DELOVNA SKUPINA ZA VARSTVO PODATKOV IZ ČLENA 29

0829/14/SL
WP216

Mnenje št. 5/2014 o anonimizacijskih tehnikah

Sprejeto 10. aprila 2014

DELOVNA SKUPINA ZA VARSTVO POSAMEZNIKOV PRI OBDELAVI OSEBNIH PODATKOV,

ustanovljena na podlagi Direktive Evropskega parlamenta in Sveta 95/46/ES z dne 24. oktobra 1995, je –

ob upoštevanju členov 29 in 30 Direktive,

ob upoštevanju svojega poslovnika –

SPREJELA NASLEDNJE MNENJE:

POVZETEK

Delovna skupina v tem mnenju analizira učinkovitost in omejitve sedanjih anonimizacijskih tehnik glede na pravni okvir EU za varstvo podatkov ter daje priporočila za obravnavanje teh tehnik ob upoštevanju preostalega tveganja za identifikacijo, ki spremlja vsako od njih.

Priznava morebitni pomen anonimizacije zlasti kot strategije za izkoriščanje prednosti „odprtih podatkov“ za posameznike in družbo na sploh, ki hkrati zmanjšuje tveganja za zadevne posameznike. Vendar so študije primerov in raziskovalne publikacije pokazale, kako težko je ustvariti resnično anonimen nabor podatkov ter hkrati ohraniti toliko osnovnih informacij, kot se zahtevajo za nalogo.

Anonimizacija je v skladu z Direktivo 95/46/ES in drugimi zadevnimi pravnimi akti EU rezultat obdelave osebnih podatkov za nepreklicno preprečitev identifikacije. Upravljavci podatkov bi morali pri tem upoštevati več elementov, med drugim vsa sredstva, „za katera se pričakuje“, da se bodo uporabila za določitev (s strani upravljavca ali katere koli tretje osebe).

Anonimizacija pomeni nadaljnjo obdelavo osebnih podatkov, zato mora izpolnjevati zahtevo glede skladnosti ob upoštevanju pravnih razlogov in okoliščin nadaljnje obdelave. Poleg tega anonimizirani podatki ne spadajo na področje uporabe zakonodaje o varstvu podatkov, vendar so lahko posamezniki, na katere se nanašajo osebni podatki, še vedno upravičeni do zaščite v skladu z drugimi določbami (kot so določbe o varstvu zaupnosti komunikacij).

V tem mnenju sta opisani glavni anonimizacijski tehniki, in sicer randomizacija in generalizacija. V njem so obravnavani zlasti dodajanje šuma, permutacija, diferencirana zasebnost, združevanje, k-anonimnost, l-raznolikost in t-podobnost. Za vsako tehniko so pojasnjeni načela, prednosti in slabosti ter pogoste napake, povezane z uporabo.

Zanesljivost posamezne tehnike je podrobneje opisana na podlagi treh meril:

(i) ali je še vedno mogoče izločiti posameznika;

(ii) ali je še vedno mogoče povezati zapise, ki se nanašajo na posameznika, in

(iii) ali je mogoče sklepati o informacijah, ki se nanašajo na posameznika.

Poznavanje glavnih prednosti in slabosti posameznih tehnik pomaga pri določitvi načina oblikovanja ustreznega anonimizacijskega postopka v danih okoliščinah.

Obravnavana je tudi psevdonimizacija, da se pojasnijo nekatere pasti in napačna prepričanja: psevdonimizacija ni anonimizacijska metoda. Z njo se samo zmanjša povezljivost nabora podatkov s prvotno identiteto posameznika, na katerega se nanašajo osebni podatki, zato je to uporaben varnostni ukrep.

V mnenju se ugotavlja, da lahko anonimizacijske tehnike dajo zagotovila glede zasebnosti in da jih je mogoče uporabljati za oblikovanje učinkovitih anonimizacijskih postopkov, vendar samo če je njihova uporaba ustrezno zasnovana – kar pomeni, da morajo biti pogoji (okoliščine) in cilj(-i) anonimizacijskega postopka jasno določeni, da se doseže ciljno usmerjena anonimizacija in hkrati zagotovijo nekateri uporabni podatki. Najboljšo rešitev bi bilo treba določiti za vsak primer posebej, po možnosti z združitvijo različnih tehnik, ob upoštevanju praktičnih priporočil iz tega mnenja.

Nazadnje, upravljavci podatkov bi morali upoštevati, da lahko anonimiziran nabor podatkov za posameznike, na katere se nanašajo osebni podatki, še vedno pomeni preostala tveganja. Anonimizacija in ponovna identifikacija sta po eni strani dejansko dejavni področji raziskav in nova odkritja se redno objavljajo, po drugi strani pa je mogoče celo anonimizirane podatke, kot so statistični podatki, uporabiti za obogatitev sedanjih profilov posameznikov, zaradi česar se postavljajo nova vprašanja v zvezi z varstvom podatkov. Anonimizacije zato ne bi smeli obravnavati kot enkratnega postopka, upravljavci podatkov pa bi morali redno ocenjevati spremljajoča tveganja.

1. Uvod

Medtem ko naprave, senzorji in omrežja ustvarjajo velike količine in nove vrste podatkov, strošek shranjevanja podatkov pa postaja zanemarljiv, so zanimanje in zahteve javnosti za ponovno uporabo teh podatkov vse večje. „Odprti podatki“ lahko zagotavljajo jasne koristi za družbo, posameznike in organizacije, vendar samo če se spoštujejo pravice vseh do varstva njihovih osebnih podatkov in zasebnega življenja.

Anonimizacija je morda dobra strategija za ohranitev koristi in zmanjšanje tveganj. Potem ko je nabor podatkov dejansko anonimiziran in posamezniki niso več določljivi, se evropska zakonodajo o varstvu podatkov ne uporablja več. Vendar so študije primerov in raziskovalne publikacije jasno pokazale, da predlog, naj se iz bogatega nabora osebnih podatkov ustvari resnično anonimen nabor podatkov ter hkrati ohrani toliko osnovnih informacij, kot se zahtevajo za nalogo, ni preprost. Na primer, nabor podatkov, ki se šteje za anonimnega, je mogoče združiti z drugim naborom podatkov, tako da je mogoče določiti enega ali več posameznikov.

Delovna skupina v tem mnenju analizira učinkovitost in omejitve sedanjih anonimizacijskih tehnik glede na pravni okvir EU za varstvo podatkov ter daje priporočila za previdno in odgovorno uporabo teh tehnik za oblikovanje anonimizacijskega postopka.

2. Opredelitev pojmov in pravna analiza

2.1 Opredelitev pojmov v pravnem okviru EU

Direktiva 95/46/ES v uvodni izjavi 26 navaja anonimizacijo, s katero se anonimizirani podatki izključijo iz področja uporabe zakonodaje o varstvu podatkov:

> *„ker se morajo načela varstva uporabljati za vse informacije v zvezi z določeno ali določljivo osebo; ker bi bilo treba za odločitev o tem, ali je oseba določljiva ali ne, upoštevati vsa sredstva, za katera se pričakuje, da jih bo uporabil bodisi upravljavec ali katera koli druga oseba za določitev take osebe; ker se načela varstva ne uporabljajo za podatke, ki so spremenjeni v anonimne tako, da posameznik, na katerega se osebni podatki nanašajo, ni več določljiv; ker so pravila ravnanja v smislu člena 27 lahko koristen instrument za usmerjanje k načinom, s katerimi se lahko podatki spremenijo v anonimne in se ohranijo v obliki, v kateri identifikacija posameznika, na katerega se osebni podatki nanašajo, ni več mogoča;“.*[1]

Natančno branje uvodne izjave 26 daje konceptualno opredelitev anonimizacije. Uvodna izjava 26 kaže, da je treba za anonimizacijo katerega koli podatka temu odvzeti dovolj elementov, tako da posameznika, na katerega se nanašajo osebni podatki, ni več mogoče določiti. Natančneje, podatki se morajo obdelati tako, da jih ni mogoče več uporabiti za določitev fizične osebe z uporabo „vseh sredstev, za katera se pričakuje, da jih bo uporabil“ bodisi upravljavec bodisi tretja oseba. Pomembno je, da mora biti obdelava podatkov nepreklicna. V Direktivi ni pojasnjeno, kako naj se takšen postopek deidentifikacije izvede ali

[1] Poleg tega je treba opozoriti, da je ta pristop uporabljen tudi v uvodni izjavi 23 osnutka uredbe EU o varstvu podatkov „za odločitev, ali je oseba določljiva, je treba upoštevati vsa sredstva, za katera se pričakuje, da jih bo za določitev posameznika uporabil bodisi upravljavec bodisi katera koli druga oseba“.

kako bi ga bilo mogoče izvesti[2]. Poudarek je na rezultatu: podatki morajo biti takšni, da posameznika, na katerega se nanašajo osebni podatki, ni mogoče določiti z „vsemi“ „verjetnimi“ in „razumnimi“ sredstvi. Uporabi se sklicevanje na pravila ravnanja kot na orodje za določitev mogočih anonimizacijskih mehanizmov ter ohranitev v obliki, v kateri identifikacija posameznika, na katerega se nanašajo osebni podatki, „ni več mogoča“. Direktiva tako jasno določa zelo visok standard.

Tudi v direktivi o e-zasebnosti (Direktiva 2002/58/ES) se „anonimizacija“ in „anonimni podatki“ obravnavajo zelo podobno. V uvodni izjavi 26 je navedeno:

> *„Podatke o prometu, uporabljene za trženje komunikacijskih storitev ali za zagotovitev storitev z dodano vrednostjo, je po opravljeni storitvi treba izbrisati ali napraviti anonimne.“*

Skladno s tem člen 6(1) določa, da:

> *„Podatki o prometu, ki se nanašajo na naročnike in uporabnike in ki jih je ponudnik javnega komunikacijskega omrežja ali javno razpoložljive elektronske komunikacijske storitve obdelal in shranil, morajo biti izbrisani ali predelani v anonimne, potem ko niso več potrebni za namen prenosa sporočila, kar ne vpliva na odstavke 2, 3 in 5 tega člena in člena 15(1).“*

Poleg tega člen 9(1) določa:

> *„Kadar se podatki o lokaciji, razen podatkov o prometu, ki se nanašajo na uporabnike ali naročnike javnih komunikacijskih omrežij ali javno razpoložljivih elektronskih komunikacijskih storitev, dajo obdelovati, se smejo takšni podatki obdelati šele potem, ko postanejo anonimni ali s privolitvijo uporabnikov ali naročnikov in to v obsegu in trajanju, ki sta potrebna za izvedbo storitve z dodano vrednostjo.“*

Temeljno načelo je, da bi moral biti rezultat anonimizacije kot tehnike, ki se uporabi za osebne podatke, glede na trenutno stanje tehnologije tako trajen, kot je izbris, tj. da ne omogoča obdelave osebnih podatkov.[3]

2.2 Pravna analiza

Na podlagi analize besedila o anonimizaciji v glavnih aktih EU o varstvu podatkov je mogoče opozoriti na štiri ključne značilnosti:

– anonimizacija je lahko rezultat obdelave osebnih podatkov, katere cilj je nepreklicno preprečiti določitev posameznika, na katerega se nanašajo osebni podatki;

– lahko se predvidi več anonimizacijskih tehnik, zakonodaja EU pa ne vsebuje predpisanega standarda;

---

[2] Ta koncept je podrobneje obravnavan na strani 8 tega mnenja.

[3] Pri tem je treba opozoriti, da je anonimizacija opredeljena tudi v mednarodnih standardih, kot je ISO 29100, in sicer kot „postopek, s katerim se osebni identifikacijski podatki (OIP) nepreklicno spremenijo tako, da upravljavec OIP sam ali v sodelovanju s katero koli drugo osebo ne more neposredno ali posredno določiti načela OIP“ (ISO 29100: 2011). Tudi za ISO je ključna nepreklicnost spremembe osebnih podatkov, ki omogoča neposredno ali posredno določitev posameznika. Iz tega vidika obstaja precejšnja podobnost z načeli in koncepti, na katerih temelji Direktiva 95/46/ES. To velja tudi za opredelitve iz nekaterih nacionalnih zakonodaj (npr. v Italiji, Nemčiji in Sloveniji), v katerih je poudarek na nezmožnosti določitve posameznika, navajajo pa se tudi „nesorazmerna prizadevanja“, potrebna za ponovno določitev (Nemčija, Slovenija). Vendar francoski zakon o varstvu podatkov določa, da podatki ostanejo osebni podatki, tudi če je izjemno težko in malo verjetno znova določiti posameznika, na katerega se nanašajo – tj. da ni določbe o preskusu „razumnosti“.

– pomembne so okoliščine: upoštevati je treba „vsa“ sredstva, „za katera se pričakuje“, da jih bodo za določitev uporabili upravljavec in tretje osebe, pri čemer je treba posebno pozornost nameniti vprašanju, kaj v zadnjem času „v trenutnem stanju tehnologije“ pomeni zveza „za katera se pričakuje“ (glede na povečanje računalniške zmogljivosti in razpoložljivih orodij);

– z anonimizacijo je neločljivo povezan dejavnik tveganja: ta dejavnik tveganja je treba upoštevati pri oceni veljavnosti vsake anonimizacijske tehnike, vključno z mogočimi uporabami podatkov, ki se „anonimizirajo“ s takšno tehniko, prav tako pa bi bilo treba oceniti stopnjo in verjetnost tega tveganja.

V tem mnenju se uporablja izraz „anonimizacijska tehnika“ in ne „anonimnost“ ali „anonimni podatki“, da se poudari preostalo implicitno tveganje za ponovno identifikacijo, ki je povezano s katerim koli tehnično-organizacijskim ukrepom, katerega cilj je ohraniti podatke „anonimne“.

2.2.1 Zakonitost anonimizacijskega postopka

Prvič, anonimizacija je tehnika, ki se uporablja za osebne podatke, da se doseže nepreklicna deidentifikacija. Zato je treba izhajati iz predpostavke, da je treba osebne podatke zbrati in obdelati v skladu z veljavno zakonodajo o ohranitvi podatkov v določljivi obliki.

V tem okviru je anonimizacijski postopek, ki pomeni obdelavo takšnih osebnih podatkov za zagotovitev njihove anonimizacije, primer „nadaljnje obdelave“. Ta obdelava mora zato izpolnjevati zahteve preskusa skladnosti v skladu s smernicami, ki jih je delovna skupina zagotovila v svojem mnenju št. 3/2013 o omejitvi namena.[4].

To pomeni, da je mogoče pravno podlago za anonimizacijo načeloma najti v katerem koli razlogu iz člena 7 (vključno z zakonitim interesom upravljavca podatkov), če so izpolnjene tudi zahteve glede kakovosti podatkov iz člena 6 Direktive ter ob ustreznem upoštevanju posebnih okoliščin in vseh dejavnikov iz mnenja delovne skupine o omejitvi namena[5].

Po drugi strani pa bi bilo treba opozoriti na določbe iz člena 6(1)(e) Direktive95/46/ES (ter členov 6(1) in 9(1) direktive o e-zasebnosti), ker kažejo na potrebo po shranjevanju osebnih podatkov „v obliki, ki dopušča identifikacijo“ le toliko časa, kolikor je potrebno za namene, za katere so bili zbrani ali za katere se obdelujejo naprej.

