Anonimizacija podataka

Vrljić, Andrea

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Andrea Vrljić

ANONIMIZACIJA PODATAKA

Diplomski rad

Voditelj rada:
prof. dr. sc. Robert Manger

Zagreb, 2020.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

„Ne boj se jer sam ja s tobom! Ne gledaj plašljivo naokolo, jer sam ja tvoj Bog!
Ja te krijepim. Ja ti pomažem. Ja te podupirem svojom jakom desnicom.“ Iz 41,10

Želim se zahvaliti svojim roditeljima, majci Danijeli i ocu Miroslavu koji su prihvatili moj život i svojim radom, ljubavlju i podrškom požrtvovno me pratili tijekom mog studija i života. Hvala braći Anti i Luki te sestri Anamariji na njihovom strpljenju i podršci.

Hvala mojoj obitelji salezijancima i svim oratorijancima, prijateljima, na svakom osmijehu, zagrljaju, saslušanju i molitvi kojom su me pratili i nosili na svojim krilima preko svake prepreke. Hvala dragom Bogu koji mi je sve to omogućio darovavši mi talente i snagu za završetak ovog studija, a početak novog razdoblja života.

Sadržaj

Uvod

Globalnim umrežavanjem društva nameće se sve veća potreba za širenjem i dijeljenjem osobnih podataka. Sve je više povijesnih podataka koji su danas dostupni u elektroničkom obliku. Povezivanjem podataka otkrivaju se dodatne informacije o osobi ili organizaciji.S ciljem zaštite osobnih podataka, često se uklanjaju ili šifriraju određeni identifikatori kao što je ime, adresa, broj mobitela. Međutim, druge karakteristike, koje nazivamo kvazi-identifikatori, često mogu biti povezani sa javno dostpunim informacijama i otkriti identitet. Povećanom zloporabom osobnih podataka pojavila se veća potreba za anonimizacijom podataka. Ovo su neki primjeri zloporabe osobnih podataka:

- Poslovne firme istraživačkog sektora prijavile su 70% sigurnosnih incidenata i 80% prijetnji koje dolaze iz unutrašnjosti firme, a 65% ih je neotkriveno.
- Vodeća zdravstvena ustanova u Europi u razdoblju od 3 godine pretrpjela je 899 prekršaja i zloporabe osobnih podataka klijenata, a najveća prijetnja sigurnosti podataka je njihovo osoblje.
- Međunarodna azijska banka kompromitirala je podatke 20 000 svojih korisnika.

Rast incidenata zlouporabe osobnih podataka rezultiralo je mnoštvom propisa o zaštiti privatnosti od strane vlada diljem svijeta kojima se štite podaci od neovlaštenih osoba, odnosno programera, testera i sl. Anonimizacijom podataka osigurava se neupotrebljivost ukradenih podataka.

U prvom poglavlju prikazan je rizik objavljivanja podataka te opisan način povezivanja zapisa. U drugom poglavlju, predstavljen je model k-anonimnosti koji je temelj sigurnosti anonimnosti pojedinca. Uz definiciju k-anonimnost, definirat ćemo i druge pojmove kao što je klasa ekvivalencije, kvazi-identifikator, pokazat ćemo korištenje generalizacije te uvesti pojam minimalne generalizacije koji pospješuje k-anonimnost. Treće poglavlje predstavlja proširenje modela k-anonimnosti do l-raznolikosti i t-bliskosti. Konkretniji primjeri i načini anonimizacije podataka prikazani su u četvrtom poglavlju. Sljedeća tri poglavlja donose nam djelomično maskiranje, i detaljnije o maskiranju s obzirom na okolinu podataka, dodatne uvjete, veze kolona te specijalizirane tehnike maskiranja iz [8]. U posljednjem poglavlju imamo primjer anonimizacije koristeći ARX alat.

Poglavlje 1

Rizik objavljivanja podataka

Prilikom objavljivanja podataka, prikupljač podataka mora garantirati da nema osjetljivih podataka i informacija o pojedincu koji se mogu razotkriti. Postoje dva tipa razotkrivanja podataka, prvi je razotkrivanje identiteta, a drugi je razotkrivanje osobina. *Razotkrivanje identiteta* je tip razotkrivanja kojim se krši anonimnost osobe, njen osobni identite. Do ovoga dolazi kada neovlaštena osoba može iz objavljenih podataka doći do originalnih podataka osobe. *Razotkrivanje osobina* je kršenja privatnosti povjerljivih informacija pojedinca. Ova dva tipa kršenja privatnosti su neovisna jedan o drugome. Ako dođe do razotrkivanja identiteta osobe,a povjerljive informacije su maskirane ne mora nužno značiti da je došlo i do razotkrivanja osobina. Isto tako je moguće da dođe do razotkrivanja osobine pojedinca, ali ne i razotkrivanja identiteta osobe.

Kako bi dočarali problem objavljivanja podataka, način na koji se vrlo lako može povezati i razotkriti identitet koristit ćemo prikaz dvije vanjske datoteke dane tablicom 1.1, prva tablica pod (1.1a) je iz baze medicinskih podataka, a druga tablica (1.1b) prikazuje glasačku listu. Iz tablice medicinske baze imamo objavljene atribute kao što je poštanski broj, datum rođenja, spol, bračni status koji također mogu biti objavljeni u drugoj tablici sa osobnim identifikacijskim podacima. Takav primjer je dan drugom tablicom, glasačke liste koja sadrži poštanski broj, datum rođenja i spol kao i prethodna tablica, uz osobne podatke. Povezivanjem tih dviju tablica možemo dobiti identitet osobe.

(a) Anonimizirani medicinski podaci

SSN	Ime	Rasa	Rođenje	Spol	Pošta	Bračni status	Zdravlje
		azijac	09/27/64	Žensko	94139	Rastavljen/a	Hipertenzija
		azijac	09/30/64	Žensko	94139	Rastavljen/a	Pretilost
		azijac	04/18/64	Muško	94139	Vjenčan/a	Bol u prsima
		azijac	04/15/64	Muško	94139	Vjenčan/a	Pretilost
		crnac	03/13/63	Muško	94138	Vjenčan/a	Hipertenzija
		crnac	03/18/63	Muško	94138	Vjenčan/a	Plitko disanje
		crnac	09/13/64	Žensko	94141	Vjenčan/a	Plitko disanje
		crnac	09/07/64	Žensko	94141	Vjenčan/a	Pretilost
		bijelac	05/14/61	Muško	94138	Slobodan/na	Bol u prsima
		bijelac	05/08/61	Muško	94138	Slobodan/na	Pretilost
		bijelac	**09/15/61**	**Žensko**	**94142**	**Udovac/ica**	**Plitko disanje**

(b) Glasačka lista

Ime	Adresa	Grad	Pošta	Dob	Spol	Stranka
.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....
Sue Carlson	**900 Market St.**	**San Francisco**	**94142**	**9/15/61**	**Žensko**	**Demokrat**
.....	.....	.....	.....	.....	.....	.....

Tablica 1.1: Povezivanje podataka vanjskih datoteka

1.1 Procjena rizika otkrivanja - povezivanje zapisa

Kontrola zaštite podataka koristi algoritam ili više algoritama za povezivanja zapisa kako bi povezala zapise iz anonimiziranog skupa podataka sa zapisima u originalnom skupu podataka. Poznavanjem stvarnog podudaranja originalnih i zaštićenih (anonimiziranih) podataka, kontrola zaštite određuje točan postotak povezanih parova te procjenjuje broj mogućih razotkrivanja neovlašenom korisniku. Ako je broj razotkrivanja prevelik, skup podataka mora biti više anonimiziran kako bi se mogao objaviti.

Jednostavno povezivanje podataka temelji se na povezivanju vrijednosti zajedničkih osobina. Osobine koje se pojavljuju u oba skupa podataka istovremeno se uspoređuju, odjednom. Par zapisa se podudara ako osobina koja se odnosi na oba zapisa ima jednake vrijednosti i ako postoje samo dva zapisa koja dijele tu vrijednost. Par zapisa se ne po-

dudara ako se razlikuju u vrijednosti osobine ili ako je više parova zapisa koji dijele istu vrijednost osobine. Za dani par zapisa x i y koristimo funkciju sličnosti *sim* koja vraća vrijednosti između 0 i 1 koje predstavljaju stupanj sličnosti između zapisa x i y.

Smatramo da:

- $sim(x, y) \in [0, 1]$ daje stupanj sličnosti para u rasponu $[0, 1]$, tako da 1 znači potpuno jednaki, a 0 znači potpuno različiti.
- $sim(x, y) = 1 \Leftrightarrow x = y$. Vrijednost funkcije sličnosti je 1 ako i samo ako su zapisi jednaki.

Postoji više načina računanja sličnosti između osobina zapisa i neki od načina računanja sličnosti koriste računanje broja umetanja, brisanja ili zamjena koje je potrebno napraviti kako bi iz jednog zapisa došli do drugoga.

Prag povezivanja zapisa

Prag povezivanja zapisa je prilagodba funkcije sličnosti na podudaranje zapisa. Umjesto da kažemo dva zapisa se podudaraju kada se sve vrijednosti zajedničke osobine podudaraju, kažemo da se dva zapisa podudaraju kada su dovoljno slična. Prag nam je potreban kako bi mogli odrediti kada su dva zapisa dovoljno slična. Kako bi odredili sličnost zapisa između podudaraju se i ne podudaraju se, koristimo varijablu vrijednosti praga t i tada slijedi:

- $sim(x, y) \geq t$: par zapisa x i y se podudara.
- $sim(x, y) < t$: par zapisa x i y se ne podudara.

Korištenje oštrog praga razlikovanja između podudaranja i ne podudaranja može biti preograničeno. Vjerojatnost pogreške procjene sličnosti može biti velika kada se vrijednost funkcije sličnosti nalazi blizu vrijednosti praga. Kako bi izbjegli pogrešku, podijelit ćemo klasifikaciju na tri skupine, podudaranje, nepodudaranje i moguće podudaranje. U skupini moguće podudaranje nalaze se parovi zapisa koji nisu čisto klasificirani kao podudaranje, a nisu klasificirani ni kao nepodudaranje. Za ovakav pristup potrebne su nam dvije pragovne varijable t_u i t_d gdje je $t_u > t_d$ i tada vrijedi:

- $sim(x, y) \geq t_u$: par zapisa x i y se podudaraju.
- $t_d \leq sim(x, y) \leq t_u$: moguće podudaranje para zapisa x i y.
- $sim(x, y) < t_d$: par zapisa x i y se ne podudaraju.

Postoji vektor funkcija sličnosti koje koristimo u pravilu povezivanja podataka što je bolje nego samo jednu osnovnu funkciju.

Vjerojatnost povezivanja zapisa

Prilikom klasificiranja zapisa u dane skupine, želimo smanjiti broj pogrešno raspoređenih parova. Promatrat ćemo dva skupa podataka X i Y koji u pravilu nisu dijsunktni, što znači da se određen zapis može nalaziti i u X i u Y. Za originalan podatak $x \in X$ pridružen mu je zapis $\alpha(x)$ i za $y \in Y$ odgovara zapis $\beta(y)$. Slijedi, ako $x = y$ onda može biti $\alpha(x) \neq \beta(y)$, i isto tako ako je $x \neq y$, može biti $\alpha(x) = \beta(y)$.

Kako bi povezali podatke iz skupova koristimo $\alpha(X)$ i $\beta(Y)$, uzimajući u obzir da je svaki par zapisa dan sa $\alpha(X) \times \beta(Y)$ i pripada jednoj od mogućnosti: podudaranje, nepodudaranje, moguće podudaranja. Skup $\alpha(X) \times \beta(Y)$ možemo podijeliti na dvije skupine, skupinu stvarnih podudaranja $M = \{(\alpha(x), \beta(y)) : x = y\}$, i skup stvarnih nepodudaranja $U = \{(\alpha(x), \beta(y)) : x \neq y\}$. Prvi korak kako bi otkrili trebaju li dva zapisa iz $\alpha(X) \times \beta(Y)$ biti povezana je izračunati vektor funkcije sličnosti. Definiramo vektor sličnosti za $\alpha(X)$ i $\beta(Y)$ sa

$$\gamma(\alpha(x), \beta(y)) = (\gamma^1(\alpha(x), \beta(y)), \ldots, \gamma^k(\alpha(x), \beta(y))). \tag{1.1}$$

Skup mogućih događaja funkcije γ naziva se prostor usporedbe i označavamo ga sa Γ. Za dani događaj $\gamma_0 \in \Gamma$ zanimaju nas dva uvjeta vjerojatnosti. Prvo je vjerojatnost γ_0 uz uvjet stvarnog podudaranja $P(\gamma_0 \mid M)$ te vjerojatnost γ_0 uz uvjet stvarnog nepodudaranja $P(\gamma_0 \mid U)$. S danim vjerojatnostima možemo izračunati njihov omjer za dani γ_0:

$$R(\gamma_0) = \frac{P(\gamma_0 \mid M)}{P(\gamma_0 \mid U)}. \tag{1.2}$$

Prilikom povezivanja promatramo funkciju $\gamma(\alpha(x), \beta(y))$, moramo odlučiti je li $((\alpha(x), \beta(y)) \in M$ ili $(\alpha(x), \beta(y)) \in U$, ali dopuštamo i opredjeljenje za moguće podudaranje. Pravilo povezivanja je korelacija prostora usporedbe Γ sa vjerojatnosnom distribucijom mogućih klasifikacija, gdje je podudaranje u oznaci L, nepodudaranje N, moguće podudaranje C.

$$\mathcal{L} : \Gamma \to \{(p_L, p_N, p_C) \in [0,1]^3 : p_L + p_N + p_C = 1\}. \tag{1.3}$$

Pravilo povezivanja možemo označiti sa dvije vjerojatnosti, vjerojatnošću krivog podudaranja $\mu = P_{\mathcal{L}}(L \mid U)$ ili sa vjerojatnošću krivog ne podudaranja $\lambda = P_{\mathcal{L}}(N \mid M)$, gdje krivo podudaranje znači da povezan par nema zapravo podudaranje, a krivo nepodudaranje znači da ne povezani par ima podudaranje. U [5] je predstavljeno optimalno pravilo povezivanja. Pravilo povezivanja je optimalno u smislu da za maksimalnu toleranciju vjerojatnosti pogreške μ i λ, pravilo ima najmanju vjerojatnost rezultata mogućeg podudaranja. Klasificira svaku moguću sličnost vektora Γ pridružujući jednom izboru podudaranja, nepodudaranja ili mogućeg podudaranja ovisno o omjeru vjerojatnosti. Za dani skup podataka $(\alpha(a), \beta(b))$

i vektor sličnosti $\gamma(\alpha(x), \beta(y))$ izračunata je i dana klasifikacija na slijedeći način:

$$\begin{cases} R(\gamma(\alpha(x), \beta(y))) \geq T_\mu & \rightarrow podudaranje \\ R(\gamma(\alpha(x), \beta(y))) \leq T_\lambda & \rightarrow nepodudaranje \\ T_\lambda < R(\gamma(\alpha(x), \beta(y))) < T_\mu & \rightarrow moguće\ podudaranje \end{cases} \tag{1.4}$$

gdje je sa T_μ i T_λ označena gornja i donja vrijednost praga. Pogreška je dana sa:

$$\mu = \sum_{\gamma \in \Gamma : R(\gamma) \geq T_\mu} P(\gamma \mid U) \tag{1.5}$$

$$\lambda = \sum_{\gamma \in \Gamma : R(\gamma) \leq T_\lambda} P(\gamma \mid M) \tag{1.6}$$

Primjer

Kod vjerojatnosti povezivanja podataka, glavna pretpostavka je da datoteke imaju neke iste atribute. Svaka od danih prikazanih tablica 1.2 sadrži 8 zapisa sa 3 varijable, Ime, Prezime i Dob. Zbog razumijevanja, tablica je prikazana tako da zapisi u istom redu se podudaraju, dok zapisi u različitim redovima se ne podudaraju. Cilj povezivanja zapisa ovog primjera je pronaći sve moguće parove te odrediti koji su došli iz istog reda, odnosno koji odgovaraju podudaranju. Sve parove smo zabilježili $(a, b) \in A \times B$.

Ime^A	Prezime^A	Dob^A	Ime^B	Prezime^B	Dob^B
Joan	Casanoves	19	Joan	Casanovas	19
Pere	Joan	17	Pere	Joan	18
J.M.	Casanovas	35	J.Manel	Casanovas	35
Juan	Garcia	53	Juan	Garcia	53
Ricardo	Garcia	14	Ricard	Garcia	14
Pere	Garcia	18	Pere	Garcia	82
Juan	Garcia	18	Juan	Garcia	18
Ricard	Tanaka	14	Ricard	Tanaka	18

Tablica 1.2: Povezivanje podataka vanjskih datoteka

Skup svih događaja Γ dane tablice je $\Gamma = \{\gamma^1 = 000, \gamma^2 = 001, \gamma^3 = 010, \gamma^4 = 011, \gamma^5 = 100, \gamma^6 = 101, \gamma^7 = 110, \gamma^8 = 111\}$. Klasifikacija parova (a, b) povezana je sa vektorom $\gamma(a, b)$. Prikaz je dan tablicom na slici 1.1.

Tablicom sa slike 1.2 prikazani su parovi te izračunate vrijednosti m_i, u_i, m_i/u_i, $\log(m_i/u_i)$, gdje je $m_i = P(\gamma^i = \gamma(a', b') \mid (a', b') \in M)$ te $u_i = P(\gamma^i = \gamma(a', b') \mid (a', b') \in U)$. Za omjer

se često koristi prirodni logaritam omjera te ćemo ga ovdje iskoristiti u izračunu $\log R(\gamma_i)$. Sada ćemo izračunati pogrešku danim jednadžbama pogreške. Pretpostavimo da su dane gornja i donja vrijednost praga, $T_\mu = 2.5$, a $T_\lambda = 1.5$. Koristeći podatke iz tablice imamo da je:

$$\mu = 0/56 + 1/56 = 1/56 = 0.0178$$
$$\lambda = 0 + 0 + 0 + 1/8 = 0.125$$

Prva vjerojatnost μ odgovara situaciji krivog povezivanja, odnosno par je klasificiran kao par podudaranja, ali ne podudara se. A druga vjerojatnost λ odgovara situaciji kada nismo povezali par, a par ima podudaranje. Prikazane tablice i detalji se mogu pronaći na [11].

