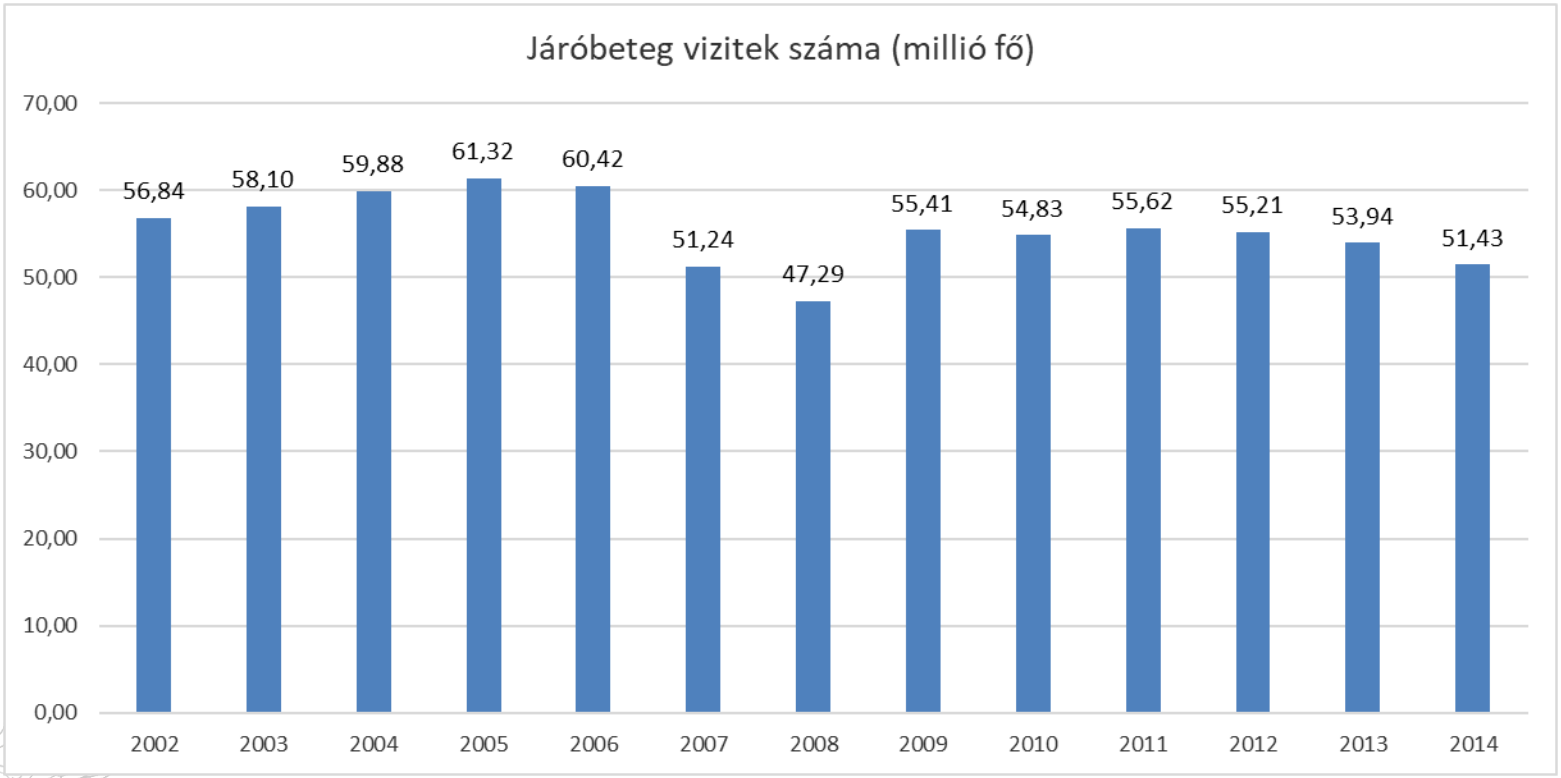
- Az Állami Egészségügyi Ellátó Központnak (ÁEEK) 2002-2014 évekre (13 év) vonatkozó járóbeteg ellátási adatokat tartalmazó adatállományból ingyenesen biztosított kutatási adatokért.