Ta določba sama po sebi odločno poudarja, da bi morali biti osebni podatki anonimizirani že „privzeto“ (ob upoštevanju različnih pravnih zahtev, kot so zahteve za podatke o prometu iz direktive o e-zasebnosti). Če želi upravljavec podatkov ohraniti takšne osebne podatke, potem ko so bili nameni prvotne ali nadaljnje obdelave doseženi, bi bilo treba anonimizacijske tehnike uporabiti tako, da se nepreklicno prepreči identifikacija.

[4] Mnenje št. 3/2013 Delovne skupine iz člena 29 je na voljo na: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf.

[5] To zlasti pomeni, da je treba opraviti vsebinsko oceno ob upoštevanju vseh zadevnih okoliščin, zlasti naslednjih ključnih dejavnikov:
(a) povezave med nameni, za katere so bili osebni podatki zbrani, in nameni nadaljnje obdelave;
(b) okoliščin, v katerih so bili osebni podatki zbrani, in razumnih pričakovanj posameznikov, na katere se nanašajo, glede njihove nadaljnje uporabe;
(c) narave osebnih podatkov in učinka nadaljnje obdelave na posameznike, na katere se nanašajo;
(d) zaščitnih ukrepov, ki jih upravljavec sprejme za zagotovitev poštene obdelave in preprečitev neupravičenega učinka na posameznike, na katere se nanašajo osebni podatki.

Zato delovna skupina meni, da je mogoče anonimizacijo kot primer nadaljnje obdelave osebnih podatkov šteti za združljivo s prvotnimi nameni obdelave, vendar le če je anonimizacijski postopek takšen, da zanesljivo zagotovi anonimizirane informacije v smislu, kot je opisan v tem dokumentu.

Poudariti je treba tudi, da mora biti anonimizacija opravljena v skladu s pravnimi omejitvami, na katere je Sodišče opozorilo v svoji odločitvi v zadevi C-553/07 (*College van burgemeester en wethouders van Rotterdam proti M. E. E. Rijkeboer*) in ki se nanašajo na potrebo po ohranitvi podatkov v določljivi obliki, da se na primer posameznikom, na katere se nanašajo osebni podatki, omogoči uveljavljanje pravic do dostopa. Sodišče Evropske unije je razsodilo, da „*morajo države članice v skladu s členom 12(a) Direktive [95/46/ES] pravico do dostopa do informacij glede prejemnikov ali vrste prejemnikov podatkov in glede vsebine posredovanih informacij določiti ne le za sedanjost, ampak tudi za preteklost. Države članice morajo določiti rok za shranjevanje teh informacij in temu ustrezno dostop do njih, pri čemer skrbijo za pravično ravnovesje med, po eni strani, interesom posameznika, na katerega se osebni podatki nanašajo, za varstvo njegove zasebnosti zlasti s pomočjo pravic in pravnih sredstev, ki jih določa Direktiva, ter, po drugi strani, bremenom, ki ga za upravljavca pomeni obveznost shranjevanja teh informacij.*“

To velja zlasti, če se upravljavec podatkov pri anonimizaciji opre na člen 7(f) Direktive 95/46/ES: zakoniti interes upravljavca podatkov mora biti vedno v ravnovesju s pravicami in temeljnimi svoboščinami posameznikov, na katere se nanašajo osebni podatki.

Na primer, preiskava, ki jo je v letih 2012 in 2013 opravil nizozemski organ za varstvo podatkov v zvezi s tehnologijami za temeljiti pregled paketov, ki jih uporabljajo štirje mobilni operaterji, je pokazala pravno podlago v skladu s členom 7(f) Direktive 95/46/ES za anonimizacijo vsebine podatkov o prometu čim prej po njihovi pridobitvi. Člen 6 direktive o e-zasebnosti dejansko določa, da morajo biti podatki o prometu, ki se nanašajo na naročnike in uporabnike in ki jih je ponudnik javnega komunikacijskega omrežja ali javno razpoložljive elektronske komunikacijske storitve obdelal in shranil, izbrisani ali predelani v anonimne, kakor hitro je to mogoče. Ker je to dovoljeno v skladu s členom 6 direktive o e-zasebnosti, obstaja v tem primeru ustrezna pravna podlaga v členu 7 direktive o varstvu podatkov. To bi bilo mogoče predstaviti tudi obratno: če način obdelave podatkov ni dovoljen v skladu s členom 6 direktive o e-zasebnosti, ne more biti pravne podlage v členu 7 direktive o varstvu podatkov.

2.2.2 Morebitna določljivost anonimiziranih podatkov

Delovna skupina je podrobno obravnavala pojem osebnih podatkov v mnenju št. 4/2007 o osebnih podatkih, ki je osredotočeno na osnovne elemente opredelitve iz člena 2(a) Direktive 95/46/ES, vključno z „določenim ali določljivim“ delom takšne opredelitve. V zvezi s tem je tudi ugotovila, da so „anonimizirani podatki […] potemtakem anonimni podatki, ki so se pred tem nanašali na določljivo osebo, vendar identifikacija ni več mogoča“.

Delovna skupina je tako že pojasnila, da je v Direktivi predlog za uporabo preskusa „sredstev […], za katera se pričakuje, da jih bo uporabil“ kot merila, ki se uporabi za oceno, ali je anonimizacijski postopek dovolj zanesljiv, tj. ali je postala identifikacija „razumno“ nemogoča. Na določljivost neposredno vplivajo posebne okoliščine določenega primera. V tehnični prilogi k temu mnenju je opravljena analiza učinka izbire najustreznejše tehnike.

Kot je bilo že poudarjeno, raziskave še potekajo, orodja in računalniška zmogljivost pa se razvijajo. Zato ni niti mogoče niti koristno zagotoviti izčrpnega seznama okoliščin, v katerih

identifikacija ni več mogoča. Vendar si nekateri ključni dejavniki zaslužijo, da se proučijo in ponazorijo.

Prvič, trditi je mogoče, da bi se morali upravljavci podatkov osredotočiti na konkretna sredstva, ki bi bila potrebna za izvedbo postopka anonimizacije v obratni smeri, zlasti ob upoštevanju stroškov ter strokovnega znanja in izkušenj, potrebnih za izvedbo navedenih sredstev ter oceno njihove verjetnosti in težavnosti. Na primer, svoja prizadevanja in stroške za anonimizacijo (v smislu potrebnega časa in virov) bi morali uskladiti z vse bolj cenovno ugodno razpoložljivostjo tehničnih sredstev za določitev posameznikov v naborih podatkov, vse večjo javno dostopnostjo drugih naborov podatkov (kot so podatki, ki so dani na voljo v zvezi s politikami „odprtih podatkov“) in številnimi primeri nepopolne anonimizacije, ki ima naknadne škodljive, včasih nepopravljive posledice za posameznike, na katere se nanašajo osebni podatki.[6] Poudariti je treba, da se lahko identifikacijsko tveganje sčasoma poveča, odvisno pa je tudi od razvoja informacijske in komunikacijske tehnologije. Pravni predpisi, če obstajajo, morajo biti zato oblikovani tehnološko nevtralno, najbolje ob upoštevanju sprememb razvojnih možnosti informacijske tehnologije.[7]

Drugič, „sredstva, za katera se pričakuje, da se bodo uporabila za določitev, ali je oseba določljiva“ so sredstva, ki jih bo uporabil „bodisi upravljavec bodisi katera koli druga oseba“. Zato je treba nujno razumeti, da če upravljavec podatkov ne izbriše prvotnih (določljivih) podatkov na ravni dogodka in izroči del tega nabora podatkov (na primer po odstranitvi ali zakritju določljivih podatkov), nastali nabor podatkov še vedno pomeni osebne podatke. Nastali nabor podatkov bi bilo mogoče opredeliti kot anonimen samo, če bi upravljavec podatkov te podatke združil tako, da posamezni dogodki ne bi bili več določljivi. Na primer, če organizacija zbira podatke o potovalnih premikih posameznikov, bi potovalni vzorci posameznikov na ravni dogodka še vedno izpolnjevali pogoje za osebne podatke za katero koli osebo, vse dokler bi imel upravljavec podatkov (ali katera koli druga oseba) še vedno dostop do prvotnih neobdelanih podatkov, tudi če so bili neposredni identifikatorji odstranjeni iz nabora podatkov, predloženega tretjim osebam. Če pa bi upravljavec podatkov izbrisal neobdelane podatke in tretjim osebam predložil samo zbirne statistične podatke na visoki ravni, kot na primer „ob ponedeljkih je na poti X 160 % več potnikov kot ob torkih“, bi se ti podatki šteli za anonimne.

Učinkovita anonimizacijska rešitev vsem osebam preprečuje izločitev posameznika v naboru podatkov, povezavo dveh zapisov v naboru podatkov (ali med dvema ločenima naboroma podatkov) in sklepanje v zvezi s katero koli informacijo v takšnem naboru podatkov. Odstranitev elementov, ki omogočajo neposredno določitev, sama po sebi na splošno torej ne zadostuje za zagotovitev, da določitev posameznika, na katerega se nanašajo osebni podatki, ni več mogoča. Pogosto je treba sprejeti dodatne ukrepe za preprečitev določitve, ki so znova odvisni od okoliščin in namenov obdelave, za katero so anonimizirani podatki namenjeni.

[6] Zanimivo je, da je v spremembah osnutka splošne uredbe o varstvu podatkov, kot jih je nedavno (21. oktobra 2013) predložil Evropski parlament, v uvodni izjavi 23 posebej navedeno: „Da bi ugotovili, ali se za ta sredstva pričakuje, da bodo uporabljena za določitev posameznika, bi bilo treba upoštevati vse objektivne dejavnike, kot so stroški določitve in čas, potreben zanjo, pri čemer se upoštevata razpoložljiva tehnologija v času obdelave in tehnološki razvoj.“

[7] Glej Mnenje 4/2007 Delovne skupine iz člena 29, str. 15.

PRIMER:
Podatki o genetskih profilih so primer osebnih podatkov, za katere lahko zaradi edinstvene narave nekaterih profilov obstaja tveganje določitve, če se uporabi samo tehnika odstranitve identitete darovalca. V literaturi je bilo že prikazano[8] , da lahko kombinacija javno dostopnih genskih virov (npr. rodoslovnih registrov, osmrtnic, rezultatov poizvedb v iskalnikih) in metapodatkov o darovalcih DNK (čas darovanja, starost in kraj stalnega prebivališča) razkrije identiteto nekaterih posameznikov, čeprav je bila navedena DNK darovana „anonimno“.

Obe skupini anonimizacijskih tehnik – randomizacija in generalizacija –[9] imata slabosti; vendar je lahko vsaka od njiju ustrezna v določenih okoliščinah, da se doseže želeni namen brez ogrožanja zasebnosti posameznikov, na katere se nanašajo osebni podatki. Jasno mora biti, da „identifikacija“ ne pomeni samo možnosti pridobitve imena in/ali naslova osebe, temveč vključuje tudi morebitno določljivost z izločitvijo, povezljivostjo in sklepanjem. Poleg tega za veljavno zakonodajo o varstvu podatkov ni pomembno, kakšni so nameni upravljavca podatkov ali prejemnika. Dokler so podatki določljivi, se uporabljajo pravila o varstvu podatkov.

Če tretja oseba obdeluje nabor podatkov, ki je bil obravnavan z anonimizacijsko tehniko (ki jih je anonimiziral in objavil prvotni upravljavec podatkov), lahko to dela zakonito in ji ni treba upoštevati zahtev za varstvo podatkov, če v prvotnem naboru podatkov ne more (neposredno ali posredno) določiti posameznikov, na katere se nanašajo osebni podatki. Vendar morajo tretje osebe vse zgoraj navedene dejavnike v zvezi z okoliščinami (vključno s posebnimi značilnostmi anonimizacijskih tehnik, ki jih je uporabil prvotni upravljavec podatkov) upoštevati pri odločitvi, kako uporabiti in zlasti združiti takšne anonimizirane podatke za svoje namene – ker lahko posledice, ki izhajajo iz tega, vključujejo različne vrste odgovornosti na njihovi strani. Če so navedeni dejavniki in značilnosti takšni, da pomenijo nesprejemljivo tveganje za določitev posameznikov, na katere se nanašajo osebni podatki, obdelava podatkov znova spada na področje uporabe zakonodaje o varstvu podatkov.

Navedeni seznam nikakor ni izčrpen, temveč zagotavlja splošne smernice o pristopu k ocenjevanju morebitne določljivosti določenega nabora podatkov, ki je bil anonimiziran v skladu z različnimi razpoložljivimi tehnikami. Za vse navedene dejavnike je mogoče upoštevati, da jih tako kot številne dejavnike tveganja ocenijo upravljavci podatkov v naborih podatkov, ki se anonimizirajo, in tretje osebe pri uporabi navedenih „anonimiziranih“ naborov podatkov za svoje namene.

2.2.3 Tveganja, povezana z uporabo anonimiziranih podatkov

Upravljavci podatkov morajo pri proučitvi uporabe anonimizacijskih tehnik upoštevati naslednja tveganja:

– Posebno past pomeni obravnavanje psevdonimiziranih podatkov kot enakovrednih anonimiziranim podatkom. V oddelku Tehnična analiza je pojasnjeno, da psevdonimiziranih podatkov ni mogoče enačiti z anonimiziranimi informacijami, ker še naprej omogočajo izločitev in povezavo posameznika, na katerega se nanašajo osebni podatki, z različnimi nabori podatkov. Za uporabo psevdonimov je verjetno, da omogoča določljivost, zato ostaja

[8] Glej John Bohannon, Genealogy Databases Enable Naming of Anonymous DNA Donors (Rodoslovne podatkovne zbirke omogočajo določitev anonimnih darovalcev DNK), Znanost, zv. 339, št. 6117. (18. januar 2013), str. 262.

[9] Glavne značilnosti in razlike teh dveh anonimizacijskih tehnik so opisane v oddelku 3 v nadaljevanju („Tehnična analiza“).

na področju uporabe pravnega sistema za varstvo podatkov. To je zlasti pomembno v okviru znanstvenih in statističnih raziskav ali raziskav preteklega stanja.[10]

PRIMER:
Značilen primer napačnih predstav v zvezi s psevdonimizacijo je znani incident družbe AOL (America On Line). Leta 2006 je bila objavljena podatkovna zbirka z več kot dvajsetimi milijoni ključnih besed za iskanje za več kot 650 000 uporabnikov v treh mesecih, pri čemer je bil edini ukrep za ohranitev zasebnosti nadomestitev identifikatorja uporabnika AOL s številčno oznako. To je povzročilo javno identifikacijo in lociranje nekaterih od njih. Psevdonimizirani iskalni niz iskalnika, zlasti če je združen z drugimi atributi, kot so naslovi IP ali drugi konfiguracijski parametri naročnika, ima zelo veliko zmogljivost identifikacije.

– Druga napaka je mnenje, da ustrezno anonimizirani podatki (ki izpolnjujejo vse navedene pogoje in merila ter po opredelitvi ne spadajo na področje uporabe direktive o varstvu podatkov) posameznikom odvzamejo kakršne koli zaščitne ukrepe – predvsem zato, ker se lahko za te podatke uporabljajo drugi zakoni. Na primer, člen 5(3) direktive o e-zasebnosti preprečuje shranjevanje vseh „informacij“ in dostop do njih (vključno z neosebnimi informacijami) na terminalski opremi brez soglasja naročnika ali uporabnika, ker je to del splošnejšega načela zaupnosti komunikacij.