$Name^{\mathbf{A}}$	$Surname^{\mathbf{A}}$	$Age^{\mathbf{A}}$	$Name^{\mathbf{B}}$	$Surname^{\mathbf{B}}$	$Age^{\mathbf{B}}$	$\gamma(a,b)$	$\gamma(a,b)$
Joan	Casanoves	19	Joan	Casanovas	19	101	γ^6
Joan	Casanoves	19	Pere	Joan	18	000	γ^1
Joan	Casanoves	19	J.Manel	Casanovas	35	010	γ^3
Joan	Casanoves	19	Juan	Garcia	53	000	γ^1
Joan	Casanoves	19	Ricard	Garcia	14	000	γ^1
Joan	Casanoves	19	Pere	Garcia	82	000	γ^1
Joan	Casanoves	19	Juan	Garcia	18	000	γ^1
Joan	Casanoves	19	Ricard	Tanaka	18	000	γ^1
Pere	Joan	17	Joan	Casanovas	19	000	γ^1
Pere	Joan	17	Pere	Joan	18	110	γ^7
Pere	Joan	17	J.Manel	Casanovas	35	000	γ^1
Pere	Joan	17	Juan	Garcia	53	000	γ^1
Pere	Joan	17	Ricard	Garcia	14	000	γ^1
Pere	Joan	17	Pere	Garcia	82	100	γ^5
Pere	Joan	17	Juan	Garcia	18	000	γ^1
Pere	Joan	17	Ricard	Tanaka	18	000	γ^1
J.M.	Casanovas	35	Joan	Casanovas	19	010	γ^3
J.M.	Casanovas	35	Pere	Joan	18	000	γ^1
J.M.	Casanovas	35	J.Manel	Casanovas	35	011	γ^4
J.M.	Casanovas	35	Juan	Garcia	53	000	γ^1
J.M.	Casanovas	35	Ricard	Garcia	14	000	γ^1
J.M.	Casanovas	35	Pere	Garcia	82	000	γ^1
J.M.	Casanovas	35	Juan	Garcia	18	000	γ^1
J.M.	Casanovas	35	Ricard	Tanaka	18	000	γ^1
Juan	Garcia	53	Joan	Casanovas	19	000	γ^1
Juan	Garcia	53	Pere	Joan	18	000	γ^1
Juan	Garcia	53	J.Manel	Casanovas	35	000	γ^1
Juan	Garcia	53	Juan	Garcia	53	111	γ^8
Juan	Garcia	53	Ricard	Garcia	14	010	γ^3
Juan	Garcia	53	Pere	Garcia	82	010	γ^3
Juan	Garcia	53	Juan	Garcia	18	110	γ^7
Juan	Garcia	53	Ricard	Tanaka	18	000	γ^1
Ricardo	Garcia	14	Joan	Casanovas	19	000	γ^1
Ricardo	Garcia	14	Pere	Joan	18	000	γ^1
Ricardo	Garcia	14	J.Manel	Casanovas	35	000	γ^1
Ricardo	Garcia	14	Juan	Garcia	53	010	γ^3
Ricardo	Garcia	14	Ricard	Garcia	14	011	γ^4
Ricardo	Garcia	14	Pere	Garcia	82	010	γ^3
Ricardo	Garcia	14	Juan	Garcia	18	010	γ^3
Ricardo	Garcia	14	Ricard	Tanaka	18	000	γ^1
Pere	Garcia	18	Joan	Casanovas	19	000	γ^1
Pere	Garcia	18	Pere	Joan	18	101	γ^6
Pere	Garcia	18	J.Manel	Casanovas	35	000	γ^1
Pere	Garcia	18	Juan	Garcia	53	010	γ^3
Pere	Garcia	18	Ricard	Garcia	14	010	γ^3
Pere	Garcia	18	Pere	Garcia	82	110	γ^7
Pere	Garcia	18	Juan	Garcia	18	011	γ^4
Pere	Garcia	18	Ricard	Tanaka	18	001	γ^2
Juan	Garcia	18	Joan	Casanovas	19	000	γ^1
Juan	Garcia	18	Pere	Joan	18	001	γ^2
Juan	Garcia	18	J.Manel	Casanovas	35	000	γ^1
Juan	Garcia	18	Juan	Garcia	53	110	γ^7
Juan	Garcia	18	Ricard	Garcia	14	010	γ^3
Juan	Garcia	18	Pere	Garcia	82	010	γ^3
Juan	Garcia	18	Juan	Garcia	18	111	γ^8
Juan	Garcia	18	Ricard	Tanaka	18	001	γ^2
Ricard	Tanaka	14	Joan	Casanovas	19	000	γ^1
Ricard	Tanaka	14	Pere	Joan	17	000	γ^1
Ricard	Tanaka	14	J.Manel	Casanovas	35	000	γ^1
Ricard	Tanaka	14	Juan	Garcia	53	000	γ^1
Ricard	Tanaka	14	Ricard	Garcia	14	101	γ^6
Ricard	Tanaka	14	Pere	Garcia	82	000	γ^1
Ricard	Tanaka	14	Juan	Garcia	18	000	γ^1
Ricard	Tanaka	14	Ricard	Tanaka	18	110	γ^7

Slika 1.1: Tablica parova A x B

$Name^{\mathbf{A}}$	$Surname^{\mathbf{A}}$	$Age^{\mathbf{A}}$	$Name^{\mathbf{B}}$	$Surname^{\mathbf{B}}$	$Age^{\mathbf{B}}$	γ^i		**M/U**	m^i	u^i	m^i/u^i	$\log(m^i/u^i)$
Juan	Garcia	53	Juan	Garcia	53	111	γ^8	**M**	2/8	0/56	∞	∞
Juan	Garcia	18	Juan	Garcia	18	111	γ^8	**M**				
Pere	Joan	17	Pere	Joan	18	110	γ^7	**M**	3/8	2/56	10.5	2.35
Juan	Garcia	53	Juan	Garcia	18	110	γ^7	**U**				
Pere	Garcia	18	Pere	Garcia	82	110	γ^7	**M**				
Juan	Garcia	18	Juan	Garcia	53	110	γ^7	**U**				
Ricard	Tanaka	14	Ricard	Tanaka	18	110	γ^7	**M**				
Joan	Casanoves	19	Joan	Casanovas	19	101	γ^6	**M**	1/8	2/56	3.5	1.25
Pere	Garcia	18	Pere	Joan	18	101	γ^6	**U**				
Ricard	Tanaka	14	Ricard	Garcia	14	101	γ^6	**U**				
Pere	Joan	17	Pere	Garcia	82	100	γ^5	**U**	0/8	1/56	0	$-\infty$
J.M.	Casanovas	35	J.Manel	Casanovas	35	011	γ^4	**M**	2/8	1/56	14	2.63
Ricardo	Garcia	14	Ricard	Garcia	14	011	γ^4	**M**				
Pere	Garcia	18	Juan	Garcia	18	011	γ^4	**U**				
Joan	Casanoves	19	J.Manel	Casanovas	35	010	γ^3	**U**	0/8	11/56	0	$-infty$
J.M.	Casanovas	35	Joan	Casanovas	19	010	γ^3	**U**				
Juan	Garcia	53	Ricard	Garcia	14	010	γ^3	**U**				
Juan	Garcia	53	Pere	Garcia	82	010	γ^3	**U**				
Ricardo	Garcia	14	Juan	Garcia	53	010	γ^3	**U**				
Ricardo	Garcia	14	Pere	Garcia	82	010	γ^3	**U**				
Ricardo	Garcia	14	Juan	Garcia	18	010	γ^3	**U**				
Pere	Garcia	18	Juan	Garcia	53	010	γ^3	**U**				
Pere	Garcia	18	Ricard	Garcia	14	010	γ^3	**U**				
Juan	Garcia	18	Ricard	Garcia	14	010	γ^3	**U**				
Juan	Garcia	18	Pere	Garcia	82	010	γ^3	**U**				
Pere	Garcia	18	Ricard	Tanaka	18	001	γ^2	**U**	0/8	3/56	0	$-\infty$
Juan	Garcia	18	Pere	Joan	18	001	γ^2	**U**				
Juan	Garcia	18	Ricard	Tanaka	18	001	γ^2	**U**				
Joan	Casanoves	19	Pere	Joan	18	000	γ^1	**U**	0/8	36/56	0	$-\infty$
Joan	Casanoves	19	Juan	Garcia	53	000	γ^1	**U**				
Joan	Casanoves	19	Ricard	Garcia	14	000	γ^1	**U**				
Joan	Casanoves	19	Pere	Garcia	82	000	γ^1	**U**				
Joan	Casanoves	19	Juan	Garcia	18	000	γ^1	**U**				
Joan	Casanoves	19	Ricard	Tanaka	18	000	γ^1	**U**				
Pere	Joan	17	Joan	Casanovas	19	000	γ^1	**U**				
Pere	Joan	17	J.Manel	Casanovas	35	000	γ^1	**U**				
Pere	Joan	17	Juan	Garcia	53	000	γ^1	**U**				
Pere	Joan	17	Ricard	Garcia	14	000	γ^1	**U**				
Pere	Joan	17	Juan	Garcia	18	000	γ^1	**U**				
Pere	Joan	17	Ricard	Tanaka	18	000	γ^1	**U**				
J.M.	Casanovas	35	Pere	Joan	18	000	γ^1	**U**				
J.M.	Casanovas	35	Juan	Garcia	53	000	γ^1	**U**				
J.M.	Casanovas	35	Ricard	Garcia	14	000	γ^1	**U**				
J.M.	Casanovas	35	Pere	Garcia	82	000	γ^1	**U**				
J.M.	Casanovas	35	Juan	Garcia	18	000	γ^1	**U**				
J.M.	Casanovas	35	Ricard	Tanaka	18	000	γ^1	**U**				
Juan	Garcia	53	Joan	Casanovas	19	000	γ^1	**U**				
Juan	Garcia	53	Pere	Joan	18	000	γ^1	**U**				
Juan	Garcia	53	J.Manel	Casanovas	35	000	γ^1	**U**				
Juan	Garcia	53	Ricard	Tanaka	18	000	γ^1	**U**				
Ricardo	Garcia	14	Joan	Casanovas	19	000	γ^1	**U**				
Ricardo	Garcia	14	Pere	Joan	18	000	γ^1	**U**				
Ricardo	Garcia	14	J.Manel	Casanovas	35	000	γ^1	**U**				
Ricardo	Garcia	14	Ricard	Tanaka	18	000	γ^1	**U**				
Pere	Garcia	18	Joan	Casanovas	19	000	γ^1	**U**				
Pere	Garcia	18	J.Manel	Casanovas	35	000	γ^1	**U**				
Juan	Garcia	18	Joan	Casanovas	19	000	γ^1	**U**				
Juan	Garcia	18	J.Manel	Casanovas	35	000	γ^1	**U**				
Ricard	Tanaka	14	Joan	Casanovas	19	000	γ^1	**U**				
Ricard	Tanaka	14	Pere	Joan	17	000	γ^1	**U**				
Ricard	Tanaka	14	J.Manel	Casanovas	35	000	γ^1	**U**				
Ricard	Tanaka	14	Juan	Garcia	53	000	γ^1	**U**				
Ricard	Tanaka	14	Pere	Garcia	82	000	γ^1	**U**				
Ricard	Tanaka	14	Juan	Garcia	18	000	γ^1	**U**				

Slika 1.2: Tablica parova A x B

Poglavlje 2

Model *k*-anonimnost

U ovom poglavlju pristupamo problemu objavljivanja personaliziranih podataka, dok želimo osigurati anonimnost pojedinca s obzirom kome su podaci dostupni. Pristup se temelji na modelu k-anonimnosti. Pokazat ćemo kako se k-anonimnost provodi pomoću generalizacije te ćemo objasniti ideju minimalne generalizacije koja ne dopušta da se prilikom pripreme anonimiziranih podataka provede veća izmjena i iskrivljenost nego što je zaista potrebno za postizanje k-anonimnosti. U ovom poglavlju ćemo uvesti pojam kvazi-identifkatora, odnosno atributa koji može biti iskorišten za povezivanje, te pojam k-anonimnosti.

Baza podataka je skup međusobno povezanih podataka, pohranjenih u vanjskoj memoriji računala. Objekti, događaji koji imaju zajednička svojstva nazivaju se entitetima, a svojstva i obilježja entiteta nazivamo atributima. Jedan primjer entiteta je student koji predstavlja skup svih studenata sa obilježjima, odnosno atributima kao što su ime, prezime, datum rođenja, JMBAG i slično.

Neka je $X(A_1, \ldots, A_n)$ tablica sa danom n-torkom, gdje je n konačan. Konačan skup atributa tablice X je $\{A_1, \ldots, A_n\}$.

Definicija 2.0.1. ***Kvazi-identifikator*** *Neka je $X(A_1, \ldots, A_n)$ tablica s atributima. Kvazi-identifikator tablice X je skup atributa $\{A_i, \ldots, A_j\} \subseteq \{A_1, \ldots, A_n\}$ čije objavljivanje mora biti kontrolirano. U oznaci QI_X označavamo skup kvazi-identifikatora tablice X, sa $\mid X \mid$ označavamo kardinalitet, odnosno broj n-torki tablice X.*

Kvazi-identifikator je skup atributa u tablici koji povezivanjem s vanjskim tablicama mogu dovesti do identificiranja pojedinca. Povezujući skup kvazi-identifikatora potencijalno se može razotkriti identitet osobe. Za primjer možemo promatrati skup kvazi-identifikatora koji uključuju dob, spol i poštanski broj u medicinskoj bazi. Svaki atribut je spremljen u zasebnoj tablici. Bolest je procijenjena kao osjetljivi podataka. Napadač može povezati tablice preko kvazi-identifikatora i otkriti bolest pojedinca,a takav primjer dali smo tablicom 1.1.

Definicija 2.0.2. ***k-anonimnost*** *Neka je $X(A_1, \ldots, A_n)$ tablica sa atributima $\{A_1, \ldots, A_n\}$ i neka je QI_X skup kvazi-identifikatora tablice X. X zadovoljava k-anonimnost ako i samo ako se svaki zapis atributa kvazi-identifikatora $QI \in QI_X$ pojavi barem k puta.*

Definicija 2.0.3. ***Klasa ekvivalencije*** *Svi zapisi koji imaju jednake vrijednosti kvazi-identifikatora nazivaju se klasa ekvivalencije.*

U k-anonimnoj tablici, svaka klasa ekvivalencije sadrži barem *k* zapisa. U jednoj klasi ekvivalencije, kvazi-identifikator grupe zapisa ima jednake vrijednosti.

Odnosno, svi zapisi koji imaju jednaku vrijednost kvazi-identifikatora nazivaju se klasa ekvivalencije. Zapis u bazi svih 17-ogodišnjih dječaka puštenih 01. siječnja 2008. godine čine klasu ekvivalencije. Veličina klase može se mijenjati. Kao primjer, uzmemo 3 zapisa 17-ogodišnjaka puštenih 01. siječnja 2008. Ako godine generaliziramo na intervale od 5 godine, tada možemo imati 8 zapisa koji se odnose na muške osobe između 16 i 20 godina koji su pušteni 01. siječnja 2008. godine.

Kako bi bolje razumjeli model *k*-anonimnosti, slijedi jedan primjer koji sadrži jedan identifikacijski broj SS, dva kvazi-identifikatora godine i poštanski broj te jedan osjetljiv podatak, bolest osobe. Primjer je prikazan u tablici 2.1.

	Identifikator	Kvazi-identifikator		Osjetljiv atribut
	SS broj	Dob	Poštanski broj	Bolest
1	1234-12-1234	21	23058	Srčana bolest
2	2345-23-2345	24	23059	Srčana bolest
3	3456-34-3456	26	23060	Virusna infekcija
4	4567-45-4567	27	23061	Virusna infekcija
5	5678-56-5678	43	23058	Bubrežni kamenac
6	6789-68-6789	43	23059	Srčana bolest
7	7890-78-7890	47	23060	Virusna infekcija
8	8901-89-8901	49	23061	Virusna infekcija
9	9012-90-9012	32	23058	Bubrežni kamenac
10	0123-12-9012	34	23059	Bubrežni kamenac
11	4321-43-4321	35	23060	AIDS
12	5432-54-5432	38	23061	AIDS

Tablica 2.1: Primjer indentifikatora, kvazi-identifikatora u medicini

U danom primjeru proveden je model 4-anonimnosti prikazan tablicom 2.2 na skupu podataka. Svaki kvazi-identifikator se pojavljuje barem 4 puta.

	Identifikator	Kvazi-identifikator		Povjerljiva osobina
	SS broj	Dob	Poštanski broj	Bolest
1	*	[20,30]	230**	Srčana bolest
2	*	[20,30]	230**	Srčana bolest
3	*	[20,30]	230**	Virusna infekcija
4	*	[20,30]	230**	Virusna infekcija
5	*	[40,50]	230**	Bubrežni kamenac
6	*	[40,50]	230**	Srčana bolest
7	*	[40,50]	230**	Virusna infekcija
8	*	[40,50]	230**	Virusna infekcija
9	*	[30,40]	230**	Bubrežni kamenac
10	*	[30,40]	230**	Bubrežni kamenac
11	*	[30,40]	230**	AIDS
12	*	[30,40]	230**	AIDS

Tablica 2.2: 4-anonimnost koristeći generalizaciju

Kao i drugi modeli, k-anonimnost također ima određene uvjete koje objavljeni podaci moraju proći kako bi spriječili njihovo razotkrivanje. Kao još jedan primjer uzet ćemo tablicu X koja zadovoljava k-anonimnost za zadani $k = 2$ te je određena kvazi-identifikatorima QI_X = {rasa, datum rođenja, spol i poštanski broj}. U danoj tablici 2.3, svaka n-torka, odnosno četvorka atributa pojavljuje se barem dva puta. U ovom primjeru vrijedi za danu četvorku, $x_1[QI_X] = x_2[QI_X]$, $x_3 = [QI_X] = x_4[QI_X]$, $x_5 = [QI_X] = x_6[QI_X]$, $x_7 = [QI_X] = x_8[QI_X] = x_9 = [QI_X]$ i $x_{10} = [QI_X] = x_{11}[QI_X]$.

