Tampereen yliopisto

Saara Saaninkoski

MIKROAINEISTOJEN ANONYMISOINTI JA PALJASTUMISRISKIN ARVIOINTI

Informaatioteknologian ja viestinnän tiedekunta
Tampereen yliopisto
Kandidaatintutkielma
Tammikuu 2021

TIIVISTELMÄ

Saara Saaninkoski: Mikroaineistojen anonymisointi ja paljastumisriskin arviointi
Kandidaatintutkielma
Tampereen yliopisto. Informaatioteknologian ja viestinnän tiedekunta
Tietojenkäsittelytieteiden tutkinto-ohjelma
Tammikuu 2021

EU:n tietosuoja-asetuksen määritelmän mukaan henkilötiedot ovat tunnistettavissa olevaan luonnolliseen henkilöön liittyviä tietoja, ja niitä koskevat yksityisyyden suojaksi säädetyt käsittelyä, tallennusta sekä julkaisua koskevat säännökset. Anonymisoinnilla tarkoitetaan henkilötiedoista johdettua tietoa, jota on muokattu siten, ettei tietoa voida välillisestikään yhdistää yksittäiseen henkilöön. Anonymisoitu tieto ei siis sisällä henkilötietoja, eikä siihen näin ollen sovelleta henkilötietoja koskevia säännöksiä.

Anonymisointiin on kehitetty monia menetelmiä, joissa tietoja esimerkiksi poistetaan, yleistetään, tai niihin lisätään kohinaa. Anonymisoinnin tavoite on yksityisyyden turvaamisen lisäksi myös säilyttää alkuperäisen aineiston tietoarvo mahdollisimman tarkkana ja muuttumattomana. Riittävän suojaustason määrittäminen aineiston tietoarvo samalla mahdollisimman hyvin säilyttäen on kuitenkin haastava optimointiongelma. Kiinnostus anonymisointimenetelmien kehittämiseen on kasvanut 2000-luvulla muun muassa tiukentuneiden tietosuojasäännösten sekä jatkuvasti kehittyvien anonymisoinnin purkutekniikoiden vuoksi. Ihmisistä kerätään lisäksi yhä kasvavassa määrin valtavia määriä dataa esimerkiksi sosiaalisessa mediassa, ja kerätyn tiedon hyödyntämiseen on niin taloudellisia kuin tieteellisiä intressejä.

Tässä kirjallisuuskatsauksessa tutustutaan yksikkötason aineistojen eli mikroaineistojen sisältämien henkilötietojen suojaamiseen anonymisoinnin avulla. Tutkielman tavoite on selvittää, mitä ovat yleisimmät anonymisointimenetelmät, millaisia rajoituksia niillä on, sekä kuinka anonymisoidun aineiston paljastumisriskiä voidaan arvioida.

Tutkielman johtopäätös on, että aineistot ovat niin monimuotoisia, ettei voida sanoa minkään menetelmän soveltuvan kaikkiin tapauksiin, vaan menetelmä tulee valita aineiston tyypin sekä loppukäyttäjän tarpeiden mukaan. Paljastumisriskin arviointi on tärkeässä roolissa, kun anonymisoinnin onnistumista tai kestävyyttä arvioidaan. Tietosuojan säilymiseksi on tärkeää, että tiedon käsittelijöillä on riittävästi asiantuntemusta eri anonymisointimenetelmien käytöstä sekä anonymisoidun aineiston yksityisyyden tason arvioinnista.

Avainsanat: anonymisointi, tietosuojamenetelmät, paljastumisriski, tunnistamisriski, tietosuoja, yksityisyys

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck –ohjelmalla.

SISÄLLYSLUETTELO

1. Johdanto 1

2. Anonymisoinnin taustaa 2

2.1. Käsitteitä 2

2.2. Tunnisteellinen tieto 3

2.3. Paljastumisriski ja informaatiokato 4

3. Yksityisyysmallit 5

3.1. K-anonymiteetti ja sen laajennokset 5

3.2. Differentiaalinen yksityisyys 8

4. Anonymisointimenetelmiä 9

4.1. Rajoittavat menetelmät 10

4.2. Satunnaistavat menetelmät 10

4.3. Synteettinen tieto 11

5. Anonymisointimenetelmien haavoittuvuuksia 12

5.1 K-anonymiteetin ja l-diversiteetin haavoittuvuuksia 12

5.2 Esimerkki puutteellisesta anonymisoinnista: The Netflix Price 14

6. Paljastumisriskin arviointi 15

6.1. Paljastumisriskin määreitä 15

6.2. Ainutlaatuisuus 16

6.3. Tietueiden yhdistämismallit 17

7. Pohdinta 18

8. Yhteenveto 19

9. Lähteet 21

1. JOHDANTO

Teknologia mahdollistaa suurten tietojoukkojen keräämisen ja analysoinnin tehokkaammin ja taloudellisemmin kuin koskaan ennen. Kerätystä datasta voidaan saada arvokasta tietoa, jota voidaan hyödyntää niin kaupallisesti kuin tutkimustarkoituksissa. Kerätty data sisältää kuitenkin usein yksittäisiin henkilöihin liittyviä tietoja, jolloin on otettava huomioon henkilötietojen käsittelyä ja säilyttämistä koskeva lainsäädäntö.

EU:n yleisessä tietosuoja-asetuksessa (EU, 2016/679) määritellään henkilötiedot sellaisiksi tiedoiksi, joiden perusteella luonnollinen henkilö voidaan tunnistaa suoraan tai välillisesti esimerkiksi tietoja yhdistelemällä. Riittävällä tasolla anonymisoidut tiedot eivät kuitenkaan ole asetuksen mukaan henkilötietoja, eivätkä asetuksen säännöt koske tällaisten tietojoukkojen käsittelyä ja säilytystä.

Anonymisointi tarkoittaa erilaisia toimenpiteitä, joita tietojoukolle suoritetaan henkilötiedoiksi luettujen tunnisteiden häivyttämiseksi. Anonymisoinnissa tärkeää on minimoida paljastumisriski eli arkaluontoisten tai tunnistavien tietojen paljastuminen, sekä maksimoida tiedon informaatioarvon pysyvyys. Riittävän suojaustason määrittäminen tietoarvo säilyttäen on haastava optimointiongelma, jota on tutkittu paljon erityisesti 2000-luvulla. Massadatan aikakaudella kiinnostus anonymisointia kohtaan on kasvussa, sillä teknisesti on helppoa kerätä mitä monimuotoisempaa tietoa, jonka hyödyntämisen sellaisenaan alati kiristyvät tietosuojasäännöt kuitenkin estävät.

Vaikka anonymisointia tutkitaan paljon, ei mitään yleispätevää keinoa tunnistavien tietojen häivyttämiseen ole keksitty. Anonymisoitavat aineistot ovat niin monimuotoisia, ettei voida sanoa minkään menetelmän sopivan kaikissa tapauksissa, vaan menetelmä tai menetelmien yhdistelmä tulee valita aineiston tyypin sekä loppukäyttäjän tarpeiden mukaan. Paljastumisriskin arviointi on tärkeässä roolissa, kun anonymisoinnin onnistumista ja kestävyyttä arvioidaan. Tietosuojan säilymiseksi onkin tärkeää, että tietojen käsittelijöillä on tietoa eri anonymisointimenetelmien soveltuvuudesta eri tietotyypeille sekä kykyä arvioida anonymisoidun aineiston paljastumisriskiä.

Tarkastelen tässä tutkielmassa anonymisointimenetelmiä mikroaineistojen eli yksikkötason aineistojen näkökulmasta. Tutkielman tavoite on selvittää, millaisia ovat yleisimmät anonymisointimenetelmät, mihin tarkoituksiin ne soveltuvat ja millaisia rajoitteita niillä on. Lisäksi tarkastelen, miten paljastumisriskiä voidaan arvioida. Tutkielman tarkoituksena on antaa käytännönläheinen katsaus anonymisoinnin toteutuksessa ja arvioinnissa käytettyihin menetelmiin tarkkojen formaalien määritelmien sijaan.

