Zasebnost podatkov v odprtih podatkovnih množicah

Žan Povše[1], Marko Kompara[1], Lili Nemec Zlatolas[1]
[1]UM FERI, Koroška cesta 46, 2000 Maribor
zan.povse@um.si, marko.kompara@um.si, lili.nemec.zlatolas@um.si

Izvleček

V raziskavi smo obravnavali zasebnost osebnih podatkov v odprtih podatkovnih množicah. Opredelili smo glavna tveganja, ki nastajajo pri objavi podatkov, ter predstavili sodobne tehnike zaščite zasebnosti, med njimi anonimizacijo in psevdonimizacijo.. Analizirali smo učinkovitost teh metod ter preostala tveganja, ki bi kljub zaščiti lahko omogočila ponovno identifikacijo posameznikov.
V empiričnem delu smo izvedli analizo izbranih slovenskih odprtih podatkovnih množic, ki je potrdila realne možnosti ponovne identifikacije in poudarila potrebo po odgovornem ravnanju s podatki. Predstavili smo primer pravilne anonimizacije kot model dobre prakse. Prikazali smo tudi pravni okvir na ravni Evropske unije in Republike Slovenije s posebnim poudarkom na Splošni uredbi o varstvu podatkov. Ugotovitve raziskave so potrdile nujnost povezovanja tehničnih in pravnih ukrepov za učinkovito in trajno zaščito zasebnosti posameznikov.

Ključne besede: odprti podatki, zasebnost podatkov, anonimizcija, GDPR, pravni okvir, zaščitne tehnike

Data Privacy in Open Data

Abstract

In our research, we examined the privacy of personal data in open data sets. We identified the main risks arising from the publication of data and presented contemporary privacy protection techniques, including anonymisation, pseudonymisation, as well as more advanced approaches. We analysed the effectiveness of these methods and the residual risks that, despite protective measures, could still enable the re-identification of individuals.
In the empirical part, we conducted an analysis of selected Slovenian open data sets, which confirmed real possibilities of re-identification and highlighted the need for responsible data management. We presented an example of proper anonymisation as a model of good practice. We also outlined the legal framework at the level of the European Union and the Republic of Slovenia, with particular emphasis on the General Data Protection Regulation. The findings confirmed the necessity of combining technical and legal measures for the effective and lasting protection of individual privacy.

Keywords: open data, data privacy, anonymization, GDPR, regulatory frame, protection tehniques

1 UVOD

V zadnjem desetletju so odprti podatki postali eno ključnih gonil znanstvenega napredka, transparentnosti in inovacij [1]. Omogočajo širšo dostopnost informacij, spodbujajo sodelovanje med institucijami ter podpirajo razvoj novih rešitev na področjih od medicine do umetne inteligence [2]. Hkrati pa odpiranje podatkovnih virov prinaša tudi številne izzive, zlasti pri varstvu osebnih podatkov in zagotavljanju skladnosti z zakonodajo. Evropski in slovenski pravni okvir GDPR ter Zakon o varstvu osebnih podatkov (ZVOP-2) določata jasna pravila za zaščito posameznikov pred zlorabo njihovih podatkov, še posebej tam, kjer so prisotni občutljivi in individualni zapisi.

V tem članku se bomo osredotočili na raziskavo slovenskih podatkovnih množic, ki vsebujejo individualne podatkovne zapise, ter preučili njihovo ustreznost za javno objavo. Poseben poudarek bomo namenili postopkom pravilne anonimizacije, saj ta predstavlja temeljni mehanizem za varno objavljanje podatkov, ne da bi pri tem ogrozili zasebnost posameznikov. Namen prispevka je pokazati, kakšno je trenutno stanje anonimizacije v slovenskih odprtih množicah z individualnimi zapisi ter prikazati načine za izboljšavo.

2 ODPRTI PODATKI

2.1 Definicija in pomen odprtih podatkov

Koncept odprtih podatkov se nanaša na idejo, da bi morali biti določeni podatki dostopni vsem in to brez specifičnih pooblastil ali finančnih zmožnosti. Podatki pa morajo imeti odprto pravico uporabe in distribucije [3]. Ti podatki se po večini nanašajo na podatke, ki jih zbirajo javni organi in uradi med svojim delovanjem, vendar so možne tudi pridobitve podatkov zasebnih podjetij ali raziskovalnih institucij.

Glavne značilnosti odprtih podatkov so:

- univerzalna dostopnost: vsakdo mora imeti pravico do dostopa, uporabe in distribucije takšnih podatkov in ne sme biti diskriminiran glede na svoj izvor ali pripadnost katerikoli skupini,
- dostopnost: podatki moraj biti v celoti dostopni preko spleta ali drugačnih podobnih virov, ki pa ne smejo presegati razumnih stroškov reprodukcije,
- strojna berljivost: podatki bi morali biti v obliki, ki omogoča enostavno obdelavo (npr. JSON, XML, CSV),
- ponovna uporaba in distribucija: podatki morajo biti na voljo tako, da dovoljujejo ponovno uporabo in distribucijo, kar tudi vključuje povezavo z drugimi podatkovnimi nizi.

Pojem podatkov je večplasten. Na gospodarski ravni predstavljajo podatki zelo dragocen vir za inovacije, razvoj novih inovativnih produktov in optimizacijo že obstoječih rešitev. Analize kažejo na gospodarski potencial, ki izhaja iz uporabe podatkov javnega sektorja [4]. Na družbeni ravni odprti podatki prispevajo k večji transparentnosti javne uprave, kar omogoča državljanom in različnim združbam vpogled v delovanje države in v njeno porabo virov. Omogočeno pa je tudi sodelovanje javnosti v reševanju družbenih problemov. Na znanstveni ravni odprti podatki omogočajo dostop raziskovalcem do velike količine uporabnih podatkov. Ta dostop raziskovalcem olajša raziskovanje ter pospešuje odkritja in sodelovanje.

Za spodbujanje objave podatkov so bile ustvarjene različne iniciative, kot so: nacionalni portali za odprte podatke (OPSI – Odprti podatki Slovenije [5]), evropski (EU Open Data Portal [3]), ameriški (Data.gov [6]) in drugi globalni centri. Ti centri služijo kot največje točke za pridobivanje, iskanje, dostop in prenos odprtih podatkovnih množic iz vseh mogočih področij.

2.2 Pravni okvir varstva osebnih podatkov (GDPR, ZVOP-2)

Kljub koristnim aspektom odprtih podatkov, ti s svojo odprtostjo ne smejo kršiti temeljne pravice posameznikov do zasebnosti. Ta pravica je v Evropski uniji dobro pravno zavarovana. Na področju imamo dva aktivna akta:

- Splošna uredba o varstvu podatkov (GDPR – General Data Protection Regulation): Uredba (EU) 2016/679 [7], ki se uporablja v vseh državah članicah EU od maja 2018 in predstavlja glavni temelj evropskega sistema varstva osebnih podatkov. GDPR določa pravila glede obdelave podatkov fizičnih oseb in pravila glede pretoka teh podatkov. Cilj uredbe je, da se uredi varnost podatkov po celotni EU in okrepi pravice vseh posameznikov do zasebnosti njihovih podatkov. Za obdelovalce osebnih podatkov so zakonsko določena posebna pravila. [7]
- Zakon o varstvu podatkov (ZVOP-2): v slovenski pravni svet prinaša določbe GDPR, ki jih Evropska Unija prepušča nacionalni uredbi. Ureja specifične vidike varstva podatkov v Sloveniji, vključno z delovanjem nacionalnega nadzornega organa (tj. informacijskega pooblaščenca). [8]

GDPR navaja, da se načela varstva podatkov ne uporabljajo za anonimne informacije oziroma informacije, ki se ne nanašajo na določljivega posameznika, ali za osebne podatke, ki so anonimizirani tako, da posameznik iz podatkov ni več določljiv (Izjava 26 GDPR [9]). Če so podatki učinkovito anonimizirani, njihova objava v javne podatke ni več v pristojnosti uredbe GDPR. Tako velika stopnja anonimizacije je velik izziv.

Če podatki niso popolnoma anonimizirani, ampak so samo psevdonimizirani (obdelani tako, da jih ni več mogoče pripisati specifičnemu posamezniku brez uporabe dodatnih informacij, ki se hranijo ločeno in zanje veljajo tehnični in organizacijski ukrepi), potem za njih veljajo vsa pravila GDPR. Psevdonimizacija velja za ustrezen varnostni ukrep, ki pomaga pri izpolnjevanju zahtev GDPR, vendar sama ne izključuje pristojnosti uredbe.

Pred objavo podatkovne množice, ki lahko vsebujejo ali zagotovo vsebujejo osebne podatke, morajo lastniki podatkov skrbno predvideti tveganja in zagotoviti skladnost z GDPR. Pogosto se pri tem vključi oceno učinkovitosti varstva podatkov – DPIA (ang. Data Protection Impact Assessment). Ocena mora biti izvedena, ne glede na to ali so bili podatki masovno obdelani ali ne.

2.3 Osebni podatki, občutljivi osebni podatki in kvazi-identifikatorji

Za pravilno razumevanje in uporabo pravnih okvirjev ter razumevanje tveganj je ključno definirati vse vrste podatkov, ki so predmet varstva [10]:

- Osebni podatek: v četrtem členu GDPR [11] je osebni podatek »katerakoli informacija v zvezi z določenim ali določljivim posameznikom; določljiv posameznik je tisti, ki ga je mogoče neposredno ali posredno določiti, zlasti z navedbo identifikatorja kot je ime, identifikacijska številka, spletni identifikator, podatek o lokaciji ali navedba dejavnikov, ki so značilni za karakteristike identitete posameznika.« Poudariti moramo, da definicija »določljivega posameznika« opisuje vsa možna sredstva, za katera je verjetno, da jih bo katerakoli oseba uporabila za identifikacijo. Pri presoji, ali so podatki osebni, moramo tudi upoštevati možnost povezave z drugimi podatkovnimi množicami ali viri podatkov.
  - Kvazi-identifikatorji (posredni identifikatorji) so atributi, ki sami po sebi običajno ne omogočajo neposredne identifikacije posameznika, vendar lahko v kombinaciji z drugimi podatki postanejo dovolj specifični, da razkrijejo identiteto. Takšni podatki so posebej problematični pri odprtih ali javno objavljenih podatkovnih množicah, saj lahko napadalci s križanjem različnih virov podatkov izvedejo uspešne napade re-identifikacije. Običajni primeri kvazi-identifikatorjev so: datum rojstva, kraj ali poštna številka, spol, poklic, nacionalnost, podatki o lastništvu ipd. Pri pripravi podatkov za objavo je zato nujno upoštevati prisotnost teh posrednih identifikatorjev in zagotoviti ustrezne metode anonimizacije, ki zmanjšajo tveganje ponovne identifikacije posameznikov.
  - Neposredni identifikatorji: podatki, ki sami po sebi omogočajo direktno identifikacijo posameznika. Primeri so: EMŠO, ime in priimek, davčna številka, fotografija obraza, številka osebnih dokumentov ipd.
- Občutljivi osebni podatki: 9. člen GDPR opredeljuje posebno kategorijo osebnih podatkov. Ti podatki so praviloma prepovedani za obdelavo, razen če so obdelani pod strogimi pogoji. To so podatki, ki razkrivajo: raso, politično mnenje, versko prepričanje, članstvo v sindikatu, genetiko, zdravstveno stanje ipd. Množice, ki vsebujejo te vrste podatkov, zahtevajo posebej visoko stopnjo previdnosti in terjajo tudi izrecno privolitev posameznika ali drugo močno pravno podlago in izjemno visoko stopnjo anonimizacije.

