Információs Technológiák és a Jövő Társadalma (FuturICT.hu)
TÁMOP-4.2.2.C-11/1/KONV-2012-0013

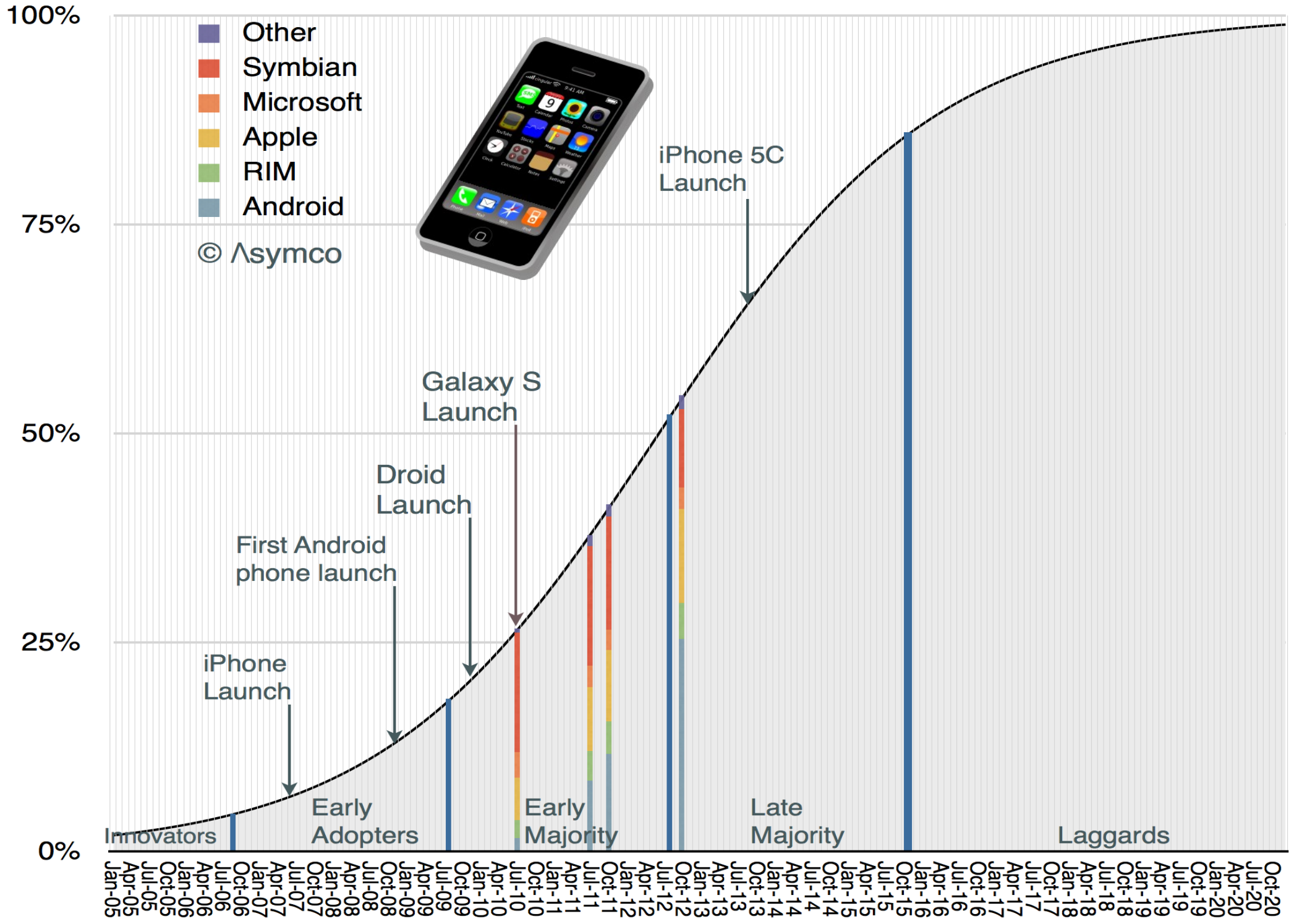
ÚJ SZÉCHENYI TERV

A magánélet védelme az elosztott adatbányászatban

Jelasity Márk

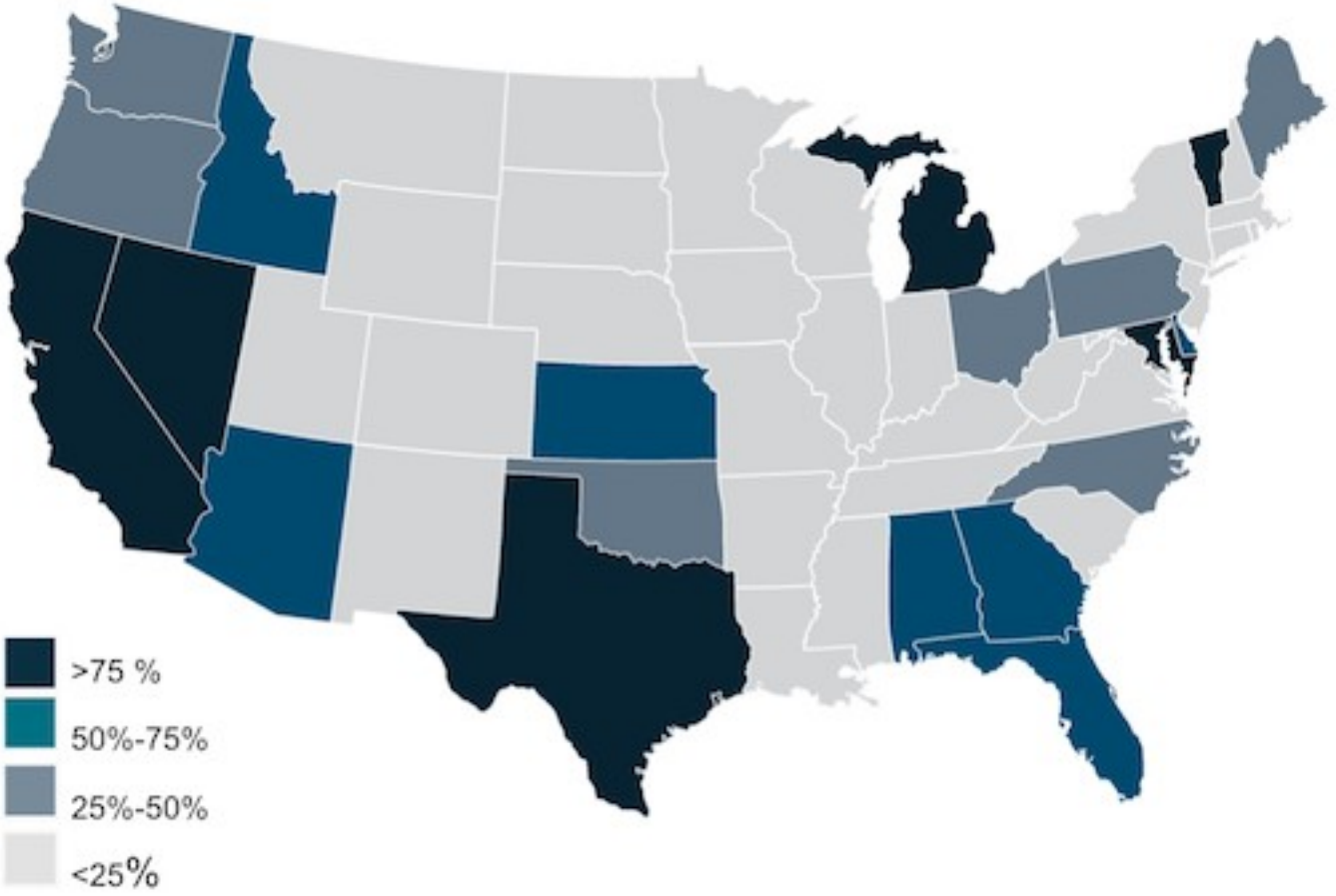
Szegedi Tudományegyetem

EU5 Smartphone Penetration



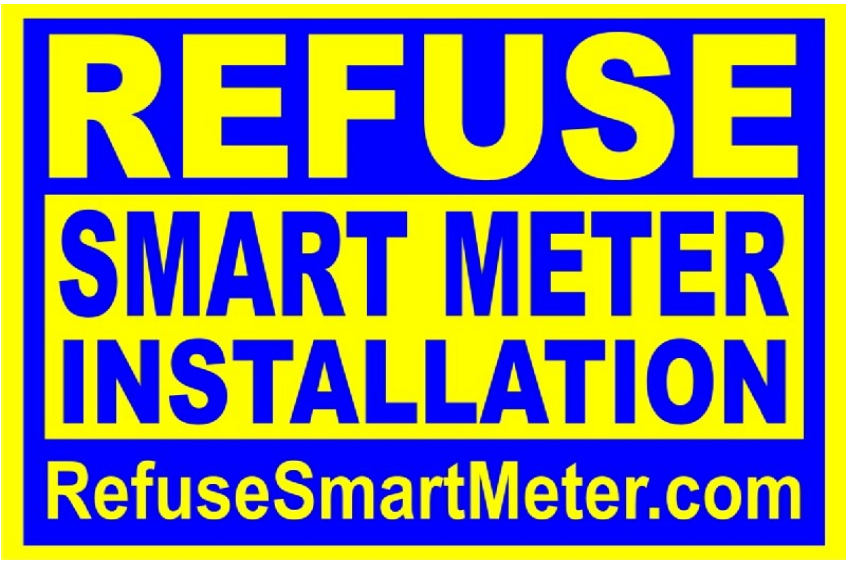
GTM RESEARCH

United States Smart Meter Penetration by 2014



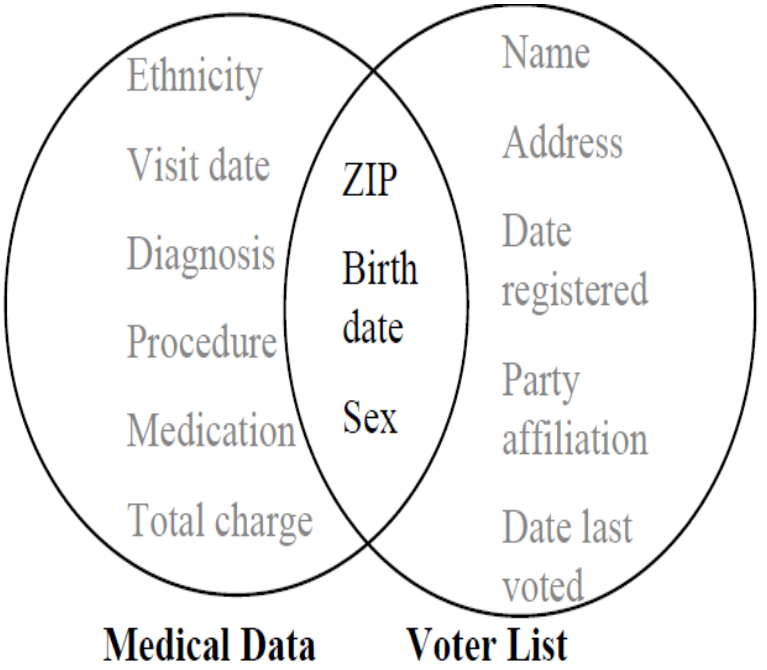
Aggodalom a magánéletért

- Az okos mérőórák érzékeny információkat ismernek
  - Mikor vagyunk otthon, sőt, mit nézünk a tévében, stb.
- Az okostelefonok az egész életünket követik
  - Szenzorok (gps, hang, kamera), alkalmazások (facebook, google, stb.)
- Ugyanakkor hasznos adat-alapú alkalmazások