– Tretja malomarnost prav tako izhaja iz neupoštevanja učinka ustrezno anonimiziranih podatkov na posameznike v nekaterih okoliščinah, zlasti pri oblikovanju profilov. Področje posameznikovega zasebnega življenja je zaščiteno s členom 8 EKČP in členom 7 Listine EU o temeljnih pravicah; čeprav se zakonodaja o varstvu podatkov morda ne uporablja več za takšne podatke, lahko uporaba anonimiziranih naborov podatkov, ki se predložijo v uporabo tretjim osebam, kot taka povzroči izgubo zasebnosti. Posebna previdnost je potrebna pri ravnanju z anonimiziranimi informacijami zlasti, če se takšne informacije uporabijo (pogosto skupaj z drugimi podatki) za sprejetje odločitev, ki vplivajo (čeprav posredno) na posameznike. Kot je bilo v tem mnenju že poudarjeno in kot je delovna skupina pojasnila zlasti v mnenju o pojmu „omejitev namena“ (Mnenje št. 3/2013)[11], bi bilo treba upravičena pričakovanja posameznikov, na katere se nanašajo osebni podatki, v zvezi z nadaljnjo obdelavo njihovih podatkov oceniti ob upoštevanju ustreznih dejavnikov, ki so povezani z okoliščinami – kot so narava odnosa med posamezniki, na katere se nanašajo osebni podatki, in upravljavci podatkov, veljavne zakonske obveznosti in preglednost postopkov obdelave.

3. Tehnična analiza, zanesljivost tehnologij in značilne napake

Obstajajo različne anonimizacijske prakse in tehnike z različnimi stopnjami zanesljivosti. V tem oddelku so obravnavane glavne točke, ki jih morajo upravljavci podatkov proučiti v zvezi z njihovo uporabo in pri tem zlasti upoštevati zagotovilo, ki ga je mogoče doseči z določeno tehniko glede na trenutno stanje tehnologije in ob upoštevanju treh bistvenih tveganj anonimizacije:

- *izločitve,* ki ustreza možnosti osamitve nekaterih ali vseh zapisov, ki določajo posameznika v naboru podatkov;
- *povezljivosti*, ki je možnost povezave vsaj dveh zapisov, ki se nanašata na istega posameznika, na katerega se nanašajo osebni podatki, ali skupino takšnih posameznikov (v isti podatkovni zbirki ali dveh različnih podatkovnih zbirkah). Če lahko napadalec (na primer s korelacijsko analizo) določi, da sta dva zapisa

[10] Glej tudi Mnenje 4/2007 Delovne skupine iz člena 29, str. 18–20.
[11] Na voljo na spletnem naslovu http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

dodeljena isti skupini posameznikov, ne more pa izločiti posameznikov v tej skupini, tehnika zagotavlja zaščito pred „izločitvijo“, ne pa tudi pred povezljivostjo;
- *sklepanja*, ki je možnost, da se z veliko verjetnostjo določi vrednost atributa na podlagi vrednosti skupine drugih atributov.

Rešitev za ta tri tveganja bi bila torej zaščita pred ponovno identifikacijo, ki se izvede z najverjetnejšimi in razumnimi sredstvi, ki jih lahko uporabita upravljavec podatkov in katera koli tretja oseba. Delovna skupina v zvezi s tem poudarja, da se tehnike za deidentifikacijo in anonimizacijo še raziskujejo ter da takšne raziskave dosledno kažejo, da nobena tehnika sama po sebi ni brez pomanjkljivosti. Na splošno obstajata dva različna pristopa k anonimizaciji: prvi temelji na ***randomizaciji***, drugi pa na ***generalizaciji***. V mnenju so obravnavani tudi drugi koncepti, kot so *psevdonimizacija, diferencirana zasebnost, l-raznolikost* in *t-podobnost*.

V tem delu mnenja so uporabljeni naslednji izrazi: nabor podatkov sestavljajo različni zapisi, ki se nanašajo na posameznike (posameznike, na katere se nanašajo osebni podatki). Vsak zapis se nanaša na enega posameznika, na katerega se nanašajo osebni podatki, in je sestavljen iz nabora vrednosti (ali „vnosov“, npr. 2013) za vsak atribut (npr. leto). Nabor podatkov je zbirka zapisov, ki se lahko oblikujejo tudi kot preglednica (ali sklop preglednic) ali kot graf z opombami/uteženi graf, kar je danes vse pogosteje. Primeri iz mnenja se nanašajo na preglednice, lahko pa se uporabijo tudi za druge grafične predstavitve zapisov. Kombinacije atributov, ki se nanašajo na posameznika ali skupino posameznikov, na katere se nanašajo osebni podatki, se lahko imenujejo kvaziidentifikatorji. Nabor podatkov lahko ima v nekaterih primerih več zapisov o istem posamezniku. „Napadalec“ je tretja oseba (ki ni upravljavec ali obdelovalec podatkov), ki po naključju ali namerno dostopa do prvotnih zapisov.

3.1 Randomizacija

Randomizacija je skupina tehnik, s katero se spremeni verodostojnost podatkov za odpravo močne povezave med podatki in posameznikom. Če so podatki dovolj negotovi, se ne morejo več nanašati na določenega posameznika. Edinstvenost posameznega zapisa se ne zmanjša samo z randomizacijo, ker vsak zapis še vedno izhaja iz enega posameznika, na katerega se nanašajo osebni podatki, vendar randomizacija zagotavlja zaščito pred napadi s sklepanjem/tveganji glede sklepanja in jo je mogoče združiti z generalizacijskimi tehnikami, da se zagotovijo trdnejša zagotovila glede zasebnosti. Za zagotovitev, da z zapisom ni mogoče določiti samo enega posameznika, se lahko zahtevajo dodatne tehnike.

3.1.1 Dodajanje šuma

Tehnika dodajanja šuma je uporabna zlasti, kadar lahko atributi precej škodljivo vplivajo na posameznike, sestavljajo pa jo spremenljivi atributi v naboru podatkov, ki tako postanejo manj točni, pri čemer se ohrani celotna porazdelitev. Opazovalec pri obdelavi nabora podatkov predpostavlja, da so vrednosti točne, vendar to velja samo do določene mere. Na primer, če je bila višina posameznika prvotno izmerjena na centimeter natančno, lahko anonimiziran nabor podatkov vsebuje višino z natančnostjo +/–10 cm. Če se ta tehnika uporablja učinkovito, tretja oseba ne more niti določiti posameznika niti popraviti podatkov ali drugače ugotoviti, kako so bili podatki spremenjeni.

Dodajanje šuma je treba po navadi združiti z drugimi anonimizacijskimi tehnikami, kot je odstranitev očitnih atributov in kvaziidentifikatorjev. Raven šuma bi morala biti odvisna od

potrebne ravni zahtevanih informacij in učinka razkritja zaščitenih atributov na zasebnost posameznikov.

3.1.1.1. Zagotovila

- *Izločitev:* Čeprav so zapisi manj zanesljivi, je še vedno mogoče izločiti zapise o posamezniku (morda nedoločljivo).
- *Povezljivost:* Zapise o istem posamezniku je še vedno mogoče povezati, vendar so ti manj zanesljivi, zato je mogoče pravi zapis povezati z umetno dodanim (tj. s „šumom“). Napačni atribut lahko v nekaterih primerih izpostavi posameznika, na katerega se nanašajo osebni podatki, znatnemu in celo večjemu tveganju kot pravilni.
- *Sklepanje:* Mogoči so napadi s sklepanjem, ki pa so manj uspešni, in nekateri napačni pozitivni rezultati (in napačni negativni rezultati).

3.1.1.2. Pogoste napake

- Dodajanje nesorazmernega šuma: Če šum semantično ne more obstati (tj. če je „nesorazmeren“ in ne upošteva logike med atributi v naboru), ga lahko napadalec, ki ima dostop do podatkovne zbirke, odstrani in v nekaterih primerih obnovi manjkajoče vnose. Poleg tega, če je nabor podatkov preveč razpršen[12], je še vedno mogoče vnesene podatke, opremljene s šumom, povezati z zunanjim virom.
- Predpostavka, da zadostuje dodajanje šuma: dodajanje šuma je dopolnilni ukrep, ki napadalcu oteži pridobitev osebnih podatkov. Če šum ni močnejši od informacij v naboru podatkov, se ne bi smelo predpostavljati, da dodajanje šuma pomeni samostojno rešitev za anonimizacijo.

3.1.1.3. Neuspešno dodajanje šuma

Zelo znan poskus ponovne identifikacije je bil opravljen v zbirki podatkov o kupcih ponudnika video vsebin Netflix. Raziskovalci so analizirani geometrične lastnosti navedene podatkovne zbirke, sestavljene iz več kot 100 milijonov ocen na lestvici od 1 do 5 za več kot 18 000 filmov, ki jih je prispevalo več kot 500 000 uporabnikov, družba pa jih je objavila po „anonimizaciji“ v skladu z notranjo politiko o zasebnosti, pri čemer so bile odstranjene vse informacije, ki so se nanašale na kupce, razen ocen in datumov. Dodan je bil šum, saj so bile ocene nekoliko višje ali nižje.

Kljub temu je bilo ugotovljeno, da je bilo mogoče 99 % zapisov o uporabnikih v naboru podatkov nedvomno določiti z osmimi ocenami in datumi s 14-dnevnimi napakami kot merilom za izbiro, medtem ko je znižanje merila za izbiro (dve oceni in tridnevna napaka) še vedno omogočalo določitev 68 % uporabnikov.[13]

3.1.2. Permutacija

Ta tehnika, ki vključuje premešane vrednosti atributov v preglednici, da se nekateri od njih umetno povežejo z različnimi posamezniki, na katere se nanašajo osebni podatki, se uporablja, kadar je treba ohraniti točno porazdelitev posameznega atributa v naboru podatkov.

[12] Ta koncept je podrobneje obravnavan v Prilogi na strani 30.

[13] Narayanan, A., in Shmatikov, V. (maj 2008). Robust de-anonymization of large sparse datasets (Zanesljiva deanonimizacija obsežnih razpršenih naborov podatkov). V *Varnost in zasebnost*, 2008. *SP 2008. Simpozij IEEE o* (str. 111–125). IEEE.

Permutacijo je mogoče šteti za posebno obliko dodajanja šuma. Pri običajni šumni tehniki se atributi spremenijo z randomiziranimi vrednostmi. Ustvarjanje sorazmernega šuma je lahko zahtevna naloga, rahlo spreminjanje vrednosti atributov pa morda ne bo zagotovilo zadostne zasebnosti. Druga možnost je, da se s permutacijskimi tehnikami spremenijo vrednosti v naboru podatkov samo z njihovo zamenjavo med dvema zapisoma. Takšna zamenjava zagotovi, da ostaneta obseg in porazdelitev vrednosti enaka, kar pa ne velja za korelacije med vrednostmi in posamezniki. Če med dvema ali več atributi obstaja logična ali statistična povezava in so bili atributi permutirani neodvisno, se takšna povezava poruši. Zato je morda pomembno, da se niz podobnih atributov permutira tako, da se ne prekine logična povezava, saj bi lahko napadalec v nasprotnem primeru določil permutirane atribute in obrnil permutacijo.

Na primer, če vzamemo podmnožico atributov v naboru medicinskih podatkov, kot so „razlogi za bolnišnično oskrbo/simptomi/odgovorni oddelek“, so vse vrednosti v večini primerov močno logično povezane in bi bilo zato mogoče ugotoviti in celo obrniti permutacijo samo ene vrednosti.

Permutacija sama po sebi, podobno kot dodajanje šuma, morda ne zagotavlja anonimizacije, zato jo je treba vedno združiti z odstranitvijo očitnih atributov/kvaziidentifikatorjev.

3.1.2.1. Zagotovila

- *Izločitev:* Tako kot pri dodajanju šuma je še vedno mogoče izločiti zapise o posamezniku, ki pa so manj zanesljivi.
- *Povezljivost:* Če permutacija vpliva na atribute in kvaziidentifikatorje, lahko prepreči „pravilno“ notranjo in zunanjo povezavo z naborom podatkov, še vedno pa omogoča „nepravilno“ povezljivost, ker je mogoče pravi vnos povezati z drugim posameznikom, na katerega se nanašajo osebni podatki.
- *Sklepanje:* Nabor podatkov še vedno omogoča sklepanje, zlasti če so atributi med seboj povezani ali v tesnem logičnem razmerju; vendar mora napadalec, ki ne ve, kateri atributi so bili permutirani, upoštevati, da njegovo sklepanje temelji na napačni hipotezi, zato je mogoče samo verjetnostno sklepanje.

3.1.2.2. Pogoste napake

- Izbira napačnega atributa: permutacija neobčutljivih in netveganih atributov ne bi veliko pripomogla k varstvu osebnih podatkov. Če bi bili občutljivi/tvegani atributi še vedno povezani s prvotnim atributom, bi lahko napadalec dejansko še vedno pridobil občutljive informacije o posameznikih.
- Naključna permutacija atributov: Če sta dva atributa med seboj tesno povezana, naključna permutacija atributov ne zagotovi trdnih zagotovil. Ta pogosta napake je ponazorjena v preglednici 1.
- Predpostavka, da permutacija zadostuje: Tako kot dodajanje šuma tudi permutacija sama po sebi ne zagotavlja anonimizacije, zato jo je treba združiti z drugimi tehnikami, kot je odstranitev očitnih atributov.

3.1.2.3. Neuspešna permutacija

To je primer, kako se z naključno permutacijo atributov dosežejo šibka zagotovila glede zasebnosti, če med različnimi atributi obstajajo logične povezave. Po poskusu anonimizacije je izjemno preprosto sklepati o dohodku posameznikov glede na delovno

mesto (in leto rojstva). Na primer, na podlagi neposrednega pregleda podatkov je mogoče trditi, da je bil izvršni direktor iz preglednice zelo verjetno rojen leta 1957 in ima najvišjo plačo, medtem ko je bil brezposelni rojen leta 1964 in ima najnižji dohodek.

Leto	Spol	Delovno mesto	Dohodek (permutiran)
1957	M	Inženir	70k
1957	M	Izvršni direktor	5k
1957	M	Brezposelni	43k
1964	M	Inženir	100k
1964	M	Vodja	45k

Preglednica 1. Primer neuspešne anonimizacije s permutacijo povezanih atributov.

3.1.3. Diferencirana zasebnost

Diferencirana zasebnost[14] spada v skupino randomizacijskih tehnik z drugačnim pristopom: medtem ko se šum dejansko vstavi pred predvideno objavo nabora podatkov, se lahko diferencirana zasebnost uporabi, ko upravljavec podatkov oblikuje anonimizirane prikaze nabora podatkov, pri čemer ohrani kopijo prvotnih podatkov. Takšni anonimizirani prikazi se po navadi oblikujejo s podmnožico poizvedb za določeno tretjo osebo. Podmnožica vključuje nekaj naključnega šuma, ki se namerno doda naknadno. Diferencirana zasebnost upravljavcu podatkov pove, koliko šuma mora dodati in v kakšni obliki, da dobi potrebna zagotovila glede zasebnosti.[15] Pri tem je še zlasti pomembno stalno spremljanje (vsaj za vsako novo poizvedbo), ker obstaja možnost določitve posameznika v naboru rezultatov poizvedbe. Vendar je treba pojasniti, da tehnike diferencirane zasebnosti ne spremenijo prvotnih podatkov, zato lahko upravljavec podatkov, dokler se ohranjajo prvotni podatki, posameznike določi na podlagi rezultatov poizvedb diferencirane zasebnosti, ob upoštevanju vseh sredstev, za katera se pričakuje, da se bodo uporabila. Tudi takšni rezultati se morajo obravnavati kot osebni podatki.