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"
```

Megjelenések



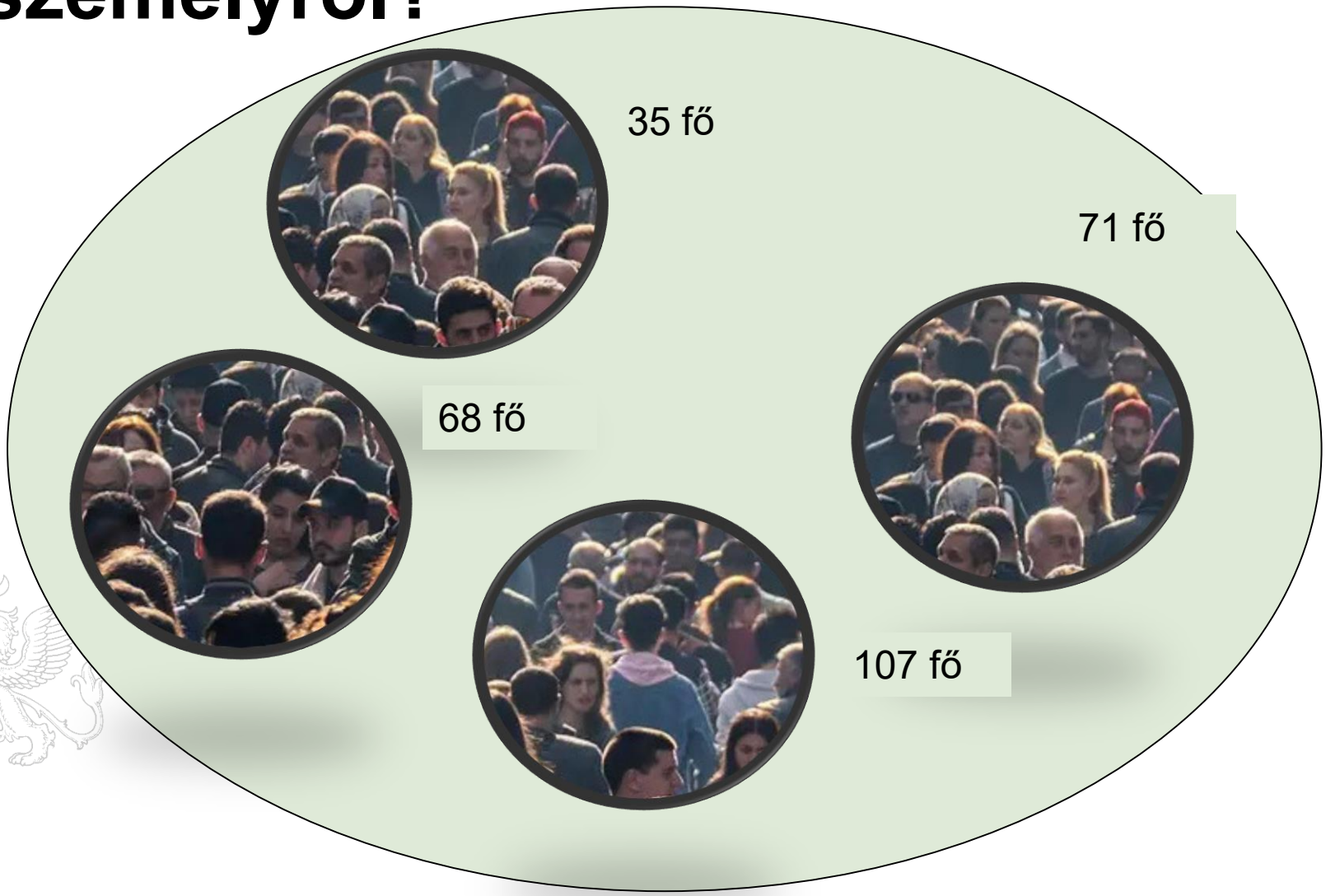
Összesen: 721 525 095 megjelenés, 12 200 284 különböző személy

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 / tripletekhez 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 tripleteket 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).

Hány bitet tudunk átlagosan egy személyről?



Hány bit információt tudunk egy személyről?

68 fő

N különböző személy megkülönböztetéséhez $\log_2(N)$ bitre van szükség.

Ha azonban $k = 68$ személyt nem tudunk egymástól megkülönböztetni, akkor $\log_2(k)$ bittel kevesebbet kell használnunk, kevesebb információt tudunk a benne található emberekről.

10111000111010100010101011

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

Ha több csoport van, akkor kiszámíthatunk egy átlagos információ tartalmat.