A magánélet védelmének problémái

- Személyes adatokat „sanitized” (tisztított) formában sem biztonságos közzétenni: kapcsolásos támadás (linkage attack)
  - Netflix and IMDB [Narayanan and Smatikov]
  - MGIC and voting register [Latanya Sweeney]
  - Stb.



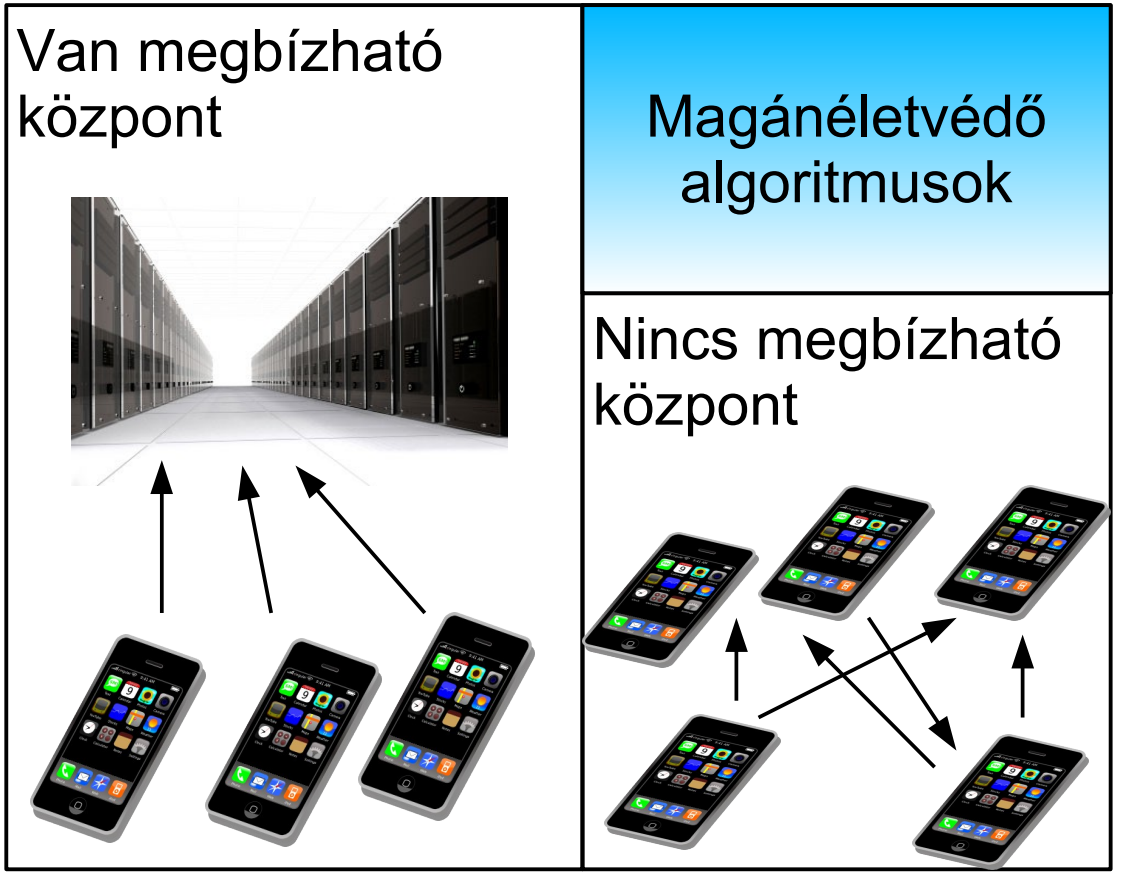
- Bizonságos lekérdezés-végrehajtás kell, és kontrollálni kell a megengedett lekérdezéseket ill. az eredményük publikálását

Rendszermodell

- Nagyon nagyszámú eszköz (sok millió)
  - Minden eszközön adat keletkezik
  - Viszonylag kis mennyiségű, azonos típusú adat van minden eszközön (horizontális adateloszlás)
- Csomagkapcsolt hálózati kommunikáció
  - Minden eszköznek címe van
  - Közvetlenül üzenhetnek egymásnak
    - **Ez nem teljesen igaz (NAT, tűzfalak)**

