PROBLEMATIKA OHRANJANJA ZASEBNOSTI PRI PODATKOVNEM RUDARJENJU DOKUMENTOV Z OBČUTLJIVIMI PODATKI

Matjaž Kragelj,
National and University Library,
Turjaška 1, 1000 Ljubljana, Slovenia
matjaz.kragelj@nuk.uni-lj.si

Mirjana Kljajić Borštnar,
University of Maribor, Faculty of Organizational Sciences,
Kranj, Slovenia,
e-mail: Mirjana.Kljajic@um.si

Alenka Brezavšček,
University of Maribor, Faculty of Organizational Sciences,
Kranj, Slovenia,
e-mail: Alenka.Brezavscek@um.si

Povzetek

V prispevku obravnavamo problem, s katerim se soočamo pri uporabi dokumentov, ki poleg vsebinskih podatkov vsebujejo tudi občutljive podatke o posamezniku, ki omogočajo njegovo razkritje, tudi ko to ni zaželeno. Med področja, kjer nastane veliko podatkov te vrste, štejemo zlasti zdravstveno varstvo, transport, kazenski pregon in nacionalno varnost, izobraževanje, sodobne internetne storitve, področje sodobnih aplikacijskih ekosistemov, internet stvari, finančni sektor in odprte podatke državne uprave. Cilj je zaščititi zasebnost subjekta ter hkrati zagotoviti kakovostne podatke za nadaljnje poglobljene analize in s tem nudenje novih znanj za naprej. Za reševanje omenjenih izzivov na področju podatkovnega rudarjenja se je razvilo posebno podpodročje, imenovano PPDM – Privacy Preserving Data Mining, ki se ukvarja z ohranjanjem zasebnosti pri tem procesu. Sistematično smo pregledali relevantno literaturo podpodročja PPDM in opisali glavne metode in tehnike. Tehnike PPDM so zasnovane tako, da zagotavljajo določeno raven zasebnosti, obenem pa ohranjajo uporabnost podatkov, da se lahko uporaba še vedno učinkovito izvaja na transformiranih podatkih. Metode, s katerimi dosegamo zaščito posameznika na eni in uporabno vrednost podatkov na drugi strani v grobem delimo na metode razprševanja podatkov, metode izkrivljanja (z uporabo anonimizacije, randomizacije, vrtenja in vnašanjem šuma v podatke) ter metode šifriranja podatkov. Za doseganje višje zaščite lahko uporabimo tudi kombinacije teh metod. Poleg pregleda metod smo podali nekaj praktičnih primerov ter našteli domene oz. področja, kjer se kaže potreba po nadaljnji analizi in ponovni uporabi podatkov, a hkrati potreba po anonimizaciji oz. prikritju lastnika (subjekta) in njegovih podatkov (atributov).

Ključne besede: podatkovno rudarjenje, osebni podatki, ohranjanje zasebnosti, metode ohranjanja zasebnosti podatkov, pregled literature, varnost podatkov

UVOD

S pojavom interneta se je pojavila potreba in možnost po orodjih za iskanje, razvrščanje in kasneje analizo zbranih digitalnih podatkov. Kot primer lahko navedemo, da sta še v prejšnjem stoletju Smith & Chang (1997), razvila spletno orodje - WebSeek, ki je (bilo) namenjeno iskanju in sortiranju slik s spleta. Dokler so bile količine podatkov obvladljive, so bile tudi metode in orodja razmeroma enostavna, v današnjem času pa potreba po zbiranju in obdelavi podatkov strmo narašča (Mendes & Vilela, 2017). V podjetju IBM ugotavljajo[1], da je bilo v letih 2012 in 2013 ustvarjeno več kot 90% vseh podatkov na svetu (Devakunchari, 2014), do leta 2025, pa bo za analizo primernih več kot 150 zetabajtov[2] (10^9 terabajtov). Z naraščanjem količine digitalnih podatkov, ki jih je potrebno obdelati, se je pojavila potreba po vpeljavi naprednejših metod, ki temeljijo na principih umetne inteligence in strojnega učenja, natančneje – podatkovnega rudarjenja. Gre za proces pridobivanja implicitnih informacij in znanj, ki bi lahko bile koristne, črpanje le-teh pa poteka bodisi iz množičnih, neurejenih, nepopolnih, nejasnih in naključnih in podobno ne nujno strukturiranih podatkovnih struktur (Sahu, Shrma & Gondhalakar, 2008).

Med digitalnimi in digitaliziranimi dokumenti, ki jih rudarimo, so tudi takšni, ki so občutljive narave in zaradi tega zahtevajo posebno skrb in previdnost pri hranjenju, obdelavi in posredovanju tretjim osebam. Neustrezno postopanje pri rudarjenju takih dokumentov lahko povzroči grožnjo zasebnosti preko razkritja identitete in s tem posredno povezanih podatkov (atributov), saj atribute navadno povezujemo z lastništvom ravno preko identitete (Shanthi & Karthikeyan, 2012). V skladu z definicijo iz slovarja Cambridge, je zasebnost definirana kot pravica posameznika, da obdrži svoje osebne podatke, zadeve in odnose v tajnosti (Singh, 2019). Definicija sledi 12. členu Splošne deklaracije o človekovih pravicah, ki pravi, da se ni dovoljeno nikomur samovoljno vmešavati v zasebno življenje, družino, dom ali dopisovanje, prav tako pa ni dovoljeno žaljenje časti in dobrega imena slehernega posameznika. Vsakdo ima pravico do pravnega varstva pred takšnim vmešavanjem ali napadi[3]. Težava, ki jo pri rudarjenju dokumentov z občutljivo vsebino zaznavamo, je naslednja: analizo želimo izvajati tako, da v popolnosti ohranjamo vrednosti in pomen podatkov, ki pomenijo atribute, a hkrati želimo zaščititi občutljive podatke o posamezniku, torej zaščititi njegovo zasebnost. Kot ugotavljajo Gokulnath et al. (2015), je ohranjanje zasebnosti pri rudarjenju občutljivih in osebnih podatkov ključnega pomena za učinkovito izvedeno podatkovno rudarjenje.