Tutkielman aluksi luvussa 2 määrittelen tutkielman kannalta relevantteja anonymisointiin ja paljastumisriskin arviointiin liittyviä käsitteitä. Luvussa 3 esittelen muutamia tunnetuimpia yksityisyysmalleja, joita on kehitetty julkaistavien aineistojen yksityisyyden vaatimusten määrittelyn avuksi. Neljännessä luvussa käsittelen kolmen tyyppisiä anonymisointimenetelmiä: rajoittavia menetelmiä, satunnaistavia menetelmiä sekä synteettistä tietoa. Luvussa 5 käsittelen eräitä anonymisointimenetelmien haavoittuvuuksia ja esittelen tapausesimerkin puutteellisesti anonymisoidun aineiston julkaisemisesta. Luvussa 6 paneudun syvemmin paljastumisriskin arvioimiseen jälkikäteen, eli kun anonymisointi on jo suoritettu. Kaksi viimeistä lukua koostuvat pohdinnasta ja yhteenvedosta.

2. ANONYMISOINNIN TAUSTAA

Tässä luvussa esittelen tässä tutkielmassa käytettyjä anonymisointiin ja paljastumisriskin arviointiin liittyviä käsitteitä. Tunnisteellisen tiedon käsitettä käsitellään omassa aliluvussa, jossa avataan myös anonymisoinnin lainsäädännöllistä kontekstia. Lisäksi esittelen paljastumisriskin sekä informaatiokadon käsitteet ja avaan anonymisoinnin määritelmää näiden määreiden optimointiongelmana.

2.1. Käsitteitä

Tunnisteellinen tieto tarkoittaa sellaista tietoa, joka voidaan suoraan tai välillisesti esimerkiksi muihin tietoihin yhdistämällä liittää yksittäiseen henkilöön. *Anonymisointi* tarkoittaa niitä toimenpiteitä, joilla näitä tunnisteellisia tietoja pyritään poistamaan tai häivyttämään, jottei julkaistusta tietoaineistosta voida tunnistaa yksittäistä henkilöä. Tunnisteellista tietoa käsitellään tarkemmin aliluvussa 2.2.

Tilastollinen tietosuoja on tilastotieteen osa-alue, joka tutkii niitä menetelmiä, joiden avulla voidaan varmistaa, ettei julkaistava tietoaineisto sisällä tunnisteellisia tietoja. Tilastolliseen tietosuojaan kuuluu paitsi ne tilastolliset menetelmät, joilla tietoja anonymisoidaan, myös ne tilastolliset menetelmät, joilla *paljastumisriskiä* voidaan arvioida. Paljastumisriski tarkoittaa sitä todennäköisyyttä, että anonymisoidusta tietojoukosta voidaan tunnistaa henkilö tai henkilön arkaluonteisia tietoja. Paljastumisriskiä käsitellään tarkemmin luvussa 2.3. *Tunnistaminen* viitataan henkilön identiteetin paljastumiseen anonymisoidusta tietojoukosta, kun taas *paljastuminen* voi viitata identiteetin paljastumisen lisäksi henkilöön liittyvän muuttujan arvon paljastumiseen.

Anonymisointimenetelmien haavoittuvuuksien yhteydessä *vastustajalla* viitataan sellaiseen tahoon, joka pyrkii paljastamaan anonymisoiduista tiedoista henkilöitä tai henkilöön liittyviä arkaluonteisia tietoja. *Ulkopuolisella tiedolla* tarkoitetaan vastustajalla olevaa tietoa, joka on saatu esimerkiksi henkilökohtaisen suhteen tai julkisen tiedon kautta.

Tässä tutkielmassa käsitellään *mikroaineistojen* eli yksikkötason aineistojen anonymisointimenetelmiä, erotuksena esimerkiksi taulukkodataan.

2.2. Tunnisteellinen tieto

EU:n yleisessä tietosuoja-asetuksessa (EU, 2016/679) määritellään henkilötiedoiksi kaikki tunnistettuun tai tunnistettavissa olevaan luonnolliseen henkilöön liittyvät tiedot. Asetuksessa tarkennetaan tunnistettavissa olevan henkilön tarkoittavan sitä, että henkilö voidaan tunnistaa ”erityisesti tunnistetietojen, kuten nimen, henkilötunnuksen, sijaintitiedon, verkkotunnistetietojen taikka yhden tai useamman hänelle tunnusomaisen fyysisen, fysiologisen, geneettisen, psyykkisen, taloudellisen, kulttuurillisen tai sosiaalisen tekijän perusteella” (EU, 2016/679, artikla 4, kohta 1). Tietosuoja-asetus rajoittaa henkilötietoja sisältävien aineistojen käsittelyä ja säilytystä.

Ne tekijät havaintokohteen tiedoista, jotka tekevät tietojoukosta tietosuoja-asetuksen mukaisesti henkilötietoja, ovat tunnisteita. Anonymisointi on tunnisteiden poistamista tai häivyttämistä siten, ettei tiedoissa esiintyvää henkilöä voida tunnistaa. Mikäli anonymisoitu tietoaineisto täyttää tämän vaatimuksen, ei sen katsota sisältävän henkilötietoja, eikä siihen näin ollen sovelleta tietosuoja-asetuksen henkilötietojen käsittelyä koskevia rajoituksia (EU, 2016/679, johdantolause). Tietosuoja-asetuksen mukaan henkilöä ei saa voida tunnistaa myöskään yhdistämällä anonymisoidun tietojoukon tietoja ulkopuolisista lähteistä löytyvään tietoon. Tämä tekee tunnisteiden identifioinnista erityisen haastavaa, sillä on vaikeaa arvioida tarkasti, mitä tämä ulkopuolinen tieto on.

Sweeneyn (2002) mukaan tunnisteet voidaan luokitella sen mukaan, miten suoraan ne viittaavat luonnolliseen henkilöön. Esimerkiksi henkilön koko nimi ja henkilötunnus ovat suoria tunnisteita, jotka riittävät yksin tunnistamaan henkilön. Ikä ja sukupuoli taas ovat epäsuoria tunnisteita, jotka eivät yksin riitä tunnistamiseen, mutta voivat yhdistettyinä mahdollistaa henkilön tunnistamisen.

2.3. Paljastumisriski ja informaatiokato

Anonymisoinnin tarkoitus on estää henkilön identiteetin tai henkilöön liittyvien tietojen paljastuminen tietoaineistojen julkaisun seurauksena, säilyttäen kuitenkin käyttäjän tarvitsemat tiedot riittävinä ja oikeellisina. Jotta tämä vaatimus voidaan täyttää, tulee ymmärtää hyödyn ja riskin suhde.

Tunnistaminen eli reidentifikaatio tarkoittaa henkilön tunnistamista anonyymiksi luokitellusta tietojoukosta. Tunnistaminen voi johtua paitsi puutteellisesti tehdystä anonymisoinnista, myös teknologian kehittymisestä tai lisääntyvästä ulkopuolisesta tiedosta. EU:n tietosuoja-asetuksen mukaan (EU, 2016/679) rekisterinpitäjä on vastuussa anonymiteetin säilymisestä, joten anonymisoinnin kestävyyttä on hyvä arvioida määräajoin.

Useita yksityisyysmalleja on kehitetty, jotta paljastumisriskiä voitaisiin paremmin hallita anonymisointia toteutettaessa. Tunnetuimpia näistä ovat *k*-anonymiteettiperiaate laajennoksineen sekä differentiaalinen yksityisyys, jotka on esitelty tarkemmin alaluvussa 2.3.

Paljastumisriskin lisäksi on luonnollisesti huomioitava tietojoukon informaatioarvon säilyminen. Mitä enemmän aineiston tietoja poistetaan tai vääristetään paljastumisriskin minimoimiseksi, sitä enemmän informaatioarvo kärsii. Domingo-Ferrerin ja Torran (2004) mukaan informaatiokadon määrä riippuu aineiston käyttötarkoituksesta. Parhaassa tilanteessa tiedetään aineiston käyttäjän tarve, ja suoritetaan anonymisointi sellaisilla metodeilla ja parametreilla, että käyttäjän tarpeen kannalta relevanttien tietojen poistaminen minimoidaan. Käyttäjää on kuitenkin usein vaikeaa määritellä tarkasti etukäteen, minkä lisäksi usean eri tavalla anonymisoidun version julkaiseminen samasta datasta lisää paljastumisen riskiä, joten informaatiokatoa tulee voida mitata yleisemmin.

Domingo-Ferrer ja Torra (2001) esittivät menetelmiä numeerisen ja luokitellun datan informaatiokadon mittaamiseen. Muuttuja on numeerinen, mikäli sillä voidaan suorittaa aritmeettisia operaatioita, kun taas luokitelluilla muuttujilla se ei ole mielekästä. Esimerkiksi ikä on numeerinen muuttuja, kun taas silmien väri on luokiteltu. Numeerisen datan informaatiokadon arviointi onnistuu tilastollisin menetelmin, kuten keskiarvon varianssia tarkastelemalla, mutta luokitelluille muuttujille tämä ei ole vaihtoehto. Luokitellun datan informaatiokadon arviointiin on esitetty muun muassa alkuperäisen ja anonymisoidun datan suoraa vertailua sekä alkuperäisen ja anonymisoidun datan ristiintaulukoiden vertailua.