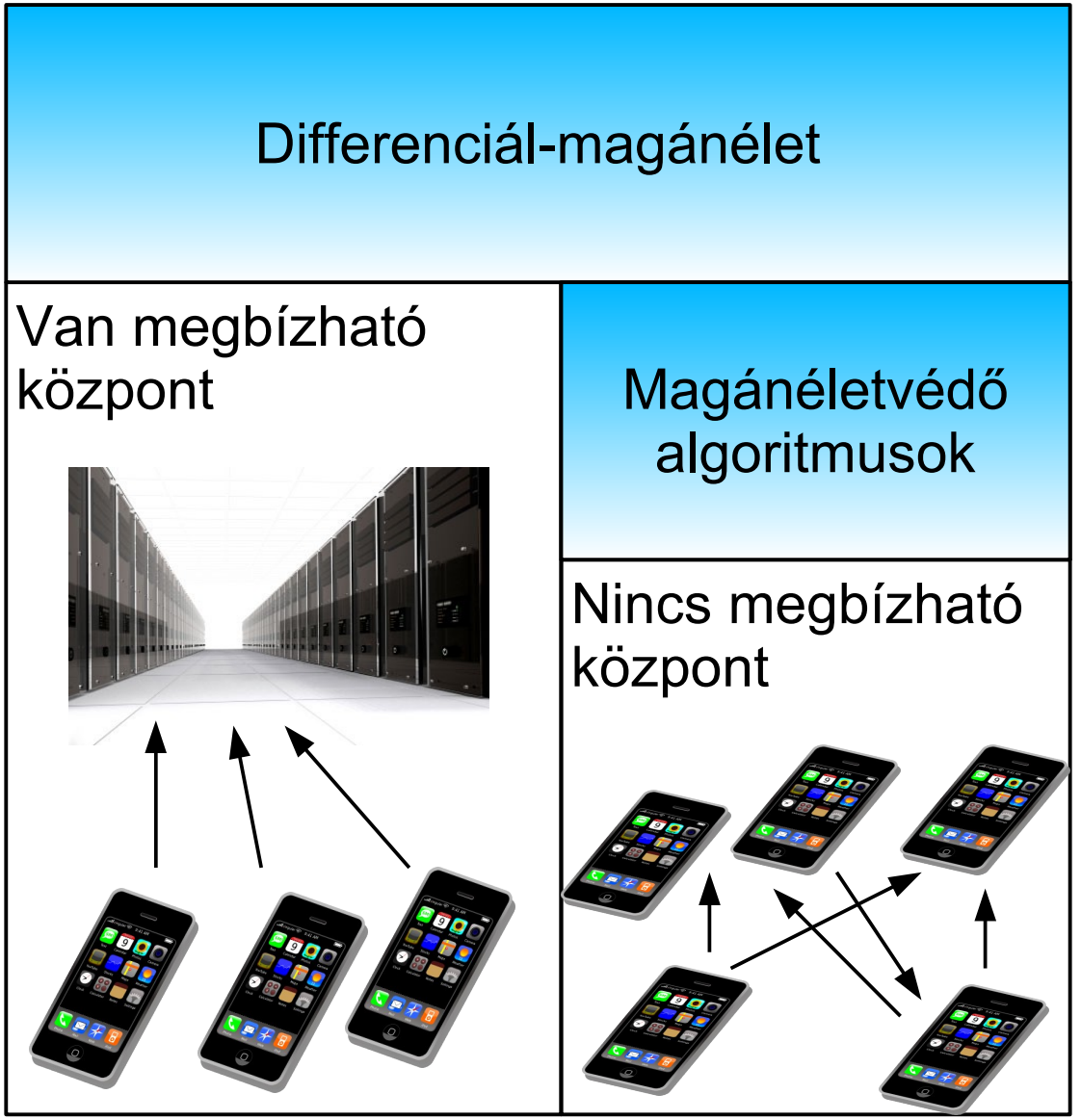
Magánélet védelmének problémái

- Lekérdezések végrehajtása
- Ha nincs megbízható központ, magánéletvédő algoritmus kell
- Kriptográfiai terület, „secure multiparty computation” algoritmusok



Magánélet védelmének problémái

- Lekérdezések kimenetének publikálása
- Kimenetből is lehet rekordokra következtetni
  - min, max
  - Több (jól megtervezett) lekérdezés elemzése



