Antti Repo

TIETOJOUKKOJEN ANONYMISOINTI JA JÄLLEENTUNNISTAMINEN

JYVÄSKYLÄN YLIOPISTO
TIETOJENKÄSITTELYTIETEIDEN LAITOS
2016

TIIVISTELMÄ

Repo, Antti
Tietojoukkojen anonymisointi ja jälleentunnistaminen
Jyväskylä: Jyväskylän yliopisto, 2016, 38s.
Tietojärjestelmätiede, kandidaatintutkielma
Ohjaaja: Halttunen, Veikko

Nykyään ihmisistä kerätään ja tallentuu massiivisia määriä henkilökohtaista dataa, mutta kyseisen datan yksityisyydensuojasta ja turvallisuudesta ei aina voida olla täysin varmoja. Kun ihmisten henkilökohtaisia tietoja, kuten sairaushistoriaa tai hoitotietoja, julkaistaan esimerkiksi tutkimuskäyttöön, tulee tiedot anonymisoida riittävällä tavalla eli käsitellä siten, ettei yksittäisiä henkilöitä kyetä tunnistamaan tiedoista. Vaikka anonymisointitekniikoita on useita ja ne voivat olla tehokkaita, eivät ne ole täydellisiä: joskus anonymisointi voi pettää ja ihmisten mahdollisesti arkaluontoisiakin tietoja voi tulla julki ja päätyä vääriin käsiin. Anonymisoinnin pettäminen voi johtua joko ihmisten tietoja sisältävän tietokannan hallinnoijan virheestä, tai vastustajan – tietojen paljastamista haluavan henkilön – aktiivisista toimista. Tämän kirjallisuuskatsauksen tarkastelun kohteena on tietojoukkojen anonymisointi sekä deanonymisointi eli jälleentunnistaminen. Tutkimuskysymyksenä on selvittää, ovatko nykyiset tietojoukkojen anonymisointitoimenpiteet riittäviä ihmisten yksityisyyden takaamiseksi tietojoukoissa, eli voidaanko anonymisointiin täysin luottaa nykyajan digitaalisessa maailmassa. Katsauksessa tutustutaan erilaisiin tietojoukkojen anonymisointitekniikoihin ja -menetelmiin sekä niiden epäonnistumisiin eli tapauksiin, joissa anonymisointi on pettänyt ja deanonymisointi on onnistunut, jolloin yksityishenkilöiden piilotettuja identiteettejä on kyetty paljastamaan. Yhteenvetona voidaan todeta, että täydellisesti anonyymiä ja samanaikaisesti hyödyllistä tietojoukkoa ei ole mahdollista luoda, sillä muun muassa täydentävää, ulkopuolista informaatiota hyväksikäyttäen yksityishenkilöistä on mahdollista paljastaa tunnistavaa henkilökohtaista informaatiota.

Asiasanat: anonymisointi, jälleentunnistaminen, deanonymisointi, tietojoukko, tietokanta, yksityisyys, tietoturva

ABSTRACT

Repo, Antti
Anonymization and reidentification of datasets
Jyväskylä: University of Jyväskylä, 2016, 38p.
Information Systems Science, Bachelor's thesis
Supervisor: Halttunen, Veikko

Nowadays, massive amounts of personal data are being collected and stored but the security of the data cannot always be guaranteed. When people's personal information such as the history of illnesses or treatments is published for example for research purposes, the data needs to be anonymized in a sufficient way so that single individuals cannot be recognized from the dataset. Although different anonymization methods are numerous and they can be efficient, sometimes the anonymization can fail and potentially sensitive information can end up in the wrong hands. The failure of anonymization can be caused by an error made by the data administrator or due to the actions of an adversary – a person who wishes to uncover anonymized information. The object of this literature review is to examine the anonymization and deanonymization of datasets. The research question seeks to find out whether current anonymization procedures are sufficient in guaranteeing the privacy of individuals and if anonymization can be fully trusted in today's digital world. I review various anonymization techniques and methods, their strengths, weaknesses, and failures, i.e. cases where the anonymization has failed and reidentification has succeeded: hidden identities of individuals have been revealed. As a conclusion it can be stated that a dataset that is simultaneously perfectly anonymous and useful cannot currently be created, because of the fact that by, for example, combining outside information with the data it is still possible to reveal personal information about individuals.

Keywords: anonymization, reidentification, deanonymization, dataset, database, privacy, data security

KUVIOT

TAULUKOT

SISÄLLYS

1 JOHDANTO

Yksityisyydensuoja on nykyään ajoittaisten digitaalisten tietovuotojen ja -murtojen vuoksi usein esillä. Ihmisistä kerätään ja tallennetaan suuria määriä dataa, monesti jopa tiedon kohteen eli yksityishenkilön sitä tiedostamatta: esimerkiksi investointialan asiantuntijayritys Skandian (2011) suorittaman kyselyn mukaan vain seitsemän prosenttia isobritannialaisista aikuisista lukee käyttöehtosopimukset ottaessaan käyttöön uutta tuotetta tai palvelua.

Elämme ”big datan” aikakautta: tietokoneiden prosessointinopeus ja tallennustila ovat kasvaneet huomattavasti ja kehitys tiedonlouhinnassa on ollut merkittävää. Datasta on tullut raaka-ainetta, jolla on suuri taloudellinen ja sosiaalinen arvo (Tene & Polonetsky, 2012). Dataa on nykyään enemmän kuin koskaan koko ”datan historian” aikana: taltioidun historian alusta vuoteen 2003 dataa oli tuotettu 5 miljardia gigatavua, vuonna 2011 dataa tuotettiin 5 miljardia gigatavua joka toinen päivä, vuonna 2013 joka kymmenes minuutti ja vuonna 2015 sama 5 miljardia gigatavua tuotettiin joka kymmenes sekunti (Smolan & Erwitt, 2012). Jo 2000-luvun alusta asti internetin tarjoamia mahdollisuuksia on hyödynnetty datan keräämiseen ja analysointiin. HTTP-pohjaiset Web 1.0 –järjestelmät, kuten hakukoneet Yahoo ja Google sekä verkkokauppayritykset kuten Amazon ja eBay, antoivat organisaatioille mahdollisuuden tuoda liiketoimintansa verkkoon ja päästä hyvin lähelle asiakkaitaan (Chen, Chiang & Storey, 2012). Evästeiden kautta saumattomasti kerättävät yksityiskohtaiset ja IP-spesifit käyttäjähaut sekä vuorovaikutuslokit ja palvelinlokit ovat nykyajan kultakaivoksia käyttäjien halujen tunnistamiseen ja uusien liiketoimintamahdollisuuksien löytämiseen (Chen et al., 2012).

Meitä koskevaa dataa on kaikkialla. Tietokoneet ja älypuhelimet ohjelmistoineen ja sensoreineen tallentavat muun muassa sijainteihin, puheluihin, viesteihin, internetin selaushistoriaan ja moniin muihin asioihin liittyvää dataa, jonka luovuttamiseen kulloisenkin ohjelmiston käyttäjät suostuvat mahdolliset käyttöehtosopimukset hyväksyessään, tietoisesti tai tietämättään. Lääkärikäynnin yhteydessä potilaasta tallennetaan tietokantoihin esimerkiksi tietoja sairauksista, hoidoista sekä elämäntavoista, ja lääketieteen tutkijat voivat jakaa tätä dataa keskenään (Ohm, 2010). Yksityishenkilön ei aina

ole helppo tai edes mahdollista tietää kenellä häntä koskevaa dataa on, mitä tietoja tämä data sisältää ja mihin sitä käytetään. Dataa hallinnoivat tahot eli käytännössä tietokantojen ylläpitäjät voivat myös julkaista dataa erilaisiin käyttötarkoituksiin, tai data voi epätarkoituksenmukaisesti vuotaa tai muuten päätyä vääriin käsiin (Ohm, 2010). Ohm (2010) puhuu uhkaavasti ”turmion tietokannasta” (eng. *database of doom*), joka ei kuitenkaan ole yksi tietty tietokanta jossain tietyssä palvelinsalissa, vaan potentiaalinen uhkakuva. ”Turmion tietokanta” käsittää kaiken yksittäisestä ihmisestä kerätyn datan, missä datan sisältävät tietokannat sitten sijaitsevatkaan. Sillä tarkoitetaan kaikkia viestejä, kuvia, julkaisuja, vierailuja, tekstejä, sijaintitietoja, henkilötietoja ynnä muita tietoja, joita yksityishenkilö on tallentanut, joita yksityishenkilöstä on tallennettu ja tallentuu erinäisiin tietokantoihin henkilön liikkuessa verkossa ja fyysisessä maailmassa. ”Turmion tietokanta” on konsepti, jonka rakentumiseen jokainen verkon käyttäjä on itse osallistunut, ja jonka sisältö vääriin käsiin joutuessaan voi kenties aiheuttaa tiedon kohteelle suurta harmia ja vaivaa tai jopa jotain vielä vakavampaa.

Suomen henkilötietolaissa tietojen suojaamisesta todetaan seuraavasti: ”Rekisterinpitäjän on toteutettava tarpeelliset tekniset ja organisatoriset toimenpiteet henkilötietojen suojaamiseksi asiattomalta pääsyltä tietoihin ja vahingossa tai laittomasti tapahtuvalta tietojen hävittämiseltä, muuttamiselta, luovuttamiselta, siirtämiseltä taikka muulta laittomalta käsittelyltä. Toimenpiteiden toteuttamisessa on otettava huomioon käytettävissä olevat tekniset mahdollisuudet, toimenpiteiden aiheuttamat kustannukset, käsiteltävien tietojen laatu, määrä ja ikä sekä käsittelyn merkitys yksityisyyden suojan kannalta” (Oikeusministeriö, 1999). Laissa ei suoranaisesti oteta kantaa siihen, mitä ”tarpeelliset tekniset ja organisatoriset toimenpiteet henkilötietojen suojaamiseksi (…)” ovat. On kuitenkin selvää, että tietokantojen ylläpitäjien tai hallinnoijien vastuulla on ylläpitämiensä tietojoukkojen yksityisyyden ja anonymiteetin turvaaminen. Aina tämä ei kuitenkaan ole ollut mahdollista reaalimaailman ja digitaalisen maailman monista muuttujista johtuen.

Tässä kandidaatintutkielmassa tarkastelen kirjallisuuskatsauksena tietojoukkojen eli sähköisiin tietokantoihin tallennettavien tietojen anonymisointia ja deanonymisointia. Deanonymisointi, tai jälleentunnistaminen, tarkoittaa anonymisoinnin purkamista tai ehkä pikemminkin anonymisoinnin epäonnistumista. Deanonymisointitapaukset ovat tapauksia, joissa jotkin tahot ovat saaneet selvitettyä ihmisten henkilö- tai muita tunnistetietoja anonyymeiksi tarkoitetuista tietojoukoista sekä omalla aktiivisella toiminnallaan että tietojoukkojen ylläpitäjien laiminlyönneistä johtuen. Tarkoituksena on selvittää, millaisia anonymisointitekniikoita on yleisesti käytössä, ovatko ne riittäviä yksityishenkilöiden anonymiteetin takaamiseksi ja onko anonymiteetti nykyisessä digitalisoituneessa maailmassa alun perinkään mahdollinen käsite. Tutkimuskysymyksenä on selvittää, ovatko nykyiset anonymisointitoimenpiteet riittäviä ihmisten yksityisyyden

takaamiseksi tietojoukoissa, eli voidaanko anonymisointiin täysin luottaa nykyajan digitaalisessa maailmassa.