2.4 Razmerje med odprtostjo podatkov in varstvom zasebnosti

Po opisu zgoraj je jasno razvidno, da obstaja razmerje med cilji odprtih podatkov in zahtevami, ki jih moramo implementirati za namene varstva osebnih podatkov. Po eni strani je prisotna želja po odprtosti podatkov in napredovanju, ki ga podatki lahko omogočijo, po drugi strani pa ti podatki ne smejo vsebovati vseh podrobnosti, zaradi katerih bi bila ogrožena zasebnost posameznikov.

Iskanje ravnotežja med tema dvema predpostavkama je kompleksen problem. Podatki, ki so preveč obdelani ali pa imajo preveč postavk izbrisanih, imajo dobro možnost, da postanejo neuporabni ali pa neoptimalni za predvidene namene. Preveč restriktiven pristop k varovanju podatkov lahko vodi k slabi uporabnosti množice za raziskave ali analize. Po drugi strani pa preveč odprt pristop k objavi lahko vodi do velikih kršitev pravic zasebnosti posameznikov v teh podatkih.

3. TVEGANJA ZA ZASEBNOST V ODPRTIH PODATKOVNIH MNOŽICAH

Objava odprtih podatkovnih množic je lahko zelo koristna, vendar prinaša velika tveganja za zasebnost posameznikov. Samo odstranjevanje podatkov, kot

so ime, priimek ali EMŠO, ne zadostuje za ustrezno zaščito zasebnosti. Napadalci lahko z uporabo različnih programov, tehnik in drugih virov izkoristijo podatke za pridobivanje identifikatorjev iz množic, s čimer napadejo zasebnost sodelujočih v množici. V naslednjem podpoglavju so predstavljena in kategorizirana glavna tveganja, ki lahko omogočajo napade.

3.1 Neposredna in posredna identifikacija

Največje osnovno tveganje predstavlja možnost identifikacije posameznika iz odprte množice. Poznamo [12] :

- Neposredna identifikacija: do nje pride, če podatkovna množica vsebuje identifikatorje, kot so EMŠO, številke dokumentov, imena, priimki in podobno. V kontekstu odprtih podatkov se od upravljalca pričakuje, da odstrani ali anonimizira vsaj te najočitnejše pokazatelje identitete posameznikov. Objava takšnih podatkov brez privolitve ali pravne podlage je huda kršitev predpisov o varovanju osebnih podatkov.
- Posredna identifikacija: visoko tveganje nastane pri objavi množic, ki vsebujejo kvazi-identifikatorje. Ti omogočajo posredno identifikacijo, tj. ko kvazi-indetifikatorji s kombiniranjem postanejo dovolj specifični, da se spremenijo v polnomočne identifikatorje. Na primer, z objavo osebnega imena in kraja bivanja v manjših krajih z lahkoto popolnoma identificiramo posameznika, zato je to kršitev varstva podatkov. Nevarnost posredne identifikacije je, da izdajatelj podatkov ne more predvideti vseh načinov, s katerimi bi lahko napadalec ustvaril polnomočne identifikatorje, lahko pa poskusi predvideti tveganje za nevarnost razkritja glede na:
  – velikost populacije, na kateri so podatki ustvarjeni,
  – dostopnost podatkov, ki bi lahko vsebovali druge kvazi-identifikatorje, s katerimi bi lahko napadalec kombiniral svoje podatke in pridobil polne identifikatorje ter
  – število in tip kvazi-identifikatorjev v množici.

3.2 Napadi re-identifikacije

Re-identifikacijski napadi (ang. Linkage Attacks) so napadi na objavljene anonimizirane podatke. Napadalci jih izvedejo s povezovanjem podatkov iz različnih virov. Uporabijo dve množici, eno z anonimiziranimi podatki in drugo, ki vsebuje neposredne identifikatorje ali kvazi-identifikatorje. Nato povežejo skupne atribute obeh množic, s čimer pridobijo osebne identifikatorje iz anonimizirane množice [13].

Primer tovrstnega napada je predstavila dr. Latanya Sweeney [14], ki je prikazala, da za identifikacijo velikega odstotka ameriške populacije zadostujejo zgolj trije kvazi-identifikatorji (poštna številka, datum rojstva in spol). V svoji raziskavi je anonimizirane zdravstvene podatke s tremi kvazi-identifikatorji in diagnozami kombinirala z javno dostopnimi volilnimi seznami ter tako identificirala takratnega guvernerja Massachusettsa in njegove osebne zdravstvene podatke.

Napade re-identifikacije se lahko izvede na več načinov:

- s povezovanjem podatkov znotraj iste množice ali s podatki iz drugih odprtih množic: prisotnost iste osebe znotraj več (odprtih) podatkovnih množic omogoča povezavo kvazi-identifikatorjev med množicami, kar razkrije obširnejši profil osebe in lahko vodi v identifikacijo osebe ali razkritje občutljivih podatkov;
- s povezovanjem z zunanjim virom: če je napadalcu omogočen dostop do zunanjih virov podatkov, kot so baze in registri podatkov s prostimi identifikatorji ali kvazi-identifikatorji, lahko uporabi te množice za povezavo oziroma kombiniranje podatkov.

Tveganja za re-identifikacijske napade se povečujejo z vse večjo dostopnostjo podatkov, ki je posledica vseh zapisov na spletu, od tistih na družbenih omrežij do različnih javnih evidenc. Tako se povečuje nabor podatkov za možnost povezave in identifikacije posameznika. Drugo tveganje je tudi v razvoju tehnologije, ki omogoča hitrejše in natančnejše povezovanje podatkov [15]. Hitro se razvijajo tudi novi algoritmi za povezovanje podatkov, kot so verjetnostno povezovalni algoritmi, ki ne potrebujejo istih kvazi-identifikatorjev, da bi našli povezavo.

3.3 Napadi sklepanja

V primeru neuspeha napadalca, da bi povezal podatke in izvedel re-identifikacijo, lahko iz množice izlušči občutljive podatke s sklepanjem (ang. Inference attacks). Ti napadi izkoriščajo korelacije med atributi in statistično verjetnostjo lastnosti podatkov. Poznamo več vrst [13]:

- sklepanje o atributih: ta napad se zgodi, ko napadalec uspešno ugotovi vrednost občutljivega

atributa (dohodek, bolezen, vera, ipd.), za tega določenega uporabnika pa že pozna identiteto oziroma vključenost v drugi podatkovni množici. Tudi ob anonimizirani množici lahko napadalec odkrije občutljive podatke, če so skupine z občutljivimi podatki in kvazi-identifikatorji enolične, zato moramo paziti, da se take skupine ne pojavljajo;

- sklepanje o članstvu: cilj tega napada ni odkrivanje vrednosti atributov, ampak informacija o prisotnosti posameznika v podatkovni množici. Podatki o prisotnosti v določenih podatkovnih množicah so lahko občutljivi podatki. Na primer, če je objavljena študija določene bolezni v obliki anonimiziranih podatkov, lahko z napadom članstva iz te množice izvemo, ali ima posameznik to bolezen. Sklepanje o članstvu pogosto izkorišča možnost, da se modeli strojnega učenja, trenirani na določenem nizu podatkov, obnašajo drugače za vnose, ki so bili del učne množice, kot za tiste, ki to niso bili.

Napadi sklepanja so najpogosteje uporabljeni in najnevarnejši pri objavi statističnih podatkov ali analiz (pogosto modelov strojnega učenja), ki so bili v osnovi občutljivi podatki. Napadalec lahko take podatke uporabi za rekonstrukcijo izvornih podatkov.

3.4 Napadi na homogenost podatkov, neenakomernost podatkov in napadi s predznanjem

Ti napadi (ang. homogenity, skewness and background knowledge attacks) so narejeni za izpodbijanje osnovnih anonimizacijskih tehnik, kot je k-anonimnost (ta zahteva, da zapis ni razločljiv do vsaj k-1 drugih zapisov) [16].

- Napad na homogenost: ta napad se uporablja, kadar vsebujejo vsi zapisi znotraj k-anonimnih zapisov (ekvivalenčna skupina, ki ima skupne vrednosti kvazi-indentifikatorjev) isto občutljivo vrednost atributa. Če napadalec ve za določenega posameznika, ki pripada tej ekvivalenčni skupini, lahko izve vrednost njegovega občutljivega atributa, čeprav ne ve, kateri zapis pripada posamezniku. Na primer, trije posamezniki (k=3) s poštno številko 2000, starostjo med 40 in 50 let, ki so moškega spola imajo v anonimizirani podatkovni množici zdravstvenega izvora povišan krvni tlak. Napadalec, ki ve za soseda Mateja, ki ustreza določenim kvazi-identifikatorjem, poveže te s tistimi iz množice in tako izve, da ima njegov sosed povišan krvni tlak. Pojav teh napadov je bil povod za ustvarjenje novih modelov kot je l-raznolikost (poglavje 4).
- Napad s predznanjem: napadalec uporabi svoje zunanje podatkovne vire, da obide določene zaščitne mehanizme. Če skupina izpolnjuje l-raznolikost (vsebuje vsaj l različnih vrednosti občutljivega atributa), lahko napadalec z znanjem zoži nabor možnih vrednosti za posameznika. Na primer, če napadalec ve, da ima sosed Matej (v ekvivalenčni skupini) zagotovo raka, in če so v l-raznoliki skupini prisotne občutljive vrednosti »rak na slinavki«, »srčna bolezen« in »dermatitis«, lahko napadalec z veliko verjetnostjo sklepa, da ima njegov sosed Matej raka na slinavki. Obstoj takih napadov je dokaz, da nobena tehnika anonimizacije ni zmožna zaščite pred vsemi načini napada z zunanjim znanjem, kar je spodbudilo razvoj modela t-bližina (poglavje 4).
- Napad na neenakomernost (ang. skewness) podatkov: izkorišča neuravnoteženo porazdelitev občutljivega atributa v podatkovni množici. Določena vrednost atributa v množici izstopa, kar napadalcu zmanjša negotovost o tem atributu. Na primer, če imamo podatkovno množico ljudi s podatki o boleznih, v kateri ima 90 % ljudi v množici gripo, lahko z visoko gotovostjo (90 %) sklepamo, da ima posameznik, ki ga poizkušamo re-identificirati in pripada tej množici, gripo, čeprav obstaja več različnih bolezni in 10 % verjetnost, da ima katero od teh [17].

3.5 Primeri znanih incidentov kršitev zasebnosti

Obstaja kar nekaj znanih incidentov, v katerih so se zgodile identifikacije posameznikov ali razkritja drugih osebnih podatkov kljub domnevni anonimizaciji. Eden najbolj izpostavljenih primerov so iskalni dnevniki podjetja AOL [18]. Podjetje je objavilo veliko anonimizirano množico iskalnih dnevnikov 650.000 uporabnikov. Čeprav so bila uporabniška imena zakrita in zamenjana z naključnimi številkami, je novinarjem New York Timesa uspelo na podlagi vsebine iskalnih poizvedb razpoznati nekatere uporabnike. Iz tega incidenta je razvidno, kako lahko na videz nepomembni podatki, če so objavljeni v dovolj veliki količini, razkrijejo identiteto posameznikov.

Podoben izziv se jezgodil pri tekmovanju Netflix prize kjer je podjetje Netflix objavilo anonimizirano množico podatkov ocen filmov svojih uporabnikov,

da bi organiziralo tekmovanje v izboljšanju sistema priporočil vsebin uporabnikom. Raziskovalca Narayanan in Shmatikov [19] sta pokazala, kako se poveže podatke te množice in množice ocen na strani IMDB, kjer jih uporabniki podajajo s pravimi imeni. Če bi napadalec poznal le nekaj filmov, ki so jih ti uporabniki ocenili na obeh platformah in datum, kdaj so podali svoje ocene, bi s povezovanjem podatkov razkril identitete drugače anonimnih ocenjevalcev na platformi Netflix. Ta raziskava je odkrila zasebne preference uporabnikov ter, ali so uporabniki naročniki Netflix-a. Tekmovanje je nato podjetje Netflix ukinilo zaradi težav z zasebnostjo uporabnikov.

