O geometriji diferenciirane zasebnosti

Luka Lodrant
Mentor: doc. dr. Aljoša Peperko

17. april 2018

Fakulteta za matematiko in fiziko

Pregled

Predstavitev problema

Priprava okolja

Končni rezultat

Predstavitev problema

Anonimizacija podatkov

- Velike statistične podatkovne baze (big data)
- Primeri:
  - Google
  - Facebook
  - Državne statistike
- Zagotavljanje zasebnosti
- Politične težave
- Rešitev: anonimizacija podatkov

Diferenciirana zasebnost

- Cynthia Dwork - 2006
- Frank McSherry, Kobbi Nissim in Adam D. Smith
  - objava podatkov brez zasebnih informacij je nemogoča
  - z majhnim številom poizvedb je bazo mogoče poustvariti
- matematični model za analizo zasebnosti
- definira zasebnost kot nekaj merljivega

Priprava okolja

Podatki

Ime	Opravil UNM kviz
Anica	0
Boštjan	1
Ciril	1
Domen	0
Ester	1

Predstavitev podatkov

Definicija

Podatkovno bazo predstavimo kot vektor v $\mathbb{R}^n$

$$A = (0, 1, 1, 0, 1)$$

Definicija

Poizvedba je linearna kombinacija členov v podatkovni bazi

$$p(A) = 1 * x_0 + 0 * x_1 + ...$$

Definicija

Mehanizem zasebnosti je naključen algoritem, ki kot vhodni podatek vzame podatkovno bazo in poizvedbo in vrne rezultat v obliki realnega števila.

Diferenciirana zasebnost

Definicija
Za ϵ-zaseben mehanizem M in dve podatkovni bazi, ki se razlikujeta je v enem členu $D1$ in $D2$ velja:

$$Pr[M(D1) \in S] \leq exp(x) \times Pr[M(D2) \in S]$$

kjer je S katerakoli podmnožica slike M.

Problem
Vsak zasebnostni mehanizem povzroči napako v izhodnih podatkih. Ali je mogoče podati dobro spodnjo in zgornjo mejo te napake v odvisnosti od željene zasebnosti?

Končni rezultat

Končni rezultat

Z uporabo geometrijskih lastnosti naše predstavitve podatkov lahko pridemo do sledeče spodnje meje.

Izrek
Naj bo $\epsilon > 0$ in $F : \mathbb{R}^n \to \mathbb{R}^d$ linearna preslikava, ki predstavlja poizvedbo. Potem ima vsak ϵ-zaseben mehanizem M napako v l_1 normi vsaj $\Omega(min(d * \sqrt{d}/\epsilon, d\sqrt{log(n/d)}/\epsilon))$