Datu publicēšanas platformas pilnveidošana saskaņā ar tehnisko specifikāciju

Rekomendācijas atvērto datu atlases kritērijiem un publicēšanai paredzēto datu depersonificēšanu

V. 1.0.1

VRAA.DPP.PS.1.0.1.2019.08.15

Dokumenta autortiesības

Šo dokumentu Valsts reģionālās attīstības aģentūras (VRAA) uzdevumā ir izstrādājis A/S "RIX Technologies" projekta „Vienotā datu telpa”, Nr. 2.2.1.1/17/I/009, ietvaros. Dokuments ir izmantojams saskaņā ar 2018. gada 4. septembra līguma ID Nr. 13-7/18/233 „Datu publicēšanas platformas pilnveidošana” nosacījumiem.

Kontaktpersonas

R. Prikulis

A/S "RIX Technologies"

Projektu vadītājs

Blaumaņa iela 5a

Rīga, LV 1011

Tālr.: +371 67142990

Fakss: +371 67142991

E-pasts: [ritvars.prikulis@rixtech.lv](mailto:ritvars.prikulis@rixtech.lv)

WWW: [www.rixtech.lv](http://www.rixtech.lv/)

I. Kovkājeva

Valsts reģionālās attīstības aģentūras

Informācijas sistēmu attīstības departamenta

Pārvaldības nodaļas

Elektronisko pakalpojumu un datu apmaiņu sektora projektu vadītāja

Alberta iela 10

Rīga, LV-1010

Tālr.: +371 66164650,

Fakss: +371 67079001

E-pasts: [inga.kovkajeva@vraa.gov.lv](mailto:inga.kovkajeva@vraa.gov.lv)

WWW:

<http://www.vraa.gov.lv>

Saturs

[1 Ievads 4](#_Toc17099783)

[1.1 Dokumenta nolūks 4](#_Toc17099784)

[1.2 Saīsinājumi 4](#_Toc17099785)

[1.3 Definīcijas par personas datiem un to apstrādi 5](#_Toc17099786)

[1.4 Saistītie dokumenti 8](#_Toc17099787)

[1. Tehniskā analīze, tehnoloģiju noturība un tipiskākās kļūdas 10](#_Toc17099788)

[1.1. Nejauša izlase 11](#_Toc17099789)

[1.1.1. Trokšņa pievienošana 11](#_Toc17099790)

[1.1.2. Permutācija 13](#_Toc17099791)

[1.1.3. Diferencēts privātums 14](#_Toc17099792)

[1.2. Vispārināšana 16](#_Toc17099793)

[1.2.1. Agregācija un K-anonimitāte 16](#_Toc17099794)

[1.2.2. Daudzveidība / pietuvinājums 18](#_Toc17099795)

[2. Pseidonimizācija 21](#_Toc17099796)

[2.1. Garantijas 22](#_Toc17099797)

[2.2. Biežāk pieļautās kļūdas 23](#_Toc17099798)

[2.3. Pseidonimizācijas trūkumi 24](#_Toc17099799)

[3. Secinājumi un rekomendācijas 25](#_Toc17099800)

[3.1. Secinājumi 25](#_Toc17099801)

[3.2. Rekomendācijas 25](#_Toc17099802)

[1.4.1 Labā depersonificēšanas prakse 26](#_Toc17099803)

# Ievads

## Dokumenta nolūks

Rekomendācijas atvērto datu atlases kritērijiem un publicēšanai paredzēto datu depersonificēšanu atbilstoši normatīvajiem aktiem par personas datu aizsardzību.

## Saīsinājumi

Dokumentā izmantotie saīsinājumi ir paskaidroti tabulā Definīcijas un saīsinājumi.

Tabula 1: Definīcijas un saīsinājumi

|  |  |
| --- | --- |
| **Definīcija vai saīsinājums** | **Skaidrojums** |
| DPP | Datu publicēšanas platforma |
| Entropija | Informācijas teorijas pamatjēdziens, kas kvantitatīvi raksturo gadījumlieluma nenoteiktību. |
| Jaucējfunkcija | Algoritms, kas pārveido dažāda garuma tekstu fiksēta garuma izvadē. |
| K-anonimitāte | K-anonimitāte ir datu kopas īpašība, ko parasti izmanto, lai aprakstītu datu kopas anonimitātes līmeni. Datu kopa ir k-anonīma, ja katra identitāti atklājošo īpašību kombinācija notiek vismaz k dažādās datu kopas rindās. |
| Kvaziidentifikators | Kvaziidentifikatori ir informācijas vienības, kas pašas par sevi nav unikāli identifikatori, bet ir pietiekami labi saistīti ar datu kopu, lai tos varētu apvienot ar citiem kvazidentifikatoriem, tādējādi izveidojot unikālu identifikatoru. |
| L-diversity | L-diversity ir uz grupām balstīta anonimizācijas forma, kuru izmanto, lai saglabātu privātumu datu kopās, samazinot datu attēlojuma detalizāciju. |
| Projekts | Datu publicēšanas platformas pilnveidošana saskaņā ar tehnisko specifikāciju |
| Pasūtītājs, VRAA | Valsts reģionālās attīstības aģentūra |
| VARAM | Vides aizsardzības un reģionālās attīstības ministrija |
| Piegādātājs, RIX | A/S "RIX Technologies" |
| Salt | Salt kriptogrāfijā norāda uz nejaušu virkni, kas tiek pievienota dotajam vienkāršajam tekstam pirms tā turpmākas apstrādes (piemēram, papildus vētība jaucējfunkcijai), lai palielinātu ievades entropiju. To bieži izmanto paroļu glabāšanai un pārsūtīšanai, lai palielinātu informācijas drošību. |
| Sistēma | Datu publicēšanas platforma |
| T-closeness | T-closeness ir uz L-diversity balstītas anonimizācijas turpmāka pilnveidošana, ko izmanto, lai saglabātu privātumu datu kopās, samazinot datu attēlojuma precizitāti. Šis samazinājums ir kompromiss, kura rezultātā tiek zaudēta datu pārvaldības vai datu ieguves algoritmu efektivitāte, lai iegūtu zināmu privātumu. T-closeness modelis paplašina L-diversity modeli, atšķirīgi apstrādājot atribūta vērtības, ņemot vērā šī atribūta datu vērtību sadalījumu. |