Ena od prednosti pristopa, ki temelji na diferencirani zasebnosti, je dejstvo, da se nabori podatkov zagotovijo pooblaščenim tretjim osebam v odgovor na določeno poizvedbo in ne z objavo posameznega nabora podatkov. Upravljavec podatkov lahko za pomoč pri pregledu ohrani seznam vseh poizvedb in zahtevkov, s čimer se tretjim osebam prepreči dostop do podatkov, za katere niso pooblaščene. Pri poizvedbah se lahko za dodatno zaščito zasebnosti uporabijo anonimizacijske tehnike, vključno z dodajanjem šuma ali nadomestitvijo. Še vedno potekajo raziskave v zvezi z določitvijo ustreznega interaktivnega mehanizma poizvedbe in odgovora, s katerim bo mogoče pravično in točno odgovoriti na vsako vprašanje (kar pomeni z manj šuma) ter hkrati ohraniti zasebnost.

Za omejitev napadov, povezanih s sklepanjem in povezljivostjo, je treba spremljati poizvedbe, ki jih izda subjekt, in informacije, pridobljene o posameznikih, na katere se nanašajo osebni podatki; zato se podatkovne zbirke, ki vključujejo „diferencialno zasebnost“, ne bi smele uporabljati v odprtokodnih iskalnikih, ki ne omogočajo sledljivosti subjektov, ki poizvedujejo.

[14] Dwork, C. (2006). Diferencirana zasebnost. V *Automata, languages and programming* (Stroji, programski jeziki in programiranje) (str. 1–12). Springer Berlin Heidelberg.
[15] Glej Ed Felten (2012). Protecting privacy by adding noise (Zaščita zasebnosti z dodajanjem šuma). URL: https://techatftc.wordpress.com/2012/06/21/protecting-privacy-by-adding-noise/.

3.1.3.1. Zagotovila

- *Izločitev:* Če so rezultati samo statistični podatki in so pravila, ki se uporabljajo za nabor, ustrezno izbrana, posameznika ne bi smelo biti mogoče izločiti na podlagi odgovorov.
- *Povezljivost:* Na podlagi večkratnih zahtevkov je morda mogoče povezati vnose, ki se nanašajo na določenega posameznika, med dvema odgovoroma.
- *Sklepanje:* Na podlagi večkratnih zahtevkov je mogoče sklepati o informacijah o posameznikih ali skupinah.

3.1.3.2. Pogoste napake

- Nezadosten šum: Cilj je zagotoviti čim manj dokazov o tem, ali je določen posameznik, na katerega se nanašajo osebni podatki, ali skupina takšnih posameznikov prispevala k naboru podatkov, da se prepreči povezovanje z osnovni informacijami. Največja težava z vidika varstva podatkov je zagotoviti ustrezno količino šuma, ki se doda pravim odgovorom, da se zaščiti zasebnost posameznikov in hkrati ohrani uporabnost objavljenih odgovorov.

3.1.3.3. Napake pri diferencirani zasebnosti

Neodvisno obravnavanje vsake poizvedbe: Kombinacija rezultatov poizvedb lahko omogoči razkritje informacij, ki naj bi ostale zaupne. Če se zgodovina poizvedb ne ohrani, lahko napadalec za podatkovno zbirko z „diferencirano zasebnostjo“ oblikuje večkratna vprašanja, ki postopno zmanjšujejo velikost dobljenega vzorca, dokler morda ni mogoče z gotovostjo ali veliko verjetnostjo določiti posebnega značaja enega posameznika, na katerega se nanašajo osebni podatki, ali skupine takšnih posameznikov. Poleg tega je treba še paziti, da se prepreči napačno razmišljanje, da so podatki za tretjo osebo anonimni, medtem ko lahko upravljavec podatkov v prvotni podatkovni zbirki še vedno določi posameznika, na katerega se nanašajo osebni podatki, ob upoštevanju vseh sredstev, za katera se pričakuje, da se bodo uporabila.

3.2. Generalizacija

Generalizacija je druga skupina anonimizacijskih tehnik. Ta pristop vključuje posplošitev ali razvodenitev atributov posameznikov, na katere se nanašajo osebni podatki, s spremembo zadevnega obsega ali velikostnega razreda (npr. regija namesto mesto, mesec namesto teden). Čeprav je mogoče z generalizacijo uspešno preprečiti izločitev posameznikov, pa ta v vseh primerih ne omogoča učinkovite anonimizacije; zahteva zlasti posebne in zapletene kvantitativne pristope za preprečitev povezljivosti in sklepanja.

3.2.1. Združevanje in k-anonimnost

Cilj tehnik združevanja in k-anonimnosti je preprečiti izločitev posameznika, na katerega se nanašajo osebni podatki, tako da se ga razvrsti v skupino vsaj z drugimi k-posamezniki. Za dosego tega se vrednosti atributov toliko posplošijo, da imajo vsi posamezniki enako vrednost. Na primer, zmanjšanje razdrobljenosti lokacije iz mesta na državo pomeni več posameznikov, na katere se nanašajo osebni podatki. Datumi rojstva posameznikov se lahko posplošijo v razpon datumov ali združijo po mesecih ali letih. Drugi numerični atributi (npr. plače, teža, višina ali odmerek zdravila) se lahko posplošijo z intervalnimi vrednostmi (npr. plača od 20 000 do 30 000 EUR). Te metode se lahko uporabijo, če bi lahko s povezavo točnih vrednosti atributov nastali kvaziidentifikatorji.

3.2.1.1. Zagotovila

- *Izločitev:* K-uporabniki imajo zdaj enake atribute, zato posameznika ni več mogoče izločiti iz skupine k-uporabnikov.
- *Povezljivost:* Čeprav je povezljivost omejena, je še vedno mogoče povezati zapise po skupinah k-uporabnikov. Verjetnost, da dva zapisa v tej skupini ustrezata istim psevdoidentifikatorjem, je 1/k (kar je lahko znatno več, kot znaša verjetnost, da so takšni vnosi nepovezljivi).
- *Sklepanje:* Glavna slabost modela k-anonimnosti je, da ne preprečuje nobene vrste napada s sklepanjem. Če so vsi k-posamezniki v isti skupini in če se ve, v katero skupino spada posameznik, je dejansko izjemno preprosto pridobiti vrednost te lastnosti.

3.2.1.2. Pogoste napake

- Pomanjkanje nekaterih kvaziidentifikatorjev: Ključni parameter pri obravnavanju k-anonimnosti je mejna vrednost k. Višja kot je vrednost k, trdnejša so zagotovila glede zasebnosti. Pogosta napaka je umetno zvišanje vrednosti k z zmanjšanjem obravnavanega nabora kvaziidentifikatorjev. Z zmanjšanjem števila kvaziidentifikatorjev je zaradi privzete možnosti identifikacije, povezane z drugimi atributi, preprosteje ustvariti skupine k-uporabnikov (zlasti če so nekateri od njih občutljivi ali imajo zelo visoko entropijo, kot na primer pri zelo redkih atributih). Ključna napaka je, če se pri izbiri atributa za posplošitev ne upoštevajo vsi kvaziidentifikatorji; če je mogoče nekatere atribute uporabiti za izločitev posameznika v skupini k, potem nekateri posamezniki niso zaščiteni z generalizacijo (glej primer v preglednici 2).
- Majhna vrednost k: Tudi usmeritev v majhno vrednost k je podobno problematična. Če je k premajhen, je pomen posameznika v skupini prevelik, napadi s sklepanjem pa so uspešnejši. Na primer, če je k = 2, potem je verjetnost, da imata dva posameznika isto lastnost, večja kot za k > 10.
- Nezdruževanje posameznikov z enakim pomenom: Tudi združevanje skupine posameznikov z neenakomerno porazdelitvijo atributov je lahko problematično. Učinek posameznikovega zapisa na nabor podatkov je različen: nekateri pomenijo pomemben del vnosov, medtem ko so prispevki drugih razmeroma nepomembni. Zato je treba zagotoviti dovolj velik k, da noben posameznik v skupini ne pomeni preveč pomembnega dela vnosov.

3.1.3.3. Neuspešna k-anonimnost

Največja težava pri k-anonimnosti je, da ne preprečuje napadov s sklepanjem. Če napadalec v naslednjem primeru ve, da je določen posameznik v naboru podatkov in da je bil rojen leta 1964, potem tudi ve, da je imel srčni napad. Če vemo, da je bil ta nabor podatkov pridobljen od francoske organizacije, potem vsi posamezniki živijo v Parizu, ker so prve tri številke pariških poštnih številk 750* .

Leto	Spol	Poštna številka	Diagnoza
1957	M	750*	Srčni napad
1957	M	750*	Holesterol
1957	M	750*	Holesterol
1964	M	750*	Srčni napad
1964	M	750*	Srčni napad

Preglednica 2. Primer slabo zasnovane k-anonimizacije.

3.2.2. L-raznolikost/T-podobnost

L-raznolikost nadgrajuje k-anonimnost, da se prepreči možnost deterministčnih napadov s sklepanjem, in sicer z zagotovitvijo, da ima vsak atribut v posameznih ekvivalenčnih razredih vsaj *l* različne vrednosti.

Eden od temeljnih ciljev je omejiti nastanek ekvivalenčnih razredov z majhno spremenljivostjo atributa, tako da ostane napadalec z osnovnimi informacijami o določenem posamezniku, na katerega se nanašajo osebni podatki, vedno precej negotov.

L-raznolikost je uporabna za zaščito podatkov pred napadi s sklepanjem, če so vrednosti atributov ustrezno porazdeljene. Vendar je treba poudariti, da s to tehniko ni mogoče preprečiti uhajanja informacij, če so atributi v razdelku neenakomerno porazdeljeni ali pripadajo ozkemu razponu vrednosti ali pomenov. Nazadnje, l-raznolikost je predmet napadov z verjetnostnim sklepanjem.

T-podobnost je izboljšana l-raznolikost, saj je njen cilj ustvariti ekvivalenčne razrede, ki so podobni prvotni razporeditvi atributov v preglednici. Ta tehnika se uporabi, če morajo podatki ostati čim bolj podobni prvotnim; zato se ekvivalenčni razred dodatno omeji, da so v vsakem takšnem razredu vsaj / različne vrednosti in da je vsaka vrednost zastopana tolikokrat, kot je to potrebo, da se upošteva prvotna porazdelitev vsakega atributa.

3.2.2.1. Zagotovila

- *Izločitev:* L-raznolikost in t-podobnost lahko podobno kot k-anonimnost preprečita izločitev zapisov, ki se nanašajo na posameznika, v podatkovni zbirki.
- *Povezljivost*: l-raznolikost in t-podobnost ne pomenita izboljšave k-anonimnosti glede nepovezljivosti. Težava je enaka kot pri vsaki skupini: verjetnost, da isti vnosi pripadajo istemu posamezniku, na katerega se nanašajo osebni podatki, je večja kot 1/N (pri čemer je N število posameznikov, na katere se nanašajo osebni podatki, v podatkovni zbirki).
- *Sklepanje:* Glavna izboljšava l-raznolikosti in t-podobnosti v primerjavi s k-anonimnostjo je, da ni več mogoče popolnoma zanesljivo izvesti napadov s sklepanjem na podatkovno zbirko z „l-raznolikostjo“ ali „t-podobnostjo“.

3.2.2.2. Pogoste napake

- Zaščita vrednosti občutljivih atributov na podlagi združevanja z drugimi občutljivimi atributi: Dve vrednosti atributa v skupini nista dovolj za zagotovitev zagotovil glede

zasebnosti. Porazdelitev občutljivih vrednosti v vsaki skupini bi morala biti dejansko podobna porazdelitvi vrednosti v skupni populaciji ali vsaj enotna v vsej skupini.

3.2.2.3. Neuspešna l-raznolikost

L-raznolikost je v spodnji preglednici zagotovljena za atribut „diagnoza“, vendar je glede na nam znan podatek, da je v tej preglednici posameznik, ki je bil rojen leta 1964, še vedno mogoče z zelo veliko verjetnostjo predpostavljati, da je imel srčni napad.

Leto	Spol	Poštna številka	Diagnoza
1957	M	750*	**Srčni napad**
1957	M	750*	**Holesterol**
1957	M	750*	**Holesterol**
1957	M	750*	**Holesterol**
1964	M	750*	**Srčni napad**
1964	M	750*	**Srčni napad**
1964	M	750*	**Srčni napad**
1964	M	750*	**Holesterol**
1964	M	750*	**Srčni napad**
1964	M	750*	**Srčni napad**
1964	M	750*	**Srčni napad**
1964	M	750*	**Srčni napad**
1964	M	750*	**Srčni napad**
1964	M	750*	**Srčni napad**

Preglednica 3. Preglednica za l-raznolikost, v kateri vrednosti „diagnoze“ niso enakomerno porazdeljene.

Ime	Datum rojstva	Spol
Smith	1964	M
Rossi	1964	M
DuPont	1964	M
Jansen	1964	M
Garcia	1964	M

Preglednica 4. Če napadalec ve, da so ti posamezniki v preglednici 3, lahko sklepa, da so imeli srčni napad.

4. Psevdonimizacija

Psevdonimizacija vključuje nadomestitev enega atributa (običajno edinstvenega atributa) v zapisu z drugim. Posredna določitev fizične osebe je torej še vedno verjetna; zato psevdonimizacija, če se uporabi samostojno, ne pomeni anonimnega nabora podatkov. Vendar je v tem mnenju kljub temu obravnavana, ker v zvezi z njeno uporabo obstajajo številne napačne predstave in napake.

S psevdonimizacijo se zmanjša povezljivost nabora podatkov s prvotno identiteto posameznika, na katerega se nanašajo osebni podatki; kot takšna je uporaben varnostni ukrep, ne pa tudi metoda anonimizacije.

Rezultat psevdonimizacije je lahko neodvisen od prvotne vrednosti (tako kot pri naključnem številu, ki ga določi upravljavec, ali priimku, ki ga izbere posameznik, na katerega se nanašajo osebni podatki), ali pa lahko izhaja iz prvotnih vrednosti atributa ali nabora atributov, na primer zgoščevalna funkcija ali šifrirna shema.

Najpogosteje se uporabljajo naslednje psevdonimizacijske tehnike:

- Šifriranje s skrivnim ključem: imetnik ključa v tem primeru izjemno preprosto vsakega posameznika, na katerega se nanašajo osebni podatki, znova določi z dešifriranjem nabora podatkov, ker so osebni podatki še vedno v naboru podatkov, čeprav v šifrirani obliki. Dešifriranje je ob predpostavki, da je bila uporabljena najsodobnejša šifrirna shema, mogoče samo s poznavanjem ključa.
- Zgoščevalna funkcija: to je funkcija, ki iz vhodnih podatkov katere koli velikosti zagotovi rezultat nespremenljive velikosti (vhodni podatek je lahko en sam atribut ali nabor atributov) in ki je ni mogoče izvesti v obratni smeri; to pomeni, da tveganje povrnitve pri šifriranju ne obstaja več. Če pa je pri zgoščevalni funkciji znan razpon vrednosti vhodnih podatkov, se lahko ti ponovijo z zgoščevalno funkcijo, da se pridobi pravilna vrednost za določeni zapis. Na primer, če je bil nabor podatkov psevdonimiziran z zgoščevanjem nacionalne identifikacijske številke, potem se lahko ta preprosto pridobi z zgoščevanjem vseh mogočih vrednosti vhodnih podatkov in primerjavo rezultata z navedenimi vrednostmi v naboru podatkov. Zgoščevalne funkcije so po navadi zasnovane tako, da jih je mogoče sorazmerno hitro izračunati in so pogosto predmet napadov z grobo silo.[16] Sestaviti je mogoče tudi predhodno izračunane preglednice, da se omogoči množična povrnitev obsežnega nabora zgoščenih vrednosti.