Anonymisointi tarkoittaa toimenpiteitä, joilla yksityishenkilöitä koskeva henkilökohtainen data suojataan siten, ettei yksityishenkilöä voida siitä tunnistaa, esimerkiksi tunnistetietoja (kuten nimi ja sosiaaliturvatunnus) poistamalla tai muuttamalla (Ohm, 2010). Pfitzmannin ja Koehntoppin (2001) määritelmän mukaan anonyymiys tarkoittaa ”tunnistamattomissa olemisen tilaa kohteiden joukossa eli anonyymijoukossa”. Esimerkiksi lääketieteellinen data anonymisoidaan silloin, kun sitä julkaistaan tutkimuskäyttöä varten, ja yritys voi anonymisoida asiakastietojaan antaessaan niitä esimerkiksi sisäisesti markkinointiosastonsa käyttöön. Ohmin (2010) mukaan data voi kuitenkin olla joko täysin anonyymia, tai sitten se voi olla hyödyllistä. Tällä hän tarkoittaa sitä, että jos datan halutaan olevan anonyymiä, siitä on poistettava yksittäisiä henkilöitä koskevia tietoja, mikä kuitenkin samalla tekee datasta vähemmän ja vähemmän hyödyllistä esimerkiksi tutkijoille. Tämä on toki hieman kärjistetysti ilmaistu, mutta voi tutkijoiden näkökulmasta olla hyvinkin totta.

Tutkielman aluksi eli toisessa pääluvussa käsittelen anonymisointia, siihen liittyviä olennaisia käsitteitä kuten *vastustajaa* ja *datan julkaisua* sekä anonymisoinnin syitä ja merkitsevyyttä. Kolmannessa luvussa luon katsauksen yleisesti käytettyihin anonymisointitekniikoihin kuten satunnaistamiseen ja yleistämiseen sekä niiden alaisuuteen kuuluviin tekniikoihin. Luvun lopuksi puhun pseudonymisoinnista, jota käsittelen erillään, sillä se ei ole anonymisoinnin keino (Article 29 Data Protection Working Party, 2014; El Emam & Álvarez, 2014), vaan oma tekniikkansa yksityisyyden suojaamiseen. Tutkielman neljäs luku käsittelee katsauksen toista pääaihetta, tietojoukkojen deanonymisointia eli jälleentunnistamista. Jälleentunnistaminen tarkoittaa toimenpiteitä, joilla anonyymiksi tarkoitetusta ja mielletystä tietojoukosta kyetään tunnistamaan yksityishenkilöitä ja heidän tietojaan. Luvussa käyn läpi esiteltyjen anonymisointitekniikoiden heikkouksia sekä tunnettuja tietovuotoja ja -murtoja, joissa anonymisointi ei ole ollut riittävää ja yksityisiksi tarkoitetut tiedot ovat paljastuneet, jolloin yksityishenkilöitä on kyetty tunnistamaan anonyymiksi mielletystä datasta. Johtopäätökset-luvun (Luku 5) lopuksi käsittelen sekä olemassa olevia että tulevaisuudessa kenties mahdollisia keinoja taata parempi anonymiteetti sekä yksityishenkilön että tietokantojen hallinnoijan tai ylläpitäjän näkökulmasta. Tutkielman viimeisessä luvussa tehdään yhteenveto tutkielmasta.

2 ANONYMISOINTI

Tässä luvussa käsittelen anonymisoinnin piiriin kuuluvia käsitteitä sekä anonymisoinnin konseptia. Luvussa esitän myös perusteet anonymisoinnin tarpeellisuudelle ja anonymisointitoimenpiteisiin johtavia syitä. Henkilökohtaisesti tunnistavaa informaatiota käsittelen omana alalukunaan, sillä se on käsitteenä anonymisoinnin puitteissa yksi tärkeimmistä.

2.1 Käsitteet

Tietojen *anonymisointi* tarkoittaa esimerkiksi tietokannassa olevien henkilöiden tunnistetietojen (kuten nimi ja sosiaaliturvatunnus) poistamista tai muuntamista niin, ettei tunnistaminen enää ole mahdollista eikä henkilöitä voida identifioida (Ohm, 2010). Tietojoukon voidaan katsoa olevan anonyymi, kun yksilön erottaminen joukosta ei enää ole mahdollista, yksilöön liittyvien tietueiden yhdistäminen ei ole mahdollista eikä yksilöstä voida tehdä päätelmiä. Osa näistä riskeistä voidaan poistaa tietyllä anonymisointitekniikalla ainakin osittain, joten yksittäisen tekniikan soveltaminen tietyssä tilanteessa edellyttää huolellista suunnittelua, kuten myös tekniikoiden yhdistelmän soveltaminen niin, että lopputuloksena saadaan vahvempi anonymiteetti. (Article 29 Data Protection Working Party, 2014.)

Anonymisoinnista on muitakin määritelmiä. Esimerkiksi kansainvälisessä standardissa ISO 29100 anonymisointi on määritelty 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).

Yksi tärkeä anonymisoinnin piiriin kuuluvista käsitteistä on *henkilökohtaisesti tunnistava informaatio* (eng. *Personally Identifiable Information*, *PII*). Tällainen informaatio voi käsittää nimiä, henkilötunnuksia ja osoitteita tai mitä tahansa dataa, jota voidaan käyttää yksityishenkilön identiteetin

tunnistamiseen tai jäljittämiseen joko sellaisenaan tai yhdistelemällä muuhun informaatioon, joka voidaan liittää tiettyyn yksityishenkilöön (Krishnamurthy & Wills, 2009). Henkilökohtaisesti tunnistavaa informaatiota käsitellään tarkemmin luvussa 2.3.

Ohmin (2010) mukaan nykyiset anonymisointitekniikat ovat kaukana täydellisistä, ja siispä hänen mielestään sanaa ”anonymisointi” ei tulisi käyttää kuvaamaan prosessia, jossa tietojoukkojen datasta pyritään puhdistamaan tunnistetiedoista. Samaa mieltä on myös Sweeney (1997): hänen mukaansa sana *anonyymi* implikoi, että dataa ei voida manipuloida tai linkittää siten, että yksityishenkilö voitaisiin tunnistaa. Sweeney (1997) käyttääkin artikkelissaan sanaa ”deidentifioida” (eng. *deidentify*). Ohmin (2010) mukaan anonymisoinnin sijaan tulisi ennemminkin käyttää esimerkiksi ilmaisua ”yrittää saavuttaa anonymiteetti”, tai sanaa ”kuurata” (eng. *scrub*). Tässä tutkielmassa käytän kuitenkin selvyyden vuoksi sanaa ”anonymisointi”.

Vastustajalla (eng. *adversary*) tarkoitetaan tietojenkäsittelytieteissä henkilöä, joka pyrkii jälleentunnistamaan jo anonymisoitua dataa (Ohm, 2010). Tietojenkäsittelytieteilijät kuvailevat anonymisointia ja jälleentunnistamista ikään kuin vastakkainasetteluun perustuvaksi peliksi, jossa tietokannan anonymisointi voidaan katsoa pelin ensimmäiseksi siirroksi (Ohm, 2010). Tietojenkäsittelytieteilijät eivät vaikuta pyrkivän moralisoimaan vastustajan toimia, eivätkä tee oletuksia vastustajan jälleentunnistamisaikeiden hyvän- tai pahantahtoisuudesta. Vastustajan määrittävä ominaisuus vaikuttaa olevan vastustaminen – hän on motivoitunut tekemään jotain, mitä datan ylläpitäjät eivät toivo tapahtuvan (Ohm, 2010). Vastustajia ovat sellaiset henkilöt, joilla on jokin motiivi jälleentunnistamiseen. Narayananin ja Shmatikovin (2009) mukaan heitä voivat olla ahdistelijat, etsivät tai tutkijat, työtoverit, työnantajat tai naapurit. Ohmin (2010) mukaan listaan voi lisätä myös poliisin, valtion turvallisuusanalyytikot, mainostajat sekä kenet tahansa muun, joka haluaa yhdistää yksityishenkilön dataan.

Ulkopuolista informaatiota voidaan kutsua myös avustavaksi tai täydentäväksi informaatioksi (eng. *auxiliary information*) tai taustatiedoiksi (eng. *background knowledge*) (Ganta, Kasiviswanathan & Smith, 2008). Kun vastustaja löytää uniikin datajäljen, hän voi linkittää kyseisen datan ulkopuoliseen informaatioon. Eri tietokantojen tietoja yhdistelemällä vastustaja voi tunnistaa yksittäisiä henkilöitä ja saada haltuunsa hyvinkin tarkkaa tietoa heistä. Monet anonymisointitekniikat olisivat täydellisiä, jos vastustaja ei tietäisi maailman ihmisistä yhtään mitään muuta jälleentunnistettavana olevan datan lisäksi. Todellisuudessa maailma on täynnä dataa ihmisistä ja uusia tietokantoja luodaan päivittäin. (Ohm, 2010.)

Vastustajat yhdistävät anonymisoitua dataa ulkopuoliseen informaatioon paljastaakseen hämärrettyjä identiteettejä (eng. *obscured identities)* (Ohm, 2010). Tällaista ulkopuolista informaatiota on esimerkiksi muiden tietokantojen data, internetin palveluiden käyttäjien itsensä paljastama informaatio, kuten Facebook-profiilin tiedot tai muut julkisesti tai muulla tavalla vastustajan ulottuvilla olevat data- ja informaatiolähteet.

Datan julkaisulla viitataan de-anonymisoinnin kontekstissa sosiaalisissa verkostoissa prosessiin, jossa vastustaja saa haltuunsa anonymisoitua ja mahdollisesti puhdistettua (eng. *sanitized*) dataa (Ding, Zhang, Wan & Gu, 2010). Ding ym. (2010) mukaan datan julkaisut voidaan jakaa kahteen kategoriaan: eksplisiittiseen datan julkaisuun ja implisiittiseen datan julkaisuun.

Eksplisiittisellä datan julkaisulla tarkoitetaan esimerkiksi sosiaalisista medioista kerätyn datan kausittaista julkaisua kolmansille osapuolille kuten mainostajille, ohjelmistokehittäjille ja datanlouhinnan tutkijoille. Eksplisiittisesti julkaistu data on yleensä anonymisoitu ja puhdistettu yksityisyyden suojelemiseksi ennen sen julkistamista. Esimerkiksi yksilöivä informaatio kuten nimet, puhelinnumerot ja sähköpostiosoitteet, joita voitaisiin käyttää yksilön tunnistamiseen, yleensä korvataan sattumanvaraisilla (mutta yksilöllisillä) tunnisteilla. Yksityisyyden turvaamisen tekniikoita kuten k-anonymiteettiä voidaan myös käyttää hyväksi. (Ding ym., 2010.)

Implisiittisellä datan julkaisulla tarkoitetaan datan tietyn tyyppistä vuotamista tai "paljastumista". Tämä tarkoittaa, että dataa ei varsinaisesti julkaista suorasti eikä epäsuorasti, vaan vastustaja kykenee saamaan yksilöivää informaatiota haltuunsa esimerkiksi www-selaimen HTTP-otsakkeiden, URI:en (Uniform Resource Identifier, merkkijono, jolla kerrotaan tietyn tiedon paikka tai yksikäsitteinen nimi) ja evästeiden kautta. (Ding ym., 2010.)