	Rasa	Rođenje	Spol	Pošta	Bolest
x_1	Crnac	1965	m	0214*	Plitko disanje
x_2	Crnac	1965	m	0214*	Bol u prsima
x_3	Crnac	1965	ž	0213*	Hipertenzija
x_4	Crnac	1965	ž	0213*	Hipertenzija
x_5	Crnac	1964	ž	0213*	Pretilost
x_6	Crnac	1964	ž	0213*	Bol u prsima
x_7	Bijelac	1964	m	0213*	Bol u prsima
x_8	Bijelac	1964	m	0213*	Pretilost
x_9	Bijelac	1964	m	0213*	Plitko disanje
x_{10}	Bijelac	1967	m	0213*	Bol u prsima
x_{11}	Bijelac	1967	m	0213*	Bol u prsima

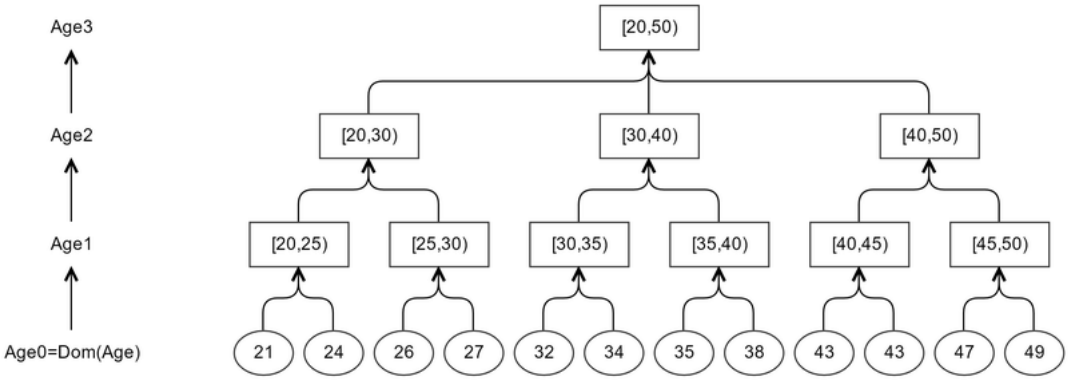
Tablica 2.3: k-anonimnost, k=2, QI={Rasa, Rođenje, Spol, Pošta}

Lema 2.0.4. *Neka je $X(A_1, \ldots, A_n)$ tablica, neka je $QI_X = (A_i, \ldots, A_j)$ kvazi-identifikator dane tablice tako da je $A_i, \ldots, A_j \subseteq A_1, \ldots, A_n$ i tablica X zadovoljava k-anonimnost. Tada se svaki niz vrijednosti u $X[A_s]$ pojavljuje barem k-puta u tablici $X(QI_X)$, gdje je $s = i, \ldots, j$.*

Proučavajući tablicu 2.3 dokaz je jasan. Tablica X sa danim kvazi-identifikatorima QI_X zadovoljava k-anonimnost, vidi se da se svaka vrijednost kvazi-identifikatora pojavljuje barem $k = 2$ puta, odnosno | X[Rasa="Crnac"] |= 6. | X[Rasa="Bijelac"] |= 5, | X[Rođenje="1964"] |= 5, | X[Rođenje="1965"] |= 4, | X[Rođenje="1967"] |= 2, | X[Spol="m"] |= 6, | X[Spol="ž"] |= 5, | X[Pošta="0213*"] |= 9,
| X[Pošta="0214*"] |= 2.

2.1 Generalizacija i skrivanje na temelju *k*-anonimnosti

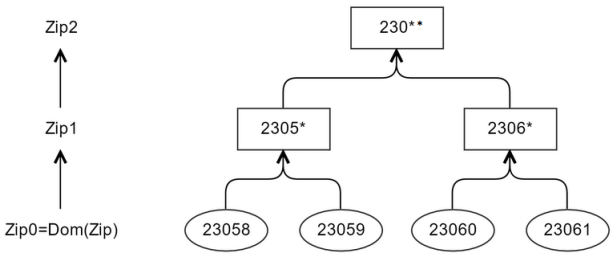
Domena atributa je skup vrijednosti koji se mogu pridružiti određenom atributu. Kako bi postigli k-anonimnost podataka, generalizacijom smanjujemo količinu vrijednosti informacija atributa. Ovo se događa preklapanjem originalnih vrijednosti atributa sa generaliziranim verzijama. Na svaki atribut možemo primjeniti nekoliko generalizacija koje su povezane i čine hijerarhiju. Prikaz jedne hijerarhije za dob osobe dana je slikom 2.1 preuzete iz [1].



Slika 2.1: Hijerarhija generalizacije za dob

Na razini *Age0* su originalne vrijednosti, na razini *Age1* su godine po intervalima od 5 godina, *Age2* su intervali od 10 godina, a *Age3* je jedan interval koji obuhvaća sve. *Age3* je općenitiji od *Age2* koji je općenitiji od *Age1*, a *Age1* od *Age0*. Kao drugi primjer možemo uzeti ZIP, odnosno poštanski broj, čija generalizacijska hijerarhija je prikazana slikom 2.2

iz [1] izvora. U primjeru poštanskog broja ulazeći u veći korak generalizacije prikrivamo još jedan broj.



Slika 2.2: Hijerarhija generalizacije ZIP-a

Definicija 2.1.1. ***Odnos generalizacija atributa*** *Neka su X^i atributi skupa podataka X. Neka su G_1 i G_2 dvije generalizacije za atribut X^i. Označimo vezu generalizacije atributa sa $\leq_{X^i}$. Koristimo zapis $G_1 \leq_{X^i} G_2$ gdje je G_2 identičan ili generalizacija G_1.*

Jednom kada je definirana hijerarhija generalizacije za svaki atribut zasebno, možemo ih kombinirati kako bi dobili generalizaciju zapisa.

Definicija 2.1.2. *Neka je X skup podataka sa danim atributima $X^1, \ldots, X^m$.* ***Generalizacija zapisa*** *je n-torka $(G_1, \ldots, G_m)$ gdje je G_i generalizacija atributa X^i.*

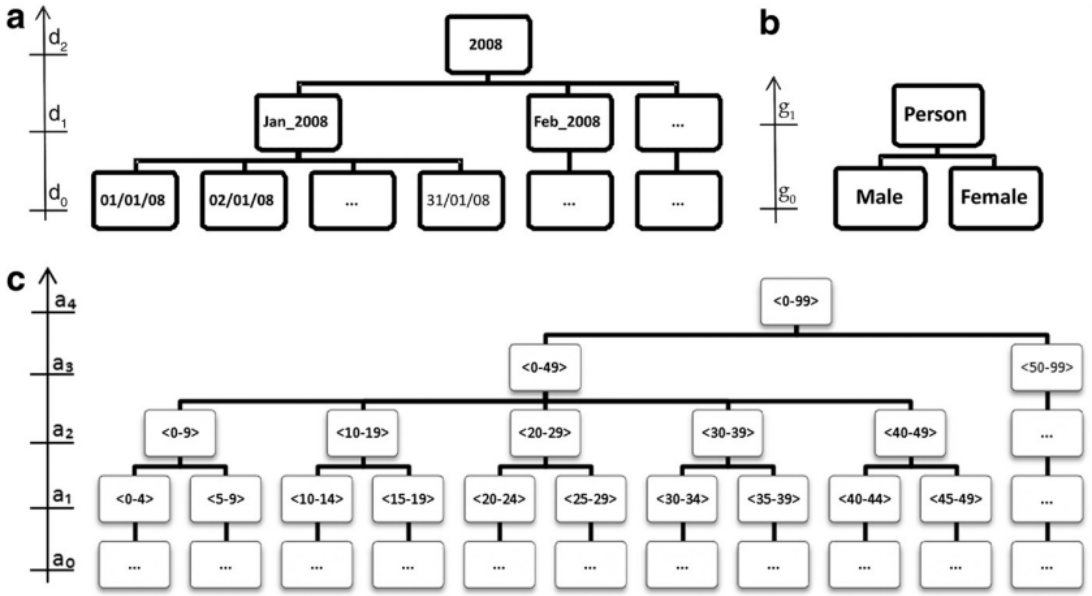
Definicija 2.1.3. *Neka je X skup podataka sa danim atributima $X^1, \ldots, X^m$. Neka su $(G_1, \ldots, G_m)$ i $(G'_1, \ldots, G'_m)$ dvije generalizacije zapisa. Odnos između generalizacija zapisa označavamo sa $\leq_X$ i koristimo zapis $(G_1, \ldots, G_m) \leq_X (G'_1, \ldots, G'_m)$ pokazujući da je G'_i ili identičan G_i ili generalizacija od G_i za svaki $i = 1, \ldots, m$.*

Cilj nam je odabrati generalizaciju zapisa tako da k-anonimnost bude zadovoljena. Generirajući k-anoniman skup podataka samo su kvazi-identifikatori generalizirani pa ćemo generalizaciju restringirati na njih. Budući da broj generalizacija utječe na količinu izgubljenih podataka, cilj nam je pronaći što manji broj generalizacija.

Definicija 2.1.4. *Neka je X skup podataka sa danim atributima $X^1, \ldots, X^m$. Neka je QI skup atributa kvazi-identifikatora i G je generalizacija zapisa za QI. Kažemo da je G* ***minimalna generalizacija zapisa*** *ako zadovoljava k-anonimnost i ako za bilo koji drugi generalizacijski zapis G' za QI takav da $G' \leq_{QI} G$ vrijedi da G' ne zadovoljava k-anonimnost.*

Neki algoritmi k-anonimizacije koriste lokalne zapise, odnosno generalizacije nisu konzistentno provedene kroz cjelokupan zapis. Praktičnije je korsititi globalne zapise gdje

ista varijabla ima isti zapis. Algoritam globalnog optimuma zadovoljava k-anonimnost dok minimizira gubitak informacija. Prevelik gubitak informacija može utjecati na netočne rezultate analize, a globalna optimizacija ublažava ovakve nedostatke. U nastavku ćemo predstaviti "OLA" (eng. Optimal Lattice Anonymization), odnosno stablo optimalne anonimizacije. Za primjer koristimo generalizaciju prikazanu slikom 2.3 iz članka [4].



Slika 2.3: Generalizacijska hijerarhija tri kvazi-identifikatora

Generalizacijska hijerarhija prikazana slikom može se prikazati kao mreža, prikazana slikome 2.4. Strijelice u rešetci prikazuju moguće generalizacijske puteve kroz mrežu. Niz koraka naziva se generalizacijska strategija. Slikom 2.5 su prikazane dvije generalizacijske strategije koje prolaze kroz čvor $< d_0, g_1, a_2 >$. Neki od danih čvorova je globalno optimalan te ga je potrebno efikasno pronaći. Sve klase ekvivalencije koje imaju manje od k zapisa su uklonjene, odnosno ne objavljuju se. Na slici 2.4, 70% podataka prikazanih čvorom $< d_0, g_0, a_0 >$ je uklonjeno jer su ti zapisi bili u malim klasama. Kod traženja optimalnog rješenja, prije ćemo izabrati stupanj većeg uklanjanja, nego stupanj veće generalizacije. Korisnik koji traži podatke odlučuje koliko uklanjanje će dopustiti i označit ćemo ga sa $MaxSup$. U našem primjeru dopuštamo $MaxSup = 5\%$, te je čvor u mreži k-anoniman ako je iznos uklanjanja manji od $MaxSup$. Čvorovi koji zadovoljavaju taj uvijet, označeni su sivo. Što više koristimo generalizaciju, manje podataka ostaje neobjavljeno odnosno ostaje skriveno, to možemo vidjeti na prikazu donjeg čvora $< d_0, g_0, a_0 >$ gdje nema generalizacije te je skriveno 70%, a na gornjem čvoru $< d_2, g_1, a_4 >$ gdje je

maksimalna generalizacija je 0% skrivanja. Ako gledamo samo uklanjanje, onda bi koristili čvor najveće generalizacije, ali to nije najbolje rješenje.

Mjera gubitka informacija

Potrebna je dodatna mjera gubitka informacija za identificiranje najmanje generaliziranog čvora s obzirom na *k*-anonimnost. Budući da smo sivom označili čvorove koji zadovoljavaju *k*-anonimnost, naše uvjete zadovoljava $< d_0, g_1, a_1 >$ s pretpostavkom da je zadovoljen omjer generalizacije i skrivanja. Međutim, ovakav pristup nije najbolji jer ne mjerimo težinu generalizacije, nije jednaka težina generalizacije spola "Muško" u "Osoba" i generalizacije godine u petogodišnji interval. Jedna od mjera gubitka informacija *Prec* u obzir uzima visinu generalizacijske hijerarhije. Za svaku varijablu, omjer koraka generalizacije i broja maksimalne generalizacije daje iznos gubitka informacija za danu varijablu. Iz dane slike, za varijablu Dob, za generalizaciju u intervalima pet godina, gubitak je $\frac{1}{4}$. Gubitak informacija cijele mreže dan je prosjekom gubitka informacija svakog kvazi-identifikatora.

Druga mjera je mjera razlikovanja ili DM od engleskog *Discernability Metric*. Mjera razlikovanja kažnjava cijeli skup podataka za svaki zapis koji se razlikuje od cijele baze. Formula DM dana je sa $DM = \sum_{f_i \geq k}(f_i)^2 + \sum_{f_i \leq k}(n \times f_i)$ gdje je f_i veličina klase ekvivalencije $i, i = 1, \ldots, Z$, gdje je Z ukupan broj klasa ekvivalencije u skupu podataka, a n je ukupan broj zapisa u skupu podataka. DM mjera nije najbolja, a više o tome možemo pronaći u [4]. Još jedna mjera gubitka informacija dana je modifikacijom mjere razlikovanja, označimo je sa DM*. Njena formula dana je sa $DM^* = \sum_{i=1}^{Z}(f_i)^2$. Ovakav oblik mjere ne daje rezultete koje bi intuitivno očekivali kada varijabla nema uniformnu razdiobu. DM* nam ukazuje da je gubitak informacija za ne uniformno distribuirane podatke mnogo veći, nego za uniformno distribuirane. Detaljnije o formuli za izračunavanje gubitka informacija za podatke koji nisu uniformno distribuirani dana je [3].

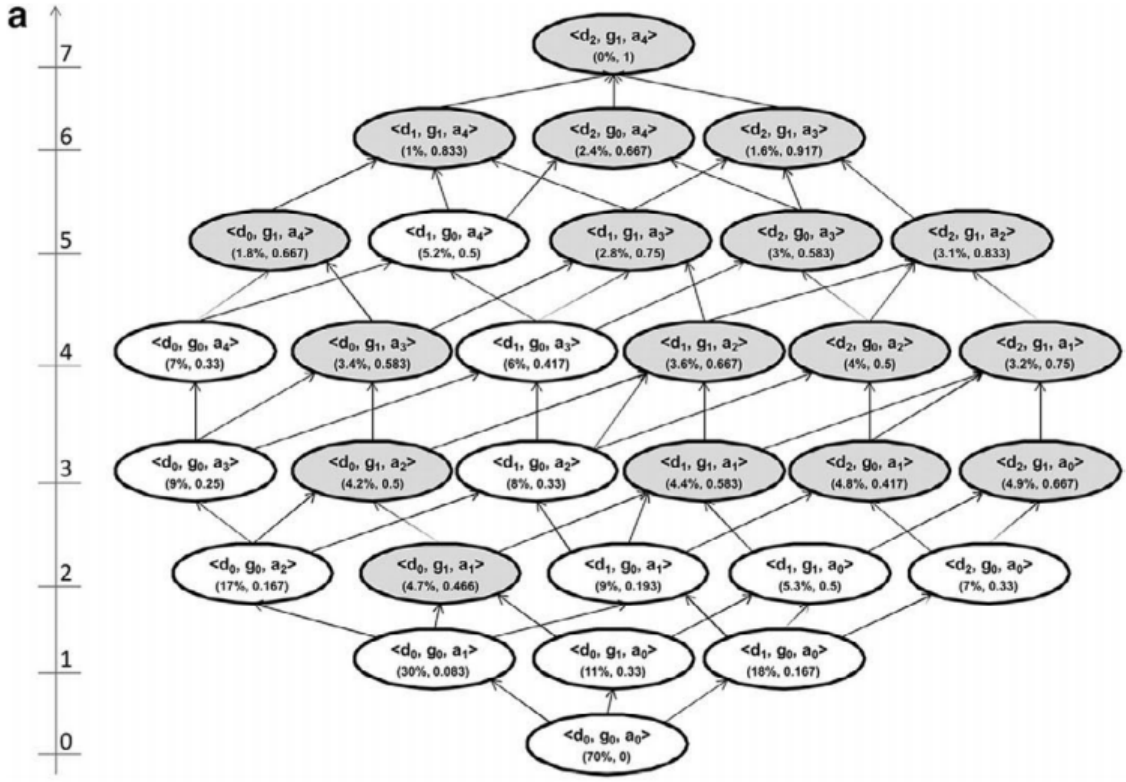
OLA algoritam

Opisat ćemo OLA algoritam (eng. Optimal Lattice Anonymization), gdje možemo koristiti jednu od gore navedenih mjera. Kao pretpostavku uzimamo da skup podataka ima više od *k* zapisa. Cilj OLA algoritma je pronalazak optimalnog čvora mreže. Optimalni čvor je *k*-anoniman i ima minimalni gubitak informacija. Algoritam se provodi u 3 koraka:

1. Za svaku generalizacijsku strategiju, provesti binarno pretraživanje kako bi pronašao sve *k*-anonimne čvorove.

2. Zadržava samo *k*-anoniman čvor najmanje visine u mreži.

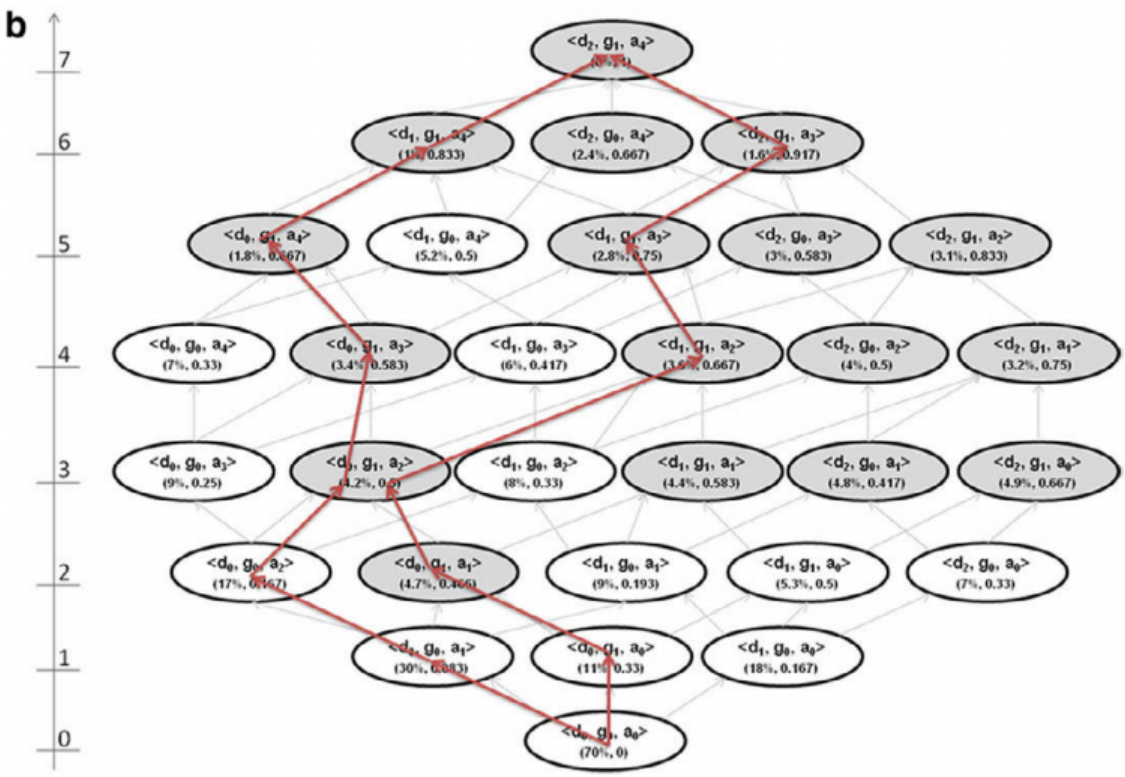
3. Kada imamo *k*-anonimne čvorove, oni su usporedivi u smislu gubitka informacija. Čvor s najmanjim gubitkom je izabran kao globalno optimalan.