  Z uporabo soljene zgoščevalne funkcije (če se atributu, ki se zgoščuje, doda naključna vrednost, za katero se uporablja izraz „sol“) je mogoče zmanjšati verjetnost pridobitve vrednosti vhodnih podatkov, še vedno pa je mogoče z razumnimi sredstvi izračunati prvotne vrednosti atributa, skrite za rezultatom soljene zgoščevalne funkcije.[17]
- Šifrirana zgoščevalna funkcija s shranjenim ključem: to je posebna zgoščevalna funkcija, ki uporablja skrivni ključ kot dodatni vhodni podatek (v tem se razlikuje od soljene zgoščevalne funkcije, saj sol po navadi ni zaupna). Upravljavec podatkov

[16] Takšni napadi so sestavljeni iz preskušanja vseh verjetnih vhodnih podatkov, da se oblikujejo preglednice povezav.

[17] Zlasti če je znana vrsta atributa (ime, številka socialnega zavarovanja, datum rojstva itd.). Če je bila izračunana vrednost večkrat zgoščena s kratko soljo, se je mogoče za dodajanje zahteve glede izračuna opreti na zgoščevalno funkcijo za izpeljavo ključa.

lahko funkcijo na atributu ponovi s skrivnim ključem, napadalec pa jo brez poznavanja ključa ponovi veliko težje, ker je število možnosti, ki jih je treba preskusiti, dovolj veliko, da to ni izvedljivo.

- Deterministično šifriranje ali šifrirana zgoščevalna funkcija z izbrisom ključa: ta tehnika je morda primerljiva z izbiro naključnega števila kot psevdonima za vsak atribut v podatkovni zbirki, čemur sledi izbris preglednice povezav. Ta rešitev omogoča[18] zmanjšanje tveganja povezljivosti med osebnimi podatki v naboru podatkov in podatki, ki se nanašajo na istega posameznika, v drugem naboru podatkov, če se uporabi drugačen psevdonim. Napadalec ob upoštevanju najsodobnejšega algoritma težko dešifrira ali ponovi funkcijo z izračunom, ker bi to pomenilo preskus vseh mogočih ključev, ker pravi ključ ni na voljo.
- Razčlenjevanje: ta tehnika se po navadi (med drugim) uporablja na finančnem področju za nadomestitev identifikacijske številke kartice z vrednostmi, ki imajo za napadalca manjšo uporabnost. Izhaja iz prejšnjih tehnik, zanjo pa je značilno, da temelji na uporabi enosmernih šifrirnih mehanizmov ali dodelitvi zaporedne ali naključno določene številke, ki ni matematično izračunana iz prvotnih podatkov, pri čemer se ta dodelitev izvede z indeksno funkcijo.

4.1. Zagotovila

- *Izločitev:* Še vedno je mogoče izločiti zapise o posameznikih, ker je posameznik še vedno določen z edinstvenim atributom, ki je rezultat psevdonimizacijske funkcije (tj. psevdonimiziran atribut).
- *Povezljivost:* Zapise je še vedno mogoče izjemno preprosto povezati med seboj z uporabo istega psevdonimiziranega atributa, ki se nanaša na istega posameznika. Tudi če se za istega posameznika, na katerega se nanašajo osebni podatki, uporabijo različni psevdonimizirani atributi, je lahko povezljivost še vedno mogoča z drugimi atributi. Če pa za določitev posameznika, na katerega se nanašajo osebni podatki, ni mogoče uporabiti nobenega drugega atributa v naboru podatkov in če je bila odpravljena vsaka povezava med prvotnim in psevdonimiziranim atributom (vključno z izbrisom prvotnih podatkov), ni očitne povezave med dvema naboroma podatkov, za katera se uporabljajo različni psevdonimizirani atributi.
- *Sklepanje:* Napadi s sklepanjem na dejansko identiteto posameznika, na katerega se nanašajo osebni podatki, so mogoči v okviru nabora podatkov ali med različnimi podatkovnimi zbirkami, v katerih se za posameznika uporablja isti psevdonimizirani atribut, ali če so psevdonimi sami po sebi razumljivi in ne zakrijejo ustrezno prvotne identitete posameznika, na katerega se nanašajo osebni podatki.

4.2. Pogoste napake

- Prepričanje, da je psevdonimiziran nabor podatkov anonimiziran: Upravljavci podatkov pogosto predpostavljajo, da odstranitev ali nadomestitev enega ali več atributov zadostuje za zagotovitev anonimnosti nabora podatkov. Številni primeri kažejo, da to ne drži; samo s spremembo identifikatorja ni mogoče preprečiti, da nekdo ne bi določil posameznika, na katerega se nanašajo osebni podatki, če v naboru podatkov ostanejo kvaziidentifikatorji ali če je z vrednostmi drugih atributov še vedno mogoče določiti posameznika. Pogosto je lahko enako preprosto določiti posameznika v psevdonimiziranem naboru podatkov kot s prvotnimi podatki. Da bi se nabor

[18] Odvisno od drugih atributov v naboru podatkov in izbrisa prvotnih podatkov.

podatkov štel za anonimen, je treba sprejeti posebne ukrepe, vključno z odstranjevanjem in posploševanjem atributov ali izbrisom prvotnih podatkov ali vsaj s premestitvijo teh podatkov na višjo zbirno raven.

- Pogoste napake pri uporabi psevdonimizacije kot tehnike za zmanjšanje povezljivosti:

  - Uporaba istega ključa v različnih podatkovnih zbirkah: odprava povezljivosti različnih naborov podatkov je zelo odvisna od uporabe šifriranega algoritma in dejstva, da posameznik v različnih okoliščinah ustreza različnim psevdonimiziranim atributom. Zato se je treba izogniti uporabi istega ključa v različnih podatkovnih zbirkah, da se omogoči zmanjšanje povezljivosti.
  - Uporaba različnih ključev („izmeničnih ključev“) za različne porabnike: uporaba različnih ključev za različne skupine uporabnikov in sprememba ključa na podlagi uporabe (npr. uporaba istega ključa za zapis desetih vnosov, ki se nanašajo na istega uporabnika) se morda zdi privlačna. Vendar lahko ta postopek, če ni ustrezno zasnovan, povzroči pojav vzorcev, ki delno zmanjšajo predvidene koristi. Na primer, izmenjava ključa v skladu s posebnimi pravili za določene posameznike bi olajšal povezljivost vnosov, ki se nanašajo na določene posameznike. Prav tako bi lahko izginotje ponavljajočih se psevdonimiziranih podatkov v podatkovni zbirki ob pojavu novih kazalo na to, da se oba zapisa nanašata na isto fizično osebo.
  - Hramba ključa: Če se skrivni ključ hrani skupaj s psevdonimiziranimi podatki in če so podatki ogroženi, lahko napadalec izjemno preprosto poveže psevdonimizirane podatke z njihovim prvotnim atributom. Enako velja, če se ključ hrani ločeno od podatkov, vendar ne varno.

4.3. Pomanjkljivosti psevdonimizacije

- *Zdravstveno varstvo*

1. **Ime, naslov, datum rojstva**	**2.** **Obdobje prejete posebne pomoči**	**3.** **Indeks telesne mase**	**6.** **Referenčna št. raziskovalne kohorte**
[redacted]	< 2 leti	15	QA5FRD4
	> 5 let	14	2B48HFG
	< 2 leti	16	RC3URPQ
	> 5 let	18	SD289K9
	< 2 leti	20	5E1FL7Q

Preglednica 5. Primer psevdonimizacije z zgoščevanjem (ime, naslov, datum rojstva), ki jo je mogoče zlahka obrniti.

Nabor podatkov je bil ustvarjen za proučitev povezave med posameznikovo telesno maso in prejemom plačila posebnega denarnega nadomestila. Prvotni nabor podatkov je vključeval ime, naslov in datum rojstva posameznika, na katerega se nanašajo osebni podatki, ki pa so bili izbrisani. Referenčna številka raziskovalne kohorte je bila določena z zgoščevalno funkcijo iz izbrisanih podatkov. Čeprav so bili ime, naslov in datum rojstva izbrisani iz preglednice, je ob poznavanju imena, naslova in datuma rojstva posameznika, na katerega se nanašajo osebni podatki, ter uporabljene zgoščevalne funkcije preprosto izračunati referenčne številke raziskovalnih kohort.

- *Družbena omrežja*

Prikazano je bilo[19] , da je mogoče iz grafov družbenih omrežij pridobiti občutljive informacije o posameznikih, kljub temu da se za takšne podatke uporabljajo „psevdonimizacijske“ tehnike. Ponudnik družbenega omrežja je napačno domneval, da je psevdonimizacija dovolj zanesljiva, da ne omogoča identifikacije po prodaji podatkov drugim podjetjem, ki jih bodo uporabila za trženje in oglaševanje. Namesto pravih imen je uporabil psevdonime, kar pa očitno ni zadostovalo za anonimizacijo uporabniških profilov, ker so povezave med različnimi posamezniki edinstvene in jih je mogoče uporabiti kot identifikator.

- *Lokacije*

Raziskovalci pri MIT[20] so pred kratkim analizirali psevdonimiziran nabor podatkov, sestavljen iz prostorsko-časovnih koordinat mobilnosti 1,5 milijona ljudi v 15 mesecih na območju v polmeru 100 km. Pokazali so, da bi bilo mogoče s štirimi lokacijami izločiti 95 % prebivalstva, samo dve lokaciji pa sta zadostovali za izločitev več kot 50 % posameznikov, na katere so se nanašali osebni podatki (ena od teh lokacij je znana in je zelo verjetno „dom“ ali „pisarna“), pri čemer je prostor za varstvo zasebnosti zelo omejen, tudi če so bile identitete posameznikov psevdonimizirane z nadomestitvijo njihovih pravih atributov […] z drugimi oznakami.

5. Sklepne ugotovitve in priporočila

5.1. Sklepne ugotovitve

V zvezi s postopki deidentifikacije in anonimizacije potekajo intenzivne raziskave, v tem mnenju pa je dosledno prikazano, da ima vsaka tehnik svoje prednosti in slabosti. V večini primerov ni mogoče zagotoviti osnovnih priporočil glede parametrov, ki jih je treba uporabiti, ker je treba vsak nabor podatkov obravnavati za vsak primer posebej.

Anonimiziran nabor podatkov lahko v številnih primerih še vedno pomeni preostalo tveganje za posameznike, na katere se nanašajo osebni podatki. Tudi če zapisa o posamezniku ni več mogoče natančno pridobiti, je morda še vedno mogoče zbrati informacije o navedenem posamezniku na podlagi drugih virov informacij, ki so na voljo (javno ali ne). Poudariti je treba, da se lahko poleg neposrednega učinka posledic neustreznega anonimizacijskega postopka (nadlegovanje, porabljen čas in občutek izgube nadzora z vključitvijo v skupino brez obvestila ali predhodnega soglasja) na posameznike, na katere se nanašajo osebni podatki, pojavijo še drugi posredni stranski učinki neustrezne anonimizacije, če napadalec po pomoti zaradi obdelave anonimiziranih podatkov vključi posameznika, na katerega se nanašajo osebni podatki, v cilj – zlasti če je napadalec zlonameren. Delovna skupina zato poudarja, da je mogoče z anonimizacijskimi tehnikami dati zagotovila glede zasebnosti, vendar samo če je njihova uporaba ustrezno zasnovana – kar pomeni, da morajo biti pogoji (okoliščine) in cilj(-i) anonimizacijskega postopka jasno določeni, da se doseže ciljna raven anonimizacije.

[19] A. Narayanan in V. Shmatikov, „De-anonymizing social networks“ (Deanonimizacija družbenih omrežij), na 30. simpoziju IEEE o varnosti in zasebnosti, 2009.
[20] Y.-A. de Montjoye, C. Hidalgo, M. Verleysen in V. Blondel, „Unique in the Crowd: The privacy bounds of human mobility,“ (Edinstveni v množici: Meje zasebnosti pri mobilnosti ljudi), Nature, št. 1376, 2013.

5.2. Priporočila

- Nekatere anonimizacijske tehnike same po sebi vključujejo omejitve. Upravljavci podatkov morajo te omejitve resno proučiti, preden uporabijo določeno tehniko za oblikovanje anonimizacijskega postopka. Upoštevati morajo namene, ki naj bi se dosegli z anonimizacijo – kot sta varstvo posameznikove zasebnosti ob objavi nabora podatkov ali dovoljenje za pridobitev informacije iz nabora podatkov.
- Nobena tehnika, opisana v tem dokumentu, ne izpolnjuje zanesljivo merila učinkovite anonimizacije (tj. brez izločitve posameznika, brez povezljivosti med zapisi, ki se nanašajo na posameznika, in brez sklepanja o posamezniku). Ker pa lahko določena tehnika v celoti ali delno odpravi nekatera od teh tveganj, je treba skrbno zasnovati uporabo posamezne tehnike v določenih okoliščinah in uporabo kombinacije navedenih tehnik kot načina za povečanje zanesljivosti rezultata.

Spodnja preglednica vsebuje pregled prednosti in slabosti obravnavanih tehnik glede na tri temeljne zahteve:

	Ali še vedno obstaja tveganje za izločitev?	**Ali še vedno obstaja tveganje za povezljivost?**	**Ali še vedno obstaja tveganje za sklepanje?**
Psevdonimizacija	Da	Da	Da
Dodajanje šuma	Da	Morda	Morda
Nadomestitev	Da	Da	Morda
Združevanje ali k-anonimnost	Ne	Da	Da
L-raznolikost	Ne	Da	Morda
Diferencirana zasebnost	Morda	Morda	Morda
Zgoščevanje/Razčlenjevanje	Da	Da	Morda

Preglednica 6. Prednosti in slabosti obravnavanih tehnik.

- Najboljšo rešitev bi bilo treba določiti za vsak primer posebej. Rešitev (tj. popoln anonimizacijski postopek), ki bi izpolnjevala vsa tri merila, bi bila varna pred identifikacijo, ki jo izvede upravljavec podatkov ali katera koli tretja oseba z najverjetnejšimi in razumnimi sredstvi.
- Če predlog ne izpolnjuje enega od meril, bi bilo treba opraviti temeljito oceno identifikacijskih tveganj. To oceno bi bilo treba predložiti organu, če se v skladu z nacionalno zakonodajo zahteva, da organ oceni ali odobri anonimizacijski postopek.

Za zmanjšanje identifikacijskih tveganj bi bilo treba upoštevati naslednje dobre prakse:

<u>Dobre anonimizacijske prakse</u>

Na splošno:

- Ne temeljijo na pristopu „objavi in pozabi“. Upravljavci podatkov bi morali glede na preostalo tveganje:
  - 1. opredeliti nova tveganja in redno znova ocenjevati preostalo oziroma preostala tveganja;
  - 2. oceniti, ali je nadzor nad opredeljenimi tveganji zadovoljiv in ustrezno prilagojen, TER

- 3. spremljati in nadzorovati tveganja.