## Definīcijas par personas datiem un to apstrādi

1. “personas dati” ir jebkura informācija, kas attiecas uz identificētu vai identificējamu fizisku personu (“datu subjekts”); identificējama fiziska persona ir tāda, kuru var tieši vai netieši identificēt, jo īpaši atsaucoties uz identifikatoru, piemēram, minētās personas vārdu, uzvārdu, identifikācijas numuru, atrašanās vietas datiem, tiešsaistes identifikatoru vai vienu vai vairākiem minētajai fiziskajai personai raksturīgiem fiziskās, fizioloģiskās, ģenētiskās, garīgās, ekonomiskās, kultūras vai sociālās identitātes faktoriem;
2. “apstrāde” ir jebkura ar personas datiem vai personas datu kopumiem veikta darbība vai darbību kopums, ko veic ar vai bez automatizētiem līdzekļiem, piemēram, vākšana, reģistrācija, organizēšana, strukturēšana, glabāšana, pielāgošana vai pārveidošana, atgūšana, aplūkošana, izmantošana, izpaušana, nosūtot, izplatot vai citādi darot tos pieejamus, saskaņošana vai kombinēšana, ierobežošana, dzēšana vai iznīcināšana;
3. “apstrādes ierobežošana” ir uzglabātu personas datu iezīmēšana ar mērķi ierobežot to apstrādi nākotnē;
4. “profilēšana” ir jebkura veida automatizēta personas datu apstrāde, kas izpaužas kā personas datu izmantošana nolūkā izvērtēt konkrētus ar fizisku personu saistītus personiskus aspektus, jo īpaši analizēt vai prognozēt aspektus saistībā ar minētās fiziskās personas sniegumu darbā, ekonomisko situāciju, veselību, personīgām vēlmēm, interesēm, uzticamību, uzvedību, atrašanās vietu vai pārvietošanos;
5. “pseidonimizācija” ir personas datu apstrāde, ko veic tādā veidā, lai personas datus vairs nav iespējams saistīt ar konkrētu datu subjektu bez papildu informācijas izmantošanas, ar noteikumu, ka šāda papildu informācija tiek turēta atsevišķi un tai piemēro tehniskus un organizatoriskus pasākumus, lai nodrošinātu, ka personas dati netiek saistīti ar identificētu vai identificējamu fizisku personu;
6. “kartotēka” ir jebkurš strukturēts personas datu kopums, kas ir pieejams saskaņā ar konkrētiem kritērijiem, neatkarīgi no tā, vai datu kopums ir centralizēts, decentralizēts vai izkliedēts, pamatojoties uz funkcionālu vai ģeogrāfisku motivāciju;
7. “pārzinis” ir fiziska vai juridiska persona, publiska iestāde, aģentūra vai cita struktūra, kas viena pati vai kopīgi ar citām nosaka personas datu apstrādes nolūkus un līdzekļus; ja šādas apstrādes nolūkus un līdzekļus nosaka ar Savienības vai dalībvalsts tiesību aktiem, pārzini vai tā iecelšanas konkrētos kritērijus var paredzēt Savienības vai dalībvalsts tiesību aktos;
8. “apstrādātājs” ir fiziska vai juridiska persona, publiska iestāde, aģentūra vai cita struktūra, kura pārziņa vārdā apstrādā personas datus;
9. “saņēmējs” ir fiziska vai juridiska persona, publiska iestāde, aģentūra vai cita struktūra, kurai izpauž personas datus – neatkarīgi no tā, vai tā ir trešā persona vai nav. Tomēr publiskas iestādes, kas var saņemt personas datus saistībā ar konkrētu izmeklēšanu saskaņā ar Savienības vai dalībvalsts tiesību aktiem, netiek uzskatītas par saņēmējiem; minēto datu apstrāde, ko veic minētās publiskās iestādes, atbilst piemērojamiem datu aizsardzības noteikumiem saskaņā ar apstrādes nolūkiem;
10. “trešā persona” ir fiziska vai juridiska persona, publiska iestāde, aģentūra vai struktūra, kura nav datu subjekts, pārzinis, apstrādātājs un personas, kuras pārziņa vai apstrādātāja tiešā pakļautībā ir pilnvarotas apstrādāt personas datus;
11. datu subjekta “piekrišana” ir jebkura brīvi sniegta, konkrēta, apzināta un viennozīmīga norāde uz datu subjekta vēlmēm, ar kuru viņš paziņojuma vai skaidri apstiprinošas darbības veidā sniedz piekrišanu savu personas datu apstrādei;
12. “personas datu aizsardzības pārkāpums” ir drošības pārkāpums, kura rezultātā notiek nejauša vai nelikumīga nosūtīto, uzglabāto vai citādi apstrādāto personas datu iznīcināšana, nozaudēšana, pārveidošana, neatļauta izpaušana vai piekļuve tiem;
13. “ģenētiskie dati” ir personas dati, kas attiecas uz fiziskas personas pārmantotajām vai iegūtajām ģenētiskajām pazīmēm, sniedz unikālu informāciju par minētās fiziskās personas fizioloģiju vai veselību un kas izriet jo īpaši no attiecīgās fiziskās personas bioloģiskā parauga analīzes;
14. “biometriskie dati” ir personas dati pēc specifiskas tehniskas apstrādes, kuri attiecas uz fiziskas personas fiziskajām, fizioloģiskajām vai uzvedības pazīmēm, kas ļauj veikt vai apstiprina minētās fiziskās personas unikālu identifikāciju, piemēram, sejas attēli vai daktiloskopijas dati;
15. “veselības dati” ir personas dati, kas saistīti ar fiziskas personas fizisko vai garīgo veselību, tostarp veselības aprūpes pakalpojumu sniegšanu, un kas atspoguļo informāciju par tās veselības stāvokli;
16. “galvenā uzņēmējdarbības vieta”:
    1. attiecībā uz pārzini, kura uzņēmējdarbības vietas atrodas vairāk nekā vienā dalībvalstī – tā galvenās pārvaldes atrašanās vieta Savienībā, ja vien lēmumi par personas datu apstrādes nolūkiem un līdzekļiem netiek pieņemti citā pārziņa uzņēmējdarbības vietā Savienībā un ja vien šai citai uzņēmējdarbības vietai nav šādu lēmumu īstenošanas pilnvaru – tādā gadījumā par galveno uzņēmējdarbības vietu uzskata to uzņēmējdarbības vietu, kurā pieņemti šādi lēmumi;
    2. attiecībā uz apstrādātāju, kura uzņēmējdarbības vietas atrodas vairāk nekā vienā dalībvalstī – tā galvenās pārvaldes atrašanās vieta Savienībā, vai, ja apstrādātājam nav galvenās pārvaldes Savienībā, – apstrādātāja uzņēmējdarbības vieta Savienībā, kur notiek galvenās apstrādes darbības saistībā ar apstrādātāja uzņēmējdarbības vietas darbībām, ciktāl uz apstrādātāju attiecas šajā regulā paredzētie konkrētie pienākumi;
17. “pārstāvis” ir fiziska vai juridiska persona, kura veic uzņēmējdarbību Savienībā un kuru pārzinis vai apstrādātājs ir iecēlis rakstiski saskaņā ar 27. pantu, un kura pārstāv pārzini vai apstrādātāju saistībā ar to attiecīgajiem pienākumiem, kas paredzēti šajā regulā;
18. “uzņēmums” ir fiziska vai juridiska persona, kas veic saimniecisku darbību, neatkarīgi no tās juridiskā statusa, tostarp partnerības vai apvienības, kas regulāri veic saimniecisku darbību;
19. „uzņēmumu grupa” ir kontrolējošais uzņēmums un tā kontrolētie uzņēmumi;
20. “saistošie uzņēmuma noteikumi” ir personas datu aizsardzības vadlīnijas, ko stingri ievēro pārzinis vai apstrādātājs, kas veic uzņēmējdarbību dalībvalsts teritorijā, attiecībā uz personas datu nosūtīšanu vai vairākkārtēju nosūtīšanu pārzinim vai apstrādātājam vienā vai vairākās trešās valstīs uzņēmumu grupā vai uzņēmējsabiedrību grupā, kas iesaistīta kopīgā saimnieciskā darbībā;
21. “uzraudzības iestāde” ir neatkarīga publiska iestāde, ko dalībvalsts izveidojusi, ievērojot 51. pantu;
22. “attiecīgā uzraudzības iestāde” ir uzraudzības iestāde, uz kuru personas datu apstrāde attiecas tāpēc, ka:
    1. pārzinis vai apstrādātājs veic uzņēmējdarbību minētās uzraudzības iestādes dalībvalsts teritorijā;
    2. apstrāde būtiski ietekmē vai var būtiski ietekmēt datu subjektus, kas dzīvo minētās uzraudzības iestādes dalībvalstī; vai
    3. sūdzība ir iesniegta minētajai uzraudzības iestādei;
23. “pārrobežu apstrāde” ir vai nu:
    1. personas datu apstrāde, kas notiek saistībā ar darbībām, kuras Savienībā veic pārziņa vai apstrādātāja uzņēmējdarbības vietās vairāk nekā vienā dalībvalstī, ja pārzinis vai apstrādātājs veic uzņēmējdarbību vairāk nekā vienā dalībvalstī; vai
    2. personas datu apstrāde, kas notiek saistībā ar darbībām, kuras Savienībā veic pārziņa vai apstrādātāja vienīgajā uzņēmējdarbības vietā, bet kas būtiski ietekmē vai var būtiski ietekmēt datu subjektus vairāk nekā vienā dalībvalstī;
24. “būtisks un motivēts iebildums” ir iebildums lēmuma projektam par to, vai šī regula ir pārkāpta, vai par to, vai iecerētā darbība attiecībā uz pārzini vai apstrādātāju atbilst šai regulai, skaidri parādot to risku nozīmīgumu, ko lēmuma projekts rada datu subjektu pamattiesībām un pamatbrīvībām un attiecīgā gadījumā – personas datu brīvai apritei Savienībā;
25. “informācijas sabiedrības pakalpojums” ir pakalpojums, kā definēts Eiropas Parlamenta un Padomes Direktīvas (ES) 2015/1535 1. panta 1. punkta b) apakšpunktā;
26. “starptautiska organizācija” ir organizācija un tai pakārtotas struktūras, kas ir starptautisko publisko tiesību subjekti, vai jebkura cita struktūra, kas ir izveidota ar divu vai vairāku valstu nolīgumu vai uz tā pamata.