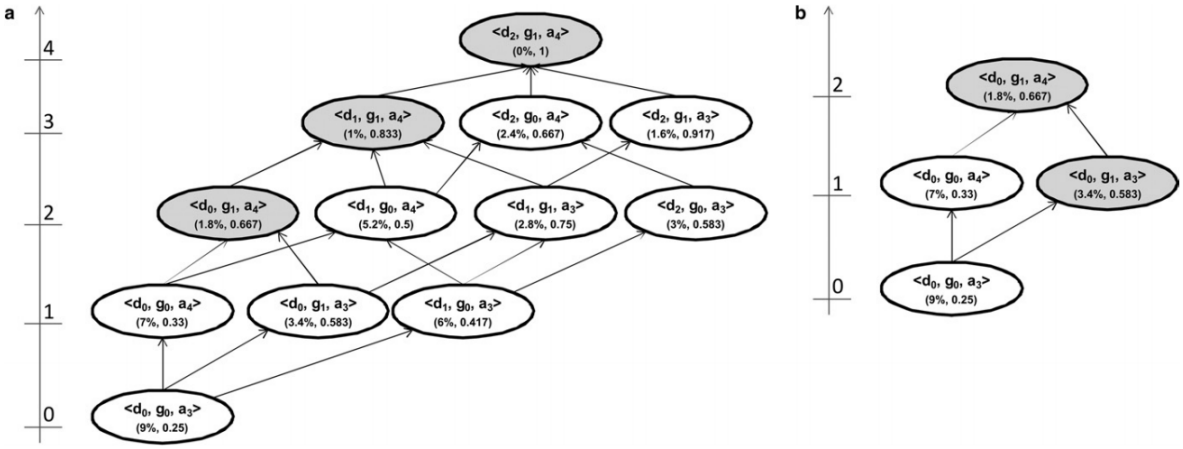
Slika 2.4: Primjer generalizacijske mreže

Procesi 1. i 3. su dugotrajni, posebno na velikim mrežama. Kako bi algoritam bio efikasan, umjesto računa, predviđanjem označava čvorove koji su k-anonimni. Predviđanje koristi dvije opcije. Ako je pronađen k-anoniman čvor N, tada svi čvorovi iste generalizacijske strategije su također k-anonimni. Umjesto da računamo njihovu k-anonimnost, onda ih samo označimo. Ako primjera radi uzmemo k-anoniman čvor $< d_0, g_1, a_2 >$, tada označimo sljedeće čvorove k-anonimnima $< d_0, g_1, a_3 >$, $< d_0, g_1, a_4 >$, $< d_1, g_1, a_4 >$, $< d_1, g_1, a_2 >$, $< d_1, g_1, a_3 >$, $< d_2, g_1, a_3 >$ i $< d_2, g_1, a_2 >$.

Kada smo dobili gornje čvorove, imamo podmrežu, ako je pronađen čvor N koji nije k-anoniman, tada svi čvorovi koji se nalaze na istoj generalizacijskog strategiji koja prolazi čvorom N nisu k-anonimni. Kao i kod k-anonimnih, umjesto računanja, možemo ih samo označiti. Primjer, izračunali smo da $< d_1, g_0, a_2 >$ nije k-anoniman, tada nisu k-anonimni ni: $< d_1, g_0, a_1 >$, $< d_0, g_0, a_2 >$, $< d_0, g_0, a_1 >$, $< d_1, g_0, a_0 >$ i $< d_0, g_0, a_0 >$. Ovakvim načinom predviđanja, znatno smanjujemo količinu računanja što omogućuje efikasnu obradu velikih mreža.



Slika 2.5: Primjer generalizacijske mreže sa potencijalnim hijerarhijama



Slika 2.6: Primjer podmreže, (a) je podmreža mreže sa slike 2.4, a (b) je podmreža mreže prikazane u (a). Sivi čvorovi su *k*-anonimni

Prolaz kroz algoritam

Algoritam implementira binarno pretraživanje kroz generalizacijsku strategiju mreže. Primjer ćemo provesti na slici 2.5. Pretraživanjem globalnog optimalnog čvora krećemo na sredini visine, $visina = 3$, prolazeći kroz čvorove počevši s lijeva. Prvi čvor koji računamo je $< d_0, g_0, a_3 >$. Računajući stupanj skrivanja dobivamo da čvor nije k-anoniman pa ga tako i označimo. Koristeći predviđanje, svi čvorovi ispod $< d_0, g_0, a_3 >$ na istoj generalizacijskoj strategiji nisu k-anonimni. Čvorovi $< d_0, g_0, a_0 >$, $< d_0, g_0, a_1 >$ i $< d_0, g_0, a_2 >$ su označeni kao ne k-anonimni bez računanja. Sada promatramo podmrežu prikazanu na slici 2.6 pod (a) kojoj je $< d_0, g_0, a_3 >$ dno, a $< d_2, g_1, a_4 >$ je vrh. Ponavljamo postupak, krećemo od sredine visine i sada je početni čvor $< d_0, g_1, a_4 >$. Računamo stupanj skrivanja za taj čvor i imamo da je k-anoniman te su i čvorovi: $< d_1, g_1, a_4 >$ i $< d_2, g_1, a_4 >$ k-anonimni. U sljedećem koraku promatramo podmrežu čiji je vrh $< d_0, g_1, a_4 >$, a dno $< d_0, g_0, a_3 >$, prikaz je dan slikom 2.6 (b). Uzimamo srednju visinu, to je prvi lijevi čvor $< d_0, g_0, a_4 >$. Računajući stupanj suzbijanja čvor nije k-anoniman, a istim postupkom nije ni čvor na nižoj visini. Nisu k-anonimni: $< d_0, g_0, a_3 >$, $< d_0, g_0, a_2 >$, $< d_0, g_0, a_1 >$ i $< d_0, g_0, a_0 >$. U najnovijoj mreži uzimamo sljedeći čvor $< d_0, g_1, a_3 >$ i računom dobivamo da je k-anoniman, te svi čvorovi u toj strategiji koji su iznad njega. k-anonimni čvorovi su: $< d_0, g_1, a_4 >$, $< d_1, g_1, a_3 >$, $< d_1, g_1, a_4 >$, $< d_2, g_1, a_3 >$ i $< d_2, g_1, a_4 >$. Vratimo se na sliku 2.6 (a) dio, i nastavljamo s idućim čvorom $< d_1, g_0, a_4 >$. Taj čvor ima stupanj skrivanja veći od dopuštenog pa nije k-anoniman i primjenjujemo isti postupak. OLA algoritam traži k-minimalno rješenje, odnosno ako pronađe k-minimalan čvor provjerava postoji li neki drugi u toj istoj strategiji. Ako pronađe drugi minimalan, prijašnje briše s liste i dodaje novi. Posljednji korak algoritma je uspoređivanje k-minimalnih čvorova s obzirom na gubitak informacija te uzimajući jedan kao globalno optimalno rješenje.

Algorithm 1: *k*-minimalni algoritam

```
1  // Ulazni podaci su korijen i vrh mreže
2  Function Main
3      S={}
4      Kmin(Bottom-Node, Top-Node)
5      Optimal= min_{x∈S}(InfoLoss(x))
6  Function Kmin (Bnode,Tnode)
7      L=Lattice(Bnode,Tnode)
8      H_H=Height(L,Tnode)
9      if H_H > 1 then
10         h = ⌊H_H/2⌋
11         for p:=1 to Width(L,h) do
12             N=Node(L,h,p)
13             if IsTaggedKAnonymous(N)==True then
14                 Kmin(Bnode,T)
15             else if IsTaggedNotKAnonymous(N)==True then
16                 Kmin(N,Tnode)
17             else if IsKAnonymous(N)=True then
18                 TagKAnonymous(N)
19                 Kmin(Bnode,N)
20             else
21                 TagNotKAnonymous(N)
22                 Kmin(N,Tnode)
23             end
24         end
25     else
26         // Slučaj dva čvora mreže
27         if IsTaggedNotKAnonymous(Bnode)==True then
28             N=Tnode else if IsKAnonymous(Bnode)==True then
29             TagKAnonymous(Bnode)
30             N=Bnode
31         else
32             TagNotKAnonymous(Bnode)
33             N=Tnode
34         end
35         S= S + N
36         CleanUp(N)
37     end
```

Inkognito algoritam

Inkognito algoritam slijedi princip prolaska od dna prema vrhu prolazeći prvu razinu kako bi pronašao optimalnu generalizaciju zapisa. Kako bi smanjio prostor pretraživanja, inkognito algoritam koristi sljedeća svojstva generalizacije i *k*-anonimnosti.

Propozicija 2.1.5. ***Svojstvo generalizacije*** *Neka je X skup podataka i neka je QI kvazi-identifikator atributa skupa X, neka su G_1 i G_2 generalizacijski zapisi QI tako da vrijedi $G_1 \leq_{QI} G_2$. Ako G_1 daje k-anonimnost za X, tada i G_2 daje k-anonimnost za X.*

Propozicija 2.1.6. ***Svojstvo klase*** *Neka je X skup podataka, neka je QI kvazi-identifikator atributa skupa X, i neka su G_1 i G_2 generalizacijski zapisi QI tako da vrijedi $G_1 \leq_{QI} G_2$. Frekvenciju dane klase ekvivalencije C skupa X koja poštuje G_2 možemo izračunati kao zbroj frekvencija klase ekvivalencije skupa X sa poštivanjem G_1 koja generalizira C.*

Propozicija 2.1.7. ***Svojstvo podskupa*** *Neka je X skup podataka, neka je QI kvazi-identifikator atributa skupa X, neka je $Q \subset QI$ podskup kvazi-identifikatora. Ako je X k-anoniman tako da poštuje Q, onda je i k-anoniman kada poštuje bilo koji podskup atributa P od Q, $P \subseteq Q$.*

Dokazi prethodnih propozicija možemo pronaći u [6].

Svojstvo podskupa nam kaže da za danu generalizaciju koja zadovoljava *k*-anonimnost, tada sve generalizacije uklanjanjem jednog od atributa također moraju zadovoljavati *k*-anonimnost. Na temelju danog, inkognito algoritam pretražuje samostalne generalizacijske atribute koji nam daju *k*-anonimnost i na taj način iterativno povećava broj generalizacijskih atributa.

Rođenje	Spol	Poštanski broj	Bolest
1/21/76	M	53715	Prehlada
4/13/786	Ž	53715	Hepatitis
2/28/76	M	53703	Bronhitis
1/21/79	M	53703	Lom ruke
4/13/86	Ž	53706	Uganuće
2/28/76	Ž	53706	Kamenac

Tablica 2.4: Podaci o bolesti

Primjer gdje možemo vidjeti svojstvo podskupa dan je tablicom 2.4. Ako je atribut ”Spol” generaliziran u ”Osoba”, te uz njega promatramo atribut ”Poštanski broj”, tada je tablica 2-anonimna. Tablica mora biti 2-anonimna ako promatramo zasebno ”Poštanski broj”, ”Spol” generaliziran u ”Osoba”. Možemo uočiti da tablica nije 2-anonimna poštujući

generalizaciju ”Spol”,”Poštanski broj”. Iz ovoga možemo zaključiti da tablica ne može biti 2-anonimna ako uzmemo skup atributa ”Rođenje”, ”Spol”, ”Poštanski broj”.

Kada traži generalizaciju veličine *i* koja zadovoljava *k*-anonimizaciju, tada koristi svojstvo generalizacije kako bi smanjio prostor pretraživanja. Jednom kada je pronađena generalizacija *G* koja zadovoljava *k*-anonimnost, tada sve iduće generalizacije također zadovoljavaju *k*-anonimnost. Za primjer ponovno uzimamo tablicu 2.4. Neka je sa S_0 označena nulta generalizacija za ”Spol”, a sa S_1 prva generalizacija, gdje muško, žensko vrijednost generalizirano u osoba. Tablica je 2-anonimna poštujući S_0 pa je 2-anonimna i za S_1, jer je S_1 generalizacija od S_0.

Kako bi algoritam smanjio trošak provjeravanja da broj frekvencija povezan sa generalizacijom zadovoljava *k*-anonimnost, inkognito algoritam koristi svojstvo klase, odnosno računa frekvencije prethodnih generalizacija.

Algorithm 2: Inkognito algoritam *k*-anonimnosti

Data: X: skup originalnih podataka
k: zahtjev anonimnosti
QI: *n* atributa kvazi-identifikatora
$(G^i_j)_j$: generalizacijska hijerarhija za atribut QI^i, za svaki $i = 1, \ldots, |QI|$
Result: Skup zapisa generalizacija koji donosi *k*-anonimnost
C_1 : {Čvorovi hijerarhijske generalizacije atributa *QI*}
E_1 : {Bridovi hijerarhijske generalizacije atributa *QI*}
queue:= prazan niz
for $i := 1, \ldots, n$ **do**
  // S_i sadrži *k*-anonimne generalizacije sa atributima *i*
  $S_i := C_i$
  roots:=čovorovi od C_i bez ulaznih bridova
  Umetni *roots* u *queue* i čuvaj sortirano po visini
  while *queue* $\neq \emptyset$ **do**
    node:= izbaci prvi element iz *queue*
    if *node nije obilježen* **then**
      if $node \in roots$ **then**
        $frequencies$:=izračunaj frekvenciju od *T* s obzirom na atribute čvora *node*
      else
        $frequencies$:= izračunaj frekvenciju od *T* s obzirom na atribute čvora *node* koristeći frekvenciju roditelja
      end
    end
    Provjeri *k*-anonimnost X s obzirom na *node* koristeći $frequencies$
    if *X je k-anoniman s obzirom na node* **then**
      označi sve direktne generalizacije čvora *node*
    else
      Izbriši čvor *node* iz S_i
      Dodaj direktnu generalizaciju čvora *node* u niz *queue* i zadrži poredak po visini
    end
  end
  //Generiraj graf svih mogućih *k*-anonimnih generalizacija sa $i + 1$ atributa
  C_{i+1}, E_{i+1} := generiran graf iz S_i i E_i
end
return S_n

Poglavlje 3

l-raznolikost i *t*-bliskost

k-anonimnost daje mogućnosti zaštite privatnosti i spriječavanje razotkrivanja, ali ne pruža dovoljno zaštite kada zapisi u grupi imaju iste vrijednosti povjerljivih atributa. *l*-raznolikost se smatra proširenjem modela *k*-anonimnosti te može pružiti veću zaštitu. Cilj modela *l*-raznolikosti je zahtjevati minimnalnu razinu raznolikosti povjerljivih atributa u određenoj *k*-anonimnoj grupi zapisa.

3.1 *l*–raznolikost

Definicija 3.1.1. ***l*-raznolikost** *Klasa ekvivalencije zadovoljava l-raznolikost ako postoji barem l "dobro zastupljenih" vrijednosti osjetljivih atributa. Skup podataka zadovoljava l-raznolikost ako svaka klasa ekvivalencije zadovoljava l-raznolikost.*

Ovakva definicija nije najspretnija jer nije precizno definirano značenje dobro zastupljen. Dane su tri interpretacije:

1. *Jasna l-raznolikost* je najjednostavniji oblik *l*-raznolikosti koji zahtjeva da svaka klasa ekvivalencije ima najmanje *l* različitih osjetljivih atributa.

2. *Entropija l-raznolikosti* je dana entropijom klase ekvivalencije S sa

$$H(s) = -\sum_{s \in S} p_S(s) \log(p_S(s))$$

gdje je $p_S(s)$ dio zapisa klase ekvivalencije S osjetljivih vrijednosti jednakih s. Tablica zadovoljava entropiju *l*-raznolikosti ako za svaku klasu ekvivalencije S imamo

$$H(S) \geq \log l. \tag{3.1}$$

3. *Rekurzivna (c,l)-raznolikost* zahtjeva gornju granicu frekvencije najzastupljenijih osjetljivih vrijednosti atributa te donju granicu najmanje čestih vrijednosti. S tom pretpostavkom $r_1, \ldots, r_m$ je rastući niz frekvencija vrijednosti osjetljivih atributa u ekvivalentnoj klasi. Klasa zadovoljava (c, l)-raznolikost ako

$$r_1 < c(r_l + r_{l+} + \cdots + r_m).$$

Rekurzivno znači da ako neka od osjetljivih vrijednosti u (c, l)-raznolikoj tablici uklonimo, ostatak tablice treba biti barem $(c, l-1)$-raznolik. 1-raznolikost je uvijek ispunjena.

Tablicom 3.1 prikazani su podaci podijeljeni u dva bloka tako da su ne osjetljivi podaci generalizirani. Svaki blok sadrži 6 pojedinaca i 3 različite vrijednosti za osjetljiv atribut "Medication", "Tamoxifen", "Pepcid", "Erythropoietin" u prvom bloku. Takva tablica je 3-raznolika.