– V okviru takšnih preostalih tveganj je treba upoštevati možnost identifikacije z neanonimiziranim delom nabora podatkov (če obstaja), zlasti če je združen z anonimiziranim delom, in morebitne povezave med atributi (npr. med podatki o geografski lokaciji in ravni blaginje).

Okoliščine:

– Namene, ki naj se dosežejo z anonimizacijo nabora podatkov, bi bilo treba jasno določiti, ker imajo ključno vlogo pri opredelitvi identifikacijskega tveganja.

– To je povezano z upoštevanjem vseh zadevnih okoliščin – na primer narave prvotnih podatkov, vzpostavljenih nadzornih mehanizmov (vključno z varnostnimi ukrepi za omejitev dostopa do naborov podatkov), velikosti vzorca (količinskih značilnosti), razpoložljivosti javnih informacijskih virov (na katere se zanašajo prejemniki), predvidene predložitve podatkov tretjim osebam (omejene, neomejene, na primer na spletu itd.).

– Ob upoštevanju privlačnosti podatkov za ciljno usmerjene napade je treba proučiti morebitne napadalce (tudi tukaj sta občutljivost informacij in narava podatkov ključna dejavnika).

Tehnični elementi:

– Upravljavci podatkov bi morali razkriti anonimizacijsko tehniko/mešanico tehnik, ki jih izvajajo, zlasti če nameravajo objaviti anonimiziran nabor podatkov.

– Očitne (npr. redke) atribute/kvaziidentifikatorje bi bilo treba odstraniti iz nabora podatkov.

– Če se uporabljajo tehnike dodajanja šuma (pri randomizaciji), bi bilo treba raven šuma, dodanega k zapisom, določiti kot funkcijo vrednosti atributa (kar pomeni, da se ne sme vnesti nesorazmerni šum), učinka atributov, ki naj bi se zaščitili, na posameznike, na katere se nanašajo osebni podatki, in/ali razpršenosti nabora podatkov.

– Pri uporabi diferencirane zasebnosti (pri randomizaciji) je treba upoštevati potrebo po spremljanju poizvedb, da se odkrijejo tiste, ki posegajo v zasebnost, ker je poseganje poizvedb v zasebnost kumulativno.

– Če se izvajajo generalizacijske tehnike, je ključno, da se upravljavec podatkov niti pri istem atributu ne omeji na eno samo merilo generalizacije; to pomeni, da bi bilo treba izbrati različne lokacijske razdrobljenosti ali različne časovne presledke. Izbor predvidenih meril je treba določiti s porazdelitvijo vrednosti atributov v določeni populaciji. Vseh porazdelitev ni mogoče posplošiti – kar pomeni, da pri generalizaciji ni mogoče uporabiti univerzalnega pristopa. Zagotoviti bi bilo treba variabilnost v ekvivalenčnih razredih; ob upoštevanju navedenih „okoliščin“ (velikosti vzorca itd.) bi bilo treba na primer izbrati določen prag, in če se ta ne doseže, bi bilo treba določen vzorec zavreči (ali pa določiti drugo merilo generalizacije).

PRILOGA
Priročnik o anonimizacijskih tehnikah

A.1. Uvod

Anonimnost se v EU razlaga različno – v nekaterih državah ustreza računalniški anonimnosti (kar pomeni, da celo upravljavec v sodelovanju s katero koli tretjo osebo računalniško težko neposredno ali posredno določi enega od posameznikov, na katerega se nanašajo osebni podatki), v drugih pa popolni anonimnosti (kar pomeni, da niti upravljavec v sodelovanju s katero koli tretjo osebo ne more neposredno ali posredno določiti enega od posameznikov, na katerega se nanašajo osebni podatki). „Anonimizacija“ kljub temu v obeh primerih ustreza postopku, s katerim se zagotovi anonimnost podatkov. Razlika je v določitvi sprejemljive ravni tveganja za ponovno identifikacijo.

Za anonimizirane podatke je mogoče predvideti različne primere uporabe, ki med drugim vključujejo družbene raziskave, statistične analize in razvoj novih storitev ali izdelkov. Včasih lahko tudi takšne splošne dejavnosti vplivajo na določene posameznike, na katere se nanašajo osebni podatki, in izničijo domnevno anonimnost obdelanih podatkov. Navesti je mogoče številne primere, vse od uvedbe ciljno usmerjenih pobud na področju trženja do izvajanja javnih ukrepov na podlagi oblikovanja profilov uporabnikov, njihovega ravnanja ali vzorcev mobilnosti[21].

Razen splošnih trditev žal ne obstaja zrela matrika za predhodno oceno časa ali prizadevanj, potrebnih za ponovno identifikacijo po obdelavi podatkov, ali kakšna druga možnost za določitev najustreznejšega postopka, ki se vzpostavi, če želi nekdo zmanjšati verjetnost, da se objavljena podatkovna zbirka nanaša na določeno skupino posameznikov, na katere se nanašajo osebni podatki.

„Umetnost anonimizacije“, kot se te prakse včasih imenujejo v znanstveni literaturi[22], je novo znanstveno področje, ki je še v povojih, obstajajo pa številne prakse za zmanjšanje možnosti identifikacije naborov podatkov; vendar je treba jasno povedati, da večina teh postopkov ne preprečuje povezave obdelanih podatkov s posamezniki, na katere se nanašajo osebni podatki. V nekaterih primerih je bilo dokazano, da je bila identifikacija naborov podatkov, ki so se šteli za anonimne, zelo uspešna, v drugih pa so se pojavili napačni pozitivni rezultati.

Na splošno obstajata dva različna pristopa: en temelji na generalizaciji atributa, drugi pa na randomizaciji. Pregled podrobnosti in domiselnosti teh praks nam zagotovi nov vpogled v možnost identifikacije podatkov in na novo pojasni sam pojem osebnih podatkov.

A.2. „Anonimizacija“ z randomizacijo

Ena od možnosti za anonimizacijo je sprememba dejanskih vrednosti, da se prepreči povezava med anonimiziranimi podatki in prvotnimi vrednostmi. Ta cilj je mogoče doseči z najrazličnejšimi metodologijami, ki med drugim vključujejo dodajanje šuma in zamenjavo podatkov (permutacijo). Poudariti je treba, da je odstranitev atributa enakovredna skrajni obliki randomizacije tega atributa (ko je atribut popolnoma prikrit s šumom).

[21] Primer je TomTom na Nizozemskem (glej primer, pojasnjen v odstavku 2.2.3).

[22] Jun Gu, Yuexian Chen, Junning Fu, Huanchun Peng, Xiaojun Ye, Synthesizing: Art of Anonymization, Database and Expert Systems Applications (Sintetiziranje: Umetnost anonimizacije, uporabe podatkovnih zbirk in naprednih sistemov), zapiski predavanj iz računalništva – Springer, zv. 6261, 2010, str. 385–399.

Cilj celotnega postopka v nekaterih okoliščinah ni objaviti randomiziran nabor podatkov, temveč odobriti dostop do podatkov s poizvedbami. Tveganje za posameznika, na katerega se nanašajo osebni podatki, v tem primeru izhaja iz verjetnosti, da bo napadalec sposoben pridobiti informacije z najrazličnejšimi poizvedbami, ne da bi upravljavec podatkov vedel za to. Za zagotovitev anonimnosti posameznikom v naboru podatkov v zvezi s tem ne bi smelo biti mogoče sklepati, da je posameznik, na katerega se nanašajo osebni podatki, prispeval k naboru podatkov, s čimer se prekine povezava s kakršnimi koli osnovnimi informacijami, ki jih morda ima napadalec.

Tveganje za ponovno identifikacijo se lahko dodatno zmanjša z ustreznim šumom, ki se doda odgovoru na poizvedbo. Ta pristop, ki se v literaturi imenuje tudi diferencirana zasebnost[23], se od opisanih pristopov razlikuje po tem, da izdajateljem podatkov v primerjavi z javno objavo zagotavlja večji nadzor nad dostopom do podatkov. Dodajanje šuma ima dva glavna cilja: prvič, zaščititi zasebnost posameznikov, na katere se nanašajo osebni podatki, v naboru podatkov in, drugič, ohraniti uporabnost objavljenih informacij. Velikost šuma mora biti zlasti sorazmerna z ravnjo poizvedovanja (preveč poizvedb o posameznikih, na katere se odgovori preveč natančno, poveča verjetnost identifikacije). Uspešno uporabo randomizacije je treba zdaj obravnavati za vsak primer posebej, pri čemer nobena tehnika ne zagotavlja popolnoma zanesljive metodologije, saj obstajajo primeri uhajanja informacij o atributih posameznika, na katerega se nanašajo osebni podatki (če je bil vključen v nabor podatkov ali ne), tudi če je upravljavec podatkov menil, da je nabor podatkov randomiziran.

Obravnavanje posebnih primerov lahko pomaga pojasniti morebitne neuspešne randomizacije kot sredstva za zagotovitev anonimnosti. Na primer, pri interaktivnem dostopu lahko poizvedbe, ki upoštevajo spoštovanje zasebnosti, pomenijo tveganje za posameznike, na katere se nanašajo osebni podatki. Če napadalec ve, da je podskupina posameznikov S v naboru podatkov, ki vsebuje informacije o pogostosti atributa A v populaciji P, je dejansko mogoče samo s poizvedovanjem z dvema vprašanjema, tj. „Koliko posameznikov v populaciji P ima atribut A?“ in „Koliko posameznikov v populaciji P, razen posameznikov, ki spadajo v podskupino S, ima atribut A?“, (iz razlike) določiti število posameznikov v podskupini S, ki imajo dejansko atribut A – deterministično ali z verjetnostnim sklepanjem. Vsekakor je lahko zasebnost posameznikov v podskupini S resno ogrožena, kar je odvisno predvsem od narave atributa A.

Prav tako je mogoče meniti, da lahko objava nabora podatkov, če posameznik, na katerega se nanašajo osebni podatki, ni v naboru podatkov, vendar je znana njegova povezava s podatki v naboru podatkov, povzroči tveganje za njegovo zasebnost. Na primer, če se ve, da „se vrednost ciljne osebe za atribut A za količino X razlikuje od povprečne vrednosti populacije“, lahko napadalec samo s tem, da upravitelja podatkovne zbirke zaprosi, naj izvede postopek, ki upošteva spoštovanje zasebnosti, za pridobitev povprečne vrednosti atributa A, natančno sklepa o osebnih podatkih, ki se nanašajo na določenega posameznika.

Vnos nekaterih relativnih nepravilnosti v dejanske vrednosti v podatkovni zbirki je postopek, ki ga je treba ustrezno zasnovati. Za zaščito zasebnosti je treba dodati dovolj šuma, hkrati pa ne preveč, da podatki ostanejo uporabni. Na primer, če je število posameznikov, na katere se nanašajo osebni podatki, z značilnim atributom zelo majhno ali če je občutljivost atributa velika, je morda bolje namesto dejanskega števila sporočiti razpon ali splošni stavek, kot je na primer „malo primerov, verjetno celo nič“. Tudi če je mehanizem za razkritje s šumom znan vnaprej, se zasebnost posameznika, na katerega se nanašajo osebni podatki, ohrani, saj ostaja

[23] Cynthia Dwork, Differential Privacy (Diferencirana zasebnost), Mednarodna konferenca o avtomatizaciji, jezikih in programiranju (ICALP) 2006, str. 1–12.

določena stopnja negotovosti. Če je netočnost ustrezno zasnovana, se lahko rezultati z vidika uporabnosti še vedno uporabijo za statistiko ali odločanje.

Randomizacijo podatkovne zbirke in dostop pri diferencirani zasebnosti je treba dodatno pojasniti. Prvič, ustrezna količina izkrivljanja se lahko zelo razlikuje glede na okoliščine (vrsta poizvedbe, velikost populacije v podatkovni zbirki, narava atributa in njegova privzeta možnost identifikacije), zato ni mogoče predvideti enake rešitve za vse. Poleg tega se lahko okoliščine sčasoma spremenijo, zato bi bilo treba v skladu z njimi spremeniti interaktivni mehanizem. Umerjanje šuma zahteva spremljanje kumulativnih tveganj za zasebnost, ki jih pomeni vsak interaktivni mehanizem za posameznike, na katere se nanašajo osebni podatki. Mehanizem za dostop do podatkov bi bilo treba zato opremiti z opozorili, če je bil dosežen proračun za „stroške za zasebnost“, posamezniki, na katere se nanašajo osebni podatki, pa bi lahko bili izpostavljeni posebnim tveganjem, če se določi nova poizvedba, ki bi upravljavcu podatkov pomagala opredeliti ustrezno raven izkrivljanja, ki jo je treba vedno znova vnesti v dejanske osebne podatke.

Po drugi strani pa bi morali upoštevati tudi primere, kadar se vrednosti atributov izbrišejo (ali spremenijo). Za obravnavanje neznačilnih vrednosti atributov se pogosto uporablja rešitev, ki vključuje izbris nabora podatkov, ki se nanašajo na neznačilne posameznike, ali neznačilnih vrednosti. V zadnjem primeru je treba zato zagotoviti, da sama odsotnost vrednosti ne postane element za določitev posameznika, na katerega se nanašajo osebni podatki.

Sledi obravnavanje randomizacije z nadomestitvijo atributa. Veliko napačno prepričanje, ki se pojavlja v zvezi z anonimizacijo, je, če se ta šteje za enako šifriranju ali šifriranju s ključem. To napačno prepričanje temelji na dveh predpostavkah – in sicer, (a) če se je šifriranje uporabilo za nekatere atribute zapisa v podatkovni zbirki (npr. za ime, naslov, datum rojstva) ali če so se ti atributi nadomestili z navidezno randomiziranim nizom, ki je rezultat postopka šifriranja s ključem, kot je zgoščevalna funkcija s ključem, potem je ta zapis „anonimiziran“, in (b) anonimizacija je učinkovitejša, če je ključ ustrezno dolg in če se uporabi najsodobnejši šifrirni algoritem. To napačno prepričanje je zelo razširjeno med upravljavci podatkov in ga je treba pojasniti, ker je povezano tudi s psevdonimizacijo in njenimi domnevno manjšimi tveganji.

Prvič, cilji teh tehnik so povsem drugačni: cilj šifriranja kot varnostne prakse je zagotoviti zaupnost komunikacijske poti med določenima stranema (med ljudmi, napravami ali deli programske/strojne opreme), da se prepreči prisluškovanje ali nenamerno razkritje. Šifriranje s ključem ustreza semantičnemu prenosu podatkov na podlagi skrivnega ključa. Po drugi strani pa je cilj anonimizacije onemogočiti identifikacijo posameznikov s preprečitvijo prikritega povezovanja atributov posameznika, na katerega se nanašajo osebni podatki.

Niti šifriranje niti šifriranje s ključem sama po sebi ne dosegata cilja glede preprečitve možnosti identifikacije posameznika, na katerega se nanašajo osebni podatki: ker so prvotni podatki, ki jih ima vsaj upravljavec, še vedno na voljo ali pa je mogoče o njih sklepati. Samo s semantičnim prenosom osebnih podatkov, kot je to pri šifriranju s ključem, se ne odpravi možnost povrnitve podatkov v njihovo prvotno strukturo – z uporabo algoritma v obratni smeri ali napadi z uporabo grobe sile, odvisno od narave shem ali rezultata kršitve varnosti osebnih podatkov. Z najsodobnejšim šifriranjem je mogoče zagotoviti višjo stopnjo varstva podatkov, kar pomeni, da je za subjekte, ki ne poznajo šifrirnega ključa, to nerazumljivo, vendar njegov rezultat ni nujno anonimizacija. Dokler so na voljo ključ ali prvotni podatki (tudi pri zaupanja vredni tretji osebi, ki je pogodbeno zavezana, da mora zagotoviti varno hrambo ključa), možnost identifikacije posameznika, na katerega se nanašajo osebni podatki, ni izključena.