## Saistītie dokumenti

Šis dokuments ir veidots balstoties uz šiem dokumentiem:

1. Tehniskā specifikācija. Tehniskā specifikācija pakalpojumu iegādei no EIS kataloga pozīcijas CI110P.18.1. „Drupal pakalpojumi” „Datu publicēšanas platformas pilnveidošana”. VARAM un SIA “Datu Sistēmas”, 2018 (04.09.2018. Līguma Nr. 13-7/18/233 “Līgums par DARBA UZDEVUMA “Datu publicēšanas platformas pilnveidošana saskaņā ar tehnisko specifikāciju” izpildi”, Pielikums Nr.1).
2. EIROPAS PARLAMENTA UN PADOMES REGULA (ES) 2016/679 “par fizisku personu aizsardzību attiecībā uz personas datu apstrādi un šādu datu brīvu apriti un ar ko atceļ Direktīvu 95/46/EK (Vispārīgā datu aizsardzības regula)”, 2016. gada 27. aprīlis.
3. Opinion 05/2014 on Anonymisation Techniques, 2014. gada 10. aprīlis, <https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf>.
4. GUIDE TO BASIC DATA ANONYMISATION TECHNIQUES, 2018. gada 25.janvāris, <https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf>.
5. PSEUDONYMIZATION VS. ANONYMIZATION AND HOW THEY HELP WITH GDPR, 2017. gada 5. janvāris, <https://www.protegrity.com/pseudonymization-vs-anonymization-help-gdpr/>
6. Differential Privacy for Collaborative Security, <http://www.cis.upenn.edu/~bcpierce/papers/eurosec2010.pdf>.
7. Arvind Narayanan, Vitaly Shmatikov: Robust De-anonymization of Large Sparse Datasets. IEEE Symposium on Security and Privacy 2008.
8. L. Backstrom, C. Dwork, and J. M. Kleinberg. Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography.
9. John Bohannon, Genealogy Databases Enable Naming of Anonymous DNA Donors, Science, Vol. 339, No. 6117 (18 January 2013), p. 262

# Tehniskā analīze, tehnoloģiju noturība un tipiskākās kļūdas

Pastāv dažādas depersonificēšanas prakses un metodes, kuras dažādos veidos risina datu depersonificēšanas jautājumu.

Šajā sadaļā tiks aplūkoti galvenie jautājumi, kas datu pārvaldītājiem un apstrādātājiem jāņem vērā, piemērojot šīs metodes, izvērtējot garantijas, ko tās sniedz, ņemot vērā pašreizējo tehnoloģiju stāvokli un trīs depersonificēšanai būtiskus riskus:

* *Izšķiršana* – iespēja izšķirt dažus vai visus ierakstus, kas identificē personas datu kopu.
* *Saistāmība* – spēja sasaistīt vismaz divus ierakstus, kas attiecas uz vienu un to pašu datu subjektu vai datu subjektu grupu (vienā vai vairākās dažādās datubāzēs). Ja uzbrucējs var konstatēt (piemēram, ar korelācijas analīzes palīdzību), ka divi ieraksti ir piešķirti vienai un tai pašai personu grupai, bet nevar izšķirt atsevišķus indivīdus šajā grupā, šī metode nodrošina pretestību pret “izšķiršanu”, bet ne pret saistāmību.
* *Secināšana* – iespēja ar ievērojamu varbūtību izsecināt atribūta vērtību no citu atribūtu vērtību kopas.

Tādējādi risinājums pret šiem trim riskiem būtu drošs pret identificēšanu ar visiem iespējamiem saprātīgiem līdzekļiem, kurus datu apstrādātājs un jebkura trešā persona var izmantot. Depersonificēšana ir nepārtrauktas izpētes temats, un šāda izpēte ir konsekventi parādījusi, ka nav nevienas metodes, kurā nebūtu neviena trūkuma. Vispārīgi runājot, ir divas atšķirīgas pieejas depersonificēšanai: pirmā ir balstīta uz **nejaušu izlasi** (*randomization*), un otrā – uz **vispārināšanu**. Aplūkoti arī citi jēdzieni kā pseidonimizācija, diferencēts privātums, daudzveidība, pietuvinājums.

Aprakstā izmantoti arī šādi jēdzieni – datu kopa sastāv no dažādiem ierakstiem, kas attiecas uz indivīdiem (datu subjektiem). Katrs ieraksts ir saistīts ar vienu datu subjektu un sastāv no vērtību kopas (vai “ierakstiem”, piemēram, 2013. gads) katram atribūtam (piemēram, gadam). Datu kopa ir ierakstu krājums, ko var veidot kā tabulu (vai tabulu kopumu), vai kā anotētu / svērtu grafiku, kas tiek pielietots arvien biežāk. Sniegtie piemēri attiecas uz tabulām, bet tie ir piemērojami arī grafiskiem ierakstu attēlojumiem. Datu subjekta vai datu subjektu grupas atribūtu kombinācijas var tikt attiecinātas uz kvaziidentifikatoriem. Dažos gadījumos datu kopā var būt dažādi ieraksti par vienu un to pašu personu. “Uzbrucējs” ir trešā persona (t.i., ne datu pārzinis vai datu apstrādātājs), kurš oriģinālajiem ierakstiem piekļūst nejauši vai tīši.

## Nejauša izlase

Nejauša izlase (*randomization*) ir metožu kopums, kas maina datu ticamību, noņemot ciešo saikni starp personu un tās datiem. Ja dati ir pietiekami neskaidri, tos vairs nevar sasaistīt ar noteiktu personu. Randomizācija pati par sevi nemazina katra ieraksta vienreizīgumu, dati joprojām tiek iegūti no viena datu subjekta, bet tā var aizsargāt pret uzbrukumiem/riskiem, un to var kombinēt ar vispārināšanas metodēm, lai nodrošinātu spēcīgākas privātuma aizsardzības garantijas. Lai nodrošinātu, ka ieraksts nevar identificēt personu, var būt nepieciešamas arī vēl papildu metodes.

### Trokšņa pievienošana

Trokšņa pievienošanas metode ir īpaši noderīga, ja atribūtiem ir nozīmīga ietekme uz personu. Metode ietver atribūtu modificēšanu datu kopā tā, lai tie būtu mazāk precīzi, vienlaikus saglabājot vispārējo sadalījumu. Datu kopas novērotājs (pētītājs) pieņem, ka dati ir patiesi, bet tas tā ir tikai līdz noteiktam līmenim. Piemēram, ja personas auguma garums sākotnēji tika mērīts līdz tuvākajam centimetram, tad depersonificētā datu kopa var saturēt garumu, kas ir precīzs tikai +/-10 cm. Pielietojot šo metodi, faktiski trešā persona nevarēs identificēt personu, kā arī nevarēs salabot datus vai noskaidrot, kā dati ir modificēti.

Trokšņa pievienošanas metodi būtu ieteicams apvienot ar citām depersonificēšanas metodēm, piemēram, acīmredzamo atribūtu un kvaziidentifikatoru noņemšanu. Trokšņa līmenim vajadzētu būt atkarīgam no informācijas apjoma un šīs informācijas ietekmes uz indivīdu privātumu, aizsargāto atribūtu atklāšanas gadījumā.

**Kad lietot šo metodi?**

Metodi pielieto priekš kvaziidentifikatoriem (parasti skaitļiem un datumiem), kas potenciāli var tikt identificēti kombinējot ar citiem datu avotiem un kur nelielas izmaiņas datu vērtībā ir pieņemamas. Šo metodi nedrīkst izmantot, ja datu precizitāte ir ļoti svarīga.

#### Garantijas

* *Izšķiršana* – joprojām ir iespējams izšķirt personas datus (varbūt neidentificējamā veidā), pat ja ieraksti ir mazāk ticami.
* *Saistāmība* – joprojām ir iespēja sasaistīt viena indivīda ierakstus, bet ieraksti ir mazāk ticami, un tādejādi reālu ierakstu var sasaistīt ar mākslīgi pievienotu (t.i., “troksni”). Jāsaprot gan, ka dažos gadījumos nepareizs attiecinājums var pakļaut datu subjektu nozīmīgam un pat augstākam riska līmenim nekā pareizs.
* *Secināšana* – iespējamo secināšanas mēģinājumu rezultātu līmenis būs zemāks, un ir ticami iegūt viltus pozitīvus (vai viltus negatīvus) rezultātus.