3. YKSITYISYYSMALLIT

Anonymisoinnin perustana on jokin yksityisyysmalli, joka määrittelee anonymisoidulle aineistolle sallittavan tunnistamisriskin rajan. Yksityisyysmallit voidaan yleistää karkeasti kahteen luokkaan. Ensimmäiseen luokkaan kuuluvat *k*-anonymiteetti ja sen laajennukset kuten *l*-diversiteetti sekä *t*-läheisyys. Toiseen luokkaan kuuluvat ε-differentiaalinen yksityisyys sekä sen muunnokset kuten joukkoon sulautumisen yksityisyysmalli ja BlowFish. Tässä luvussa esittelen muutaman esimerkin kummastakin yksityisyysmalliluokasta.

3.1. *K*-anonymiteetti ja sen laajennokset

Samarati ja Sweeney (1998) kehittivät *k*-anonymiteetin käsitteen tutkiessaan tapoja, joilla tietojoukko voidaan anonymisoida säilyttäen alkuperäisen aineiston tiedot mahdollisimman tarkasti. *K*-anonymiteettimallissa tietojoukon attribuutit jaetaan suoriin tunnisteisiin, epäsuoriin tunnisteisiin sekä arkaluontoisiin attribuutteihin. *K*-anonymiteetin määritelmän mukaan tietojoukko toteuttaa *k*-anonymiteetin, jos jokaista tietuetta kohden on vähintään $k - 1$ muuta tietuetta, joiden epäsuorien tunnisteiden arvot ovat samat. Tämä saavutetaan poistamalla tai yleistämällä attribuuttien arvoja. Parametrin k arvoksi ei ole määritelty kaikille tietojoukoille pätevää lukua, vaan se määritellään tapauskohtaisesti. Mitä suurempi k on, sitä hankalampi anonymisointi on purkaa. Toisaalta on myös otettava huomioon, että mitä enemmän muuttujia joudutaan poistamaan tai yleistämään, sitä enemmän datan tietoarvo kärsii.

Sweeneyn (2002) mukaan arvojen yleistämisellä on useita tasoja, jotka voidaan arvottaa sen mukaan, kuinka paljon yleistämisoperaatio vaikuttaa anonymisoidun datan käyttökelpoisuuteen. Mitä matalammalla ollaan yleistyshierarkiassa, sitä lähempänä ollaan alkuperäisen datan tarkkuutta. Esimerkiksi postinumeron yleistyshierarkia matalimmasta korkeimpaan on 12345, $1234\ *$, $123\ **$, $12\ ***$, $1\ ****$, $*****$. Käytettäessä *k*-anonymiteettiperiaatetta tietojoukkojen anonymisointiin, on siis tärkeää paitsi identifioida epäsuorat tunnisteet ja arvioida sopiva k:n arvo käsiteltävälle tietojoukolle, myös punnita, mikä on kullekin tunnisteelle matalin *k*-anonymiteetin toteuttava yleistystaso.

Havainnollistetaan menetelmää seuraavan esimerkin avulla. Taulukossa 1 on esitetty potilastietoja sisältävä kuvitteellinen tietojoukko, jossa on 5 attribuuttia ja 8 tietuetta. Aloitetaan tunnistamalla tietojoukosta attribuuttien tyypit. Tässä tietojoukossa suora

tunniste on nimi, kun taas ikä, sukupuoli ja postinumero ovat epäsuoria tunnisteita. Diagnoosi taas on arkaluontoinen tieto, jota ei haluta pystyttävän yhdistämään yksittäiseen henkilöön.

TAULUKKO 1. Potilastietoja sisältävä kuvitteellinen otos

Nimi	Ikä	Sukupuoli	Postinumero	Diagnoosi
Mikko Mallikas	40	Mies	33310	Masennus
Pekka Kokkola	57	Mies	33410	Sydänsairaus
Linda Mökkönen	33	Nainen	34410	Sydänsairaus
Senja Kataja	34	Nainen	34240	Sydänsairaus
Mira Raittinen	31	Nainen	34410	Sydänsairaus
Petri Länsivuo	27	Mies	33230	Keliakia
Heikki Lundberg	29	Mies	33730	Syöpä
Petri Saarinen	29	Mies	33310	Masennus

Käsitellään ensin suorat tunnisteet. Tässä tietojoukossa on järkevintä poistaa nimiattribuutin arvo. Se olisi mahdollista myös yleistää esimerkiksi pelkäksi etunimeksi, mutta tällainen menettely kasvattaisi turhaan paljastumisriskiä parantamatta kuitenkaan aineiston tietoarvoa. Käsitellään seuraavaksi epäsuorat tunnisteet. Valitaan k:n arvoksi 2, eli yleistetään epäsuorien tunnisteattribuuttien arvoja siten, että yksikään rivi ei ole uniikki epäsuorien tunnisteiden suhteen. Käyttämällä matalinta mahdollista yleistyshierarkiaa, voimme yleistää ikäattribuutin arvot kolmeen luokkaan, ja postinumeron käsittämään kaksi ensimmäistä numeroa.

TAULUKKO 2. Iän, sukupuolen ja postinumeron suhteen 2-anonyymi otos

Nimi	Ikä	Sukupuoli	Postinumero	Diagnoosi
*	≥ 40	Mies	33***	Masennus
*	≥ 40	Mies	33***	Sydänsairaus
*	29 < ikä < 40	Nainen	34***	Sydänsairaus
*	29 < ikä < 40	Nainen	34***	Sydänsairaus
*	29 < ikä < 40	Nainen	34***	Sydänsairaus
*	≤ 29	Mies	33***	Keliakia
*	≤ 29	Mies	33***	Syöpä
*	≤ 29	Mies	33***	Masennus

Tarkastellaan seuraavaksi, miten taulukon 1 tietojoukko saadaan 3-anonymisoitua. Taulukossa 2 postinumero on 3-anonyymi sukupuolen suhteen, muttei iän suhteen. Ikää taas ei voida yleistää siten, että tietojoukko olisi iän ja sukupuolen suhteen 3-anonyymi. On siis poistettava joko ikä- tai sukupuoliattribuutti. Mikäli poistetaan sukupuoliattribuutti, tulee iän lisäksi myös postinumero yleistää. Valitaan sukupuoli poistettavaksi ja yleistetään ikä ja postinumero. Taulukossa 3 on esitetty tulokseksi saatu iän, sukupuolen ja postinumeron suhteen 3-anonyymi tietojoukko. Huomataan, että 3-anonyymissa tietojoukossa informaatioarvo on jo selvästi 2-anonyymia tietojoukkoa heikompi.

TAULUKKO 3. Iän, sukupuolen ja postinumeron suhteen 3-anonyymi otos

Nimi	Ikä	Sukupuoli	Postinumero	Diagnoosi
*	≥ 30	*	3****	Masennus
*	≥ 30	*	3****	Sydänsairaus
*	≥ 30	*	3****	Sydänsairaus
*	≥ 30	*	3****	Sydänsairaus
*	≥ 30	*	3****	Sydänsairaus
*	< 30	*	3****	Keliakia
*	< 30	*	3****	Syöpä
*	< 30	*	3****	Masennus

K-anonymiteettiperiaatteen käytöllä on monia hyviä puolia ja *k*-anonymiteettiin perustuvia menetelmiä käytetäänkin laajalti. *K*-anonymiteetin etuna on, ettei dataan tarvitse lisätä kohinaa, joten alkuperäinen tieto pysyy oikeellisena. Lisäksi dataa muokataan systemaattisesti, joten vastaanottajalle voidaan tietosuojaa vaarantamatta kertoa, mitä muuttujia on muokattu tai poistettu.