V drugem poglavju bomo navedli glavna področja, kjer se ustvarjajo podatki in dokumenti z občutljivimi atributi. Med temi področji veljajo podatki o zdravstvenem varstvu za najpomembnejše, a so hkrati najbolj občutljivi, saj vsebujejo vse zasebne podatke, ki so informacije o pacientu, kot so bolezni, zdravljenje, recept, ime, naslov itd. Takšna zbirka podatkov katerekoli zdravstvene organizacije je dovzetna za različne napade (Singh, 2019). Veliki podatki (ang. big data) ponujajo epidemiologom, zdravnikom in strokovnjakom za

---

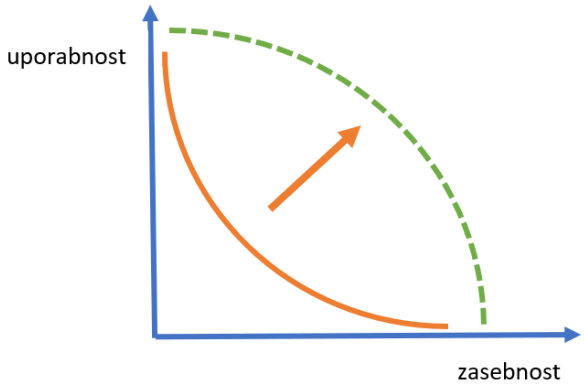
[1] 2.5 quintillion bytes of data created every day. How does CPG & Retail manage it?, Vir: https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/, dostopano: januar 2020.

[2] https://www.ibm.com/blogs/services/2020/05/28/how-to-manage-complexity-and-realize-the-value-of-big-data/ , dostopano: september 2021

[3] Splošna deklaracija človekovih pravic, https://www.gov.si/assets/ministrstva/MZZ/Dokumenti/multilaterala/mednarodno-pravo/bf55533011/Splosna-deklaracija-clovekovih-pravic.pdf dostopano: april 2021

zdravstveno politiko veliko priložnost za presojo na podlagi analize dostopnih podatkov, ki bo sčasoma dvignila raven oskrbe bolnikov (SA, 2018).

Kot prikazuje slika 1, je izziv iskanje metodologij in tehnik za zaščito zasebnosti in nerazkrivanje občutljivih podatkov na eni strani in nudenje kvalitetnih podatkov o uporabnikih za raziskave, analizo in ustvarjanje dodane vrednosti z novimi znanji na drugi strani. K reševanju tega izziva je potrebno pristopiti skrbno, saj lahko poseganje v same podatke in njihovo preoblikovanje okrni njihovo uporabnost, kar lahko vodi v napačne interpretacije zavajajoče informacije ter neustrezne odločitve (Mendes & Vilela, 2017).



Slika 1: Iskanje ravnovesja med zagotavljanjem zasebnosti uporabnikov in izvajanjem kvalitetnega nudenja podatkov o uporabnikih
Vir: povzeto po (Cranor, Rabin, Shmatikov, Vadhan, & Weitzner, 2016a)

Za reševanje te vrste izzivov se je na področju podatkovnega rudarjenja razvilo posebno podpodročje, ki se ukvarja z ohranjanjem zasebnosti pri rudarjenju dokumentov z občutljivimi podatki. To področje se je v tuji literaturi uveljavilo pod imenom "Privacy preserving data mining" – PPDM (Agrawal & Srikant, 2000a; Lindell & Pinkas, 2000a). Metodologije PPDM so zasnovane tako, da zagotavljajo določeno raven zasebnosti, obenem pa ohranjajo uporabnost podatkov tako, da lahko rudarjenje še vedno učinkovito izvajamo na transformiranih podatkih. Da gre za zelo pomembno področje, je opozorila tudi Evropska komisija v "Mnenju številka 5" (EK, 2014). PPDM temelji na uporabi različnih metod, ki prispevajo k ohranjanju zasebnosti, kot npr. anonimizacija, randomizacija, uporaba permutacij, vnašanje šuma v podatke, kriptografske tehnike in druge (Shanthi & Karthikeyan, 2012). Ker je področje razmeroma novo, a hkrati zanimivo, posebno danes, pri novih izzivih v zdravstvu (Covid19), bomo v nadaljevanju pripravili sistematičen pregled metod, ki se z izbrano problematiko ukvarjajo.