#### Biežāk pieļautās kļūdas

* *Nekonsekventa trokšņa pievienošana* – ja troksnis nav semantiski dzīvotspējīgs (t.i., tas ir “ārpus mēroga” un nerespektē loģiku starp atribūtiem kopā), tad uzbrucējs, piekļūstot datubāzei, varēs izfiltrēt troksni un atsevišķos gadījumos arī reģenerēt trūkstošos ierakstus. Turklāt, ja datu kopā ir pārāk izkliedēta, pastāv iespēja sasaistīt ar troksni apstrādātos ierakstus ar ārēju avotu.
* *Pieņēmums, ka pietiek tikai ar trokšņa pievienošanu* – trokšņa pievienošana ir papildinošs pasākums, kas nodrošina to, ka uzbrucējam ir grūtāk iegūt personas datus. Ja vien troksnis nav augstāks kā datu kopā iekļautā informācija, nevajadzētu trokšņa pievienošanu uzskatīt par pašpietiekamu depersonifikācijas risinājumu.

#### Kļūmes trokšņa pievienošanā

Ļoti slavens re-identifikācijas eksperiments ir tas, ko klientu datubāzē veic video satura nodrošinātāja Netflix. Pētnieki ir analizējuši šīs datu bāzes, kas sastāv no vairāk kā 100 miljoniem reitingu atzīmju skalā no 1 līdz 5 par vairāk nekā 18.000 filmām, kurus snieguši vairāk kā 500.0000 lietotāju, īpašības un struktūru. Uzņēmums publiskoja datu bāzi pēc depersonificēšanas saskaņā ar savu iekšējo privātuma politiku, noņemot visu informāciju par klientiem un atstājot tikai novērtējumus un datumus. Trokšņa pievienošanas rezultātā reitingi nedaudz paaugstinājās vai samazinājās.

Neraugoties uz to, tika konstatēts, ka 99% lietotāju ierakstu ir unikāli identificējami kā atlases kritērijus izmantojot 8 vērtējumus un datumus ar 14 dienu nobīdi, bet atlases kritēriju samazināšana (2 vērtējumi un trīs dienu nobīde) joprojām ļāva identificēt 68% lietotāju.

Tabula 2:Neefektīvs trokšņa piemērs – pārāk mazs trolsņa līmenis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Gads** | **Dzimums** | **Amats** | **Garums (cm)** | **Garums r troksni** |
| 1957 | V | Inženieris | 180 | 181 |
| 1957 | V | Izpilddirektors | 150 | 149 |
| 1957 | V | Bezdarbnieks | 170 | 169 |
| 1964 | V | Inženieris | 190 | 191 |
| 1964 | V | Vadītājs | 150 | 151 |

### Permutācija

Šī metode ietver atribūtu vērtību jaukšanu tabulā tā, lai dažas no tām ir mākslīgi saistītas ar citiem datu subjektiem, un tas ir noderīgi, ja ir svarīgi saglabāt precīzu katra atribūta sadalījumu datu kopā. Permutāciju var uzskatīt par īpašu trokšņa pievienošanas formu. Klasiskajā trokšņa pievienošanas tehnikā atribūti tiek modificēti ar nejaušām vērtībām. Pastāvīga trokšņa ģenerēšana var būt sarežģīts uzdevums, un atribūtu vērtību maiņa var nebūt pietiekama privātuma nodroši­nā­šanai. Kā alternatīvs paņēmiens ir permutācija, kas izmaina vērtības datu kopā, vienkārši mainot tās no viena ieraksta uz citu. Šāda apmaiņa nodrošinās, ka vērtību diapazons un sadalījums nemainīsies, bet mainīsies atbilstība jeb korelācija starp vērtībām un indivīdiem. Ja diviem vai vairāk atribūtiem ir loģiskas saites vai statistiska atbilstība, un tie tiek neatkarīgi pārkārtoti, šīs saites tiek iznīcinātas. Tādēļ var būt svarīgi permutēt jeb mainīt saistītu atribūtu kopu tā, lai nelauztu loģiskās saistības, pretējā gadījumā uzbrucējs var identificēt permutētos atribūtus un atjaunot permutāciju.

Piemēram, ja aplūkojam atribūtu apakškopu medicīniskajā datu kopā, piemēram, “hospitalizācijas iemesli / simptomi / atbildīgā nodaļa”, vairumā gadījumu vērtības saista spēcīgas loģiskas saites, un tikai vienas vērtības permutācija varētu tikt atklāta un pat atjaunota.

Tāpat kā trokšņa pievienošana, arī permutācija pati par sevi nevar radīt anonimitāti, un būtu vienmēr apvienojama ar acīmredzamo atribūtu noņemšanu.

**Kad lietot šo metodi?**

Metodi lieto, ja turpmākajā datu analīzē ir jāapskata tikai apkopotie dati, vai analīze ir iekšējo atribūtu līmenī - citiem vārdiem sakot, nav nepieciešama analīze par saitēm starp atribūtiem ierakstu līmenī.

#### Garantijas

* *Izšķiršana* – līdzīgi kā ar trokšņa pievienošanu, joprojām ir iespējams izšķirt indivīdu ierakstus, bet ieraksti ir mazāk ticami.
* *Saistāmība* – atribūtiem un kvaziidentifikatoriem pielietojot permutāciju, var novērst “pareizu” atribūtu saistīšanu gan iekšēji, gan ārēji ar datu kopu, bet tas tomēr atļauj “nepareizu” saistāmība, jo reālu ierakstu var saistīt ar citu datu subjektu.
* *Secināšana* – No datu kopas joprojām var izdarīt secinājumus, jo īpaši, ja atribūti korelē vai ir spēcīgas loģiskās saites; tomēr nezinot, kuri atribūti ir bijuši permutēti, uzbrucējam ir jārēķinās, ka viņa secinājums balstās uz kļūdainu hipotēzi, un līdz ar to iespējams tikai hipotētisks pieņēmums.

#### Biežāk pieļautās kļūdas

* *Nepareiza atribūta izvēle* – nesensitīvu vai maz riskantu atribūtu permutācija nedod būtisku labumu personas datu aizsardzībai. Ja sensitīvie un/vai riskantie atribūti joprojām ir saistīti ar oriģinālo atribūtu, uzbrucējam joprojām ir iespēja iegūt sensitīvu personu informāciju.
* *Atribūtu nejauša permutācija* – ja divi atribūti ir cieši saistīti, tad nejauša atribūtu permutācija nedos drošu garantiju.
* *Pieņēmums, ka pietiek tikai ar permutāciju* – līdzīgi kā trokšņa pievienošana, permutācija pati par sevi nenodrošina anonimitāti, un tā būtu jāapvieno ar citiem paņēmieniem, tādiem kā acīmredzamo atribūtu noņemšanu.

#### Permutācijas kļūmes

Piemērā parādīts, kā nejauša atribūtu permutācija rada sliktu privātuma garantiju, ja starp dažādiem atribūtiem pastāv loģiskas saites. Pēc depersonificēšanas mēģinājuma ir diezgan triviāli atjaunot informāciju par katra indivīda ienākumiem atkarībā no amata (un dzimšanas gada). Piemēram, veicot tiešu datu pārbaudi, var apgalvot, ka izpilddirektors, visticamāk, dzimis 1957. gadā, un viņam ir vislielākā alga, kamēr bezdarbnieks dzimis 1964. gadā, un viņam ir viszemākie ienākumi.

Tabula 3:Neefektīvs anonimizācijas piemērs, koriģējot savstarpēji saistītos atribūtus

|  |  |  |  |
| --- | --- | --- | --- |
| **Gads** | **Dzimums** | **Amats** | **Ienākumi (permutēti)** |
| 1957 | V | Inženieris | 2500 |
| 1957 | V | Izpilddirektors | 100 |
| 1957 | V | Bezdarbnieks | 1800 |
| 1964 | V | Inženieris | 3000 |
| 1964 | V | Vadītājs | 1500 |

### Diferencēts privātums

Diferencētā privātuma metode ietilpst nejaušas izlases metožu saimē ar atšķirīgu pieeju – tā paredz depersonificētu datu bāzu skatu ģenerēšanu, saglabājot oriģinālo datu kopiju. Šādi depersonificēti skati parasti tiek ģenerēti, izmantojot pieprasījumu kopu, konkrētai trešajai personai. Pieprasījumu kopa ietver arī dažus trokšņus, kas apzināti pievienoti *post factum*. Metode parāda datu pārzinim, cik lielu troksni un kādā formā nepieciešams pievienot, lai iegūtu nepieciešamās privātuma garantijas. Šajā kontekstā īpaši svarīgi veikt nepārtrauktu uzraudzību (vismaz katram jaunam pieprasījumam), lai pieprasījumu rezultātos nerastos iespēja identificēt personu. Jāprecizē, ka diferencēta privātuma pieeja nemaina sākotnējos datus, un tādējādi datu apstrādātājs, izmantojot visus iespējamos izmantojamos līdzekļus, var diferencēta privātuma pieprasījumu rezultātos identificēt personas. Šādi rezultāti arī ir uzskatāmi par personas datiem.

