TIETOSUOJATYÖRYHMÄ

0829/14/FI
WP216

Lausunto 5/2014 anonymisointitekniikoista

annettu 10. huhtikuuta 2014

TIETOSUOJATYÖRYHMÄ, joka

on perustettu 24 päivänä lokakuuta 1995 annetulla Euroopan parlamentin ja neuvoston direktiivillä 95/46/EY,

ottaa huomioon mainitun direktiivin 29 ja 30 artiklan,

ottaa huomioon työjärjestyksensä,

ON ANTANUT SEURAAVAN LAUSUNNON:

TIIVISTELMÄ

Tietosuojatyöryhmä analysoi tässä lausunnossa anonymisointitekniikkojen tehokkuutta ja rajoituksia EU:n tietosuojalainsäädäntöön nähden ja antaa suosituksia tekniikkojen käsittelemiseksi siten, että otetaan huomioon kuhunkin tekniikkaan olennaisesti liittyvä tunnistamisen jäännösriski.

Tietosuojatyöryhmä myöntää anonymisoinnin mahdollisen arvon erityisesti strategiana, jolla yksilöt ja yhteiskunta laajemmin voivat hyödyntää avointa dataa asianomaisille yksilöille aiheutuvia riskejä samalla lieventäen. Tapaustutkimuksissa ja tutkimusjulkaisuissa on kuitenkin osoitettu, kuinka vaikeaa on luoda todella anonyymi tietoaineisto ja samalla säilyttää tehtävän edellyttämä määrä perustietoja.

Direktiivin 95/46/EY ja muiden asiaankuuluvien EU:n säädösten mukaan anonymisointi tapahtuu käsittelemällä henkilötietoja siten, että henkilön tunnistaminen estyy peruuttamattomasti. Rekisterinpitäjän olisi näin tehdessään otettava huomioon kaikki ”kohtuudella toteutettavissa” olevat keinot, joita joko rekisterinpitäjä tai jokin kolmas osapuoli voi tunnistamiseen käyttää.

Anonymisointi on henkilötietojen myöhempää käsittelyä, ja sen on siten täytettävä yhteensopivuusvaatimus henkilötietojen myöhemmän käsittelyn oikeudellisten perusteiden ja käsittelyolosuhteiden osalta. Sitä paitsi vaikka anonyymit tiedot eivät kuulu tietosuojalainsäädännön piiriin, rekisteröidyt voivat siitä huolimatta olla oikeutettuja suojaan muiden säännösten (kuten viestinnän luottamuksellisuutta koskevat säännökset) nojalla.

Tässä lausunnossa kuvataan tärkeimpiä anonymisointitekniikoita eli satunnaistamista ja luokituksen karkeistamista. Lausunnossa käsitellään erityisesti kohinan lisäämistä, permutaatiota, differentiaalia yksityisyyttä, aggregointia, k-anonymiteettia, l-diversiteettiä ja t-läheisyyttä. Siinä selitetään niiden periaatteita, vahvuuksia ja heikkouksia sekä kunkin tekniikan käyttöön liittyviä tavallisia virheitä ja puutteita.

Lausunnossa pohditaan kunkin tekniikan vahvuutta kolmen kriteerin perusteella:

(i) onko yksilö edelleen mahdollista erottaa joukosta,

(ii) onko tietojen yhdistäminen yksilöön edelleen mahdollista, ja

(iii) voidaanko yksilöä koskevat tiedot päätellä.

Kunkin tekniikan tärkeimpien vahvuuksien ja heikkouksien tunteminen auttaa suunnittelemaan kussakin tilanteessa asianmukaisen anonymisointimenetelmän.

Lisäksi käsitellään peitenimillä suojaamista joidenkin piilevien vaarojen selkeyttämiseksi ja harhakäsitysten hälventämiseksi: peitenimillä suojaaminen ei ole anonymisointimenetelmä. Sillä pelkästään vähennetään mahdollisuutta yhdistää tietoaineisto rekisteröidyn alkuperäiseen henkilöllisyyteen, ja sellaisena se on hyödyllinen turvatoimi.

Lausunnossa todetaan lopuksi, että anonymisointitekniikoilla voidaan antaa takeet tietosuojasta ja niitä voidaan käyttää tuottamaan tehokkaita anonymisointiprosesseja ainoastaan, jos niiden soveltaminen on suunniteltu asianmukaisesti. Sillä tarkoitetaan, että anonymisointiprosessin edellytykset (tausta) ja tavoite tai tavoitteet on määritettävä selkeästi, jotta tavoiteltu anonymisointi saavutetaan ja samalla tuotetaan hyödyllisiä tietoja. Paras

mahdollinen ratkaisu olisi valittava tapauskohtaisesti käyttämällä mahdollisesti eri tekniikoiden yhdistelmää ja ottamalla huomioon tässä lausunnossa annetut käytännön suositukset.

Rekisterinpitäjien olisi niin ikään otettava huomioon se, että anonyymi tietoaineisto voi edelleen muodostaa jäännösriskin rekisteröidylle. Toisaalta anonymisointi ja uudelleentunnistaminen ovat aktiivisia tutkimusaloja, ja uusia tutkimustuloksia julkaistaan säännöllisesti, ja toisaalta jopa anonyymeja tietoja, kuten tilastoja, voidaan käyttää rikastamaan olemassa olevia henkilöprofiileja, jolloin syntyy uusia tietosuojaongelmia. Anonymisointia ei tulisikaan pitää kertaluonteisena tehtävänä, vaan rekisterinpitäjien olisi säännöllisesti arvioitava uudelleen siihen liittyviä riskejä.

1 Johdanto

Laitteet, anturit ja verkostot luovat suuria määriä uudenlaisia tietoja, ja tietojen tallennus on yhä edullisempaa, joten tietojen uudelleenkäyttö vastaa aiempaa useammin yleistä etua ja sen julkinen kysyntä kasvaa. Avoin data voi tuottaa yhteiskunnalle, yksilöille ja organisaatioille selkeitä hyötyjä, mutta ainoastaan jos kunnioitetaan jokaisen oikeutta henkilötietojen ja yksityiselämän suojaan.

Anonymisointi voi olla hyvä strategia, jolla hyödyt säilytetään ja riskejä lievennetään. Sen jälkeen, kun tietoaineisto on tosiasiassa anonymisoitu eikä yksilöitä voida enää tunnistaa, EU:n tietosuojalainsäädäntöä ei enää sovelleta. Tapaustutkimusten ja tutkimusjulkaisujen perusteella on kuitenkin selvää, että todella anonyymin tietoaineiston luominen aineistosta, joka sisältää runsaasti henkilötietoja, ei ole yksinkertainen tehtävä, jos samalla on tarkoitus säilyttää mahdollisimman suuri osa kulloiseenkin tehtävään tarvittavista perustiedoista. Anonyyminä pidetty tietoaineisto voidaan esimerkiksi yhdistää toiseen tietoaineistoon siten, että yksi tai useampi yksilö on tunnistettavissa.

Tietosuojatyöryhmä analysoi tässä lausunnossa anonymisointitekniikkojen tehokkuutta ja rajoituksia EU:n tietosuojalainsäädäntöön nähden ja antaa suosituksia siitä, miten anonymisointiprosessia voidaan kehittää käyttämällä näitä tekniikkoja varovasti ja vastuullisesti.

2 Määritelmät ja oikeudellinen analyysi

2.1. EU:n lainsäädännössä annetut määritelmät

Direktiivin 95/46/EY johdanto-osan 26 kappaleessa viitataan anonymisointiin anonyymien tietojen sulkemiseksi tietosuojalainsäädännön soveltamisalan ulkopuolelle:

> *”tietosuojaa koskevia periaatteita on sovellettava kaikkiin tunnistettua tai tunnistettavissa olevaa henkilöä koskeviin tietoihin; sen määrittämiseksi, onko henkilö tunnistettavissa, olisi otettava huomioon kaikki kohtuullisesti toteutettavissa olevat keinot, joita joko rekisterinpitäjä tai joku muu voi kyseisen henkilön tunnistamiseksi käyttää; tietosuojaa koskevia periaatteita ei sovelleta tietoihin, jotka on tehty anonyymeiksi siten, ettei rekisteröity enää ole tunnistettavissa; 27 artiklassa tarkoitettuja käytännesääntöjä voidaan käyttää ohjeena, kun pohditaan keinoja tietojen anonyymeiksi tekemistä varten ja niiden säilyttämiseksi muodossa, josta rekisteröidyn henkilöllisyys ei käy ilmi”.*[1]

Johdanto-osan 26 kappaletta tarkasti tutkimalla voidaan muodostaa anonymisoinnin käsitteellinen määritelmä. Johdanto-osan 26 kappale merkitsee, että tietojen anonymisoimiseksi niistä on poistettava riittävästi elementtejä, jotta rekisteröityä ei enää ole mahdollista tunnistaa. Tarkemmin sanoen tietoja on käsiteltävä siten, ettei rekisterinpitäjä tai kolmas osapuoli voi enää käyttää tietoja luonnollisen henkilön tunnistamiseen, vaikka otetaan

[1] On lisäksi huomattava, että tätä lähestymistapaa on noudatettu EU:n tietosuoja-asetusehdotuksen johdanto-osan 23 kappaleessa: ”Sen määrittämiseksi, onko luonnollinen henkilö tunnistettavissa, olisi otettava huomioon kaikki keinot, joita joko rekisterinpitäjä tai muu henkilö voi kohtuuden rajoissa käyttää mainitun henkilön tunnistamiseksi.”

huomioon ”kaikki kohtuullisesti toteutettavissa olevat keinot”. Käsittelyn peruuttamattomuus on tärkeä tekijä. Direktiivissä ei täsmennetä, kuinka tällainen tunnistettavuuden poistaminen olisi suoritettava tai voitaisiin suorittaa.[2] Painopiste on lopputuloksessa: tiedot eivät saa mahdollistaa rekisteröidyn tunnistamista minkään ”kohtuullisesti toteutettavissa” olevan keinon avulla. Käytännesääntöihin viitataan välineenä, jossa voidaan esittää mahdolliset keinot tietojen anonymisoimiseksi sekä niiden säilyttämiseksi muodossa, josta rekisteröidyn henkilöllisyys ”ei käy ilmi”. Direktiivissä asetetaan siten selvästi erittäin korkeat vaatimukset.

Myös sähköisen viestinnän tietosuojadirektiivissä (direktiivi 2002/58/EY) viitataan anonymisointiin ja nimettömiin tietoihin pitkälti samassa merkityksessä. Sen johdanto-osan 26 kappaleessa todetaan seuraavasti:

> *”Viestintäpalvelujen markkinointiin tai lisäarvopalvelujen tarjoamiseen käytettävät liikennetiedot olisi myös poistettava tai tehtävä nimettömiksi palvelun tarjoamisen jälkeen.”*

Vastaavasti direktiivin 6 artiklan 1 kohdan mukaan

> *”tilaajia ja käyttäjiä koskevat liikennetiedot, jotka yleisen viestintäverkon tai yleisesti saatavilla olevien sähköisten viestintäpalvelujen tarjoaja käsittelee ja tallentaa, on poistettava tai tehtävä nimettömiksi, kun niitä ei enää tarvita viestinnän välittämiseen, sanotun kuitenkaan rajoittamatta tämän artiklan 2, 3 ja 5 kohdan sekä 15 artiklan 1 kohdan soveltamista”.*

Lisäksi 9 artiklan 1 kohdassa todetaan seuraavasti:

> *”Jos yleisten viestintäverkkojen tai yleisesti saatavilla olevien sähköisten viestintäpalvelujen käyttäjien tai tilaajien muita paikkatietoja kuin liikennetietoja voidaan käsitellä, näitä tietoja saa käsitellä vain silloin, kun ne on tehty nimettömiksi, tai jos käyttäjät tai tilaajat ovat antaneet siihen suostumuksensa, ja tietoja saa käsitellä ainoastaan siinä määrin ja niin kauan kuin lisäarvopalvelujen tarjoaminen edellyttää.”*

Taustalla vaikuttava perustelu on, että henkilötietoihin sovellettavan anonymisointitekniikan lopputuloksen olisi tekniikan kehityksen nykyvaiheessa oltava yhtä pysyvä kuin tietojen poistamisen, eli sen on tehtävä henkilötietojen käsittely mahdottomaksi.[3]

2.2 Oikeudellinen analyysi

EU:n tärkeimmässä tietosuojalainsäädännössä anonymisoinnista käytettyjen ilmaisujen analyysin perusteella voidaan korostaa neljää päätekijää:

---

[2] Käsitettä pohditaan tarkemmin tämän lausunnon sivulla 8.

[3] Tässä yhteydessä on syytä muistaa, että anonymisointi on määritelty myös kansainvälisissä standardeissa, kuten standardissa ISO 29100, menetelmäksi, jossa henkilökohtaisesti tunnistettavat tiedot muutetaan peruuttamattomasti niin, että henkilötietorekisterin pitäjä ei enää yksin tai yhteistyössä jonkin muun osapuolen kanssa voi suoraan tai välillisesti tunnistaa henkilötietojen kohdetta (ISO 29100:2011). Henkilötietojen muuttaminen peruuttamattomasti, jotta suora tai välillinen tunnistaminen ei ole mahdollista, on keskeistä myös ISO-standardissa. Tältä kannalta katsottuna se on huomattavan yhdenmukainen direktiivin 95/46/EY perusperiaatteiden ja -käsitteiden kanssa. Sama koskee joissakin kansallisissa laeissa (esim. Italiassa, Saksassa ja Sloveniassa) annettuja määritelmiä, joissa painopisteenä on tunnistamattomuus ja joissa viitataan uudelleentunnistamisen edellyttämiin kohtuuttomiin ponnistuksiin (Saksa ja Slovenia). Ranskan tietosuojalaissa kuitenkin säädetään, että tiedot ovat henkilötietoja silloinkin kun rekisteröidyn uudelleentunnistaminen on äärimmäisen vaikeaa ja epätodennäköistä – toisin sanoen kohtuullisuustestiä ei ole käytetty.

– Anonymisointi voi olla tulosta henkilötietojen käsittelystä, jonka tavoitteena on peruuttamattomasti estää rekisteröidyn tunnistaminen.

– Useat anonymisointitekniikat ovat mahdollisia, koska EU:n lainsäädännössä ei ole asiaa koskevia säännöksiä.

– Kontekstuaalisiin elementteihin on tärkeää kiinnittää huomiota: On otettava huomioon ”kaikki” keinot, joita rekisterinpitäjä ja kolmannet osapuolet voivat ”kohtuullisesti” toteuttaa. On erityisesti kiinnitettävä huomiota siihen, mitä on nykytekniikan kehityksen valossa pidettävä ”kohtuullisesti toteutettavissa” olevana (ottaen huomioon tietokoneiden teho ja käytettävissä olevien välineiden lisääntyminen).

– Anonymisointiin liittyy aina riskitekijä, joka on otettava huomioon kunkin anonymisointitekniikan pätevyyden arvioinnissa, ja myös tällaisen tekniikan avulla anonymisoitujen tietojen mahdolliset käytöt on otettava huomioon. Riskin vakavuus ja todennäköisyys on arvioitava.

Tässä lausunnossa käytetään käsitettä ”anonymisointitekniikka” ”nimettömyyden” tai ”anonyymien tietojen” sijasta; tarkoituksena on kiinnittää huomio uudelleentunnistamisen jäännösriskiin, joka liittyy luontaisesti jokaiseen tietojen anonymisointiin tähtäävään teknis-organisatoriseen toimenpiteeseen.

2.2.1 Anonymisointiprosessin lainmukaisuus

Anonymisointi on ensinnäkin tekniikka, jota sovelletaan henkilötietoihin tunnistetietojen poistamiseksi peruuttamattomasti. Tästä syystä aloitusolettaman mukaan henkilötiedot on pitänyt kerätä ja niitä on pitänyt käsitellä tunnistetiedoilla varustettujen tietojen säilyttämiseen sovellettavan lainsäädännön mukaisesti.

Anonymisointiprosessi, jolla tarkoitetaan henkilötietojen käsittelyä siten, että niiden tunnistetiedot poistetaan, on tässä mielessä esimerkki ”myöhemmästä käsittelystä”. Sellaisena käsittelyn on vastattava käsittelytarkoituksen rajoittamista koskevassa tietosuojatyöryhmän lausunnossa nro 3/2013 annettuja ohjeita[4].

Tällä tarkoitetaan, että periaatteessa anonymisoinnin oikeudellisen perustan on oltava jokin 7 artiklan mukaisista perusteista (rekisterinpitäjän oikeutettu intressi mukaan luettuna) siten, että myös direktiivin 6 artiklan mukaiset tietojen laatua koskevat vaatimukset täyttyvät. Samalla erityisolosuhteet ja kaikki henkilötietojen käsittelytarkoituksen rajoittamista koskevassa tietosuojatyöryhmän lausunnossa[5] mainitut tekijät on otettava asianmukaisesti huomioon.

