Bom začel s predstavitvijo… Danes bom predstavil svojo diplomsko nalogo z naslovom »Splošna definicija diferencirane zasebnosti«. Najprej na hitro ponovimo, kaj sploh je »diferencirana zasebnost«. Gre za matematično definicijo oz. koncept zasebnosti pri javni objavi ter rudarjenju podatkov. Vemo, da živimo v dobi podatkov; podjetja ter ustanove zbirajo ogromne količine podatkov o svojih uporabnikih in postavlja se vprašanje kako lahko objavljajo in pridobivajo te podatke na načine, ki ne bi ogrožali zasebnosti posameznikov. Za lažje razumevanje samo še kratek primer: ameriške bolnišnice so v preteklem desetletju objavile podatke o svojih pacientih (želeli so omogočiti dostop raziskovalcem v namene statističnih analiz), pri tem pa so uporabile ti. metodo anonimizacije (odstranile so vse eksplicitne identifikatorje, kot so ime, emšo itd.). Znanstevniki so te podatki združili z bazo iz volilnega sistema in na ta način jim je uspelo identificirati večino pacientov v zdravstevni bazi. Podobnih primerov je bilo še precej, kar je pokazalo, da uporabljene metode niso dovolj dobre in zato se je pojavila potreba po bo sistematičnem matematičnem pristopu k doseganju zasebnosti. In odgovor na to je diferencirana zasebnost.

V prvem delu diplomske naloge opišem model, ki omogoča obravnavo različnih vrst podatkov hkrati (pogosto se namreč diferencirana zasebnost definira le za eno vrsto podatkov, npr numerične). O tem sem velik govoril že na dolgi predstavitvi, zato bom ta del danes samo na hitro preletel. V drugem delu pa se bolj posvetimo funkcijski podatkom, konkretno za primer jedrne cenilke za gostoto skonstruiramo diferencirano zaseben mehanizem.

SOSEDNJE BAZE:

Želimo razkriti lastnosti celotne populacije pri tem pa ohraniti zasebnost konkretnega posameznika.

POIZVEDBA:

Ne objavi se javno več celotna baza, ampak se podatke varno shrani na nekem serverju in se omogoči znanstvenikom dostop samo prek ti. poizvedb.

ODZIVNI MEHANIZEM:

Izkaže se, da smo izpostavljeni kršenju zasebnosti, tudi v primeru determinističnih poizvedb (npr. v primeru da ima napadalec dodatne informacije, kot v zgledu iz uvoda, kjer so uporabili dodatno podatkovno bazo), zato se moramo vsake toliko časa »zlagati« za ohranitev zasebnosti. Pravi odgovor podamo le z neko verjetnostjo, zato govorimo o prehodu iz determinističnih poizvedb na probabilistične.