Viena no diferencētā privātuma pieejas priekšrocībām ir tā, ka dati, kas tiek sniegti pilnvarotām trešajām personām, tiek iegūti, izpildot specifiskus pieprasījumus, un nevis tiek izplatīta visa datu bāze. Auditēšanas nolūkos datu pārvaldnieks pārvalda visu vaicājumu un pieprasījumu informāciju, nodrošinot, ka trešās personas nepiekļūst datiem, kuriem tās nav autorizētas. Pieprasījumiem var tikt piemēroti arī depersonificēšanas paņēmieni, tostarp troksnis vai aizstāšana, lai pastiprinātu privātuma aizsardzību. Tomēr izveidot labu interaktīvu pieprasījumu un atbilžu mehānismu, kas spēj sniegt pietiekami precīzu informāciju uz jebkuru pieprasījumu (pēc iespējas mazāk trokšņainā veidā), vienlaikus saglabājot privātumu, joprojām ir aktuāls pētniecības jautājums.

Lai ierobežotu izsecināšanas un saistāmības uzbrukumus, nepieciešams nepārtraukti uzraudzīt pieprasījumus un ar tiem izgūto informāciju par datu subjektiem, līdz ar to “diferencēta privātuma” datubāzes nevajadzētu izvietot uz atvērtām meklētājprogrammām, kas nepiedāvā vaicājumu un izgūto ierakstu izsekojamību.

**Kad lietot šo metodi?**

Metodi lieto, ja nepieciešams dažādiem datu izguvējiem nodrošināt dažādus rezultātus. Metode nav pielietojama statistu datu anonimizēšanai.

#### Garantijas

* *Izšķiršana:* Ja tiek iegūti tikai statistiski dati un pielietotie pieprasījumi ir veidoti saprātīgi, izgūtajiem datiem nevajadzētu būt tādiem, pēc kuriem var identificēt personu.
* *Saistāmība:* Izmantojot vairākus pieprasījumus, pastāv iespēja saistīt divus atbilžu ierakstus ar vienu konkrētu personu.
* *Secināšana:* Izmantojot vairākus pieprasījumus, ir iespējams izsecināt informāciju par indivīdiem vai grupām.

Piemērus un rīkus metodes lietošanai var apskatīt šajā tīmekļa vietnē - <https://privacytools.seas.harvard.edu/differential-privacy>.

#### Biežāk pieļautās kļūdas

* *Nepietiekama trokšņa pievienošana:* Lai novērstu sasaisti ar oriģinālajiem datiem, svarīgi ir panākt pēc iespējas minimālākus pierādījumus par to, ka izgūtie dati attiecas uz datu subjektu vai datu subjektu grupu. Galvenā problēma no datu aizsardzības skatījuma ir spēja radīt tādu pievienojamā trokšņa daudzumu, kas aizsargātu personu privātumu un tai pašā laikā saglabātu izgūto datu lietderību.

#### Kļūmes diferencētā privātuma pielietošanā

Katra pieprasījuma neatkarīga apstrāde: Pieprasījumu rezultātu kombinācija var ļaut atklāt slepenu informāciju. Ja pieprasījumu vēsture netiek saglabāta, tad uzbrucējs var pielietot vairākus pieprasījumus “diferencētā privātuma” datu bāzei, kas pakāpeniski samazina iegūto rezultātu izkliedes amplitūdu, līdz ar ļoti lietu varbūtību var izdalīt vienu datu subjektu vai subjektu kopu. Bez tam svarīgi nepieļaut kļūdainu pieņēmumu, ka dati ir anonīmi trešajai personai, kamēr datu apstrādātājs, izmantojot sev pieejamos līdzekļus, joprojām var identificēt personu sākotnējā datubāzē.

## Vispārināšana

Vispārināšana (*generalization*) pārstāv otru depersonificēšanas metožu saimi. Šī pieeja paredz datu subjektu raksturlielumu vispārināšanu jeb “atšķaidīšanu”, mainot atbilstošo skalu vai detalizētību (t.i., reģions nevis pilsēta, mēnesis nevis nedēļa). Lai arī vispārināšana var būt efektīva, lai novērstu izšķiršanu, tā tomēr nenodrošina efektīvu depersonificēšanu pilnīgi visos gadījumos, un pārsvarā tā prasa īpašu un sarežģītu kvantitatīvu pieeju, lai novērstu saistāmību un secināšanu.

### Agregācija un K-anonimitāte

Agregācijas un K-anonimitātes *(K–anonymity*) metožu mērķis ir novērst datu subjekta izšķiršanu, sagrupējot tos ar vismaz k citām personām. Lai to panāktu, atribūtu vērtības tiek vispārinātas tādā mērā, ka vairākiem indivīdiem vērtība ir vienāda. Piemēram, samazinot atrašanās vietas detalizētību no pilsētas uz valsti, vienā kopā tiek iekļauts lielāks skaits datu subjektu. Dzimšanas datumus var vispārināt datumu diapazonā vai grupēt pēc mēneša vai gada. Citi skaitliskie atribūti (piemēram, alga, svars, augums u.c.) var tikt vispārināti vērtību intervālos (piemēram, alga 2000 – 3000 EUR). Šādu pieeju lietderīgi izmantot, ja atribūtu precīzo vērtību korelācijas var radīt kvaziidentifikatorus.

**Kad lietot šo metodi?**

Metodi lieto, kad atsevišķi ieraksti nav nepieciešami un apkopotie dati ir

pietiekams mērķim.

#### Garantijas

* *Izšķiršana:* Tā kā viens atribūts tiek piemērots k personām, nebūs vairs iespējams izšķirt vienu atsevišķu personu starp k personām.
* *Saistāmība:* Lai gan saistāmība ir ierobežota, tomēr joprojām ir iespējams saistīt ierakstus pa grupām no k personām. Tad šīs grupas ietvaros varbūtība, ka divi ieraksti atbilst tiem pašiem pseidoidentifikatoriem ir 1/k (kas ir ievērojami lielāka nekā varbūtība, ka šie ieraksti ir nesaistāmi).
* *Secināšana:* Galvenais K-anonimitātes modeļa trūkums ir tas, ka tas neizslēdz nevienu secināšanas uzbrukuma veidu. Un tiešām, ja visi k indivīdi ietilpst vienā grupā, un ir zināms, kurai grupai indivīds pieder, ir triviāli izgūt šī atribūta vērtību.

#### Biežāk pieļautās kļūdas

* *Dažu kvaziidentifikatoru iztrūkums:* Kritiskais parametrs, apsverot K- anonimitātes pielietošanu, ir k slieksnis. Jo lielāka k vērtība, jo spēcīgākas ir privātuma garantijas. Tipiska kļūda ir mākslīgi palielināt k vērtību, samazinot atbilstošo kvaziidentifikatoru skaitu. Kvaziidentifikatoru samazināšana atvieglo indivīdu kopņu veidošanu, jo var pastāvēt raksturīga identifikācija saistībā ar citiem atribūtiem (sevišķi, ja atribūti ir sensitīvi, vai tiem piemīt augsta entropija, kā tas ir ļoti retu atribūtu gadījumā). Kritiska kļūda, izvēloties vispārināšanas atribūtus, ir neņemt vērā visus kvaziidentifi­katorus. Ja indivīdu var izšķirt pēc dažiem atribūtiem, tad vispārināšana nespēj nodrošināt anonimitāti.
* *Maza k vērtība:* Tikpat problemātiska var būt mazas k vērtības pielietošana. Ja k vērtība ir pārāk maza, jebkura indivīda svars kopā ir pārāk nozīmīgs, un izsecināšanas uzbrukumiem ir labāks rezultāts. Piemēram, ja k=2, tad iespēja, ka diviem indivīdiem piemīt viena un tā pati īpašība, ir lielāka nekā tad, ja k>10.
* *Indivīdu ar vienādu svaru nesagrupēšana vienā grupā:* Problēmas var radīt arī indivīdu ar nevienmērīgu atribūtu sadalījumu grupēšana vienā kopā. Indivīdu ierakstu ietekme datu kopā atšķirsies: daži pārstāvēs nozīmīgu ierakstu daļu, kamēr citi – stipri mazāku. Tādēļ ir būtiski pārliecināties, ka k ir pietiekoši augsts, lai neviens indivīds nepārstāv pārāk nozīmīgu daļu no ierakstiem kopnē.

