„Infokommunikációs technológiák és a jövő társadalma (FuturICT.hu)” projekt

TÁMOP-4.2.2.C-11/1/KONV-2012-0013

Egészségügyi adatbázisok tisztességes anonimizálása

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

IME, XI. Országos Egészségügyi Infokommunikációs Konferencia
2013. május 29. Budapest

Nemzeti Fejlesztési Ügynökség
www.ujszechenyiterv.gov.hu
06 40 638 638

MAGYARORSZÁG MEGÚJUL

A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

Előzmények

- A TEA (Tételes Egészségügyi Adattár) létrehozása 2004-ben a 76/2000. (VIII. 19.) számú ESzCsM rendelettel
- Az adatvédelmi biztos 1301/A/2006-9. számú állásfoglalása (négy egészségügyi adatvédelmi problémára hívta fel a figyelmet – de egyiket sem fogadta meg a tárca)
- Az Alkotmánybírósághoz benyújtott 937/B/2006 indítványom ügyében hozott elutasító határozat (az OEP csak személyazonosításra alkalmatlan adatokat továbbíthat – vizsgálat nélkül elfogadta az AB)
- Paul Ohm: Broken promises of privacy – cikk a születési dátum, nem, irányítószám alapján történő személyazonosításról (87,1%)
- Kanadában is folytak hasonló kutatások (80%)

Mit jelent az anonimizálás?

- Az anonim a görög *ἀνωνυμία* (anonymia) szóból ered, jelentése név nélkül.
- A cél, hogy az érintett személyisége rejtve maradjon, ne fedje fel a kilétét.
- Az infokommunikáció korában az **anonimizálás** azt jelenti, hogy minden olyan adatot el kell távolítani, amely ahhoz vezethet, hogy az érintett azonosítani lehessen.
- A személynév eltávolítása nem elegendő.
- Az emberek azonosíthatók nem csak a nevük alapján, hanem pl. munkahely, munkakör, munkahelyi vezető; vagy pontos lakóhelyük; vagy születési dátum, iskola, osztályfőnök; kórház, orvos, beavatkozás, dátum adatok alapján is.

Azonosító	Munkakör	Születési dátum	Betegség
10784343	tanár	1965. május 3.	HIV+
13453453	könyvelő	1946. június 2.	rák
53353534	járművezető	1964. augusztus 17.	szifilisz+

Find us on Facebook

	Hobbi	Születési dátum	Munkakör
David Simon	hegymászás	1964. augusztus 17.	járművezető
John Smith	vitorlázás	1965. május 3.	tanár
Jackie Chan	búvárkodás	1946. június 2.	könyvelő

A gondatlan anonimizálás veszélyei

- Az anonimizálás káros hatásait nem lehet meg nem történtté tenni.
- A már közzé tett adatokat nem lehet visszavonni.
- Olyan károkat okoz, amelyeket nem lehet jóvátenni, orvosolni.
- Egy jövőbeli kockázat (az újra azonosítás kockázata folyamatosan fenyegeti az érintetteteket).
- **Nem tekinthető tisztességes adatkezelésnek.**
- Elizabeth France (angol adatvédelmi biztos, 1998) : az anonimizálás is adatfeldolgozás és csak törvényi felhatalmazás alapján hajtható végre.

Előnyök és hátrányok

- Előnyök:
  - Az adatok feldolgozhatók, anélkül, hogy az érintetteket sértenék
  - Az adatvédelmi törvényt nem kell alkalmazni
  - Nem merülnek fel etikai kérdések
  - Az adatok megoszthatók, eladhatók
- Hátrányok:
  - Az emberek és cégek (munkahelyek) egyre több információt tesznek fel magukról az Internetre
  - Nem tudjuk megjósolni a jövőt. Legközelebb milyen információt fognak nyilvánosságra hozni.
  - Egyes cégek leszüretelik a nyilvános információkat a webről (neveket, fényképeket, születési dátumokat, lakóhelyet, iskolákat stb.)
  - Ez megteremti az iparszerű újraazonosítását az állítólag anonim adatoknak.