K-anonymiteetti ei kuitenkaan estä arkaluontoisen attribuutin arvon päättelemistä esimerkiksi tilanteessa, jossa henkilön tiedetään kuuluvan tiettyyn *k*-anonyymiin osajoukkoon. Machanavajjhala et al. (2006) kehittivätkin *l*-diversiteetin käsitteen täydentämään tätä *k*-anonymiteetin puutetta. *K*-anonyymin tietojoukon sanotaan toteuttavan *l*-diversiteetin, jos kullakin joukon *k*-anonyymilla osajoukolla on vähintään l eri arvoa kullekin arkaluontoiselle attribuutille. Arkaluontoisen attribuutin arvoilla on siis oltava riittävästi keskinäistä vaihtelua, jottei sitä voida yksiselitteisesti liittää *k*-anonyymiin osajoukkoon. Esimerkiksi taulukon 2 tiedoista voi vastustaja päätellä tietyllä 31-vuoti-

aalla naishenkilöllä olevan sydänsairaus, mikäli on tiedossa, että hänen tietonsa löytyvät taulukosta. Sen sijaan taulukon 3 tiedoista ei sairautta voida yksiselitteisesti päätellä, sillä myös masennus on mahdollinen diagnoosi; taulukko 3 toteuttaa 2-diversiteetin.

K-anonymiteettiperiaatteen jatkeeksi on kehitetty muitakin malleja, joissa pyritään parantamaan ja tarkentamaan jo olemassa olevia malleja. Esimerkiksi *t*-läheisyys (Li et al., 2007) täydentää *l*-diversiteettiä asettamalla vaatimuksen, että *k*-anonyymin, *l*-diversiteetin toteuttavan osajoukon arkaluontoisen attribuutin arvojen jakauman poikkeama koko tietojoukon vastaavasta jakaumasta on enintään t. Parametri t saa arvon väliltä [0, 1], jossa arvolla 0 osajoukon tilastolliset ominaisuudet vastaavat täydellisesti koko tietojoukon tilastollisia ominaisuuksia arkaluontoisen attribuutin suhteen, kun taas vastaavasti arvo 1 tarkoittaa täydellistä poikkeavuutta. *T*-läheisyys suojaa esimerkiksi tilastollisesti harvinaisten arvojen paljastumiselta eli ns. vinoushyökkäykseltä sekä huomioi tietoarvon säilymisen pelkkää *l*-diversiteettiä paremmin.

3.2. Differentiaalinen yksityisyys

Differentiaalinen yksityisyys (Dwork, 2006) poikkeaa *k*-anonymiteettimallista olennaisesti siten, että tavoitteena ei ole anonymisoida tietojoukkoa kokonaisuudessaan, vaan tuottaa käyttäjän tarpeen mukaan anonyymeja näkymiä tietojoukosta. Differentiaalisen yksityisyyden periaate on, että kyselyn tulos ei riipu siitä, onko yksittäinen tietue tietojoukossa vai ei. Tästä seuraa, että yksittäisten henkilöiden tietoja ei voi kyselyjen tulosten perusteella paljastua. Näkymien anonymisointi toteutetaan usein lisäämällä satunnaista kohinaa alkuperäisen tiedon sekaan muuttamatta alkuperäisiä tietoja.

Dwork (2006) määritteli artikkelissaan ε-differentiaalisen yksityisyyden käsitteen seuraavasti: olkoon ε positiivinen reaaliluku ja K satunnaistettu funktio, joka palauttaa kyselyn tuloksen sekä kohinaa. Merkitään lisäksi funktion K kuvajoukkoa $Im\,K$. Funktio K tuottaa ε-differentiaalisesti yksityisen tuloksen, mikäli kaikilla tietojoukoilla D ja D', jotka eroavat yhden tietueen suhteen, sekä $\forall S \subseteq Im\,K$ pätee, että

$$\Pr(K(D) \in S) \leq \exp(\varepsilon) \cdot Pr(K(D') \in S).$$

ε-differentiaalinen yksityisyys siis toteutuu, kun satunnaisen D:hen kohdistuvan kyselyn tulos todennäköisyys on enintään tai sama kuin e^{ε} kertaa satunnaisen D':hen kohdistuvan kyselyn tuloksen todennäköisyys. Tästä seuraa, että mitä pienempi ε:n arvo

on, sitä parempi on suoja tunnistamista vastaan. Toisaalta myös informaatiokato kasvaa ε:n pienentyessä: esimerkiksi 0-differentiaalisesti yksityinen tulos on pelkkää kohinaa, sillä $\varepsilon = 0 \rightarrow \Pr\big(K(D)\big) = \Pr\big(K(D')\big)$.

ε-differentiaalinen yksityisyys voi tarjota vahvan suojan tunnistamista vastaan, mutta se säilyttää infomaatioarvon huonosti. Sarathyn ja Muralidharin (2011) mukaan numeerisen datan muuntaminen ε-differentiaalisesti yksityiseksi kohinaa lisäämällä vaatii niin suuren määrän kohinaa, että on kyseenalaista, onko tulokseksi saadulla datalla enää mitään käyttöarvoa. Differentiaalisesta yksityisyydestä onkin laadittu uusia, vähemmän rajoittavia versioita, joista seuraavaksi esittelen lyhyesti joukkoon sulautumisen yksityisyysmallin sekä Blowfishin.

Joukkoon sulautumisen yksityisyysmallissa (engl. crowd-blending privacy, Gehrke et al. 2012) kevennetään differentiaalisen yksityisyyden vaatimusta, jotta informaatioarvo säilyisi alkuperäistä mallia paremmin. Tietojoukon sanotaan toteuttavan joukkoon sulautumisen yksityisyyden parametrilla k, mikäli kukin yksittäinen tietue i sulautuu k:n muun tietueen j joukkoon eli satunnaistetun funktion K tulos muuttuu korkeintaan e^{ε}-kertaiseksi, mikäli tietue i korvataan millä tahansa tietueella j. Toisin sanoen joukkoon sulautumisen yksityisyysmalli on kuten ε-differentiaalinen yksityisyys, mutta sen sijaan, että koko tietojoukon vaadittaisiin olevan differentiaalisesti yksityinen, muodostetaan k kokoisia differentiaalisesti yksityisiä joukkoja. Joukkoon sulautumisen malli yhdistää siis *k*-anonymiteetin idean differentiaaliseen yksityisyyteen.

Blowfish (He et al. 2014) on malli, jossa yleistetään differentiaalisen yksityisen vaatimusta. Siinä yksityisyyden ehto on sama kuin differentiaalisessa yksityisyydessä, mutta naapurijoukkojen määritelmä ei ole niin tiukka. Kun differentiaalisen yksityisyyden määritelmän mukaan joukko D sekä naapurijoukko D' eroavat yhden tietueen suhteen, Blowfish-mallissa sallitaan mikä tahansa naapuruuden määritelmä. Blowfish on joustava malli, joka säilyttää alkuperäisen informaatioarvon differentiaalista yksityisyyttä paremmin.

4. ANONYMISOINTIMENETELMIÄ

Anonymisointimenetelmät voidaan jakaa kahteen pääluokkaan: rajoittaviin (engl. non-perturbative) sekä satunnaistaviin (engl. perturbative) menetelmiin (Hundepool et al., 2012, s.3). Rajoittavilla menetelmillä tietoja poistetaan kokonaan tai vähennetään niiden yksityiskohtaisuutta luokittelemalla ne karkeammin. Satunnaistavilla menetelmillä

sen sijaan dataa muokataan siten, ettei voida olla varmoja, onko tietojoukosta satunnaisesti valittu arvo henkilöön liittyvää vai onko siihen lisätty satunnaisvaihtelua. Edellä mainittujen lisäksi käsittelen menetelmää, jossa luodaan synteettistä dataa alkuperäisen tietojoukon ominaisuudet säilyttäen.

4.1. Rajoittavat menetelmät

Rajoittavissa menetelmissä yksityisyyttä lisätään karkeistamalla tiedon tarkkuutta, mutta alkuperäiseen dataan ei lisätä mitään. Tästä seuraa, että tiedon tarkkuus heikkenee, mutta alkuperäinen tieto ei kuitenkaan vääristy anonymisointiprosessissa. Rajoittavia menetelmiä ovat muun muassa seuraavat (Hundepool et al., 2012, s. 51–53):

- Otanta on rajoittava menetelmä, jossa aineistosta julkaistaan vain rajattu otos. Jos otoskoko on pieni suhteessa tietojoukon kokoon, voi tämä menetelmä yksinään riittää luokitellun datan anonymisointiin, mikäli otos ei ole tietojoukon tasolla uniikki. Yleensä tämä tarkoittaa, että numeerista dataa ei tällä menetelmällä voida anonymisoida, sillä silloin ei-uniikin otoksen saaminen on vaikeaa.
- Yleistämisessä tietojen tarkkuutta vähennetään, jotta yksittäisen ihmisen tunnistaminen vaikeutuu. Luokiteltujen muuttujien kohdalla tämä tarkoittaa luokkien yhdistämistä uudeksi, yleisemmäksi luokaksi. Numeeristen muuttujien arvot taas yleistetään intervalleihin. Mikäli muuttujien arvot ovat järjestettävissä, voidaan ne yleistää järjestysluokkiin.
- Lokaali rajoittaminen tarkoittaa yksittäisten attribuuttien arvojen poistamista, jotta yksittäinen tietue ei olisi epäsuorilta tunnisteiltaan liian erottuva. Tätä käytetään erityisesti tietojoukossa harvinaisten arvojen häivyttämiseen.