Pogosto citiran je tudi primer podatkov Massachusetts GIC raziskovalke dr. Latanye Sweeney [14], ki je povezala množico zdravstvenih podatkov z volilnimi seznami in tako pridobila občutljive podatke takratnega guvernerja. To je klasičen primer re-identifikacije z uporabo javnih podatkov in kvazi-identifikatorjev.

Avstralska vlada (MBS/PBS) [20] je tudi imela probleme, ko je leta 2018 objavila obsežen anonimiziran podatkovni vzorec, ki je vseboval podatke o zdravniških storitvah (MBS) in zdravilih, izdanih na recept (PBS). Ta podatkovna množica je vsebovala podatke približno 10 % populacije. Raziskovalci Univerze v Melbournu so kmalu po objavi teh podatkov odkrili napako v metodi psevdonimizacije, ki je omogočila relativno enostavno identifikacijo zdravnikov in kasneje tudi pacientov [21]. Podatkovna množica je bila po razkritju kmalu odstranjena.

Iz teh primerov je razvidno, da se lahko v primeru nepravilne ali nepremišljene objave podatkov škodi zasebnosti posameznikov. Pri izvajanju zadostnih ukrepov je treba nakloniti posebno pozornost uporabi in anonimizaciji tovrstnih podatkov, da se prepreči napake in zmanjša varnostna tveganja.

4 TEHNIKE ZA ZAŠČITO OSEBNIH PODATKOV V ODPRTIH PODATKOVNIH MNOŽICAH

V nadaljevanju so prikazane tehnike anonimizacije. Vsaka tehnika je opisana s prikazom njenega delovanja ter naštetimi prednostmi in slabostmi njene uporabe.

4.1 Osnovni pristopi: Agregacija, supresija, generalizacija

To so najosnovnejše tehnike, ki pri obdelavi predstavljajo prvi korak za anonimizacijo množice in so namenjene zmanjševanju podrobnosti v podatkih. So le en del kompleksnega procesa anonimizacije.

Pri agregaciji [22] se uporablja združevanje podatkov v skupine, objavi pa se le podatke v agregiranem stanju. Največkrat se uporabljajo povprečja, vsote in števila. Na primer, pri objavljanju dohodkov v nekem mestu se objavi le povprečna plača tistega geografskega področja ali povprečje posamezne skupine znotraj področja.
Supresija [23] je tehnika odstranjevanja podatkov iz množice. Obstaja več vrst:

- supresija zapisov: odstranjevanje zapisov, ki predstavljajo tveganje za razkritje identitete (zapisi posameznikov, ki bistveno odstopajo od ostalih in jih zato ni mogoče vključiti v skupine z zadostno stopnjo anonimnosti),
- in supresija celic oziroma atributov: odstranjujemo le določene atribute ali vrednosti, ki bi lahko vsebovale občutljive podatke (zamenjava vrednosti z * ali z oznako »izpuščeno«). To tehniko uporabljamo v kombinaciji z drugimi anonimizacijskimi tehnikami, kot so k-anonimnost za zapise, ki jih ne moremo generalizirati.

Generalizacija [24] je metoda, pri kateri natančne vrednosti atributov zamenjamo z manj specifičnimi zapisi. Natančno vrednost dohodka zamenjamo z razponom oziroma razredom. Na primer: natančno višino človeka (npr. 192 centimetrov) zamenjamo z razponom (npr. 190-195 centimetrov). Generalizacija ima tudi posebno obliko, ki se ji reče ang. top-coding in bottom-coding, kjer se vrednosti z določenimi mejami združi v eno kategorijo (npr. spol = moški, starost<20).

4.2 Psevdonimizacija

Psevdonimizacija [25] je tehnika, ki jo GDPR (člen 4 postavka 5) [11] izrecno omenja kot varnostni ukrep. Definira jo kot »obdelavo osebnih podatkov tako, da osebnih podatkov ni več mogoče pripisati specifičnemu posamezniku brez uporabe dodatnih informacij, pod pogojem, da se te dodatne informacije hranijo ločeno ter zanje veljajo tehnični in organizacijski ukrepi za zagotavljanje, da se osebni podatki ne pripišejo določenemu ali določljivemu posamezniku.«

Pri procesu psevdonimizacije se specifične vrednosti zamenjajo z identifikatorji ali s t.i. psevdonimi, ki so umetno ustvarjeni za vsako vrednost. Ti podatki so tako obdelani, da jih je mogoče ponovno identificirati zgolj z uporabo posebne tabele podatkov, ki povezuje psevdonim z realno vrednostjo

identifikatorja. Te tabele oziroma vrednosti morajo biti dobro varovane v ločenih shrambah.

4.3 Modeli anonimizacije

Osnovni modeli zagotavljajo določeno raven zaščite, ki ne zadostuje za objavo podatkov, zato so bili ustvarjeni formalnejši pristopi k anonimizaciji. Ti zagotavljajo višjo stopnjo anonimizacije in varnosti pred raznimi načini napadov re-identifikacije. Formalni modeli so narejeni na osnovi supresije kvazi-identifikatorjev in generalizacije.

4.3.1 K-anonimnost

K-anonimnost je eden izmed najbolj poznanih algoritmov anonimizacije. Cilj te metode je zagotoviti, da nobenega unikatnega posameznika v množici ni mogoče razlikovati od vsaj k-1 drugih posameznikov na podlagi njihovih kvazi-identifikatorjev [14]. Enostavneje, vsaka kombinacija vrednosti kvazi-identifikatorjev se mora v množici pojaviti vsaj k-krat. Skupine s podobnimi kvazi-identifikatorji se imenujejo ekvivalenčne skupine.

K-anonimnost se doseže z generalizacijo in opcijsko supresijo. Ta postopka se izvaja, dokler vsaka ekvivalenčna skupina ne vsebuje vsaj k število variant. Število k si izberemo sami, pri čemer moramo sprejeti kompromis; večji k namreč nudi večjo zaščito pred napadi re-identifikacje, vendar zahteva tudi večjo stopnjo generalizacije in supresije, ki zmanjša informacijsko vrednost podatkovne množice.

4.3.2 L-raznolikost

L-raznolikost [26] je razširitev k-anonimnosti z namenom boljše zaščite pred napadom na homogenost. L-raznolikost poleg k-anonimnosti zahteva, da vsaka ekvivalenčna skupina vsebuje vsaj l »dobrih različnih vrednosti« občutljivega atributa. To so dovolj raznolike vrednosti, da so napadi sklepanja in napadi re-identifikacije na njih oteženi ali onemogočeni.

Poznamo več različnih pogledov na »dobre različne vrednosti«:

- Rekurzivna (c, l)-raznolikost: metoda preprečuje da bil odstotek specifičnega l-občutljivega atributa večji od določenega. S tem preprečuje, da bi bile nekatere vrednosti pogostejše od drugih.
- Razločna l-raznolikost: najlažja verzija te metode, ki zahteva vsaj l različnih vrednosti atributov v vsaki ekvivalenčni skupini.
- Entropijska l-raznolikost: zahteva, da je entropijska porazdelitev občutljivega atributa v množici v vsaki skupini vsaj logaritem od l. S to raznolikostjo dosežemo enakomernejšo porazdelitev vrednosti.

4.3.3 T-bližina

T-bližina [27] je bila zasnovana zaradi slabosti l-raznolikosti in k-anonimnosti, še posebej zaradi napadov podobnosti in napadov na neenakomernost podatkov. Model deluje tako, da so porazdelitve vrednosti občutljivega atributa v ekvivalenčnih skupinah blizu porazdelitvi tega atributa v izvirni množici. Bližina se v tej metodi meri z razdaljo t občutljivega atributa v ekvivalenčni množici od atributa v izvirni množici, ki ne sme presegati meje t.

4.3.4 Diferencialna zasebnost

Diferencialna zasebnost [28] predstavlja pristop, ki ni podoben ostalim pristopom (tj. skrivanjem posameznikov v skupine), ampak temelji na omejevanju količine informacij o posamezniku, ki jih je mogoče pridobiti iz rezultata analize ali objavljenih podatkov. Je matematično zahtevna definicija zasebnosti, ki kot rezultat analize ali poizvedbe poda popolnoma enak odgovor, ne glede na to, ali so posameznikovi podatki vključeni ali ne.

Ta zasebnost se doseže z dodajanjem umerjenega naključnega šuma, bodisi v proces poizvedbe bodisi v same podatke, preden so objavljeni. Količina šuma, ki je dodan, je odvisna od parametra epsilon (e^{ε}), imenovanega tudi proračun zasebnosti (ang. Privacy Budget). Manjši e^{ε} pomeni več zasebnosti (več šuma) in obratno. Diferencialna zasebnost zagotavlja, da za dve naključni množici in $M1$ in $M2$, ki se razlikujeta le v enem zapisu, in naključno množico C, velja: $P(ZasebniIzr(M1) \in C) \leq e^{\varepsilon} * P(ZasebniIzr(M2) \in C) + \delta$. Tukaj P predstavlja verjetnost, da bo ($ZasebniIzr(M1)$ vključen v množico izhodov C, e^{ε} pa predstavlja Eulerjevo število, ki je matematična konstanta (približno 2.71828). Parameter δ predstavlja odstopanja, pri katerih zaščita zasebnosti ni zagotovljena. Pogosto se uporablja poenostavljena zasebnost, kjer je $\delta = 0$.

Mehanizmi za doseganje diferencialne zasebnosti so:

- Gaussov mehanizem: doda šum z Gaussovo porazdelitvijo k numeričnim rezultatom. Najpogosteje se uporablja s kombinacijo (e^{ε}, δ) – DZ (diferencialna zasebnost).
- Eksponentni mehanizem: se uporablja največ pri

numeričnih rezultatih, kjer se verjetnost odgovora določi na njegovo kvaliteto in e^{ϵ}. Lahko pa ga tudi uporabimo za kategorične podatke, kjer določimo podatek z uporabo posebne funkcije.
- Laplaceov mehanizem: dodajamo šum z Laplaceovo porazdelitvijo k numeričnim rezultatom poizvedb (npr. povprečja). Koliko šuma bo dodanega, je odvisno od občutljivosti poizvedbe in od e^{ϵ}.
- Naključni odgovor: uporabljamo pri kategoričnih in numeričnih podatkih tako, da mešamo odgovore na ravni populacije. Če nekdo odgovori z odgovorom, nam tega odgovora ni treba dodati v zapis iste osebe, temveč v drug zapis, kar storimo pri vseh posameznikih iz populacije. Tako je na ravni populacije še vedno ista vrednost podatkov, le identitete posameznikov so zakrite.

Obstajajo tudi različni modeli diferencialne zasebnosti:
- Lokalna diferencialna zasebnost: vsak uporabnik v svoje podatke doda šum, preden jih pošlje centralni agregacijski enoti. S tem pristopom se zagotovi večja zaščita, ampak nastane v podatkih več šuma za isto raven zasebnosti. Ta metoda je varnejša, saj agregator nikoli ne vidi osebnih podatkov posameznika.
- Centralizirana diferencialna zasebnost: določi se zaupanja vrednega skrbnika, ki se mu omogoči dostop do vseh podatkov. Skrbnik izvaja poizvedbe in dodaja šum k rezultatom, preden so podatki objavljeni.

4.4 Generiranje sintetičnih podatkov

Naslednja možnost za zaščito osebnih podatkov je ustvarjanje popolnoma sintetičnih podatkov [29], ki po analitičnem in vzorčnem stanju popolnoma posnemajo vzorce izvirne množice, ampak ne vsebujejo nobene realne informacije iz nje.