PODROČJA UPORABE PPDM

V tem razdelku bomo navedli področja, kjer prihaja do potreb po rudarjenju podatkov/dokumentov, ki so po svoji naravi lahko občutljivi (Cranor et al., 2016). Čeprav avtorji navajajo domene uporab v ZDA, lahko izpostavljena področja uporabe preslikamo tudi v našo regijo. Področja, kjer je potreba po uporabi metod PPDM pogosta in intenzivna, so naslednja:

- **Zdravstveno varstvo**. Informatizacija procesov v zdravstvu lahko močno izboljša zdravstvene storitve, vključno z možnostjo natančnejše diagnostike, bolj omogočanjem bolj

prilagojene in usklajene oskrbe, hitrejšega razvoja novih načinov zdravljenja, učinkovitejšega zdravljenja ob nižjih stroških. Izziv predstavlja razkritje občutljivih zdravstvenih podatkov, širjenje le-teh (legalno ali nelegalno), kar lahko privede do različnih neprijetnih in uporabniku škodljivih situacij (npr. do diskriminacije pri zaposlovanju[4]).

- **Transport**. Koristi informacijske tehnologije pri prevozu so lahko v zmanjšanju zastojev, preprečevanju nesreč, zmanjšanju smrti in poškodb, povečanju učinkovitosti porabe goriva, itd. Skrbi glede zasebnosti izvirajo iz možnosti sledenja gibanj posameznikov preko navigacijskih sistemov, cestnih senzorjev, prometnih kamer, zbiranja podatkov v avtomobilu in komunikacije med avtomobili.
- **Kazenski pregon in nacionalna varnost**. Organi pregona in obveščevalne agencije zbirajo in analizirajo različne vrste podatkov (kazenski zapisi in dopolnilne informacije) z namenom ustvarjanja "virtualne slike" posameznika, ki pomaga pri reševanju kriminalnih dejanj, preprečevanju napadov in sledenju teroristom. Glede ohranjanja zasebnosti so glavne skrbi te, da organizacije kot npr. policija množično zbirajo informacije o splošni populaciji, kar povečuje možnost nedovoljene uporabe in s tem niža kvaliteto učinka nadzora, zaradi nezakonitega odtekanja podatkov. Kot primer lahko navedemo primer delovanja slovenske policije[5].
- **Izobraževanje.** Informacijska tehnologija in podatki o izobraževanju lahko izboljšajo izobraževanje z nudenjem prilagodljivih in prilagojenih vsebin in spletnih tečajev. Tveganje glede zasebnosti izhaja iz občutljivosti podatkov o angažiranosti in uspešnosti uporabnikov (učenci, dijaki, študenti).
- **Sodobne internetne storitve.** Iskalniki, družbena omrežja, spletne video storitve in spletni trgovci imajo dostop do bogatega niza podatkov, ki jih je mogoče uporabiti za koristne namene, vključno z napredno personalizacijo vsebine in povezovanjem z drugimi (navadno poslovnimi) subjekti. Zaskrbljenost se kaže pri uporabi, zlorabi in skupni rabi podatkov za namene izven področja namembnosti.
- **Sodobni aplikacijski ekosistemi.** Naprave, kot so pametni telefoni, spletni brskalniki, pametne ure in njihove aplikacije, uporabnikom zagotavljajo veliko uporabnost (npr. pri športnih aktivnostih, zaradi vgrajenih GPS naprav in pedimetrov), zabavo in funkcionalnost. Kot potencialno težavo lahko navedemo sledljivost uporabnika (zaradi GPS podatkov, ki se zbirajo v aplikaciji). Izziv predstavlja zagotovitev, da aplikacije spoštujejo zasebnost in varnost njihovih uporabnikov.
- **Internet stvari.** Pametna mesta, pametne zgradbe, pametni domovi, pametni hladilniki, televizije ipd. omogočajo izboljšanje življenjskih razmer, produktivnosti in kakovosti življenja. Vendar pa se lahko isti podatki uporabijo za sledenje, kdaj so posamezniki doma, katere TV programe gledajo, katera spletna mesta obiskujejo, njihov urnik spanja in drugo vedenje. Tveganje predstavlja izkoriščanje takšnih podatkov za druge namene, kot npr. zavarovalne police (na voljo so informacije o prehrambnih navadah, aktivnostih in s tem tveganjih – profiliranje uporabnikovih navad), nezaželeno oglaševanje ali kriminal.
- **Finančni sektor.** Podatki finančnih institucij lahko regulatorjem pomagajo pri oceni skladnosti in omogočijo analizo trendov ter opozorijo na nevarnosti, kot npr. prihajajoča

---

[4] Tri tedne po tem, ko je Nydia Velázquez zmagala, kot kandidatka Demokratske stranke v New Yorku za predstavnico v Ameriškem predstavniškem domu, je nekdo iz bolnišnice St. Claire v New Yorku, preko faksa poslal Velázquezino zdravstveno kartoteko časopisu New York Post. V dokumentu je bila podrobno opisana oskrba pacientke, ki je tam pristala zaradi poskusa samomora. Poskus samomora se je zgodil nekaj let pred volitvami, na katerih je zmagala. Povzeto po (Wu and Velázquez, 2000).

[5] Avstrijska nevladna organizacija AlgorithmWatch je v svojem poročilu analize policijske uporabe tehnologije za prepoznavanje obrazov zapisala, da slovenska policija od leta 2014 uporablja doma razvito tehnologijo za prepoznavanje obrazov. Kot navajajo, gre za problem regulacije te tehnologije. Informacijska pooblaščenka je tako med leti 2015 in 2019 izdala več negativnih mnenj Ministrstvu za notranje zadeve. Slovenska policija in biometrijske metode nadzora https://www.eticen.it/2019/12/12/slovenska-policija-in-biometrijske-metode-nadzora, dostopano: januar 2020

finančna kriza. Vendar so finančni podatki občutljivi ne le na ravni posameznih strank temveč tudi na ravni institucij, saj razkrivajo lastniške informacije o strategijah in tržnih deležih.