#### Kļūmes K-anonimitātes pielietošanā

Galvenā K-anonimitātes problēma ir, ka tā neaizsargā pret secināšanas uzbrukumiem. Ja uzbrucējs zina, ka datu kopā ir noteikta persona, kas ir dzimusi 1964. gadā, kā arī viņš zina, ka indivīdam bija sirdslēkme. Turklāt, ja mēs zinām, ka šī datu kopa ir iegūta no Latvijas atklāto datu portāla, tad var var pieņemt, ka Indeksi būs no Latvijas, kā arī, ka katrs indivīds dzīvo bijušajā Cēsu rajonā (pirmie divi cipari indeksā).

Tabula 4:Vāji izstrādātas k-anonimizācijas piemērs

|  |  |  |  |
| --- | --- | --- | --- |
| **Gads** | **Dzimums** | **Indeks** | **Diagnoze** |
| 1957 | V | 41\* | Sirdslēkme |
| 1957 | V | 41\* | Paaugsitnāts asinsspiedins |
| 1957 | V | 41\* | Paaugsitnāts asinsspiedins |
| 1964 | V | 41\* | Sirdslēkme |
| 1964 | V | 41\* | Sirdslēkme |

### Daudzveidība / pietuvinājums

Daudzveidība (*L-diversity model*) paplašina K-anonimitāti, lai nodrošinātu, ka secināšanas uzbrukumi vairs nav iespējami, nodrošinot, ka katrā ekvivalences klasē katram atribūtam ir vismaz l dažādas vērības.

Viens no pamatmērķiem, kas jāsasniedz, ir ierobežot ekvivalenču klašu ar vāju atribūtu mainīgumu rašanos – tā, lai uzbrucējs ar zināšanām par konkrēto datu subjektu tomēr vienmēr saskartos ar būtisku nenoteiktību.

Daudzveidība ir noderīga, aizsargājot no secināšanas uzbrukumiem datus, kuru vērtības ir pietiekami izkliedētas. Tomēr jāuzsver, ka šī metode nevar novērst informācijas noplūdi, ja atribūti kopā ir nevienmērīgi izkliedēti, vai arī vērtību vai semantiskās nozīmes diapazons ir neliels. Tādēļ daudzveidība var būt pakļauta iespējamiem secināšanas uzbrukumiem.

Pietuvinājums (*T-closeness model*) ir daudzveidības precizējums, kura mērķis ir radīt līdzvērtīgas klases, kas atgādina sākotnējo atribūtu sadalījumu tabulā. Šī tehnika ir noderīga, kad ir svarīgi saglabāt pēc iespējas lielāku datu atbilstību oriģinālam. Šim nolūkam uz ekvivalences klasi attiecina papildu ierobežojumus, tā, lai katrai ekvivalences klasei ne tikai eksistē l dažādas vērtības, bet arī, lai katra vērtība tiktu attēlota tik daudz reizes, cik nepieciešams, lai atspoguļotu katra atribūta sākotnējo sadalījumu.

**Kad lietot šo metodi?**

Metodi lieto, kad atsevišķi ieraksti nav nepieciešami un apkopotie dati ir

pietiekams mērķim.

#### Garantijas

* *Izšķiršana:* Līdzīgi kā K-anonimitāte, daudzveidība un pietuvinājums nodrošina, ka ieraksti, kas attiecas uz personu, datu bāzē nevar tikt izšķirti.
* *Saistāmība*: Attiecībā pret K-anonimitāti daudzveidība un pietuvinājums neveic uzlabojumus saistāmības jautājumā. Problēma ir tā pati, kas jebkurai kopai: iespēja, ka ieraksti pieder datu subjektam ir augstāka kā 1/N (kur N ir datu subjektu skaits datubāzē).
* *Secināšana:* Galvenais daudzveidības un pietuvinājuma uzlabojums salīdzinot ar K-anonimitāti ir tāds, ka nav vairāk iespējams veikt secināšanas uzbrukumus “daudzveidīgai” un “pietuvinātai” datu bāzei ar 100% ticamību.

#### Biežāk pieļautās kļūdas

*Sensitīvu atribūtu vērtību aizsargāšana, jaucot tos ar citiem sensitīviem atribūtiem*: Nepietiek ar divām atribūta vērtībām kopnē, lai nodrošinātu konfidencialitātes garantijas. Patiesībā sensitīvo vērtību izplatībai katrā kopnē vajadzētu atbilst šo vērtību izplatībai kopējā populācijā, vai vismaz tam vajadzētu būt vienādam visā kopnē.

#### Kļūmes daudzveidības pielietošanā

Zemāk esošajā tabulā l-diveristy tiek nodrošināta attiecībā uz atribūtu “Diagnoze” atribūtu, tomēr, ja uzbrucēju interesē 1964. gadā dzimis indivīds, tad joprojām ir iespējams izdarīt ar ļoti lielu varbūtību pieņēmumu, ka viņam ir bijusi sirdslēkme.

Tabula 5: L-diversity tabula, kurā “Diagnozes” vērtības nav vienmērīgi sadalītas

|  |  |  |  |
| --- | --- | --- | --- |
| **Gads** | **Dzimums** | **Indeks** | **Diagnoze** |
| 1957 | V | 41\* | Sirdslēkme |
| 1957 | V | 41\* | Paaugsitnāts asinsspiedins |
| 1957 | V | 41\* | Paaugsitnāts asinsspiedins |
| 1964 | V | 41\* | Sirdslēkme |
| 1964 | V | 41\* | Sirdslēkme |
| 1964 | V | 41\* | Sirdslēkme |
| 1964 | V | 41\* | Sirdslēkme |
| 1964 | V | 41\* | Sirdslēkme |
| 1964 | V | 41\* | Paaugsitnāts asinsspiedins |
| 1964 | V | 41\* | Sirdslēkme |
| 1964 | V | 41\* | Sirdslēkme |
| 1964 | V | 41\* | Sirdslēkme |
| 1964 | V | 41\* | Sirdslēkme |
| 1964 | V | 41\* | Sirdslēkme |
| 1964 | V | 41\* | Sirdslēkme |

# Pseidonimizācija

Pseidonimizācija ir viena atribūta (parasti unikāla) aizvietošana ierakstā ar citu. Līdz ar to fiziska persona joprojām ir netieši identificējama, un pseidonimizācija, ja lietota viena pati, nenodrošina depersonifikāciju (šī tēma šajās vadlīnijās tiek apskatīta tikai daudzo pārpratumu un kļūdu tās lietošanā dēļ). Detalizētu informāciju, par to var atrast - https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/what-is-personal-data/what-is-personal-data/.

Pseidonimizācija samazina datu kopas saistāmību ar sākotnējo datu subjekta identitāti, tādejādi tā ir noderīgs drošības pasākums, bet ne depersonicēšanas metode.

Pseidonimizācijas rezultāts var būt neatkarīgs no sākotnējās vērtības (tāpat kā nejaušs datu apstrādātāja ģenerēts skaitlis vai izvēlēts personas uzvārds) vai tas var būt atvasināts no atribūta vai atribūtu kopas sākotnējām vērtībām, izmantojot jaucējfunkciju vai šifrēšanas shēmu.