Poznamo generacijo z uporabo strojnega učenja, na primer z generativnimi nasprotniškimi mrežami (ang. Generative Adversarial Networks), ki so naučene generirati realistične podatke. Poznamo tudi pristop generacije na podlagi statističnih modelov, ki temelji na vzorčenju iz porazdelitve atributov.

Možno je generiranje s kombinacijo sintetičnih podatkov z diferencialno zasebnostjo. Najprej se model uči na podatkih z zagotovljeno diferencialno porazdelitvijo, nato pa se z njim generira sintetične podatke. S tem pristopom omogočimo visoko raven zasebnosti.

4.5 Primerjava tehnik: Prednosti, slabosti in področja uporabe

Izbira tehnike ali kombinacije tehnik za zagotavljanje zaščite je odvisna od več dejavnikov:
- kolikšna stopnja zasebnosti je zahtevana glede na pravne zahteve in ocene tveganja,
- narave podatkov (števila zapisov, atributov, kvazi-identifikatorjev, itd.),
- virov na razpolago in znanja upravljalcev,
- potencialnih napadov in znanja napadalcev,
- želene stopnja uporabnosti podatkov po uporabi anonimizacijske metode.

Za zagotavljanje anonimizacije pri objavljanju podatkovnih množic obstaja širok nabor tehnik in modelov. Pri izbiri tehnike nobena ne izstopa iz množice, vendar je vsaka dobra za svoje področje uporabe in vse zahtevajo isti kompromis; za večjo stopnjo zaščite tvegamo večjo izgubo ali popačenost podatkov. Osnovni modeli so preprosti in nudijo hitro, a slabo zaščito. Modeli kot so k-anonimnost, l-raznolikost in t-bližina nudijo večjo odpornost pred napadi, vendar imajo določene pomanjkljivosti. Diferencialna zasebnost skuša povečati zaščito, pri čemer doda šum v podatke, kar zmanjša njihovo uporabnost. Sintetični podatki ne objavijo nobenega osebnega podatka, ampak so računsko zahtevni za izvedbo. Najpogosteje se uporablja kombinacija več tehnik za zagotovitev največje odpornosti pred napadi. Najpomembnejše je zavedanje tveganja in občutljivosti podatkov, ki jih objavljamo, poznati pa moramo tudi tehnike in pristope, ki bodo najučinkoviteje zaščitile podatke, ki jih nameravamo objaviti.

5 METODOLOGIJA RAZISKAVE

Raziskavo smo zasnovali kot sistematični pregled in empirično analizo javno dostopnih podatkovnih zbirk z vidika tveganj za ponovno identifikacijo posameznikov.

Najprej smo oblikovali raziskovalna vprašanja:
- V kolikšni meri izbrane odprte podatkovne zbirke omogočajo ponovno identifikacijo posameznikov?
- Katere tehnike anonimizacije oziroma psevdonimizacije so bile uporabljene in kako učinkovite so?
- Ali izbrane zbirke ustrezajo sodobnim standardom varstva osebnih podatkov?

V prvem koraku smo izvedli sistematični pregled slovenskega portala odprtih podatkov (OPSI),

kjer so objavljene podatkovne zbirke različnih državnih organov in institucij. Pri tem smo uporabili kriterije izbora, ki so vključevali:

- prisotnost podatkov na ravni posameznika (individualni zapisi),
- dostopnost podatkov brez omejitev,
- vsebovanje potencialnih kvazi-identifikatorjev (npr. starost, spol, geografska lokacija),
- zbirke, ki predstavljajo različne tipe tveganj: področje kriminalitete (občutljivi osebni podatki), register vozil (tehnični unikatni identifikatorji) in kmetijstvo (javna dostopnost zaradi transparentnosti).

Na podlagi teh kriterijev smo izbrali tri podatkovne zbirke za nadaljnjo analizo. Izbor je bil namenoma usmerjen v zbirke, pri katerih obstaja povečano tveganje za ponovno identifikacijo zaradi strukture podatkov.

V drugem koraku smo nad izbranimi zbirkami izvedli eksperimente ponovne identifikacije. Ti so temeljili na kombiniranju kvazi-identifikatorjev ter povezovanju z zunanjimi, javno dostopnimi viri podatkov. Na ta način smo simulirali realistične scenarije napadov in ocenili verjetnost uspešne reidentifikacije posameznikov.

V tretjem koraku smo analizirali uporabljene pristope varovanja podatkov, pri čemer smo dosledno razlikovali med: anonimizacijo (nepovratna odstranitev identifikacijskih elementov) in psevdonimizacijo (zamenjava identifikatorjev z nadomestnimi vrednostmi, ob možnosti ponovne povezave).

Stopnjo zaščite podatkov smo ovrednotili tudi z vidika uveljavljenih modelov, kot so k-anonimnost, l-raznolikost in t-približek. Preverili smo tudi status obravnavanih podatkovnih zbirk (npr. ali so bile zaradi ugotovljenih pomanjkljivosti anonimizacije naknadno umaknjene ali spremenjene) ter to ustrezno vključili v interpretacijo rezultatov. Na podlagi ugotovitev smo oblikovali priporočila za izboljšanje praks objavljanja odprtih podatkov, z namenom zmanjšanja tveganj za razkritje identitete posameznikov.

6 ANALIZA PRIMEROV ODPRTIH PODATKOVNIH MNOŽIC

V tem poglavju smo uporabili tehnike in postopke, opisane v prejšnjih poglavjih, za analiziranje konkretnih primerov podatkovnih množic. Naš cilj je bil pregledati podatkovne množice in analizirati uporabljene anonimizacijske tehnike, oceniti tveganja ter pregledati možnost ponovne identifikacije posameznikov v množici. Vsaka analizirana množica omogoča globlji vpogled v trenutno stanje in prakse obdelave ter objave podatkovnih množic. Z rezultati analize smo dobili podlago za preverjanje trenutnega stanja Slovenskih odprtih množic.

6.1 Analiza 1. primera: podatki o slovenskih kaznivih dejanjih

Prvi primer analize zajema javno dostopno podatkovno množico kaznivih dejanj, zabeleženih v Sloveniji leta 2023. Množica vsebuje 94.816 zapisov, pri čemer vsak zapis predstavlja posameznika, vpletenega v kaznivo dejanje – bodisi kot storilca bodisi kot žrtev. Množica vključuje širok nabor atributov, med drugim: zaporedno številko kaznivega dejanja, časovno umestitev (mesec, dan v tednu, uro storitve), policijsko upravo (PU) storitve, podatke o povratništvu, opis in poglavje kaznivega dejanja, označbe gospodarskega, organiziranega ali mladoletniškega kriminala, podatke o poskusu dejanja, kriminalistične oznake, uporabljena sredstva, upravno enoto, opis kraja dogodka, leto in vrsto zaključnega dokumenta, podatke o osebi (zaporedna številka, vloga, starostni razred, spol, državljanstvo) ter dodatne atribute, kot so poškodba, vpliv alkohola ali mamil, vključenost v organizirano združbo in ocenjena škoda.

Na podlagi sistematičnega pregleda smo prepoznali dva neposredna identifikatorja, devet kvazi-identifikatorjev in tri občutljive atribute. Takšna kombinacija kaže na visoko tveganje za kršitev zasebnosti, saj omogoča možnost posredne prepoznave posameznikov, zlasti v primeru povezovanja z zunanjimi viri.

V anonimizacijskem pregledu nismo zaznali znakov uporabe naprednih anonimizacijskih tehnik. Prisotna je zgolj osnovna supresija, izvedena z odstranitvijo nekaterih atributov, medtem ko metod, kot so generalizacija ali agregacija, množica ne kaže. Takšna struktura podatkov omogoča enostavno povezovanje posameznih zapisov z drugimi javno dostopnimi informacijami.

Analitične ugotovitve kažejo, da je tveganje za re-identifikacijo visoko zaradi:

- nepopolne uporabe anonimizacijskih metod,
- visokih podrobnosti o času, kraju in vrsti kaznivih dejanj,
- prisotnosti občutljivih atributov, ki bi lahko povzročili znatno škodo posameznikom v primeru razkritja.

Za oceno dejanske izvedljivosti re-identifikacije smo izvedli kontroliran poskus re-identifikacije, ki je bil izveden v raziskovalne in izobraževalne namene, skladno z etičnimi načeli varstva osebnih podatkov. Postopek je temeljil izključno na uporabi javno dostopnih virov brez neposrednega vpogleda v neanonimizirane osebne podatke.

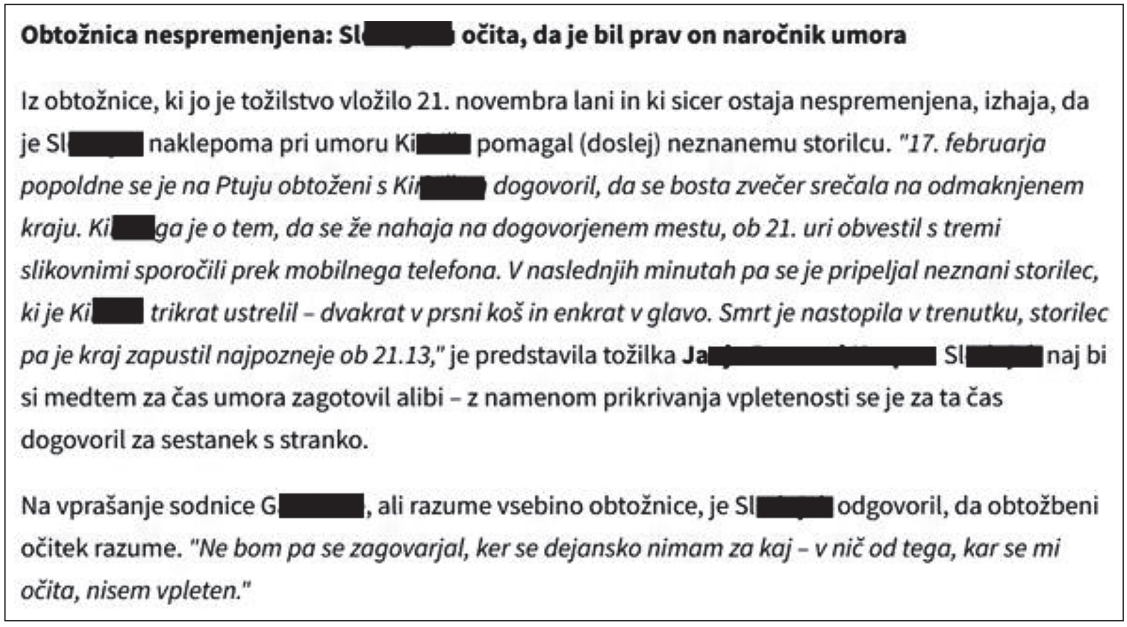
V okviru poskusa smo identificirali zapis (Tabela 1), ki se je nanašal na redko kaznivo dejanje – umor, storjen leta 2023 na Ptuju. Zapis je vseboval naslednje kvazi-identifikatorje: čas (četrtek zvečer), kraj (Ptuj) in uporabljeno sredstvo (strelno orožje). Ujemanje teh atributov smo primerjali z novinarskim člankom, objavljenim na javno dostopnem spletnem portalu (Slika 1), ki je opisoval isti dogodek. Članek je navajal dodatne informacije, med drugim ime in priimek storilca. Povezava med podatki iz članka in kvazi-identifikatorji v množici je omogočila uspešno ponovno identifikacijo posameznika.

Analiza primera potrjuje, da lahko redkost dogodkov (npr. umorov) ter kombinacija časovnih in prostorskih podatkov bistveno povečata tveganje za re-identifikacijo tudi v navidezno anonimiziranih množicah. Takšni atributi imajo močno diskriminativno moč in omogočajo natančno povezovanje z javnimi viri.