- **Odprti podatki državne uprave**. Vlade na vseh ravneh sproščajo velike količine podatkov, da bi povečale zaupanje in preglednost ter omogočile inovativne aplikacije. Vendar se te objave podatkov pogosto nanašajo na občutljive informacije o državljanih. Lahko navedemo nekaj primerov takšnih spletnih storitev pri nas - eDavki[6], eUprava[7], eVem[8], eZdravje[9] in drugi.

METODE ZA ZAŠČITO OBČUTLJIVIH PODATKOV, KI PRI PODATKOVNEM RUDARJENJU OMOGOČAJO OHRANJANJE ZASEBNOSTI

Analitika velikih podatkov (ang. big data) sestoji iz petih stopenj oz. faz, in sicer: pridobivanje podatkov, shranjevanje podatkov, upravljanje s podatki, analiza podatkov ter vizualizacija podatkov in poročanje. Pri dveh od teh se soočamo z ohranjanjem zasebnosti: **shranjevanje podatkov** in, **upravljanje s podatki** (Pawar, Ahirrao, & Churi, n.d.; SA, 2018; Vassakis, Petrakis, & Kopanakis, 2018). Podatki, ki jih pridobivamo, so lahko strukturirani, delno strukturirani ali pa gre za nestrukturirane podatke. Podatke lahko pridobimo iz ustnih virov (intervju, telefonski pogovor) ali pisnih virov (npr. anketa, izvid, vprašalnik, diagnoza, mnenje). Pogosto so viri podatkov tudi slikovni ali multimedijski (npr. magnetna resonanca, računalniška tomografija, ultrazvok itd.). Ne nazadnje v dobi interneta lahko podatke pridobimo tudi iz spletnih anket, vprašalnikov, poskusov. Pogosto podatke že sami shranjujemo, posredujemo in s tem ponujamo v varne ali ne - oblačne storitve (ang. cloud services). Gre za podatke, pridobljene iz pametnih naprav, telefonov z uporabo aplikacij, kot so npr. Drive, Training Peaks, Polar Flow, Strava, in podobne.

Glavni nalogi uporabe PPDM, kot ju navedejo Xu, Jiang, Wang, Yuan, & Ren (2014) sta soočanje in razreševanje problematike neprimernosti neposredne uporabe občutljivih, surovih podatkov (npr. številka osebne izkaznice, mobilnega telefona) za rudarjenje ter potreba po izključitvi občutljivih rezultatov rudarjenja, katerih razkritje bi povzročilo kršitev zasebnosti. Pionirsko delo, opis prvih metod rudarjenja občutljivih podatkov na tem področju je opisano v člankih (Agrawal & Srikant, 2000b; Lindell & Pinkas, 2000b).

Pri rudarjenju podatkov z namenom zaščite zasebnosti se uporabljajo različne metode. Usmerjene so predvsem v omejevanje dostopa in uporabe občutljivih podatkov za nadaljnjo analizo, ki bi sicer lahko identificirali posameznika. Po Abdul, Aldeen, Salleh, & Razzaque (2015), Qi & Zong (2012) in Taneja, Khanna, & Tilwalia, (2016) med glavne metode umeščamo:

- **Razprševanje podatkov** (ang. partitioning): podatke distribuiramo po eni ali več podatkovnih baz.
- **Izkrivljanje podatkov** (ang. data distortion, perturbation): pri tem načinu posegamo v podatke, ki jih želimo uporabiti in katerih vrednost je tista, ki jo želimo zaščititi. Sem umeščamo **anonimizacijo, randomizacijo, vrtenje, vnašanje šuma** v podatke.
- **Kriptografske tehnike šifriranja** (ang. cryptographic technique): gre za različne, pogosto računsko potratne metode, kjer se podatki s pomočjo ustreznega šifrirnega

[6] Edavki, https://edavki.durs.si/EdavkiPortal/OpenPortal/CommonPages/Opdynp/PageA.aspx
[7] eUprava, https://e-uprava.gov.si
[8] eVem, http://evem.gov.si/evem/drzavljani/zacetna.evem
[9] eZdravje, https://zvem.ezdrav.si/e-zdravje

algoritma (simetričnega ali asimetričnega) in šifrirnega ključa pretvorijo v neberljivo obliko.

Razprševanje podatkov

Pri procesu hranjenja podatkov že lahko govorimo o varnosti in eni od tehnik zagotavljanja anonimnosti. O razprševanju podatkov (ang. partitioning) govorimo, kadar podatke, shranjene v eni podatkovni bazi, porazdelimo (razpršimo) v več podatkovnih baz. Podatke lahko razpršimo horizontalno, vertikalno ali funkcionalno. Vse omenjeno lahko počnemo centralizirano (na enem mestu), ali pa distribuirano (na več lokacijah).

Pri horizontalni razpršitvi podatkov pridobimo predvsem na razširljivosti (ang. scalability) in učinkovitosti v smislu hitrejšega dostopa do podatkov (ang. performance), na varnosti (ang. security) pa precej manj, saj so v posamezni relaciji (povezava med dvema ali več entitetami) zbrani vsi atributi. Iz vidika varnosti je poskrbljeno zgolj za to, da niso vsi podatki o vseh subjektih zbrani na enem mestu, ampak razpršeni po več podatkovnih bazah.