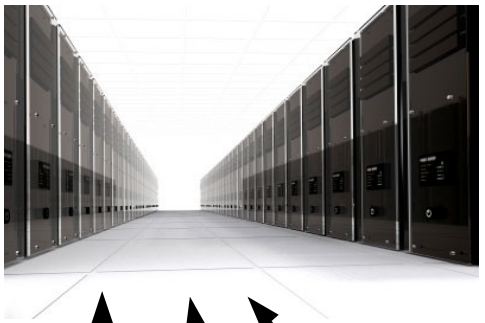
Magánélet védelmének problémái

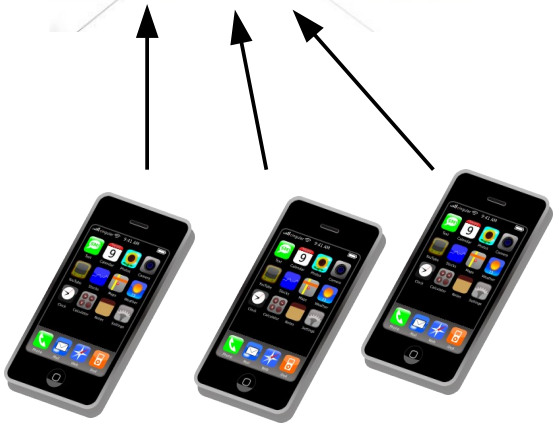
Lekérdezés-kimenetek publikálása

Lekérdezések végrehajtása

Differenciál-magánélet

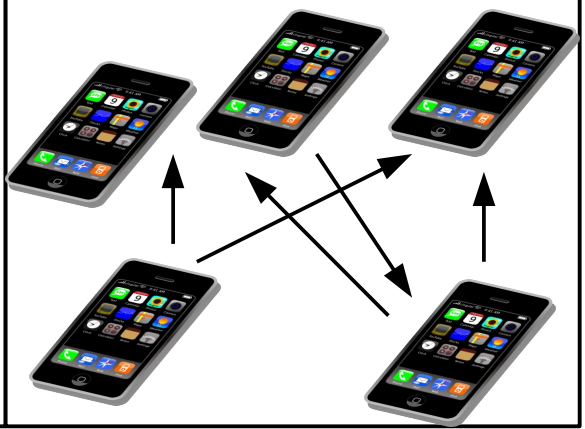
Van megbízható központ





Magánéletvédő algoritmusok

Nincs megbízható központ



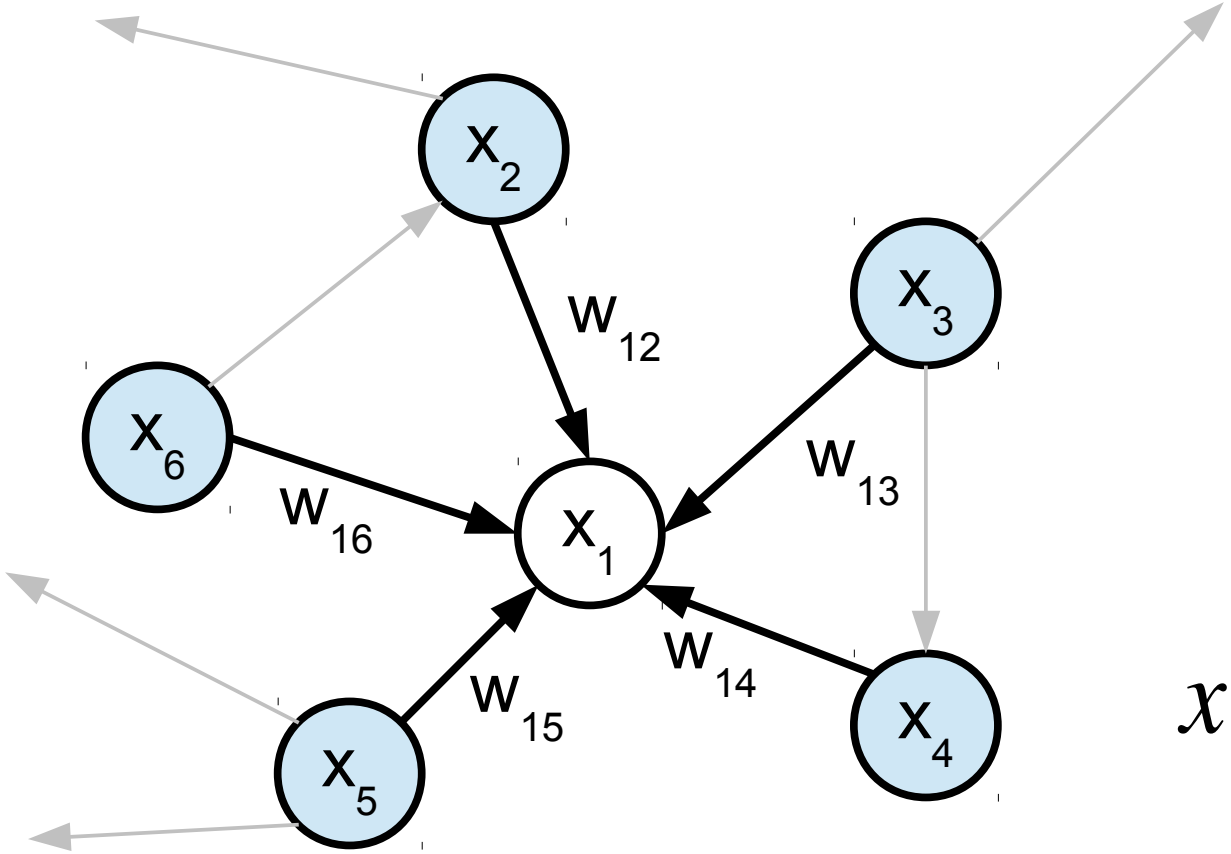
Elosztott hatványiteráció

- sajátvektor alkalmazásai
  - fontossági rangsor, gráfvizualizáció, bizalom mértéke, stb.
- A hatványiteráció az egyik legegyszerűbb módszer (a normalizálást nem mutatjuk)

$$x_i \sim \sum_{i=1}^{n} a_{ij} x_j$$

$$x_i^{(m+1)} = \sum_{i=1}^{n} a_{ij} x_j^{(m)}$$

Elosztott adatok gráf alakban



$$x_i^{(m+1)} = \sum_{i=1}^{n} a_{ij} x_j^{(m)}$$

Ötlet

- Az x_i csúcs minden j be-szomszédja titokmegosztó (secret sharing) algoritmussal megosztja a $w_{ij} x_j$ értéket a többi be-szomszéddal
- A megkapott titokrészletekből egy aggregált értéket elküldenek x_i-nek, amely ezekből rekonstruálja az összeget
- Milyen titokmegosztó algoritmus kell?
  - Robusztus, hatékony, olcsó, egyszerű, és ahol a fenti aggregálás megoldható

K-ból k titokmegosztás véletlen összeggel

- A titok $s \in F$.
- k-1 véletlen számot húzunk F-ből: ezek $s_2,\ldots,s_k$.
- $s_1=s-s_2-\ldots-s_k$.
- A titok részei $s_1,\ldots,s_k$, ezeket osztjuk szét
- Rekonstruálás: mind a k részösszeg kell, $s=s_1+\ldots s_k$.
- Vannak n-ből k módszerek, pl Shamir módszere (polinomok konstans együtthatója)

Alkalmazás a hatványiterációban

- Az x_j csúcs a $w_{ij}x_j$ értékhez tartozó $s_2,..,s_k$ értéket elküldi k-1 be-szomszédnak
- Az x_j csúcs a megkapott titokrészleteket összeadja és elküldi x_i-nek
- Az x_i csúcs összeadja az összes bejövő üzenetet, ami a keresett összeget adja

Tulajdonságai

- Biztonságos a „semi-honest adversary” modellben
- Üzenetek száma O(kn+k) egy csúcs egy iterációjában
- Véletlen részek maradhatnak több iteráción át (O(nk/m+n) ha m iteráción át)
- K lehet különböző minden be-szomszédra
- Kaotikus iterációt is lehetővé tesz
  - A hibatűréshez még számos részlet kell amit most nem tárgylunk