Zaključek analize potrjuje, da množica ne izkazuje ustrezne anonimizacije in da predstavlja visoko tveganje za ponovno identifikacijo. Uporaba osnovnih ali naprednejših metod, kot so k-anonimnost, l--raznolikost in t-bližina, bi bila nujna za zmanjšanje verjetnosti razkritja posameznikov, obenem pa bi ohranila raziskovalno vrednost podatkov. Primer jasno ponazarja, da tudi uradno anonimizirani nabori podatkov lahko predstavljajo tveganje, če vsebujejo redke dogodke in kombinacije atributov, ki omogočajo posredno identifikacijo.

Tabela 1: **Zapis re-identifikacije prve množice**

Zaporedna številka KD	**18646**
Mesec storitve	22022
Ura storitve	21:00-21:59
Dan v tednu	Četrtek
Policijska enota	PU Maribor
Povratnik	Ne
Opis kaznivega dejanja	KZ12/116*/1// – Umor
Poglavje KD	Kaznivo dejanje zoper življenje in telo
Gospodarski kriminal	Splošna kazniva dejanja
Organiziran kriminal	(ni podatka)
Mladoletniška kriminaliteta	(ni podatka)
Poskus KD	Ne
Kriminalistična oznaka 1	Ni oznake
Kriminalistična oznaka 2	(prazno)
Kriminalistična oznaka 3	(prazno)
Uporabljeno sredstvo 1	236 – Pištola
Uporabljeno sredstvo 2	(prazno)
Uporabljeno sredstvo 3	(prazno)
Uporabljeno sredstvo 4	(prazno)
Upravna enota	Ptuj
Opis kraja dogodka	Območje prometa
Leto zaključnega dokumenta	2022
Vrsta zaključnega dokumenta	Ovadba
Zaporedna številka osebe	18646002
Vrsta osebe	Ovadenec / osumljenec
Starostni razred	34–44 let
Spol	Moški
Državljanstvo	Slovensko
Poškodba	(prazno)
Vpliv alkohola	Ne
Vpliv mamil	Ne
Organizirana združba	Ne
Škoda	Brez



Obtožnica nespremenjena: Sl[redacted] očita, da je bil prav on naročnik umora

Iz obtožnice, ki jo je tožilstvo vložilo 21. novembra lani in ki sicer ostaja nespremenjena, izhaja, da je Sl[redacted] naklepoma pri umoru Ki[redacted] pomagal (doslej) neznanemu storilcu. *"17. februarja popoldne se je na Ptuju obtoženi s Ki[redacted] dogovoril, da se bosta zvečer srečala na odmaknjenem kraju. Ki[redacted]ga je o tem, da se že nahaja na dogovorjenem mestu, ob 21. uri obvestil s tremi slikovnimi sporočili prek mobilnega telefona. V naslednjih minutah pa se je pripeljal neznani storilec, ki je Ki[redacted] trikrat ustrelil – dvakrat v prsni koš in enkrat v glavo. Smrt je nastopila v trenutku, storilec pa je kraj zapustil najpozneje ob 21.13,"* je predstavila tožilka **Ja[redacted] Sl[redacted]** naj bi si medtem za čas umora zagotovil alibi – z namenom prikrivanja vpletenosti se je za ta čas dogovoril za sestanek s stranko.

Na vprašanje sodnice G[redacted], ali razume vsebino obtožnice, je Sl[redacted] odgovoril, da obtožbeni očitek razume. *"Ne bom pa se zagovarjal, ker se dejansko nimam za kaj – v nič od tega, kar se mi očita, nisem vpleten."*

Slika 1: **Članek ujemajoč z zapisom iz množice**

Analiza prvega primera slovenske podatkovne množice kaže na velike izzive pri zasebnosti posameznikov, katerih podatki so objavljeni v odprtih množicah, in opozarja na potrebo po boljši zaščiti podatkov, preden so ti objavljeni.

6.2 Analiza 2. primera: podatki o prvi registraciji avtomobilov

Kot drugi primer smo analizirali javno dostopno podatkovno množico prvih registracij vozil v Sloveniji za leto 2023. Množica vsebuje 12.890 zapisov, pri čemer vsak zapis predstavlja posamezno registracijo vozila. Vsebovani atributi vključujejo med drugim: datum prve registracije vozila, datum prve registracije v Sloveniji, oznako potrdila o skladnosti, datum in status vozila, registrsko območje ter podvrsto registrske tablice, izvajalno enoto prve registracije, starost uporabnika vozila, spol lastnika (kadar gre za fizično osebo), tip lastnika (pravna ali fizična oseba), znamko in tovarniško oznako vozila, državo izvora, datum izdaje COC dokumenta ter VIN številko in identifikacijsko kodo vozila.

Na podlagi pregleda strukture podatkov smo ugotovili prisotnost enega neposrednega identifikatorja (VIN številka) ter štirih kvazi-identifikatorjev, ki bi lahko v kombinaciji omogočili prepoznavo posameznika. Analiza anonimnosti množice je pokazala, da podatki ne kažejo znakov sistematične anonimizacije. Edina prepoznana tehnika, ki bi lahko bila uporabljena, je supresija z odstranitvijo določenih atributov, medtem ko metod, kot so generalizacija ali agregacija nismo zaznali.

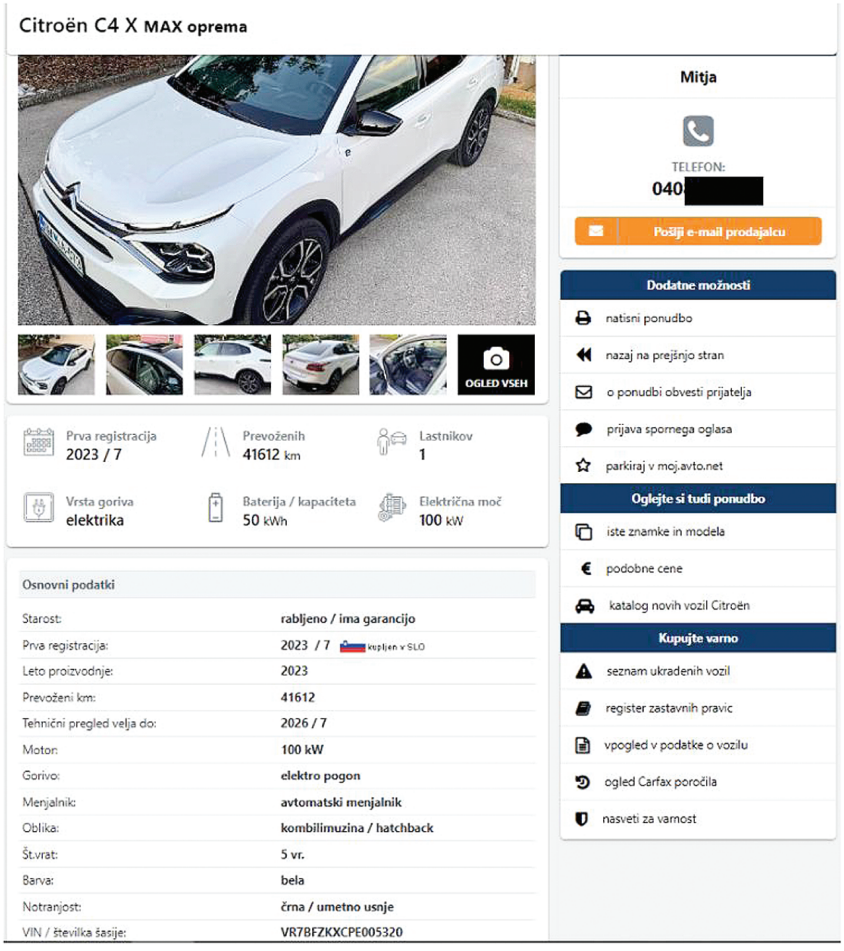
Zaradi ohranjenih identifikacijskih lastnosti smo ocenili, da je tveganje za re-identifikacijo precejšnje, saj je mogoče posamezne zapise povezati z zunanjimi javno dostopnimi viri. V ta namen smo izvedli kontroliran poskus re-identifikacije, katerega namen je bil eksperimentalno potrditi obstoj tveganja. Poskus je bil izveden izključno z uporabo javno dostopnih informacij in v skladu z etičnimi načeli raziskovanja.

V okviru poskusa smo izbrali en zapis, pri katerem smo VIN številko uporabili kot povezovalni atribut. S pomočjo javno dostopnih spletnih oglasov za prodajo vozil smo poiskali ujemanje tovarniške številke (VIN) z vozilom v oglasu. Oglas je vseboval dodatne informacije, kot so telefonska številka in ime prodajalca, ki smo jih nato povezali z ustreznimi atributi v podatkovni množici. Na ta način je bilo mogoče določiti identiteto posameznika, kar predstavlja uspešno izvedeno re-identifikacijo. Kot primer uspešne ponovne identifikacije smo analizirali zapis iz podatkovne množice (Tabela 2), ki je vseboval informacije o avtu Citroen C4 MAX. Zapis je omenjal avto s specifično tovarniško oznako katero smo na portalu avto.net [5] našli v enem izmed oglasov. Podatke iz množice smo primerjali z oglasom (Slika 2)

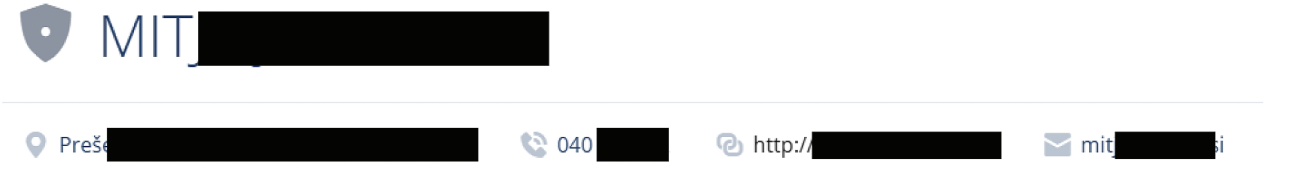
in našli, da se tudi drugi podatki iz množice ujemajo. Za identifikacijo posameznika pa smo uporabili telefnosko številko objavitelja oglasa ter njegovo ime preko katere smo s pomočjo imenika ter iskalnika našli naslov ter polno ime osebe, ki ima ta avto v lasti (Slika 3). S tem primerom vidimo, da lahko objave neposrednih identifikatorjev hitro vodijo v re-identifikacijo posameznika v množici tudi, če ne kažejo neposredno na posameznika.