Pri vertikalni razpršitvi gre za nasproten proces. Hitrost dostopa in uporabe podatkov pada, saj je treba atribute iz več podatkovnih baz med sabo združiti. Pri tem pristopu imamo atribute razdeljene v več skupin, vsaka izmed skupin pa je v svoji podatkovni bazi. Navadno ima vsaka relacija skupni ključ, ki povezuje podatke med seboj.

Pri funkcionalni razpršitvi ločujemo podatke glede na funkcijo, oz. uporabo[10] (Panse & Paikrao, 2017; Qi & Zong, 2012).

Omeniti je treba, da z razprševanjem samih podatkov ne spreminjamo, lahko zgolj omejimo dostop do njih. Pri vertikalni razpršitvi lahko uporabniku ponudimo dostop do delov podatkov (atributov, ki jih potrebuje), uporabnik pa si ne more ustvariti "celotne slike", ker je dostop do celotne vsebine omejen. Za prikaz celotne slike moramo združiti podatke iz različnih podatkovnih baz oz. sistemov.

Izkrivljanje podatkov

V skupino izkrivljanja podatkov (ang. data perturbation) sodijo metode, tehnike in algoritmi, ki podatke spreminjajo, ali pa vanje dodajajo šum. Največ literature s področja PPDM je posvečeno ravno temu segmentu, ki je tudi najbolj kompleksen. Večinoma gre za matematične metode, ki posegajo v podatke in jih z uporabo vektorjev, matrik, faktorjev – spreminjajo. Na začetku je bilo implementiranih nekaj metod, ki so temeljile zgolj na naključnem seštevanju in množenju, a niso bile imune na praktično nobeno vrsto napada (Upadhyay, Sharma, Sharma, Bharadwaj, & Seeja, 2018).

Cilj izkrivljanja podatkov je ponuditi informacije, ki jih je mogoče uporabiti za rudarjenje na način, da ostane prikrita identifikacija lastnika (subjekta) atributov. Atribute v grobem delimo v tri skupine: *identifikacijski atributi* (identificirajo subjekt), *javni* ali *kvazi atributi* (njihove vrednosti lahko pridobimo tudi v drugih, javnih bazah podatkov, kot je npr. volilni imenik, podatki v profilu na socialnih omrežjih (letnik ali starost, kraj, naslov, itd.) ter *privatni atributi*, ki (v primeru bolnišničnega kartona) opisujejo stanje bolnika, bolezen, zdravljenje. Cilj je zagotoviti dostop do privatnih atributov, s pomočjo katerih bi radi ugotovili povezave,

[10] Horizontal, vertical, and functional data partitioning, https://docs.microsoft.com/en-us/azure/architecture/best-practices/data-partitioning, dostopano januar 2010

odvisnosti in s tem prišli do novih spoznanj in znanja, hkrati pa zavarovali informacije, ki identificirajo posameznika (Gionis & Tassa, 2009).

Medtem, ko pri metodah razprševanja in šifriranja podatke skrivamo, delimo, distribuiramo, jih s pomočjo anonimizacijskih tehnik **izkrivljamo** - z namenom nudenja nadaljnji uporabi. Obstoječe metode na področju izkrivljanja podatkov opisujejo Sachan, Roy, & Arun (2013) in Qi & Zong (2012), ki med metodami omenjajo metodo **k-anonimnosti, generalizacijo, klasifikacijo, gručenje, povezovalna pravíla, porazdeljeno ohranjanje zasebnosti, l-raznolikost, t-podobnost, randomizacijo in drevesno razvrščanje.**

Metode, ki uporabljajo algoritme dodajanja šuma, permutacij in randomizacijske tehnike, imajo prednosti, kot so npr. neodvisno izvajanje skozi vse vrednosti atributov (neodvisnost) in ohranjanje statistične natančnosti po rekonstrukciji originalnih podatkov. Med slabosti pa umeščamo zmanjšano uporabnost atributov pri generalizaciji v intervale, skrivanje robnih podatkov, kar zahteva visoko uporabo šuma in s tem znižuje uporabnost podatkov (Mendes & Vilela, 2017). V članku Li & Sarkar (2006) ponujata izboljšan pristop spreminjanja podatkov (dodajanja šuma), kot je zgolj povečevanje / zmanjševanje numeričnih vrednosti atributa za isti faktor, vrednost ali rotacijo. Ta temelji na razvrščanju v drevesa in na uporabi največje variance vrednosti med atributi.

Generalizacije se poslužujemo v primeru, ko uporabnik podatkov ne potrebuje natančnih vrednosti atributov in lahko le-te posplošimo, npr. v intervale, ki niso nujno vedno enako široki. Kot primer - pri osebnem dohodku lahko za nižje vrednosti uporabimo ožji, pri višjih vrednostih pa širši interval. Namesto prave vrednosti ponudimo interval, na katerem vrednost leži. Temu načinu spreminjanja rečemo *interval vrednostnih razredov* (ang. value-class interval). Drugi način je *vrednostno izkrivljanje* (ang. value distortion). Gre za dodajanje šuma, saj namesto vrednosti x_i ponudimo spremenjeno vrednost $Z_{(i)} = x_i + r$, kjer je r naključna vrednost iz nekega intervala $[-a, +a]$, ali pa iz normalne (Gaussove) porazdelitve (Agrawal & Srikant, 2000a). Pri izkrivljanju podatkov s to metodo, govorimo o izkrivljanju podatkov z dodatnim šumom, multiplikativnim šumom ali permutacijo (EK, 2014; Upadhyay et al., 2018). Kot šum si lahko predstavljamo podatek, ki je spremenjen na nivoju vseh elementov - npr. podatek o teži, višini pacientov, je za vse paciente spremenjen za določen faktor.