Osredotočanje zgolj na zanesljivost šifrirnega mehanizma kot ukrepa stopnje „anonimizacije“ nabora podatkov je zavajajoče, ker na splošno varnost šifrirnega mehanizma ali zgoščevalne funkcije vplivajo številni drugi tehnični in organizacijski dejavniki. V literaturi poročajo o številnih uspešnih napadih, ki so popolnoma obšli algoritem, ker so izkoristili neustrezno hrambo ključev (npr. obstoj manj varnega privzetega načina) ali druge človeške dejavnike (npr. šibko geslo za povrnitev ključa). Izbrana šifrirna shema z določeno velikostjo ključa je nenazadnje zasnovana za zagotovitev zaupnosti v določenem obdobju (večini sedanjih ključev bo treba spremeniti velikost okoli leta 2020), medtem ko anonimizacijski postopek ne bi smel biti časovno omejen.

Zato je treba zdaj podrobneje proučiti omejitve randomizacije atributov (ali njihovo nadomestitev in odstranitev), ob upoštevanju različnih primerov neuspešne anonimizacije z randomizacijo iz zadnjih let ter razlogov za neuspešnost.

Zelo znan primer, ki vključuje objavo slabo anonimiziranega nabora podatkov, je nagrada Neflix[24]. Vzemimo za primer splošni zapis v podatkovni zbirki, v katerem je bilo randomiziranih več atributov za posameznika, na katerega se nanašajo osebni podatki: vsak zapis je mogoče razdeliti še na dva podzapisa: {randomizirane atribute, nešifrirane atribute}, pri čemer so lahko nešifrirani atributi katera koli kombinacija domnevno neosebnih podatkov. Na podlagi nabora podatkov nagrade Netflix je mogoče sprejeti posebno ugotovitev, ki temelji na mnenju, da je mogoče vsak zapis predstaviti s točko v večdimenzionalnem prostoru, kjer je vsak nešifrirani atribut koordinata. S to tehniko je mogoče vsak nabor podatkov obravnavati kot konstelacijo točk v takšnem večdimenzionalnem prostoru, ki lahko kaže znake visoke stopnje razpršenosti, kar pomeni, da so lahko točke oddaljene druga od druge. Dejansko so lahko tako oddaljene, da vsaka regija po razdelitvi prostora na širše regije vsebuje samo en zapis. Tudi z dodajanjem šuma ni mogoče doseči, da bi bili zapisi dovolj blizu, da bi bili v isti večdimenzionalni regiji. Na primer, v poskusu Netflix so bili zapisi dovolj edinstveni z ocenami samo osmih filmov, izraženih v razmiku 14 dni. Po dodanem šumu k ocenam in datumom ni bilo mogoče ugotoviti združevanja regij. Z drugimi besedami, ista izbira samo osmih ocenjenih filmov je sestavljala prstni odtis izraženih ocen, ki ni bil enak niti za dva posameznika, na katera so se nanašali osebni podatki, iz podatkovne zbirke. Raziskovalci so na podlagi te geometrične ugotovitve povezali domnevno anonimni nabor podatkov Netflix z drugo javno podatkovno zbirko z ocenami filmov (IMDB) in tako odkrili uporabnike, ki so izrazili ocene za iste filme v istih časovnih presledkih. Ker je bila za večino uporabnikov vzpostavljena povezava v razmerju ena proti ena, je bilo mogoče pomožne informacije, pridobljene iz podatkovne zbirke IMDB, uvoziti v objavljeni nabor podatkov Netflix in tako z identitetami obogatiti vse domnevno anonimizirane zapise.

Poudariti je treba, da je to splošna lastnost: preostali del vsake randomizirane podatkovne zbirke ima še vedno veliko zmožnost identifikacije, odvisno od redkosti kombinacije ostalih atributov. To je opozorilo, ki bi ga morali upravljavci podatkov upoštevati vedno, ko izberejo randomizacijo kot svoj način za dosego ciljno usmerjene anonimizacije.

V številnih takšnih poskusih ponovne identifikacije je bil uporabljen podoben pristop projeciranja dveh podatkovnih zbirk v isti podprostor. To je zelo zmogljiva metodologija ponovne identifikacije, ki se je v zadnjem času veliko uporabljala na različnih področjih. Na

[24] Arvind Narayanan in Vitaliju Shmatikov: Robust de-anonymization of large sparse datasets (Zanesljiva deanonimizacija obsežnih razpršenih naborov podatkov). Simpozij IEEE o varnosti in zasebnosti 2008: str. 111–125

primer, v poskusu identifikacije, izvedenem zoper družbeno omrežje[25], je bil uporabljen socialni graf uporabnikov, psevdonimiziranih z oznakami. V tem primeru so bili atributi, uporabljeni za identifikacijo, seznam stikov vsakega uporabnika, saj je bilo prikazano, da je zelo majhna verjetnost, da bi imela dva posameznika enaka seznama. Na podlagi te intuitivne predpostavke je bilo ugotovljeno, da podgraf notranjih povezav zelo omejenega števila vozlišč sestavlja topološki prstni odtis, ki ga je treba pridobiti in je skrit v omrežju, in da je mogoče po določitvi tega podomrežja določiti velik del celotnega družbenega omrežja. Samo še nekaj podatkov o uspešnosti podobnega napada: prikazano je bilo, da lahko uporaba manj kot desetih vozlišč (ki lahko povzročijo milijon različnih podomrežnih konfiguracij, od katerih lahko vsaka sestavlja topološki prstni odtis) povzroči, da je družbeno omrežje z več kot štirimi milijoni psevdonimiziranih vozlišč in 70 milijoni povezav dovzetno za napade za ponovno identifikacijo, ogrožena pa je lahko tudi zasebnost velikega števila povezav. Poudariti je treba, da ta pristop k ponovni identifikaciji ni prilagojen posebnim okoliščinam družbenih omrežij, temveč je dovolj splošen, da ga je mogoče prilagoditi drugim podatkovnim zbirkam, v katerih so zabeležene povezave med uporabniki (npr. telefonski stiki, dopisovanje po e-pošti, strani za zmenke itd.).

Druga možnost identifikacije domnevno anonimnega zapisa temelji na analizi sloga pisanja (stilometriji)[26]. Razvitih je bilo že več algoritmov za določitev metrike na podlagi analiziranega besedila, ki med drugim vključujejo pogostost uporabe določene besede, pojavljanje posebnih slovničnih struktur in vrsto ločil. Z vsemi temi značilnostmi je mogoče domnevno anonimno besedilo povezati s slogom pisanja določenega avtorja. Raziskovalci so pridobili slog pisanja iz več kot 100 000 blogov in danes se lahko avtor prispevka samodejno določi s skoraj 80-odstotno natančnostjo; natančnost te tehnike naj bi se še povečala tudi z drugimi znamenji, kot so kraj ali drugi metapodatki iz besedila.

Zmožnost identifikacije s semantiko zapisa (tj. ostalega, nerandomiziranega dela zapisa) je vprašanje, ki bi mu morali raziskovalna skupnost in industrija nameniti več pozornosti. Nedavna povrnitev identitet darovalcev DNK (leta 2013)[27] kaže, da je bil od znanega incidenta družbe AOL, ki se je zgodil leta 2006 – ko je bila javno objavljena podatkovna zbirka z več kot dvajsetimi milijoni iskalnih ključnih besed za več kot 650 000 uporabnikov v treh mesecih –, dosežen majhen napredek. To je povzročilo identifikacijo in določitev kraja številnih uporabnikov AOL.

Podatki o kraju so druga skupina podatkov, ki se redko anonimizira samo z odstranitvijo identitet posameznikov, na katere se nanašajo osebni podatki, ali z delnim šifriranjem nekaterih atributov. Vzorci mobilnosti ljudi so lahko dovolj edinstveni, da lahko semantični del podatkov o kraju (kraji, v katerih je bil posameznik, na katerega se nanašajo osebni podatki, ob določenem času) tudi brez drugih atributov razkrije številne značilnosti

[25] L. Backstrom, C. Dwork in J. M. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography (Zakaj obstaja r3579x: anonimizirana družbena omrežja, prikriti vzorci in strukturna steganografija), zapiski iz 16. mednarodne konference o svetovnem spletu WWW'07, str. 181–190. (2007).

[26] http://33bits.org/2012/02/20/is-writing-style-sufficient-to-deanonymize-material-posted-online/.

[27] Zlasti pomemben primer občutljivih podatkov so genetski podatki, pri katerih obstaja tveganje za ponovno identifikacijo, če je edini mehanizem, ki se uporabi za njihovo „anonimizacijo“, odstranitev identitete darovalcev. Glej primer iz odstavka 2.2.2 zgoraj. Glej tudi John Bohannon, Genealogy Databases Enable Naming of Anonymous DNA Donors (Rodoslovne podatkovne zbirke omogočajo določitev anonimnih darovalcev DNK), Science, zv. 339, št. 6117 (18. januar 2013), str. 262.

posameznika, na katerega se nanašajo osebni podatki[28]. To je bilo večkrat dokazano v reprezentativnih znanstvenih raziskavah[29].

V zvezi s tem je treba opozoriti na uporabo psevdonimov kot načina za zagotovitev ustrezne zaščite posameznikov, na katere se nanašajo osebni podatki, pred uhajanjem identitet ali atributov. Če psevdonimizacija temelji na nadomestitvi identitete z drugo edinstveno kodo, je domneva, da to pomeni zanesljivo deidentifikacijo, preprosta ter ne upošteva zapletenosti identifikacijskih metodologij in različnih okoliščin, v katerih jih je mogoče uporabiti.

A.3 „Anonimizacija“ z generalizacijo

Pristop, ki temelji na generalizaciji atributa, je mogoče pojasniti s preprostim primerom.

Vzemimo primer, ko se upravljavec podatkov odloči za objavo preproste preglednice, ki vsebuje tri informacije ali atribute: identifikacijsko številko, ki je edinstvena za vsak zapis; opredelitev kraja, ki posameznika, na katerega se nanašajo osebni podatki, povezuje s krajem, v katerem živi; in opredelitev lastnosti, iz katere je razvidna lastnost navedenega posameznika, na katerega se nanašajo osebni podatki; predpostavljamo pa lahko tudi, da je ta lastnost ena od dveh različnih vrednosti s splošnima oznakama {P1, P2}:

Serijska številka	**Identifikator kraja**	**Lastnost**
#1	Rim	P1
#2	Madrid	P1
#3	London	P2
#4	Pariz	P1
#5	Barcelona	P1
#6	Milano	P2
#7	New York	P2
#8	Berlin	P1

Preglednica A1. Primer posameznikov, na katere se nanašajo osebni podatki, zbranih po kraju ter lastnostih P1 in P2.

Če je nekdo, ki se imenuje napadalec, že vnaprej seznanjen z podatkom, da je določen posameznik, na katerega se nanašajo osebni podatki (ciljna oseba) in ki živi v Milanu, vključen v preglednico, potem po pregledu preglednice lahko izve, da ima ta posameznik tudi lastnost P2, ker je #6 edini posameznik, na katerega se nanašajo osebni podatki, z navedenim identifikatorjem kraja.

Ta zelo preprost primer kaže glavne elemente vsakega identifikacijskega postopka, ki se uporabi za nabor podatkov, za katere naj bi bil domnevno uporabljen anonimizacijski postopek. Obstaja namreč napadalec, ki je (slučajno ali namerno) imel osnovne informacije o nekaterih ali vseh posameznikih, na katere se nanašajo osebni podatki, v naboru podatkov.

[28] To vprašanje je obravnavano v nekaterih nacionalnih zakonodajah. Na primer, v Franciji se objavljeni statistični podatki o kraju anonimizirajo z generalizacijo in permutacijo. INSEE zato objavi statistične podatke, ki so generalizirani z združevanjem vseh podatkov na območju 40 000 kvadratnih metrov. Nabor podatkov je dovolj razdrobljen, da se ohrani uporabnost podatkov, permutacije pa preprečijo napade za deanonimizacijo na redko poseljenih območjih. Združevanje te skupine podatkov in njihova permutacija na splošno zagotavljata trdna zagotovila za zaščito proti napadom s sklepanjem in napadom za deanonimizacijo (http://www.insee.fr/en/).

[29] de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M. in Blondel, V. D., Unique in the Crowd: The privacy bounds of human mobility, (Edinstveni v množici: Meje zasebnosti pri mobilnosti ljudi). Nature. 3, 1376 (2013).

Cilj napadalca je te osnovne informacije povezati s podatki v objavljenem naboru podatkov, da bi pridobil jasnejšo sliko o značilnostih navedenih posameznikov, na katere se nanašajo osebni podatki.

Upravljavec podatkov bi se lahko za to, da bi bilo povezovanje s katerimi koli osnovnimi informacijami manj učinkovito ali manj neposredno, osredotočil na identifikator kraja in tako mesto, v katerem živijo posamezniki, na katere se nanašajo osebni podatki, nadomestil s širšim območjem, kot je na primer država. Preglednica bi bila v tem primeru naslednja.

Serijska številka	**Identifikator kraja**	**Lastnost**
#1	Italija	P1
#2	Španija	P1
#3	Združeno kraljestvo	P2
#4	Francija	P1
#5	Španija	P1
#6	Italija	P2
#7	ZDA	P2
#8	Nemčija	P1

Preglednica A2. Generalizacija preglednice A1 po nacionalnosti.

Napadalcu njegove osnovne informacije o identificiranem posamezniku, na katerega se nanašajo osebni podatki (npr. „ciljna oseba živi v Rimu in je v preglednici“), v skladu s tem novim združevanjem podatkov ne omogočajo, da bi sprejel kakršno koli jasno ugotovitev o njegovi lastnosti: ker imata oba Italijana v preglednici različne lastnosti, P1 oziroma P2. Napadalcu ostane 50-odstotna negotovost glede lastnosti ciljnega subjekta. To je preprost primer, ki kaže učinek generalizacije na izvajanje anonimizacije. Čeprav je mogoče s to generalizacijsko potezo učinkovito prepoloviti verjetnost za določitev ciljne osebe iz Italije, pa te poteze ni mogoče učinkovito uporabiti za ciljno osebo iz drugih krajev (npr. iz ZDA).

Poleg tega lahko napadalec še vedno izve informacije o ciljni osebi iz Španije. Če ima napadalec osnovno informacijo, kot je „ciljna oseba živi v Madridu in je v preglednici“ ali „ciljna oseba živi v Barceloni in je v preglednici“, lahko s popolno gotovostjo sklepa, da ima ciljna oseba lastnost P1. Generalizacija torej ne more zagotoviti enake ravni zasebnosti ali zaščite pred napadi s sklepanjem za celotno populacijo v naboru podatkov.

Ob takšnem razmišljanju bi bilo mogoče sklepati, da bi lahko močnejša generalizacija morda pomagala preprečiti kakršno koli povezovanje – na primer generalizacija po celinah. Preglednica bi bila v tem primeru naslednja.