Tabela 2: **Zapis ujemajočega zapisa drugega primera**

B-Datum prve registracije vozila	27.06.2023
2A-Datum prve registracije vozila v SLO	27.06.2023
1K-Oznaka potrdila o skladnosti	SA
Datum statusa vozila	27.06.2023
Status vozila (id)	1
Status vozila (opis)	registrirano
Izvajalna enota prve registracije	REGIA GROUP d.d., PE MOSTE
4A-Registrsko obmocje tablice prve registracije	LJUBLJANA
4A-Podvrsta tablice prve registracije	NT
C-Spol uporabnika (če gre za fizično osebo)	P
C-Ali je uporabnik tudi lastnik vozila	DA
C1.3-Upravna enota uporabnika (oznaka)	009
C1.3-Upravna enota uporabnika (opis)	GROSUPLJE
C1.3-Občina uporabnika (oznaka)	032
C1.3-Občina uporabnika (opis)	Grosuplje
C2-Ali je lastnik pravna ali fizična oseba	P
D.1-Znamka	CITROEN
D.2-Tovarniška oznaka	BFZKXC-B0MA00
D.3-Komercialna oznaka	C4 X / 100 / ELEKT.
D.4.2-Država (opis)	Francija
D.4.2-Država (koda)	FRA
D.5-Datum izdaje COC dokumenta	26.06.2023
E-VIN	VR7BFZKXCPE023004
F.1-Največja tehnično dovoljena masa vozila	2040
G-Masa vozila	1715
J-Kategorija in vrsta vozila (oznaka)	M1
J-Kategorija in vrsta vozila (opis)	osebni avtomobil
K-Homologacijska oznaka vozila	e9*2007/46*6816*15
L-Število osi	2
M-Medosje	2670
M.1-Zadnji previs	1050,0
N-Razporeditev najv. tehnično dovoljene mase na osi	1020/1070
O.1.3-Priklopnik s centralno osjo	–1
O.2-Nezavirano priklopno vozilo	–1
P.1.3-Vrsta goriva (opis)	Ni goriva
P.1.5-Oznaka motorja	PERMANENTI MAGN. SINHRON MOTOR
P.2.1-Tip	100
P.2.3-Delovna napetost	400
P.2.4-Pogonske baterije	Lithium-ion 400V
P.2.5-Oznaka motorja	ZK01
R-Barva vozila (opis)	navaden – BELA – SREDNJA
S.1-Število sedežev	5
T-Najvišja hitrost	150
V.1-CO	68
V.7-CO2	0
V.9-Podatek o okoljevarstveni kategoriji vozila	715/2007*2018/1832AX (EURO)
X-Oblika nadgradnje (oznaka)	AA
X-Oblika nadgradnje (opis)	limuzina
Y.1-Dolžina	4600
Y.2-Širina	1800
Y.3-Višina	1525
Z.1-Dovoljene pnevmatike in platišča	195/60 R18 96H 6.50J18 ET26, 215/60 R17 96H 6.50J17 ET32, 215/65 R16 98H 6.50J16 ET32
Datum izdaje potrdila o skladnosti	26.06.2023
Okoljevarstvena oznaka	Ni relevantno
Komercialna oznaka do prvega	C4 X



Slika 2: **Ujemajoč oglas z zapisom iz množice**



Slika 3: **Pordrobnejša re-identifikacija oglasa**

Ta primer kaže, da lahko, kot smo storili že pri prejšnji množici, brez dobre anonimizacije izvedemo re-identifikacijo in pridobimo podatke iz množic, ki naj bi bile varne.

6.3 Analiza 3. primera: certifikati pridelovalcev

Kot tretji primer smo analizirali javno dostopno podatkovno množico o izdanih varnostnih in strokovnih certifikatih v Republiki Sloveniji. Množica vsebuje 8.472 zapisov, pri čemer vsak zapis predstavlja posamezno izdajo certifikata določenemu podjetju ali posamezniku. Vsebovani atributi vključujejo: ime certifikacijskega organa, naziv certifikata, področje certifikacije, številko in datum izdaje certifikata, veljavnost, status (aktiven/potekel), ime podjetja ali nosilca, kraj oziroma sedež podjetja, panogo dejavnosti ter vrsto standarda (npr. ISO 9001, ISO 14001, ISO/IEC 27001 ipd.).

Analiza strukture podatkovne množice je pokazala, da ta vsebuje atribute, ki lahko v določenih primerih omogočajo posredno identifikacijo posameznikov, predvsem kadar gre za samostojne podjetnike ali fizične osebe, ki nastopajo kot nosilci certifikatov. Prisotnost atributov, kot so naziv podjetja ali osebe, kraj, datum izdaje in vrsta certifikata, predstavlja kombinacijo potencialnih kvazi-identifikatorjev, ki bi v povezavi z javno dostopnimi viri lahko omogočili prepoznavo posameznika.

Pri pregledu anonimizacijskih značilnosti množice nismo zaznali uporabe nobene izmed standardnih tehnik anonimizacije, kot so generalizacija ali razne agregacije. Edini ukrep, ki bi ga bilo mogoče delno prepoznati, je supresija pri določenih atributih (npr. izpuščene identifikacijske številke). Kljub temu struktura množice ohranja dovolj natančne informacije, da bi bilo ob določenih okoliščinah mogoče razkriti identiteto posameznika.

V tem primeru nismo izvajali postopka re-identifikacije, saj množica ni bila objavljena z oznako anonimiziranih podatkov. Namen analize je bil predvsem opozoriti na dejstvo, da se lahko tudi v odprtih podatkovnih zbirkah, ki so primarno namenjene preglednosti in informiranju javnosti, nahajajo osebni podatki ali podatki, ki omogočajo njihovo posredno razkritje. Takšne situacije poudarjajo pomen ustrezne presoje, ali je objava določenih atributov res nujna, ter potrebo po dosledni uporabi tehnik anonimizacije v primerih, kjer to zahteva zakonodaja ali etična odgovornost upravljavcev podatkov.

6.4 Povzetek pregledov

Analiza treh različnih podatkovnih množic je pokazala različne stopnje tveganja za zasebnost posameznikov in obstoječe ali potencialne možnosti re-identifikacije.

Primer 1 – Kazniva dejanja: Množica vsebuje podrobne zapise o posameznikih, ki so vpleteni v kazniva dejanja, vključno z demografskimi podatki in kontekstom dogodkov. Prisotnost neposrednih identifikatorjev, kvazi-identifikatorjev in občutljivih atributov omogoča visoko tveganje za re-identifikacijo. Poskus povezovanja podatkov z javno dostopnimi viri je pokazal, da je identifikacija posameznikov praktično izvedljiva, kar jasno kaže na potrebo po ustrezni anonimizaciji pri obdelavi in objavi takšnih podatkov.

Primer 2 – Registracije vozil: Množica vsebuje zapise o posameznih registracijah vozil, kjer so prisotni neposredni in kvazi-identifikatorji (VIN številka, starost lastnika, status vozila, registrska območja). Analiza je pokazala odsotnost sistematične anonimizacije, tveganje za re-identifikacijo pa je bilo potrjeno z eksperimentalnim poskusom, pri katerem smo posamezne zapise povezali z javno dostopnimi oglasi za prodajo vozil.

Primer 3 – Certifikati kmetijskih in strokovnih pridelovalcev: Množica vsebuje javno objavljene podatke o prejemnikih certifikatov, pri čemer so neposredni identifikatorji (ime in naziv nosilca certifikata) namenoma prisotni. Kljub temu, da re-identifikacije nismo izvajali, analiza kaže, da je kombinacija atributov potencialno dovolj natančna za identifikacijo posameznika ali podjetja. Ta primer poudarja, da je v odprtih podatkih prisotnost osebnih ali posredno identificirajočih atributov smiselna le, če je to v skladu z namenom objave in zakonodajo.

Skupna ugotovitev vseh treh pregledov je, da podatkovne množice v praksi pogosto vsebujejo osebne podatke ali podatke, ki omogočajo identifikacijo posameznikov. Stopnja tveganja se razlikuje glede na naravo in namen množice, prisotnost neposrednih identifikatorjev, kvazi-identifikatorjev ter občutljivih atributov. Analiza potrjuje, da je uporaba ustreznih tehnik anonimizacije ključna za zmanjšanje možnosti re-identifikacije, predvsem, ko želimo zaščititi zasebnost posameznikov.

7 PRIKAZ MOČNEJŠE ANONIMIZACIJE

V tem poglavju bomo predstavili pristop k pravilni anonimizaciji podatkovne množice policijskih prekškov, ki je bila obravnavana kot prvi primer v našem pregledu. Kot smo ugotovili v analizi, je množica vsebovala neposredne identifikatorje, kvazi-identifikatorje in občutljive podatke, kar je ustvarjalo visoko tveganje za ponovno identifikacijo posameznikov.

Cilj tega poglavja je prikazati metodologijo sistematične anonimizacije, ki zmanjšuje tveganje za re-identifikacijo, hkrati pa ohranja uporabnost podatkov za analitične in raziskovalne namene. Posebna pozornost je namenjena ohranjanju strukture podatkov in možnosti analiziranja agregiranih podatkov, pri čemer se uporabi kombinacija tehnik, kot so supresija, generalizacija, psevdonimizacija in, po potrebi, agregacija podatkov.

7.1 Postopek anonimizacije

Identifikacija neposrednih in kvazi-identifikatorjev: neposredne identifikatorje (npr. zaporedna številka osebe, ime, priimek) smo identificirali in označili kot občutljive atribute. Kvazi-identifikatorji, kot so starostni razred, spol, lokacija (uprava, občina) ter časovni atributi (mesec, dan, ura), so bili evidentirani za nadaljnjo obdelavo.

Supresija neposrednih identifikatorjev: vse atribute, ki bi omogočali neposredno identifikacijo posameznika, smo odstranili ali psevdonimizirali. Zaporedne številke in imena smo nadomestili z anonimnimi oznakami (npr. ID001, ID002) brez možnosti povratne povezave do originalnih podatkov.

Generalizacija kvazi-identifikatorjev: starost smo zamenjali s starostnimi razredi (npr. 18–25, 26–35, 36–45), da bi zmanjšali natančnost in hkrati ohranili možnost statistične obdelave. Lokacijske atribute smo znižali na višjo administrativno raven (npr. policijska uprava namesto občine ali naselja). Časovne atribute (ura, dan) smo agregirali v širše intervale (npr. jutro, popoldan, večer).

Psevdonimizacija občutljivih atributov: atributi, kot so opis kaznivega dejanja ali status prekrška, so ostali v množici, vendar smo jih po potrebi kodirali ali kategorizirali, da bi preprečili neposredno prepoznavanje specifičnih primerov.

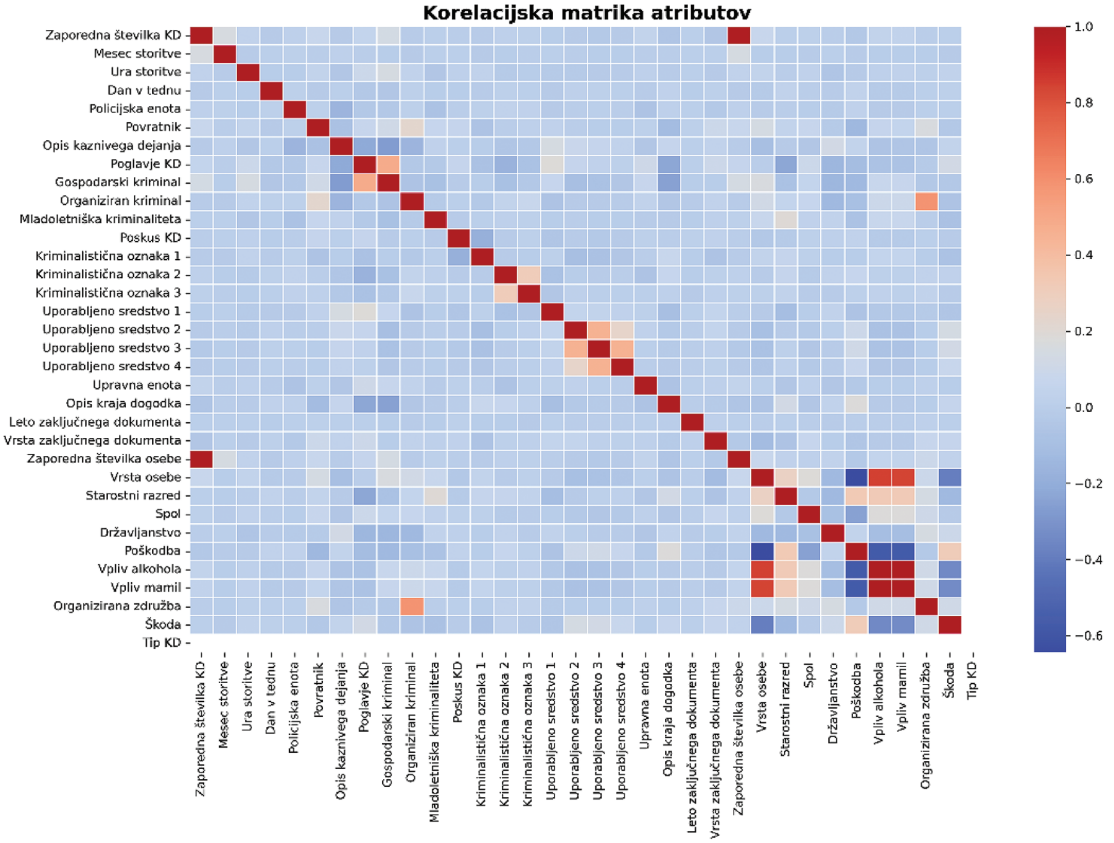
Agregacija in pregled doslednosti: v primeru, da je kombinacija kvazi-identifikatorjev še vedno omogočala edinstvene zapise, smo zapise združili ali agregirali (npr. manjši podatkovni sklopi so bili združeni v skupine). Preverili smo konsistentnost celotne množice, da ni prišlo do izgube pomembnih informacij za analizo, hkrati pa smo zmanjšali tveganje za re-identifikacijo.