Tietokannoissa säilytetään *tietojoukkoa*, joka koostuu erilaisista *tallenteista*, jotka liittyvät *yksilöihin* eli yksityishenkilöihin, datan kohteisiin. Jokainen tallenne liittyy yhteen datan kohteeseen ja koostuu joukosta arvoja (tai syötteistä, esimerkiksi 2013), jotka on annettu eri attribuuteille (esimerkiksi vuosi). Tietoaineisto on kokoelma tallenteita, jotka voidaan muotoilla vaihtoehtoisesti taulukoksi tai joukoksi taulukoita, tai selityksin varustetuiksi graafeiksi. Tutkielmassa esitetyt esimerkit liittyvät taulukoihin. Datan kohteisiin tai niiden ryhmiin liittyviä attribuuttien yhdistelmiä voidaan kutsua kvasitunnisteiksi tai näennäistunnisteiksi. Joissain tapauksissa tietojoukossa voi olla useita tallenteita samasta yksilöstä. (Article 29 Data Protection Working Party, 2014).

Datan kontrolloija on taho, joka säilyttää ja ylläpitää dataa tai tietoaineistoa (esimerkiksi tietokantojen ylläpitäjä), ja joka määrittelee mihin tarkoitukseen ja miten henkilökohtaista dataa käytetään tai tulee käyttää. Taho voi olla yksi tai useampi henkilö tai esimerkiksi organisaatio. (Information Commissioner's Office, 2016.)

2.2 Anonymisoinnin tarpeellisuus ja syyt

Nykyään erilaiset laitteet, sensorit ja verkot tallentavat sekä luovat suuria määriä monenlaista dataa ja datan tallentamisen hinta alkaa olla käytännössä merkityksetön. Tämä on luonut kasvavaa kiinnostusta ja kysyntää tällaisen datan uudelleenkäyttöön. "Avoin data" voi tuoda selviä hyötyjä yhteiskunnalle, yksilöille ja organisaatioille, mutta vain jos kaikkien oikeutta

henkilökohtaisen datan ja yksityiselämän suojaamiseen kunnioitetaan. (Article 29 Data Protection Working Party, 2014.)

Monet organisaatiot keräävät ja julkaisevat lisääntyvissä määrin dataa eli tietojoukkoja (monesti taulukkomuodossa), jotka sisältävät aggregoimattomia eli ryhmittelemättömiä tai koostamattomia tietoja yksittäisistä henkilöistä. Näissä tietojoukoissa olevat tiedot voivat olla sairaaloiden potilastietoja, äänestäjätietoja, väestönlaskennallisia tietoja tai esimerkiksi asiakastietoja. Tällainen data on arvokas informaation lähde niin julkisten varojen käytön suunnittelulle, lääketieteelliselle tutkimukselle kuin trendien analysoinnillekin. Jos datasta kuitenkin voitaisiin tunnistaa yksittäisiä henkilöitä, heidän henkilökohtaiset tietonsa (kuten terveydentila) paljastuisivat, mikä ei ole hyväksyttävää. (Machanavajjhala, Gehrke, Kifer & Venkitasubramaniam, 2007.) Aiottaessa julkaista ihmisten henkilökohtaisia tietoja sisältävää dataa onkin varauduttava ”pahimpaan” ja tehtävä se oletus, että vastustajalla voi olla käytössään lähes mitä tahansa lisä- ja taustatietoa (Martin, Kifer, Machanavajjhala, Gehrke & Halpern, 2007).

Kun järjestelmien tai tietokantojen ylläpitäjät haluavat julkaista dataa, he anonymisoivat sen suojellakseen datan kohteiden (yleensä yksityishenkilöiden) yksityisyyttä. Dataa jaetaan yleensä kolmelle joukolle, joista ensimmäinen on kolmannen osapuolen toimijat: esimerkiksi terveystutkijat voivat jakaa potilasdataa keskenään, verkkosivustot myyvät transaktiodataa mainostajille ja puhelinoperaattorit voivat joutua luovuttamaan puhelutietoja viranomaisille. (Ohm, 2010.)

Toinen joukko ovat tietokantojen ylläpitäjät, jotka voivat joskus julkaista anonymisoitua dataa julkiselle yleisölle. Ylläpitäjät ovat tehneet tätä viime aikoina kasvavissa määrin, esimerkiksi osallistaakseen yleisöä niin sanottuun joukkoistamiseen – toimintaan, jossa pyritään valjastamaan suuri joukko vapaaehtoisia, jotka osaavat analysoida dataa tehokkaammin ja läpikotaisemmin kuin pieni määrä palkattuja työntekijöitä. (Ohm, 2010.)

Kolmas joukko ovat tietokantojen ylläpitäjät, jotka voivat julkaista dataa muiden käyttöön organisaationsa sisällä. Eritoten suurten organisaatioiden sisällä datan kerääjät voivat haluta suojella datan kohteiden yksityisyyttä muilta ihmisiltä organisaatiossa. Esimerkiksi suuret pankit voivat haluta jakaa dataa markkinointiosastojensa kanssa, mutta vain asiakkaan yksityisyydensuojan perusteella suoritetun anonymisoinnin jälkeen. (Ohm, 2010.)

Lessig (2006) listaa neljä käyttäytymisen sääntelijää: normit ja etiikka, markkinat, arkkitehtuuri (tietokantojen tapauksessa koodi) ja laki. Ohmin (2010) mukaan nämä neljä sääntelijää ajavat tietokantojen ylläpitäjät anonymisoimaan dataansa. Anonymisoinnin normit ja etiikka vaikuttavat usein parhaita toimintatapoja mallintavien anonymisointia yksityisyyden suojaamiseksi suosittelevien dokumenttien taustalla. Ohmin (2010) mukaan data voi olla joko hyödyllistä tai täydellisen anonyymiä, muttei koskaan molempia. Tällä hän tarkoittaa sitä, että mitä enemmän yksityiskohtaista tietoa on esimerkiksi tutkijoiden käyttöön julkaistavassa potilastietodatassa (kuten potilaan ikä,

sukupuoli, osoite tai postinumero, sairaudet, oireet ja niin edelleen.), sitä hyödyllisempää se on datan tutkijoille sen monipuolisuuden vuoksi. Toisaalta on huomattava, että mitä tarkemmin henkilöitä yksilöivää tietoa julkaistussa datassa on, sitä helpompaa eri toimijoille on tunnistaa siitä yksittäisiä henkilöitä ja väärinkäyttää tietoja. (Ohm, 2010.)

Datan anonymisoinnin perimmäisenä tarkoituksena on ihmisen yksityisyydensuojan takaaminen. Se ei ole yhdentekevä toimenpide; henkilötietodatan anonymisointia edellytetään esimerkiksi dataa tutkimuskäyttöön julkaistaessa, ja tietyt anonymisointitoimenpiteet on säädetty pakollisiksi monen maan laissa. Tällainen lainsäädäntö on hyvä varotoimi. Yksittäisiin henkilöihin liittyvää dataa kerätään ja luodaan nykyään valtavia määriä, eikä oikeastaan kukaan voi tarkalleen tietää, mitä kaikkea tietoa hänestä on kerätty ja missä sitä säilytetään. Ohm (2010) puhuukin ”turmion tietokannoista”, joilla hän tarkoittaa juuri sitä kaikkea ihmisistä kerättyä arkaluontoista tietoa, joka vääriin käsiin joutuessaan voi tuottaa mittavaa harmia tiedon kohteille. Konseptina ”turmion tietokanta” on mielenkiintoinen, vaikka on huomattava, että kyse ei ole yhdestä tietokannasta missään tietyssä konesalissa, vaan useista arkaluontoista tietoa sisältävistä tietokannoista ympäri maata tai maailmaa. Yhdistelyhyökkäyksellä näiden tietojen linkittäminen toisiinsa on toki teoriassa mahdollista potentiaalisen vastustajan löytäessä päällekkäisyyksiä eri tietokantojen datassa, ja vaikka reaalimaailman todennäköisyydet tällaiselle tapahtumalle ovat matalat, on se silti mahdollista.

2.3 Henkilökohtaisesti tunnistava informaatio

Henkilökohtaisesti tunnistavan informaation (eng. *Personally Identifiable Information*, *PII*) määritelmä ei nykyään ole täysin selvä tai helposti määriteltävissä alati kasvavan informaation määrän ja laadun vuoksi. Krishnamurthy ja Wills (2009) määrittelevät henkilökohtaisesti tunnistavan informaation ”informaatioksi, jota voidaan käyttää yksityishenkilön henkilöllisyyden erottelemiseen tai jäljittämiseen joko sellaisenaan tai yhdistelemällä muuhun informaatioon, joka on yhdistettävissä tiettyyn yksityishenkilöön.” Tällaista informaatiota voivat olla esimerkiksi sosiaaliturvatunnus, etu- ja sukunimi, syntymäaika ja –paikka, vanhempien nimet, osoite, puhelinnumero, valokuva (esimerkiksi sosiaalisessa mediassa) ja niin edelleen (Krishnamurthy & Wills, 2009). Narayanan ja Shmatikov (2010) toteavat, että vaikka jotkin attribuutit voivat yksinään olla henkilökohtaisesti tunnistavia, voi kuitenkin mikä tahansa attribuutti yhdessä muiden attribuuttien kanssa olla tunnistava. Näin ollen he laajentaisivatkin henkilökohtaisesti tunnistavan informaation määritelmää koskemaan käytännössä mitä tahansa informaatiota, jota voidaan käyttää erottamaan yksi henkilö toisesta.

Nykyään käytännössä lähes kaikkea verkossa ihmisistä kerättyä informaatiota voidaan pitää henkilökohtaisesti tunnistavana: verkon

selaushistoria, hakuhistoria, kulutustottumukset, sähköpostiosoitteet ja niin edelleen. Siitä, ovatko Internet Protocol (IP) –osoitteet (IP-verkkoihin kytkettyjen verkkosovittimien yksilöimiseen käytettävät numerosarjat) henkilökohtaisesti tunnistavaa informaatiota, on väitelty pitkään. Konsensus Yhdysvalloissa vaikuttaa kuitenkin olevan, että IP-osoitteet eivät sellaisenaan ole henkilökohtaisesti tunnistavaa informaatiota, mutta sitä voidaan pitää sellaisena laajennettuna tai yhdistettynä muuhun informaatioon. Euroopassa taas IP-osoitteet katsotaan henkilötiedoiksi (Article 29 Data Protection Working Party, 2007), kuten esimerkiksi Suomessa eräässä vuoden 2006 Tietosuojalautakunnan (2006) päätöksessä. Ohm (2010) esittääkin, että kaikki informaatio voi olla henkilökohtaisesti tunnistavaa sellaiselle toimijalle, jolla on pääsy sopivaan ulkopuoliseen informaatioon.

2.4 Yhteenveto

Yhteenvetona luvusta voidaan todeta datan anonymisoinnin perimmäisenä tarkoituksena olevan ihmisten yksityisyydensuojan säilyttäminen tai ainakin siihen pyrkiminen. Henkilötietodatan anonymisointia edellytetään dataa esimerkiksi tutkimuskäyttöön julkaistaessa, ja tietyt anonymisointitoimenpiteet on säädetty pakollisiksi monen maan laissa. Lakien lisäksi myös normit, etiikka, markkinat ja arkkitehtuuri (tietokantojen tapauksessa koodi) luovat omat haasteensa ja vaatimuksensa henkilötietodatan anonymisoinnille (Ohm, 2010).