K-anonymiteettiin perustuvat menetelmät ovat useimmiten rajoittavia. Esimerkiksi Samaratin ja Sweeneyn (1998) alkuperäinen menetelmä käytti yleistämisen ja lokaalin rajoittamisen yhdistelmää.

4.2. Satunnaistavat menetelmät

Satunnaistavissa menetelmissä dataan lisätään satunnaisvaihtelua, jotta tunnisteita ja arkaluontoisia tietoja saadaan häivytettyä. Satunnaistavat menetelmät perustuvat oletukseen, että satunnaisvaihtelun lisääminen ei muuta tietojoukon tilastollisia ominaisuuksia, joten informaatioarvo ei heikkene (Hundepool et al., 2012, s. 33). Satunnaistavia menetelmiä ovat muiden muassa seuraavat (Hundepool et al., 2012, s. 54–74):

- Satunnaisen kohinan lisäämisessä arkaluontoisten attribuuttien arvoja muunnetaan, jottei niiden alkuperäinen arvo paljastuisi. Yksinkertaisimmillaan tämä tarkoittaa esimerkiksi satunnaisen luvun lisäämistä kuhunkin arvoon siten, että alkuperäisen tietojoukon tilastolliset ominaisuudet pysyvät mahdollisimman muuttumattomina. Tämä menetelmä soveltuu vain numeeriselle datalle.
- Mikroaggregaatiossa tietojoukon tietueet jaetaan vähintään k tietueen ryhmiin, minkä jälkeen julkaistaan keskiarvoinen tietue kustakin ryhmästä. Ryhmät muodostetaan samankaltaisuuden perusteella, sillä mitä enemmän ryhmän jäsenet ovat samankaltaisia, sitä vähemmän informaatiokatoa tapahtuu keskiarvoista tietuetta laskettaessa. Mikroaggregaatiomenetelmiä on olemassa sekä numeeriselle että luokitellulle datalle.
- Arvojen vaihtaminen (engl. data swapping) on menetelmä, jossa yksittäisten tietueiden attribuuttien arvoja vaihdetaan keskenään siten, että arvojen frekvenssit pysyvät samoina. Arvojen vaihtaminen on suunniteltu luokitellun datan satunnaistamiseen.
- PRAM (post-randomisation method) on menetelmä, jossa muuttujien arvoja muutetaan tai jätetään ennalleen siirtymämatriisin todennäköisyyksien perusteella. PRAM-menetelmällä voidaan suojata ainoastaan luokiteltuja muuttujia.

Differentiaalisen yksityisyyden saavuttamiseen käytetään satunnaistavia menetelmiä. Esimerkiksi Dwork (2006) käyttää satunnaisen kohinan lisäämistä ε-differentiaalisen yksityisyyden saavuttamiseen.

4.3. Synteettinen tieto

Rubin (1993) esitti, että henkilötietojen suojaamiseksi alkuperäisestä tietojoukosta voitaisiin luoda kokonaan uusi, synteettinen tietojoukko, joka vastaa valituilta ominaisuuksiltaan alkuperäistä tietojoukkoa. Oletuksena on, että koska kokonaan synteettinen tietojoukko ei sisällä oikeita henkilötietoja, on yksittäisen ihmisen tunnistaminen teoriassa mahdotonta. Käytännössä on kuitenkin mahdollista, joskin hyvin epätodennäköistä, että synteettinen tietue vastaa sattumalta alkuperäistä tietuetta niin tarkasti, että tunnistaminen tapahtuu. Kokonaan synteettisen datan käyttöarvo on lisäksi rajallinen, sillä se luodaan tiettyä tarkoitusta varten, vain tietyt, ennalta valitut ominaisuudet säilyttäen, eikä alkuperäisen tietojoukon muita ominaisuuksia ole mahdollista tarkas-

tella synteettisestä datasta. Kokonaan synteettisen datan rajoitteisiin on vastattu kehittämällä menetelmiä osittain synteettisen datan ja niin sanotun hybrididatan luomiseksi.

Kun vain valitut tiedot - esimerkiksi tiettyjen arkaluntoisten tai epäsuorasti tunnistavien muuttujien arvot, tai tietyt kirjaukset - korvataan synteettisillä tiedoilla, on kyseessä osittain synteettinen data (Hundepool et al., 2012, s. 78). Osittain synteettisen datan paljastumisriski on kokonaan synteettistä dataa korkeampi ja altis yhdistämishyökkäykselle (luku 6.3), sillä se sisältää jonkin verran tosia tietoja.

Jos alkuperäisen tietojoukon tietueita ja synteettisesti luotuja tietueita yhdistetään, syntyy niin kutsuttua hybrididataa (Dandekar et al., 2002). Tämän menetelmän etuna on, että käsittelijä voi vaikuttaa synteettisten tietueiden yksittäisiin ominaisuuksiin enemmän, kuin jos tietojoukko olisi kokonaan synteettinen. Synteettisen datan määrää hybrididatassa voidaan säädellä, mikä tekee menetelmästä joustavan ja tarjoaa käsittelijälle mahdollisuuden vaikuttaa tietojoukon informaatiokadon tasoon.

5. ANONYMISOINTIMENETELMIEN HAAVOITTUVUUKSIA

Tietosuoja voi vaarantua anonymisoinnin kontekstissa kahdella tavalla: henkilön identiteetti voi paljastua tai henkilöön liittyvä ominaisuus voi paljastua. Identiteetin paljastuminen tarkoittaa sitä, että vastustaja pystyy yhdistämään julkaistun tietojoukon yksittäiseen henkilöön. Ominaisuuden paljastuminen taas tarkoittaa sitä, että vastustaja saa tietojoukosta lisää tietoa yksittäisestä henkilöstä. Identiteetin paljastumisesta seuraa usein myös ominaisuuksien paljastumista (Li et al., 2007, s.106). Tässä luvussa esittelen *k*-anonymiteetille ja *l*-diversiteetille tyypillisiä haavoittuvuuksia sekä esitän case-esimerkin epäonnistuneesta anonymisoinnista.

5.1 *K*-anonymiteetin ja *l*-diversiteetin haavoittuvuuksia

K-anonymiteetti antaa suojan identiteetin paljastumista vastaan, mutta ei suojaa ominaisuuksien paljastumiselta. Mikäli tietojoukon *k*-anonyymissa osajoukossa ei ole riittävästi vaihtelua arkaluontoisen attribuutin arvoissa, voi vastustaja saada arkaluontoisen tiedon selville, mikäli onnistuu päättelemään henkilön tietojen olevan osajoukossa. Machanavajjhala et al. (2007) nimesivät tämän tunnistamistavan *homogeenisyyshyökkäykseksi*. Esitetään homogeenisyyshyökkäys esimerkin avulla: Topi ja Liisa

ovat naapureita. Eräänä päivänä Liisa sairastuu niin pahasti, että hänet viedään ambulanssilla sairaalaan. Topi näkee ambulanssin ja uteliaana miehenä haluaa selvittää, mistä syystä Liisa sairastui. Oletetaan, että sairaala julkaisee anonyymeja potilastietoja ja luvussa 3 esitetty taulukko 2 kuvastaa eräänä päivänä lääkärin vastaanotolla käyneiden henkilöiden 2-anonymisoituja tietoja. Koska Liisa on nainen, pystyy Topi päättelemään, että Liisa vietiin sairaalaan sydänsairauden vuoksi.

Vastustaja voi myös onnistua päättelemään aiemman tietonsa avulla, mikä tietue sisältää tietyn henkilön tietoja. Tätä tunnistamistapaa Machanavajjhala et al. (2007) kutsuvat *taustatietohyökkäykseksi*. Esitetään esimerkki taustatietohyökkäyksestä jatkamalla edellistä esimerkkiä: Topilla on 40-vuotias setä, jonka Topi tietää käyneen samana päivänä samassa sairaalassa kuin Liisa. Iän perusteella Topi voi päätellä, että sedän tiedot sisältävä tietue on jompikumpi taulukon 2 kahdesta ensimmäisestä rivistä. Topi tietää myös, että hänen suvussaan ei isän puolelta ole kenelläkään sydänsairauksia. Koska sydänsairaudet ovat perinnöllisiä, Topi voi kohtuullisella varmuudella päätellä, että setä sairastaa masennusta.