A Tételes Egészségügyi Adattár

- Az OEP által összegyűjtött elszámolási adatokból származik.
- Az OEP negyedévente küldi az új adatokat TEA adattárba, amelyből egy-egy példány jelenleg a GYEMSZI-nél és az OTH-nál is megtalálható
- Az OEP kicseréli a TAJ azonosítókat egy pszeudo-TAJ azonosítóra amely ugyanúgy személyes azonosításra alkalmas, így az egy személyre vonatkozó ellátások adatai összekapcsolhatók. Lényegében minden magyar állampolgár megtalálható az adattárban.
- A járó- és fekvőbeteg ellátások, valamint a vénykiváltások adatai vannak az adattárban.
- A megőrzési idő nincs meghatározva, élethosszig tart.
- Tartalmazza a páciens lakóhelyének az irányítószámát, a páciens nemét és születési dátumát is.
- A nem támogatott vények adatai is szerepelnek benne (legalábbis 2009-ig).
- Nincs független adatvédelmi és orvosi etikai felügyelet.

Statisztikai adatbázis a kockázat elemzéshez

- Az állami népességnyilvántartásból származó statisztikai adatok.
- Tartalmazza az irányítószámot, nemet, születési dátumot minden magyar lakóhellyel rendelkező állampolgárról (10 004 090 fő).
- P-ikrek (pseudo ikrek): olyan személyek, akik ugyanabban az irányítószám körzetben laknak, azonos neműek, és azonos napon születettek. Ha egyéb adat nem áll rendelkezésre, akkor megkülönböztethetetlenek.
- A legnagyobb klón 11 P-ikerből áll (1 klón, 1975), majd további 12 klón tartalmaz 10 P-ikert, stb.

1011;1989.01.23.;N;2

1011;1989.02.01.;N;1

1011;1989.03.11.;N;1

(8 million lines)

…

Postai irányítószámok

Falvak és városok

Lakosság	Irányítószám körzetek száma	Teljes lakosság	P_1	P_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%
Összesen:	3110	10004090	78,426%	94,001%