Monesti dataa pyritään anonymisoimaan poistamalla siitä henkilökohtaisesti tunnistava informaatio, eli selvät tunnisteet, kuten henkilöiden nimet, syntymäpäivät ja osoitteet (Krishnamurthy & Wills, 2009). Henkilökohtaisesti tunnistavan informaation määritelmä kuitenkin elää ja oikeastaan laajenee jatkuvasti, joten rajan vetäminen siihen, mikä on henkilökohtaisesti tunnistavaa informaatiota ja mikä ei, on hankalaa. On myös huomattava, että tällaisen informaation poistaminen henkilötietodatasta ei kuitenkaan vielä tee datasta anonyymiä. Anonymisointiin onkin olemassa useita tekniikoita, joilla tietojoukoissa esiintyviä identiteettejä pyritään hämärtämään. Datan anonymisoinnin perimmäisenä tarkoituksena on pyrkimys taata yksityishenkilöille yksityisyydensuojaa, sillä nykyään ihmisistä kerätään ja luodaan valtavia määriä dataa, eikä oikeastaan kukaan voi tarkalleen tietää, mitä kaikkea tietoa hänestä on kerätty, missä sitä säilytetään ja mitä sillä tai sille voidaan tehdä. Ohm (2010) puhuukin ”turmion tietokannoista”, joilla hän tarkoittaa juuri sitä kaikkea ihmisistä kerättyä arkaluontoista tietoa, joka vääriin käsiin joutuessaan voi tuottaa suurta harmia tiedon kohteille eli yksityishenkilöille. Anonymisoinnilla on siksi tällaisten ”turmion tietokantojen” riskipotentiaalin pienentämisessä suuri rooli.

3 ANONYMISOINTITEKNIIKOITA

Tämä luku käsittelee muutamia laajemmalti käytössä olevia tietojoukkojen anonymisointitekniikoita sekä pseudonymisointia, jota ei lasketa anonymisoinnin keinoksi (El Emam & Álvarez, 2014). Anonymisointi voi olla hyvä strategia tietojoukkojen hyötyjen säilyttämiseen ja riskien minimoimiseen.

On kuitenkin selvää tapaustutkimusten ja tutkimusjulkaisujen kautta, että todella anonyymin tietojoukon luominen monipuolista henkilökohtaista dataa sisältävästä tietojoukosta samalla mahdollisimman paljon hyödyllistä informaatiota säilyttäen ei ole helppo tehtävä. Hyvinkin anonymisoitu tietoaineisto voidaan esimerkiksi yhdistää johonkin toiseen tietoaineistoon, ja tätä kautta tietojoukossa olevia yksityishenkilöitä on mahdollista tunnistaa ja yksilöidä. (Ohm, 2010; Article 29 Data Protection Working Party, 2014.)

3.1 Yleisesti käytetyt tekniikat

Anonymisointi on jatkuvasti tutkimuksen kohteena, mutta mikään datan anonymisointiratkaisu ei tällä hetkellä ole täydellinen. Tekniikat eivät siis välttämättä toimi yksin, vaan vaativat muita tekniikoita rinnalleen ollakseen tehokas kokonaisuus. Siksi onkin käytettävä olemassa olevien ratkaisujen yhdistelmiä. Anonymisointiin on useita lähestymistapoja, ja EU:n riippumaton neuvoa-antava tietosuojatyöryhmä on tunnistanut kaksi lähestymistapaa: ensimmäinen perustuu *satunnaistamiseen* (eng. *randomization*) ja toinen *yleistämiseen* (eng. *generalization*). Nämä lähestymistavat sisältävät erilaisia anonymisointitekniikoita, kuten *kohinan lisääminen*, *permutaatio*, *differentiaalinen yksityisyys*, *koostaminen* tai *k-anonymiteetti* (Article 29 Data Protection Working Party, 2014). Ohm (2010) käsittelee artikkelissaan samoja tekniikoita, mutta hieman eroavilla nimityksillä sekä luokituksella. Käsittelen näitä tekniikoita, sillä ne ovat löydetyn kirjallisuuden perusteella nykyisin käytetyimmät anonymisointitekniikat.

Kohinan lisääminen koostuu tietojoukon attribuuttien muokkaamisesta epätarkemmiksi, samalla kuitenkin kokonaisjakauman säilyttäen. Kohinan lisääminen toimii lisäämällä tai kertomalla sattumanvarainen numero luottamuksellisiin kvantitatiivisiin attribuutteihin (Mivule, 2013). Jos tietojoukossa henkilön pituus on alun perin ilmoitettu sentin tarkkuudella (esimerkiksi 177), kohinan lisäämisen avulla anonymisoidussa tietoaineistossa pituus voidaan esimerkiksi ilmoittaa vain kymmenen senttimetrin tarkkuudella (esimerkiksi 180). Koska tietojoukon kokonaisjakauma samana säilyy kohinan lisäämisen jälkeen, voi tietojoukosta edelleen olla mahdollista tunnistaa yksittäisiä henkilöitä. Eräs tunnettu tällainen tapaus oli Netflixin ”tietovuoto”, jota käsittelen luvussa 4.2.

Permutaatiota voidaan pitää yhtenä kohinan lisäämisen erikoismuotona. Esimerkiksi tietojoukossa, jossa on henkilö A, jonka pituus-attribuutin arvo on 175 ja henkilö B, jonka sama attribuutti on 180, voitaisiin käyttää permutaatiotekniikkaa ja vaihtaa pituus-attribuuttien paikkaa, eli osoittaa henkilön B pituus henkilöön A ja päinvastoin. Tällainen paikkojen vaihtelu varmistaa, että arvojen vaihteluväli ja jakauma pysyvät samoina mutta korrelaatiot yksilöiden ja arvojen välillä eivät (Article 29 Data Protection Working Party, 2014). Permutaatiolla on kuitenkin puutteensa, sillä se ei välttämättä estä esimerkiksi päättelyhyökkäyksiä. Jos permutoidussa tietojoukossa olisi muun muassa kahden henkilön, vuonna 1980 syntyneen huoltomiehen sekä vuonna 1957 syntyneen toimitusjohtajan vuositulot, joista toimitusjohtajan tulot olisivat 40,000 euroa ja huoltomiehen 100,000 euroa vuodessa, ei olisi kovin vaikeaa päätellä, kenelle vuositulot todennäköisesti oikeasti kuuluvat.

Differentiaalinen yksityisyys kuuluu satunnaistamisen tekniikoiden perheeseen, mutta erilaisella lähestymistavalla: kun kohinan lisääminen tehdään etukäteen ennen tietojoukon suunniteltua julkaisua, differentiaalista yksityisyyttä voidaan käyttää juuri silloin siinä vaiheessa, kun datan kontrolloija generoi tietojoukosta anonymisoituja näkymiä, samalla kopion alkuperäisestä datasta säilyttäen. Tällaiset anonymisoidut näkymät generoitaisiin tyypillisesti joidenkin tiettyjen kyselyiden tuloksena jotain tiettyä kolmatta osapuolta varten. Tällä tavoin tietojoukon tuleva käyttäjä ei voi tietää, onko dataa käsitelty vai ei, sillä hänellä ei ole pääsyä alkuperäiseen dataan (Mivule, 2013). Differentiaalisen yksityisyyden heikkoutena on kyselyjen tulosten yhdisteleminen ja jokaisen kyselyn toisistaan riippumaton käsittely, eli jos kolmannen osapuolen (mahdollisen hyökkääjän) kyselyhistoriaa ei säilytetä, voi hyökkääjä suunnitella kysymysten sarjan, jolla voidaan saada esiin yksittäisen rekisteröidyn tai rekisteröityjen ryhmän ominaispiirre. Tässä tapauksessa tietojoukkojen hallinnoijalla on siis merkittävä rooli yksityisyydensuojan säilyttämisessä.

Differentiaalinen yksityisyys ottaa kantaa paradoksiin, jossa halutaan välttyä paljastamasta mitään tietoa yksilöstä, samalla paljastaen hyödyllistä informaatiota kokonaisesta väestöstä (Dwork & Roth, 2014). Lääketieteellisestä tietokannasta voi käydä ilmi, että tupakointi aiheuttaa syöpää, mikä taas

vaikuttaa vakuutusyhtiön näkemykseen pitkäaikaistupakoijan pitkän aikavälin terveydenhuoltokuluista. Jossain mielessä voidaankin sanoa, että tämä analyysi on vahingoittanut tupakoivaa henkilöä, sillä hänen vakuutusmaksunsa voivat nousta, jos vakuutusyhtiö tietää tämän tupakoijaksi. Analyysistä voi olla myös hyötyä hänelle: saatuaan tietoa tupakoinnin riskeistä, hän voi ryhtyä lopettamaan tupakoinnin (Dwork & Roth, 2014). Dwork ja Roth (2014) pohtivat, onko tupakoijan yksityisyyttä loukattu tai vaarannettu analyysillä ja onko hänen tietojansa ”vuodettu”. Differentiaalisen yksityisyyden näkökulmasta asia ei ole näin, perustellen siten että analyysin vaikutus tupakoijaan on *riippumaton siitä, oliko hän osallisena analyysissä vai ei*. Tupakoijaan vaikuttavat analyysissä *saavutetut lopputulokset*, ei hänen mukanaolonsa tai poissaolonsa tietojoukossa. (Dwork & Roth, 2014.)

Koostaminen ja *k-anonymiteetti* pyrkivät estämään tiedon kohteen tunnistamista tietojoukosta ryhmittämällä ne ainakin *k* muun yksityishenkilön kanssa, jolloin jokainen tallenne on erottamaton ainakin *k*-1 muusta tallenteesta (Machanavajjhala ym., 2007). Esimerkiksi sijainti voidaan yleistää kaupungista maahan, jolloin joukkoon sisältyy suurempi määrä tietojoukkoon rekisteröityjä henkilöitä. Syntymäajat voidaan yleistää ajanjaksoiksi tai ryhmitellä kuukausittain tai vuosittain. Muiden numeeristen arvojen (kuten pituuden, painon tai palkan) luokitusta voidaan yleistää tarkan lukeman sijaan jollekin välille, esimerkiksi palkka välille 20,000 - 30,000 euroa (Article 29 Data Protection Working Party, 2014). K-anonymiteetin ongelma on se, ettei sekään estä päättelyhyökkäyksiä. Jos tietojoukossa on tarpeeksi attribuutteja, voi niistä olla yleistämisestä huolimatta mahdollista tehdä päätelmiä yksityishenkilöiden identiteeteistä (Article 29 Data Protection Working Party, 2014).

Article 29 Data Protection Working Party (2014) ottaa lausunnossaan huomioon myös *pseudonymisoinnin* eli peitenimillä suojaamisen, mutta sitä he eivät kuitenkaan laske anonymisoinnin keinoksi. Samaa sanovat myös El Emam ja Álvarez (2014).

Ohm (2010) keskittyy artikkelissaan laajaan ja mielestään tärkeään anonymisointitekniikoiden osajoukkoon, jota hän kutsuu ”julkista ja unohda” –anonymisoinniksi (eng. *release-and-forget anonymization*). Tällä tarkoitetaan sitä, että näitä tekniikoita käyttävä datan hallinnoija julkaisee tietojoukkoja ja sen jälkeen ”unohtaa” asian, eli hän ei millään tavoin pyri seuraamaan, mitä julkaistulle datalle tapahtuu julkaisun jälkeen. Ennen julkaisua hän kuitenkin pyrkii anonymisoimaan dataa muokkaamalla sitä. (Ohm, 2010.)