	Ne osjetljivi		Osjetljivi
	Godine	Pošta	Lijekovi
1	≤60	75***	Tamoxifen
7	≤60	75***	Tamoxifen
5	≤60	75***	Pepcid
2	≤60	75***	Tamoxifen
12	≤ 60	75***	Tamoxifen
9	≤ 60	75***	Erythropietin
3	≥ 60	75***	Captopril
6	≥60	75***	Synthroid
11	≥60	75***	Synthroid
4	≥60	75***	Synthroid
8	≥60	75***	Pepcid
10	≥60	75***	Pepcid

Tablica 3.1: 3-raznolikost

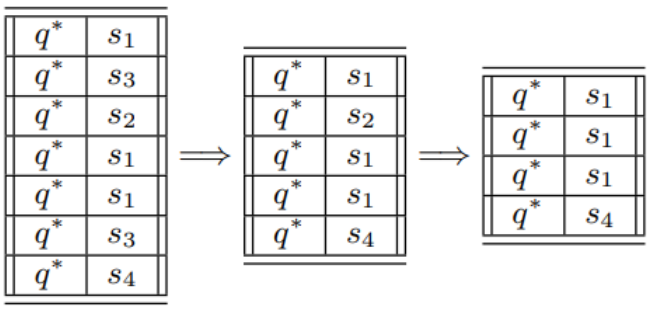
Za entropiju, uzimamo istu tablicu. Po definiciji entropije slijedi:

$$H_1 = -(\frac{4}{6} \cdot \log(\frac{4}{6}) + 2 \cdot \frac{1}{6} \cdot \log(\frac{1}{6})) \approx 0.378$$

$$H_2 = -(\frac{3}{6} \cdot \log(\frac{3}{6}) + \frac{2}{6} \cdot \log(\frac{1}{6}) + \frac{1}{6} \cdot \log(\frac{1}{6})) \approx 0.439$$

Kako bi ispunili uvjet potrebno je provjeriti $H_1 \geq \log(l)$ i $H_2 \geq \log(l)$, gdje je minimalna entropija H_1 koja određuje l. U ovom slučaju slijedi $H_1 \geq \log(l) \Leftrightarrow 10^{H_1} \approx 2.387$. Stoga

je dana tablica barem 2.3-raznolika, što znači da svaki blok sadrži barem dvije različite osjetljive vrijednosti.



Slika 3.1: (c,*l*)-raznolikost

Slikom 3.1 su prikazani fiktivni podaci. Moguće osjetljive vrijednosti S su s_1, s_2, s_3, s_4, gdje su osjetljivi podaci generalizirani sa q*. Iz tablice vidimo frekvencije $n_1 = 3, n_2 = 2, n_3 = 1, n_4 = 1$, odakle dobivamo $l = 3 - 1$, najveća frekvencija manje 1. Neka je konstanta $c = 2$. Tada je tablica $(2, 3)$-raznolika, ako vrijedi $n_1 < c(n_3 + n_4)$, i vidimo da vrijedi za $c = 2$, $3 < 2 \cdot 2 = 4$. Druga najzastupljenija vrijednost je s_3, nju uklonimo. Novo nastala tablica treba biti $(2, 2)$-raznolika jer mora biti zadovoljen uvjet $n_1 < c(n_3 + n_4)$. Ovaj slučaj prikazan je srednjom tablicom na slici 3.1. Uklanjajući drugi osjetljiv podatak, dobivamo $(2, 1)$-raznolikost. Kako je 1-raznolikost uvijek zadovoljena, smatramo da vrijedi $(2, 3)$-raznolikost.

3.2 *t*–bliskost

Model *t*-bliskosti proširuje model *l*-raznolikosti uzimajući u obzir raspodjelu osjetljivog atributa u tablici i u klasi. Ovom metodom možemo zaštititi osjetljive atribute, ali ne i otkrivanje identiteta.

Definicija 3.2.1. ***t-bliskost*** *Klasa ekvivalencije zadovoljava t-bliskost ako udaljenost distribucije osjetljivih atributa u klasi i distribucije atributa u cijeloj bazi nije veća od granice t. Kažemo da skup podataka zadovoljava t-bliskost ako svaka klasa ekvivalencije zadovoljava t-bliskost.*

Udaljenost među distribucijama računamo sa $EMD(P, Q)$ koja računa trošak transformacije iz distribucije P u distribuciju Q mijenjajući težinu vjerojatnosti. EMD dolazi iz engleskog *earth mover's distance*.

Definicija 3.2.2. ***EMD udaljenost***
Neka je $\{v_1, \ldots, v_r\}$ *skup vrijednosti i neka su* $P = (p_1, \ldots, p_r)$ *i* $Q = (q_1, \ldots, q_r)$ *razdiobe vjerojatnosti, gdje su* p_i *i* q_i *vjerojatnosti* P *i* Q *dodijeljene* v_i*.* $EMD(P, Q)$ *računa minimalnu cijenu trasporta od* P *do* Q *pa ovisi koliko je težine prenešeno i koliko daleko. Sa* d_{ij} *označimo udaljenost između vrijednosti* v_i *i* v_j*, sa* f_{ij} *označimo težinu koju prenosimo između* v_i *i* v_j*,* $EMD(P, Q)$ *možemo računati sa:*

$$EMD(P, Q) = \min_{f_{ij}} \sum_{i=1}^{r} \sum_{j=1}^{r} d_{ij} f_{ij}$$

tako da vrijedi:

$$f_{ij} \geq 0 \qquad 1 \leq i, j \leq r$$

$$p_i - \sum_{j=1}^{r} f_{ij} + \sum_{j=1}^{r} f_{ji} = q_i \qquad 1 \leq i \leq m$$

$$\sum_{i=1}^{m} \sum_{j=1}^{m} f_{ij} = \sum_{i=1}^{m} p_i = \sum_{j=1}^{m} q_i = 1.$$

Računanje EMD za numeričke atribute

Za numeričke atribute, prirodno se nameće kao mjeru za računanje koristiti udaljenost.

Definicija 3.2.3. ***Uređena udaljenost***
Neka je $\{v_1, \ldots, v_r\}$ *je skup vrijednosti tako da su elementi sortirani rastući. Uređena udaljenost dana je sa*

$$uređena\_udaljenost(v_i, v_j) = \frac{| \, rang(v_i) - rang(v_j) \, |}{r - 1} = \frac{| \, i - j \, |}{r - 1}$$

gdje rang predstavlja redni broj elementa u sortiranom nizu.

Neka su $P = (p_1, \ldots, p_r)$ i $Q = (q_1, \ldots, q_r)$ vjerojatnosne distribucije za $\{v_1, \ldots, v_r\}$. EMD za P i Q možemo izračunati na sljedeći način

$$EMD(P, Q) = \frac{1}{r - 1} \sum_{i=1}^{r} \sum_{j=1}^{i} | \, (p_i - q_i) \, | \, . \tag{3.2}$$

Računanje EMD za kategorijske atribute

Dok je za numeričke atribute vrlo jasna definicija udaljenosti, za kategorijske atribute postoji nekoliko udaljenosti ovisno o tipu kategorije koju promatramo. Redni kategorijski

atribut možemo promatrati kao numerički atribut. Za takve atribute koristimo prethodno navedenu formulu *uređena_udaljenost*. Za nominalne atribute ne postoji veza vrijednosti različitih atributa. U tom slučaju ćemo koristiti klasičnu udaljenost, koju ćemo za bilo koje dvije različite vrijednosti atributa postaviti na 1. Ako su $P = (p_1, \ldots, p_r)$ i $Q = (q_1, \ldots, q_r)$ vjerojatnosne distribucije za dane vrijednosti nominalnih atributa, EMD računamo kao

$$EMD(P, Q) = \frac{1}{2} \sum_{i=1}^{r} \mid p_i - q_i \mid . \tag{3.3}$$

Lema 3.2.4. *Neka je X skup podataka i neka je Y generazlizacija od X koja zadovoljava t-bliskost. Ako je Z sljedeća generalizacija od Y, tada Z također zadovoljava t-bliskost u skladu sa X.*

Dokaz Svaka klasa ekvivalencije generalizacije Z je unija skupa klasa ekvivalencije generalizacije Y i svaka klasa ekvivalencije u Y zadovoljava t-bliskost. Tada svaka klasa ekvivalencije u Z također zadovoljava t-bliskost te skup podataka X zadovoljava t-bliskost generalizacijom Z. □

Lema 3.2.5. *Neka je X skup podataka. Neka je Y skup podataka koji zadovoljavaju t-bliskost s obzirom na X. Ako je Z skup podataka dobiven od Y uklanjajući određene atribute, tada je Z također t-blizak uzimajući u obzir X.*

Dokaz Klasa ekvivalencije Y zadovoljava t-bliskost. Z je unija skupa klasa ekvivalencije od Y. Svaka klasa ekvivalencije generalizacije Z također zadovoljava t-bliskost. Tada skup podataka X zadovoljava t-bliskost s obzirom na Z. □

Za više detalja o računanju EMD-a za nominalne kategorije atributa možemo pronaći u [7] te više informacija o primjeru možemo pronaći u [9].

Računanje EMD numeričkih atributa

Neka je $Q_1 = \{14, 27, 88, 101\}$ i neka je $P_1 = \{14, 88\}$, tako da je $P_1 \subseteq Q_1$ tako da je P_1 ekvivalentna klasa Q_1. Distribucije danih skupova su $\mathbb{Q}_1 = \{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\}$, $\mathbb{P}_1 = \{\frac{1}{2}, 0, \frac{1}{2}, 0\}$. Jednadžbom 3.2 računamo trošak transformacije iz distribucije P_1 u distribuciju $\mathbb{Q}_1$, gdje je r veličina skupa. Iz dane jednadžbe 3.2, stoga slijedi t-bliskost.

$$
\begin{aligned}
EMD(\mathbb{P}_1, \mathbb{Q}_1) = & \frac{1}{4-1}\left[\left|\frac{1}{2}-\frac{1}{4}\right| + \right. \\
& \left|\left(\frac{1}{2}-\frac{1}{4}\right)+\left(0-\frac{1}{4}\right)\right| + \\
& \left|\left(\frac{1}{2}-\frac{1}{4}\right)+\left(0-\frac{1}{4}\right)+\left(\frac{1}{2}-\frac{1}{4}\right)\right| + \\
& \left.\left|\left(\frac{1}{2}-\frac{1}{4}\right)+\left(0-\frac{1}{4}\right)+\left(\frac{1}{2}-\frac{1}{4}\right)\right|\right] \\
& \approx 0.1667.
\end{aligned}
$$

Ako imamo više od jedne klase ekvivalencije u danoj tablici, t-bliskost tablice je maksimum EMD-a pojedinačne klase ekvivalencije.

Računanje EMD kategorijskih atributa

EMD kategorijskih atributa računamo danom formulom 3.3. U ovom primjeru koristit ćemo podatke prikazane u tablici 3.2.

Adresa	Zona	Incident
*	2C	Nestanak struje
*	2C	Nestanak struje
*	2C	Nestanak struje
*	4F	Krađa
*	4F	Požar
*	4F	Smrtni slučaj
*	4F	Požar
*	9A	Popravak pločnika
*	9A	Nestanak struje
*	3B	Pesticidi
*	3B	Nestanak struje
*	3B	Popravak pločnika
*	3B	Sadnja drveća
*	3B	Popravak pločnika

Tablica 3.2: Anonimizirana tablica

Skup vrijednosti zastupljenih u tablici je Q_2 = {Nestanak struje, Krađa, Požar, Smrtni slučaj, Popravak pločnika, Pesticidi, Sadnja drveća }. Distribucija skupa Q_2 je jednaka $\mathbb{Q}_2 = \{\frac{5}{14}, \frac{1}{14}, \frac{2}{14}, \frac{1}{14}, \frac{3}{14}, \frac{1}{14}, \frac{1}{14}\}$. Prvu klasu ekvivalencije označimo sa $P_{2,1}$ = {Nestanak struje, Nestanak struje, Nestanak struje}, a distribucija je $\mathbb{P}_{2,1} = \{\frac{3}{3}, 0, 0, 0, 0, 0, 0\}$. Sljedeće

klase ekvivalencije su $P_{2,2}$ = {Krađa, Požar, Smrtni slučaj, Požar}, $\mathbb{P}_{2,2} = \{0, \frac{1}{4}, \frac{2}{4}, \frac{1}{4}, 0, 0, 0\}$, $P_{2,3}$ = {Popravak pločnika, Nestanak struje}, $\mathbb{P}_{2,3} = \{\frac{1}{2}, 0, 0, 0, \frac{1}{2}, 0, 0\}$, $P_{2,4}$ = {Pesticidi, Nestanak struje, Popravak pločnika, Sadnja drveća, Popravak pločnika} sa distribucijom $\mathbb{P}_{2,4} = \{\frac{1}{5}, 0, 0, 0, \frac{2}{5}, \frac{1}{5}, \frac{1}{5}\}$. Računamo udaljenost za svaku klasu ekvivalencije zasebno pa slijedi:

$$\begin{aligned} EMD(\mathbb{P}_{2,1}, \mathbb{Q}_2) = & \frac{1}{2}\left[\left|\frac{3}{3} - \frac{5}{14}\right| + \left|0 - \frac{1}{14}\right| + \left|0 - \frac{2}{14}\right| + \right. \\ & \left. \left|0 - \frac{1}{14}\right| + \left|0 - \frac{3}{14}\right| + \left|0 - \frac{1}{14}\right| + \left|0 - \frac{1}{14}\right|\right] \\ & \approx 0.6429. \end{aligned}$$

Na istin način dobivamo: $EMD(\mathbb{P}_{2,2}, \mathbb{Q}_2) \approx 0.7143$, $EMD(\mathbb{P}_{2,3}, \mathbb{Q}_2) \approx 0.4286$ i $EMD(\mathbb{P}_{2,4}, \mathbb{Q}_2) \approx 0.4429$. Maksimum dobivenih vrijednosti 0.7143 je t-bliskost dane tablice.

Više primjera nalazi se u izdanom članku [2].

Poglavlje 4

Osnovne tehnike anonimizacije

Tehnike anonimiziranja podataka pomažu pri uklanjanju osjetljivih informacija iz podataka. Tehnike mogu biti u okvirima statistike, algoritmi ili izrađene po narudžbi te moraju osigurati da vrsta podatka ostane nepromijenjena, odnosno da izmijenjena i originalna vrijednost budu istog tipa.

4.1 Supstitucija

Primjenom supstitucije ulazni, odnosno originalni podaci nasumičnim odabirom su uvijek zamijenjeni sa odgovarajućom zamjenom. Tip ulaznoga podatka može biti numerički, alfanumerički ili u obliku datuma. Ova tehnika može biti primijenjena na kompletan unos podataka ili na određen dio te se tehnika može primijeniti uz dodatne uvjete. Primjer supstitucije gdje je svaki fonetski znak danog ulaznog podatka zamijenjen nasumičnim odabirom drugog znaka te tako generirajući maskirane podatake prikazan je sa Tablica 4.1

	VRIJEDNOST	OPIS
Ulazni podataka	LONDON01	Tip podatka i njegova dužina je sačuvana. Zamjenski znakovi su nasumično odabrani koristeći algoritam.
Izlazni podataka	ABMEQD12	U ovom primjeru, znak ”O” u ”LONDON01” zamijenjen drugačijim znakom, primjer ”B” i ”Q”, slično ”N” sa ”M” i ”D”.

Tablica 4.1: Nasumična supstitucija

Tehnika supstitucije je korisna u slučaju čuvanja tipa i formata maskiranih podataka te kada nam nije potrebno da podaci izgledaju stvarno, odnosno realno. Kada želimo da nam podaci izgledaju stvarno, onda koristimo varijante supstitucije te ćemo objasniti

”supstituciju riječi” ili ”potraga zamjene”.
Ovaj oblik tehnike obuhvaća vanjsku datoteku ili repozitorij koji sadrži listu vrijednosti. Vrijednosti iz vanjske datoteke ili repozitorija biti će korištene za maskiranje originalnih podataka. Tablica 4.2 prikazuje kako se odvija supstitucija riječi. Vrijednost Marko unutar osjetljivo pohranjenih podataka zamijenjena je sa vrijednošću iz vanjske datoteke, odnosno sa David.

	VRIJEDNOST	OPIS
Ulazni podatak	Marko	Dužina riječi ne mora biti sačuvana.
Izlazni podatak	David	S obzirom na alat koji se koristi, ime je nasumično odabrano iz vanjske datoteke.

Tablica 4.2: Supstitucija riječi

Ovakva tehnika anonimizacije, odnosno maskiranja podataka najčešće se koristi kada je potrebno podatke prikazati realistično, posebno imena. Nedostatak ovoj tehnici je maskiranje velikog skupa podataka jer veliki skup podataka zahtjeva i veliku vanjsku datoteku sa vrijednostima koje koristimo za zamijene.

4.2 Miješanje

Tehnika miješanja uključuje preuređenje podataka premiještanjem različitih redova unutar istog stupca kao što je prikazano u tablicama 4.3 i 4.4. U prikazanoj tablici osjetljiv dani podatak je broj računa korisnika koji je anonimiziran korištenjem premještanja redova unutar kolone.

TRANSAKCIJSKI BROJ	BROJ RAČUNA	DATUM	IZNOS	STANJE
100001	6001298791	23-03-2011	2000	10000
100002	6001298891	02-02-2012	200	7800
100003	6011398801	01-03-2012	788	9880
100004	6014546781	10-10-2011	2055	2990

Tablica 4.3: Tablica transakcija prije anonimizacije

Na ovaj način miješanjem redova osigurali smo da se broj računa korisnika ne može povezati sa korisnikovim stanjem računa ili transakcijskim iznosom. Ovakvo preuređivanje može biti nasumično, ali se moramo osigurati da određena kolona nije povezana sa drugom kolonom u razotkrivanju osobnih podataka. Miješanje podataka čuva tip podatka jer nema

TRANSAKCIJSKI BROJ	BROJ RAČUNA	DATUM	IZNOS	STANJE
100001	6014546781	23-03-2011	2000	10000
100002	6001298791	02-02-2012	200	7800
100003	6001298891	01-03-2012	788	9880
100004	6011398801	10-10-2011	2055	2990

Tablica 4.4: Tablica transakcija nakon anonimizacije korištenjem miješanja

unošenja novih, to je prilično lagana tehnika za implementiranje. Tehnika miješanja redova je puno efikasnija kada se primjenjuje na tablice sa velikim brojem zapisa i podataka u odnosu na male tablice. Sa malim skupom podataka vrlo je lako utvrditi originalne osjetljive podatke, stoga miješanje ne treba koristiti na malom broju podataka.
Grupno miješanje je varijanta miješanja gdje se grupa kolona zajedno premiješta. Jedan primjer grupnog miješanja podataka dan je tablicom korisnikove adrese gdje je potrebno osigurati da adresa djeluje stvarno, ali da se ne može identificirati. Stvarni podaci prikazani su tablicom 4.5 te je na njoj provedeno grupno miješanje kolona grad, država, poštanski broj.