Serijska številka	Identifikator kraja	Lastnost
#1	Evropa	P1
#2	Evropa	P1
#3	Evropa	P2
#4	Evropa	P1
#5	Evropa	P1
#6	Evropa	P2
#7	Severna Amerika	P2
#8	Evropa	P1

Preglednica A3. Generalizacija preglednice A1 po celinah.

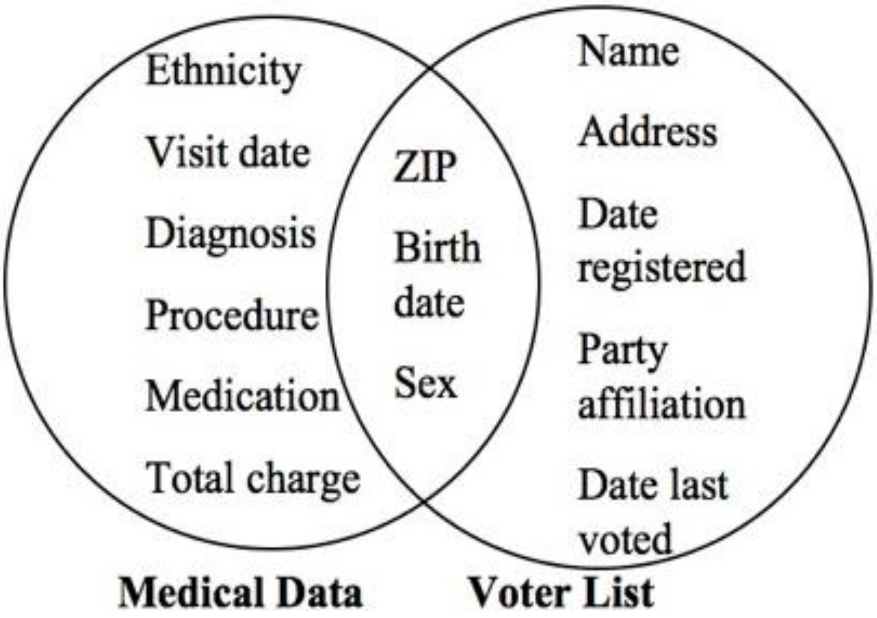
S takšnim združevanjem bi bili vsi posamezniki, na katere se nanašajo osebni podatki, iz preglednice zaščiteni pred napadi glede povezovanja in identifikacije, razen tisti, ki živijo v ZDA, vsaka osnovna informacija, kot je „ciljna oseba živi v Madridu in je v preglednici“ ali „ciljna oseba živi v Milanu in je v preglednici“, pa bi vodila do določene stopnje verjetnosti glede lastnosti, ki velja za določenega posameznika, na katerega se nanašajo osebni podatki (P1 z 71,4-odstotno verjetnostjo in P2 z 28,6-odstotno verjetnostjo), in ne do neposrednega povezovanja. Tudi ta nadaljnja generalizacija negativno vpliva na očitno in popolno izgubo informacij: preglednica ne omogoča odkritja morebitnih povezav med lastnostmi in krajem, in sicer ali morda za določen kraj obstaja večja verjetnost, da povzroči katero od lastnosti, ker zagotavlja samo tako imenovane „obrobne“ porazdelitve, tj. absolutno verjetnost pojava lastnosti P1 in P2 v populaciji (v našem primeru 62,5 % oziroma 37,5 %) in za vsako celino (kot je bilo navedeno, 71,4 % oziroma 28,6 % v Evropi ter 100 % in 0 % v Severni Ameriki).

Primer kaže tudi, da izvajanje generalizacije vpliva na dejansko uporabnost podatkov. Danes je na voljo nekaj inženirskih orodij za predhodno določitev najustreznejše ravni generalizacije atributov (tj. pred objavo nabora podatkov), da se zmanjšajo tveganja za identifikacijo posameznikov iz preglednice, na katere se nanašajo osebni podatki, brez prevelikega učinka na uporabnost objavljenih podatkov.

K-anonimnost

Poskus preprečitve napadov s povezovanjem, ki temelji na generalizaciji atributov, je znan kot k-anonimnost. Ta praksa izhaja iz poskusa ponovne identifikacije, opravljenega konec devetdesetih let prejšnjega stoletja, ko je zasebna ameriška družba iz zdravstvenega sektorja javno objavila domnevno anonimiziran nabor podatkov. Ta anonimizacija je zajemala odstranitev imen posameznikov, na katere so se nanašali osebni podatki, vendar je nabor podatkov še vedno vseboval zdravstvene podatke in druge atribute, kot so poštna številka (identifikator kraja, v katerem so živeli), spol in poln datum rojstva. Isti trojček {poštna številka, spol in polni datum rojstva} je bil vključen tudi v druge javno dostopne registre (npr. v volilni imenik), zato je lahko akademski raziskovalec z njim identiteto določenih posameznikov, na katere so se nanašali osebni podatki, povezal z atributi v objavljenem naboru podatkov. Napadalec (raziskovalec) je lahko imel naslednje osnovne informacije: „Vem, da je posameznik, na katerega se nanašajo osebni podatki, iz volilnega imenika z določenim trojčkom {poštna številka, spol, polni datum rojstva} edinstven. V objavljenem

naboru podatkov je zapis z navedenim trojčkom“. Empirično je bilo ugotovljeno[30] , da je bila velika večina (več kot 80 %) posameznikov, na katere se nanašajo osebni podatki, iz javnih registrov, uporabljenih v tem znanstvenem poskusu, nedvomno povezana z zadevnim trojčkom, ki je omogočil identifikacijo. Podatki torej tudi v tem primeru niso bili ustrezno anonimizirani.



Prikaz A1. Ponovna identifikacija s povezovanjem podatkov.

Za zmanjšanje učinkovitosti podobnih napadov s povezovanjem se je trdilo, da bi morali upravljavci najprej pregledati nabor podatkov in združiti atribute, za katere se pričakuje, da jih bo napadalec uporabil, da bi objavljeno preglednico povezal z drugim pomožnim virom; vsaka skupina bi morala vključevati vsaj *k* enakih kombinacij generaliziranih atributov (tj. morala bi predstavljati ekvivalenčni razred atributov). Nabori podatkov bi se tako morali objaviti šele po razdelitvi v takšne homogene skupine. Atributi, izbrani za generalizacijo, so v literaturi znani kot kvaziidentifikatorji, saj bi njihovo poznavanje v nešifrirani obliki pomenilo neposredno identifikacijo posameznikov, na katere se nanašajo osebni podatki.

Številni identifikacijski poskusi so pokazali slabost neustrezno zasnovanih k-anonimiziranih preglednic. To se lahko na primer zgodi, ker so drugi atributi v ekvivalenčnem razredu enaki (kot je to v ekvivalenčnem razredu španskih posameznikov, na katere se nanašajo osebni podatki, v primeru iz preglednice A2), ker je njihova porazdelitev zelo neenakomerna, ker je določeni atribut zelo razširjen, ker je število zapisov v ekvivalenčnem razredu zelo majhno, kar v obeh primerih omogoča verjetnostno sklepanje, ali ker med nešifriranimi atributi v ekvivalenčnih razredih ni pomembnih „semantičnih“ razlik (npr. kvantitativni ukrep takšnih atributov je lahko dejansko različen, numerično pa zelo podoben, ali pa lahko atributi spadajo v razpon semantično podobnih atributov, npr. isti razpon kreditnega tveganja ali ista skupina bolezenskih stanj), tako da iz nabora podatkov še vedno lahko uhaja velika količina informacij o posameznikih, na katere se nanašajo osebni podatki, za napade s povezovanjem[31]. Tukaj je treba poudariti, da je v primeru, če so podatki razpršeni (če se na

[30] L. Sweeney. Weaving Technology and Policy Together to Maintain Confidentiality (Povezovanje tehnologije in politike za ohranitev zaupnosti). Journal of Law, Medicine & Ethics, 25, št. 2 in 3 (1997): str. 98–110.

[31] Poudariti je treba, da je mogoče povezave določiti tudi po združitvi zapisov podatkov v skupine po atributih. Če upravljavec podatkov pozna vrste povezav, ki jih želi preveriti, lahko izbere atribute, ki so najpomembnejši. Na primer, rezultati raziskave PEW niso predmet napadov s podrobnim sklepanjem in so še vedno zelo uporabni

primer določena lastnost na geografskem območju redko pojavlja) in s prvim združevanjem ni mogoče združiti podatkov z zadostnim številom primerov različnih lastnosti (če je na primer na geografskem območju še vedno mogoče krajevno določiti malo primerov redkih lastnosti), potrebno dodatno združevanje atributov, da se doseže ciljno usmerjena anonimizacija.

L-raznolikost

V preteklosti so bile na podlagi teh ugotovitev predlagane različice k-anonimnosti in nekatera tehnična merila za izboljšanje izvajanja anonimizacije z generalizacijo, katerih cilj je zmanjšanje tveganj napadov s povezovanjem. Temeljijo na verjetnih lastnostih naborov podatkov. Natančneje, 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. To pomeni enako, kot če bi rekli, da bi moral nabor podatkov (ali razdelek) imeti minimalno število primerov izbrane lastnosti: s to potezo se morda zmanjša tveganje ponovne identifikacije. To je cilj anonimizacijskega postopka z l-raznolikostjo. Primer tega postopka je naveden v preglednicah A4 (prvotni podatki) in A5 (rezultati obdelave podatkov). Očitno je, da se z generalizacijo atributov ob ustrezni obdelavi identifikatorja kraja in starosti posameznikov v preglednici A4 znatno poveča negotovost v zvezi z dejanskimi atributi vsakega posameznika, na katerega se nanašajo osebni podatki, iz raziskave. Na primer, tudi če napadalec ve, da posameznik, na katerega se nanašajo osebni podatki, spada v prvi ekvivalenčni razred, ne more nadalje ugotoviti, ali ima oseba lastnost X, Y ali Z, ker v navedenem razredu (in v vsakem drugem ekvivalenčnem razredu) obstaja vsaj en zapis s takšno lastnostjo.

Serijska številka	Identifikator kraja	Starost	Lastnost
1	111	38	X
2	122	39	X
3	122	31	Y
4	111	33	Y
5	231	60	Z
6	231	65	X
7	233	57	Y
8	233	59	Y
9	111	41	Z
10	111	47	Z
11	122	46	Z
12	122	45	Z

Preglednica A4. Preglednica s posamezniki, združenimi po kraju, starosti in treh lastnostih: X, Y in Z.

za ugotavljanje povezav med demografskimi značilnostmi in interesi (http://www.pewinternet.org/Reports/2013/Anonymity-online.aspx).

Serijska številka	Identifikator kraja	Starost	Lastnost
1	11*	< 50	X
4	11*	< 50	Y
9	11*	< 50	Z
10	11*	< 50	Z
5	23*	> 50	Z
6	23*	> 50	X
7	23*	> 50	Y
8	23*	> 50	Y
2	12*	< 50	X
3	12*	< 50	Y
11	12*	< 50	Z
12	12*	< 50	Z

Preglednica A5. Primer preglednice A4 v različici z 1-raznolikostjo.

T-podobnost:

Poseben primer atributov v razdelku, ki so neenakomerno porazdeljeni ali pripadajo majhnemu razponu vrednosti ali semantičnih pomenov, je obravnavan s pristopom, imenovanim t-podobnost. To je nadaljnje izboljšanje anonimizacije z generalizacijo, ki v praksi pomeni ureditev podatkov, da se čim bolj dosežejo ekvivalenčni razredi, ki upoštevajo prvotno porazdelitev atributov v prvotnem naboru podatkov. Za to se uporabi naslednji dvostopenjski postopek. Preglednica A6 je prvotna podatkovna zbirka, ki vključuje nešifrirane zapise posameznikov, na katere se nanašajo osebni podatki, združene po kraju, starosti, plači in dveh skupinah semantično podobnih lastnosti: (X1, X2, X3) oziroma (Y1, Y2, Y3) (npr. razredi podobnih kreditnih tveganj, podobne bolezni). Najprej se za preglednico izvede *l-raznolikost*, pri čemer je $l = 1$ (preglednica A7), kar je mogoče doseči z združitvijo zapisov v semantično podobne ekvivalenčne razrede in neustrezno ciljno usmerjeno anonimizacijo; nato se podatki obdelajo, da se pridobi t-podobnost (preglednica A8) in večja variabilnost v vsakem razdelku. V drugi fazi se v vsak ekvivalenčni razred dejansko vključijo zapisi iz obeh skupin lastnosti. Opozoriti je treba, da imata identifikatorja za kraj in starost v različnih stopnjah postopka različno razdrobljenost: po pomeni, da se lahko za vsak atribut zahtevajo različna merila generalizacije, da se zagotovi ciljno usmerjena anonimizacija, to pa od upravljavcev podatkov zahteva posebna tehnična prizadevanja in ustrezne izračune.

Serijska številka	Identifikator kraja	Starost	Plača	Lastnost
1	1127	29	30K	X1
2	1112	22	32K	X2
3	1128	27	35K	X3
4	1215	43	50K	X2
5	1219	52	120K	Y1
6	1216	47	60K	Y2
7	1115	30	55K	Y2
8	1123	36	100K	Y3
9	1117	32	110K	X3

Preglednica A6. Preglednica s posamezniki, združenimi po kraju, starosti, plačah in dveh skupinah lastnosti.

Serijska številka	Identifikator kraja	Starost	Plača	Lastnost
1	11**	2*	30K	X1
2	11**	2*	32K	X2
3	11**	2*	35K	X3
4	121*	> 40	50K	X2
5	121*	> 40	120K	Y1
6	121*	> 40	60K	Y2
7	11**	3*	55K	Y2
8	11**	3*	100K	Y3
9	11**	3*	110K	X3

Preglednica A7. Različica preglednice A6 z *l*-raznolikostjo.

Serijska številka	Identifikator kraja	Starost	Plača	Lastnost
1	112*	< 40	30K	X1
3	112*	< 40	35K	X3
8	112*	< 40	100K	Y3
4	121*	> 40	50K	X2
5	121*	> 40	120K	Y1
6	121*	> 40	60K	Y2
2	111*	< 40	32K	X2
7	111*	< 40	55K	Y2
9	111*	< 40	110K	X3

Preglednica A8. Različica preglednice A6 s *t*-podobnostjo.

Jasno je treba poudariti, da je mogoče cilj generalizacije atributov posameznikov, na katere se nanašajo osebni podatki, s tako obzirnostjo včasih doseči samo za majhno število zapisov in ne vse od njih. Dobre prakse bi morale zagotoviti, da vsak ekvivalenčni razred vsebuje več posameznikov in da napad s sklepanjem ni več mogoč. Ta pristop vsekakor zahteva, da morajo upravljavci podatkov opraviti poglobljeno oceno razpoložljivih podatkov, vključno s kombinatorično oceno različnih drugih možnosti (npr. amplitude različnega razpona, različna krajevna ali starostna razdrobljenost itd.). Z drugimi besedami, upravljavci podatkov anonimizacije z generalizacijo ne morejo preprosto doseči v prvem poskusu, v katerem želijo analitične vrednosti atributov v zapisu nadomestiti z razponi, saj so za to potrebni nekoliko natančneje opredeljeni kvantitativni pristopi – kot sta ocena entropije atributov v vsakem razdelku ali merjenje razdalje med prvotno porazdelitvijo atributov in porazdelitvijo v vsakem ekvivalenčnem razredu.