Ohm (2010) keskittyy julkista ja unohda –tekniikoihin kahdesta syystä: ne ovat laajasti käytössä ja ne ovat usein viallisia. Ohmin (2010) mukaan monet viimeaikaiset jälleentunnistamisen saralla otetut kehitysaskeleet ovat kohdistuneet nimenomaan julkista ja unohda –tekniikoihin. Seuraavaksi käsittelen muutamia yleisiä tekniikoita, jota varten tarvitaan yksinkertaistettu ja hypoteettinen sairaalan vierailu- ja vaivatietokanta. Tämä tietokanta on esitetty taulukkomuodossa (taulukko 1).

TAULUKKO 1 Alkuperäinen (anonymisoimaton) data

Nimi	Kansalaisuus	Syntymäpäivä	Sukupuoli	Postinumero	Vaiva
Samuli	Suomi	20/9/1965	Mies	02141	Hengenahdistus
Daniel	Suomi	14/2/1965	Mies	02141	Rintakipu
Kati	Suomi	23/10/1965	Nainen	02138	Silmäkipu
Maria	Suomi	24/8/1965	Nainen	02138	Kähisevä ääni
Helena	Suomi	7/11/1964	Nainen	02138	Nivelkipu
Jaana	Suomi	1/12/1964	Nainen	02138	Rintakipu
Åke	Ruotsi	23/10/1964	Mies	02138	Hengenahdistus
Ulla	Ruotsi	15/3/1965	Nainen	02139	Korkea verenpaine
Sven	Ruotsi	13/8/1964	Mies	02139	Nivelkipu
Håkan	Ruotsi	5/5/1964	Mies	02139	Kuume
Pelle	Ruotsi	13/2/1967	Mies	02138	Oksentelu
Matias	Ruotsi	21/3/1967	Mies	02138	Selkäkipu

Suojatakseen taulukossa olevien ihmisten yksityisyyttä, Ohm (2010) esittää sairaalan käyttävän neljää tekniikkaa: tunnistavan informaation erottaminen (eng. *singling out identifying information*), vaimentaminen (eng. *suppression*), yleistäminen (eng. *generalization*) ja koostaminen (eng. *aggregation*).

Tunnistavan informaation erottaminen: datan hallinnoija erottelee datasta sellaiset kentät, joiden avulla voitaisiin kyetä tunnistamaan yksittäisiä henkilöitä. Tällaisia kenttiä ovat paitsi yleisesti tiedetyt tunnisteet kuten nimi ja sosiaaliturvatunnus, myös eri kenttien yhdistelmät, jotka yhdessä saattaisivat voida liittää taulukossa olevan tallenteen potilaan identiteettiin. Joskus datan hallinnoija valitsee potentiaalisesti tunnistavat kentät itse, joko intuitiivisesti (eristämällä tunnistavilta vaikuttavia datatyyppejä) tai analyyttisesti (etsimällä uniikkeja attribuutteja datasta). Esimerkiksi esitetyssä tietojoukossa (taulukko 1) kellään kahdella henkilöllä ei ole sama syntymäpäivä. Tästä johtuen datan hallinnoijan on suhtauduttava syntymäpäivään yksilöivänä tunnisteena. Jos hän ei tekisi niin, kuka tahansa Åken syntymäpäivän (ja hänen sairaalavierailustansa) tietävä voisi löytää Åken anonymisoidusta datasta. (Ohm, 2010.)

Vaimentaminen: seuraavaksi datan hallinnoija päättää (tai on saanut ohjeistuksen), että nimi, syntymäpäivä, sukupuoli ja postinumero ovat potentiaalisia yksilöiviä tunnisteita. Hän voi vaimentaa ne poistamalla ne taulukosta kokonaan. Tuloksena olisi hyvin paljon suppeampi tietojoukko (taulukko 2). (Ohm, 2010.)

TAULUKKO 2 Neljän tunnisteen vaimentaminen

Kansalaisuus	Vaiva
Suomi	Hengenahdistus
Suomi	Rintakipu
Suomi	Silmäkipu
Suomi	Kähisevä ääni
Suomi	Nivelkipu
Suomi	Rintakipu
Ruotsi	Hengenahdistus
Ruotsi	Korkea verenpaine
Ruotsi	Nivelkipu
Ruotsi	Kuume
Ruotsi	Oksentelu
Ruotsi	Selkäkipu

Tietojoukon tämän version kanssa yksityisyydestä tuskin koituu ongelmia: vaikka tiedettäisiin Åken syntymäpäivä, sukupuoli, postinumero ja kansalaisuus, hänen vaivaansa ei silti kyettäisi selvittämään. Toisaalta tämä aggressiivinen vaimentaminen on tehnyt datasta tutkijoille lähes hyödytöntä. (Ohm, 2010.)

Yleistäminen: paremman tasapainon saavuttamiseksi hyödyllisyyden ja yksityisyyden välillä, datan tunnisteita voitaisiin yleistää niiden vaimentamisen sijaan. Tämä tarkoittaa tunnisteiden arvojen muokkaamista niiden poistamisen sijaan yksityisyyden parantamiseksi hyödyllisyys säilyttäen. Esimerkiksi nimikenttä voidaan vaimentaa, yleistää syntymäpäivä pelkän vuoden tarkkuustasolle ja yleistää postinumero vain kolmen ensimmäisen numeron tarkkuuteen (taulukko 3). (Ohm, 2010.)

TAULUKKO 3 Yleistetty data

Kansalaisuus	Syntymävuosi	Sukupuoli	Postinumero	Vaiva
Suomi	1965	Mies	021*	Hengenahdistus
Suomi	1965	Mies	021*	Rintakipu
Suomi	1965	Nainen	021*	Silmäkipu
Suomi	1965	Nainen	021*	Kähisevä ääni
Suomi	1964	Nainen	021*	Nivelkipu
Suomi	1964	Nainen	021*	Rintakipu
Ruotsi	1964	Mies	021*	Hengenahdistus
Ruotsi	1965	Nainen	021*	Korkea verenpaine
Ruotsi	1964	Mies	021*	Nivelkipu
Ruotsi	1964	Mies	021*	Kuume
Ruotsi	1967	Mies	021*	Oksentelu
Ruotsi	1967	Mies	021*	Selkäkipu

Vaikka joku tietäisi Åken syntymäpäivän, postinumeron, sukupuolen ja kansalaisuuden, olisi hänen hyvin vaikea eristää tietojoukosta juuri Åken ilmoittama vaiva. Tallenteita tästä tietojoukosta (taulukko 3) on vaikeampi tunnistaa kuin alkuperäisestä datasta (taulukko 1), mutta esimerkiksi tutkijoille tämän tietojoukon data on huomattavasti hyödyllisempää kuin vaimennettu data (taulukko 2). (Ohm, 2010.)

Koostaminen: data-analyytikoille usein riittävät pelkät yhteenvedot tilastoista raa'an datan sijaan. Jos tutkijoiden tarvitsee vain tietää, kuinka moni mies valitti hengenahdistuksesta, datan hallinnoijat voisivat julkaista koostetun tilaston tiedoista (taulukko 4). (Ohm, 2010.)

TAULUKKO 4 Koostettu tilasto

Miehet hengenahdistus	2

Åke oli toinen miehistä, joita tämä tilasto (taulukko 4) kuvastaa, sillä hän oli valittanut hengenahdistuksesta. Kuitenkin, ilman runsasta määrää täydentävää informaatiota, kukaan ei voisi sitä selvittää: hänen yksityisyytensä on siis turvassa. (Ohm, 2010.)

Article 29 Data Protection Working Party (2014) tunnistaa ehdotuksessaan kolme anonymisointiin liittyvää riskiä: *erottaminen joukosta* eli mahdollisuus eristää datasta osa tai kaikki tallenteet, jotka tunnistavat yksittäisen henkilön, y*hdisteltävyys* eli mahdollisuus yhdistää vähintään kaksi samaan kohteeseen tai kohteisiin liittyvää tallennetta joko samassa tietokannassa tai eri tietokannoissa sekä *pääteltävyys* eli mahdollisuus hyvällä todennäköisyydellä päätellä attribuutin arvo toisen attribuuttijoukon perusteella. Nämä kolme riskiä ovat anonymisoinnin kannalta tärkeimmät, ja eri tekniikoiden käytöllä niihin voidaan varautua.

3.2 Pseudonymisointi

Pseudonymisointi eli peitenimillä suojaaminen tarkoittaa anonymisoinnin piirissä yhden (tyypillisesti uniikin) attribuutin korvaamista tallenteessa toisella, esimerkiksi käyttäjänimen muuttamista toiseksi (Lundin & Jonsson, 2000). Luonnollinen henkilö voidaan todennäköisesti siksi edelleen tunnistaa epäsuorasti, eli pseudonymisointi ei yksinään takaa anonyymiä tietoaineistoa. Pseudonymisointia on esimerkiksi käyttäjänimien korvaaminen datassa jollakin generoidulla tai juoksevalla numerosarjalla. Tämä ei kuitenkaan poista yhteyttä käyttäjää koskevan datan ja tämän numerosarjalla korvatun nimimerkin välillä.

Pseudonymisointia ei pidetä anonymisoinnin keinona (El Emam & Álvarez, 2014). Se vain vaikeuttaa datajoukon yhdistettävyyttä alkuperäisen datan kohteen identiteettiin, mutta on kuitenkin tällä tavoin hyödyllinen turvallisuustoimenpide (Article 29 Data Protection Working Party, 2014). Yksinään se ei kuitenkaan ole riittävä takaamaan sitä, että jälleentunnistamisen

riski pysyy pienenä. Tämä johtuu siitä, että pseudonymisointia ei käytetä epäsuoriin tunnisteisiin, vaan yleensä suoriin (El Emam & Álvarez, 2014). Suurin osa tunnetuista ja onnistuneista jälleentunnistamishyökkäyksistä kohdistui pseudonymisoituun dataan (El Emam & Álvarez, 2014).

Pseudonymisoinnilla voidaan varmistaa, että käyttäjä joka toimii yhden tai useamman pseudonyymin takana voi käyttää resursseja tai palveluita paljastamatta identiteettiään keinotekoisten nimien käytön ansiosta (Tinabo, Mtenzi & O'Shea, 2009). Joissain tilanteissa kuitenkin on mahdollista kääntää pseudonyymit käyttäjien identiteeteiksi (Fischer-Hübner, 2001). Tämä on erittäin tärkeää, jos esimerkiksi sairaalassa potilaan täytyisi saada tietynlaista hoitoa tutkimustyön tuloksena (Tinabo ym., 2009).

3.3 Yhteenveto

Anonyymin tietojoukon luominen monipuolista henkilökohtaista dataa sisältävästä tietojoukosta samalla mahdollisimman paljon hyödyllistä informaatiota säilyttäen ei ole helppo tehtävä. Hyvinkin anonymisoitu tietoaineisto voidaan esimerkiksi yhdistää johonkin toiseen tietoaineistoon, ja tätä kautta tietojoukossa olevia yksityishenkilöitä on mahdollista tunnistaa ja yksilöidä (Ohm, 2010).

Luvussa käsittelin tekniikoita satunnaistamisen ja yleistämisen ryhmistä. Jokaisella tekniikalla on käyttökohteensa ja hyötynsä – mutta myös heikkoutensa – ja yksittäin käytettynä ne eivät tarjoa riittäviä takeita yksityisyydensuojan säilyttämisessä. Useimmat tekniikat ovat heikkoja ulkopuolista eli täydentävää informaatiota hallussaan pitävän vastustajan sekä päättelyhyökkäysten edessä. Anonymisointitekniikoita onkin siksi käytettävä yhdistellen sekä tapauskohtaisesti parhaan mahdollisen anonymisointituloksen saavuttamiseksi ja yksityisyydensuojan takaamiseksi. Toimiva tekniikoiden yhdistelmä voi jossain määrin vastata joukosta erottamisen, yhdisteltävyyden ja pääteltävyyden riskeihin, muttei välttämättä poista niitä kokonaan.