L-diversiteetti kehitettiinkin suojaksi homogeenisyyshyökkäystä ja taustatietohyökkäystä vastaan (Machanavajjhala et al., 2007). Li ja muut (2007) huomasivat kuitenkin, että *l*-diversiteetti ei ole riittävä suoja tunnistamista vastaan silloin, kun arkaluontoisen attribuutin arvon suhteellinen frekvenssi *k*-anonyymin osajoukon sisällä poikkeaa arvon frekvenssistä koko tietojoukossa. Tämä *vinoushyökkäykseksi* nimetty tietojen paljastumistapa on erityisen suuri riski silloin, kun osajoukossa on harvinaisia attribuutteja. Otetaan esimerkiksi jokin harvinainen sairaus, jonka esiintyvyys tietojoukossa on 1 %. Anonymisoidaan tietojoukko siten, että osajoukot ovat *k*-anonyymeja ja toteuttavat *l*-diversiteetin. Oletetaan, että tällöin eräässä osajoukossa edellä mainittua harvinaista sairautta esiintyy joka toisessa kirjauksessa. Tällöin tämän attribuutin paljastumisen todennäköisyys kasvaa 1 %:sta 50 %:iin.

Li et al. (2007) huomauttivat myös, että *l*-diversiteetti ei huomioi arkaluontoisten attribuuttien arvojen mahdollista semanttista samankaltaisuutta. *Samankaltaisuushyökkäyksessä* (Li et al., 2007) vastustaja saa tietoa henkilöstä, kun hänet on luokiteltu osajoukkoon, jonka arkaluontoiset attribuutit ovat merkitykseltään samankaltaisia. Palataan esimerkkiin, jossa Topi päätteli Liisalla olevan sydänsairaus, sillä sydänsairaus esiintyi kaikissa Liisan tiedot sisältävän *k*-anonyymin osajoukon kirjauksissa. Jos osajoukon diagnoosiattribuutin arvot olisivatkin olleet ”rytmihäiriö”, ”sepelvaltimotauti” ja ”sydämen vajaatoiminta”, täyttäisi osajoukko *l*-diversiteetin. Tämä ei kuitenkaan estä

Topia päättelemästä Liisan sairautta, sillä kaikki arvoista viittaavat jonkinlaiseen sydänvikaan. Li et al. (2007) kehittivät *t*-läheisyyden suojaamaan vinoushyökkäykseltä sekä samankaltaisuushyökkäykseltä.

5.2 Esimerkki puutteellisesta anonymisoinnista: The Netflix Price

Eräs tunnetuimpia esimerkkejä epäonnistuneesta anonymisoinnista on suoratoistopalvelu Netflixin vuonna 2006 julkaisema Netflix Price -aineisto, joka sisälsi sata miljoonaa elokuva-arviota lähes puolelta miljoonalta käyttäjältä kuuden vuoden ajalta. Yksittäisestä kirjauksesta ilmenee, millä arvosanalla ja minä päivänä pseudonymisoitu käyttäjä oli arvioinut tietyn elokuvan. Julkaisun syynä oli Netflixin julistama kilpailu, jossa luvattiin miljoonan dollarin palkinto sille, joka pystyisi kehittämään suosittelualgoritmin, joka parantaisi Netflixin silloista elokuvasuosittelujärjestelmää vähintään kymmenellä prosentilla. Julkaistussa aineistossa käyttäjänimi oli korvattu uniikilla tunnisteella, josta käyttäjää ei pystytä suoraan tunnistamaan, mutta joka säilyttää yhteyden käyttäjän ja arvioitujen elokuvien välillä. Lisäksi aineisto anonymisoitiin lisäämällä kohinaa: arvioiden arvoja sekä arviointipäivämääriä muutettiin satunnaisesti. Netflix ilmoitti lisäksi, että julkaistu tietojoukko sisälsi vain alle kymmenesosan alkuperäisen tietojoukon kirjauksista. (Narayanan & Shmatikov, 2008)

Jo kaksi viikkoa aineiston julkaisun jälkeen Texasin yliopiston tutkijat Arvind Narayanan ja Vitaly Shmatikov ilmoittivat pystyvänsä tunnistamaan yksittäisten henkilöiden tietoja aineistosta. Narayanan ja Shmatikov osoittivat, että hyvin pieni määrä aiempaa taustatietoa riittää tunnistamaan henkilön anonymisoidusta tietojoukosta. He testasivat tunnistamisalgoritmiaan yhdistämällä Internet Movie Databasen käyttäjäarviotietoja Netflixin aineistoon yhteneväisyyksien löytämiseksi, mutta taustatiedon lähde voi olla mikä vain. Kokeellisen tutkimuksen tulos oli, että Narayananin ja Shmatikovin algoritmi pystyi tunnistamaan käyttäjän 68 % tapauksista, kun taustatietona on kaksi arviota ja päivämäärää, joille sallittiin kolmen päivän virhe. Jos taustatietona on 8 arvioita ja päivämäärää 14 päivän virheolettamalla, kirjauksista tunnistettiin 99 %. (Narayanan & Shmatikov, 2008)

Narayananin ja Shmatikovin mukaan Netflix-aineiston anonymisointimenetelmässä oli useita virheitä. Ensinnäkään julkaistu otos ei vaikuttanut olevan satunnaisotos, vaan otoksessa tunnuttiin suosivan yli 20 elokuvaa arvioineita käyttäjiä. Mitä enemmän yksittäisestä käyttäjästä on tietoja, sen enemmän kasvaa todennäköisyys, että yksittäisen käyttäjän arvioiden yhdistelmä on uniikki, ja sen helpompi käyttäjä on tunnistaa.

Toiseksi satunnaisuuden määrä aineistossa ei ollut riittävä. Narayananin ja Shmatikovin onnistui verrata pientä osaa Netflixin alkuperäisestä, satunnaistamattomasta tietojoukosta julkaistuun, satunnaistettuun tietojoukkoon. Kahden käyttäjän otoksella tulos oli, että toisella käyttäjällä 1/306 arviosta oli satunnaistettu ja toisella 5/229. Tästä ei toki voida varmasti päätellä, mikä on satunnaistettujen kirjausten määrä koko aineistossa, mutta tämän otoksen perusteella satunnaistamista olisi vaadittu huomattavasti enemmän. Kolmanneksi *k*-anonymiteettiä ei ollut huomioitu lainkaan aineiston anonymisoinnissa. (Narayanan & Shmatikov, 2008)

Narayananin ja Shmatikovin tutkimus osoittaa, että satunnaistaminen ei takaa automaattisesti yksityisyyttä, vaan se pitää suunnitella ja toteuttaa huolella. Luvussa 3.2. esitelty ε-differentiaalinen yksityisyys tarjoaa työkalun satunnaistetulla menetelmällä anonymisoidun tietojoukon yksityisyyden arviointiin. Satunnaistavat ja yleistävät menetelmät eivät myöskään ole toisiaan poissulkevia – satunnaistetun Netflix-aineiston yksityisyyttä olisi voitu lisätä varmistamalla myös *k*-anonymiteetin toteutuminen.

6. PALJASTUMISRISKIN ARVIOINTI

Anonymisoinnin tarkoitus on estää henkilön identiteetin tai henkilöön liittyvien tietojen paljastuminen tietoaineistojen julkaisun seurauksena, säilyttäen kuitenkin käyttäjän tarvitsemat tiedot riittävinä ja oikeellisina. Jotta tämä vaatimus voidaan täyttää, tulee tietoarvon säilymiselle sekä paljastumisriskille määritellä mittarit. Lisäksi, koska paljastumisriski on aina olemassa, tulee määritellä kullekin julkaistavalle tietojoukolle sopiva riskin taso. Tässä luvussa esittelen muutamia tunnettuja menetelmiä ja määreitä paljastumisriskin arviontiin.