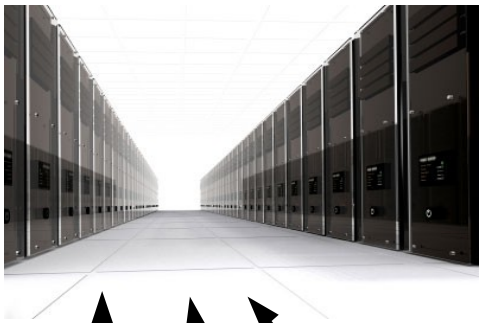
Magánélet védelmének problémái

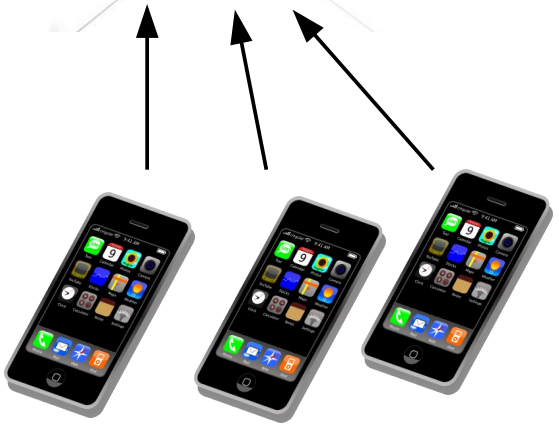
Lekérdezés-kimenetek publikálása

Lekérdezések végrehajtása

Differenciál-magánélet

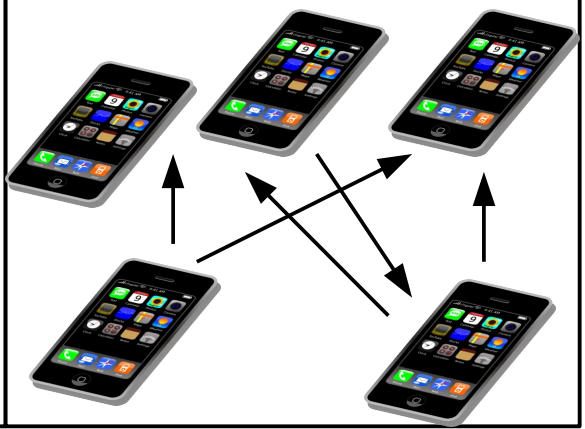
Van megbízható központ





Magánéletvédő algoritmusok

Nincs megbízható központ

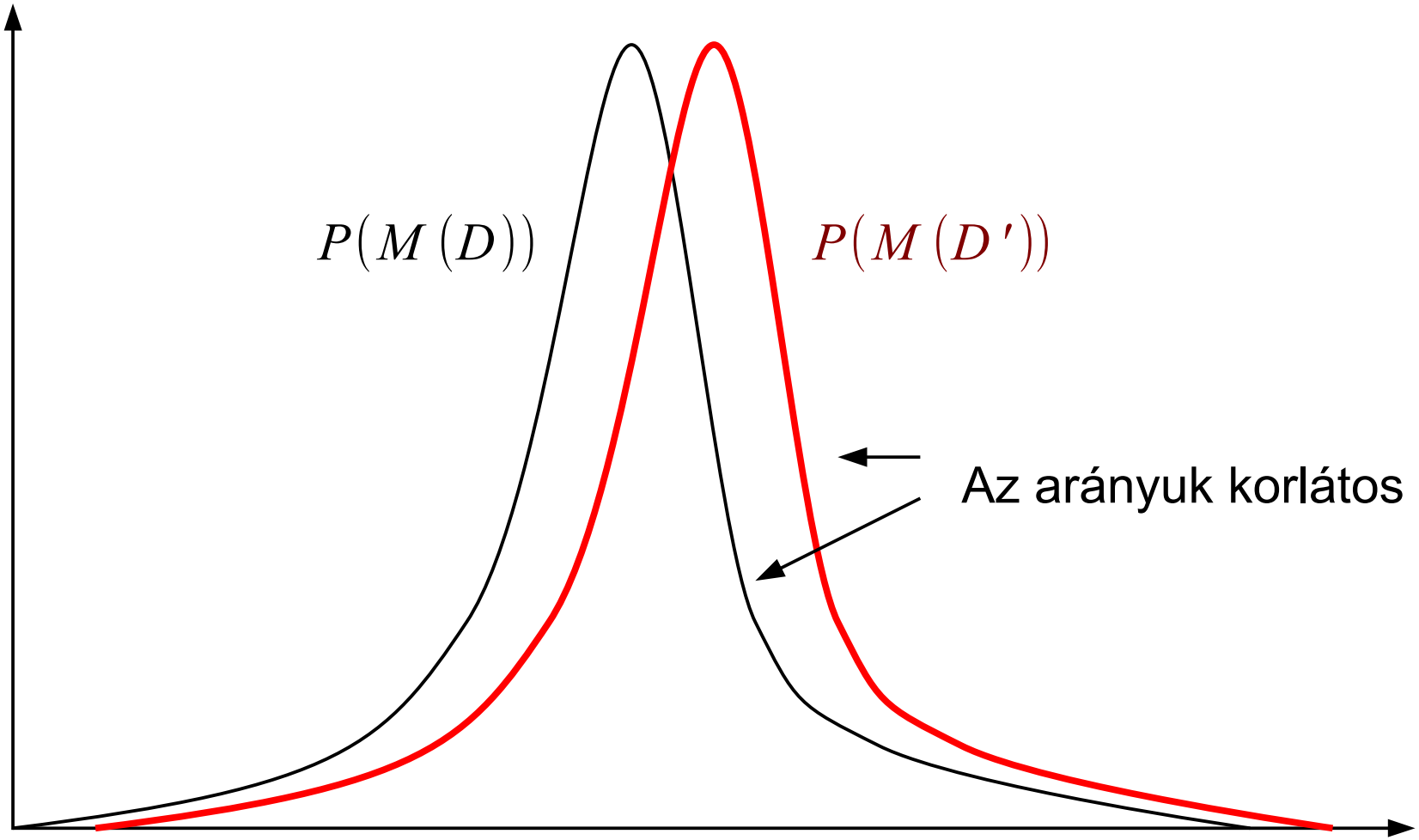


differenciál-magánélet

- Rögzítsünk egy D adatbázist
- Legyen M(D) a lekérdezés kimenete
  - M(D) véletlen változó, a véletlen bitek a az M() függvényből jönnek, D konstans
- Legyen D' olyan adatbázis, amely D-től csak egy rekordban különbözik
- M ε-differenciáltan magán (ε-differentially private), ha

$$P(M(D) \in S) \leqslant \exp(\varepsilon) P(M(D') \in S)$$

differenciál-magánélet



Kompozicionalitás

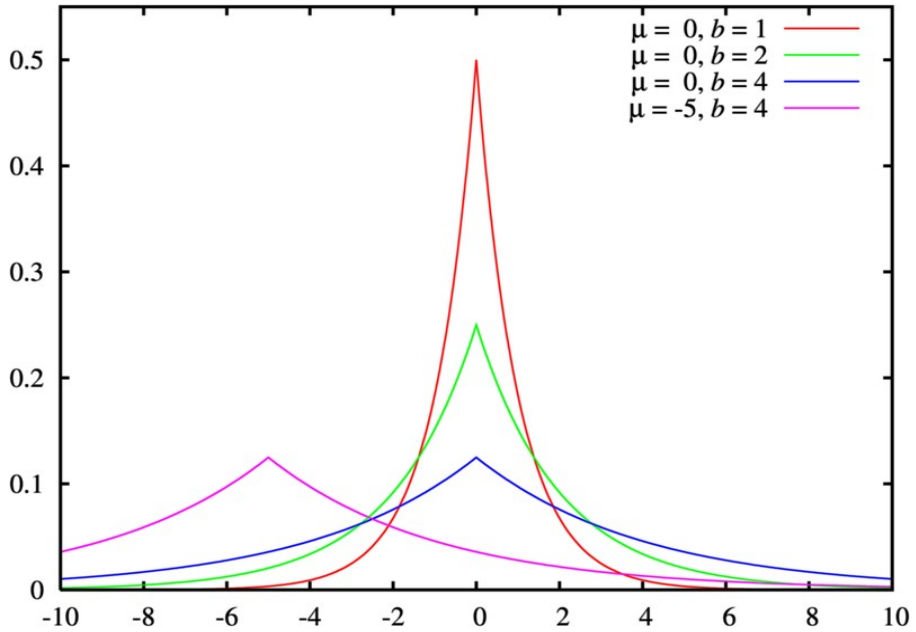
- Ha M_1 és M_2 ε_1- illetve ε_2-differenciáltan magán és független, akkor M_1 és M_2 publikálása $\varepsilon_1+\varepsilon_2$-differenciáltan magán
  - „privacy budget”: vagy véges lekérdezés, vagy növekvő pontatlanság
  - Ha a budget kimerül, több lekérdezés nem megengedett (az adatot el kell dobni)

A differenciál-magánélet egy megvalósítása

- Globális érzékenység:
- Laplace eloszlás:
- Adding noise to deterministic query (Y ~ Laplace(0,Δg/ε)):