OSOBNI BROJ	ADRESA	GRAD	DRŽAVA	POŠTANSKI BROJ
100001	606, Edison Square	Phoenix	AZ	85097
100002	202, Thomas St	Edison	NJ	08817
100003	4, King Street	Waltham	MA	02455
100004	555, Jefferson Plaza	Chicago	IL	60604

Tablica 4.5: Adresa korisnika prije anonimizacije

OSOBNI BROJ	ADRESA	GRAD	DRŽAVA	POŠTANSKI BROJ
100001	606, Edison Square	Waltham	MA	02455
100002	202, Thomas St	Chicago	IL	60604
100003	4, King Street	Phoenix	AZ	85097
100004	555, Jefferson Plaza	Edison	NJ	08817

Tablica 4.6: Adresa korisnika nakon anonimizacije

Nakon anonimizacije grupnim miješanjem, podaci su prikazani tablicom 4.6. Ovim primjerom Waltham se nalazi u državi Massachussets(MA) gdje je poštanski broj 02455 i

Phoenix je u AZ određen poštanskim brojem 85097 i kao takvi će krajnjem korisniku dati osjećaj da je adresa točna, iako nije u potpunosti.
Kada je potrebno anonimizirati grupu informacija kao što je adresa korisnika u više kolona kao što su adresa, grad, država, pošanski broj tada je dobro koristiti grupno miješanje unutar kolone. Kao i kod općenitog miješanja i grupno miješanje je učinkovitije korištenjem veće količine podataka.

4.3 Varijacija broja

Ova statistička tehnika uključuje generiranje broja između donje i gornje granice danog skupa podataka ili varijacija postojećih nemaskiranih brojeva. Ako je generirani broj izvan zadanih granica, odnosno ako je veći od gornje granice odnosno manji od donje granice, anonimizirani broj biti će isti kao gornja granica, odnosno kao donja granica. Primjer na kojem možemo provesti ovu metodu je bankovni polog korisnika prikazan u tablici 4.7. Anonimizaciju ćemo provesti tehnikom varijacije broja na koloni polog tako da je gornja granica određena sa 25 000, a donja granica sa 15 000 te povećavajući ostale iznose za 1000. Anonimizirani podaci prikazani su tablicom 4.8.

OSOBNI BROJ	POLOG
100001	25 000
100002	13 000
100003	20 000
100004	18 000

Tablica 4.7: Korisnikov polog prije anonimizacije

OSOBNI BROJ	POLOG
100001	25 000
100002	15 000
100003	21 000
100004	19 000

Tablica 4.8: Korisnikov polog nakon anonimizacije

Nasumična varijacija broja je također tehnika varijacije broja koja može biti ojačana ovisno o generiranju nasumičnog broja. Ovaj primjer provodimo na prijašnjem primjeru tablice 4.7. Kao u prethodnom primjeru, donja granica je 15 000, a gornja 25 000 s dodatnom dodjelom broja. Ako se dodijeljen broj nalazi u granicama od 0 do 4, tada se iznos pologa smanjuje za 1000, a ako je broj u rasponu od 5 do 9, tada se polog povećava za

1000, pazeći da ostane u zadanim granicama. Primjenjujući navedene uvjete rezultat je dan tablicom 4.9.

OSOBNI BROJ	POLOG	NASUMIČAN BROJ
100001	25 000	6
100002	15 000	3
100003	21 000	7
100004	17 000	2

Tablica 4.9: Korisnikov polog nakon anonimizacije

Mnogi alati za anonimizaciju podataka dopuštaju korisniku da sam odredi donju granicu, gornju granicu te operator kojim izmjenjujemo vrijednosti.
Ljudski potencijali koriste ovakvu tehniku pri maskiranju plaće zaposlenika ili banke primjenjujući ovakav oblik anonimizaciju na uplate, iznose i druge transakcije svojih korisnika. Ovakva tehnika koristi se isključivo na numeričkim podacima.

4.4 Varijacija datuma

Varijacija datuma je također statistička tehnika koja uključuje generiranje datuma ili varijacija već danog nemaskiranog datuma koji se nalazi u zadanim granicama. Ako je varijacija novog datuma ispod ili iznad dane granice, tada anonimizirani datum postaje upravo donja granica, odnosno gornja. Za primjer ćemo promotriti tablicu sa prikazanim datumima rođenja, gdje je donja granica dana sa 1. siječanj 1975. i gornja granica 1. siječanj 1979. godine te ima povećanje sa 30 dana. Podaci prije anonimizacije prikazani su tablicom 4.10, a nakon korištenja varijante datuma dobili smo podatke u tablici 4.11.

OSOBNI BROJ	DATUM ROĐENJA
100001	1-siječanj-1978
100002	30-rujan-1974
100003	30-rujan-1978
100004	1-listopad-1980

Tablica 4.10: Datum rođenja korisnika prije anonimizacije

Tehnika variranja datuma može biti snažnija koristeći nasumični odabir broja ili datuma koji koristimo pri dodavanju odnosno funkciji koju provodimo prilikom anonimizacije. Većina alata koja podržava ovakvu tehniku anonimizacije dopušta korisniku da odabere gornju i donju granicu te artimetički operator po kojem će se datum mijenjati.

OSOBNI BROJ	DATUM ROĐENJA
100001	31-siječanj-1978
100002	1-siječanj-1975
100003	30-listopad-1978
100004	1-siječanj-1979

Tablica 4.11: Datum rođenja korisnika nakon anonimizacije

Varijaciju datuma koriste odjeli kao što su ljudski resursi kako bi zaštitili datume rođenja svojih zaposlenika te sustavi banaka koji koriste datume rođenja svojih klijenata. Kao što sam naziv tehnike kaže, može se koristiti isključivo na podacima koji su tipa datum.

4.5 Poništavanje

Tehnika poništavanja ili nuliranja je vrlo jednostavan način anonimizacije zamjenjujući kolonu podataka osjetljivih vrijednosti sa NULL vrijednosti. Ovakav način anonimizacije koristan je samo u određenim situacijama. Jedan od uvjeta ove tehnike je nemogućnost korištenja na kolonama čiji se podaci ne mogu prikazati kao NULL vrijednosti, odnosno ne mogu postići NULL vrijednost. Primjer gdje kolona ne može biti NULL je spol osobe koji je uvijek Muško ili Žensko. Usporedno s tim možemo imati kolonu Komentar u kojoj se može nalaziti bilo koja vrijednosti te ju je moguće postaviti na NULL te provesti trenutnu tehniku. Zamjena simbolom je podtehnika poništavanja, umjesto postavljanja vrijednosti na NULL, možemo koristiti bilo koji drugi simobl kao što je razmak, ”D”, ”N” za zamjenu vrijednosti. Za primjer možemo uzeti bazu koja treba biti testirana za određeni projekt, a sadrži kolonu ”Počinio kazneno djelo” koja je osjetljiv podatak. Ako u primjeru kolona sadrži odgovore sa ”D”, ”N”, kako bi zaštitili privatnost korisnika onda možemo cijelu kolonu zamijeniti sa vrijednosti ”N”.
Drugi način i primjer kako zaštititi osobe sa invaliditetom, odnosno uklanjanje identiteta zaposlenih osoba određujući postotka zaposlenika koji će biti označeni sa ”D” kao počinitelji, a ostatku zaposlenika pridružen je ”N”. Prvo su originalni, odnosno nemaskirani podaci prikazani tablicom 4.12.

OSOBNI BROJ	REGIJA	INVALIDITET
100001	Srednjizapad	D
100002	Srednjizapad	N
100003	Srednjizapad	D
100004	Srednjizapad	D
100005	Sjever	N
100006	Sjever	D
100007	Srednjizapad	N
100008	Srednjizapad	N
100009	Sjever	N
100010	Srednjizapad	N
100011	Srednjizapad	N
100012	Srednjizapad	N
100013	Srednjizapad	N
100014	Sjever	N

Tablica 4.12: Invaliditet zaposlenika

Prvi korak je izmiješati redove kolone Regija, a nakon toga odrediti 10% zaposlenika koji će posebno biti obilježeni sa ”D” u koloni Invaliditet. Nakon miješanja kolone Regija rezultat je dan tablicom 4.13.

OSOBNI BROJ	REGIJA	INVALIDITET
100001	Sjever	D
100002	Srednjizapad	N
100003	Sjever	D
100004	Srednjizapad	D
100005	Sjever	N
100006	Srednjizapad	D
100007	Sjever	N
100008	Srednjizapad	N
100009	Srednjizapad	N
100010	Srednjizapad	N
100011	Srednjizapad	N
100012	Srednjizapad	N
100013	Srednjizapad	N
100014	Srednjizapad	N

Tablica 4.13: Invaliditet zaposlenika sa promiješanom regijom

Nakon miješanja redova u koloni Regija, odabiremo u našem slučaju 10% zaposlenika koji će biti označeni u koloni Invaliditet sa ”D”, tako anonimizirani podaci dani su u tablici 4.14.

OSOBNI BROJ	REGIJA	INVALIDITET
100001	Sjever	N
100002	Srednjizapad	N
100003	Sjever	D
100004	Srednjizapad	N
100005	Sjever	N
100006	Srednjizapad	N
100007	Sjever	N
100008	Srednjizapad	D
100009	Srednjizapad	N
100010	Srednjizapad	N
100011	Srednjizapad	N
100012	Srednjizapad	N
100013	Srednjizapad	N
100014	Srednjizapad	N

Tablica 4.14: Anonimizirani podaci sa 10% invaliditeta zaposlenika

U nekim slučajevima, samostalno uklanjanje podataka je dopustivo. Primjer jednog uklanjanja je stvaranje kopije, odnosno podskupa originalnih podataka koji čine bazu za programere i testere. Vlasnik baze na taj način može odrediti da svi podaci o performansama zaposlenika sadržanim u koloni ”Komentar” moraju biti izbrisani. Također postoji i tehnika brisanja koja se provodi ako su zadovoljeni određeni uvjeti. Kao primjer možemo uzeti bazu u kojoj su spremljeni podaci korisnika. Baza sadrži kolonu ”Žalba” gdje je zabilježeno je li se korisnik predao žalbu ili nije te kolonu ”Komentar” u kojoj su opisani detalji žalbe. Uvjet brisanja podataka iz određene kolone, odnosno kolone ”Komentar”, može biti ako je odgovor ”DA” u koloni ”Žalba” onda izbriši podatke iz kolone ”Komentar”.

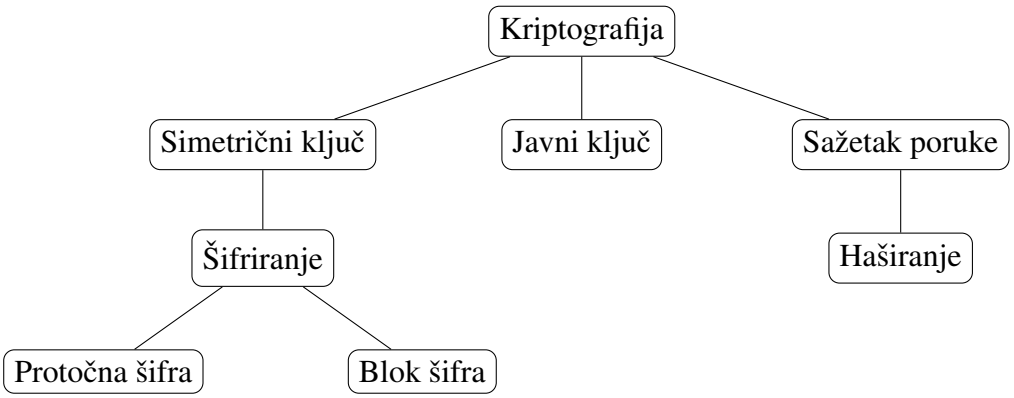
4.6 Maskiranje simbolom

Ovo je jedna od najpopularnijih i najzastupljenijih tehnika koja se koristi za zaštitu podataka. Koristeći ovu tehniku mijenjamo cijelu vrijednost ili dio vrijednosti sa simbolima kao što su X, *, & ili neki drugi sličan. Dužina maskiranih podataka ostaje nepromijenjena. Maskiranje simbolima često kostimo prilikom zaštite brojeva kreditnih kartica, telefonskih brojeva. Broj kreditne kartice 4455 3230 0010 5169 je primjer koji može biti anonimiziran

po dijelovima sa simbolom X. Tako maskirani broj kartice mogao bi izgledati kao XXXX XXXX XXXX 5169.

4.7 Kriptografija

Kriptografija je znanstvena disciplina koja se bavi proučavanjem metoda za slanje poruka u takvom obliku da ih samo onaj kome su namijenjene može pročitati. Kriptografija uglavnom uključuje korištenje termina kao što su otvoreni tekst, odnosno to je nešifrirani tekst koji pošiljatelj želi poslati primaocu tako da originalni tekst šifriramo, šifrirani tekst ili šifrat, šifra, simetrični ključ i mnoge druge. Simetrični ključ je ključ koji koristimo za šifriranje i dešifriranje. Kriptografija je relevantna za anonimizaciju podataka kada je ključna tajnost i integritet podataka te je alternativan način realističnog prikaza podataka. Tehnike kriptografije su relevantnije za dinamička maskiranja i za integracijska testna okruženja u usporedbi sa statičkim testnim okruženjem. Tehnike kriptografije prikazane slikom 4.1 dijele se na simetrični ključ, javni ključ, sažetak poruke.

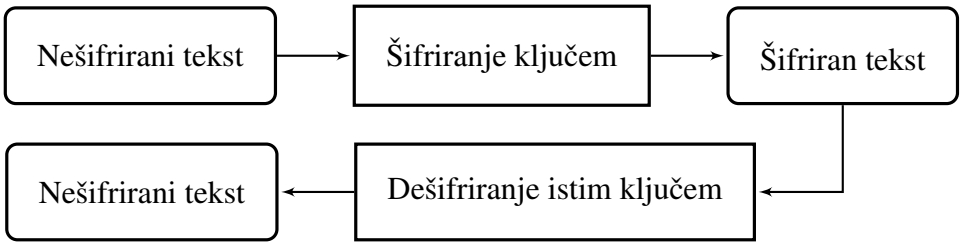


Slika 4.1: Tehnike kriptografije

Simetrični ključ

Tehnika imetričnog ključa uključuje korištenje istog ključa za anonimizaciju podataka kao i za deanonimizaciju. Originalni podaci nazivaju se otvoreni tekst, a nakon šifriranja pomoću danog ključa naziva se šifrat. Ovdje razlikujemo blokovne šifre, kod kojih se obrađuje jedan po jedan blok elemenata otvorenog teksta koristeći jedan te isti ključ K, te protočne šifre (eng. stream cipher) kod koji se elementi otvorenog teksta obrađuju jedan po jedan koristeći pritom niz ključeva (eng. keystream) koji se paralelno generiraju. Proces korištenja simetričnog ključa dan je slikom 4.2. Kod simetričnih kriptosustava, ključ

za dešifriranje se može izračunati poznavajući ključ za šifriranje i obratno. Uglavnom u praksi, ti ključevi su najčešće identični. Sigurnost ovih kriptosustava leži u tajnosti ključa.



Slika 4.2: Proces simetričnog ključa

Protočna šifra

Protočna šifra uzima bit otvorenog teksta i djeluje sa slijednim ključem na taj bit provodeći ga u šifrirani bit. Iako je tehnika protočnog šifriranja vrlo popularna, isto tako je i vrlo ranjiva s obzirom na napade. Jedan od najpoznatijih algoritama korištenih za generiranje slijedne šifre je RC4.

Blokovna šifra

Prilikom blokovnog šifriranja podaci su podijeljeni u blokove bitova, npr. 64-bitne blokove i svaki blok je zasebno šifriran. Dvije implementacije blokovnog šifriranja su DES (engl. data encryption bits) i Trostruki DES. Trostrukti DES je sigurniji algoritam jer se 64-bitni blok podataka šifrira tri puta koristeći tri različita ključa. AES (engl. advanced encryption standard) je relativno novi algoritam koji se koristi za veće blokove od 128 bitova i kao takav zamijenio je DES. Simetrični ključ koji se koristi u AES može biti 128, 192 ili 256 bitni ključ.

Jedna od najbitnijih primjena DES-a je kod banaka i novčanih transakcija, koristio za šifriranje PIN-ova kartica te je bio popularan u uporabi u civilnim satelitskim komunikacijama. Unatoč ogromnoj popularnosti kriptografije kao tehnike zaštite podataka, postoje i neki nedostatci kao u slučaju kada je potrebno podatke prikazati stvarnima. Ključ koji se koristi također mora biti pod kontroliranim pristupom, dostupan samo određenom osoblju te mora biti spremljen na sigurno mjesto.

Javni ključ

U prethodnom poglavlju prikazali smo načine šifriranja koji koriste tajni ključ koji mora biti čuvan na dobrom i sigurnom mjestu. To je i veliki nedostatak jer zahtjeva da pošiljatelj

i primatelj poruke budu u mogućnosti razmijeniti tajni ključ preko nekog sigurnog komunikacijskog kanala. Unatoč tomu u stvarnom svijetu javni ključ nije zamjena za simetični ključ i simetrično šifriranje, već se javni ključ koristi za šifriranje ključeva. Podaci se od mjesta A do mjesta B šalju korištenjem simetričnog ključa koji se razmijenjuje korištenjem javnog ključa. Takav način šifriranja zove se hibridni sustav. Osnovni razlog zašto se javni ključ ne koristi za šifriranje poruka, jest da su algoritmi s javnim ključem puno sporiji (oko 1000 puta) od modernih simetričnih algoritama. RSA je algoritam koji koristi gore opisan način kriptiranja, ali se ne koristi mnogo u anonimizaciji podataka.

Sažetak poruke

Tehnika sažimanja poruke uključuje korištenje hash-funkcije. Hash-funkcija koja uzima poruku fiksne duljine i generira izlaznu šifriranu poruku koju nazivamo hash-rezultat, hash-vrijednost ili jednostavno hash. Tipovi sažimanja poruke koji postoje s obzirom na ključ su:

Nepovratan što znači da ulazni podaci koji su anonimizirani pomoću hash-funkcije ne smiju biti ponovno dohvatljivi, vraćeni u originalni oblik.

Bez kolizije, vrijednosti hash-funkcije moraju osigurati da ne postoji kolizija u šifriranju, odnosno da dvije različite ulazne vrijednosti nemaju istu izlaznu vrijednost.