6.1. Paljastumisriskin määreitä

Paljastumisriskin voidaan katsoa olevan totuusarvollinen tai mitallinen määre (Torra, 2017, s. 114). Luvussa 3 esitellyt yksityisyysmallit, kuten *k*-anonymiteetti ja differentiaalinen yksityisyys, ovat esimerkkejä totuusarvollisista paljastumisriskin määreistä. Mallit määrittelevät itsessään suurimman sallitun paljastumisriskin tason, joka joko toteutuu tai ei toteudu. Differentiaalisen yksityisyyden mallissa tätä tasoa kuvastaa parametri ε ja *k*-anonymiteettimallissa parametri k. On kuitenkin huomioitava, että differentiaalinen yksityisyys on ominaisuuksien paljastumisen estämiseen keskittyvä malli,

kun taas *k*-anonymiteetissa identiteetin paljastumisen estäminen on keskiössä. Tämän vuoksi nämä mallit eivät välttämättä anna riittävää kokonaiskuvaa.

Tietueen ainutlaatuisuus (engl. uniqueness) sekä tunnistaminen (engl. reidentification) ovat mitallisia paljastumisriskin määreitä. Ainutlaatuisuus kuvastaa todennäköisyyttä tunnistaa yksittäisen tietueen tietojen kuuluvan tietylle henkilölle sen perusteella, kuinka harvinainen on sen muuttujien arvojen yhdistelmä. Ainutlaatuisuutta ei ole mielekästä arvioida satunnaistettujen menetelmien yhteydessä, sillä kohinan lisäämisestä seuraa, ettei alkuperäisessä tietojoukossa oleva harvinainen yhdistelmä hyvin todennäköisesti esiinny anonymisoidussa tietojoukossa. Tunnistaminen tarkoittaa arviota niiden tietueiden lukumäärästä anonymisoidussa tietojoukossa, joista vastustaja voi tunnistaa henkilön identiteetin tai ominaisuuksia. Tätä lukumäärää voidaan arvioida tietueiden yhdistämismallien avulla. (Torra, 2017, s. 121)

Seuraavissa aliluvuissa käsitellään ainutlaatuisuuteen perustuvaa riskiarviota sekä onnistuneiden tunnistusten määrän arviointia tietueiden yhdistämismallien avulla.

6.2. Ainutlaatuisuus

Fienberg & Makov (1998) määrittelivät tietueen ainutlaatuisuuden tietueen muuttujien arvojen yhdistelmän harvinaisuudeksi koko tietojoukossa tai sen otoksessa. Koska tietue on sitä helpompi yhdistää yksittäiseen henkilöön, mitä ainutlaatuisempi se on, voidaan ainutlaatuisuutta käyttää paljastumisriskin arviointiin. Mikäli on julkaistu koko tietojoukko, on ainutlaatuisuutta helppo arvioida: tietueen ainutlaatuisuus on sitä suurempi, mitä harvemmin sen valittujen attribuuttien arvojen yhdistelmä esiintyy tietojoukossa.

Jos taas koko tietojoukosta on julkaistu vain otos, voidaan ainutlaatuisuus laskea todennäköisyytenä, että otoksessa ainutlaatuinen tietue (engl. sample unique) on myös tietojoukossa ainutlaatuinen (engl. population unique). Merkitään seuraavassa otostasolla ainutlaatuista tietuetta SU ja tietojoukon tasolla ainutlaatuista tietuetta PU.

Elamir (2004, s. 275) määrittelee ainutlaatuisuuden otoksessa seuraavasti:

$$\Pr(PU|SU) = \frac{\Pr\,(PU, SU)}{\Pr\,(SU)} = \frac{\sum_j I(F_j = 1, f_j = 1)}{\sum_j I(f_j = 1)},$$

jossa $\mathrm{j} = 1, \ldots, \mathrm{J}$ kuvastaa otoksessa esiintyviä attribuuttien arvoja, F_j on niiden tietueiden määrä alkuperäisessä tietojoukossa, joissa arvo j esiintyy, f_j on vastaava frekvenssi otoksessa ja I merkitsee otoksen kardinaliteettia. Koska oletetaan, että F ei ole

hyökkääjän tiedossa, frekvenssit F_1 ... F_j mallinnetaan Poissonin jakauman avulla. Mikäli otanta on ainoa käytetty anonymisointimenetelmä, niin käytännössä paljastumisriski $Pr(PU|SU)$ on matala yleensä vain, jos otos on merkittävästi tietojoukkoa pienempi.

6.3. Tietueiden yhdistämismallit

Paljastumisriskiä voidaan mitata niiden tietueiden suhteellisella määrällä, jotka vastustaja pystyy tunnistamaan. Tätä määrää voidaan arvioida tietueiden yhdistämismallien (engl. record linkage models) avulla. Tietueiden yhdistämismalleissa lähdetään alkuoletuksesta, jossa vastustajalla on jotain aiempaa tietoa, sekä käytössään mahdollisimman tehokkaat välineet, joilla pyrkii hallussaan olevan tietojoukon sekä anonymisoidun tietojoukon tietueita yhdistämällä paljastamaan tietoja (Domingo-Ferrer & Torra, 2003). Tietueiden yhdistämismallit voidaan jakaa kahteen pääluokkaan, välimatkaan perustuvaan malliin sekä todennäköisyyksiin perustuvaan malliin.

Välimatkaan perustuva tietueiden yhdistäminen koostuu kahden eri tiedostoissa sijaitsevan tietueen välisen välimatkan laskemiseen. Anonymisoidun tietojoukon sekä vastustajan tiedossa olevan tietojoukon tietueista etsitään ne parit, joiden välimatka on lyhin. Tiettyjä tietueita tai muuttujia voidaan painottaa. Välimatkan laskemiseen perustuvasta menetelmästä on useita versioita, joissa välimatkan laskemiseen käytetty funktio vaihtelee. Mallin toimivuus riippuukin suuresti käytetystä välimatkafunktiosta. (Torra, 2017, s. 163)

Todennäköisyyteen perustuvassa mallissa tietueiden väliltä etsitään yhteyksiä valitun todennäköisyysmallin mukaan. Todennäköisyyteen perustuvat menetelmät ovat välimatkaan perustuvia menetelmiä monimutkaisempia toteuttaa. Todennäköisyyteen perustuvassa mallissa tietueita tai muuttujia ei voida painottaa kuten välimatkaan perustuvassa mallissa. (Torra, 2017, s. 142)

Tietueiden yhdistämismallien vahvuus on, että niillä voidaan arvioida hyvin monenlaisten tietojoukkojen, jopa synteettisten tietojoukkojen, tunnistamisriskiä (Winkler, 2004). Yhdistämismallien haasteena on arvioida, mitä tietoa vastustajalla on saatavilla. Kokeellisissa asetelmissa vastustajan tiedoiksi oletetaan usein jokin alkuperäisen tietojoukon osajoukko tai jopa koko alkuperäinen tietojoukko. Yhdistämismalleja voidaan käyttää myös tunnistamisriskin ylärajan arvioimiseen olettamalla tilanne, jossa vastus-

tajalla on käytössään suurin mahdollinen määrä ulkopuolista tietoa sekä paras mahdollinen reidentifikaatioalgoritmi parhaalla mahdollisella parametrisoinnilla. (Torra, 2017, s. 125–126)

7. POHDINTA

Anonymisointi perustuu paljastumisriskin minimoimiseen mahdollisimman vähällä informaatiokadolla. Anonymisoinnin lähtökohtana on yleensä jokin yksityisyysmalli, joka määrittelee hyväksyttävän paljastumisriskin tason. Yksityisyysmallit voidaan jakaa kahteen pääluokkaan: *k*-anonymiteettiin sekä differentiaaliseen yksityisyyteen perustuviin malleihin. *K*-anonymiteettia ja differentiaalista yksityisyyttä on vaikeaa suoraan vertailla, sillä niitä käytetään usein eri tarkoituksiin. *K*-anonymiteettiperiaatteeseen pohjautuvia menetelmiä käytetään yleensä, jos halutaan julkaista koko tietojoukko tai otos siitä, kun taas differentiaalisen yksityisyyden menetelmät luovat tietojoukosta tarpeen mukaan näkymiä, eikä koko tietojoukkoa ole tarkoituskaan julkaista.

Narayanan & Shmatikov (2010) kritisoivat artikkelissaan *k*-anonymiteettiin perustuvien menetelmien niin sanottua ”julkaise ja unohda”-toimintatapaa. Heidän mukaansa julkaistun aineiston anonymiteettia tulisi uudelleenarvioida säännöllisin väliajoin, sillä uutta tietoa tulee jatkuvasti ja yhdisteltävyysriski kasvaa. Heidän mukaansa differentiaalinen yksityisyys on tämän vuoksi *k*-anonymiteettia kestävämpi ratkaisu. Differentiaalinen yksityisyys on kuitenkin tietosuoja-asetuksen kannalta ongelmallinen silloin, jos kaikkiin tietueisiin ei lisätä kohinaa. EU:n tietosuojaryhmä (2014) pohtikin lausunnossaan (WP216), että koska alkuperäisiä tietoja ei muuteta, on rekisterinpitäjällä mahdollisuus tunnistaa yksittäinen henkilö differentiaalisesti yksityisistä näkymistä. Tämän vuoksi näkymät eivät tietosuojalain näkökulmasta ole anonyymeja, vaan niitä on pidettävä henkilötietoina.