$$\Delta g = \max_{D, D' \text{szomszéd}} |g(D) - g(D')|$$

$$f(x|\mu, \beta) = \frac{1}{2b} \exp\left(\frac{-|x-\mu|}{b}\right)$$



$$M(D) = Y + g(D)$$

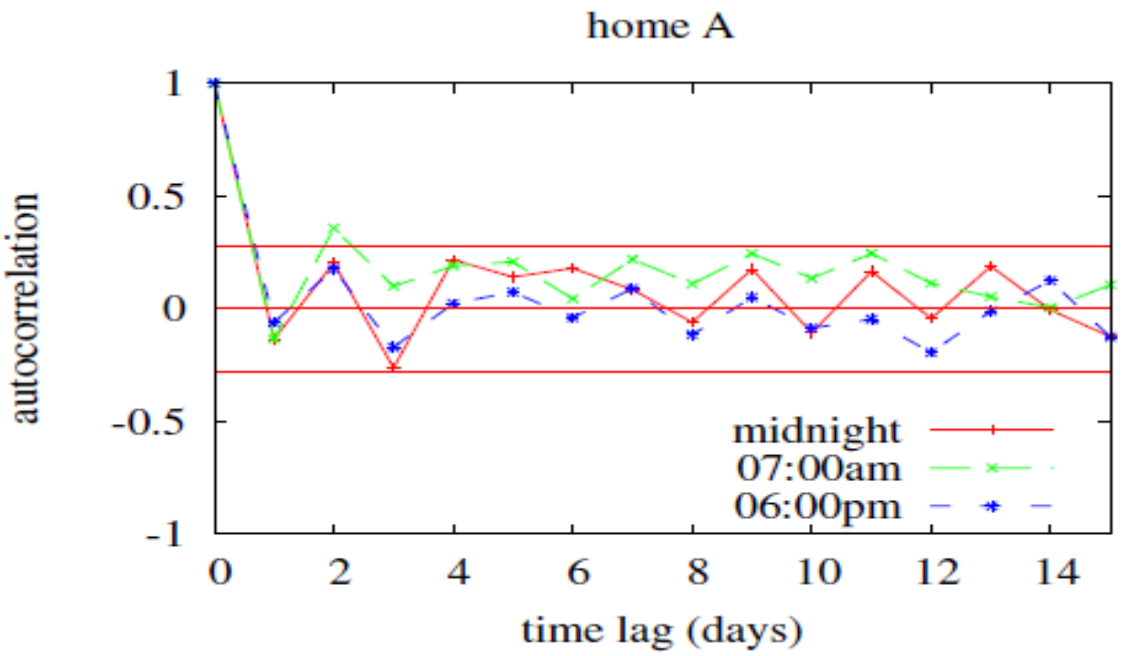
Okos mérőórák

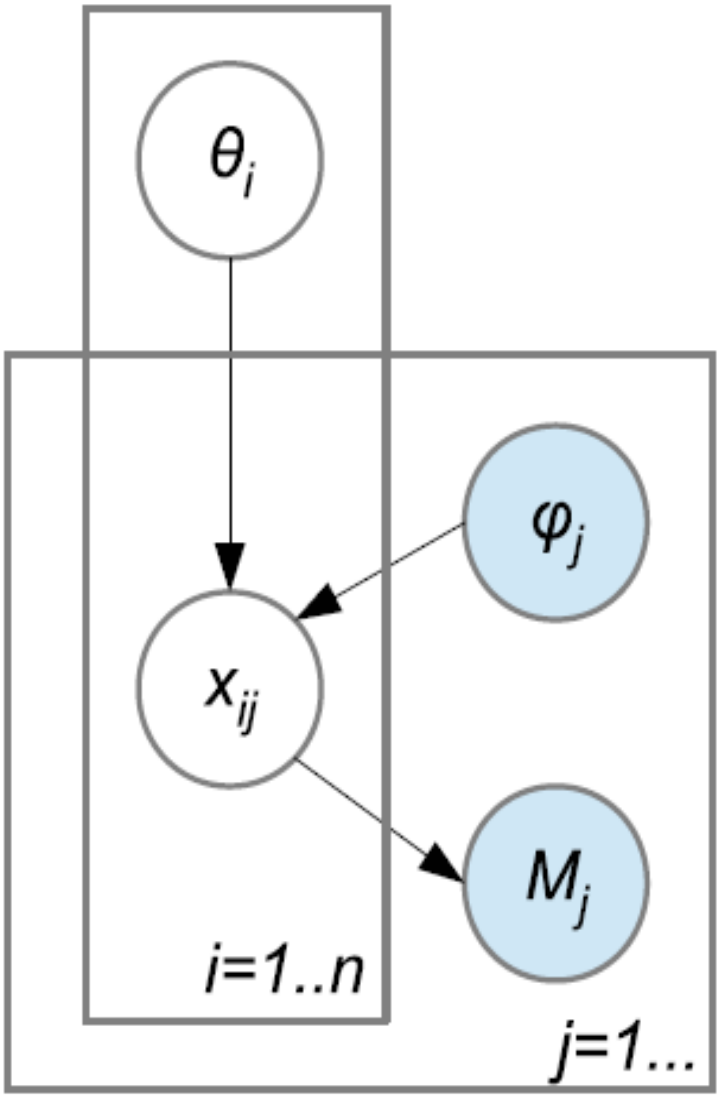
- Okos mérőórák hálózata
- Feladat:
  - Összfogyasztás folyamatos követése
  - Fogyasztás vezérlése az árak manipulációjával
- Probléma:
  - Nem bízunk a szolgáltatóban (sem)
  - A folyamatos megfigyelés során a fogyasztás eloszlásának statikus paraméterei kiderülnek, még ha minden mérés differenciálisan magán is

$$M(D) = Y + \sum x_i$$

Valószínűségi modell

- Xij: az i. óra j. időpontban
- Feltesszük, hogy az időben szomszédos mérések függetlenek!





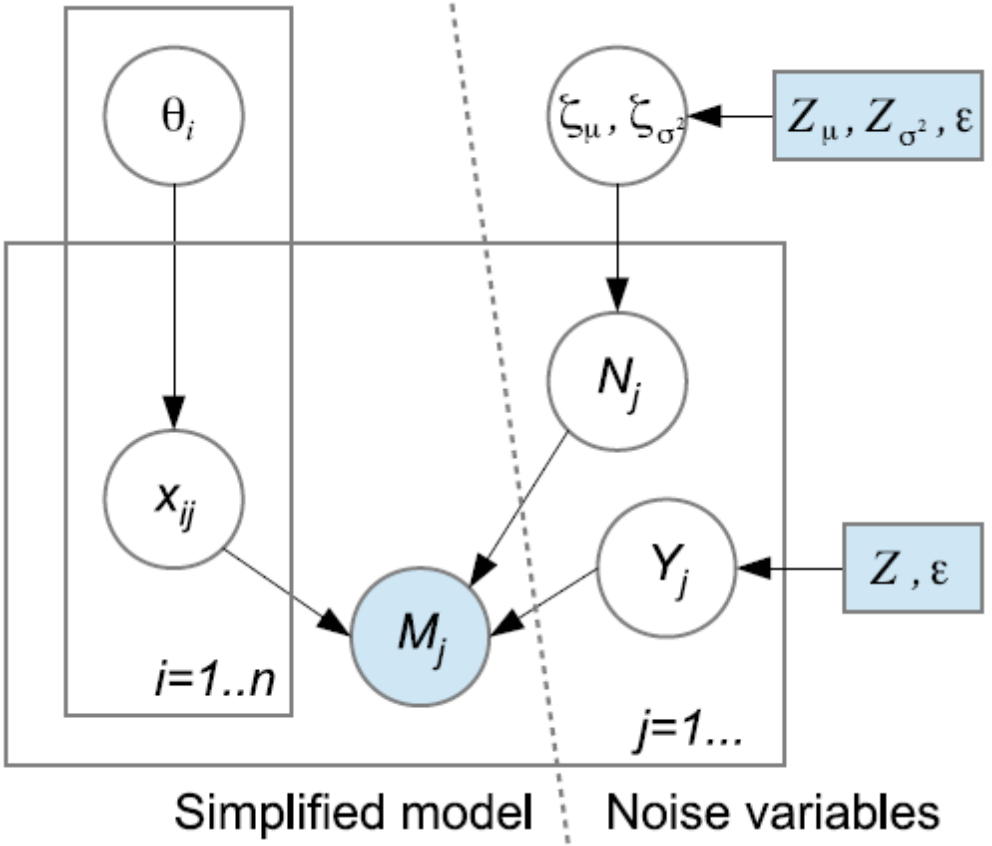
Eloszlási differenciál-magánélet

- Az eloszlásparaméterek felett értelmezzük a differenciál-magánélet fogalmát
- A konkrét lekérdezés helyett végtelen számú lekérdezést (illetve a lekérdezés eloszlását) vesszük
- A korábbi valószínűségi modellt tesszük fel

$$P(\theta_1, \ldots, \theta_n | \, (M_j)_{j=1}^{\infty}) \leq P(\theta'_1, \ldots, \theta'_n | \, (M_j)_{j=1}^{\infty}) \cdot \exp(\varepsilon)$$

Normális eloszlás

- Ha x_{ij} normális eloszlású, az összeg is az
  - Stabil eloszlásokra általában igaz
- Globális érzékenység a paraméterek terében