Toisaalta on kiinnitettävä huomiota direktiivin 95/46/EY 6 artiklan 1 kohdan e alakohdan säännöksiin. Niiden mukaan henkilötiedot saa säilyttää muodossa, josta rekisteröity on

[4] Tietosuojatyöryhmän lausunto 3/2013 (*Opinion 03/2013 on purpose limitation*), saatavana verkko-osoitteessa: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

[5] Tällä tarkoitetaan erityisesti, että kaikkien asiaan kuuluvien olosuhteiden valossa on tehtävä perusteellinen arviointi, jossa otetaan huomioon varsinkin seuraavat keskeiset tekijät:
a) henkilötietojen keruun tarkoituksen ja niiden myöhemmän käsittelyn tarkoituksen suhde
b) yhteys, jossa henkilötiedot on kerätty, ja rekisteröityjen kohtuulliset odotukset niiden myöhemmästä käytöstä
c) henkilötietojen luonne ja myöhemmän käsittelyn vaikutus rekisteröityihin
d) rekisterinpitäjän suojatoimet asianmukaisen tietojenkäsittelyn takaamiseksi ja rekisteröityihin kohdistuvien asiattomien vaikutusten estämiseksi.

tunnistettavissa, ainoastaan sen ajan kuin on tarpeen niiden tarkoitusten toteuttamista varten, joita varten tiedot kerättiin tai joissa niitä myöhemmin käsitellään. Sama vaatimus on ilmaistu sähköisen viestinnän tietosuojadirektiivin 6 artiklan 1 kohdassa ja 9 artiklan 1 kohdassa.

Kyseisessä säännöksessä tuodaan voimakkaasti esiin vaatimus, että henkilötiedot olisi oletusarvoisesti vähintään anonymisoitava, ellei lainmukaisista vaatimuksista muuta johdu. Tällaisia ovat esimerkiksi sähköisen viestinnän tietosuojadirektiivin liikennetietoja koskevat vaatimukset. Jos henkilötiedot halutaan säilyttää sen jälkeen, kun alkuperäisen tai myöhemmän käsittelyn tarkoitus on saavutettu, rekisterinpitäjän olisi käytettävä anonymisointitekniikoita, joilla tunnistaminen estetään peruuttamattomasti.

Näin ollen anonymisointi on tietosuojatyöryhmän näkemyksen mukaan henkilötietojen myöhempää käsittelyä, jonka voidaan katsoa olevan henkilötietojen alkuperäisen käsittelyn tarkoituksen kanssa yhteensopivaa ainoastaan sillä ehdolla, että anonymisointiprosessilla voidaan luotettavasti tuottaa anonyymeja tietoja tässä asiakirjassa tarkoitetulla tavalla.

Lisäksi on korostettava, että anonymisoinnissa on noudatettava oikeudellisia rajoituksia, joista Euroopan yhteisöjen tuomioistuin on muistuttanut asiassa C-553/07 (College van burgemeester en wethouders van Rotterdam v. M.E.E. Rijkeboer) antamassaan tuomiossa. Asia liittyy tarpeeseen säilyttää tiedot tunnistettavassa muodossa muun muassa, jotta rekisteröidyt voivat käyttää tiedonsaantioikeuttaan. Yhteisöjen tuomioistuimen tuomiossa todetaan: ”Direktiivin [95/46] 12 artiklan a alakohdassa velvoitetaan jäsenvaltiot säätämään paitsi nykyisyyteen myös menneeseen aikaan kohdistuvasta, vastaanottajia tai vastaanottajaryhmiä sekä luovutettujen tietojen sisältöä koskevasta tiedonsaantioikeudesta. Jäsenvaltioiden tehtävänä on asettaa määräaika näiden tietojen säilyttämiselle ja säätää niitä koskevasta vastaavasta tiedonsaantioikeudesta, joilla saavutetaan oikea tasapaino yhtäältä intressin, joka rekisteröidyllä on yksityiselämänsä suojaamiseen muun muassa direktiivissä säädettyjen puuttumis- ja oikeussuojakeinojen avulla, ja toisaalta sen taakan välillä, jota näiden tietojen säilyttämisvelvollisuus merkitsee rekisterinpitäjälle.”

Tällä on merkitystä erityisesti, jos rekisterinpitäjä nojautuu anonymisoinnissa direktiivin 95/46/EY 7 artiklan f alakohtaan: rekisterinpitäjän oikeutettu intressi on aina saatettava tasapainoon rekisteröidyn oikeuksien ja perusvapauksien kanssa.

Esimerkiksi Alankomaiden tietosuojaviranomainen tutki vuosina 2012–2013 neljän matkapuhelinoperaattorin käyttämää DPI-tekniikkaa eli pakettien syväluotausta. Tutkimuksen mukaan liikennetietojen sisällön anonymisoinnille niin pian kuin mahdollista tietojen keruun jälkeen oli oikeudellinen perusta direktiivin 95/46/EY 7 artiklan f alakohdassa. Sähköisen viestinnän tietosuojadirektiivin 6 artiklassa säädetäänkin, että tilaajia ja käyttäjiä koskevat liikennetiedot, joita yleisen viestintäverkon tai yleisesti saatavilla olevien sähköisten viestintäpalvelujen tarjoaja käsittelee ja tallentaa, on poistettava tai tehtävä nimettömiksi mahdollisimman nopeasti. Tässä tapauksessa tietosuojadirektiivin 7 artiklan mukainen vastaava oikeudellinen perusta on olemassa, koska se sallitaan sähköisen viestinnän tietosuojadirektiivin 6 artiklassa. Asia voitaisiin esittää toisinkin päin: jos tietojenkäsittelyn tyyppi ei ole sallittu sähköisen viestinnän tietosuojadirektiivin 6 artiklassa, sille ei voi olla oikeudellista perustetta tietosuojadirektiivin 7 artiklassa.

2.2.2 Anonyymien tietojen mahdollinen tunnistettavuus

Tietosuojatyöryhmä on käsitellyt henkilötietojen käsitettä perusteellisesti lausunnossa 4/2007. Lausunnossa on painotettu direktiivin 95/46/EY 2 artiklan a kohdassa annetun määritelmän osatekijöitä, myös määritelmän ilmausta ”tunnistettua tai tunnistettavissa olevaa”. Tässä

yhteydessä tietosuojatyöryhmä myös totesi seuraavasti: ”’Anonyymeiksi tehdyt’ tiedot ovat siten anonyymeja tietoja, jotka aiemmin liittyivät tunnistettavissa olevaan henkilöön, mutta joiden avulla tunnistaminen ei enää ole mahdollista.”

Tietosuojatyöryhmä on näin ollen jo selventänyt, että direktiivissä kehotetaan käyttämään ”kohtuullisesti toteutettavissa olevia keinoja” perusteena, jota on sovellettava, kun arvioidaan, onko anonymisointiprosessi riittävän vahva eli onko tunnistamisesta tullut ”kohtuullisesti” mahdotonta. Kunkin tapauksen erityisolosuhteet ja tausta vaikuttavat suoraan tunnistettavuuteen. Tämän lausunnon teknisessä liitteessä analysoidaan asianmukaisimman tekniikan valinnan vaikutuksia.

Kuten jo edellä korostettiin, tutkimus, välineet ja tietokoneiden teho kehittyvät. Siitä syystä ei ole mahdollista eikä edes hyödyllistä luetella tyhjentävästi olosuhteita, joissa tunnistaminen ei ole enää mahdollista. Joitakin keskeisiä tekijöitä on kuitenkin syytä ottaa huomioon ja havainnollistaa.

Ensinnäkin voidaan väittää, että rekisterinpitäjien olisi keskityttävä käytännön keinoihin, joita tarvitaan anonymisointitekniikan peruuttamiseen, erityisesti kustannuksiin ja tietämykseen, joita keinojen toteuttamiseen tarvitaan, sekä niiden todennäköisyyteen ja vakavuuteen. Rekisterinpitäjien olisi esimerkiksi anonymisointityötään ja sen kustannuksia (sekä vaadittavaa aikaa että resursseja) harkitessaan otettava huomioon, että käytettävissä on yhä enemmän edullisia teknisiä keinoja tunnistaa yksilöitä tietoaineistoista, että julkisesti saatavissa on aiempaa enemmän muita tietoaineistoja (muun muassa avoimen datan periaatteen mukaisesti julkaistuja) ja että on olemassa monia esimerkkejä, joissa epätäydellinen anonymisointi on vaikuttanut myöhemmin haitallisesti ja toisinaan korjaamattomasti rekisteröityihin.[6] On huomattava, että tunnistamisriski saattaa ajan mittaan kasvaa ja että se riippuu myös tieto- ja viestintätekniikan kehityksestä. Siitä syystä mahdolliset säädökset on muotoiltava teknologisesti neutraalilla tavalla, ja ihannetapauksessa niissä otetaan huomioon tietotekniikan kehittymispotentiaalissa tapahtuvat muutokset.[7]

Toiseksi ”joko rekisterinpitäjä tai joku muu” käyttää ”kohtuullisesti toteutettavissa” olevia keinoja ”sen määrittämiseksi, onko henkilö tunnistettavissa”. Näin ollen on ratkaisevan tärkeää ymmärtää, että jos rekisterinpitäjä ei poista alkuperäisiä (tunnistettavissa olevia) tietoja tapahtumatasolla ja luovuttaa tästä tietoaineistosta osan edelleen (esimerkiksi tunnistettavien tietojen poistamisen tai peittämisen jälkeen), tuloksena olevassa tietoaineistossa on edelleen henkilötietoja. Ainoastaan, jos rekisterinpitäjä aggregoi tiedot tasolle, jossa yksittäisiä tapahtumia ei ole mahdollista tunnistaa, tuloksena saatavaa tietoaineistoa voidaan pitää anonyymina. Jos organisaatio esimerkiksi kerää tietoja yksittäisistä matkoista, yksittäiset matkustusmallit ovat edelleen minkä tahansa osapuolen kannalta henkilötietoja niin kauan kuin rekisterinpitäjällä (tai jollakulla muulla) on pääsy alkuperäisiin käsittelemättömiin tietoihin, vaikka suorat tunnisteet olisikin poistettu kolmansille osapuolille toimitetusta aineistosta. Jos rekisterinpitäjä sen sijaan poistaa käsittelemättömät tiedot ja luovuttaa kolmansille osapuolille vain korkealla tasolla aggregoidut tiedot, kuten ”maanantaisin reitillä X on 160 prosenttia enemmän matkustajia kuin tiistaisin”, tietoja voidaan pitää anonyymeina.

[6] On kiinnostavaa todeta, että Euroopan parlamentin äskettäin (21. lokakuuta 2013) yleistä tietosuoja-asetusta koskevaan ehdotukseen tekemissä tarkistuksissa mainitaan johdanto-osan 23 kappaleessa erityisesti seuraava: ”Sen varmistamiseksi, käytetäänkö henkilön tunnistamiseksi kohtuuden rajoissa olevia keinoja, olisi otettava huomioon kaikki objektiiviset tekijät, kuten tunnistamisesta aiheutuvat kulut ja siihen tarvittava aika sekä käsittelyajankohtana käytettävissä oleva teknologia että tekninen kehitys.”

[7] Ks. tietosuojatyöryhmän lausunto 4/2007, s. 15.

Tehokas anonymisointiratkaisu estää kaikkia osapuolia erottamasta yksilöä tietoaineistosta, yhdistämästä kahta tietuetta tietoaineiston sisällä (tai kahden erillisen tietoaineiston välillä) ja päättelemästä mitään tietoja tällaisessa tietoaineistossa. Siitä syystä suorien tunniste-elementtien poistaminen ei yleensä riitä varmistamaan, että rekisteröidyn tunnistaminen ei ole mahdollista. Usein tunnistamisen estämiseksi tarvitaan lisätoimenpiteitä, jotka nytkin riippuvat anonyymien tietojen käsittelyn asiayhteydestä ja tarkoituksesta.

ESIMERKKI:
Geneettinen dataprofiili on esimerkki henkilötiedoista, jotka ovat vaarassa olla tunnistettavia, jos ainoa käytetty tekniikka on luovuttajan henkilöyden poistaminen, koska tietyt profiilit ovat luonteeltaan ainutlaatuisia. Kirjallisuudessa on jo osoitettu,[8] että tiettyjen yksilöiden henkilöys voidaan paljastaa yhdistämällä julkisesti saatavilla olevia geneettisiä resursseja (kuten sukutauluja, kuolinilmoituksia ja hakukonetuloksia) ja luovuttajien dna:ta koskevat metatiedot (luovutusajankohta, ikä, asuinpaikka), vaikka dna olisikin luovutettu ”nimettömänä”.

Kummallakin anonymisointitekniikoiden ryhmällä – tietojen satunnaistamisella ja luokituksen karkeistamisella –[9] on puutteensa, mutta tietyissä olosuhteissa kumpikin niistä saattaa sopia tavoitellun tarkoituksen saavuttamiseen vaarantamatta rekisteröityjen tietosuojaa. On tehtävä selväksi, että ”tunnistamisella” ei tarkoiteta yksinomaan mahdollisuutta selvittää henkilön nimi ja/tai osoite, vaan siihen kuuluvat myös mahdollinen tunnistettavuus joukosta erottamalla, yhdistettävyys ja päättely. Rekisterinpitäjän tai tietojen vastaanottajan aikomukset eivät liioin vaikuta tietosuojalain soveltamiseen, vaan tietosuojasääntöjä sovelletaan aina, kun tiedot ovat tunnistettavia.

Kolmas osapuoli voi käsitellä anonymisointitekniikalla käsiteltyä (alkuperäisen rekisterinpitäjän anonymisoimaa ja luovuttamaa) tietoaineistoa laillisesti tarvitsematta ottaa huomioon tietosuojavaatimuksia edellyttäen, että se ei voi (suoraan tai välillisesti) tunnistaa alkuperäisen tietoaineiston rekisteröityjä. Kolmansien osapuolten on kuitenkin otettava huomioon edellä mainitut kontekstuaaliset ja olosuhteisiin liittyvät tekijät (myös alkuperäisen rekisterinpitäjän soveltaman anonymisointitekniikan erityispiirteet) ratkaistessaan, kuinka ne käyttävät ja etenkin yhdistelevät tällaisia anonyymeja tietoja omia tarkoituksiaan varten, koska ne saattavat joutua käsittelyn seurauksena erilaisiin vahingonkorvausvastuisiin. Jos kyseiset tekijät ja piirteet tuovat mukanaan rekisteröityjen tunnistamisriskin, jota ei voida hyväksyä, tietojenkäsittely kuuluu taas tietosuojalainsäädännön alaisuuteen.

Edellä annettu luettelo ei ole millään muotoa tarkoitettu tyhjentäväksi, vaan tarkoituksena on antaa yleisiä ohjeita siitä, miten käytettävissä olevilla eri tekniikoilla anonymisoidun tietoaineiston tunnistettavuusmahdollisuuksia olisi arvioitava. Kaikkia edellä mainittuja tekijöitä voidaan pitää riskitekijöinä, jotka sekä tietoaineistoja anonymisoivan rekisterinpitäjän että kyseisiä anonyymeja tietoaineistoja omiin tarkoituksiinsa käyttävän kolmannen osapuolen on otettava huomioon.

2.2.3 Anonyymien tietojen käytön riskit

Rekisterinpitäjien on otettava huomioon seuraavat riskit, kun ne harkitsevat anonymisointitekniikoiden käyttöä:

[8] Ks. Bohannon, John: ”Genealogy Databases Enable Naming of Anonymous DNA Donors”, *Science*, nide 339, nro 6117, 18.1.2013, s. 262.
[9] Edellä mainittujen kahden anonymisointitekniikan tärkeimpiä piirteitä ja eroja kuvataan jäljempänä 3 jaksossa (Tekninen analyysi).

– Erityinen salakuoppa on pitää peitenimillä suojattuja tietoja anonyymeina tietoina. Teknistä analyysia käsittelevässä jaksossa selitetään, että peitenimillä suojattuja tietoja ei voida rinnastaa anonyymeihin tietoihin, koska niiden avulla yksilö voidaan edelleen erottaa joukosta ja yhdistää eri tietoaineistoissa. Peitenimillä suojaaminen mahdollistaa todennäköisesti tunnistettavuuden, ja siksi peitenimillä suojatut tiedot kuuluvat edelleen tietosuojalainsäädännön piiriin. Tällä on merkitystä erityisesti tieteellisen, tilastollisen tai historiantutkimuksen yhteydessä.[10]

ESIMERKKI:
Tyyppiesimerkki peitenimillä suojaukseen liittyvistä harhakäsityksistä on tunnettu AOL-tapaus (America On Line). Vuonna 2006 julkaistiin yli 650 000:ta käyttäjää kolmen kuukauden ajanjaksolla koskeva tietokanta, joka sisälsi 20 miljoonaa avainhakusanaa, ja ainoa käytetty tietosuojatoimenpide oli AOL:n käyttäjätunnuksen korvaaminen numeromääreellä. Tämä johti joidenkin käyttäjien julkiseen tunnistamiseen ja paikantamiseen. Peitenimillä suojatut hakukoneiden kyselymerkkijonot ovat erittäin vahva tunnistetekijä, varsinkin jos niihin liittyy muita määreitä, kuten IP-osoitteita tai muita asiakasparametreja.