k-anonimizacijo uvrščamo v drugo skupino anonimizacijskih tehnik. Cilj tehnike je anonimizirati člana množice ali skupine, na način generalizacije vrednosti atributov (npr. mesto ali poštno številko z regijo, višino zaokrožiti na desetice, starost v interval, itd.). Pristop k-anonimnosti sta prva predlagala Samarati & Sweeney v (1998) in Sweeney (2002). Cilj postopka k-anonimizacije je vključitev vsakega posameznika, o katerem podatki so nam na voljo v večjo skupino, s k-posamezniki in s tem povečati negotovost identifikacije posameznika. Tehnika k-anonimizacija ni imuna na uporabo posrednega znanja (ang. background knowledge), ki ga ima napadalec. To nastane zaradi slabo definiranih intervalov ali znanja, ki ga ima napadalec o posamezniku, katerega podatke preiskuje (npr. zaradi atributov, ki niso skriti). Mendes & Vilela (2017) navajata, da so prednosti metode k-anonimizacija enostavnost definicije protokola in velik nabor obstoječih algoritmov za doseg k-anonimizacije, kot slabost pa predvidevanje, da vsak zapis predstavlja podatke o edinstvenem posamezniku. Če temu ni tako, se razred enakovrednosti s k zapisi ne poveže nujno s k različnimi posamezniki. Prav tako privatni (občutljivi) atributi ne pridejo v poštev za anonimizacijo v primeru, če imajo vsi podatki razreda isto vrednost.

Nadgradnja k-anonimnosti, je **l-raznolikost**. Dodana je še ena omejitev, in sicer, da se vsak atribut v ekvivalenčnem razredu pojavi vsaj l-krat, tako da je napadalec vedno precej negotov glede atributov, tudi če ima osnovne informacije o določenem posamezniku, na katerega se nanašajo osebni podatki (Machanavajjhala, Gehrke, Kifer, & Venkitasubramaniam, 2006). Med slabosti metode je treba poudariti predvsem dejstvo, da je zahtevna za implementacijo (težko doseči primerno obliko), poleg tega pa se napadalec, v primeru, da so občutljivi atributi nekega razreda enaki, nauči / izve vrednost tega atributa za določenega posameznika (Mendes & Vilela, 2017; Vasa & Modi, 2018).

Čeprav model l-raznolikost učinkovito rešuje težave, ki obstajajo v modelu k-anonimnosti, se model ne more upreti napadom na podobnost (ang. similarity attacks). To pomeni, da je delež vrednosti občutljivega atributa prevelik. V tem primeru ima napadalec veliko verjetnost, da bo pridobil zasebnost posameznika. Zato je znanstvenik Li Ning Hui predlagal model **t-podobnost (ang. t-closeness)**. Ta zahteva, da vrednost razlike med porazdelitvijo vrednosti občutljivih atributov v enakovrednih razredih in porazdelitvijo atributa v celotni podatkovni tabeli ni večja od t. Če je na primer občutljivi atribut številski, l-raznolikost ne upošteva, da so si nekatere vrednosti lahko zelo podobne, da so si blizu, kar rešuje metoda t-podobnost. Ta določa, da mora biti porazdelitev občutljivega atributa v vsakem ekvivalenčnem razredu podobna porazdelitvi v celotni tabeli. To lahko prepreči napade na podobnost in dodatno reši težave, ki obstajajo v modelu l-raznolikosti. Model t-podobnost je velja za najboljši anonimni model varovanja zasebnosti (EK, 2014; Hao & Ya-Bin, 2017; Machanavajjhala et al., 2006; Quirós, Alonso, Díaz, & Montes, 2015).

Šifriranje podatkov

Šifriranje občutljivih podatkov (atributov) je dober pristop k procesu varovanju podatkov, saj ne spreminja podatkov, ne prihaja do izgube (generalizacije), ali šuma. Med slabosti štejemo predvsem težavno implementacijo pri velikih zbirkah podatkov, poleg tega pa rezultat (originalni podatki) odkriva tako javne, kot skrite atribute (Pinkas, 2002; Taneja et al., 2016).

Metode šifriranja podatkov lahko uporabimo v kombinaciji z razprševanjem podatkov, in sicer na dva načina. Podatki so lahko razdeljeni vertikalno skozi več podatkovnih baz (med več lastniki), kjer so šifrirani zgolj zasebni podatki, lahko pa so šifrirani vsi podatki. V praksi se je razvil model "Ohranjanje varnosti na podlagi kontrole dostopa preko vlog" (ang. privacy preserving role based access control approach – PRBAC), ki kombinira vertikalno razprševanje podatkov in tehnologijo šifriranja za dostop do podatkov, ki jih deli na javne in zasebne (Vasudevan, Sukanya, & Aarthi, 2008).

Slabost metode je časovna zahtevnost (enkripcija/dekripcija) ter čas, potreben za sestavljanje relacije (iz n podatkovnih baz). Postopek je moč pohitriti z uvedbo horizontalnega razprševanja podatkov, tedaj je iskana relacija v celoti v eni izmed podatkovnih baz, a je takšna realizacija manj varna (Vasudevan et al., 2008).