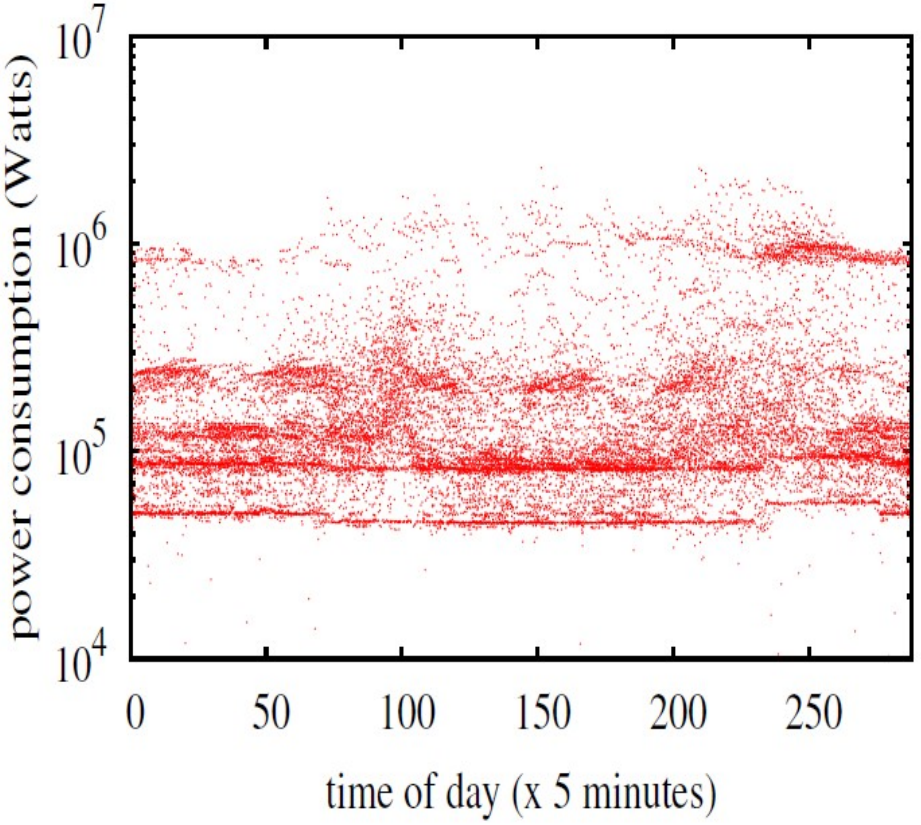


$$N_j \sim \mathcal{N}(\zeta_\mu, \zeta_{\sigma^2})$$

$$N_j + \sum_{i=1}^{n} x_{ij} \sim \mathcal{N}(\zeta_\mu + \sum_{i=1}^{n} \mu_i, \zeta_{\sigma^2} + \sum_{i=1}^{n} \sigma_i^2)$$

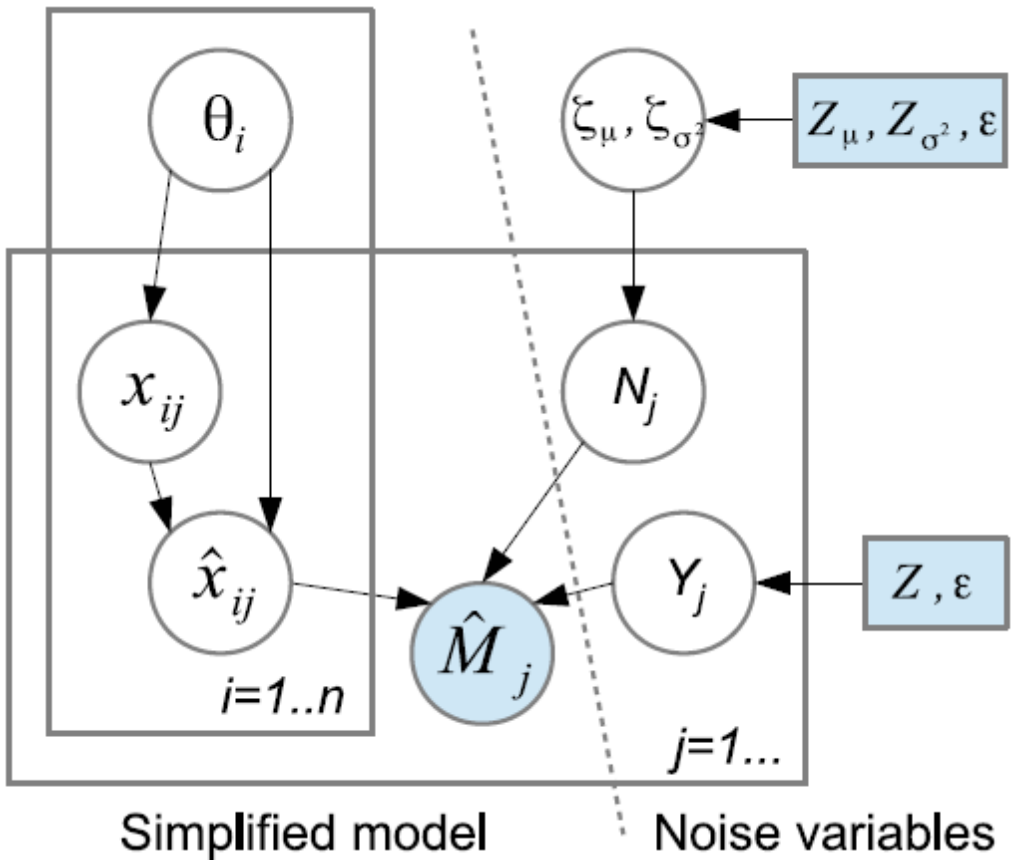
Normális eloszlás

- De x_{ij} nem normális eloszlású
- Az eloszlást transzformálhatjuk
  - Normálissá
  - Vagy Bernoullivá



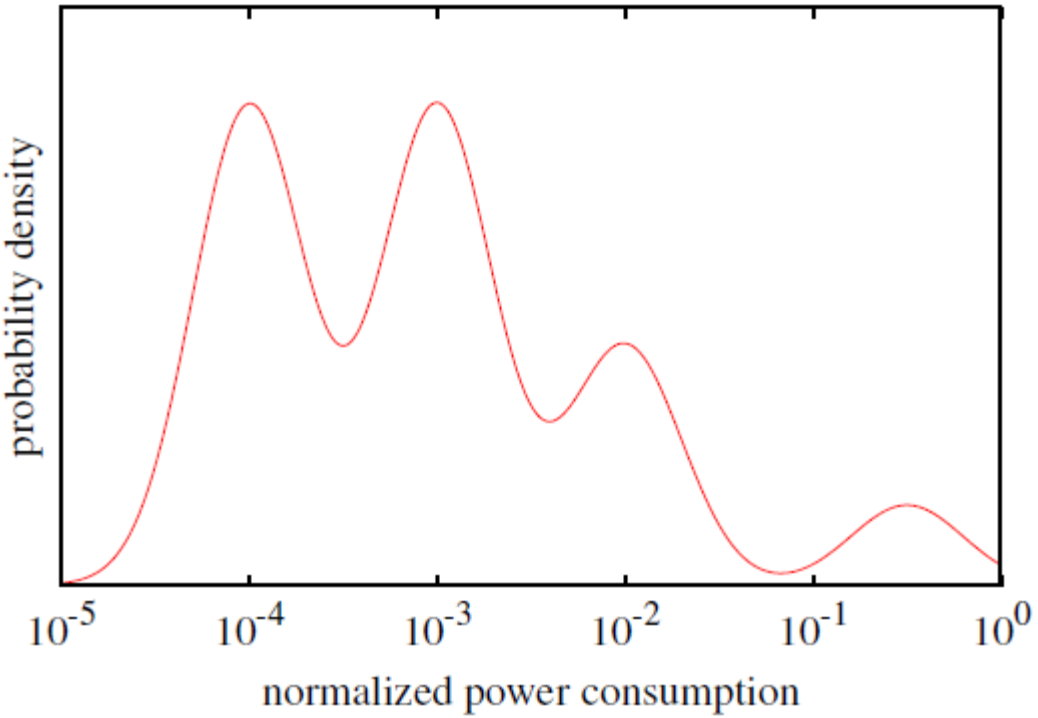
Transzformáció normális eloszlássá

- Ha θ_i ismert lokálisan, az új normális eloszlású változót x_{ij} eloszlásával azonos várható értékkel és szórással, x_{ij}-vel azonos kvantilist reprezentálva számoljuk



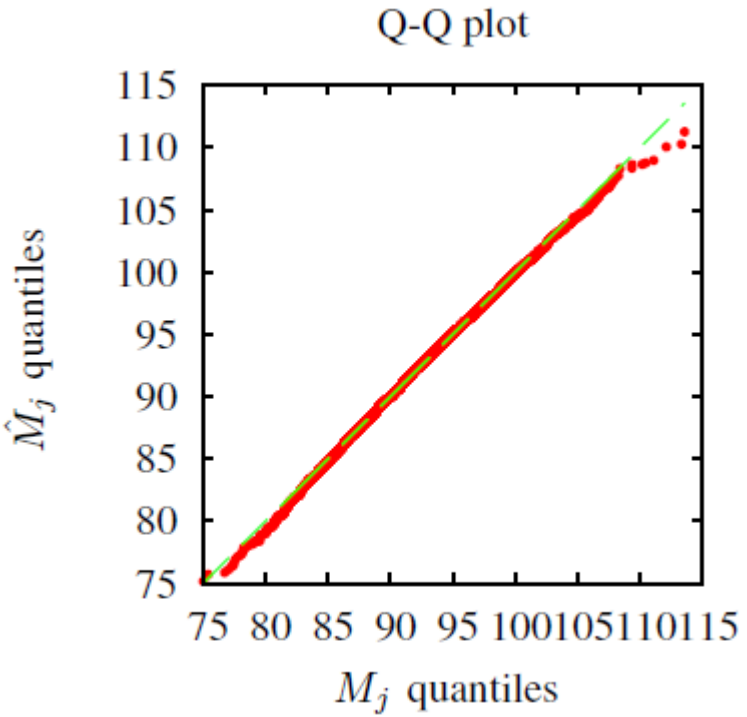
$$N_j \sim \mathcal{N}(\zeta_\mu, \zeta_{\sigma^2})$$