– Toinen virhe on katsoa, että jos tiedot on asianmukaisesti anonymisoitu, yksilöllä ei ole enää oikeutta mihinkään suojatoimiin (koska kaikki edellä mainitut edellytykset ja perusteet on täytetty eivätkä tiedot enää kuulu tietosuojadirektiivin soveltamisalaan) – ennen kaikkea sen takia, että tietojen käyttöön saatetaan soveltaa jotain muuta lainsäädäntöä. Esimerkiksi sähköisen viestinnän tietosuojadirektiivin 5 artiklan 3 kohdassa kielletään kaikenlaisten ”tietojen” (myös muiden kuin henkilötietojen) tallentaminen ja päätelaitteelle tallennettujen tietojen käyttäminen ilman tilaajan tai käyttäjän suostumusta, koska se kuuluu laajempaan viestinnän luottamuksellisuutta koskevan periaatteen piiriin.

– Kolmas laiminlyönti voi syntyä siitä, ettei oteta huomioon asianmukaisesti anonymisoitujen tietojen vaikutusta yksilöihin tietyissä olosuhteissa, erityisesti profiloinnin yhteydessä. Yksilön yksityiselämä on suojattu Euroopan ihmisoikeussopimuksen 8 artiklassa ja EU:n perusoikeuskirjan 7 artiklassa. Vaikka tietosuojalakeja ei sovelleta anonyymeihin tietoaineistoihin, jotka on annettu kolmansien osapuolten käyttöön, niistä voi silti seurata yksityisyyden loukkaus. Anonyymeihin tietoihin on suhtauduttava erityisen varovasti aina, kun niitä (usein yhdessä muiden tietojen kanssa) käytetään tehtäessä yksilöihin – välillisestikin – vaikuttavia päätöksiä. Sitä, miten rekisteröidyt oikeutetusti voivat odottaa tietojaan myöhemmin käsiteltävän, on arvioitava asiayhteyden perusteella, kuten tietosuojatyöryhmä on tässä lausunnossa jo huomauttanut ja käsittelytarkoituksen rajoittamista koskevassa lausunnossa 3/2013[11] selventänyt. Muun muassa rekisteröityjen ja rekisterinpitäjien suhde, sovellettavat lakisääteiset velvoitteet ja käsittelyoperaatioiden avoimuus on otettava huomioon.

3 Tekninen analyysi, tekniikoiden vahvuudet ja tyypilliset virheet

Anonymisointikäytäntöjen ja -tekniikoiden vahvuus vaihtelee. Tässä jaksossa käsitellään tärkeimpiä näkökohtia, jotka rekisterinpitäjän on otettava huomioon soveltaessaan niitä. Tällaisia ovat etenkin kunkin tekniikan tarjoamat takeet teknologisen kehityksen nykyvaiheessa ja anonymisoinnin kolme olennaista riskiä:

[10] Ks. myös tietosuojatyöryhmän lausunto 4/2007, s. 18–20.
[11] Saatavana verkko-osoitteessa http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

- *Erottaminen joukosta* tarkoittaa mahdollisuutta eristää jokin tai kaikki tietueet, joilla yksilö tunnistetaan tietoaineistosta.
- *Yhdistettävyys* tarkoittaa mahdollisuutta yhdistää vähintään kaksi tietuetta, jotka koskevat samaa rekisteröityä tai rekisteröityjen ryhmää (joko samassa tietokannassa tai kahdessa eri tietokannassa). Jos hyökkääjä voi (esim. korrelaatioanalyysin avulla) todeta, että kaksi tietuetta koskee samaa yksilöiden ryhmää, mutta ei pysty erottamaan yksilöitä kyseisestä ryhmästä, tekniikka tarjoaa suojan joukosta erottamista vastaan, mutta ei yhdistettävyyttä vastaan.
- *Päättely* tarkoittaa mahdollisuutta päätellä attribuutin arvo muiden attribuuttien sarjan arvoista huomattavan todennäköisesti.

Ratkaisu, joka antaa suojan näitä kolmea riskiä vastaan, muodostaa vahvan suojan uudelleentunnistamiselta, jonka rekisterinpitäjä ja muu kolmas osapuoli voisivat tehdä käyttämällä kohtuullisesti toteutettavissa olevia keinoja. Tässä yhteydessä tietosuojatyöryhmä korostaa, että tunnistetietojen poistamiseen ja anonymisointiin käytettäviä tekniikoita tutkitaan jatkuvasti ja kaikkien tutkimustulosten mukaan mikään tekniikka ei itsessään ole täydellinen. Anonymisointi perustuu laajasti ottaen kahteen tekniikkaryhmään: ***satunnaistamiseen*** ja ***luokituksen karkeistamiseen***. Lausunnossa käsitellään myös muita käsitteitä, kuten *peitenimillä suojaamista, differentiaalista yksityisyyttä*, *l-diversiteettiä* ja *t-läheisyyttä*.

Lausunnon tässä jaksossa käytetään seuraavaa termistöä: tietoaineisto koostuu yksilöihin (rekisteröityihin) liittyvistä eri tietueista. Jokainen tietue liittyy yhteen rekisteröityyn, ja sillä on erilaisia arvoja (tai kirjauksia, esim. 2013) kutakin attribuuttia (esim. vuotta) varten. Tietoaineisto on kokoelma tietueita, joista voidaan muodostaa vaihtoehtoisesti joko taulukko (tai taulukoiden sarja) tai annotoitu taikka – nykyään yhä tavallisempi – painotettu graafi. Lausunnossa annetut esimerkit koskevat taulukoita, mutta niitä voidaan soveltaa myös muunlaisiin tietueiden graafisiin esityksiin. Rekisteröityyn tai rekisteröityjen ryhmään liittyvien attribuuttien yhdistelmiä voidaan kutsua kvasitunnisteiksi. Joissakin tapauksessa tietoaineistossa voi olla monia samaa yksilöä koskevia tietueita. ’Hyökkääjällä’ tarkoitetaan kolmatta osapuolta (muuta kuin rekisterinpitäjää tai henkilötietojen käsittelijää), joka käyttää alkuperäisiä tietueita joko vahingossa tai tahallaan.

3.1 Satunnaistaminen

Satunnaistamisella tarkoitetaan tekniikkojen ryhmää, jolla muutetaan tietojen totuudenmukaisuutta, jotta tietojen ja yksilön vahva yhteys voidaan poistaa. Jos tiedot ovat riittävän epävarmoja, niitä ei enää voi yhdistää tiettyyn yksilöön. Satunnaistaminen ei sinällään vähennä tietueiden ainutlaatuisuutta, koska jokainen tietue voidaan edelleen johtaa yhdestä ainoasta rekisteröidystä, mutta se voi suojata päättelyhyökkäyksiltä/-riskeiltä. Se voidaan myös yhdistää luokituksen karkeistamiseen tietosuojatakeiden vahvistamiseksi. Lisätekniikat saattavat olla tarpeen sen varmistamiseksi, että tietueesta ei voida tunnistaa yksittäistä yksilöä.

3.1.1 Kohinan lisääminen

Kohinan lisääminen on erityisen hyödyllistä, jos attribuutit voivat aiheuttaa yksilöille merkittävää haittaa. Sillä tarkoitetaan tietoaineiston attribuuttien tarkkuuden vähentämistä siten, että kokonaisjakauma säilyy. Tietoaineistoa käsittelevä havainnoitsija otaksuu, että arvot ovat tarkkoja, mutta tämä pitää paikkansa vain tietyssä määrin. Jos yksilön pituus on esimerkiksi alun perin mitattu senttimetrin tarkkuudella, anonymisoidussa tietoaineistossa pituus voi olla ilmoitettu vain kymmenen senttimetrin tarkkuudella. Jos tätä tekniikkaa

sovelletaan tehokkaasti, kolmas osapuoli ei pysty tunnistamaan yksilöä eikä pysty korjaamaan tietoja tai muulla tavoin havaitsemaan, miten tietoja on muutettu.

Kohinan lisäämiseen on yleensä yhdistettävä muita anonymisointitekniikoita, kuten ilmeisten attribuuttien ja kvasitunnisteiden poistaminen. Kohinan taso riippuu tarvittavien tietojen tasosta ja suojattujen attribuuttien paljastumisen vaikutuksesta yksilön yksityisyyteen.

3.1.1.1 Takeet

- *Erottaminen joukosta:* Yksilön tietueet on edelleen mahdollista erottaa joukosta (kenties tunnistamattomalla tavalla), vaikka tietueet eivät olekaan yhtä luotettavia.
- *Yhdistettävyys:* Saman yksilön tietueet on edelleen mahdollista yhdistää, mutta tietueet eivät ole yhtä luotettavia ja siksi todellinen tietue saatetaan yhdistää keinotekoisesti lisättyyn tietueeseen (eli kohinaan). Joissakin tapauksissa väärä attribuutti saattaa altistaa rekisteröidyn merkittävälle ja ehkä suuremmalle riskille kuin oikea.
- *Päättely:* Päättelyhyökkäykset voivat olla mahdollisia, mutta onnistumisaste on pienempi ja väärät positiiviset (sekä väärät negatiiviset) päätelmät ovat todennäköisiä.

3.1.1.2 Tavallisia virheitä

- Kohinaa lisätään epäjohdonmukaisesti: Jos kohina ei ole semanttisesti uskottavaa (eli se on suhteetonta eikä noudata sarjan attribuuttien logiikkaa), tietokantaan pääsevä hyökkääjä voi suodattaa kohinan pois ja pystyy joissakin tapauksissa tuottamaan uudelleen puuttuvat kirjaukset. Jos tietoaineisto on liian niukka[12], kohinaisten kirjausten yhdistäminen ulkoiseen lähteeseen voi lisäksi edelleen olla mahdollista.
- Oletetaan kohinan lisäämisen riittävän: Kohinan lisääminen on täydentävä toimenpide, jonka ansiosta hyökkääjän on vaikeampi saada henkilötietoja. Ellei kohina ole suurempi kuin tietoaineiston sisältämät tiedot, ei pidä olettaa, että kohinan lisääminen olisi yksinään riittävä anonymisointiratkaisu.

3.1.1.3 Kohinan lisäämisen puutteet

Erittäin kuuluisa esimerkki uudelleentunnistamiskokeesta on tehty videosisältöjen tarjoaja Neflixin asiakastietokannassa. Tutkijat analysoivat tietokannan geometriset ominaisuudet. Tietokannassa oli yli 100 miljoonaa lähes 500 000 käyttäjän asteikolla 1–5 antamaa arviointia yli 18 000 elokuvasta. Yhtiö julkaisi arvioinnit sen jälkeen, kun ne oli ”anonymisoitu” sisäisen tietosuojakäytännön mukaisesti siten, että kaikki asiakkaiden tunnistetiedot poistettiin arviointeja ja päivämääriä lukuun ottamatta. Kohinaa lisättiin parantamalla tai heikentämällä arviointeja hieman.

Tästä huolimatta todettiin, että tietoaineistossa olevista käyttäjien tietueista 99 prosenttia oli mahdollista ainutkertaisesti tunnistaa käyttämällä valintaperusteina kahdeksaa arviointia ja päivämääriä 14 päivän virheolettamalla. Valintakriteerien alentaminen (kaksi arviointia ja kolmen päivän virheolettama) mahdollisti edelleen sen, että 68 prosenttia käyttäjistä tunnistettiin.[13]

[12] Käsitettä pohditaan tarkemmin liitteessä sivulla 30.

[13] Narayanan, Arvind ja Shmatikov, Vitaly: ”Robust de-anonymization of large sparse datasets”. Symposiumijulkaisussa *IEEE Symposium on Security and Privacy, 2008.* IEEE 2008, toukokuu, s. 111–125.

3.1.2 Permutaatio

Tässä tekniikassa attribuuttien arvoja siirrellään taulukossa siten, että jotkin niistä liitetään keinotekoisesti eri rekisteröityyn. Tekniikka on hyödyllinen, jos on tärkeää säilyttää tietoaineiston jokaisen attribuutin tarkka jakauma.

Permutaatiota voidaan pitää kohinan lisäämisen erityismuotona. Klassisessa kohinan lisäämisessä attribuutteja muokataan satunnaistetuilla arvoilla. Johdonmukaisen kohinan tuottaminen saattaa olla vaikea tehtävä, eikä attribuuttien arvojen vähäinen muuttaminen välttämättä tarjoa riittävää tietosuojaa. Permutaatiossa tietoaineiston arvoja sen sijaan muutetaan yksinkertaisesti siirtämällä ne yhdestä tietueesta toiseen. Vaihdolla varmistetaan, että arvojen vaihteluala ja jakauma säilyvät entisellään, mutta arvojen ja yksilöiden välinen korrelaatio häviää. Jos kahdella tai useammalla attribuutilla on looginen suhde tai tilastollinen korrelaatio, ja niiden arvot vaihdetaan toisistaan riippumatta, suhde tuhotaan. Siitä syystä voi olla tärkeää valita permutaatioon toisiinsa liittyvien attribuuttien joukko, jotta looginen suhde ei katkea. Muussa tapauksessa hyökkääjä voi tunnistaa attribuutit, joiden arvot on vaihdettu, ja peruuttaa permutaation.

Esimerkiksi lääketieteellisen tietoaineiston osajoukossa ”sairaalalähetteen syyt / oireet / vastaava osasto” arvojen välillä on useimmissa tapauksissa voimakas looginen suhde. Tällöin vain yhden arvon permutaatio havaittaisiin ja se voitaisiin jopa peruuttaa.

Samoin kuin kohinan lisääminen, permutaatio ei yksinään tarjoa anonymiteettia, vaan siihen olisi aina yhdistettävä ilmeisten attribuuttien / kvasitunnisteiden poistaminen.

3.1.2.1 Takeet

- *Erottaminen joukosta:* Samoin kuin kohinan lisäyksessä, yksilön tietueet on edelleen mahdollista erottaa joukosta, mutta tietueet eivät ole yhtä luotettavia.
- *Yhdistettävyys:* Jos permutaatio vaikuttaa attribuutteihin ja kvasitunnisteisiin, se voi estää attribuuttien oikean yhdistämisen sekä sisäisesti että ulkoisesti tietoaineistoon, mutta se mahdollistaa edelleen väärän yhdistettävyyden, koska todellinen kirjaus saatetaan liittää eri rekisteröityyn.
- *Päättely:* Tietoaineistosta voidaan edelleen tehdä päätelmiä, varsinkin jos attribuuteilla on korrelaatioita tai vahvoja loogisia suhteita. Hyökkääjä ei kuitenkaan voi tietää, minkä attribuuttien arvot on vaihdettu. Siksi hyökkääjän on otettava huomioon, että päätelmät voivat perustua vääriin olettamiin ja että vain probabilistinen päättely on mahdollista.

3.1.2.2 Tavallisia virheitä

- Valitaan väärä attribuutti: Muiden kuin arkaluonteisten tai riskialttiiden attribuuttien permutaatio ei merkittävästi paranna henkilötietojen suojaa. Jos arkaluonteiset/riskialttiit attribuutit liittyvät yhä alkuperäiseen attribuuttiin, hyökkääjä pystyy edelleen saamaan yksilöistä arkaluonteista tietoa.
- Attribuuttien arvoja vaihdetaan havaintojen välillä satunnaisesti: Jos kaksi attribuuttia korreloi voimakkaasti, attribuuttien satunnainen permutaatio ei tarjoa vahvoja takeita. Tätä tavallista virhettä on havainnollistettu taulukossa 1.

- Oletetaan permutaation riittävän: Permutaatio ei sinällään tarjoa anonymiteettia sen enempää kuin kohinan lisääminen. Siihen olisi yhdistettävä muita tekniikoita, kuten ilmeisten attribuuttien poistaminen.

3.1.2.3 Permutaation puutteet

Esimerkissä osoitetaan, että attribuuttien satunnainen permutaatio antaa heikot tietosuojatakeet, jos eri attribuuttien välillä on loogisia yhteyksiä. Anonymisointiyrityksen jälkeen on yksinkertaista päätellä jokaisen yksilön tulot työpaikan (ja syntymävuoden) perusteella. Pelkästään tietoja tarkastelemalla voidaan esimerkiksi väittää, että taulukossa oleva toimitusjohtaja syntyi todennäköisesti vuonna 1957 ja hänellä on korkein palkka, kun taas työtön on syntynyt vuonna 1964 ja hänellä on pienimmät tulot.

Vuosi	Sukupuoli	Työ	Tulot (arvot vaihdettu)
1957	M	Insinööri	70 000
1957	M	Toimitus-johtaja	5 000
1957	M	Työtön	43 000
1964	M	Insinööri	100 000
1964	M	Johtaja	45 000

Taulukko 1. Esimerkki tehottomasta anonymisoinnista, joka on tehty vaihtamalla keskenään korreloivien attribuuttien arvoja.