Pri uporabi *razprševanja podatkov* in *šifriranja*, ali uporabi kombinacije teh dveh metod, podatki ostajajo v obliki, kot so nastali. S tem, ko jih ne spreminjamo ali izkrivljamo, ostajajo za nadaljnjo analizo potencialno najustreznejši, a bi pri uporabi lahko prišlo do kršitve ohranjanja zasebnosti, predvsem v primeru, če bi do podatkov lahko dostopali nelegalno ali nepooblaščeno.

Model PRBAC in vertikalno, ter funkcionalno porazdeljene podatkovne baze težavo delno odpravljajo, saj je v prvem primeru zagotovljen dostop do podatkov preko vlog dostopa, v drugem pa do pridobitve zgolj dela podatkov (dela atributov relacije). Nobeden od teh dveh

pristopov ni primeren za nudenje podatkov v množično uporabo, npr. za podatkovno rudarjenje, saj ni poskrbljeno za anonimizacijo podatkov, oz. za zakritje povezave med lastnikom podatkom in njegovimi atributi v relaciji drugače, kot s serviranjem zgolj dela podatkov končnemu uporabniku.

ZAKLJUČEK

Pri iskanju ravnovesja med nudenjem podatkov za analizo, z namenom ustvarjanja dodane vrednosti in znanj je potrebno poskrbeti za zaščito identitete in interesov posameznika po anonimnosti. Pri pregledu literature smo zasledili, da se za reševanje omenjenih izzivov na področju podatkovnega rudarjenja razvilo posebno podpodročje, ki se ukvarja z ohranjanjem zasebnosti pri tem procesu. Metode, s katerimi dosegamo zaščito posameznika na eni in uporabno vrednost podatkov na drugi strani v grobem delimo na metode *razprševanja podatkov, metode izkrivljanja* in *metode šifriranja podatkov*. Za doseganje višje zaščite lahko uporabimo tudi kombinacije teh metod.

Za primere iz prakse, kjer bi uporaba tovrstnih metod pripomogla k boljši uporabi (javnih) podatkov in informacij javnega značaja, omenimo Zakon o dostopu do informacij javnega značaja (ZDIJZ-NPB10)[11]. Ta v šestem členu navaja izjeme, kjer dostop do takšnih podatkov ni dovoljen, večinoma zaradi posledic razkritja in s tem kršitve varstva osebnih, ali drugih podatkov. Pri Splošni uredbi o varstvu podatkov (ang. general data protection regulation - GDPR[12]), uredba v 17. členu določa Pravico do izbrisa (»pravico do pozabe«), ki določa pogoje, pod katerimi lahko posameznik zahteva izbris osebnih podatkov iz dokumentov, ter 28. člen, kjer določa pristojnosti in omejitve obdelovalca podatkov. Za omogočanje dostopa do dokumentov takšne narave, sta predlagana šifriranje in revizija dostopov (Evans, 2017), ter izkrivljanje podatkov. Skladno z GDPR se lahko dokumenti objavijo po odstranitvi vseh podatkov, ki identificirajo posameznika (Broen, Trangucci, & Zelner, 2021).

Največ izboljšav in novih metod in algoritmov smo zasledili ravno za področje spreminjanja podatkov (ang. data perturbation). To je edini pristop, ki podatke transformira, a pri tem ohranja (skozi funkcijo transformacije) možnost rekonstrukcije, podatki pa ohranjajo visoko uporabno vrednost za nadaljnjo uporabo – podatkovno rudarjenje.

VIRI IN LITERATURA

Abdul, Y., Aldeen, A. S., Salleh, M., & Razzaque, M. A. (2015). A comprehensive review on privacy preserving data mining. *SpringerPlus*. https://doi.org/10.1186/s40064-015-1481-x

Agrawal, R., & Srikant, R. (2000a). Privacy-preserving data mining. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, *29*(2), 439–450. https://doi.org/10.1145/335191.335438

[11] Zakon o dostopu do informacij javnega značaja, https://zakonodaja.com/zakon/zdijz, dostopano september 2021

[12] GDPR, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02016R0679-20160504,

Agrawal, R., & Srikant, R. (2000b). Privacy-preserving data mining. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, *29*(2), 439–450. https://doi.org/10.1145/335191.335438

Broen, K., Trangucci, R., & Zelner, J. (2021). Measuring the impact of spatial perturbations on the relationship between data privacy and validity of descriptive statistics. *International Journal of Health Geographics*, *20*(1), 1–17. https://doi.org/10.1186/s12942-020-00256-8

Cranor, L., Rabin, T., Shmatikov, V., Vadhan, S., & Weitzner, D. (2016a). Towards a Privacy Research Roadmap for the Computing Community. *Http://Cra.Org/Ccc/Resources/Ccc-Led-Whitepapers/*. Retrieved from http://arxiv.org/abs/1604.03160

Cranor, L., Rabin, T., Shmatikov, V., Vadhan, S., & Weitzner, D. (2016b). Towards a Privacy Research Roadmap for the Computing Community. *Http://Cra.Org/Ccc/Resources/Ccc-Led-Whitepapers/*.