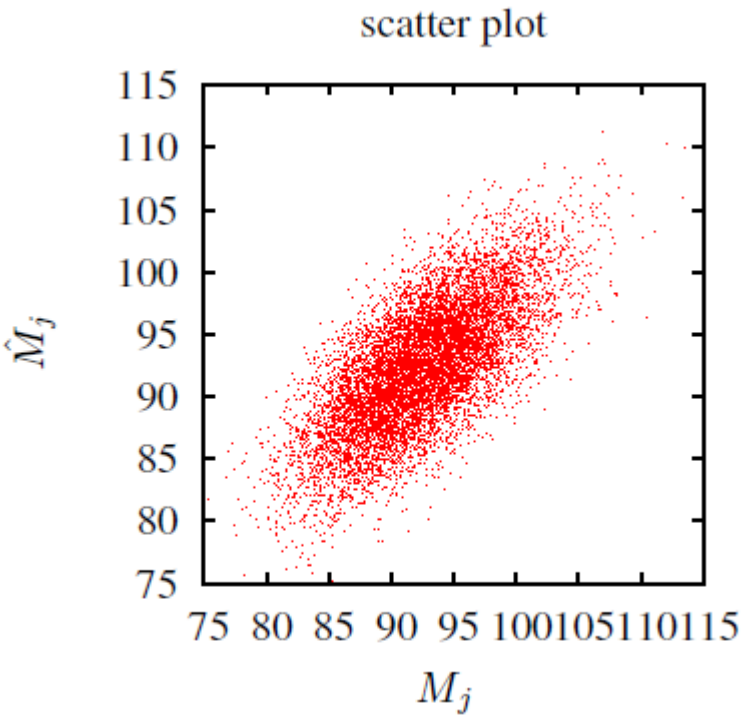
$$N_j + \sum_{i=1}^{n} x_{ij} \sim \mathcal{N}(\zeta_\mu + \sum_{i=1}^{n} \mu_i, \zeta_{\sigma^2} + \sum_{i=1}^{n} \sigma_i^2)$$



Eredeti eloszlás

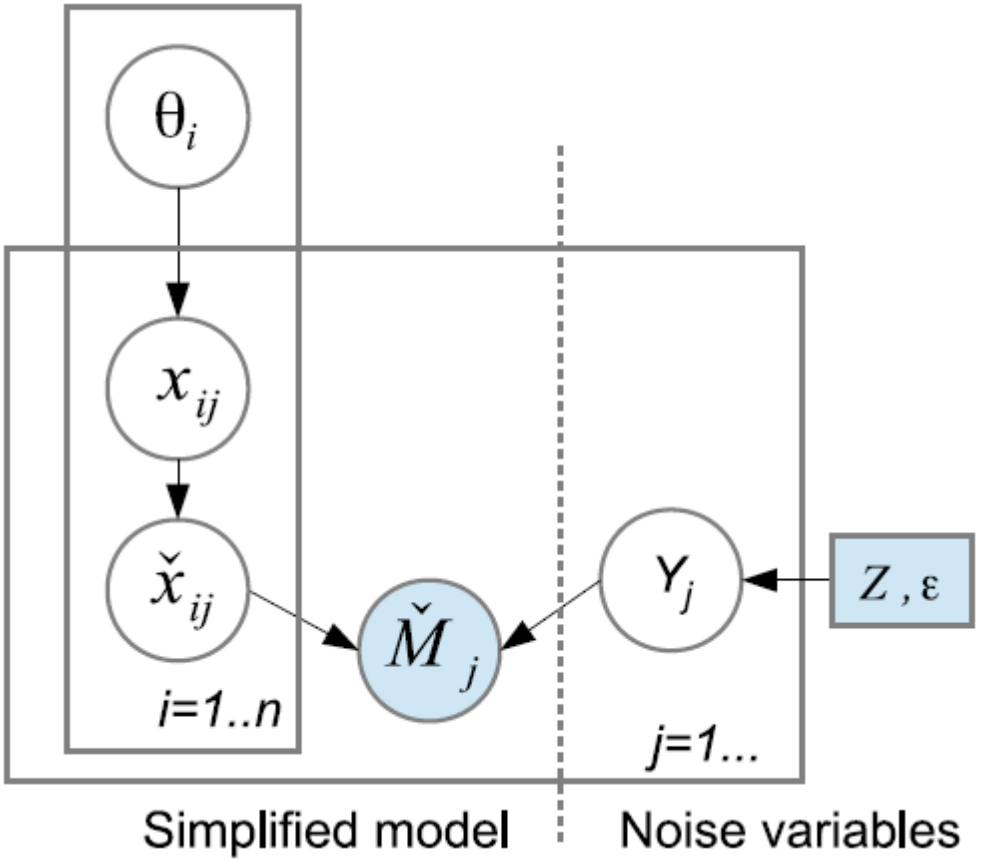
Viszony a transzformált eloszlással





Transzformálás Bernoulli eloszlássá

- Θ_i-t nem kell ismerni
- az új Bernoulli eloszlású változó x_{ij}-vel azonos várható értékű
- A várható érték nem védett



$$\check{x}_{ij} \sim \text{Bernoulli}(x_{ij})$$

Hozzáadott zaj

- Ez a módszer több zajt eredményez
- Nem szükségképpen extra zaj!
- Pl. a korábbi eloszlásra alkalmazva n=1000 mérőórára 1.9-szeresére növekszik a zaj

$$\mathrm{stdev}\left[\sum_{i=1}^{n} \check{x}_{ij}\right] = \sqrt{\sum_{i=1}^{n} x_{ij}(1-x_{ij})} \leq \frac{\sqrt{n}}{2}$$

Megjegyzések

- Felhasznált „fekete doboz” komponensek
  - Elosztott biztonságos összeglekérdezés
  - Elosztott biztonságos zaj generálás
- Fontos feltevés: időbeli függetlenség
  - Egy órát csak elegendő késleltetés után olvasunk le (mintavételezés)
  - Ez segíti a vezérlési hurok stabilitását is

Konklúziók

- Nem elég anonimizálni
- Nem elég, ha csak aggregált információt adunk közre
  - Pedig már ezt sem triviális megvalósítani (privacy preserving data mining)
- Jelenlegi módszerek nagyon jelentősen korlátozzák a felhasználást (privacy budget)
  - Valószínűleg túlbiztosítottak
- A folyamatos megfigyelés nagyon fontos de további problémákat vet fel
  - A hozzájárulásunk erre vonatkozik

Publikációk

- Márk Jelasity, Geoffrey Canright, and Kenth Engø-Monsen. Asynchronous distributed power iteration with gossip-based normalization. In Anne-Marie Kermarrec, Luc Bougé, and Thierry Priol, editors, Euro-Par 2007, volume 4641 of Lecture Notes in Computer Science, pages 514–525. Springer-Verlag, 2007. (doi:10.1007/978-3-540-74466-5_55)
- Juan A. M. Naranjo, Leocadio G. Casado, and Márk Jelasity. Asynchronous privacy-preserving iterative computation on peer-to-peer networks. Computing, 94(8–10):763–782, 2012. (doi:10.1007/s00607-012-0200-5)
- Márk Jelasity and Kenneth P. Birman. Distributional Differential Privacy for Large-Scale Smart Metering. submitted