Evalvacija tveganja: po izvedbi anonimizacije smo ovrednotili k-anonimnost, l-raznolikost in druge meritve tveganja za re-identifikacijo. Rezultati so pokazali, da je bilo tveganje bistveno znižano, hkrati pa so podatki ohranili uporabnost za raziskovalne in statistične analize.

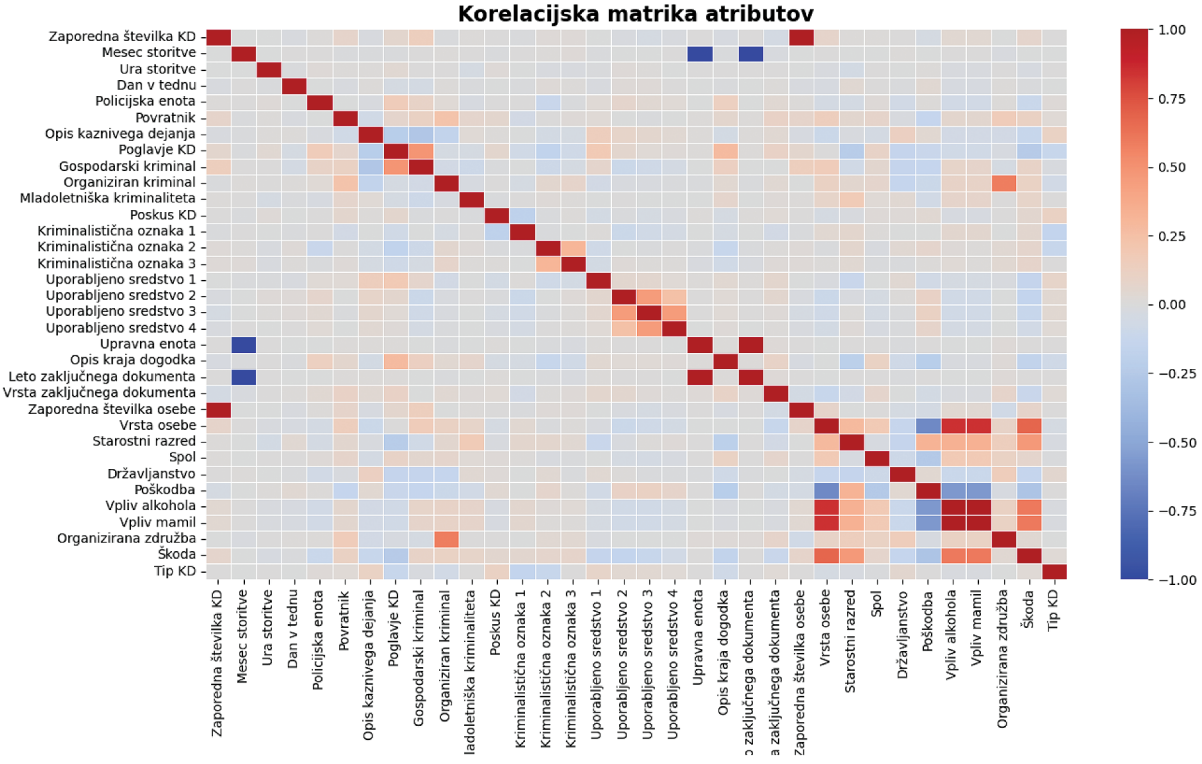
7.2 Prikaz anonimizacije

V prikazu anonimizacije smo izvedli postopek anonimizacije nad podatkovno množico kriminalnih dejanj, pri čemer smo želeli hkrati ohraniti uporabnost podatkov in zagotoviti ustrezno raven zaščite zasebnosti. Za oceno vpliva anonimizacije na analitično vrednost podatkov smo izračunali in primerjali korelacijske metrike podatkov pred anonimizacijo in po njej. Rezultati, predstavljeni na Sliki 4 in Sliki 5, prikazujejo, v kolikšni meri je bila struktura medsebojnih povezav med atributi ohranjena ter kako so se posamezne relacije spremenile zaradi uporabljenih anonimizacijskih metod.

Za ponazoritev zaščite zasebnosti pa smo izvedli tudi preverjanje odpornosti anonimizirane množice na poskus ponovne identifikacije. Na ta način smo iz novo anonimizirane množice izločili vse zapise, ki bi se lahko ujemali z dejanjem, uporabljenim v re-identifikacijskem primeru, ter jih prikazali v Tabeli 3. Rezultat jasno kaže, da anonimizacija učinkovito preprečuje enolično prepoznavanje posameznih zapisov, saj so se podatki združili v večje skupine z istimi atributi, kar otežuje identifikacijo posameznika. S tem prikazom smo tako pokazali dvojni vidik kakovostne anonimizacije: ohranitev analitične uporabnosti podatkov ter hkrati zanesljivo zaščito pred napadi re-identifikacije.



Slika 4: **Korelacijska matrika pred anonimizacijo**



Slika 5: **Korelacijska metrika atributov po anonimizaciji**

Tabela 3: **Zapisi ujemajoči z člankom po anonimizaciji**

Atribut	Vrednosti (ločene z vejico)
Zaporedna številka KD	KD041951, KD041951, KD013801, KD013801
Mesec storitve	(ni podatka), (ni podatka), (ni podatka), (ni podatka)
Ura storitve	Popoldan (12–18), Popoldan (12–18), Popoldan (12–18), Popoldan (12–18)
Dan v tednu	Četrtek, Četrtek, Četrtek, Četrtek
Policijska enota	Regionalna, Regionalna, Regionalna, Regionalna
Povratnik	Ne, Ne, Ne, Ne
Opis kaznivega dejanja	KZ08/116-/4// – Umor, KZ08/116-/4// – Umor, KZ12/116*/1// – Umor, KZ12/116*/1// – Umor
Poglavje KD	KD zoper življenje in telo, KD zoper življenje in telo, KD zoper življenje in telo, KD zoper življenje in telo
Gospodarski kriminal	Splošna, Splošna, Splošna, Splošna
Organiziran kriminal	(ni podatka), (ni podatka), (ni podatka), (ni podatka)
Mladoletniška kriminaliteta	(ni podatka), (ni podatka), (ni podatka), (ni podatka)
Poskus KD	Da, Da, Ne, Ne
Kriminalistična oznaka 1	Umor in samomor, Umor in samomor, Umor in samomor, Umor in samomor
Kriminalistična oznaka 2	(prazno), (prazno), (prazno), (prazno)
Kriminalistična oznaka 3	(prazno), (prazno), (prazno), (prazno)
Uporabljeno sredstvo 1	221 – Orožje, 221 – Orožje, 201 – Nož, 201 – Nož
Uporabljeno sredstvo 2	(prazno), (prazno), (prazno), (prazno)
Uporabljeno sredstvo 3	(prazno), (prazno), (prazno), (prazno)
Uporabljeno sredstvo 4	(prazno), (prazno), (prazno), (prazno)
Upravna enota	Regija, Regija, Regija, Regija
Opis kraja dogodka	Drugo, Drugo, Drugo, Drugo
Leto zaključnega dokumenta	2023, 2023, 2023, 2023
Vrsta zaključnega dokumenta	Poročilo, Poročilo, Poročilo, Poročilo
Zaporedna številka osebe	OSEBA072107, OSEBA072108, OSEBA024992, OSEBA024993
Vrsta osebe	Žrtev, Ostalo, Ostalo, Žrtev
Starostni razred	18–34, 55+, 18–34, 55+
Spol	Moški, Moški, Moški, Ženski
Državljanstvo	Slovensko, Slovensko, Slovensko, Slovensko
Poškodba	Lahka telesna poškodba, (prazno), (prazno), Smrtna poškodba
Vpliv alkohola	NN, Ne, Ne, NN
Vpliv mamil	NN, Ne, Ne, NN
Organizirana združba	Ne, Ne, Ne, Ne
Škoda	Brez škode, Brez škode, Brez škode, Brez škode

8 SKLEP

Na podlagi izvedene analize treh različnih podatkovnih množic in prikaza postopka pravilne anonimizacije smo ugotovili, da se v praksi še vedno pogosto pojavljajo primeri, kjer odprti ali deljeni podatki vsebujejo osebne oziroma posredno prepoznavne informacije. Posebej zaskrbljujoče je, da kar dva od treh analiziranih primerov ne kažeta znakov sistematično izvedene anonimizacije, kar bistveno povečuje tveganje za ponovno identifikacijo posameznikov.

V vseh analiziranih primerih smo zaznali pomanjkljivosti pri uporabi anonimizacijskih tehnik. V prvem primeru, ki se je nanašal na kazniva dejanja, smo ugotovili, da kombinacija demografskih in

kontekstualnih podatkov omogoča visoko verjetnost re-identifikacije, zlasti ob povezovanju z javno dostopnimi viri. V drugem primeru, kjer smo analizirali prve registracije vozil, smo s praktično izvedenim poskusom re-identifikacije pokazali, da je že en sam neposredni identifikator (npr. VIN številka) zadosten za ponovno prepoznavo posameznika. Tretji primer, ki je obravnaval javno objavljene certifikate ekoloških pridelovalcev, pa izpostavlja pomembno razliko med zakonito objavo osebnih podatkov zaradi transparentnosti in dejanskim tveganjem za zasebnost, ki kljub temu ostaja prisotno.

Na podlagi zastavljenih raziskovalnih vprašanj smo prišli do naslednjih ugotovitev:

- V kolikšni meri izbrane odprte podatkovne zbirke omogočajo ponovno identifikacijo posameznikov? Analiza je pokazala, da izbrane podatkovne zbirke v veliki meri omogočajo ponovno identifikacijo, predvsem zaradi prisotnosti kvazi-identifikatorjev in v nekaterih primerih celo neposrednih identifikatorjev. Tveganje se dodatno poveča ob povezovanju z zunanjimi javno dostopnimi viri.
- Katere tehnike anonimizacije oziroma psevdonimizacije so bile uporabljene in kako učinkovite so? V večini analiziranih primerov tehnike anonimizacije niso bile ustrezno uporabljene ali pa sploh niso bile prisotne. Kjer so bile uporabljene, so bile pogosto nezadostne (npr. delna odstranitev podatkov brez celovitega pristopa), kar ni preprečilo možnosti re-identifikacije.
- Ali izbrane zbirke ustrezajo sodobnim standardom varstva osebnih podatkov? Ugotovili smo, da analizirane zbirke v večini primerov ne dosegajo sodobnih standardov varstva osebnih podatkov, saj ne zagotavljajo zadostne ravni anonimizacije in ne vključujejo sistematične ocene tveganja.

Na podlagi ugotovitev predlagamo, da bi moralo biti ob objavi odprtih podatkovnih zbirk obvezno vključeno tudi jasno opisano:

- katere anonimizacijske ali psevdonimizacijske tehnike so bile uporabljene,
- kakšna je bila ocena tveganja za ponovno identifikacijo,
- ter katere omejitve anonimizacije ostajajo prisotne.