3.1.3 Differentiaalinen yksityisyys

Differentiaalinen yksityisyys[14] kuuluu satunnaistamistekniikoihin, mutta lähestymistapa on erilainen: kohinan lisääminen tehdään ennen kuin tietoaineisto on määrä julkaista, mutta differentiaalista yksityisyyttä voidaan käyttää, kun rekisterinpitäjä tuottaa tietoaineistosta anonyymeja näkymiä ja säilyttää samalla kopion alkuperäisistä tiedoista. Anonyymeja näkymiä tuotetaan tyypillisesti kyselyjen alajoukon avulla tietylle kolmannelle osapuolelle. Alajoukko sisältää tarkoituksellisesti jälkikäteen lisättyä satunnaista kohinaa. Differentiaalinen yksityisyys kertoo rekisterinpitäjälle, kuinka paljon kohinaa on lisättävä ja missä muodossa, jotta tarvittavat tietosuojatakeet saavutetaan.[15] Tässä yhteydessä on erityisen tärkeää seurata jatkuvasti (ainakin jokaisen uuden kyselyn kohdalla), onko yksilö mahdollista tunnistaa kyselyn tulosjoukosta. On kuitenkin selvennettävä, että differentiaalisen yksityisyyden tekniikoilla alkuperäisiä tietoja ei muuteta. Rekisterinpitäjä pystyy tämän vuoksi tunnistamaan yksilöt differentiaalista yksityisyyttä koskevien kyselyjen tuloksista, kunhan alkuperäiset tiedot ovat tallella ja kun otetaan huomioon kohtuullisesti käytettävissä olevat keinot. Siksi myös tällaisia tuloksia on pidettävä henkilötietoina.

Differentiaaliseen yksityisyyteen perustuvan lähestymistavan etuja on se, että tietoaineistot luovutetaan vastauksena erityiskyselyyn kolmansille osapuolille, jotka ovat saaneet luvan niiden käyttöön. Yksittäistä tietoaineistoa ei näin ollen luovuteta. Rekisterinpitäjä voi säilyttää luettelon kaikista kyselyistä ja pyynnöistä. Tämä helpottaa tarkastusta sen varmistamiseksi, että kolmannet osapuolet eivät saa tietoja, joiden käyttöön niillä ei ole lupaa. Myös kyselyyn voidaan tietosuojan parantamiseksi soveltaa anonymisointitekniikoita, kuten kohinan

[14] Dwork, Cynthia: ”Differential privacy”. Teoksessa *Automata, languages and programming*. *33rd International Colloquium, ICALP*. Springer Berlin Heidelberg 2006, s. 1–12.
[15] Vrt. Felten, Ed: ”Protecting privacy by adding noise” (2012). URL: https://techatftc.wordpress.com/2012/06/21/protecting-privacy-by-adding-noise/

lisäämistä tai tietojen korvaamista toisilla. Tutkimustyö hyvän interaktiivisen kysely-vastausmekanismin kehittämiseksi jatkuu edelleen. Mekanismin avulla olisi voitava samaan aikaan sekä vastata kysymyksiin suhteellisen tarkasti (eli mahdollisimman vähäisellä kohinalla) että säilyttää yksityisyyden suoja.

Päättely- ja yhdistettävyyshyökkäysten rajoittamiseksi on välttämätöntä seurata kunkin tahon tekemiä kyselyjä ja tarkkailla, mitä tietoja se saa rekisteröidyistä. Differentiaalisella yksityisyydellä suojattuja tietokantoja ei siis pidä käyttää avoimissa hakukoneissa, joissa kyselyjä tekeviä tahoja ei voida jäljittää.

3.1.3.1 Takeet

- *Erottaminen joukosta:* Jos tietoaineistoon sovellettavat säännöt valitaan hyvin ja tuotoksena on vain tilastoja, vastausten perusteella ei pitäisi olla mahdollista erottaa yksilöä joukosta.
- *Yhdistettävyys:* Käyttämällä useita pyyntöjä voi olla mahdollista yhdistää kahdessa vastauksessa olevat tiettyyn yksilöön liittyvät kirjaukset.
- *Päättely:* Yksilöjä tai ryhmiä koskevia tietoja on mahdollista päätellä käyttämällä useita tietopyyntöjä.

3.1.3.2 Tavallisia virheitä

- Kohinaa ei lisätä riittävästi: Taustatietoihin yhdistäminen voidaan estää antamalla mahdollisimman vähän tietoja siitä, onko tietty rekisteröity tai rekisteröityjen ryhmä mukana tietoaineistossa. Tietosuojan kannalta suurin vaikeus on lisätä todellisiin vastauksiin oikea määrä kohinaa, jotta voidaan suojata yksilön yksityisyyttä ja samalla säilyttää luovutettujen vastausten hyödyllisyys.

3.1.3.3 Differentiaalisen yksityisyyden puutteet

Jokaisen kyselyn riippumaton käsittely: Kyselyjen tuloksia yhdistämällä voidaan paljastaa luottamuksellisiksi tarkoitettuja tietoja. Jos kyselyhistoriaa ei säilytetä, hyökkääjä voi suunnitella differentiaalisesti suojattuun tietokantaan kysymyssarjan, jolla vähennetään asteittain tuotetun otoksen kokoa, kunnes saadaan esiin yksittäisen rekisteröidyn tai rekisteröityjen ryhmän ominaispiirre deterministisesti tai erittäin suurella todennäköisyydellä. Lisäksi on varottava kuvittelemasta, että tiedot ovat kolmannelle osapuolelle anonyymejä, jos rekisterinpitäjä voi edelleen tunnistaa rekisteröidyn alkuperäisessä tietokannassa ottamalla huomioon kaikki kohtuullisesti toteutettavissa olevat keinot.

3.2 Luokituksen karkeistaminen

Luokituksen karkeistaminen muodostaa toisen anonymisointitekniikoiden ryhmän. Lähestymistavassa rekisteröidyn attribuuttien luokitusta karkeistetaan eli yleistetään muuttamalla mittakaavaa tai suuruusluokkaa (esim. alue kaupungin sijasta, kuukausi viikon sijasta). Vaikka luokituksen karkeistamisella voidaan tehokkaasti estää joukosta erottaminen, se ei tarjoa kaikissa tapauksissa tehokasta anonymiteettia. Varsinkin yhdistettävyyden ja päättelyn estämiseen tarvitaan monimutkaista kvantitatiivista erityislähestymistapaa.

3.2.1 Aggregointi ja k-anonymiteetti

Aggregoinnin ja k-anonymiteettitekniikan tarkoituksena on estää rekisteröidyn erottaminen joukosta ryhmittämällä rekisteröidyn kanssa samaan ryhmään vähintään k yksilöä. Tätä varten attribuuttien arvojen luokitusta karkeistetaan siinä määrin, että jokaisella yksilöllä on sama arvo. Jos esimerkiksi sijainnin arvon luokitus karkeistetaan kaupungista maahan, luokkaan sisältyy suurempi määrä rekisteröityjä. Yksilölliset syntymäajat voidaan yleistää ajanjaksoiksi tai ryhmitellä kuukausittain tai vuosittain. Muiden numeeristen arvojen (kuten palkan, painon, pituuden tai lääkeannoksen koon) luokitusta voidaan karkeistaa käyttämällä rajakohta-arvoja (esim. palkka 20 000–30 000 euroa). Aggregointia ja k-anonymiteettia voidaan käyttää, jos attribuuttien täsmällisten arvojen korrelointi voi luoda kvasitunnisteita.

3.2.1.1 Takeet

- *Erottaminen joukosta:* Koska k käyttäjällä on nyt samat attribuutit, yksilön erottamisen k käyttäjän ryhmästä ei pitäisi olla mahdollista.
- *Yhdistettävyys:* Vaikka yhdistettävyys on rajoitettua, on edelleen mahdollista yhdistää tietueita k käyttäjän ryhmiin. Ryhmän sisällä todennäköisyys, että kaksi tietuetta vastaa samaa pseudotunnistetta on 1/k (mikä saattaa olla merkittävästi korkeampi kuin todennäköisyys, että tällaisia kirjauksia ei voida yhdistää).
- *Päättely:* K-anonymiteettimallin tärkein puute on, että se ei estä minkääntyyppisiä päättelyhyökkäyksiä. Jos kaikki k yksilöä kuuluvat samaan ryhmään ja tiedetään, mihin ryhmään yksilö kuuluu, on helppoa saada esiin ominaisuuden arvo.

3.2.1.2 Tavallisia virheitä

- Kvasitunnisteita jää huomaamatta: K-anonymiteetin kriittinen parametri on k:n kynnysarvo. Mitä suurempi k:n arvo on, sitä vahvemmat tietosuojatakeet menetelmä antaa. Tavallinen virhe on kasvattaa keinotekoisesti k:n arvoa pienentämällä huomioon otettua kvasitunnisteiden joukkoa. Kvasitunnisteiden vähentäminen helpottaa k käyttäjän klustereiden muodostamista muihin attribuutteihin olennaisesti kuuluvan tunnistamispotentiaalin takia (varsinkin, jos jotkin niistä ovat arkaluonteisia tai niillä on erittäin korkea entropia, kuten erittäin harvinaisten attribuuttien tapauksessa). Jos kaikkia kvasitunnisteita ei oteta huomioon, kun karkeistettavaa attribuuttia valitaan, tehdään ratkaiseva virhe. Jos joitakin attribuutteja voidaan käyttää erottamaan yksilö k käyttäjän klusterista, luokituksen karkeistamisella ei pystytä suojaamaan kaikkia yksilöitä (ks. esimerkki taulukossa 2).
- K:lla on pieni arvo: K:n pienen arvon tavoittelu on yhtä lailla ongelmallista. Jos k on liian pieni, yksilön painoarvo klusterissa on liian merkittävä, ja päättelyhyökkäysten mahdollisuus onnistua on suurempi. Jos esimerkiksi k = 2, todennäköisyys, että kahdella yksilöllä on sama ominaisuus, on suurempi kuin jos k > 10.
- Ryhmiteltävillä yksilöillä ei ole samaa painoarvoa: Ongelmia voi syntyä myös, jos yhteen joukkoon ryhmitellään yksilöitä, joiden attribuuttien jakauma on epätasainen. Yksilön tietueen vaikutus tietoaineistoon vaihtelee: toiset edustavat merkittävää osuutta kirjauksista, kun taas toisten panos jää suhteellisen vaatimattomaksi. Tästä syystä on tärkeää varmistaa, että k on riittävän suuri, jotta mikään yksilö ei edusta liian merkittävää osuutta klusterin kirjauksista.

3.1.3.3 K-anonymiteetin puutteet

Tärkein k-anonymiteettiin liittyvä ongelma on, ettei se estä päättelyhyökkäyksiä. Jos hyökkääjä seuraavassa esimerkissä tietää, että tietty yksilö on tietoaineistossa ja että hän on syntynyt vuonna 1964, hän tietää myös, että yksilöllä on ollut sydänkohtaus. Jos lisäksi tiedetään, että tietoaineisto on saatu ranskalaiselta organisaatiolta, tiedetään myös, että jokainen yksilö asuu Pariisissa, koska Pariisin postinumeroiden kolme ensimmäistä numeroa ovat 750*.

Vuosi	Sukupuoli	Postinro	Diagnoosi
1957	M	750*	Sydänkohtaus
1957	M	750*	Kolesteroli
1957	M	750*	Kolesteroli
1964	M	750*	Sydänkohtaus
1964	M	750*	Sydänkohtaus

Taulukko 2. Esimerkki huonosti suunnitellusta k-anonymiteetista.

3.2.2 L-diversiteetti/t-läheisyys

K-anonymiteettia on kehitetty edelleen l-diversiteetillä deterministisen päättelyhyökkäysten estämiseksi. Sillä varmistetaan, että jokaisella attribuutilla on kussakin ekvivalenssiluokassa vähintään *l* eri arvoa.

Eräs perustavoite on rajoittaa sellaisten ekvivalenssiluokkien määrää, joissa attribuuttien vaihtelevuus on heikkoa, jotta hyökkääjä, jolla on taustatietoa tietystä rekisteröidystä, jää aina huomattavan epävarmaksi.

Tietoja voidaan suojata päättelyhyökkäyksiltä l-diversiteetin avulla, kun attribuuttien arvot vaihtelevat riittävästi. On kuitenkin korostettava, että tekniikalla ei voida estää tietojen vuotamista, jos osioon kuuluvat attribuutit ovat jakautuneet epätasaisesti tai kuuluvat pieneen arvoalueeseen tai merkitysryhmään. Sitä paitsi l-diversiteettiin voidaan kohdistaa probabilistisia päättelyhyökkäyksiä.

T-läheisyys on l-diversiteetin jatkokehittelyä sikäli, että sillä pyritään luomaan ekvivalenssiluokka, jossa attribuuttien jakauma muistuttaa taulukon attribuuttien alkuperäistä jakaumaa. Tekniikkaa käytetään, kun on tärkeää säilyttää tiedot mahdollisimman lähellä alkuperäisiä. Sen takia ekvivalenssiluokkaan kohdistetaan lisärajoitus: kussakin ekvivalenssiluokassa on oltava vähintään *l* eri arvoa, minkä lisäksi jokaisen arvon on esiinnyttävä niin monta kertaa, että se vastaa kunkin attribuutin alkuperäistä jakaumaa.

3.2.2.1 Takeet

- *Erottaminen joukosta:* Kuten k-anonymiteetilla, l-diversiteetillä ja t-läheisyydellä voidaan varmistaa, että yksilöön liittyviä attribuutteja ei voida erottaa tietokannasta.
- *Yhdistettävyys:* L-diversiteetti ja t-läheisyys eivät ole parannuksia k-anonymiteettiin nähden yhdistettävyyden osalta. Ongelma on sama kuin missä tahansa klusterissa: todennäköisyys, että samat kirjaukset kuuluvat samalle rekisteröidylle, on suurempi kuin 1/N (jossa N on rekisteröityjen määrä tietokannassa).

- *Päättely:* L-diversiteetin ja t-läheisyyden tärkein parannus k-anonymiteettiin nähden on, että päättelyhyökkäykset tietokantaa vastaan eivät ole sataprosenttisen varmoja, jos tietokantaan on sovellettu l-diversiteettiä tai t-läheisyyttä.

3.2.2.2 Tavallisia virheitä

- Arkaluonteisten attribuuttien arvot on suojattu sekoittamalla ne muiden arkaluonteisten attribuuttien kanssa: Tietosuojaa ei voida taata sillä, että klusterissa on kaksi attribuutin arvoa. Arkaluonteisten arvojen jakauman pitäisi kussakin klusterissa muistuttaa kyseisten arvojen jakaumaa koko väestössä tai sen pitäisi vähintään olla yhdenmukainen koko klusterissa.

3.2.2.3 L-diversiteetin puutteet

Jäljempänä olevassa taulukossa l-diversiteettiä on sovellettu attribuuttiin ”Diagnoosi”. Jos kuitenkin tiedetään, että taulukossa on vuonna 1964 syntynyt henkilö, on edelleen mahdollista olettaa erittäin suurella todennäköisyydellä, että hänellä on ollut sydänkohtaus.

Vuosi	Sukupuoli	Postinro	Diagnoosi
1957	M	750*	**Sydänkohtaus**
1957	M	750*	**Kolesteroli**
1957	M	750*	**Kolesteroli**
1957	M	750*	**Kolesteroli**
1964	M	750*	**Sydänkohtaus**
1964	M	750*	**Sydänkohtaus**
1964	M	750*	**Sydänkohtaus**
1964	M	750*	**Kolesteroli**
1964	M	750*	**Sydänkohtaus**
1964	M	750*	**Sydänkohtaus**
1964	M	750*	**Sydänkohtaus**
1964	M	750*	**Sydänkohtaus**
1964	M	750*	**Sydänkohtaus**
1964	M	750*	**Sydänkohtaus**

Taulukko 3. L-diversiteetin mukainen taulukko, jossa diagnoosin arvot eivät ole jakautuneet yhdenmukaisesti.

Nimi	Syntymäaika	Sukupuoli
Smith	1964	M
Rossi	1964	M
Dupont	1964	M
Jansen	1964	M
Garcia	1964	M

Taulukko 4. Jos hyökkääjä tietää, että kyseiset henkilöt ovat taulukossa 3, hän voi päätellä, että heillä on ollut sydänkohtaus.

4. Peitenimillä suojaaminen

Peitenimillä suojaamisella tarkoitetaan tietueen yhden (tavallisesti ainutkertaisen) attribuutin korvaamista toisella. Luonnollinen henkilö on sen vuoksi edelleen todennäköisesti tunnistettavissa välillisesti, joten peitenimillä suojaaminen ei yksinään johda anonyymin tietoaineiston luomiseen. Sitä käsitellään kuitenkin tässä lausunnossa sen käyttöön liittyvien monien harhakäsitysten ja virheiden takia.

Peitenimillä suojaaminen vähentää tietoaineiston yhdistettävyyttä rekisteröidyn alkuperäiseen identiteettiin, ja sellaisena se on hyödyllinen turvatoimi mutta ei anonymisointimenetelmä.