Visbiežāk lietotās pseidonimizācijas metodes ir šādas:

* Šifrēšana ar šifra atslēgu*:* Atslēgas turētājs var vienkāršā veidā identificēt katru datu subjektu, veicot datu kopas atšifrēšanu, jo personas dati joprojām ir datu bāzē, lai arī šifrētā veidā. Pieņemot, ka tiek lietota mūsdienīga šifrēšanas shēma, atšifrēšana iespējama tikai, zinot atslēgu.
* Jaucējfunkcija: Funkcija, kas atgriež fiksēta izmēra rezultātu no jebkura lieluma ievades (ievadē var būt viens atribūts vai atribūtu kopa), un to nevar atgriezt atpakaļ, tas nozīmē, ka ar šifrēšanu iespējamais atjaunošanas risks vairs nepastāv. Tomēr, ja jaucējfunkcijas ievades vērtību diapazons ir zināms, tās var tikt atkārtoti laistas caur jaucējfunkciju, lai noteiktu pareizu vērtību konkrētam ierakstam. Piemēram, ja datu kopa ir pseidonimizēta, jaucot personu identifikācijas numurus, tad datus var iegūt, pielietojot jaucējfunkciju visām iespējamām vērtībām un salīdzinot rezultātu ar vērtībām datu kopā. Jaucējfunkcijas parasti tiek izstrādātas tā, lai tās darbotos salīdzinoši ātri, tādēļ tās var būt pakļautas uzbrukumiem, ja tiek pielietotas nepareizā veidā. Var izveidot iepriekš izskaitļotas tabulas, kas ļauj veikt masīvu jaukšanu lielam datu apjomam. Pievienojot jaukšanas funkcijai nejaušu papildu vērtību (salt), var samazināt iespējamību iegūt ievades vērtības, tomēr arī tas pilnībā nenovērš iespēju aprēķināt sākotnējās, aiz jaukšanas funkcijas noslēptās, atribūtu vērtības, īpaši, ja atribūtu vērtības ir zināmas – vārdi, dzimšanas datumi utt.
* Jaucējfunkcija ar atslēgu: Īpaša jaucējfunkcija, kura izmanto slepenu atslēgu kā papildu ievades parametru (tā ir šīs pieejas atšķirība no “Salt”, kura parasti nav slepena). Datu apstrādātājs var atkārtoti palaist funkciju, izmantojot slepeno atslēgu, bet uzbrucējam ir daudz grūtāk izmantot funkciju, nezinot atslēgu, jo pārbaudāmo iespēju daudzums ir tik liels, ka tas jau kļūst praktiski neiespējami.
* Deterministiska šifrēšana vai jaukšanas funkcija ar atslēgas dzēšanu: šo tehniku var pielīdzināt nejauša skaita pseidonīmu atlasīšanai katram datu bāzes atribūtam, pēc tam dzēšot atbilstības tabulu. Šis risinājums ļauj samazināt saistīšanas risku starp personas datiem datu kopā un tās pašas personas datiem citā datu kopā, kur pielietots atšķirīgs pseidonīms. Uzbrucējam būs grūti atšifrēt funkciju, jo tas nozīmētu visu iespējamo atslēgu pārbaudi, pieņemot, ka atslēga nav zināma.
* Tokenizācija: Šo metodi parasti izmanto finanšu sektorā, kur karšu numuri tiek aizstāti ar vērtībām, kuras samazina uzbrukumu lietderību. Tās iegūtas, izmantojot vienvirziena šifrēšanas mehānismu, izmantojot indeksu funkciju, ietverot secīgus vai nejauši ģenerētus numurus, kas nav matemātiski atvasināti no sākotnējiem datiem.

**Kad lietot šo metodi?**

Pseidonimizāciju lieto, ja datu vērtības ir unikāli jāmaskē, kā arī ja nav nepieciešams saglabāt oriģinālo atribūtu vērtības vai jebkuru citu netiešo informāciju.

## Garantijas

* *Izšķiršana:* Joprojām ir iespējams izšķirt individuālus ierakstus, jo tie ir identificēti ar unikālu kodu, kas iegūts ar pseidonimizācijas funkciju.
* *Saistāmība*: Ierakstu sasaiste veicama vienkārši, izmantojot pseidonimizētā atribūta atsauci uz pašu indivīdu. Pat, ja datu subjektam tiek izmantoti dažādi pseidonimizēti atribūti, savietojamība joprojām var būt iespējama, izmantojot citus atribūtus. Tikai tad, ja neviens datu kopas atribūts nevar tikt izmantots, lai identificētu datu subjektu, un, ja visas saites starp sākotnējo un pseidonimizēto atribūtu ir likvidētas (tostarp, izdzēšot oriģinālu), nebūs acīmredzamas savstarpējas saites starp datu kopām.
* *Secināšana:* Ir iespējami secināšanas uzbrukumi datu subjekta reālajai identitātei datu bāzē vai dažādās datu bāzēs, kurās tiek izmantots viens un tas pats pseidonimizētais atribūts indivīdam, vai arī, ja pseidonīmi ir pašsaprotami un pietiekami nemaskē oriģinālo personas identitāti.

## Biežāk pieļautās kļūdas

* Uzskats, ka pseidonimizēti dati ir depersonificēti*:* Datu pārvaldītāji bieži uzskata, ka vienu vai dažu atribūtu noņemšana vai aizvietošana ir pietiekama, lai dati būtu anonīmi. Daudzi piemēri liecina, ka tas tā nav: vienkārši ID samainīšana nenodrošina, ka datu subjekts nevarētu tikt identificēts, ja datu bāzē paliek kvaziidentifikatori, vai arī, ja citu atribūtu vērtības joprojām spēj identificēt indivīdu. Daudzos gadījumos pseidonimizētā datu kopā personu identificēt var būt tikpat viegli kā sākotnējos datos. Lai uzskatītu datus par anonīmiem, jāveic papildu pasākumi, tostarp atribūtu noņemšana un vispārināšana, vai sākotnējo datu dzēšana, vai vismaz agregācija.
* Biežāk pieļautās kļūdas, pielietojot pseidonimizāciju kā saistāmības samazināšanas metodi:
  + Vienas un tās pašas atslēgas izmantošana vairākās datu bāzēs: dažādu datu kopu saistāmības novēršana ļoti atkarīga no šifrēšanas algoritma izmantošanas un fakta, ka viens indivīds atbildīs dažādiem pseidonimizētiem atribūtiem dažādos kontekstos. Tādēļ ir svarīgi neizmantot vienas un tās pašas šifra atslēgas dažādās datu kopās.
  + Dažādu šifra atslēgu (pēc rotācijas principa) izmantošana dažādām personām: Lai arī varētu šķist vilinoši izmantot dažādas atslēgas dažādiem personām un mainīt jeb rotēt tās pēc noteikta algoritma, piemēram, 10 ierakstu šifrēšanas. Tomēr, ja šī sistēma nav pietiekami atstrādāta un kļūst kā shēma, paredzētais drošums var samazināties. Piemēram, izmantojot atslēgu saskaņā ar kādiem noteikumiem attiecībā pret konkrētām personām, var sekmēt ierakstu sasaisti ar konkrēto personu. Tāpat arī periodiska datu pazušana no datu bāzes laikā, kad parādās jauni ieraksti, var signalizēt, ka abi ieraksti atbilst vienai personai.
  + Šifra atslēgas glabāšana: ja šifra atslēga tiek glabāta kopā ar šifrētajiem datiem, tad uzbrucējs var vienkāršā veidā sasaistīt šifrētos datus ar oriģinālajiem ierakstiem. Tas pats var notikt, ja atslēga tiek glabāta atsevišķi no datiem, bet ne drošā veidā.

## Pseidonimizācijas trūkumi

Ir izveidota datu kopa, lai pārbaudītu attiecības starp personas svaru un saņemto atalgojumu. Sākotnējā datu kopā bija iekļauti personas vārds, uzvārds adrese un dzimšanas datums, bet tas ir izdzēsts. Pētījuma grupas atsauces numurs tika izveidots no izdzēstiem datiem, izmantojot hash funkciju. Lai gan vārds, uzvārds adrese un dzimšanas datums tika izdzēsti no tabulas, vērtības ir iespējams aprēķināt zinot vismaz kādu no personām, par ko ir attiecīgais ieraksts.