4 DEANONYMISOINTI

Tässä luvussa luon katsauksen tietojoukkojen deanonymisointiin eli jälleentunnistamiseen. Käsittelen tapauksia, joissa tietojoukkojen anonymisointi on pettänyt joko tietojoukkojen ylläpitäjistä johtuen tai vastustajien aktiivisten toimien vuoksi, jolloin anonyymeiksi mielletyistä tietojoukoista on tavalla tai toisella kyetty tunnistamaan yksittäisiä henkilöitä. Ensimmäinen tarkasteltava tapaus on AOL:n datan julkaisu, toinen koskee Sweeneyn (2000) tutkimusta, ja kolmas Netflixin järjestämää kilpailua. Toisin kuin anonymisointiin, deanonymisoinnin piiriin ei niinkään kuulu joukkoa hallittuja tekniikoita, vaan monimuotoisia ja kaoottisiakin keinoja selvittää tietoja datasta. Toisin sanoen, hieman yleistäenkin, deanonymisointiin voidaan suhtautua anonymisoinnin epäonnistumisena.

4.1 Deanonymisoinnin määritelmä

Datan anonymisoinnin perusajatus on selkeä: dataa muokataan siten, ettei siitä voida enää tunnistaa yksittäisiä henkilöitä. Ensin datasta poistetaan selkeät henkilökohtaiset tunnisteet, kuten nimet ja sosiaaliturvatunnukset. Seuraavaksi muokataan muuta informaatiota sisältäviä kategorioita, jotka toimivat tunnisteina kyseisessä kontekstissa: sairaalat poistavat lähiomaisten nimet, koulut poistavat opiskelijoiden tunnistenumerot ja pankit tilinumerot. Tämän ratkaisun olisi tarkoitus olla paras kaikille osapuolille: analyytikoille data on edelleen hyödyllistä, mutta häikäilemättömien markkinoijien ja pahansuopien identiteettivarkaiden on mahdotonta tunnistaa datasta yksittäisiä henkilöitä. Ainakin teoriassa data on nyt anonymisoitu ja lainsäätäjät ja kriitikot pysyvät tyytyväisinä. Todellisuudessa näin ei kuitenkaan ole. (Ohm, 2010.)

Deanonymisointi, tai jälleentunnistaminen, tarkoittaa toimenpiteitä, joilla anonyymiksi tarkoitetusta tietojoukosta pyritään tunnistamaan yksityishenkilöitä ja heidän tietojaan. Deanonymisointi tuli terminä tunnetuksi vuonna 2006 kahden tutkijan, Arvind Narayananin ja Vitaly Shmatikovin,

ottaessa osaa Netflixin järjestämään kilpailuun, jossa osanottajien tehtävänä oli kehittää Netflixille tarkempi elokuvien suosittelualgoritmi. Deanonymisointi on monesti tarkoituksenmukaista toimintaa: ihmisten anonymisoidut identiteetit harvoin paljastuvat vahingossa, ilman jonkin toimijan aktiivisia edesottamuksia. Henkilöä, joka pyrkii deanonymisoimaan tietojoukkoja, kutsutaan vastustajaksi. Vastustajan motiivit voivat olla esimerkiksi vahingon tekeminen ja ilkivalta, kokeilemisenhalu, poliittiset syyt tai lähes mikä tahansa. Heitä voivat olla ahdistelijat, etsivät tai tutkijat, työtoverit, työnantajat tai naapurit (Narayanan & Shmatikov, 2009). Ohmin (2010) mukaan heitä voivat myös olla poliisit, valtion turvallisuusanalyytikot, mainostajat sekä kenet tahansa muun, joka haluaa yhdistää yksityishenkilön dataan.

4.2 Kolme deanonymisointitapausta – AOL, Massachusetts ja Netflix

Ensimmäinen tapaus koskee AOL-palvelua, joka on hyvä esimerkki pseudonymisoinnin ongelmista, joista kerrotaan lisää luvussa 4.3. Vuonna 2006 America Online (nykyisin ja jälkeenpäin AOL), internet-palveluntarjoaja ja hakukone, julkaisi hankkeen nimeltään ”AOL Research”, jonka tarkoituksena oli luoda tietä avoimelle tutkimusyhteistyölle. AOL Research julkisti sivuillaan vapaaseen käyttöön 650,000 käyttäjän kolmen kuukauden ajanjaksolla AOL:n hakukoneella tekemät 20 miljoonaa hakukyselyä (Arrington, 2006). Internetin käyttäytymistutkijat olivat haltioissaan (Hafner, 2006), sillä aikaisemmin hakukoneiden ylläpitäjät olivat kohdelleet tämän kaltaista informaatiota tiukasti varjeltuna salaisuutena (Ohm, 2010).

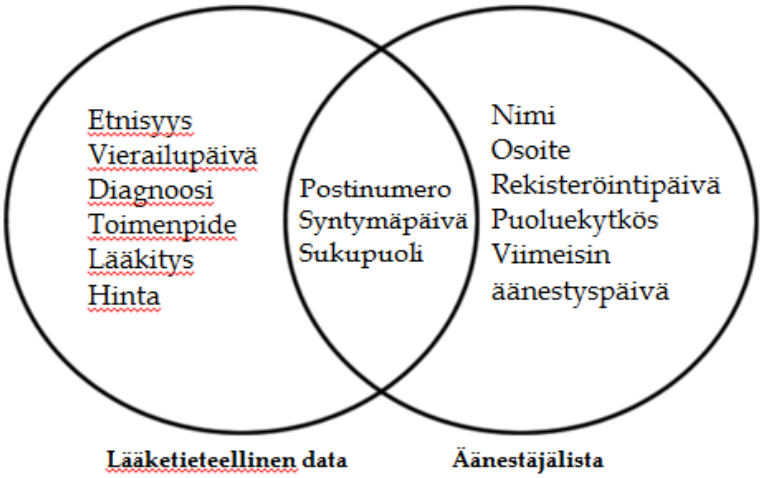
Ennen julkaisua AOL oli yrittänyt anonymisoida dataa yksityisyyden suojaamiseksi: se poisti selkeitä tunnisteita, kuten käyttäjien nimimerkit sekä IP-osoitteet. Säilyttääkseen datan hyödyllisyyttä tutkimuksen kannalta, AOL kuitenkin korvasi nämä tunnisteet uniikeilla tunnistenumeroilla, joiden avulla tutkijat pystyivät korreloimaan tehdyt haut yksittäisiin käyttäjiin. (Ohm, 2010.)

Datan julkaisua seuraavina päivinä verkon käyttäjät kahlasivat datan läpi, tuoden kerta toisensa jälkeen esille tietovuodon luonteen ja laajuuden. Heitä motivoivat joko halu tunnistaa yksityishenkilöitä tiedoista tai löytää erityisen viihdyttäviä tai järkyttäviä hakusanoja (Ohm, 2010). Siinä he onnistuivatkin, ja monien keskustelupalstaviestien ja uutisraporttien tuloksena tapaus sai laajaa huomiota osakseen. Hakusanojen joukossa oli ihmisten nimiä, osoitteita, sosiaaliturvatunnuksia, sekä huumausaineisiin ja pornoon liittyviä hakuja (Arrington, 2006). Yhden tietyn käyttäjän hakuhistoriasta löytyi kymmeniä hyvin hätkähdyttäviä ja makaabereja hakusanoja kuten ”how to kill your wife” ja ”pictures of dead people” (Frind, 2006).

Samalla kun monet internetin käyttäjät tuomitsivat AOL:n datan julkaisun, muutama bloggaaja vasta-argumentoi ettei julkaisu loukkaa kenenkään yksityisyyttä, sillä oikeita yksittäisiä henkilöitä ei ollut yhdistetty

hakukyselyihin. Tämä väittely hiljeni nopeasti, kun New York Timesin toimittajat Barbano ja Zeller (2006) löysivät vihjeitä käyttäjä 4417749:n identiteettiin tämän tekemien hakujen (asuinaluetietoja ja useita ihmisiä joiden sukunimi on ”Arnold”) perusteella. He kykenivät jäljittämään Thelma Arnoldin, 62-vuotiaan lesken Lilburnista, Georgiasta, joka myönsi tehneensä haut, joihin sisältyi myös hakuja kuten ”numb fingers” (puutuneet sormet), ”60 single men” (60 sinkkumiehet) ja ”dog that urinates on everything” (koira joka virtsaa kaikkialle). AOL irtisanoi datan julkaisseen tutkijan sekä hänen esimiehensä, ja Chief Technology Officer –tittelillä työskennellyt Maureen Govern irtisanoutui. (Ohm, 2010.)

Toinen tunnettu jälleentunnistamistapaus koski postinumeroa, sukupuolta sekä syntymäpäivää. Sweeneyn (2000) tutkimuksessa arvioidaan, että 87 % Yhdysvaltojen populaatiosta voidaan tunnistaa yksilöllisesti käyttäen näitä kolmea harmittomalta vaikuttavaa attribuuttia. Itse asiassa Sweeney (2000) käytti näitä kolmea attribuuttia Yhdysvalloissa Massachusettsissa äänestäjiksi rekisteröityneiden ihmisten tietokannan (jonka Sweeney osti laillisesti 20 dollarin hintaan, ja joka sisälsi rekisteröityneen nimen, sukupuolen, postinumeron ja syntymäpäivän) yhdistämiseksi anonyymiksi miellettyyn potilastietodataan. Ohmin (2010) mukaan monet Yhdysvaltojen osavaltiot antavat, myyvät tai muuten julkaisevat potilas- tai hoitotietodataa tutkijoille ja teollisuuden käyttöön. Tämä potilastietodata sisälsi potilaan sukupuolen, postinumeron, syntymäajan sekä diagnoosin. Myöhemmässä tutkimuksessaan Sweeney (2002) kertoo, kuinka hän tällaisella “yhdistelyhyökkäyksellä” kykeni tunnistamaan ja yksilöimään Massachusettsin kuvernöörin hoitotiedot potilastietodatasta (kuvio 1). Toisin sanoen, anonyymiksi tarkoitetusta tietojoukosta voi olla mahdollista yksilöidä ja tunnistaa henkilöitä käyttäen hyväksi muita tietojoukkoja tai muuta saatavilla olevaa informaatiota.



KUVIO 1 Jälleentunnistaminen dataa yhdistelemällä

Kolmas tunnettu jälleentunnistamistapaus koski videopalveluntarjoaja Netflixin asiakastietokantaa vuonna 2006. Netflix on vuonna 1997 Yhdysvalloissa perustettu verkossa toimiva elokuvien ja televisiosarjojen

maksullinen suoratoistopalvelu. Palvelun avulla käyttäjät voivat katsoa haluamiansa elokuvia ja sarjoja sekä antaa niille arvosanan. Vuonna 2014 palvelulla oli 62 miljoonaa käyttäjää yli 50 maassa.