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

Hány bitet tudunk átlagosan egy személyről (entrópia)

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

$E(I) = 68/281 * \log_2(281/68) + 35/281 * \log_2(281/35) + 107/281 * \log_2(281/107) + 71/281 * \log_2(281/71) = 1{,}90153$

A k-anonimitás és az entrópia

- Az entrópia szoros kapcsolatban áll a k-anonimitással, és teljesül rá, hogy ha egy adatállomány k-anonim, akkor az entrópia kisebb, mint $E(I) <= \log_2(N/k)$. Az előbbi példánál maradva, ha az adatállomány 35-anonim, akkor az entrópiának (1,9) kisebbnek kell lennie mint $\log_2(281/35) = 3{,}005$.
- Az is látható, hogy egy valódi adatállományban a megkülönböztethetetlen személyek csoportjainak számossága változik, ezért sokkal inkább lehet egy becsült (várható) *k* értéket meghatározni, ami átrendezéssel kapható:

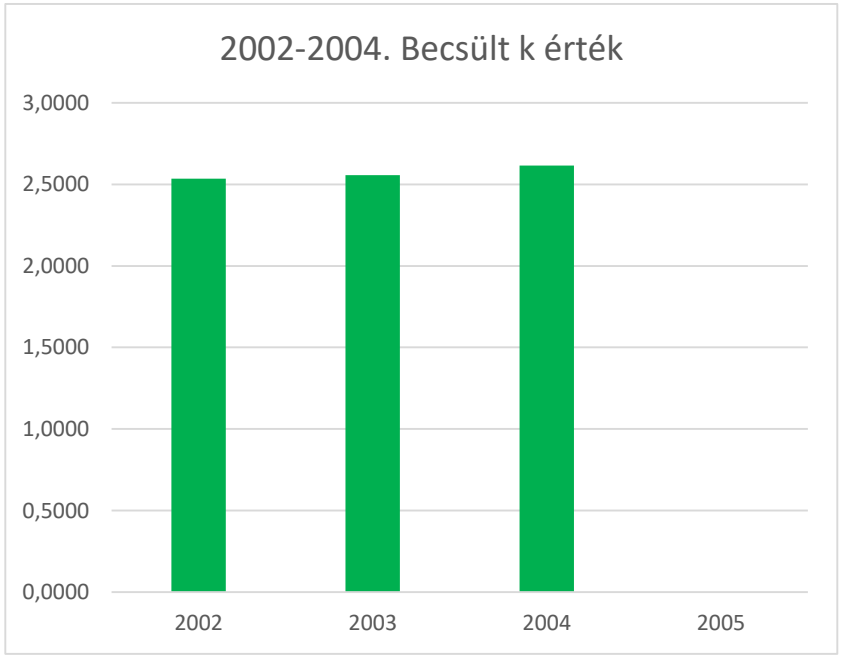
$$k = \frac{N}{2^{Entropy}}$$

- Ez valami olyasmit jelent, hogy egy véletlen módon kiválasztott személlyel egy csoportban várhatóan *k* személy lesz.
- Alexin, Z.: Entropy based approach to personal data (Slides), 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.)

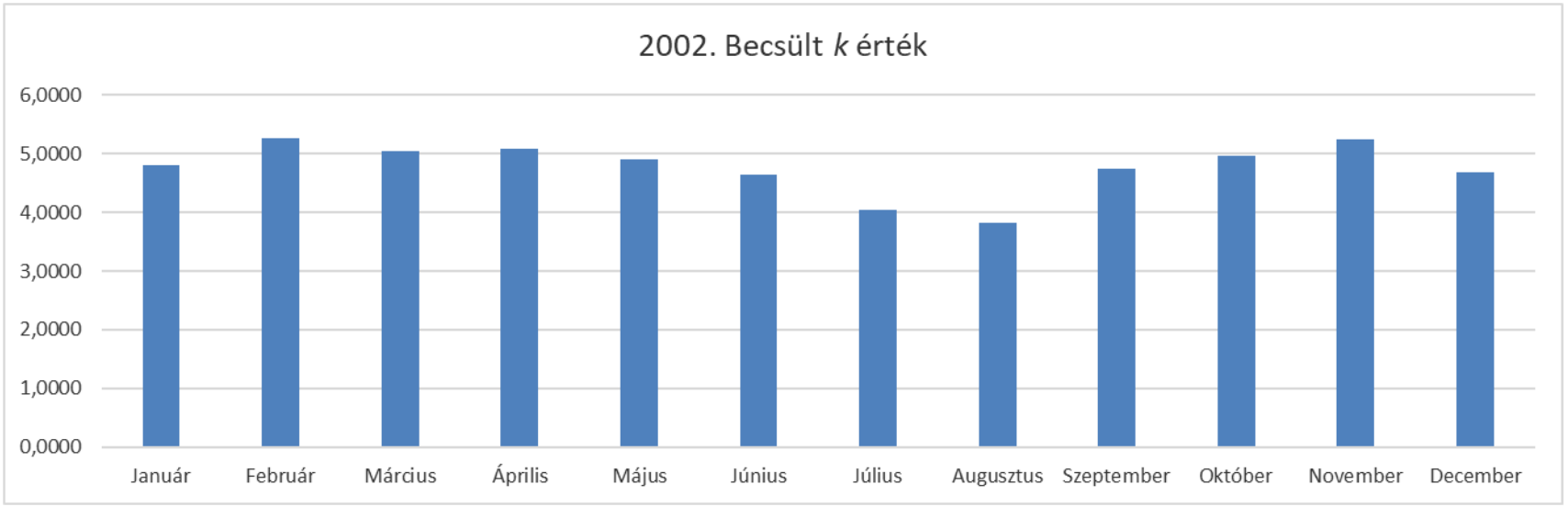
2002-2004 évek adatai párokra

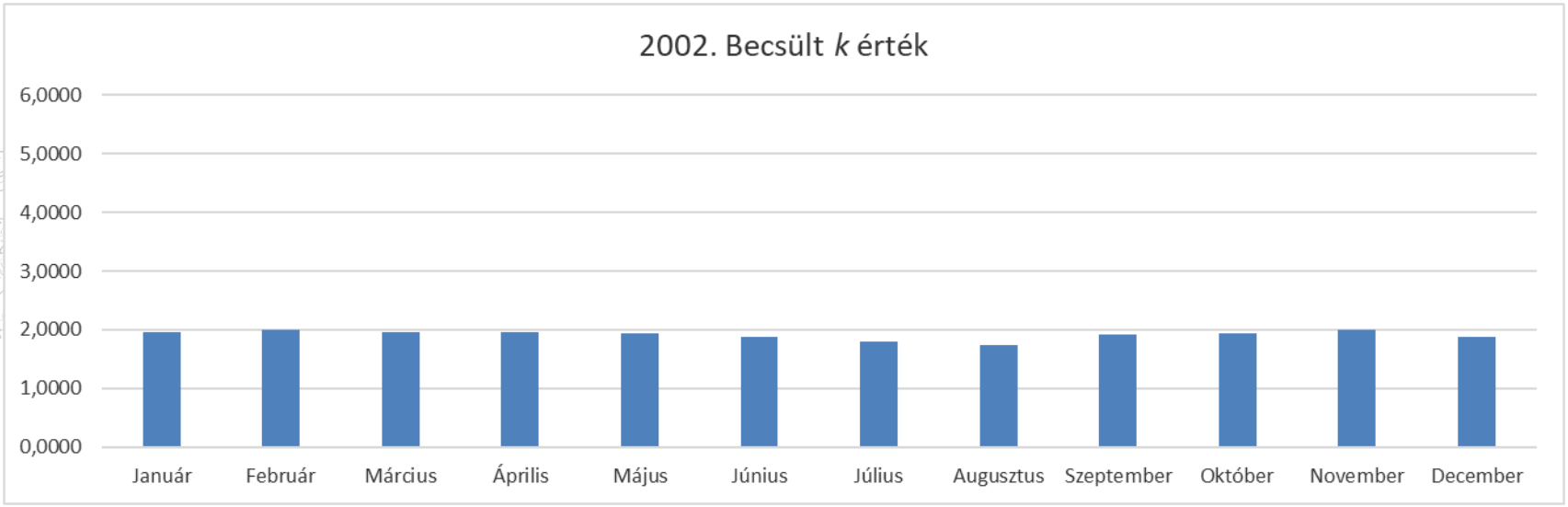
Maximális entrópia	34,250899
Számított entrópia	32,890720
Vizitpárok száma	20 443 157 017
Becsült k érték	2,5671694

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



Budapest és a vidék





UNIVERSITAS SCIENTIARUM SZEGEDIENSIS
SZEGEDI TUDOMÁNYEGYETEM
Szoftverfejlesztés Tanszék

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

UNIVERSITAS SCIENTIARUM SZEGEDIENSIS
SZEGEDI TUDOMÁNYEGYETEM
Szoftverfejlesztés Tanszék

Köszönöm a figyelmet!