Tabula 6: Pseidonimizācijas piemērs ar jaucējfunkciju (vārds, uzvārds, adrese dzimšanas datums), kur var viegli atgriezt sākotnējos datus

|  |  |  |  |
| --- | --- | --- | --- |
| **Vārds, uzvārds adrese dzimšanas datums** | **Alga** | **Ķermeņa masas index** | **References nummurs** |
|  | 720 | 15 | TEYRHHRH2 |
|  | 1000 | 14 | FJTYKYJUT |
|  | 560 | 16 | TJJETJNNR |
|  | 450 | 17 | $FGHH556G |
|  | 400 | 19 | 56FHNHRHA |

Praksē ir arī pierādījies, ka no sociālajiem tīkliem, neskatoties uz šifrēšanu, var tikt iegūti personas dati. Vietnes turētājs kļūdaini uzskatīja, ka pielietojis pietiekami spēcīgu šifrēšanu, pirms datus pārdeva citiem uzņēmumiem reklāmas un marketinga nolūkiem. Īsto personas vārdu vietā tika lietoti segvārdi, bet tas viennozīmīgi nebija pietiekami, lai aizsargātu lietotāju profilus, jo saites starp lietotājiem ir unikālas un var tikt izmantotas kā identifikators.

# Secinājumi un rekomendācijas

## Secinājumi

Depersonificēšanas un anonimizēšanas metodes ir intensīvas izpētes objekts, un katrai no tām ir savas priekšrocības un trūkumi. Vairākumā gadījumu nav iespējams vispārīgi rekomendēt izmantojamo parametru minimumu, jo katra datu kopa ir jāanalizē atsevišķi.

Daudzos gadījumos tomēr arī datu depersonificēšana joprojām var pilnībā nenodrošināt personas datu aizsardzību. Jo arī tad, ja vairs nav iespējams izgūt personas informāciju ierakstu veidā no datu bāzes, pastāv iespēja iegūt personas informāciju, izmantojot citus avotus (publiskus vai nepubliskus). Jāvērš uzmanība, ka papildu tiešai ietekmei uz datu subjektiem, kas var kļūt iespējama nepilnīgas depersonifikācijas dēļ (dažādu traucējumu, laika trūkuma dēļ vai nepilnīgas kontroles pār datiem dēļ), var rasties arī citi sliktas depersonifikācijas blakus efekti, kad uzbrucējs ir izvēlējies šo datu subjektu kā mērķi, īpaši, ja uzbrucēja nodomi ir ļaunprātīgi. Tādēļ ir svarīgi apzināties, ka depersonifikācijas metodes var nodrošināt privātuma garantijas, bet tikai tad, ja to pielietojums ir izstrādāts atbilstoši, kas nozīmē, ka ir skaidri noteikti priekšnoteikumi (konteksts), un depersonifikācijas procesa mērķi ir skaidri definēti, lai sasniegtu noteikto anonimitātes līmeni.

## Rekomendācijas

* Dažām depersonifikācijas metodēm piemīt raksturīgi ierobežojumi. Šie ierobežojumi nopietni jāapsver, pirms kāda metode tiek izmantota depersonificēšanai. Tie jāanalizē kontekstā ar depersonificēšanas mērķi – piemēram, personu privātuma aizsardzība, publicējot datus vai ļaujot izgūt daļu personu informācijas no datu bāzes.
* Neviena no minētām metodēm viena pati negarantē efektīvas depersonifikācijas visu kritēriju izpildi (t.i., neiespējamību izšķirt individuālu ierakstu, nespēju saistīt viena subjekta vairākus ierakstus, nespēju izsecināt ar datu subjektu saistītu informāciju). Tomēr, lai arī katra metode šos riskus pilnībā vai daļēji novērš, ir nepieciešama rūpīga katra individuāla gadījuma analīze, kā arī šo metožu kombināciju pielietošana, lai sasniegtu maksimālo rezultātu.

Tabulā zemāk sniegts pārskats par depersonifikācijas metožu stiprajām un vājajām pusēm attiecībā pret trīs depersonifikācijas pamatprasībām, ja tiks lietota attiecīgā metode:

Tabula 7: Apskatīto metožu stiprās un vājās puses

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Vai vēljoprojām pastāv izšķiršanas risks?** | **Vai vēljoprojām pastāv saistāmības risks?** | **Vai vēljoprojām pastāv izsecināšanas risks?** |
| Pseidonimizācija | Jā | Jā | Jā |
| Trokšņa pievienošana | Jā | Var nebūt | Var nebūt |
| Aizvietošana | Jā | Jā | Var nebūt |
| Agregācija un K-anonimitāte | Nē | Jā | Jā |
| Daudzveidība | Nē | Jā | Var nebūt |
| Diferencēts privātums | Var nebūt | Var nebūt | Var nebūt |
| Šifrēšana/ Tokenizācija | Jā | Jā | Var nebūt |

* Optimālais risinājums katram gadījumam jāizvērtē atsevišķi. Risinājumam (t.i., pilnīga depersonificēšana) jānodrošina visas trīs pamatprasības un jābūt noturīgam pret identifikācijas mēģinājumiem visiem iespējamiem un tica­miem veidiem, kurus var izmantot datu apstrādātājs un jebkura trešā persona.
* Ja rezultāts neatbilst kādam no kritērijiem, obligāti jāveic rūpīga identificēšanas risku izvērtēšana, nepieciešamības gadījumā iesaistot kopetentās iestādes.

Lai minimizētu identificēšanas riskus, ņemami vērā šādi labas prakses piemēri:

### Labā depersonificēšanas prakse

*Vispārīgi*:

* Nepaļauties uz “realizēt un aizmirst” pieeju. Ņemot vērā esošos riskus, datu apstrādātājiem:
  + Regulāri jāidentificē iespējami jauni riski un jāpārvērtē esošie,
  + Jānovērtē, vai identificēto risku kontrole ir pietiekama, un atbilstoši jākoriģē tā,
  + Riski jāuzrauga un jākontrolē nepārtraukti.
* Kā vienu no riskiem uzskatīt un ņemt vērā ne-anonimizētās datu daļas (ja tāda ir) identifikācijas iespējas, sevišķi, ja tā tiek kombinēta ar anonimizētiem datiem, kā arī iespējamo korelāciju starp atribūtiem (kā piemēram, ģeogrāfiskās atrašanās vietas un turības līmeņa korelāciju).

*Kontekstuālie elementi:*

* Datu depersonificēšanas mērķi jādefinē precīzi, jo tiem ir būtiska nozīme identifikācijas risku noteikšanā.
* Datu depersonificēšanā nedrīks aizmirst par attiecīgo kontekstuālo elementu izvērtēšanu – t.i., oriģinālo datu īpašības, kontroles mehānismi (tostarp drošības pasākumi, lai ierobežotu piekļuvi datu kopām), datu bāzes lielums (kvantitatīvie rādītāji), publisko informācijas resursu pieejamība (kurus izmanto saņēmējs), paredzētais datu nodošanas veids trešajām personām.
* Jāapsver iespējamie uzbrucēji, ņemot vērā potenciālo uzbrukumu mērķinformāciju (informācijas sensitivitāte un datu īpašības būs galvenie faktori šajā sakarā).

*Tehniskie elementi:*

* Datu apstrādātājiem jāatrod depersonifikācijas metode/metožu sajaukums, kas tiks pielietotas, īpaši, ja ir plānots izplatīt datus.
* No datu kopas jāizkļauj acīmredzamie (piemēram, reti) atribūti/ kvaziidentifikatori.
* Ja tiek pielietotas trokšņa pievienošanas metodes (nejaušas izlases metodē), ierakstiem pievienojamais trokšņa līmenis jānosaka kā atribūta vērtības funkcija, (t.i., nav jāpievieno pārāk liels vai mazs troksnis), analizējot tā ietekmi uz aizsargājamiem datiem un datu kopas īpašībām.
* Izmantojot diferencētā privātuma metodi (nejaušas izlases metodē), nepieciešams monitorēt pieprasījumus, lai atklātu privātumu apdraudošos.
* Ja tiek pielietotas vispārināšanas metodes, datu apstrādātājam būtiski ir neierobežot datu aizsardzību tikai ar vienu vispārināšanas kritēriju pat attiecībā uz vienu un to pašu atribūtu, kas nozīmē, ka piemēram, jāizvēlas dažādas atrašanās vietas detalizācijas un dažādi laika intervāli. Piemērojamā kritērija izvēli nosaka atribūta vērtību izkliede populācijā. Ne visus atribūtus var vispārināt, t.i., vispārināšanu nevar pielietot pieejā “viens der visam”. Jānodrošina mainīgums ekvivalences klasēs, piemēram, saskaņā ar augstāk minētajiem kontekstuālajiem elementiem (t.i., datu bāzes lielums, ierakstu skaits u.c.). Ja datu definētais lielums netiek sasniegts, tad attiecīgais vispārināšanas piemērs nebūtu izmantojams un lietojami citi kritērji.