Peitenimillä suojaamisen tulos voi olla riippumaton alkuperäisestä arvosta, kuten tapahtuu, jos rekisterinpitäjä tuottaa satunnaisen numeron tai rekisteröity valitsee sukunimen. Se voidaan myös johtaa attribuutin tai attribuuttien joukon alkuperäisistä arvoista esimerkiksi käyttämällä tiivistysfunktiota tai salausjärjestelmää.

Eniten käytettyjä peitenimellä suojaamistekniikoita:

- Salaus salaisella avaimella: Tässä tapauksessa avaimen haltija voi helposti tunnistaa uudelleen jokaisen rekisteröidyn purkamalla tietoaineiston salauksen, koska henkilötiedot ovat edelleen tietoaineistossa, vaikkakin salatussa muodossa. Jos oletetaan, että on käytetty uusinta salaustekniikkaa, salauksen purku on mahdollista ainoastaan, jos tuntee avaimen.
- Tiivistysfunktio (hash-funktio): Tällä tarkoitetaan peruuttamatonta funktiota, jolla syötteen koosta riippumatta palautetaan kiinteän kokoinen tuotos. Syöte voi olla yksi ainoa attribuutti tai attribuuttien joukko. Tällöin salaukseen liittyvää peruuttamisriskiä ei ole. Jos tiivistysfunktion syötteen arvojen vaihteluväli kuitenkin tunnetaan, tietyn tietueen oikea arvo voidaan johtaa toistamalla tiivistysfunktio. Jos tietoaineisto on esimerkiksi suojattu peitenimillä tiivistämällä kansallinen henkilötunnus, se voidaan johtaa yksinkertaisesti tiivistämällä kaikki mahdolliset syötteen arvot ja vertaamalla tulosta tietoaineiston arvoihin. Tiivistysfunktiot on yleensä suunniteltu siten, että ne voidaan laskea suhteellisen nopeasti, jolloin ne ovat alttiita väsytysmenetelmähyökkäyksille.[16] Myös esikäsiteltyjä taulukoita voidaan luoda, jolloin voidaan kerralla palauttaa suuri määrä tiivistettyjä arvoja.

[16] Tällaisissa hyökkäyksissä kokeillaan kaikkia todennäköisiä syötteitä vastaavuustaulukon luomiseksi.

Todennäköisyyttä, että syötteen arvo voidaan päätellä, voidaan vähentää käyttämällä niin sanottua suolattua tiivistysfunktiota (jolloin tiivistettävään attribuuttiin lisätään satunnaisarvo eli ”suolaa”). Suolatun tiivistysfunktion avulla piilotetun attribuutin alkuperäisen arvon laskeminen saattaa silti olla kohtuullisesti toteutettavissa.[17]

- Avaimeen perustuva tiivistysfunktio, jossa avain tallennetaan: Tällä tarkoitetaan erityistä tiivistysfunktiota, jossa käytetään salaista avainta lisäsyötteenä (tämä eroaa suolatusta tiivistysfunktiosta siinä, että ”suola” ei yleensä ole salainen). Rekisterinpitäjä voi toistaa funktion attribuutissa käyttämällä salausavainta, mutta hyökkääjän on paljon vaikeampi toistaa funktio tietämättä avainta, koska testattavien mahdollisuuksien lukumäärä on riittävän suuri, jotta se on käytännössä hankalaa.
- Deterministinen salaus eli avaimeen perustuva tiivistysfunktio, jossa avain poistetaan: Tekniikka vastaa sitä, että tietokannassa kullekin attribuutille valitaan peitenimeksi satunnainen numero, ja sen jälkeen vastaavuustaulukko poistetaan. Ratkaisun avulla on mahdollista[18] vähentää tietoaineistossa olevien henkilötietojen yhdistettävyyttä samaa yksilöä koskeviin tietoihin toisessa tietoaineistossa, jossa käytetään eri peitenimeä. Jos käytetään uusimman tekniikan mukaista algoritmia, hyökkääjän on laskentateknisesti vaikea purkaa salaus tai toistaa funktio, koska se edellyttäisi jokaisen mahdollisen avaimen testaamista; avainhan ei ole käytettävissä.
- Alkioiden korvaus eli tokenisaatio: Tekniikkaa käytetään tavallisesti (vaikkakaan ei yksinomaan) rahoitusalalla, jossa kortin tunnistenumerot korvataan arvoilla, joista on vähän hyötyä hyökkääjälle. Se perustuu edellä esitettyihin tekniikoihin, ja siinä käytetään tyypillisesti yhdensuuntaista salausmekanismia tai menetelmää, jossa arvolle osoitetaan indeksifunktion avulla järjestysnumero tai satunnaisesti tuotettu numero, jota ei voida laskea alkuperäisten tietojen perusteella.

4.1 Takeet

- *Erottaminen joukosta:* Yksilön tietueet on edelleen mahdollista erottaa joukosta, koska yksilön tunnisteena on ainutkertainen attribuutti, joka on saatu peitenimifunktiolla (= peitenimellä suojattu attribuutti).
- *Yhdistettävyys:* Tietueiden yhdistäminen on edelleen helppoa, jos samalla peitenimellä suojattua attribuuttia käytetään viittaamaan samaan yksilöön. Vaikka samasta rekisteröidystä käytettäisiin eri peitenimillä suojattuja attribuutteja, yhdistettävyys voi silti olla mahdollista muiden attribuuttien avulla. Eri peitenimillä suojattuja attribuutteja käyttävien kahden tietoaineiston välillä ei ole ilmeisiä ristiinviittauksia ainoastaan siinä tapauksessa, että mitään muuta tietoaineiston attribuuttia ei voida käyttää rekisteröidyn tunnistamiseen ja että jokainen yhteys alkuperäisen attribuutin ja peitenimellä suojatun attribuutin väliltä on poistettu (myös poistamalla alkuperäiset tiedot).
- *Päättely:* Rekisteröidyn todellista henkilöyttä koskevat päättelyhyökkäykset ovat mahdollisia tietoaineistossa tai eri tietokantojen välillä, jos yksilöstä käytetään samaa peitenimellä suojattua attribuuttia tai jos peitenimet ovat itsestään selviä eivätkä peitä rekisteröidyn alkuperäistä henkilöyttä asianmukaisesti.

[17] Näin on varsinkin, jos attribuutin laji on tiedossa (nimi, sosiaaliturvatunnus, syntymäaika jne.). Laskenta-ajan lisäämiseksi voidaan käyttää avaimeen perustuvaa tiivistysfunktiota, jossa laskettu arvo tiivistetään useita kertoja käyttämällä ”lyhyttä suolaa” (eli vain muutaman merkin pituista satunnaisarvoa).

[18] Tämä riippuu tietoaineiston muista attribuuteista ja alkuperäisten tietojen poistamisesta.

4.2 Tavallisia virheitä

- Kuvitellaan, että peitenimillä suojattu tietoaineisto on anonyymi: Rekisterinpitäjät olettavat usein, että yhden tai useamman attribuutin poistaminen tai korvaaminen toisella riittää tekemään tietoaineistosta anonyymin. Monet esimerkit osoittavat, ettei näin ole; pelkkä tunnisteen muuttaminen ei estä tunnistamasta rekisteröityä, jos tietoaineistoon jää kvasitunnisteita tai jos yksilön tunnistaminen on mahdollista muiden attribuuttien arvojen avulla. Monissa tapauksissa voi olla yhtä helppoa tunnistaa yksilö peitenimillä suojatusta tietoaineistosta kuin alkuperäisistä tiedoista. Tietoaineiston anonymisointi vaatii lisätoimia, kuten attribuuttien poistamista ja luokituksen karkeistamista tai alkuperäisten tietojen poistamista tai ainakin aggregointia erittäin korkealle tasolle.
- Tavallisia virheitä, kun peitenimillä suojaamista käytetään yhdistettävyyden estämiseen:

    - Saman avaimen käyttäminen eri tietokannoissa: Eri tietokantojen välisen yhdistettävyyden poistaminen edellyttää vahvasti, että käytetään koodattua algoritmia ja että yhtä yksilöä vastaavat eri yhteyksissä eri peitenimillä suojatut attribuutit. Yhdistettävyyden vähentämiseksi on siis tärkeää olla käyttämättä samaa avainta eri tietokannoissa.
    - Eri avainten käyttö eri käyttäjillä (”kiertävät avaimet”): Saattaa olla houkuttelevaa käyttää eri käyttäjäryhmillä eri avaimia ja muuttaa avainta käyttökohtaisesti (esimerkiksi käytetään samaa avainta tallentamaan samaa käyttäjää koskevat kymmenen kirjausta). Jos operaatiota ei ole suunniteltu asianmukaisesti, se saattaa kuitenkin aiheuttaa säännönmukaisuuksien muodostumista, mikä osittain vähentää aiottuja hyötyjä. Jos avainta esimerkiksi kierrätetään tiettyjen sääntöjen mukaisesti tietyillä yksilöillä, mainittuja yksilöitä vastaavien kirjausten yhdistäminen helpottuu. Myös peitenimellä suojatun tiedon katoaminen tietokannasta toistuvasti samaan aikaan, kun uusi tieto ilmestyy, voi paljastaa, että molemmat tietueet koskevat samaa luonnollista henkilöä.
    - Avaimen säilyttäminen: Jos salainen avain tallennetaan yhdessä peitenimellä suojattujen tietojen kanssa ja tietoturva vaarantuu, hyökkääjä voi helposti yhdistää peitenimellä suojatut tiedot niiden alkuperäiseen attribuuttiin. Sama koskee tilannetta, jossa avain tallennetaan erikseen muulla kuin turvallisella tavalla.

4.3 Peitenimillä suojauksen puutteita

- *Terveydenhuolto*

1. Nimi, osoite ja syntymäaika	2. Erityistukijakso	3. Painoindeksi	6. Tutkimuskohortin viitenro
[redacted]	< 2 vuotta	15	QA5FRD4
	> 5 vuotta	14	2B48HFG
	< 2 vuotta	16	RC3URPQ
	> 5 vuotta	18	SD289K9
	< 2 vuotta	20	5E1FL7Q

Taulukko 5. Esimerkki tiivistämällä tehdystä peitenimellä suojaamisesta (nimi, osoite, syntymäaika), joka voidaan helposti peruuttaa.

Tietoaineisto on luotu, jotta voidaan tutkia henkilön painon ja erityistukietuuden saamisen välistä suhdetta. Alkuperäiseen tietoaineistoon on sisältynyt rekisteröidyn nimi, osoite ja syntymäaika, mutta nämä tiedot on poistettu. Tutkimuskohortin viitenumero on tuotettu poistetuista tiedoista tiivistysfunktion avulla. Vaikka taulukosta on poistettu nimi, osoite ja syntymäaika, tutkimuskohortin viitenumeroiden laskeminen on helppoa, jos yhden rekisteröidyn nimi, osoite ja syntymäaika ovat tiedossa käytetyn tiivistysfunktion lisäksi.

- *Sosiaaliset verkostot*

  On osoitettu[19], että sosiaalisten verkostojen kaavioista on mahdollista saada yksittäisiä henkilöitä koskevia arkaluonteisia tietoja huolimatta tällaisiin tietoihin sovelletuista peitenimitekniikoista. Sosiaalisen verkoston tarjoaja oli myynyt tiedot muille yrityksille markkinointi- ja mainontatarkoituksiin ja oletti virheellisesti, että peitenimillä suojaaminen olisi vahva tapa estää tunnistaminen. Tarjoaja käytti todellisten nimien sijasta peitenimiä, mutta se ei selvästikään riittänyt anonymisoimaan käyttäjäprofiileja, koska eri yksilöiden väliset suhteet ovat ainutlaatuisia ja niitä voidaan käyttää tunnisteina.

- *Sijainnit*

  MIT:n tutkijat[20] analysoivat äskettäin peitenimillä suojatun tietoaineiston, joka käsitti 1,5 miljoonan ihmisen spatio-temporaalisen liikkuvuuden koordinaatteja sadan kilometrin säteellä 15 kuukauden aikana. Tutkijat osoittivat, että 95 prosenttia populaatiosta voitiin erottaa joukosta neljän sijaintipisteen avulla, ja pelkästään kaksi pistettä riitti erottamaan joukosta yli 50 prosenttia rekisteröidyistä (yksi tällaisista tiedossa olevista pisteistä on erittäin todennäköisesti ”koti” tai ”työpaikka”). Tällöin yksityisyyden suojalle jäi hyvin rajoitettu tila, vaikka yksilöiden henkilöydet oli suojattu peitenimillä korvaamalla todelliset attribuutit muilla merkinnöillä.

[19] Narayanan, Arvind ja Shmatikov, Vitaly: ”De-anonymizing social networks”. Symposiumijulkaisussa *30th IEEE Symposium on Security and Privacy*. IEEE, 2009.
[20] de Montjoye, Yves-Alexandre, Hidalgo, César A., Verleysen, Michel ja Blondel Vincent D.: ”Unique in the Crowd: The privacy bounds of human mobility”. *Scientific Reports 3*, artikkelinro 1376, Nature 2013.

5 Päätelmät ja suositukset

5.1 Päätelmät

Tunnistetietojen poistamiseen ja anonymisointiin käytettäviä tekniikoita tutkitaan tiiviisti, ja tässä asiakirjassa on johdonmukaisesti osoitettu, että kullakin tekniikalla on etunsa ja haittansa. Useimmissa tapauksissa ei ole mahdollista antaa käytettäviä parametreja koskevia vähimmäissuosituksia, koska jokaista tietoaineistoa on käsiteltävä tapauskohtaisesti.

Monissa tapauksissa anonyymit tietoaineistot voivat kuitenkin muodostaa rekisteröidylle jäännösriskin. Vaikka yksilön täsmällisen tietueen esiin saaminen ei olisikaan mahdollista, voi silti olla mahdollista koota yksilöä koskevia tietoja muiden (julkisesti tai muutoin) saatavilla olevien tietolähteiden avulla. Heikolla anonymisoinnilla saattaa olla rekisteröityyn suoria vaikutuksia: harmia, ajanhukkaa ja hallinnan menettämisen tunne, koska rekisteröity on sisällytetty klusteriin tietämättään tai ilman ennakkosuostumusta. On syytä korostaa, että sen lisäksi muita, välillisiä sivuvaikutuksia saattaa syntyä aina, kun hyökkääjä sisällyttää rekisteröidyn erehdyksessä kohteeseen anonyymien tietojen käsittelyn seurauksena – varsinkin, jos hyökkääjällä on pahantahtoiset aikomukset. Tästä syystä tietosuojatyöryhmä korostaa, että anonymisointitekniikoilla voidaan antaa takeet tietosuojasta ainoastaan, jos niiden soveltaminen on suunniteltu asianmukaisesti. Sillä tarkoitetaan, että anonymisointiprosessin edellytykset (tausta) ja tavoite tai tavoitteet on määritettävä selkeästi, jotta tavoiteltu anonyymiystaso saavutetaan.

5.2 Suositukset

– Joillakin anonymisointitekniikoilla on luontaisia rajoituksia. Rajoituksia on vakavasti harkittava, ennen kuin rekisterinpitäjä käyttää tiettyä tekniikkaa anonymisointiprosessin muotoiluun. Anonymisoinnin tarkoitus on otettava huomioon. Tarkoituksena voi olla turvata yksilön tietosuoja, kun tietoaineisto julkaistaan, tai mahdollistaa tietyn tiedon esiin saaminen tietoaineistosta.

– Mikään tässä asiakirjassa kuvatuista tekniikoista ei varmuudella täytä tehokkaan anonymisoinnin kriteerejä (eli yksilön erottaminen joukosta ei ole mahdollista, yksilöön liittyvien tietueiden yhdistäminen ei ole mahdollista eikä yksilöstä voida tehdä päätelmiä). Osa näistä riskeistä voidaan kuitenkin poistaa tietyllä tekniikalla kokonaan tai osittain, joten yksittäisen tekniikan soveltaminen tietyssä tilanteessa edellyttää huolellista suunnittelua, samoin tekniikoiden yhdistelmän soveltaminen siten, että tuloksena saadaan vahvempi anonymiteetti.

Jäljempänä olevassa taulukossa esitetään yleiskatsaus tekniikoiden vahvuuksiin ja heikkouksiin kolmen perusvaatimuksen kannalta:

	Onko joukosta erottaminen edelleen riski?	**Onko yhdistettävyys edelleen riski?**	**Onko päättely edelleen riski?**
Peitenimillä suojaaminen	Kyllä	Kyllä	Kyllä
Kohinan lisääminen	Kyllä	Ehkä ei	Ehkä ei
Korvaaminen	Kyllä	Kyllä	Ehkä ei
Aggregointi tai k-anonymiteetti	Ei	Kyllä	Kyllä
L-diversiteetti	Ei	Kyllä	Ehkä ei
Differentiaalinen yksityisyys	Ehkä ei	Ehkä ei	Ehkä ei
Tiivistysfunktio/tokenisaatio	Kyllä	Kyllä	Ehkä ei

Taulukko 6. Tekniikoiden vahvuudet ja heikkoudet.