Netflix julkaisi vuonna 2006 vapaaseen käyttöön 100 miljoonan tallenteen tietojoukon, joka sisälsi lähes puolen miljoonan Netflix-käyttäjän elokuva-arviot kuuden vuoden ajalta. Julkaisun syynä oli Netflixin kilpailu, jonka pääpalkintona oli miljoona dollaria sille, joka kehittäisi vähintään 10 prosenttia silloista järjestelmää tehokkaamman elokuvien suosittelualgoritmin. Jokaiseen tallenteeseen kuului arvioidun elokuvan nimi, annettu arvio (yhdestä viiteen tähteä) ja arvion päivämäärä. Netflix anonymisoi tallenteet poistamalla tunnistetietoja kuten käyttäjänimimerkit, mutta asetti kuitenkin kaikille käyttäjille uniikin tunnisteen, jotta yhteys arvostelusta toiseen säilyisi (tämä toimintatapa voitaneen laskea pseudonymisoinnin piiriin kuuluvaksi). Tällä tavoin tutkijat kykenivät kertomaan käyttäjän numero 1005 antaneen elokuvalle Godzilla arvosanan 4 marraskuun kolmantena päivänä vuonna 2003 ja elokuvalle Reservoir Dogs arvosanan 5 maaliskuun viidentenä vuonna 2005. Tietoaineistoon oli lisätty kohinaa, sillä arvioita oli nostettu tai laskettu hieman (Article 29 Data Protection Working Party, 2014).

Netflixin motiivi datan julkaisuun oli taloudellinen: sen yksi tärkeä ominaisuus on kyky tehdä tarkkoja elokuvasuosituksia käyttäjilleen. Jos Netflix päättelee esimerkiksi, että käyttäjät pitivät elokuvasta Godzilla, tulisivat he luultavasti myös pitämään elokuvasta Cloverfield. Näin Netflix voi tehdä tarkkoja suosituksia, millä käyttäjät saadaan palaamaan palvelun pariin uudestaan ja uudestaan. (Ohm, 2010.)

Huolimatta anonymisointitoimenpiteistä, Ohmin (2010) mukaan kaksi viikkoa datan julkaisun jälkeen kaksi tutkijaa Texasin yliopistosta, Narayanan ja Shmatikov (2008), ilmoittivat, että hyökkääjä, joka tietää vain vähän yksittäisestä käyttäjästä voi helposti tunnistaa kyseisen käyttäjän tallenteen jos se esiintyy Netflixin julkaisemassa tietojoukossa. Jos vastustaja tietää noin kahden viikon tarkkuudella, milloin tietokannassa esiintyvä käyttäjä on arvioinut kuusi elokuvaa (olivat ne sitten enemmän tai vähemmän tunnettuja), voi hän tunnistaa käyttäjän 99 prosenttia ajasta (Ohm, 2010).

Kaksi vuotta myöhemmin julkaistussa tutkimuksessaan Narayanan ja Shmatikov (2008) käyttivät deanonymisointimetodeitaan julkaistuun Netflix Prize –tietojoukkoon, joka sisälsi 500 000 rekisteröityneen Netflix-käyttäjän anonyymit elokuva-arviot. He vertasivat Netflixin elokuva-arviodataa samankaltaiseen dataan Internet Movie Database -palvelussa (IMDb), joka on elokuvia käsittelevä sivusto, jossa käyttäjät voivat myös arvioida elokuvia. Toisin kuin Netflix, IMDb julkaisee nämä arviot julkisesti sivustollaan. Vertaamalla Netflixissä annettuja arvioita IMDb:n arvioihin, Narayanan ja Shmatikov (2008) kykenivät 50 IMDb-käyttäjän otannalla (he olisivat halunneet käyttää IMDb:n koko arvostelutietokantaa, mutta pelkäsivät IMDb:n käyttöehtojen kieltävän sen) tunnistamaan kaksi käyttäjää lähes täydellisellä tilastollisella tarkkuudella.

Käyttämällä Internet Movie Database -verkkosivua taustatietojen lähteenä, tutkijat kykenivät tunnistamaan Netflixin käyttäjien ilmeiset poliittiset kannat sekä muuta potentiaalisesti arkaluontoista informaatiota perustuen käyttäjien arvostelemiin elokuviin ja niiden teemoihin (Narayanan & Shmatikov, 2008). Ennen tätä tapausta tuskin kukaan olisi luokitellut näitä attribuutteja henkilökohtaisesti tunnistavaksi informaatioksi nimien, henkilötunnusten tai osoitteiden tapaan. Huolimatta tutkimuksen tuloksista, yritykset ovat tapauksen jälkeenkin julkaisseet tämänkaltaista arkaluontoiseen dataan liittyvää informaatiota oletetusti anonymisoiduissa tietojoukoissa vailla minkäänlaisia seuraamuksia. (Ohm, 2010.)

4.3 Pseudonymisoinnin ongelmia

Pseudonymisointi tarkoittaa yhden attribuutin (esimerkiksi käyttäjänimen) korvaamista tallenteessa toisella (Lundin & Jonsson, 2000). Tästä johtuen pseudonymisoinnin jälkeen on edelleen mahdollista yksilöidä yksityishenkilöiden tallenteet tietojoukosta, sillä yksilöiden tunnisteina on edelleen pseudonymisoinnin tuloksena uniikki attribuutti, kuten esimerkiksi ID-numero. Oleellista on, ettei tunnistetta voida assosioida alkuperäisen tunnistedatan kanssa jotain salaisuutta (kuten pseudonymisointiin käytettyä tiivistefunktiota) tietämättä (Riedl, Grascher, Fenz & Neubauer, 2008).

Yhdisteltävyys on edelleen kohtalaisen yksinkertaista sellaisten tallenteiden välillä, jotka käyttävät samaa pseudonymisoitua attribuuttia viitatakseen samaan yksityishenkilöön. Vaikka eri pseudonymisoituja attribuutteja käytettäisiin samaan datan kohteeseen, yhdisteltävyys voi edelleen olla mahdollista muiden attribuuttien avulla. Vain jos mitään muuta attribuuttia ei voida käyttää datan kohteen tunnistamiseen ja vain jos jokainen yhteys alkuperäisen attribuutin ja pseudonymisoidun attribuutin välillä on eliminoitu (mukaan lukien alkuperäisen datan poistaminen), itsestäänselviä ristiviitteitä kahden eri pseudonymisoituja attribuutteja käyttävän tietojoukon välillä ei ole. (Article 29 Data Protection Working Party, 2014.)

Päättelyhyökkäykset datan kohteen todellisen identiteetin selvittämiseksi ovat edelleen mahdollisia saman tietojoukon sisäisesti tai eri tietokantojen välillä, jotka käyttävät samaa pseudonymisoitua attribuuttia yksilölle, tai jos pseudonyymit ovat itsestäänselviä eivätkä peitä alkuperäistä datan kohteen identiteettiä asianmukaisesti (Article 29 Data Protection Working Party, 2014). Esimerkiksi terveydenhuollollisessa datassa

Yleinen erehdys on pseudonymisoidun tietojoukon anonyymiksi mieltäminen: datan kontrolloijat olettavat usein, että yhden tai useamman attribuutin poistaminen tai korvaaminen riittää tietojoukon anonymisoimiseksi. Monet esimerkit ovat kuitenkin osoittaneet tämän olevan väärin; pelkkä ID:n muuntelu ei estä jotakuta tunnistamasta datan kohdetta, jos kvasitunniseita edelleen esiintyy tietojoukossa, tai jos muiden attribuuttien arvot kykenevät tunnistamaan yksilön. Monissa tapauksissa voi olla yhtä helppoa tunnistaa

yksityishenkilö pseudonymisoidusta tietojoukosta kuin alkuperäisestä datastakin. Ylimääräisiä askelia tulisi ottaa, jotta tietoaineisto voidaan katsoa anonymisoiduksi. Näitä askelia ovat muun muassa attribuuttien poistaminen ja yleistäminen tai alkuperäisen datan poistaminen, tai ainakin sen tuominen hyvin korkeasti koostetulle tasolle. (Article 29 Data Protection Working Party, 2014.)

Seuraavaksi käsittelen kolmea pseudonymisoinnin epäonnistumisen esimerkkitapausta. Ensimmäisenä on esimerkki terveydenhuollosta, toiseksi sosiaalisista verkostoista internetissä ja kolmanneksi fyysisiin sijainteihin liittyvä tapaus.

TAULUKKO 5 Esimerkki tiivistämällä tehdystä pseudonymisoinnista, joka voidaan peruuttaa helposti

Nimi, osoite, syntymäaika	Erityistukijakso	Painoindeksi	Tutkimuskohortin viitenumero
	< 2 vuotta	15	QA5FRD4
	> 5 vuotta	14	2B48HFG
	< 2 vuotta	16	RC3URPQ
	> 5 vuotta	18	SD289K9
	< 2 vuotta	20	5E1FL7Q

Terveydenhuolto: tietojoukon (taulukko 5) tietoaineisto on luotu, jotta voitaisiin tutkia suhdetta henkilön painon ja erityistukien saamisen välillä. Alkuperäinen tietoaineisto sisälsi datan kohteen nimen, osoitteen ja syntymäajan, mutta nämä tiedot on poistettu taulukosta. Tutkimuskohortin viitenumero on luotu poistetusta datasta käyttämällä tiivistysfunktiota. Vaikka taulukosta on poistettu nimi, osoite ja syntymäaika, on tutkimuskohortin viitenumeroiden laskeminen silti helppoa, jos yhdenkin rekisteröidyn datan kohteen nimi, osoite ja syntymäaika ovat tiedossa laskemisessa käytetyn tiivistysfunktion lisäksi (Article 29 Data Protection Working Party, 2014). Tällä tavoin suoritettu pseudonymisointi ei välttämättä takaa riittävää yksityisyyttä tiedon kohteille, sillä esimerkiksi dataa ylläpitävän organisaation sisällä olevat henkilöt kykenisivät pääsemään käsiksi dataan (Riedl ym., 2008).

Sosiaaliset verkostot: Narayanan ja Shmatikov (2009) ovat osoittaneet, että yksittäisistä henkilöistä voidaan saada selville arkaluonteista informaatiota sosiaalisten verkostojen kuvaajista, huolimatta dataan käytetyistä pseudonymisointitekniikoista. He hankkivat sosiaalisen median palvelu Twitterin koko sosiaalisen kuvaajan ja karsivat sen nimettömiksi, identiteetittömiksi ihmisiä kuvaaviksi ”solmuiksi”, jotka olivat yhteydessä solmuihin, jotka kuvasivat Twitterin ”seuraa” (eng. *follow*) -suhteita. Ohmin (2010) mukaan Narayanan ja Shmatikov (2009) sitten vertasivat tätä kuvaajaa

Flickeristä (valokuvien jakamispalvelu) kerättyyn julkiseen dataan. Selvisi, että kymmenet tuhannet Twitterin käyttäjät ovat myös Flickerin käyttäjiä, ja tutkijat käyttivät Flickerin ”kontakti”- kuvaajan ja Twitterin ”seuraa”- kuvaajan samankaltaisuuksia jälleentunnistaakseen monia anonymisoituja Twitterin käyttäjien identiteettejä. Tällä tekniikalla he kykenivät tunnistamaan kolmasosan molempiin palveluihin rekisteröityneiden käyttäjien käyttäjä- tai koko nimistä. He osoittivat, että kolmasosa käyttäjistä, jotka ovat varmuudella sekä Flickr:in että Twitterin käyttäjiä, voidaan tunnistaa Twitterin anonyymissä kuvaajassa vain 12 prosentin virhemarginaalilla, huolimatta siitä että palveluiden käyttäjien suhteiden päällekkäisyys on alle 15 prosenttia. (Ohm, 2010.)