Takšna praksa bi bistveno povečala transparentnost obdelave podatkov ter omogočila boljše razumevanje tveganj za končne uporabnike.

S prikazom postopka močne anonimizacije smo pokazali, da lahko s kombinirano uporabo metod supresije, generalizacije, agregacije in psevdonimizacije bistveno zmanjšamo tveganje ponovne identifikacije, ne da bi pri tem popolnoma izgubili uporabno vrednost podatkov. Ugotovili smo, da mora biti anonimizacija izvedena sistematično in na podlagi ocene tveganja, pri čemer je treba upoštevati značilnosti posamezne podatkovne množice.

Omejitve raziskave so v omejenem številu (treh) podatkovnih zbirk, kar omejuje možnost posploševanja rezultatov. Poleg tega so bili eksperimenti re-identifikacije izvedeni na podlagi javno dostopnih virov, kar pomeni, da bi lahko z uporabo dodatnih virov tveganje še dodatno naraslo. Prav tako analiza ni vključevala kvantitativnega merjenja stopnje anonimnosti, temveč je temeljila predvsem na kvalitativni oceni tveganja.

V prihodnje bi bilo smiselno raziskavo razširiti na večje število podatkovnih zbirk ter vključiti avtomatizirane metode za zaznavanje tveganj re-identifikacije. Nadaljnje delo bi lahko vključevalo tudi razvoj orodij za ocenjevanje tveganja ter primerjalno analizo učinkovitosti različnih anonimizacijskih tehnik.

Za organizacije, ki se bodo v prihodnje srečevale s podobnimi izzivi, predlagamo naslednje ukrepe:

- predhodno ocenjevanje tveganja za re-identifikacijo pred objavo podatkov,
- izbiro ustreznih tehnik anonimizacije glede na tip podatkov,
- redno posodabljanje postopkov anonimizacije,
- ter vzpostavljanje ravnovesja med transparentnostjo in varstvom zasebnosti.

Pravilna anonimizacija je ključni element pri zagotavljanju skladnosti z zakonodajo o varstvu osebnih podatkov in pomemben dejavnik pri ohranjanju zaupanja javnosti v odprte podatke. S skrbno načrtovanimi postopki je mogoče doseči ravnotežje med dostopnostjo informacij in zaščito zasebnosti posameznikov.

9 VIRI

[1] Y. G. C. Z. Marijn Janssen, »Understanding the evolution of open government data research: towards open data sustainability and smartness,« 1 April 2021. [Elektronski]. Available: https://www.researchgate.net/publication/351175130_Understanding_the_evolution_of_open_government_data_research_towards_open_data_sustainability_and_smartness. [Poskus dostopa 22 April 2026].

[2] M. A. R. T. Di Wang, »Open access data policies and technologies,« 1 Marec 2025. [Elektronski]. Available: https://www.sciencedirect.com/science/article/pii/S2543925125000014. [Poskus dostopa 22 April 2025].

[3] Europska unija, »EU Open Data Portal,« 10 april 2025. [Elektronski]. Available: https://data.europa.eu/en. [Poskus dostopa 26 Maj 2026].

[4] A. O. E. C. Fatemeh Ahmadi Zeleti, »Exploring the economic value of open government data,« 1 November 2016. [Elektronski]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0740624X16300077. [Poskus dostopa 22 Maj 2026].

[5] Slovenija, »OPSI,« 10 April 2025. [Elektronski]. Available: https://podatki.gov.si/. [Poskus dostopa 26 Maj 2026].

[6] Amerika, »DATA.GOV,« USA, https://data.gov/.

[7] Parlament, Evropski, »UREDBA (EU) 2016/679 EVROPSKEGA PARLAMENTA IN SVETA,« 27 April 2016. [Elektronski]. Available: https://eur-lex.europa.eu/legal-content/SL/TXT/?uri=CELEX:32016R0679. [Poskus dostopa 7 Avgust 2025].

[8] Slovenija, »Pravno informacijski sistem republike slovenije,« 15 december 2022. [Elektronski]. Available: https://pisrs.si/pregledPredpisa?id=ZAKO7959. [Poskus dostopa 26 April 2026].

[9] DPO Europe GmbH, »Uvodna izjava 26,« GDPR text, 27 Maj 2026. [Elektronski]. Available: https://gdpr-text.com/sl/read/recital-26/. [Poskus dostopa 25 Junij 2025].

[10] K. Gardhouse, »Direct and Indirect Personal Identifiers: What are they?,« PRIVATEAI, 3 April 2023. [Elektronski]. Available: https://www.private-ai.com/en/2023/04/08/personal-identifiers/. [Poskus dostopa 6 Maj 2025].

[11] DPO Europe GmbH, »Člen 4 SUVP (GDPR). Opredelitev pojmov,« GDPR TEXT, 27 Maj 2026. [Elektronski]. Available: https://gdpr-text.com/sl/read/article-4/. [Poskus dostopa 23 Julij 2025].

[12] ISSS, »Direct Identifier and Indirect Identifier in Data Privacy,« ISSS UK, 2 Februar 2024. [Elektronski]. Available: https://www.isss.org.uk/2024/02/02/direct-identifier-and-indirect-identifier-in-data-privacy/. [Poskus dostopa 5 Maj 2025].

[13] J. Powar, »SoK: Managing risks of linkage attacks on data privacy,« 1 Februar 2023. [Elektronski]. Available: https://petsymposium.org/popets/2023/popets-2023-0043.pdf. [Poskus dostopa 5 Maj 2025].

[14] S. Latanya, »k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY,« 1 Maj 2002. [Elektronski]. Available: https://epic.org/wp-content/uploads/privacy/reidentification/Sweeney_Article.pdf. [Poskus dostopa 26 April 2026].

[15] G. M. F. O. N. S. Medjahed Amina Fatima Zohra, »Advancing Heterogeneous Data Integration,« 1 Januar 2025. [Elektronski]. Available: https://www.sciencedirect.com/science/article/pii/S1877050925005964. [Poskus dostopa 22 April 2025].

[16] Q. Wang, »An Enhanced K-Anonymity Model against,« October 2011. [Elektronski]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=48555d692f13621f3db89904873e744c248e6c18#page=67. [Poskus dostopa 5 Maj 2025].

[17] N. L. L. Suresh Venkatasubramanian, »Ninghui Li Tiancheng Li,« Department of Computer Science, Purdue University, 26 April 2026. [Elektronski]. Available: https://personal.utdallas.edu/~mxk055100/courses/privacy08f_files/tcloseness.pdf. [Poskus dostopa 19 julij 2025].

[18] Proofpoint, »Throw Back Hack: The Infamous AOL Data Leak,« Proofpoint, 21 Marec 2021. [Elektronski]. Available: https://www.proofpoint.com/us/blog/insider-threat-management/throw-back-hack-infamous-aol-data-leak. [Poskus dostopa 5 Maj 2025].

[19] A. Narayanan, »Robust De-anonymization of Large Sparse Datasets,« IEEE, 1 Januar 2008. [Elektronski]. Available: https://ieeexplore.ieee.org/document/4531148. [Poskus dostopa 8 Maj 2025].

[20] Avstralija, »MBS/PBS data publication,« 23 Marec 2018. [Elektronski]. Available: https://www.oaic.gov.au/privacy/privacy-assessments-and-decisions/privacy-decisions/investigation-reports/mbspbs-data-publication. [Poskus dostopa 5 Maj 2025].

[21] Australska vlada, »MBS/PBS data publication,« Australina goverment, 23 Marec 2018. [Elektronski]. Available: https://www.oaic.gov.au/privacy/privacy-assessments-and-decisions/privacy-decisions/investigation-reports/mbspbs-data-publication. [Poskus dostopa 25 Junij 2025].

[22] U. H. Khan, »All You Need to Know About Data Aggregation,« Astera, 3 Julij 2024. [Elektronski]. Available: https://www.astera.com/type/blog/data-aggregation/. [Poskus dostopa 6 Maj 2025].

[23] Connecticut State, »Open Data Aggregation and Suppression Guidelines,« Connecticut State, 26 Maj 2026. [Elektronski]. Available: https://portal.ct.gov/-/media/ct-data/data_aggregation_guidance_v2.pdf. [Poskus dostopa 6 Maj 2025].

[24] Informatica, »Understanding Data Generalization & Advanced De-identification Techniques,« Informatica, 23 November 2023. [Elektronski]. Available: https://www.informatica.com/blogs/data-generalization-advanced-de-identification.html. [Poskus dostopa 6 Maj 2025].

[25] Evropska unija, »Guidelines 01/2025 on Pseudonymisation,« Evropa, 21 Januar 2025. [Elektronski]. Available: https://www.edpb.europa.eu/system/files/2025-01/edpb_guidelines_202501_pseudonymisation_en.pdf. [Poskus dostopa 6 Maj 2025].

[26] J. G. D. K. M. V. A. Machanavajjhala, »CS rochester,« 7 April 2007. [Elektronski]. Available: https://www.cs.rochester.edu/u/muthuv/ldiversity-TKDD.pdf. [Poskus dostopa 27 Maj 2026].

[27] N. Li, »CS purdue,« 1 Januar 2001. [Elektronski]. Available: https://www.cs.purdue.edu/homes/ninghui/papers/t_closeness_icde07.pdf. [Poskus dostopa 20 julji 2025].

[28] A. R. Cynthia Dwork, »Differential privacy,« 1 Januar 2014. [Elektronski]. Available: https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf. [Poskus dostopa 5 Maj 2025].

[29] K2view, »What is Synthetic Data Generation?,« K2view, 23 Marec 2025. [Elektronski]. Available: https://www.k2view.com/what-is-synthetic-data-generation/. [Poskus dostopa 7 Maj 2025].

▪

Žan Povše je zaposlen na Fakulteti za elektrotehniko, računalništvo in informatiko, Univerze v Mariboru, kjer deluje na področju raziskovanja. Na isti fakulteti je uspešno zaključil magistrski študij ter svoje strokovno in raziskovalno delo nadaljuje z aktivnim sodelovanjem pri raziskovalnih projektih.

▪

Marko Kompara je raziskovalec in asistent na Fakulteti za elektortehniko, računalništvo in informatiko, kjer je tudi pridobil doktorat z disertacijo na temo protokolov za dogovarjanje ključev v omejenih okoljih. Od takrat je bil in je še vedno vključen v številke projekte na podoročju kibernetske varnosti (npr. CyberSec4Europe, Cyber F-IT, RUKIV, RPUP, AKADIMOS). Njegovi raziskovalni interesi so zasebnost, kriptografija, brezžične komunikacije in varnost informacijskih sistemov.

▪

Lili Nemec Zlatolas je izredna profesorica na Univerzi v Mariboru, Fakulteti za elektrotehniko, računalništvo in informatiko, njeno raziskovalno področje pa obsega kibernetsko varnost. Aktivna je pri predmetih informacijske varnosti, osnov programiranja in podatkovnih baz. Je dejavna v različnih mednarodnih in nacionalnih projektih, med drugim v Horizon Europe projektu WEM Cyber, Erasmus+ projektu SafeWeb, sodelovala pa je tudi v projektu Horizon 2020 CONCORDIA, enem največjih evropskih projektov na področju kibernetske varnosti. Prav tako je sodelovala pri nacionalnem projektu RUKIV – Razvoj programov usposabljanj za kibernetsko varnost, namenjenem izobraževanju in krepitvi kompetenc na področju kibernetske varnosti. Sodeluje tudi v delovnih skupinah Diversity & Inclusion pri organizaciji Informatics Europe in skupini DiverseIT pri CEPIS, ter združenju WOMEN4CYBER, kjer prispeva k spodbujanju raznolikosti, enakosti in vključevanja v informatiki ter IKT.