– Ihanteellinen ratkaisu olisi valittava tapauskohtaisesti. Ratkaisu (eli täydellinen anonymisointiprosessi) antaa suojan kaikkia kolmea riskiä vastaan. Se suojaa vahvasti siltä mahdollisuudelta, että rekisterinpitäjä tai muu kolmas osapuoli voisi tunnistaa rekisteröidyn käyttämällä kohtuullisesti toteutettavissa olevia keinoja.
– Jos ehdotus ei täytä jotakin näistä kriteereistä, tunnistamisriskit olisi arvioitava perusteellisesti. Arviointi olisi toimitettava viranomaiselle, jos kansallisessa laissa edellytetään, että viranomainen arvioi anonymisointiprosessin tai antaa siihen luvan.

Tunnistamisriskien vähentämiseksi olisi otettava huomioon seuraavat hyvät käytännöt:

Hyvät anonymisointikäytännöt

Yleistä:
– Älä käytä ”julkaise ja unohda” -lähestymistapaa. Tunnistamisen jäännösriski huomioon ottaen rekisterinpitäjän olisi
    - 1. tunnistettava uudet riskit ja arvioitava jäännösriski(t) uudelleen säännöllisesti
    - 2. arvioitava, onko tunnistettujen riskien valvonta riittävää, ja mukautettava tekniikkaa arvioinnin tulosten mukaisesti JA
    - 3. seurattava ja valvottava riskejä.
– Osana jäännösriskiä on otettava huomioon tietoaineiston (mahdollisesti) anonymisoimattoman osan aiheuttama tunnistamispotentiaali, varsinkin yhdistettynä anonyymiin osaan, samoin kuin attribuuttien välisten mahdollisten korrelaatioiden aiheuttama tunnistamispotentiaali (kuten maantieteellisen sijainnin ja vauraustason välinen korrelaatio).

Kontekstuaaliset elementit:
– Tietoaineiston anonymisoinnin tarkoitus on esitettävä selkeästi, koska se on keskeinen tekijä tunnistamisriskin määrittämisessä.
– Samassa yhteydessä on otettava huomioon kaikki asiaan vaikuttavat kontekstuaaliset elementit, kuten alkuperäisten tietojen luonne, käytössä olevat valvontamekanismit (myös turvatoimet tietoaineistojen käytön rajoittamiseksi), otoksen koko (kvantitatiiviset ominaisuudet), julkisten tietolähteiden saatavuus (vastaanottajien käytössä) ja tietojen suunniteltu luovuttaminen kolmansille osapuolille (rajoitettu, rajoittamaton esim. internetissä jne.).
– Mahdolliset hyökkääjät olisi otettava huomioon, samoin kuin se, ovatko tiedot houkutteleva hyökkäyskohde. Tietojen arkaluonteisuus ja tietojen luonne ovat myös tässä suhteessa avaintekijöitä.

Tekniset elementit:

– Rekisterinpitäjän olisi ilmoitettava, mitä anonymisointitekniikkaa tai tekniikkojen yhdistelmää on käytetty, varsinkin, jos se aikoo julkaista anonyymin tietoaineiston.
– Tietoaineistosta olisi poistettava ilmeiset (esim. harvinaiset) attribuutit / kvasitunnisteet.
– Jos (satunnaistamisessa) käytetään kohinan lisäämistä, tietueisiin lisättävän kohinan taso olisi määritettävä sen perusteella, mikä on attribuutin arvo (turhan laajaa kohinaa ei pidä lisätä) tai suojattavien attribuuttien vaikutus rekisteröityihin ja/tai kuinka niukka tietoaineisto on kyseessä.
– Jos (satunnaistamisessa) käytetään differentiaalista yksityisyyttä, kyselyjä on seurattava, jotta tietosuojaa loukkaavat kyselyt havaitaan, koska niiden tunkeutuvuus kasvaa kumulatiivisesti.

– Jos käytetään luokituksen karkeistamista, on perustavan tärkeää, että rekisterinpitäjä ei rajoitu yhteen karkeistuskriteeriin edes saman attribuutin kohdalla. Toisin sanoen olisi valittava erilaisia sijaintikarkeuksia tai erilaisia aikavälejä. Sovellettavan kriteerin valinnan on perustuttava attribuutin arvojen jakaumaan annetussa populaatiossa. Kaikkia jakaumia ei voida karkeistaa – toisin sanoen tekniikka ei tarjoa mitään yhtä ainoaa oikeaa ratkaisua kaikkiin tilanteisiin. Vaihtelevuus ekvivalenssiluokkien sisällä olisi varmistettava. Olisi esimerkiksi valittava erityinen kynnysarvo edellä mainittujen kontekstuaalisten elementtien perusteella (otoksen koko jne.), ja jos kynnysarvoa ei saavuteta, kyseinen otos olisi hylättävä (tai asetettava erilainen karkeistuskriteeri).

LIITE
Perustietoja anonymisointitekniikoista

A.1 Johdanto

Anonymiteettia tulkitaan eri puolilla EU:ta eri tavoin. Joissakin maissa se vastaa tietokoneistettua anonymiteettia (eli jopa rekisterinpitäjän yhteistyössä jonkun muun kanssa pitää olla vaikea tietokoneistetusti tunnistaa rekisteröityjä suoraan tai välillisesti) ja toisissa maissa täydellistä anonymiteettia (ts. jopa rekisterinpitäjän yhteistyössä jonkun muun kanssa pitää olla mahdotonta tunnistaa rekisteröityjä suoraan tai välillisesti). Anonymisoinnilla tarkoitetaan kuitenkin kummassakin tapauksessa prosessia, jolla tiedoista tehdään anonyymeja. Ero on siinä, millainen uudelleentunnistamisen riskitaso hyväksytään.

Anonyymeilla tiedoilla on erilaisia käyttötarkoituksia sosiaalisista tutkimuksista ja tilastollisista analyyseista uusien palvelujen/tuotteiden kehittämiseen. Toisinaan jopa tällaisilla yleishyödyllisillä toiminnoilla voi olla vaikutusta tiettyihin rekisteröityihin, mikä mitätöi käsiteltyjen tietojen oletetun anonymisoinnin. Tästä on olemassa monia esimerkkejä markkinointikampanjoiden käynnistämisestä julkisten toimenpiteiden toteuttamiseen käyttäjien profiloinnin, käyttäytymisen tai liikkuvuusmallien perusteella.[21]

Yleisten lausumien ohella ei valitettavasti ole olemassa kypsää mittaustapaa, jolla voitaisiin etukäteen arvioida uudelleentunnistamiseen käsittelyn jälkeen tarvittava aika tai työmäärä tai vaihtoehtoisesti valita sopivin menettely, jos halutaan vähentää sitä todennäköisyyttä, että julkaistu tietokanta voidaan liittää tunnistettuun rekisteröityjen joukkoon.

”Anonymisoinnin taito”, kuten siihen toisinaan viitataan tieteellisessä kirjallisuudessa[22], on uusi tieteenala, joka on vielä alkuvaiheissaan, ja tietoaineistojen tunnistamisen heikentämiseen käytetään monenlaisia toimintatapoja. On kuitenkin selvästi todettava, että suurin osa niistä ei estä yhdistämästä käsiteltyjä tietoja rekisteröityihin. Joissakin olosuhteissa anonyymeina pidettyjen tietoaineistojen tunnistaminen on osoittautunut hyvin menestyksekkääksi, ja toisissa tilanteissa on saatu virheosumia.

Käytettävissä on laajasti ottaen kaksi toimintamallia: toinen perustuu attribuuttien luokituksen karkeistamiseen, toinen satunnaistamiseen. Toimintamallien yksityiskohtien ja ominaispiirteiden läpikäyminen auttaa ymmärtämään tietojen tunnistamispotentiaalia uudella tavalla ja antaa uutta valoa itse henkilötietojen käsitteeseenkin.

A.2 Anonymisointi satunnaistamalla

Eräs anonymisoinnin vaihtoehto on muuttaa todellisia arvoja, jotta anonyymien tietojen ja alkuperäisten arvojen yhdistäminen estyy. Tavoite voidaan saavuttaa useilla erilaisilla menetelmillä kohinan lisäämisestä arvojen vaihtamiseen havaintojen välillä (permutaatioon). On korostettava, että attribuutin poistaminen vastaa kyseisen attribuutin satunnaistamisen äärimuotoa (jolloin attribuutti peittyy kokonaan kohinaan).

Joissakin olosuhteissa kokonaiskäsittelyn tavoitteena ei niinkään ole satunnaistetun tietoaineiston luovuttaminen kuin tietojen antaminen käyttöön kyselyiden kautta. Tällöin rekisteröidyn riski perustuu todennäköisyyteen, että hyökkääjä voi saada tietoja käyttämällä

[21] Esimerkkinä on TomTom Alankomaissa (ks. 2.2.3 kohdassa selitetty esimerkki).

[22] Gu, Jun, Chen, Yuexian, Fu, Junning, Peng, Huanchun ja Ye, Xiaojun: ”Synthesizing: Art of Anonymization”, Database and Expert Systems Applications -konferenssijulkaisu, *Lecture Notes in Computer Science*, nide 6261, Springer 2010, s. 385–399.

erillisten kyselyiden sarjaa rekisterinpitäjän tietämättä. Jotta yksilöiden anonymiteetti tietoaineistossa voidaan taata, on oltava mahdotonta päätellä, että rekisteröidyn tietoja on tietoaineistossa, jolloin yhteys hyökkääjän hallussa mahdollisesti oleviin taustatietoihin katkeaa.

Jos kyselyn vastaukseen lisätään tarvittava määrä kohinaa, uudelleentunnistamisen riski pienenee. Tämä lähestymistapa, jota kutsutaan kirjallisuudessa myös differentiaaliseksi yksityisyydeksi[23], eroaa edellä kuvatuista siinä, että rekisterinpitäjä voi paremmin valvoa tietojen käyttöä verrattuna yleiseen julkaisemiseen. Kohinan lisäämisellä on kaksi päätavoitetta: suojella tietoaineistossa olevien rekisteröityjen yksityisyyttä ja säilyttää luovutettujen tietojen käyttökelpoisuus. Kohinan määrän on erityisesti oltava oikeassa suhteessa kyselyjen määrään. Jos liian moniin yksilöitä koskeviin kyselyihin vastataan liian yksityiskohtaisesti, tunnistamisen todennäköisyys kasvaa. Satunnaistamisen onnistunutta soveltamista on nykypäivänä harkittava tapauskohtaisesti, sillä mikään tekniikka ei tarjoa idioottivarmaa menetelmää; tietoaineistoon sisältyvän tai sen ulkopuolisen rekisteröidyn attribuutteja koskevista tietovuodoista on esimerkkejä silloinkin, kun rekisterinpitäjä on pitänyt tietoaineistoa satunnaistettuna.

Asiaa saattavat valaista esimerkit satunnaistamisen mahdollisista puutteista anonymisointimenetelmänä. Yksityisyyttä kunnioittavina pidetyt kyselyt saattavat esimerkiksi interaktiivisessa käytössä muodostaa riskin rekisteröidyille. Jos hyökkääjä tietää, että yksilöiden alaryhmä S on tietoaineistossa, joka sisältää tietoja attribuutti A:n esiintymisestä populaatiossa P, voi pelkästään kaksi kysymystä esittämällä olla mahdollista määrittää (eron perusteella), kuinka monella yksilöllä alaryhmässä S on tosiasiassa attribuutti A – joko deterministisesti tai todennäköisyyteen perustuvalla päättelyllä. Kysymykset ovat ”Kuinka monella yksilöllä populaatiossa P on attribuutti A?” ja ”Kuinka monella yksilöllä populaatiossa P on attribuutti A, lukuun ottamatta alaryhmään S kuuluvia yksilöitä?” Joka tapauksessa alaryhmään S kuuluvien yksilöiden yksityisyys voi vaarantua vakavasti, varsinkin attribuutin A luonteen mukaan.

Voidaan myös katsoa, että jos rekisteröity ei ole tietoaineistossa, mutta hänen suhteensa tietoaineistossa oleviin tietoihin tunnetaan, tietoaineiston julkistaminen saattaa olla riski hänen yksityisyytensä kannalta. Jos esimerkiksi tiedetään, että kohteena olevan yksilön attribuutin A arvo eroaa määrällä X populaation keskiarvosta, hyökkääjä voi täsmälleen päätellä rekisteröityä koskevan henkilötiedon pelkästään pyytämällä tietokannan hoitajaa suorittamaan yksityisyyttä kunnioittavan operaation laskea attribuutin A keskiarvo.

Suhteellisten epätarkkuuksien lisääminen tietokannan todellisiin arvoihin on operaatio, joka on suunniteltava huolella. Yksityisyyden suojaamiseksi on lisättävä riittävästi kohinaa mutta samalla tarpeeksi vähän, jotta tietojen hyödyllisyys säilyy. Jos esimerkiksi sellaisten rekisteröityjen määrä on hyvin pieni, joilla on epätavallinen attribuutti, tai attribuutti on erittäin arkaluonteinen, voi olla parempi ilmoittaa vaihteluväli tai yleisluonteinen lause, kuten ”pieni määrä tapauksia, mahdollisesti ei yhtään” sen sijaan, että ilmoitetaan todellinen määrä. Tällä tavoin rekisteröidyn tietosuoja säilyy, vaikka kohinainen paljastusmekanismi tiedetään etukäteen, koska tietty määrä epävarmuutta jää jäljelle. Tulokset ovat edelleen hyödyllisiä tilastointia tai päätöksentekoa varten, jos epätarkkuus on huolella suunniteltu.

Tietokannan satunnaistaminen ja differentiaalisen yksityisyyden tarjoava pääsy vaativat lisäharkintaa. Ensinnäkin vääristymien oikea määrä voi vaihdella merkittävästi asiayhteyden

[23] Dwork, Cynthia: ”Differential Privacy”. Teoksessa *Automata, Languages and Programming. 33rd International Colloquium, ICALP 2006*, s. 1–12.

mukaan (kyselyn tyyppi, tietokannan populaation koko, attribuutin luonne ja siihen liittyvä tunnistamispotentiaali), eikä kaiken kattava ratkaisu ole mahdollinen. Lisäksi konteksti voi ajan mittaan muuttua, ja interaktiivista mekanismia olisi muutettava vastaavasti. Kohinan kalibrointi edellyttää, että seurataan kumulatiivisia tietosuojariskejä, joita interaktiivinen mekanismi aina aiheuttaa rekisteröidyille. Tietojen käyttömekanismeihin olisi liitettävä varoitusjärjestelmä, kun tietosuojan ”riskiraja” on saavutettu ja rekisteröidyt saattavat altistua erityisriskeille, jos uusi kysely tehdään. Tämä auttaisi rekisterinpitäjää määrittämään aina asianmukaisen tason, jolla todellisia henkilötietoja on vääristettävä.

Toisaalta on otettava huomioon myös tapaus, jossa attribuutin arvot poistetaan (tai niitä muutetaan). Yleisesti käytetty ratkaisu käsitellä attribuuttien epätyypillisiä arvoja on poistaa joko epätyypillisiin yksilöihin liittyvät tiedot tai epätyypilliset arvot. Viimeksi mainitussa tapauksessa on tärkeää varmistaa, että arvon puuttuminen ei itsessään muutu elementiksi, jonka avulla rekisteröity voidaan tunnistaa.

Seuraavaksi käsitellään satunnaistamista attribuutin korvaamisen avulla. Huomattava väärinymmärrys on samastaa anonymisointi salaukseen tai koodaukseen. Tämä harhakäsitys perustuu kahteen olettamaan: ensinnäkin, a) että tietue on anonyymi sen jälkeen, kun tietokannassa olevan tietueen joihinkin attribuutteihin (esim. nimi, osoite, syntymäaika) on sovellettu salausta tai kyseiset attribuutit on korvattu näennäisen satunnaisella merkkijonolla koodausoperaation, kuten avaimeen perustuvan tiivistysfunktion, avulla, ja toiseksi, b) että anonymisointi on tehokkaampaa, jos avaimen pituus on asianmukainen ja salausalgoritmi edustaa uusinta tekniikkaa. Harhakäsitys on rekisterinpitäjien keskuudessa laajalle levinnyt, joten se on syytä poistaa, samoin kuin harhakäsitykset, jotka liittyvät peitenimillä suojaamiseen ja sen väitettyihin muita tekniikoita vähäisempiin riskeihin.

Ensinnäkin kyseisten tekniikoiden tavoitteet ovat jyrkästi erilaiset: Salaus tietoturvakäytäntönä pyrkii suojaamaan viestintäkanavan luottamuksellisuutta tunnistettujen osapuolten (ihmisten, laitteiden tai ohjelmistojen/laitteistojen) välillä salakuuntelun tai tietojen tahattoman paljastamisen välttämiseksi. Koodauksella tarkoitetaan tietojen semanttista kääntämistä salaisen avaimen mukaan. Toisaalta anonymisoinnin tavoite on välttää yksilöiden tunnistaminen estämällä attribuuttien piilotettu yhdistäminen rekisteröityyn.