Sopivan yksityisyysmallin valintaan vaikuttaa myös se, onko tärkeämpää minimoida paljastumisriski vai informaatiokato, sekä se, kuinka paljon käsittelijä haluaa itse kontrolloida paljastumisriskin tai informaatiokadon määrää. Anonymisointimenetelmän valinta taas riippuu paitsi valitusta yksityisyysmallista, myös alkuperäisen tietojoukon muuttujien arvojen tietotyypeistä. Taulukossa 4 on esitetty tässä tutkielmassa esitellyt anonymisointimenetelmät luokiteltuna menetelmän tyypin sekä soveltuvan datatyypin mukaan.

TAULUKKO 4: Anonymisointimenetelmiä menetelmätyypin ja soveltuvan datatyypin mukaan

	Numeerinen muuttuja	Luokiteltu muuttuja
Rajoittavat menetelmät (*k*-anonymiteetti)	Yleistäminen Lokaali rajoittaminen	Otanta Yleistäminen Lokaali rajoittaminen
Satunnaistavat menetelmät (differentiaalinen yksityisyys)	Kohinan lisääminen Mikroaggregaatio	Mikroaggregaatio Arvojen vaihtaminen PRAM

Kun anonymisointi on toteutettu vastaamaan tietyn mallin vaatimuksia, voidaan sen laatua vielä arvioida paljastumisriskin avulla. Ainutlaatuisuus on yksinkertainen paljastumisriskin mittari, joka soveltuu käytettäväksi silloin, kun tietojoukon muuttujat ovat luokiteltuja, eikä niihin ole lisätty kohinaa. Tietueiden yhdistämiseen perustuvat mittarit soveltuvat sen sijaan lähes kaiken tyyppiselle datalle, joskin ne ovat huomattavasti monimutkaisempia toteuttaa.

Tunnisteellinen tieto, jota käsiteltiin toisessa luvussa, on anonymisoinnin tärkeimpiä käsitteitä. Lisääntyvä julkinen data sekä laskentamenetelmien kehittyminen väistämättä laajentavat tunnisteellisen tiedon määritelmää: se, mikä eilen ei ollut tunnisteellista, voi tänään lisääntyneen tiedon valossa sitä olla. On tärkeää, että tiedosta vastuussa olevalla julkaisijalla on työkalut, joilla paljastumisriskiä voidaan arvioida myös jälkikäteen.

8. YHTEENVETO

Tämän tutkielman tavoitteena on selvittää, mitkä ovat mikrodatan anonymisoinnissa yleisimmin käytetyt menetelmät, mihin tarkoituksiin ne soveltuvat ja millaisia rajoitteita niillä on. Lisäksi halusin selvittää, miten paljastumisriskiä voidaan arvioida. Anonymisointimenetelmiä on kehitetty useita, joiden soveltuvuus tietylle aineistolle riippuu muun muassa aineiston rakenteesta sekä muuttujien arvojen datatyypeistä.

Paljastumisriskin sekä informaatiokadon optimointi on haastava ongelma, jota tutkitaan paljon. Nykyään lisähaastetta tuo ulkopuolisen tiedon suuri määrä, joka on myös huomioitava paljastumisriskiä arvioitaessa. Se, painotetaanko anonymisoitaessa paljastumisriskiä vai informaatiokatoa, riippuu muun muassa loppukäyttäjän tarpeista, julkaisun laajuudesta sekä mahdollisen tunnistuksen aiheuttaman vahingon vakavuudesta.

Tilastollinen tietosuoja ja anonymisointi kasvattanevat merkitystään tiukentuvien tietosuojasäännösten sekä datan helpon saatavuuden johdosta. Kerättyjen tietojen hyödyntäminen laillisesti vaatii, että anonymisointi on suoritettu riittävän laadukkaasti ja kestävästi. Tietosuojan säilymiseksi onkin tärkeää, että tietojen käsittelijöillä on asiantuntemusta eri anonymisointimenetelmien soveltuvuudesta eri datatyypeille sekä kykyä arvioida anonymisoidun aineiston paljastumisriskiä.

9. LÄHTEET

Dandekar, R. Domingo-Ferrer, J. Sebé, F. (2002). LHS-Based Hybrid Microdata vs Rank Swapping and Microaggregation for Numeric Microdata Protection. Inference Control in Statistical Databases. Lecture Notes in Computer Science, vol. 2316, s. 153-162.

Domingo-Ferrer, J. Torra, V. (2001). Disclosure control methods and information loss for microdata. Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, s. 91-110.

Domingo-Ferrer, J. Torra, V. (2003). Disclosure Risk Assessment in Statistical Microdata Protection via Advanced Record Linkage. Statistics and Computing 13, s. 343-353.

Domingo-Ferrer, J. Torra, V. (2004). Disclosure risk assessment in statistical data protection. Journal of Computational and Applied Mathematics, vol. 164-165, s. 285-293.

Dwork, C. (2006). Differential Privacy. Automata, Languages and Programming. Lecture Notes in Computer Science, vol. 4052, s. 1–12.

Elamir, E. A. H. (2004). Analysis of Re-identification Risk Based on Log-Linear Models. Privacy in Statistical Databases. Lecture Notes in Computer Science, vol. 3050, s.273-281.

Euroopan parlamentti ja neuvosto. Asetus (EU) 2016/679 luonnollisten henkilöiden suojelusta henkilötietojen käsittelyssä sekä näiden tietojen vapaasta liikkuvuudesta ja direktiivin 95/46/EY kumoamisesta (yleinen tietosuoja-asetus).

Euroopan Unionin Tietosuojaryhmän Lausunto 5/2014 anonymisointitekniikoista. Annettu 10. huhtikuuta 2014. (WP 216).

Fienberg, S.E. Makov, U. (1998). Confidentiality, Uniqueness and Disclosure Limitation for Categorical Data. Journal of Official Statistics 14(4), s. 385-397.

Gehrke, J. Hay, M. Lui, E. Pass, R. (2012). Crowd-Blending Privacy. Advances in Cryptology (CRYPTO 2012). Lecture Notes in Computer Science vol. 7417, s. 479-496.

He, X. Machanavajjhala, A. Ding, B. (2014). Blowfish privacy: tuning privacy-utility trade-offs using policies. Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, s. 1447-1458.

Hundepool, A. Domingo-Ferrer, J. Franconi, L. Giessing, S. Nordhult, E. S. Spicer, K. de Wolf, P. (2012). Statistical Disclosure Control. Wiley.

Li, N. Li, T. Ventakatasubramanian, S. (2007). t-Closeness: Privacy Beyond k-anonymity and l-diversity. IEEE 23rd International Conference on Data Engineering, s. 106-115.

Machanavajjhala, A. Kifer, D. Gehrke, J. (2007). L-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data 1(1).

Narayanan, A. Shmatikov, V. (2008). Robust De-anonymization of Large Datasets. 2008 IEEE Symposium on Security and Privacy, s. 111-125.

Narayanan, A. Shmatikov, V. (2010). Myths and Fallacies of “Personally Identifiable Information”. Communications of the ACM, 53(6), s. 24-26.

Rubin, D.B. (1993). Discussion Statistical Disclosure Limitation. Journal of Official Statistics, 9 (2), s. 461-468.

Samarati, P. Sweeney, L. (1998). Protecting Privacy when Disclosing Information: k-anonymity and its Enforcement through Generalization and Suppression. Technical Report SRI-CSL-98-04, Computer Science Laboratory, SRI International.

Sarathy, R. Muralidhar, K. (2011). Evaluating Laplace Noise Addition to Satisfy Differential Privacy for Numeric Data. Transactions on Data Privacy, vol. 4, s. 1-17.

Sweeney, L. (2002). Achieving k-anonymity Privacy Protection Using Generalization and Suppression. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5), s. 571-588.

Torra, V. (2017). Data Privacy: Foundations, New Developments and the Big Data Challenge. Springer International.

Winkler, W.E. (2004). Re-identification Methods for Masked Microdata. Privacy in Statistical Databases. Lecture Notes in Computer Science, vol. 3050, s. 216-230.