Deterministično, odnosno jedinstvenost hash-funkcije za isti uzorak, za isti ulazni skup podataka izlazna vrijednost mora biti jednaka.

Najpoznatiji algoritmi koji se odnose na sažimanje poruka su MD5 i SHA-2. MD5 algoritam za ulazni skup podataka uzima poruku proizvoljne duljine i generira izlaznu poruku kao 128-bitni otisak ili sažetak poruke. Tijekom proteklih godina, MD5 se pokazivao kao dosta ranjiv na koliziju. Zbog problema kolizije, uz MD5 najčešće je korišten SHA-1 algoritam za izračunavanje. Uz SHA-1 postoji cijela familija algoritama SHA, SHA-256, SHA-384, i SHA-512 koji se razlikuju po duljini sažetka poruke. SHA-2 je algoritam visoke razine sigurnosti koji je prihvaćen od strane američkih saveznih agencija.

Ovakav način više koristi SSL (engl. Secure Socket Layer), odnosno protokol za prijenos zaštitno kodiranih podataka, a manje se koristi u dinamičkim okolinama.

Poglavlje 5

Djelomična osjetljivost i maskiranje

Proučavajući bazu podataka, tipove podataka koji se unose i koji trebaju biti zaštićeni, možemo uočiti da u određenim područjima, kolonama nisu svi znakovi jednako osjetljivi te da se ne moraju svi anonimizirati. Jedan primjer je kreditni broj kartice koji se sastoji od 12 znamenaka te se posljednje 4 znamenke ne moraju uzeti u obzir. Kao primjer možemo uzeti i JMBG (MBO), odnosno jedinstveni matični broj građana koji na jedinstven način povezuje podatke u službenim evidencijama. JMBG, trinaesteroznamenkasti broj je sastavljen od podataka koje se vežu za određenu osobu, datumom rođenja, regiji rođenja, spolu i kontrolnom broju. Za primjer uzmemo da spol nije osjetljiv podatak te možemo zahtjevati anonimizaciju samo prvih devet znamenaka. U ovakvim slučajevima je potrebno parcijalno odnosno djelomično maskiranje. Slijedećim tablicama prikazani su primjeri djelomičnog, odnosno parijcalno maskiranja.

Nešifrirani broj kreditne kartice	1234567890123456
Djelomično šifrirani broj kreditne kartice	**********3456

Tablica 5.1: Primjer djelomičnog maskiranja

Nešifrirani JMBG	0308964384007
Djelomično šifrirani JMBG	**********007

Tablica 5.2: Primjer djelomičnog maskiranja

Poglavlje 6

Maskiranje s obzirom na vanjske utjecaje

Postoje određene situacije kada se podaci ne mogu promatrati samostalno kao osjetljivi, nego postaju osjetljivi zbog ovisnosti o drugim podacima. Osjetljivost podataka ovisi o više kolona u jednoj tablici ili čak o drugoj ili više njih. Dan je sljedeći primjer kada se koristi maskiranje podataka čije kolone ovise o nekoj drugoj koloni.

U bazi maloprodajne trgovine nalazi se tablica Artikl koja sadrži svu robu trgovine te u tablici postoji stupac Cijena i Datum isporuke dobavljača. Kako bi u ovom slučaju uspjeli sačuvati povjerljivost i tajnost cijene proizvoda u posljednjih 30 dana, zbog ovisnosti Cijene o datumu provodimo takvo maskiranje. S obzirom da želimo sačuvati tajnost u posljednjih 30 dana, ako se Datum isporuke dobavljača nalazi unutar proteklih 30 dana, tada u koloni Cijena za određen proizvod postavljamo dogovorenu vrijednost 200. U ovom primjeru za današnji datum ćemo promatrati 15. veljače čime vidimo da je prvi navedeni artikl unutar proteklih 30 dana te ga je potrebno maskirati.

Proizvodi

Cijena	Datum isporuke
180.59	01.02.2019.
230.00	05.01.2019.

Proizvodi

Cijena	Datum isporuke
200	01.02.2019.
230.00	05.01.2019.

Tablica 6.1: Cijena nakon maskiranja

Poglavlje 7

Pomoćne tehnike anonimizacije

Pomoćne tehnike anonimizacije koriste zapravo jednu ili više osnovnih tehnika anonimizacije za generiranje vrijednosti za anonimiziranje koja je posebno prilagođena za određenu situaciju. Specifične situacije su sa podacima određenog formata kao što je SSN (eng. Social Security Number), JMBG, broj mobitela, e-mail. Prilagođavanje tehnika anonimizacije omogućuje čuvanje formata podatka čak i nakon anonimizacije ili nakon utjecaja jedne ili više osnovnih tehnika anonimizacije kao što su zamjene ili kriptiranje. Ako imamo podatke kao što je telefonski broj, maskirana vrijednost također mora sadržavati razmak na istom mjesto kao i originalna vrijednost, kao prikaz u tablici 7.1.

Nemaskirani broj	Maskirani broj
510-555-4432	201-678-3865

Tablica 7.1: Telefonski broj

Sada ćemo prikazati posebne situacije koje su prilagođene formatu podatka te načine na koje je moguće provesti anonimizaciju.

Tehnika maskiranja SSN formata

SSN je američki broj radnika, jedinstveni identifikacijski broj koji se izdaje i državljanima i nedržavljanima u Sjednjenim Američkim Državama. SSN je broj koji se sastoji od 11 simbola, odnosno od 9 brojeva i 2 razdjelnika, odnosno razmaka koja broj dijele na 3 segmenta. Uporabom pomoćnih tehnika anonimizacije generirat će se nova SSN vrijednost broja koja će imati isti oblik kao i originalna vrijednost i izgledat će stvarno i još uvijek biti neiskorišten SSN.

Područje		Broj grupe		Serijski broj
999	-	99	-	1565

Tablica 7.2: Primjer djelomičnog maskiranja

Format stringa - tehnika maskiranja

Kod maskiranja često se dogodi da određeni niz znakova ima pridružen format koji je generiran te ga je potrebno takvog i zadržati prilikom anonimizacije. Kao primjer možemo uzeti string koji je generiran tako da u sebi sadrži podstring "NY" sa dodatna tri slova ispred NY i dodatna tri broja iza NY. Tehnika za ovakve primjere mora identificirati znakove koji moraju biti maskirani u stringu i nasumično generirati znakove koji će zamijeniti originalne znakove. U danom primjeru podstring NY ostaje nakon maskiranja na istom mjestu kao u originalnom. Navedeni primjer prikazujemo tablicom 7.3.

	Vrijednosti
Ulazna vrijednost	AbcNY123
Ciljani format za određen primjer $*$ označavaju mjesta koja moraju biti maskirana	***NY***
Izlazna vrijednost koristeći maskiranje formata stringa	a@bNY104

Tablica 7.3: Primjer formata stringa

Numerički format - tehnika maskiranja

Mnogo puta smo se mogli susresti sa numeričkim vrijednostima koje su dana određenim formatom pa ih treba i maskirati u tom formatu. Maskiranje numeričkog formata se vrlo često koristi. Kao primjer možemo uzeti broj osobne iskaznice građanina New York-a čiji se broj sastoji od NY na početku te 6 nasumično odabranih brojeva. Ovakvom tehnikom potrebno je identificirati mjesta na kojima se nalaze brojevi te isti unutar stringa moraju biti maskirani, odnosno zamijenjeni nasumičnim generiranjem novih brojeva. U ovom slučaju NY kao mjesto stanovanja građanina ostaje uvijek prikazan na istom mjestu unutar stringa. Ovaj primjer je prikazan tablicom 7.4.

Format kreditne kartice - tehnika maskiranja

Kreditne kartice su u visokorizičnoj skupini podataka koje treba zaštititi. Kako bi zaštititli korisnika ovakva tehnika posebno implementirana za kreditne kartice generira brojeve koji

	Vrijednosti
Ulazna vrijednost	NY899045
Ciljani format za određen primjer $*$ označavaju mjesta koja moraju biti maskirana	NY******
Izlazna vrijednost koristeći maskiranje formata stringa	NY044532

Tablica 7.4: Primjer numeričkog formata

odaju realnost broja kreditne kartice. U ovoj tehnici se uglavnom maskira određeni dio broja kreditne kartice, tako da i dalje izgleda stvarno.

Telefonski broj - tehnika maskiranja

Prilikom maskiranja telefonskog broja, također je dobro i korisno ako anonimizirani podaci izgledaju realno. Ovaj primjer ćemo provesti na američkom tipu broja. Tipično maskiranje telefonskog broja koji sadrži 12 znakova provodi se na način da prve tri znamenke su nemaskirane, a ostale jesu uz dodatak da se na četvrtom i osmom mjestu nalazi ”-” kao separator. Primjer broja dan je tablicom 7.5.

Broj znaka	1	2	3	4	5	6	7	8	9	10	11	12
Nemaskirani broj				-				-				
Algoritam maskiranja	Nemaskirani			-	Anonimizirani brojevi							

Tablica 7.5: Format telefonskog broja

E-mail - tehnika maskiranja

U današnjoj svakodnevici mnogi koriste i više od jednog računa elektronske pošte odnosno e-maila za primanje i slanje povjerljivih informacija pa je i želja za krađom mailova veća. E-mail je sastavljen od više dijelova od korisničkog imena odnosno primatelja, domene elektroničke pošte i nastavka. Prikazat ćemo u tablici primjer maskiranja e-maila. Promatrajući primjer možemo reći da se za korisničko ime rezervira 3 do 15 znakova, za domenu rezerviramo 3 do 10 znakova i za završetak 2 do 8, svi skupa se povežu u cjelinu anonimizirane vrijednosti. Ovakvim pristupom omogućujemo da e-mail korisnika ostane tajan, ali tako anonimiziran izgleda stvarno.

Ovakve tehnike anonimizacije uglavnom se koriste i najisplatljivije su kada generirani anonimizirani podaci trebaju izgledati realno, a originalni podaci trebaju biti sačuvani. U

	e-mail	@	DOMENA	.	NASTAVAK
Nemaskiran	ram.singh	@	coffeebar	.	com
Maskiran	john.devon	@	Syro	.	com

Tablica 7.6: Format e-maila

slučaju da nam je potreban anonimiziran SSN, ali da izgleda stvarno, SSN tehnika maskiranja može generirati valjani SSN.

Poglavlje 8

Primjer

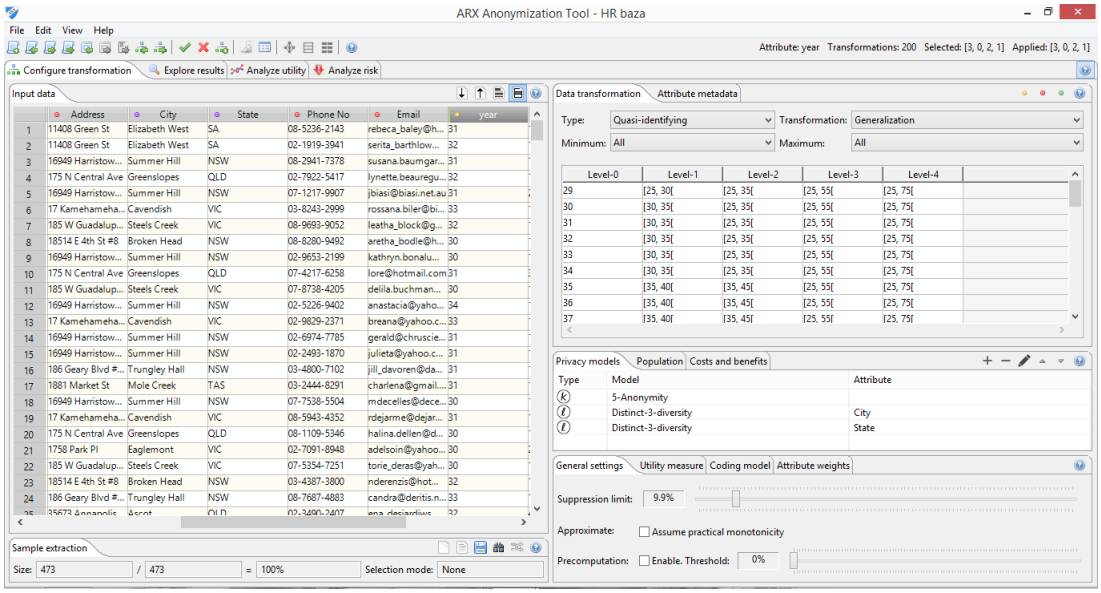
Za primjer koristimo danu bazu podataka, koja se može pronaći na [13], a izvor je sa stranice *My Excel Templates* na [10]. Dane su tri xlsx tablice, jedna je s originalnim podacima "Original", druga je "HR" koja je pripremljena za zaposlenika ljudskih resursa te posljednja koje ja anonimizirana za javnost. Za anonimizaciju tablice koristili smo alat ARX.

ARX je open source softver za anonimizaciju osobnih podataka. Alat je dostupan besplatno i može se skinuti na sljedećem linku: `https://arx.deidentifier.org/downloads/`. Ovim alatom omogućeno je korištenje različitih metoda anonimizacije, što omogućuje direktno uklanjanje atributa te primjenjivanje pravila na kvazi-identifikatore. ARX je prikladan za velike skupove podataka.

Provedena anonimizacija i hijerarhija se nalazi na [12]. Tablica podataka sadrži kolone *ID, First Name, Last Name, Gender, Address, City, State, Phone No, Email, year, Jobcat, Salary, OIB*. Prilikom prve anonimizacije koja je dana excel tablicom HR maknuli smo *Address i OIB*, a *Year i Salary* generalizirali smo pomoću alata ARX prikazanog slikom 8.1.

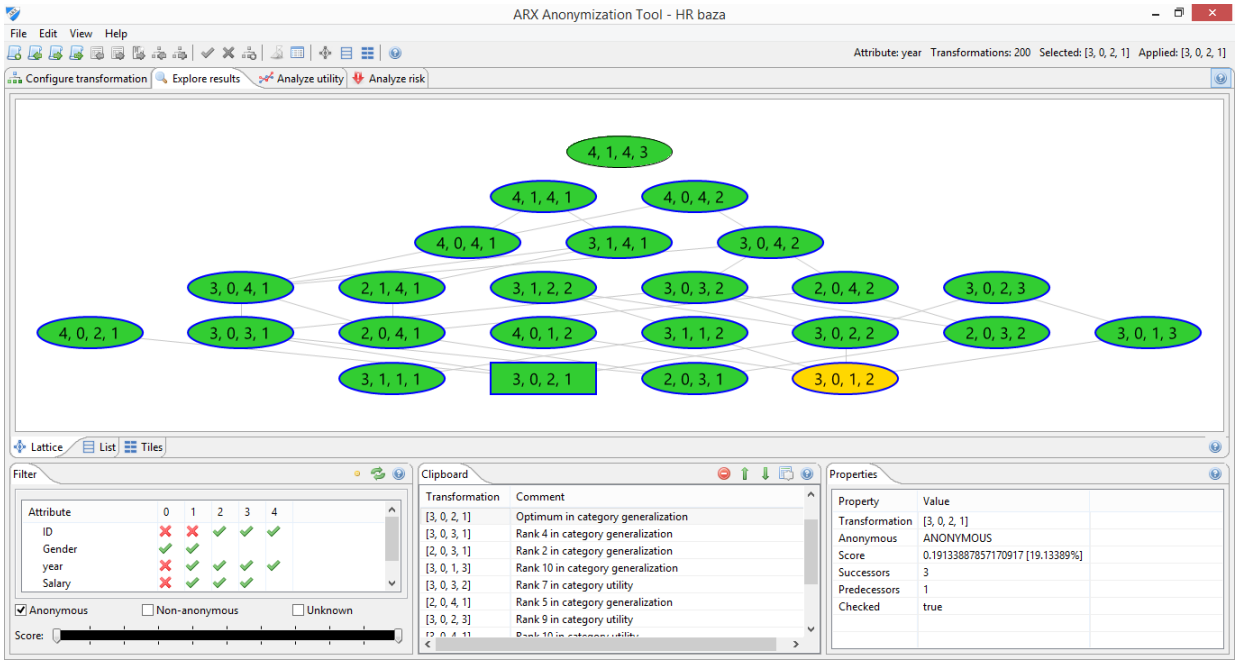
Za drugu anonimizaciju, pripremu tablice za javnost, maknuli smo dodatne osjetljive podatke kao što *Address, Phone No, Email, OIB* te smo proveli određen stupanj generalizacijske hijerarhije za *ID, Year i Salary*. Nakon određivanja tipa svakog atributa, te hijerarhije dobivamo hijerarhijsko stablo slikom 8.2. Na prikazanom stablu možemo

odabrati koju hijerarhiju želimo koristiti, u našem primjeru u tablici za javnost koristili smo generalizaciju (3,0,2,1). Ovim alatom možemo napraviti generalizaciju kako želimo, prikladnu za krajnjeg korisnika kako bi se zaštitili određeni podaci, a krajnji korisnik dobio podatke koje je moguće analizirati i dobiti potrebne povratne informacije.

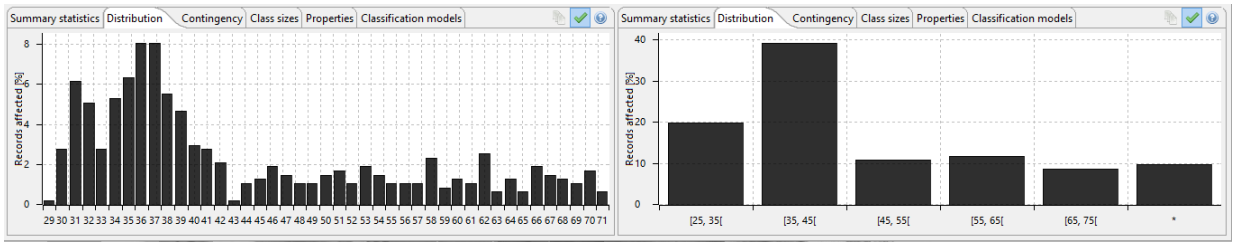


Slika 8.1: Prikaz ARX alata

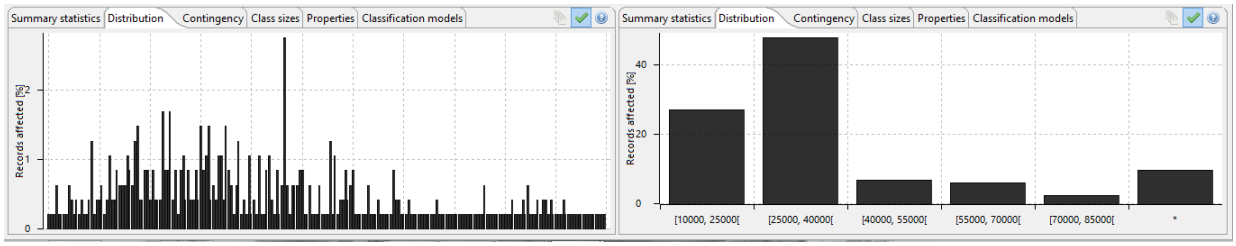
Program ARX sadrži alate koji nam mogu dati analizu podataka, tablice, distribucije podataka. U našem primjeru, možemo vidjeti distribucijske tablice za kvazi-identifikator *Year* i *Salary* prikazanih sa slikama 8.3, 8.4.