Salaus tai koodaaminen ei sellaisenaan sovellu rekisteröidyn tunnistamattomaksi tekemiseen: ainakin rekisterinpitäjällä alkuperäiset tiedot ovat edelleen käytettävissä tai johdettavissa. Jos henkilötiedot vain käännetään semanttisesti, kuten koodauksessa tapahtuu, se ei poista mahdollisuutta palauttaa tiedot alkuperäiseen muotoonsa – joko soveltamalla algoritmia päinvastaisessa suunnassa tai väsytysmenetelmähyökkäyksillä järjestelmien luonteen mukaan taikka tietomurron seurauksena. Uusinta tekniikkaa käyttävällä salauksella voidaan varmistaa tietojen korkeahkon asteen suojaus. Toisin sanoen tiedot ovat käsittämättömiä tahoille, jotka eivät tunne salausavainta, mutta salaus ei välttämättä johda anonymiteettiin. Rekisteröidyn tunnistamismahdollisuutta ei voida poistaa niin kauan kuin avain tai alkuperäiset tiedot ovat saatavilla (edes silloin, kun luotettu kolmas osapuoli on sopimuksella sitoutunut tarjoamaan turvallista avain-escrow-palvelua).

On harhaanjohtavaa keskittyä pelkästään salausmekanismin vahvuuteen tapana mitata anonymisointiastetta, koska monet muut tekniset ja organisatoriset tekijät vaikuttavat salausmekanismin tai tiivistysfunktion kokonaisturvallisuuteen. Kirjallisuudessa on raportoitu monia onnistuneita hyökkäyksiä, joissa algoritmi on onnistuttu kokonaan ohittamaan joko hyödyntämällä avainten hallinnan heikkouksia (esim. vähemmän turvallisen oletustilan olemassaolo) tai muilla inhimillisillä tekijöillä (esim. avainten palautukseen käytetyt heikot salasanat). Lopuksi on huomattava, että tiettyä avaimen kokoa käyttävä salausjärjestelmä on

suunniteltu varmistamaan luottamuksellisuus määrätyn ajan (useimpien nykyisin käytössä olevien avainten kokoa on muutettava vuoden 2020 paikkeilla), kun taas anonymiteettia ei pidä rajoittaa ajallisesti.

Seuraavaksi on syytä pohtia attribuuttien satunnaistamisen (tai korvaamisen tai poistamisen) rajoituksia käyttämällä esimerkkinä viime vuosina esiin tulleita epäonnistumisia, joissa on käytetty anonymisointia satunnaistamalla, ja tarkastelemalla niiden syitä.

Netflix-palkinto[24] on hyvin tunnettu esimerkki heikosti anonymisoidun tietoaineiston julkaisemisesta. Vaikka tietokannassa olevassa geneerisessä tietueessa osa rekisteröityyn liittyvistä attribuuteista on satunnaistettu, se voidaan silti osittaa kahdeksi alatietueeksi seuraavasti: {satunnaistetut attribuutit, puhtaat eli satunnaistamattomat attribuutit}, jossa puhtaat attribuutit saattavat olla mitä tahansa oletettavasti muita kuin henkilötietoja sisältävien tietojen yhdistelmiä. Netflix-palkintoa koskevassa tietoaineistossa voidaan tehdä se erityishavainto, että kutakin tietuetta voi edustaa piste moniulotteisessa tilassa, jossa jokainen puhdas attribuutti on koordinaatti. Tällä tekniikalla kutakin tietoaineistoa voidaan tarkastella pisteiden konstellaationa moniulotteisessa tilassa, joka voi olla erittäin harva, eli pisteet voivat olla kaukana toisistaan. Ne voivat olla jopa niin kaukana toisistaan, että sen jälkeen, kun tila on ositettu laajoiksi alueiksi, kullakin alueella on vain yksi tietue. Edes kohinaa lisäämällä tietueita ei saada niin lähelle toisiaan, että ne jakaisivat saman moniulotteisen alueen. Esimerkiksi Netflixin tapauksessa tietueet olivat riittävän ainutlaatuisia, koska 14 päivän välillä oli annettu vain kahdeksan arviointia. Alueiden välillä ei voitu havaita päällekkäisyyksiä sen jälkeen, kun sekä arvioihin että päivämääriin oli lisätty kohinaa. Toisin sanoen sama kahdeksan arvioidun elokuvan valikoima muodosti annettujen arvioiden sormenjäljen, joka ei ollut yhteinen kahdelle tietokannassa olevalle rekisteröidylle. Tutkijat yhdistivät tämän geometrisen havainnon perusteella oletettavasti anonyymit Netflixin tiedot toiseen elokuva-arviointeja sisältävään julkiseen tietokantaan (IMDB) ja löysivät siten käyttäjät, jotka olivat antaneet samoja elokuvia koskevia arvioita samalla aikavälillä. Koska suurin osa käyttäjistä vastasi toisiaan yksi yhteen, IMDB:n tietokannasta saadut lisätiedot voitiin tuoda Netflixin julkaistuun tietoaineistoon ja kaikkiin oletettavasti anonyymeihin tietueisiin voitiin liittää identiteetti.

On tärkeää korostaa, että tämä on yleinen ominaisuus: minkä tahansa satunnaistetun tietokannan jäännösosalla on edelleen erittäin korkea tunnistamispotentiaali jäännösattribuuttien yhdistelmän harvinaisuudesta riippuen. Rekisterinpitäjien olisi aina pidettävä tämä varoitus mielessään, kun satunnaistaminen valitaan tavaksi saavuttaa tavoiteltu anonymiteetti.

Monissa tämäntyyppisissä uudelleentunnistuskokeiluissa on noudatettu samanlaista lähestymistapaa ja projisoitu kaksi tietokantaa samaan alatilaan. Kyseessä on erittäin voimakas uudelleentunnistusmenetelmä, jota on viime aikoina sovellettu paljon eri aloilla. Esimerkiksi sosiaalista verkostoa vastaan suunnatussa tunnistuskokeilussa[25] käytettiin hyväksi sosiaalista kaaviota käyttäjistä, jotka oli suojattu peitenimillä tunnusten avulla. Tässä tapauksessa tunnistamiseen käytetty attribuutti oli kunkin käyttäjän yhteystietoluettelo, koska todennäköisyys, että kahdella yksilöllä on identtinen yhteystietoluettelo, on erittäin pieni. Tämän intuitiivisen olettaman perusteella todettiin, että hyvin rajoitetusta solmujen määrästä

[24] Narayanan, Arvind ja Shmatikov, Vitaly: ”Robust De-anonymization of Large Sparse Datasets”. Symposiumijulkaisussa *IEEE Symposium on Security and Privacy 2008*. IEEE 2008, s. 111–125.
[25] Backstrom, Lars, Dwork, Cynthia ja Kleinberg, Jon M.: ”Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography”, *Proceedings of the 16th International Conference on World Wide Web WWW'07*. ACM 2007, s. 181–190.

muodostuva sisäisten yhteyksien alakaavio on verkostoon piilotettu topologinen sormenjälki ja että suuri osuus koko sosiaalisesta verkostosta voidaan tunnistaa sen jälkeen, kun tämä alaverkosto on tunnistettu. Samanlaisten hyökkäysten suorituskyvystä voidaan antaa joitakin lukuja: on osoitettu, että yli neljä miljoonaa peitenimellä suojattua solmua ja 70 miljoonaa yhteyttä käsittävä sosiaalinen verkosto voidaan altistaa uudelleentunnistushyökkäyksille käyttämällä alle kymmentä solmua (joiden avulla voidaan muodostaa miljoona erilaista aliverkon konfiguraatiota, ja niistä jokainen muodostaa mahdollisesti topologisen sormenjäljen), ja tietoturva voi vaarantua erittäin suuressa määrässä yhteyksiä. On korostettava, että tätä uudelleentunnistamismenetelmää ei ole räätälöity erityisesti sosiaalisille verkostoille, vaan se on riittävän yleinen mukautettavaksi muihin tietokantoihin, joissa tallennetaan käyttäjien välisiä suhteita (kuten puhelinyhteyksiä, sähköpostikirjeenvaihtoa, treffisivustoja jne.).

Toinen tapa tunnistaa oletetusti anonyymi tietue perustuu kirjoitustyylin analyysiin (stylometria)[26]. Jäsennellyn tekstin metriikan tutkimiseksi on jo kehitetty erilaisia algoritmeja, muun muassa tietyn sanan käytön tiheys, tiettyjen kieliopillisten rakenteiden toistuminen ja välimerkitystyyppi. Kaikkien näiden ominaisuuksien avulla oletetusti anonyymi teksti voidaan yhdistää tunnistetun kirjoittajan kirjoitustyyliin. Tutkijat ovat poimineet kirjoitustyylin yli 100 000 blogista, ja he pystyvät nyt automaattisesti tunnistamaan blogimerkinnän kirjoittajan jo lähes 80 prosentin tarkkuudella. Tekniikan tarkkuuden odotetaan entisestään kasvavan, kun hyödynnetään myös muita signaaleja, kuten sijaintia tai muuta tekstiin sisältyvää metadataa.

Tutkimusyhteisön ja toimialan on paneuduttava tarkemmin tunnistamispotentiaaliin, joka liittyy tietueen semantiikan (eli tietueen satunnaistamattoman jäännösosan) hyödyntämiseen. Tuore tapaus (2013), jossa palautettiin dna:n luovuttajien henkilöyksiä,[27] osoittaa, että edistystä on tapahtunut vain vähän kuuluisan AOL-tapauksen jälkeen (2006). Silloin julkaistiin tietokanta, joka sisälsi yli 650 000 käyttäjän 20 miljoonaa avainhakusanaa kolmen kuukauden ajanjaksolta. Julkaiseminen johti usean AOL-käyttäjän tunnistamiseen ja paikantamiseen.

Sijaintitiedot ovat toinen tietojen ryhmä, joka pystytään harvoin anonymisoimaan pelkästään poistamalla rekisteröityjen henkilöydet tai salaamalla osittain joitakin attribuutteja. Ihmisten liikkumismallit saattavat olla riittävän ainutlaatuisia, jotta pelkästään sijaintitietojen semanttisen osan avulla (paikat, joissa rekisteröity oli tiettyyn aikaan) voidaan jopa ilman muita attribuutteja paljastaa monia rekisteröidyn ominaisuuksia.[28] Tämä on osoitettu edustavissa akateemisissa tutkimuksissa monesti.[29]

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

[27] Geneettiset tiedot ovat erityisen merkittävä esimerkki arkaluonteisista tiedoista, joita koskee uudelleentunnistamisen riski, jos ainoa menettely niiden anonymisointiin on luovuttajien henkilöyden poistaminen. Ks. edellä 2.2.2 kohdassa mainittu esimerkki. Ks. myös Bohannon, John: ”Genealogy Databases Enable Naming of Anonymous DNA Donors”, *Science*, nide 339, nro 6117, 18.1.2013, s. 262.

[28] Kysymys on otettu huomioon joidenkin jäsenvaltioiden lainsäädännössä. Esimerkiksi Ranskassa julkaistut sijaintitilastot anonymisoidaan luokitusta karkeistamalla ja vaihtamalla arvoja havaintojen välillä. INSEE julkaiseekin tilastot karkeistettuna aggregoimalla kaikki tiedot 40 000 neliömetrin alalle. Tietoaineiston rakeisuus riittää turvaamaan tietojen hyödyllisyyden, ja permutaatiot estävät anonymisoinnin purkamishyökkäykset harvaan asutuilla alueilla. Tämän ryhmän tietojen aggregointi ja arvojen vaihtaminen havaintojen välillä tarjoavat yleensä vahvat takeet päättely- ja anonymisoinnin purkamishyökkäyksiä vastaan (http://www.insee.fr/en/).

[29] de Montjoye, Yves-Alexandre, Hidalgo, César A., Verleysen, Michel ja Blondel Vincent D.: ”Unique in the Crowd: The privacy bounds of human mobility”. *Scientific Reports* 3, artikkelinro 1376, Nature 2013.

Tässä yhteydessä on välttämätöntä varoittaa peitenimien käytöstä riittävänä tapana suojata rekisteröityjä henkilöyden tai attribuuttien vuotoja vastaan. Jos peitenimillä suojaaminen perustuu henkilöyden korvaamiseen toisella ainutkertaisella koodilla, on naiivia olettaa, että se muodostaa vahvan suojan tunnistamista vastaan; tällöin ei oteta huomioon tunnistamismenetelmien monimutkaisuutta ja niitä moninaisia yhteyksiä, joissa menetelmiä voidaan soveltaa.

A.3 Anonymisointi karkeistamalla luokitusta

Attribuuttien luokituksen karkeistamiseen perustuvaa lähestymistapaa voidaan selventää yksinkertaisella esimerkillä.

Oletetaan, että rekisterinpitäjä päättää julkaista yksinkertaisen taulukon, jossa on kolme tietoa eli attribuuttia: tunnistenumero on ainutkertainen kullekin tietueelle, sijaintitunniste yhdistää rekisteröidyn paikkaan, jossa hän asuu, ja ominaisuustunniste osoittaa ominaisuuden, joka rekisteröidyllä on. Oletetaan lisäksi, että ominaisuus on toinen kahdesta erillisestä arvosta, joita yleensä merkitään seuraavasti: {P1, P2}.

Sarjatunniste	Sijaintitunniste	Ominaisuus
nro 1	Rooma	P1
nro 2	Madrid	P1
nro 3	Lontoo	P2
nro 4	Pariisi	P1
nro 5	Barcelona	P1
nro 6	Milano	P2
nro 7	New York	P2
nro 8	Berliini	P1

Taulukko A1. Otos rekisteröityjä, jotka on koottu sijainnin sekä ominaisuuksien P1 ja P2 perusteella.

Jos joku, jota tässä kutsutaan hyökkääjäksi, tietää ennakolta, että tietty rekisteröity (kohde), joka asuu Milanossa, on mukana taulukossa, hän saa taulukkoa tutkimalla tietää, että koska nro 6 on ainoa rekisteröity, jolla on kyseinen sijaintitunniste, hänellä on myös ominaisuus P2.

Tästä erittäin alkeellisesta esimerkistä käyvät ilmi tärkeimmät oletetusti anonymisoituun tietoaineistoon kohdistuvan tunnistusprosessin elementit. Hyökkääjä, jolla on joko sattumalta tai tarkoituksellisesti taustatietoja jostakin tai kaikista tietoaineistossa olevista rekisteröidyistä, pyrkii yhdistämään taustatiedot julkaistussa tietoaineistossa oleviin tietoihin saadakseen selkeämmän kuvan rekisteröityjen ominaisuuksista.

Jotta tietoja ei voisi yhdistää taustatietoihin yhtä tehokkaasti tai välittömästi, rekisterinpitäjä voi käsitellä sijaintitunnistetta ja korvata rekisteröidyn kotikaupungin laajemmalla alueella, kuten maalla. Tällä tavoin taulukko näyttäisi seuraavalta:

Sarjatunniste	Sijaintitunniste	Ominaisuus
nro 1	Italia	P1
nro 2	Espanja	P1
nro 3	Yhdistynyt kuningaskunta	P2
nro 4	Ranska	P1
nro 5	Espanja	P1
nro 6	Italia	P2
nro 7	Yhdysvallat	P2
nro 8	Saksa	P1

Taulukko A2. Taulukko A1:n luokituksen karkeistaminen asuinmaahan.

Tietojen luokituksen uuden karkeistamisen ansiosta hyökkääjä ei voi taustatietojensa (esim. ”kohde asuu Roomassa ja on mukana taulukossa”) perusteella tehdä selviä päätelmiä tunnistetun rekisteröidyn ominaisuuksista. Tämä johtuu siitä, että taulukossa olevilla kahdella italialaisella on eri ominaisuudet, toisella P1, toisella P2. Hyökkääjä jää puolittain epävarmaksi kohdeyksikön ominaisuudesta. Tämä yksinkertainen esimerkki osoittaa luokituksen karkeistamisen merkityksen anonymisoinnissa. Vaikka tällainen luokituksen karkeistaminen puolittaa tehokkaasti todennäköisyyden tunnistaa italialainen kohde, se ei suojaisi muista sijaintipaikoista (kuten Yhdysvalloista) olevia kohteita.

Lisäksi hyökkääjä voi silti saada tietoja espanjalaisesta kohteesta. Jos hyökkääjällä on taustatietona ”kohde asuu Madridissa ja on taulukossa” tai ”kohde asuu Barcelonassa ja on taulukossa”, hän voi sadan prosentin varmuudella päätellä, että kohteella on ominaisuus P1. Tästä syystä luokituksen karkeistaminen ei tuota samaa tietosuojan tasoa tietoaineiston koko populaatiossa eikä torju/estä/vaikeuta päättelyhyökkäyksiä koko populaatiota vastaan.