Devakunchari, R. (2014). Analysis on big data over the years. *International Journal of Scientific and Research Publications*. Retrieved from www.ijsrp.org

EK. (2014). Mnenje št. 5/2014 o anonimizacijskih tehnikah. Evropska komisija, Delovna skupina za varstvo podatkov člena 29. Retrieved from https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_sl.pdf

Evans, C. (2017). HOW GDPR WILL SHAKE UP DATA STORAGE. *Computer Weekly*, 25–28.

Gionis, A., & Tassa, T. (2009). K-anonymization with minimal loss of information. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/TKDE.2008.129

Hao, G., & Ya-Bin, X. (2017). Research on privacy preserving method based on T-closeness model. In *2017 3rd IEEE International Conference on Computer and Communications (ICCC)* (pp. 1455–1459). https://doi.org/10.1109/CompComm.2017.8322783

Li, X. B., & Sarkar, S. (2006). A tree-based data perturbation approach for privacy-preserving data mining. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/TKDE.2006.136

Lindell, Y., & Pinkas, B. (2000a). Privacy Preserving Data Mining, 36–54.

Lindell, Y., & Pinkas, B. (2000b). Privacy Preserving Data Mining BT - Advances in Cryptology — CRYPTO 2000. In M. Bellare (Ed.) (pp. 36–54). Berlin, Heidelberg: Springer Berlin Heidelberg.

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006). ℓ-Diversity: Privacy beyond k-anonymity. *Proceedings - International Conference on Data Engineering*. https://doi.org/10.1109/ICDE.2006.1

Mendes, R., & Vilela, J. P. (2017). Privacy-Preserving Data Mining: Methods, Metrics, and Applications. *IEEE Access*. https://doi.org/10.1109/ACCESS.2017.2706947

Panse, P. P., & Paikrao, P. L. (2017). Survey of Privacy Preserving Techniques and Upcoming Techniques : A Review, *6*(2), 1798–1802.

Pawar, A., Ahirrao, S., & Churi, P. P. (n.d.). Anonymization Techniques for Protecting Privacy : A Survey.

Pinkas, B. (2002). Cryptographic Techniques for Privacy-Preserving Data Mining. *SIGKDD Explor. Newsl.*, *4*(2), 12–19. https://doi.org/10.1145/772862.772865

Qi, X., & Zong, M. (2012). An Overview of Privacy Preserving Data Mining. *Procedia Environmental Sciences*, *12*(Part B), 1341–1347. Retrieved from http://10.0.3.248/j.proenv.2012.01.432

Quirós, P. (1), Alonso, P. (1), Díaz, I. (2), & Montes, S. (3). (2015). Protecting data: a fuzzy approach. *International Journal of Computer Mathematics*, *92*(9), 1989–2000. https://doi.org/10.1080/00207160.2014.928700

SA, S. (2018). Big Data in Healthcare Management: A Review of Literature. *American Journal of Theoretical and Applied Business*. https://doi.org/10.11648/j.ajtab.20180402.14

Sachan, A., Roy, D., & Arun, P. V. (2013). An Analysis of Privacy Preservation Techniques in Data Mining BT - Advances in Computing and Information Technology. In N. Meghanathan, D. Nagamalai, & N. Chaki (Eds.) (pp. 119–128). Berlin, Heidelberg: Springer Berlin Heidelberg.

Sahu, H., Shrma, S., & Gondhalakar, S. (2008). A Brief Overview on Data Mining Survey. *Ijctee*.

Samarati, P., & Sweeney, L. (1998). Generalizing data to provide anonymity when disclosing information (abstract). In *PODS '98*.

Shanthi, A. S., & Karthikeyan, M. (2012). A review on privacy preserving data mining. *2012 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2012*. https://doi.org/10.1109/ICCIC.2012.6510302

Singh, A. (2019). Data Publishing Techniques andPrivacy Preserving. *International Journal for Information Security Research*, *9*(3), 1–23.

Smith, J. R., & Chang, S.-F. (1997). New visual information in the form of images Visually Searching the Web for Content, 12–20. https://doi.org/10.1080/10413200.2012.704621

Sweeney, L. (2002). K-Anonymity: A Model for Protecting Privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, *10*(5), 557–570. https://doi.org/10.1142/S0218488502001648

Taneja, S., Khanna, S., & Tilwalia, S. (2016). A Review on Privacy Preserving Data Mining: Techniques and Research Challenges.

Upadhyay, S., Sharma, C., Sharma, P., Bharadwaj, P., & Seeja, K. R. (2018). Privacy preserving data mining with 3-D rotation transformation. *Journal of King Saud University - Computer and Information Sciences*, *30*(4), 524–530. https://doi.org/10.1016/j.jksuci.2016.11.009

Vasa, J., & Modi, P. (2018). Review of Different Privacy Preserving Techniques in PPDP. *International Journal of Engineering Trends and Technology*. https://doi.org/10.14445/22315381/ijett-v59p242

Vassakis, K., Petrakis, E., & Kopanakis, I. (2018). Big Data Analytics: Applications, Prospects and Challenges (pp. 3–20). https://doi.org/10.1007/978-3-319-67925-9_1

Vasudevan, L., Sukanya, S. E. D., & Aarthi, N. (2008). Privacy preserving data mining using cryptographic role based access control approach. *Imecs 2008: International Multiconference of Engineers and Computer Scientists, Vols I and Ii*.

Xu, L., Jiang, C., Wang, J., Yuan, J., & Ren, Y. (2014). Information security in big data: Privacy and data mining. *IEEE Access*, *2*(January), 1151–1178. https://doi.org/10.1109/ACCESS.2014.2362522