Slika 8.2: Dana generalizacijska hijerarhija



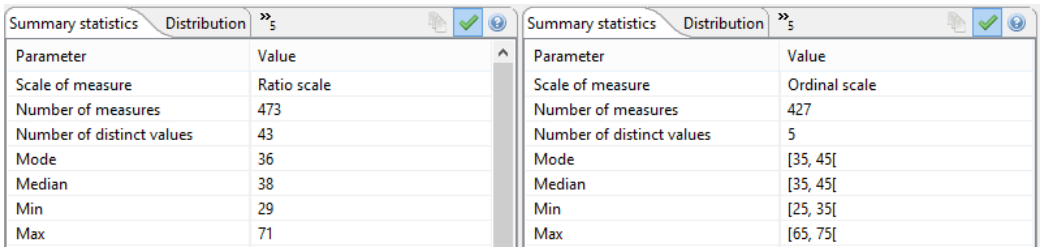
Slika 8.3: Distribucija Year prije i nakon anonimizacije



Slika 8.4: Distribucija Salary prije i nakon anonimizacije

Na prikazanim slikama, možemo vidjeti da se krivulja ponaša podjednako. Na prikazu

godina i kod neanonimiziranih podataka vidimo da je naviše u dobi do početka 40-ih, a nakon toga opadaju, odnosno ponašanje je slično na oba grafička prikaza. Isto tako na grafu za Salary, distribucija se ponaša slično. Već iz samog grafičkog prikaza anonimiziranih podataka, korisnik može zaključiti kakvu plaću većina ljudi ima.



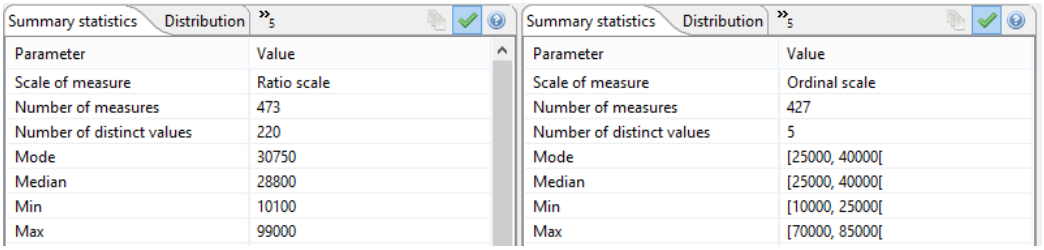
Summary statistics | Distribution

Parameter	Value
Scale of measure	Ratio scale
Number of measures	473
Number of distinct values	43
Mode	36
Median	38
Min	29
Max	71

Summary statistics | Distribution

Parameter	Value
Scale of measure	Ordinal scale
Number of measures	427
Number of distinct values	5
Mode	[35, 45[
Median	[35, 45[
Min	[25, 35[
Max	[65, 75[

Slika 8.5: Summary statistic Year



Summary statistics | Distribution

Parameter	Value
Scale of measure	Ratio scale
Number of measures	473
Number of distinct values	220
Mode	30750
Median	28800
Min	10100
Max	99000

Summary statistics | Distribution

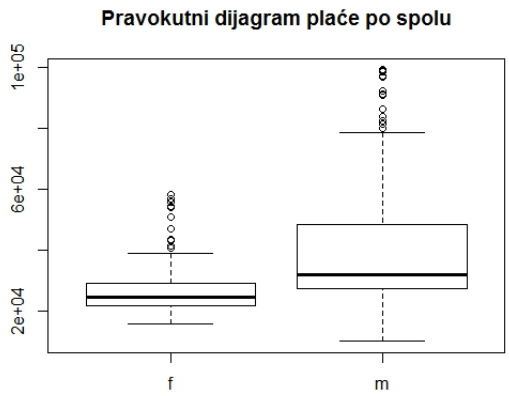
Parameter	Value
Scale of measure	Ordinal scale
Number of measures	427
Number of distinct values	5
Mode	[25000, 40000[
Median	[25000, 40000[
Min	[10000, 25000[
Max	[70000, 85000[

Slika 8.6: Summary statistic Salary

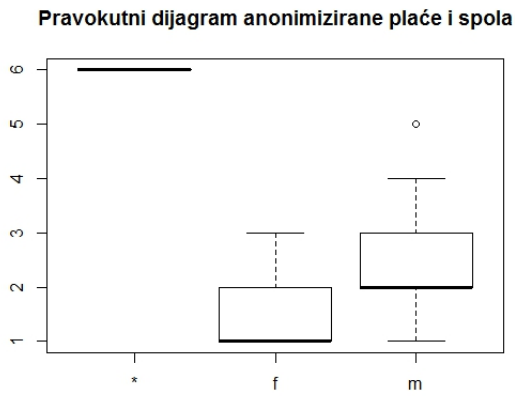
Uz grafički prikaz, postoji i "Summary statistics" prikazan slikom 8.5, 8.6. Iz danih slika imamo da je Median za godine 38, a Median za generalizirane podatke je interval [35,45], te možemo vidjeti da nam 38 odgovara dobivenom intervalu. Isto tako za plaću, Median originalnih podataka je 28000, a Median generaliziranih podataka plaće je interval [25000,40000], te vidimo da 28000 odgovara dobivenom intervalu.

Iz prikazanog dijagama možemo vidjeti da se visina plaće s obzirom na spol ponaša jednako kao i kod anonimiziranih, odnosno generaliziranih podataka. Kod anonimiziranih podataka na slici 8.7 pod (b) brojevi od 1 do 6 predstavljaju sljedeće, 1 = [10000,25000], 2 = [25000,40000], 3=[40000,55000], 4=[55000,70000], 5=[70000,85000], 6=[85000,100000]. Iz danog prikaza vidimo da se npr. plaća unutar kvartilnog pravokutnika originalnih podataka kreće od 20000 do 30000, a to odgovara intervalima označenima sa 1 i 2.

Koristeći alat R, ispitali smo jesu li varijabla spol i varijabla plaća nezavisne. Nulta hipoteza, H_0 je varijable su nezavisne. Testovi su prikazan slikom 8.8. U oba slučaja smo dobili p-vrijednost jednaku 0.00049 što nam kazuje da odbacujemo nultu hipotezu, odnosno varijabla Gender i Salary nisu nezavisne. U navedenom primjeru smo dobili različite



(a) Originalni podaci



(b) Anonimizirani podaci

Slika 8.7: Boxplot plaće i spola

```
> chisq.test(table(t$Gender,t$Salary),simulate.p.value=TRUE, B=2000)

        Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  table(t$Gender, t$Salary)
X-squared = 271.45, df = NA, p-value = 0.0004998

> chisq.test(table(t2$Gender,t2$sal),simulate.p.value=TRUE, B=2000)

        Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  table(t2$Gender, t2$sal)
X-squared = 636.94, df = NA, p-value = 0.0004998
```

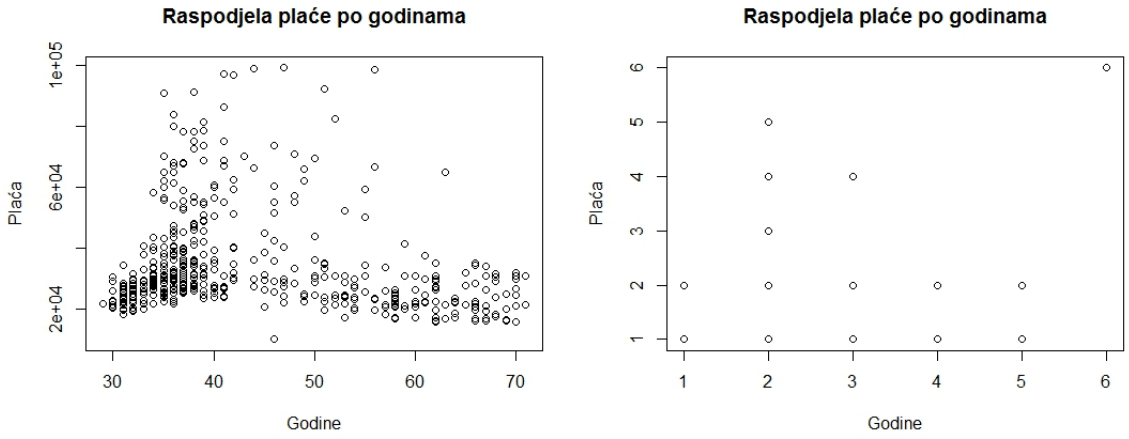
Slika 8.8: Test nezavisnosti Gender i Salary

ishode testova. Prilikom anonimizacije, intervali su dosta veliki te se gubi dosta informacija, te imamo drugačiji ishod testa. Ako smanjimo intervale sa 15 000 na veličinu 5 000, u ovisnosi s originalnim godinama dobit ćemo da su varijable nezavisne kao što daju i originalni podaci. Vidimo da se anonimizacija treba prilagoditi potrebama ili provesti na drugačiji način, u ovom slučaju prilagodbom intervala.

Uz gornji test, napravili smo još jedan test na varijablama Salary i Year prikazan slikom 8.9. U ovom slučaju prvi test nam daje p vrijednost 0.7441 što bi značilo da ne odbacujemo nultu hipotezu, odnosno da su varijable nezavisne, dok u drugom slučaju imamo da je p vrijednost 0.0004998 što ukazuje na odbacivanje nulte hipoteze, varijable nisu nezavisne.

```
> chisq.test(table(t$year,t$Salary),simulate.p.value=TRUE, B=2000) #p=0.0004998

        Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  table(t$year, t$Salary)
X-squared = 8998.9, df = NA, p-value = 0.7441

> chisq.test(table(t2$y.code,t2$sal),simulate.p.value=TRUE, B=2000) #p=0.0004998

        Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  table(t2$y.code, t2$sal)
X-squared = 658.71, df = NA, p-value = 0.0004998
```

Slika 8.9: Test nezavisnosti Year i Salary



(a) Originalni podaci

(b) Anonimizirani podaci

Slika 8.10: Grafički prikaz Salary i Year

Slikom 8.10 prikazana je raspodjela plaća po godinama. Iz prikaza bi mogli zaključiti da su podaci jednako raspršeni, te bi iz grafičkog prikaza mogli pretpostaviti da su varijable zavisne, ali testom smo pokazali da za slučaj (a) imamo da su nezavisne, a slučaj (b) nam p vrijednost daje da su zavisne. Ovim možemo zaključiti kao što sam grafički prikaz nije pouzdan, da i test prilikom grupiranja nije pouzdan. U prikazu (b) gdje su nam podaci generalizirani pod brojem 6 nalaze se podaci koji nisu dodijeljeni ni jednom intervalu, nego su označeni * kao skriveni. Prilikom naše generalizacije imali smo 46 zapisa koji su skriveni. Količinu skrivenih zapisa, veličine klasa, broj klasa i zapisa dan je slikom 8.11. Zavisnost originalnih i anonimizranih varijabli se razlikuje, što nas navodi na zaključak da ovakva anonimizacija za dane podatke nije uspjela te ju je potrebno provesti na drugačiji način, odnosno primjeniti druge tehnike ili druge parametre za istu tehniku. Podaci su

dobiveni programom ARX, te prilikom anonimizacije možemo postaviti koliki postotak skrivanja dopuštamo.

Summary statistics | Distribution | Distribution (table) | Contingency | Class sizes | Properties | »1

Measure	Value (incl. suppressed)	Value (excl. suppressed)
Average class size	19.40909 (4.1034%)	19.40909 (4.54545%)
Maximal class size	85 (17.9704%)	85 (19.90632%)
Minimal class size	5 (1.05708%)	5 (1.17096%)
Suppressed records	46 (9.72516%)	0
Number of classes	22	22
Number of records	473	427

Slika 8.11: Veličina klasa

Bibliografija

[1] Josep Domingo-Ferrer, David Sánchez i Jordi Soria-Comas, *Database Anonymization Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*, Morgan & cLaypool, San Rafael CA, 2016.

[2] Richard Dosselmann, Mehdi Sadeqi i Howard J. Hamilton, *A tutorial on Computing t-Closeness*, Department of Computer Science (2019), https://arxiv.org/pdf/1911.11212v1.pdf.

[3] K.El Emam, F. K. Dankar, R. Issa, E. Jonkers, D.Amyot, E. Cogo, J.P. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, T. Roffey i J. Bottomley, *A globally optimal k-anonimity method for the de-identification of health data, Appendix D: formula used*, American Medical Informatics Association **16**, br. 5.

[4] Khaled El Emam, F. K. Dankar, R. Issa, E. Jonkers, D.Amyot, E. Cogo, J.P. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, T. Roffey i J. Bottomley, *A globally optimal k-anonimity method for the de-identification of health data*, American Medical Informatics Association **16** (2009), br. 5, 670—-682, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2744718/pdf/670.S1067502709001236.main.pdf.

[5] I. P. Fellegi i A. B. Sunter., *A theory for record linkage*, Journal of the American Statistical Association **64** (2012), br. 328, 1183–1210, https://www.tandfonline.com/doi/abs/10.1080/01621459.1969.10501049.

[6] Kristen LeFevre, David J. DeWitt i Raghu Ramakrishnan, *Incognito: Efficient Full-Domain K-Anonymity*, Association for Computing Machinery (2005), 49–60, https://dl.acm.org/doi/pdf/10.1145/1066157.1066164?download=true.

[7] N. Li, T. Li i S. Venkatasubramanian, *t -closeness: privacy beyond k-anonymity and l -diversity*, 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey **1526** (2007), br. 23, 106–115, https://doi.org/10.1109/ICDE11332.2006.

[8] B. Raghunathan, *Complete Book Of Data Anonymization - From Planning To Implementation*, CRC Press, Boca Baton FL, 2013.

[9] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez i S. Martinez, *Enhancing data utility in differential privacy via microaggregation-based k-anonymity*, The VLDB Journal **23** (2014), br. 5, 771–794, https://dl.acm.org/doi/pdf/10.1007/s00778-014-0351-4?download=true.

[10] My Excele Templates, *Employee Database Managment*, http://myexceltemplates.com/employee-database-management/.

[11] Vincenc Torra i Josep Domingo-Ferrer, *Record linkage methods for multidatabase data mining*, Information Fusion in Data Mining **123** (2003), br. 1, 101–132, https://www.researchgate.net/profile/Josep_Domingo-Ferrer/publication/243784444_Record_linkage_methods_for_multi-database_data_mining/links/53d57f090cf2a7fbb2ea5a5e/Record-linkage-methods-for-multi-database-data-mining.pdf.

[12] A. Vrljić, *ARX*, https://drive.google.com/open?id=1QaxnxlwN6pv0qwTwIuISkOOyHCMui8ev.

[13] Vrljić, *Excel baza*, https://docs.google.com/spreadsheets/d/1Ccs4cFsG32yMMGoRy9sewmg_3cmFBOBIfdgFy8wiBzU/edit?usp=sharing.

Sažetak

U današnjem društvu sve je veća potreba za javno dostupnim podacima koji bi se koristili u daljnjoj obradi. Povećanjem krađa i zloupotrebe osobnih podataka korisnika i zaposlenika, rezultiralo je uvođenjem propisa zaštite podataka od strane vlada i izvršnih vlasti. Razvijene su razne metode tzv. anonimizacije podataka kojima se u okvirima statistike i računarstva nastoji riješiti problem zaštite podataka.

Anonimizacijske tehnike oblikuju podatke iz originalnog oblika u anonimiziran. Klasične anonimizacijske tehnike uključuju zamjenu, brojčano odstupanje, poništavanje, maskiranje znakom, miješanje podataka, te tehnike kriptografije kao šifriranje, hashiranje. Većina ovih tehnika ima daljnje varijacije.

Iako vanjski hakerski napadi na poduzeće mogu biti spriječeni putem mreže i sigurnosnim mehanizmom, sprječavanje zloupotrebe osjetljivih podataka ostvarivo je programom anonimizacije podataka obuhvaćajući upravljanje, procesiranje, obučavanje, alate, tehnike, sigurnost i pravila o privatnosti. Anonimizacijom podataka omogućujemo zadržavanje podataka koji se ne mogu povezati sa pojedincem, ali omogućuje daljnju upotrebu za analize koje se koriste u današnjem svijetu radi predviđanja, znanstvenih istraživanja. Analiza anonimiziranih podataka ovisi o stupnju anonimizacije. Različitim korisnicima su bitni različiti podaci pa su stoga za jednu bazu potrebne i drugačije anonimizacije, sa različitim stupnjevima sigurnosti. Kako bi se mogli analizirati određeni podaci, potrebno je obaviti anonimizaciju koja će davati dovoljno informacija, ali da ne ugrožavajući osobne podatke.

Summary

In today's world, there is great demand on sharing of person-specific data for further analysis. Increasing theft and misuse of personal data influenced the development of data anonymisation, which solves data protection issue combining statistics and computer science. Anonymisation techniques format data from original to anonymous. Typical anonymisation techniques include substitution, shuffling, nulling, character masking, number variance, hashing and their variants. Using data anonymisation enables saving the data which is not linkable with an individual but is usable for data analysis. Analysis of anonymised data depends on the degree of anonymisation. Different users have different requests that require several anonymisation for one database. To analyse data, anonymised data must keep enough information and at the same time not compromise personal data.

Životopis

Rodena sam 31. svibnja 1994. u Slavonskom Brodu. Živim u Starom Slatiniku, pohađala sam Osnovnu školu "Ivan Mažuranić", Sibinj. Nakon osnovne škole, završila sam Klasičnu gimnaziju fra Marijan Lanosović s pravom javnosti u Slavonskom Brodu. Preddiplomski studij matematike upisala sam 2013. godine na Prirodoslovno-matematičkom fakultetu u Zagrebu na kojem sam 2017. godine stekla titulu univ. bacc. math. i upisala diplomski studij Matematička statistika. Tijekom studiranja u župi "Duha Svetoga" započela sam volontiranje podučavajući djecu ljepotama matematike te studentski radila.