Tämän perusteella tulee helposti mieleen, että luokituksen vielä vahvempi karkeistaminen auttaisi estämään tietojen yhdistämisen – vaikkapa karkeistamalla luokitus maanosan tasolle. Tällä tavoin taulukko näyttäisi seuraavalta:

Sarjatunniste	Sijaintitunniste	Ominaisuus
nro 1	Eurooppa	P1
nro 2	Eurooppa	P1
nro 3	Eurooppa	P2
nro 4	Eurooppa	P1
nro 5	Eurooppa	P1
nro 6	Eurooppa	P2
nro 7	Pohjois-Amerikka	P2
nro 8	Eurooppa	P1

Taulukko A3. Taulukko A1:n luokituksen karkeistaminen maanosiin.

Tällä tavoin aggregoituina kaikki taulukossa olevat rekisteröidyt olisi suojattu yhdistämis- ja tunnistushyökkäyksiltä lukuun ottamatta yhtä, joka asuu Yhdysvalloissa. Tällöin taustatiedot, kuten ”kohde asuu Madridissa ja on taulukossa” tai ”kohde asuu Milanossa ja on taulukossa”, johtaisivat jonkinasteiseen todennäköisyyteen tietyn rekisteröidyn ominaisuudesta, ei suoraan

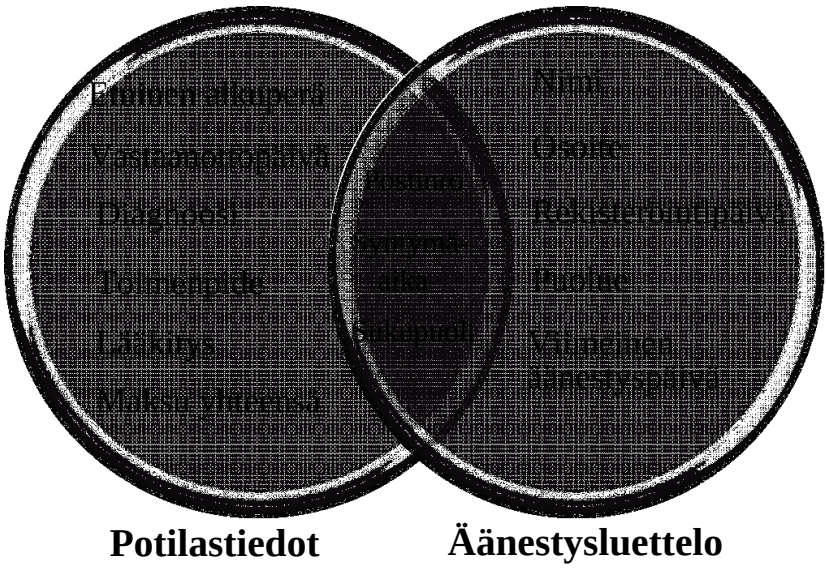
yhdistettävyyteen (P1:n todennäköisyys 71,4 prosenttia ja P2:n todennäköisyys 28,6 prosenttia). Tällainen luokituksen karkeistaminen tapahtuu ilmeisen ja perinpohjaisen tietojen menettämisen kustannuksella: taulukon avulla ei voida havaita mahdollista korrelaatiota ominaisuuksien ja sijainnin välillä, eli sitä, aiheuttaako tietty sijainti kenties jommankumman ominaisuuksista suuremmalla todennäköisyydellä. Taulukosta saa vain niin kutsutun marginaalisen jakauman eli ominaisuuksien P1 ja P2 esiintymisen absoluuttisen todennäköisyyden koko populaatiossa (esimerkissämme ensin mainitulla 62,5 prosenttia ja viimeksi mainitulla 37,5 prosenttia) ja kussakin maanosassa (Euroopassa kuten sanottua 71,4 prosenttia ja 28,6 prosenttia ja Pohjois-Amerikassa vastaavasti 100 prosenttia ja 0 prosenttia).

Esimerkki osoittaa myös, että luokituksen karkeistaminen vaikuttaa tietojen käytettävyyteen. Nykyään on jo käytettävissä joitakin suunnitteluvälineitä, joiden avulla voidaan etukäteen (eli ennen tietojen luovutusta) tutkia, mikä on asianmukaisin luokituksen karkeustaso, jotta rekisteröityjen tunnistamisriskiä voidaan pienentää vaikuttamatta liikaa luovutettujen tietojen hyödyllisyyteen.

K-anonymiteetti

Pyrkimystä estää yhdistettävyyshyökkäykset karkeistamalla attribuuttien luokitus kutsutaan k-anonymiteetiksi. Käytäntö juontuu uudelleentunnistamiskokeilusta 1990-luvun lopulla, jolloin terveydenhuoltoalalla toimiva yhdysvaltalainen yksityinen yritys julkaisi oletetusti anonyymin tietoaineiston. Anonymisointi oli tehty poistamalla rekisteröityjen nimet, mutta tietoaineisto sisälsi edelleen terveystietoja ja muita attribuutteja, kuten postinumeron (sijaintitunnisteet rekisteröityjen asuinpaikoista), sukupuolen ja täydellisen syntymäajan. Samat kolme tietoa {postinumero, sukupuoli, täydellinen syntymäaika} oli sisällytetty myös muihin julkisesti saatavilla oleviin rekistereihin (kuten äänestysluetteloihin), joten akateeminen tutkija pystyi käyttämään niitä yhdistääkseen tiettyjen rekisteröityjen henkilöyden julkaistun tietoaineiston attribuutteihin. Hyökkääjällä (tutkijalla) saattoi olla seuraavat taustatiedot: ”Tiedän, että äänestysluettelossa oleva rekisteröity, johon liittyy tietyt kolme tietoa {postinumero, sukupuoli, syntymäaika}, on ainutlaatuinen. Julkaistussa tietoaineistossa on tietue, jolla on kyseiset kolme tietoa.” Empiirisesti havaittiin[30], että suuri enemmistö (yli 80 prosenttia) tässä kokeessa käytetyssä julkisessa rekisterissä olevista rekisteröidyistä liittyi yksiselitteisesti tiettyyn kolmen tiedon ryhmään, jolloin tunnistaminen oli mahdollista. Näin ollen tietoja ei tässä tapauksessa ollut anonymisoitu asianmukaisesti.

[30] Sweeney, Latanya: ”Weaving Technology and Policy Together to Maintain Confidentiality”. *Journal of Law, Medicine & Ethics*, nide 25, nro 2–3, 1997, s. 98–110.



Kuva A1. Uudelleentunnistaminen tietoja yhdistämällä.

On esitetty, että vastaavien yhdistämishyökkäysten tehon vähentämiseksi rekisterinpitäjien olisi ensin tarkastettava tietoaineisto ja ryhmitettävä attribuutit, joiden avulla hyökkääjä voi kohtuuden rajoissa yhdistää julkistetun taulukon muuhun lisätiedonlähteeseen. Jokaisessa ryhmässä olisi oltava ainakin *k* identtistä karkeistettujen attribuuttien yhdistelmää (eli sen pitäisi edustaa attribuuttien ekvivalenssiluokkaa). Tietoaineistot olisi julkaistava vasta sen jälkeen, kun ne on ositettu tällaisiin homogeenisiin ryhmiin. Luokituksen karkeistukseen valittuja attribuutteja kutsutaan kirjallisuudessa yleisesti kvasitunnisteiksi, koska niiden tunteminen salaamattomassa muodossa aiheuttaisi rekisteröityjen välittömän tunnistamisen.

Monet tunnistamiskokeet ovat osoittaneet huonosti suunniteltujen k-anonymisoitujen taulukoiden heikkoudet. Näin voi käydä muun muassa, koska ekvivalenssiluokan muut attribuutit ovat identtisiä (kuten taulukossa A2 annetussa esimerkissä espanjalaisten rekisteröityjen ekvivalenssiluokka) tai niiden jakauma on hyvin epätasainen siten, että tietty attribuutti on erittäin vallitseva, tai koska ekvivalenssiluokan tietueiden lukumäärä on hyvin pieni. Kummassakin tapauksessa todennäköisyyteen perustuva päättely on mahdollinen. Voi olla myös, että ekvivalenssiluokkien salaamattomien attribuuttien välillä ei ole mitään merkittävää semanttista eroa. Tällaiset attribuutit saattavat esimerkiksi olla määrältään tosiasiassa erilaisia mutta numeerisesti hyvin lähellä toisiaan, tai attribuutit voivat kuulua semanttisesti samanlaisiin attribuuttiryhmiin, kuten samaan luottoriskin vaihteluväliin tai samaan tautiopilliseen ryhmään. Näin ollen tietoaineistosta voi yhdistämishyökkäyksissä edelleen vuotaa suuri määrä rekisteröityjä koskevia tietoja[31]. Tässä on tärkeää huomata, että aina kun tiedot ovat harvalukuisia (esim. tiettyä ominaisuutta esiintyy maantieteellisellä alueella vähän) eikä ensimmäinen aggregointi pysty ryhmittämään tietoja riittävään määrään eri ominaisuuksien esiintymiä (esim. vain muutamaa ominaisuutta esiintyy edelleen

[31] On korostettava, että korrelaatioita voidaan tehdä myös sen jälkeen, kun tiedot on ryhmitelty attribuuteittain. Kun rekisterinpitäjä tietää, minkätyyppisiä korrelaatioita hän haluaa tarkastaa, hän voi valita kaikkein merkityksellisimmät attribuutit. Esimerkiksi PEW:n tutkimustuloksiin ei voida kohdistaa hienojakoisia päättelyhyökkäyksiä, mutta ne ovat silti erittäin hyödyllisiä korrelaatioiden löytämiseksi väestörakenteen ja mielenkiinnon kohteiden välillä (http://www.pewinternet.org/Reports/2013/Anonymity-online.aspx).

maantieteellisellä alueella pieni määrä), attribuuttien aggregointia on lisättävä, jotta tavoiteltu anonymiteetti saavutetaan.

L-diversiteetti

Edellä kuvattujen havaintojen perusteella on vuosien mittaan ehdotettu k-anonymiteetin muunnoksia, ja luokituksia karkeistamalla toteutettavaa anonymisointia varten on kehitetty joitakin suunnittelukriteerejä, joiden tavoitteena on vähentää yhdistämishyökkäysten riskejä. Ne perustuvat tietoaineistojen probabilistisiin ominaisuuksiin. Menetelmässä lisätään uusi rajoite eli se, että ekvivalenssiluokan jokaisen attribuutin on esiinnyttävä vähintään *l* kertaa, jotta hyökkääjä jää aina attribuuttien suhteen huomattavan epävarmaksi, vaikka hänellä olisikin taustatietoa tietystä rekisteröidystä. Tämä tarkoittaa, että tietoaineistossa (tai sen osassa) on oltava vähimmäismäärä valitun ominaisuuden esiintymiä, jolloin uudelleentunnistamisen riski lievenee. Se on *l*-diversiteettiä käyttävän anonymisoinnin tavoite. Esimerkki tästä käytännöstä annetaan taulukoissa A4 (alkuperäiset tiedot) ja A5 (käsittelyn tulos). On ilmeistä, että attribuuttien luokituksen karkeistus lisää merkittävästi epävarmuutta jokaisen kyselyyn osallistuvan rekisteröidyn todellisista attribuuteista, jos taulukossa A4 olevien yksilöiden sijaintitunnisteet ja ikäluokitus suunnitellaan asianmukaisesti. Vaikka hyökkääjä esimerkiksi tietäisi, että rekisteröity kuuluu ensimmäiseen ekvivalenssiluokkaan, hän ei voi olla varma, onko henkilöllä ominaisuus X, Y tai Z, koska kyseisessä luokassa (ja kaikissa muissakin ekvivalenssiluokissa) on vähintään yksi tietue, jolla on nämä ominaisuudet.

Sarjanumero	Sijaintitunniste	Ikä	Ominaisuus
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

Taulukko A4. Taulukko, jossa yksilöt on ryhmitelty sijainnin, iän ja kolmen ominaisuuden X, Y ja Z perusteella.

Sarjanumero	Sijaintitunniste	Ikä	Ominaisuus
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

Taulukko A5. Esimerkki taulukon A4 versiosta, jossa on sovellettu l-diversiteettiä.

T-läheisyys

T-läheisyydeksi kutsutun lähestymistavan avulla käsitellään erityistapausta, jossa osion attribuutit ovat jakautuneet epätasaisesti tai kuuluvat arvojen tai semanttisten merkitysten pieneen vaihteluväliin. Se merkitsee lisäparannusta anonymisointiin luokitusta karkeistamalla, ja siinä tiedot järjestellään ekvivalenssiluokiksi, jotka kuvastavat mahdollisimman tarkoin attribuuttien alkuperäistä jakaumaa alkuperäisessä tietoaineistossa. Tässä tarkoituksessa käytetään periaatteessa seuraavaa kaksivaiheista menettelyä. Taulukko A6 on alkuperäinen tietokanta. Se sisältää rekisteröityjä koskevat salaamattomat tietueet, jotka on ryhmitelty sijainnin, iän, palkan ja kahden semanttisesti samanlaisen ominaisuuksien ryhmän perusteella. Niitä ovat ryhmät (X1, X2, X3) ja (Y1, Y2, Y3) (esim. samanlaiset luottoriskiluokat, samanlaiset sairaudet). Taulukkoon sovelletaan ensin *l*-diversiteettiä siten, että $l = 1$ (taulukko A7). Tietueet ryhmitetään semanttisesti samanlaisiin ekvivalenssiluokkiin, joihin kohdennetaan heikko anonymisointi. Sen jälkeen taulukkoa käsitellään t-läheisyyden saavuttamiseksi (taulukko A8), jolloin jokaisessa osiossa on enemmän vaihtelua. Itse asiassa toisessa vaiheessa jokaiseen ekvivalenssiluokkaan sisällytetään tietueita kummastakin ominaisuuksien ryhmästä. On syytä huomata, että sijaintitunnisteen ja iän rakeisuus vaihtelee prosessin eri vaiheissa. Tämä tarkoittaa, että jokainen attribuutti saattaa vaatia erilaiset karkeistamiskriteerit, jotta tavoiteltu anonymiteetti saavutetaan, ja tämä vuorostaan edellyttää, että rekisterinpitäjät suunnittelevat prosessin erikseen ja hyödyntävät laskentateknisiä resurssejaan asianmukaisesti.

Sarjanumero	Sijaintitunniste	Ikä	Palkka	Ominaisuus
1	1127	29	30 000	X1
2	1112	22	32 000	X2
3	1128	27	35 000	X3
4	1215	43	50 000	X2
5	1219	52	120 000	Y1
6	1216	47	60 000	Y2
7	1115	30	55 000	Y2
8	1123	36	100 000	Y3
9	1117	32	110 000	X3

Taulukko A6. Taulukko, jossa yksilöt on ryhmitelty sijainnin, iän, palkan ja kahden ominaisuusryhmän perusteella.

Sarjanumero	Sijaintitunniste	Ikä	Palkka	Ominaisuus
1	11**	2*	30 000	X1
2	11**	2*	32 000	X2
3	11**	2*	35 000	X3
4	121*	> 40	50 000	X2
5	121*	> 40	120 000	Y1
6	121*	> 40	60 000	Y2
7	11**	3*	55 000	Y2
8	11**	3*	100 000	Y3
9	11**	3*	110 000	X3

Taulukko A7. Taulukon A6 versio, jossa on sovellettu *l*-diversiteettiä.

Sarjanumero	Sijaintitunniste	Ikä	Palkka	Ominaisuus
1	112*	< 40	30 000	X1
3	112*	< 40	35 000	X3
8	112*	< 40	100 000	Y3
4	121*	> 40	50 000	X2
5	121*	> 40	120 000	Y1
6	121*	> 40	60 000	Y2
2	111*	< 40	32 000	X2
7	111*	< 40	55 000	Y2
9	111*	< 40	110 000	X3

Taulukko A8. Taulukon A6 versio, jossa on sovellettu *t*-läheisyyttä.

On selkeästi todettava, että rekisteröityjen attribuuttien luokituksen karkeistamisen tavoite näin pitkälle kehitetyllä tavalla voidaan toisinaan saavuttaa vain pienessä määrässä tietueita mutta ei kaikissa. Hyvillä käytännöillä olisi varmistettava, että jokainen ekvivalenssiluokka sisältää monia yksilöitä eivätkä päättelyhyökkäykset ole mahdollisia. Lähestymistapa edellyttää joka tapauksessa sitä, että rekisterinpitäjät arvioivat syvällisesti saatavilla olevia tietoja ja eri vaihtoehtojen yhdistelmiä (esim. eri laajuisia vaihteluvälejä, erilaista sijainnin tai iän rakeisuutta jne.). Anonymiteettia ei toisin sanoen voida saavuttaa, jos luokitusta karkeistetaan siten, että rekisterinpitäjät oikopäätä ja summittaisesti yrittävät korvata tietueen attribuuttien analyyttiset arvot vaihteluväleillä. Sen sijaan tarvitaan yksityiskohtaisia kvantitatiivisia lähestymistapoja. On esimerkiksi arvioitava attribuuttien entropia kussakin osiossa tai mitattava alkuperäisten attribuuttien jakaumien ja kunkin ekvivalenssiluokan jakauman etäisyys.