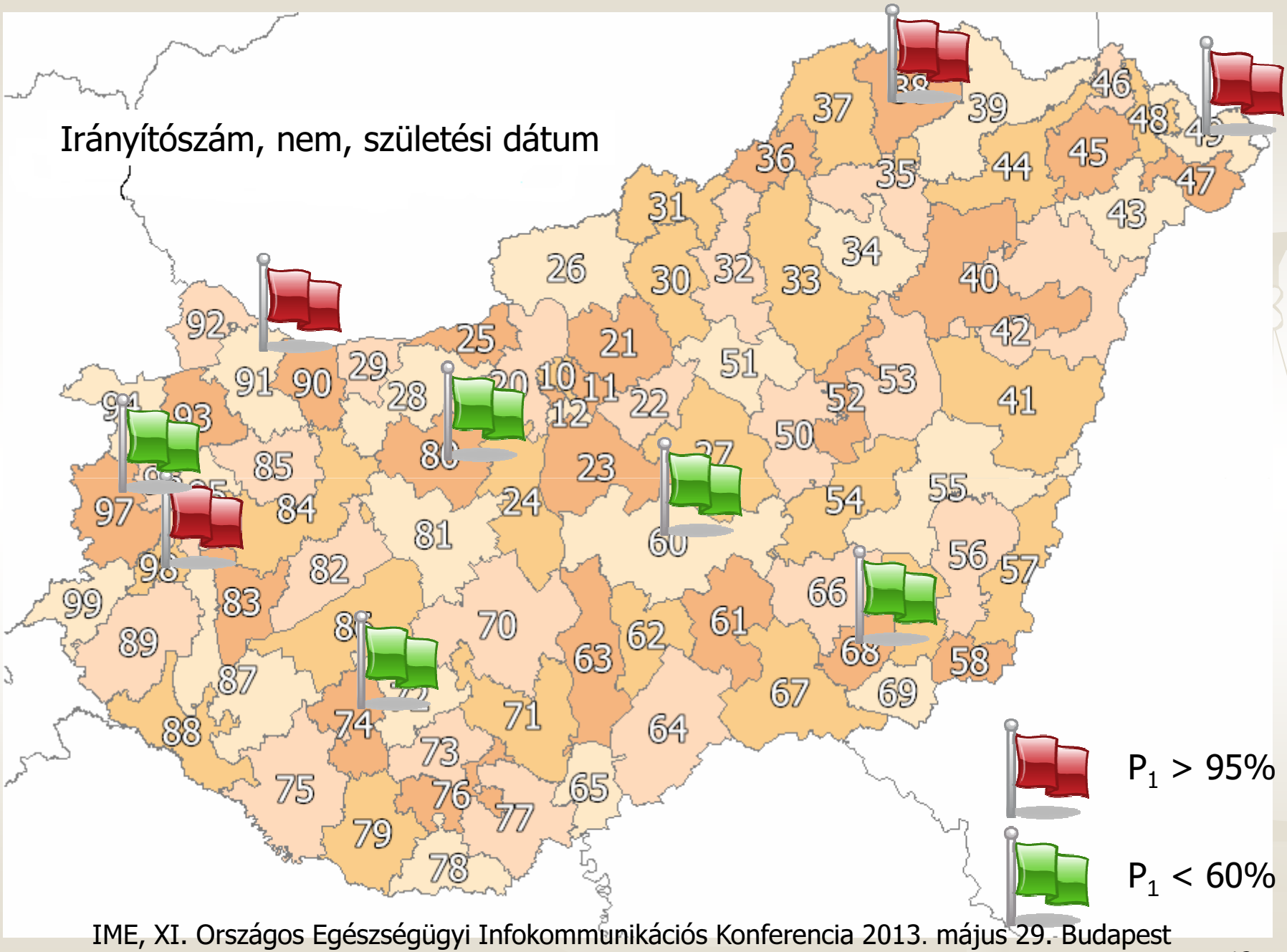
Egy személy akkor egyértelműen azonosítható, ha nincs P-ikertestvére.
P_1 = az azonosítás valószínűsége = A P-egykék száma osztva az összes személy számával.

Ha két személy közül mindig ki tudjuk számítani azt az egyet, amelyet azonosítani akarunk (egyéb információ alapján).

P_2 = (a P-egykék száma + P-kettes ikrek száma) / összes személy száma.

P-ikrek eloszlása

ZIP	Régió	1	2	3	4	P_1	P_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%


Irányítószám, nem, születési dátum
P_1 > 95%
P_1 < 60%

Az azonosítás kockázatának csökkentése

- Születési dátum helyett csak év, hónap: P_1 = 14,995%, P_2 = 27,679%
- Születési dátum helyett csak év: P_1 = 0,037%, P_2 = 0,081%
- Irányítószám első három számjegye: P_1 = 57,859%, P_2 = 82,090%
- Irányítószám első két számjegye: P_1 = 14,814%, P_2 = 33,565%
- Születési dátum helyett csak év, hónap és az irányítószám első három számjegye: P_1 = 1,853%, P_2 = 5,017%

Összefoglalás

- Egyes személyeket a törvény kötelez arra, hogy közzé tegyék életrajzukat és vagyonnyilatkozatukat. Ebben szerepel a születési dátum, lakóhely (tudósok, politikusok).
- Az üvegzseb törvény előírja a parlamenti képviselők számára, hogy tegyék közzé életrajzukat és vagyonnyilatkozatukat.
- Híres színészek, miniszterek lakóhelye (csak a város) sokszor elhangzik egy TV műsorban.
- Budapest kis körzetekre van osztva, úgy viselkedik mint egy nagyobb falu vagy kisváros.
- Sokkal nagyobb gondosság lenne szükséges egy ilyen adatbázis létrehozásakor és használatakor!

Köszönöm a figyelmet!