Sijainnit: MIT:n (Massachusetts Institute of Technology) tutkijat (de Montjoye, Hidalgo, Verleysen & Blondel, 2013) analysoivat pseudonymisoidun tietojoukon, joka käsitti 1,5 miljoonan ihmisen spatio-temporaalisen liikkuvuuden koordinaatteja sadan kilometrin säteellä 15 kuukauden aikana. Tutkijat osoittivat, että 95 % väestöstä voidaan yksilöidä neljän sijaintipisteen avulla, ja että vain kaksi pistettä riittää yksilöimään yli 50 % datan kohteista (yksi tällainen piste on hyvin suurella todennäköisyydellä “koti” tai “työpaikka”). Näin ollen yksityisyyden suojaamiselle jää hyvin vähän tilaa, vaikka yksilöiden identiteetit olisivat pseudonymisoitu korvaamalla niiden todelliset attribuutit eri nimikkeillä. (Article 29 Data Protection Working Party, 2014.)

Tietojoukoissa olevia kenttiä tai attribuutteja luokitellaan yleensä joko suoriksi tunnisteiksi, epäsuoriksi tunnisteiksi tai ”muiksi”. Suoria ja epäsuoria tunnisteita voidaan käyttää tallenteiden jälleentunnistamiseen. Pseudonymisointia käytetään yleensä suorien ja uniikkien tunnisteiden, kuten sosiaaliturvatunnusten tai pankkikorttien numeroiden suojaamiseen. Sitä ei siksi siis pidetä anonymisoinnin keinona. (El Emam & Álvarez, 2014.)

4.4 Yhteenveto

Tässä luvussa käsittelin tietojoukkojen deanonymisointia eli jälleentunnistamista. Se tarkoittaa toimenpiteitä, joilla anonymisoidusta tietojoukosta pyritään tunnistamaan yksittäisiä henkilöitä ja heidän tietojaan. Deanonymisointitoimenpiteet eivät varsinaisesti ole tiettyjä tekniikoita, vaan monimuotoisia ja kaoottisiakin keinoja selvittää tietoja datasta sekä ensisijaisesti anonymisointitekniikoiden haavoittuvuuksien hyväksikäyttöä. Se on tarkoituksenmukaista toimintaa, jolla vastustajaksi kutsuttu henkilö tai taho pyrkii saamaan haltuunsa ihmisten henkilökohtaisia tietoja. Motiivina hänellä voi olla esimerkiksi vahingonteko ja ilkivalta, kokeilunhalu, poliittiset syyt tai lähes mitä tahansa muuta (Ohm, 2010).

Yhteenvetona voidaan sanoa, että huolimatta anonymisointitoimenpiteistä on deanonymisointi ainakin tähän asti ollut olemassa oleva riski. Useat tapaukset ja tutkimukset ovat osoittaneet useimpien anonymisointitekniikoiden

olevan epätäydellisiä, sillä ne eivät ole aina kyenneet estämään yksittäisten henkilöiden tietojen paljastumista anonymisoiduksi mielletystä datasta. Esimerkiksi päättelyhyökkäykset sekä täydentävää informaatiota hyödyntävät hyökkäykset ovat olleet melko yleisiä ja hyvin onnistuneita, sillä niitä on vaikea estää yleisesti käytössä olevilla anonymisointitekniikoilla.

5 JOHTOPÄÄTÖKSET

Tutkielmassa tarkastellut anonymisoinnin keinot jakautuvat yhden näkökulman mukaan kahteen ryhmään: satunnaistamisen tekniikoihin ja yleistämisen tekniikoihin (Article 29 Working Party, 2014). Kolmantena, erillisenä tekniikkana käsittelin pseudonymisointia, mutta se ei ole anonymisoinnin keino (El Emam & Álvarez, 2014). Ohm (2010) listaa joukon menetelmiä, jotka vastaavat Article 29 Working Partyn (2014) esittämiä tekniikoita, mutta hän kutsuu niitä ”julkista ja unohda” –anonymisoinniksi. Kaikilla näillä tekniikoilla on omat heikkoutensa ja vahvuutensa (taulukko 6), ja siksi niitä tulisikin käyttää toisiinsa yhdistellen.

Informaation julkaisemisen haasteena on julkaista mahdollisimman paljon hyödyllistä informaatiota, samalla halutut yksityisyyden ja turvallisuuden takeet täyttäen. Dataa julkaisevien tahojen on mahdotonta tietää, mitä taustatietoja datan jälleentunnistamista haluavalla vastustajalla mahdollisesti on käytettävissään. Siksi datan julkaisemista mietittäessä onkin varauduttava ”pahimpaan” ja oletettava, että vastustajalla voi olla käytössään lähes mitä tahansa lisä- ja taustatietoa (Martin ym., 2007).

Article 29 Data Protection Working Party (2014) tunnistaa ehdotuksessaan kolme anonymisointiin liittyvää riskiä (taulukko 6): *erottaminen joukosta* eli mahdollisuus eristää datasta osa tai kaikki tallenteet, jotka tunnistavat yksittäisen henkilön, *yhdisteltävyys* eli mahdollisuus yhdistää vähintään kaksi samaan kohteeseen tai kohteisiin liittyvää tallennetta joko samassa tietokannassa tai eri tietokannoissa sekä *pääteltävyys* eli mahdollisuus hyvällä todennäköisyydellä päätellä attribuutin arvo toisen attribuuttijoukon perusteella. Nämä kolme riskiä ovat anonymisoinnin kannalta tärkeimmät. Mikään tutkielmassa käsitellyistä anonymisointitekniikoista ei kuitenkaan täydellä varmuudella täytä tehokkaan anonymisoinnin kriteereitä, mutta osa esitellyistä riskeistä voidaan poistaa tietyllä tekniikalla kokonaan tai osittain. (Article 29 Working Party, 2014.)

TAULUKKO 6 Käsiteltyjen tekniikoiden vahvuudet ja heikkoudet

	Onko yksilöitävyys vielä riski?	Onko yhdisteltävyys vielä riski?	Onko pääteltävyys vielä riski?
Pseudonymisointi	Kyllä	Kyllä	Kyllä
Kohinan lisääminen	Kyllä	Ei ehkä	Ei ehkä
Koostaminen tai k-anonymiteetti	Ei	Kyllä	Kyllä
Permutaatio	Ehkä	Ehkä	Kyllä
Differentiaalinen yksityisyys	Ei ehkä	Ei ehkä	Ei ehkä

Artiklan 29 työryhmän lausunnon (2014) mukaan tietoaineistojen ja tietokantojen anonymisointitekniikat voivat taata yksityisyydenturvaa tiettyyn pisteeseen saakka ja niiden avulla voidaan luoda tehokkaita anonymisointiprosesseja, mutta vain jos niitä sovelletaan oikeaoppisesti. Tämä tarkoittaa sitä, että anonymisointiprosessin ennakkovaatimukset (konteksti - mihin käyttöön dataa anonymisoidaan) ja tavoitteet tulee asettaa selkeästi, jotta tavoiteltu anonymisointi saavutettaisiin ja anonymisoitu data olisi vielä hyödyllistä. Ohm (2010) toteaakin, että data voi olla joko täysin anonyymiä tai hyödyllistä, mutta ei molempia. Anonymisoinnin tarkoitus on otettava huomioon: tarkoituksena voi olla esimerkiksi yksilön tietosuojan turvaaminen tietoaineistoa julkaistaessa, tai tietyn tiedon tietoaineistosta esiin saamisen mahdollistaminen. Optimaalinen ratkaisu tulisi päättää tapauskohtaisesti ja käyttäen eri tekniikoiden kombinaatioita, sillä joillakin anonymisointitekniikoilla on luontaisia rajoituksia. (Article 29 Data Protection Working Party, 2014.)

Article 29 Data Protection Working Partyn (2014) lausunnon mukaan anonymisoinnilla voidaan saavuttaa takeita yksityisyydestä, mikä pitää jossain määrin paikkansa, mutta on silti lopputulemana toisenlainen kuin mihin Ohm (2010) artikkelissaan päätyy. Datan haltijoiden tulisi ottaa huomioon, että anonymisoitu tietoaineisto voi vielä jälkeenpäin luoda riskejä datan kohteille.

Anonymisointi ja jälleentunnistaminen ovat aktiivisia tutkimuksen kohdealueita ja uusia löydöksiä julkaistaan säännöllisesti, mutta toisaalta jopa anonymisoitua dataa, kuten tilastoja, voidaan käyttää yksityishenkilöiden olemassa olevien profiilien rikastamiseen, luoden uusia tietoturvakysymyksiä. Tästä johtuen anonymisointiin ei tulisi suhtautua vain yhtenä suoritteena, vaan jatkuvana toimintana, jossa datan kontrolloija (esimerkiksi tietokannan ylläpitäjä) seuraa ja määrittää riskejä ja uhkia säännöllisesti (Article 29 Data Protection Working Party, 2014). Jälleentunnistamisen helppous kasvaa lähes eksponentiaalisesti uusia, täydentävää informaatiota sisältäviä tietojoukkoja paljastettaessa. Ohm (2010) mainitsee, että kun hämärretty identiteetti on kyetty

tunnistamaan esimerkiksi kahta tietojoukkoa hyväksikäyttäen, on kerta toisensa jälkeen helpompaa tunnistaa identiteetti muissakin tietojoukoissa aiempaa informaatiota hyväksikäyttäen.

Ohm (2010) summaa muutamia keinoja yksityisyyden suojaamiseksi tulevaisuudessa paremmin kuin hänen mainitsemansa ”julkaiset ja unohda” -tekniikat. Yksi tapa olisi joustaa joko ”julkaise”- tai ”unohda” –vaatimuksesta. Jotkin datan hallinnojat eivät koskaan julkaise raakaa dataa, vaan julkaisevat vain koostettuja tilastoja. Tällä tavoin jälleentunnistamista haluavan olisi hyvin vaikea tunnistaa esimerkiksi johonkin kyselyyn osaa ottaneita ihmisiä, saati sitten heidän vastauksiaan (Ohm, 2010). Toinen tapa olisi suosia interaktiivisia tekniikoita. Tutkijat esittäisivät kysymyksiä datan hallinnoijille dataan liittyen, ja he etsisivät vastauksen tietojoukoistaan.

Valitettavasti näillä toimintatavoilla on taipumus olla jäykempiä kuin tavanomaiset anonymisointikeinot. Interaktiivisuus vaatii jatkuvaa osallistumista datan hallinnoijilta, mikä nostaa analyysien hintaa ja laskee uusien analyysien tekemisen tahtia. Koska tutkijoiden täytyy lähettää kyselynsä datan hallinnoijille ja odottaa vastausta, he eivät voi testata teoriaa teorian perään haluamallaan nopeudella (Ohm, 2010).

Ehkä mielenkiintoisimpana ratkaisuna Ohm (2010) ehdottaa jälleentunnistamisen kieltämistä. Lainsäätäjät voisivat perustella kieltoa hyvin suoraviivaisesti: anonymisoimalla datan, datan hallinnoija ilmaisee tarkoituksensa suojella tiedon kohteidensa yksityisyyttä, jotka saattavat vain tämän johdosta myöntyä tarjoamaan dataansa kyseiselle hallinnoijalle. Jälleentunnistamaan pyrkivä vastustaja kumoaa tämän ilmaistun tarkoituksen ja torpedoi tiedon kohteen antaman suostumuksen niin radikaalisti, että ehkä tarvitaan jälleentunnistamisen kieltävä laki. Tällainen kielto varmasti kuitenkin epäonnistuu, sillä sitä olisi mahdoton valvoa. Oletetaan esimerkiksi, että Amazon.com anonymisoi asiakasrekisterinsä ja välittää sen markkinointiyritykselle, joka lupaa ettei se pyri jälleentunnistamaan ihmisiä datasta, vaikka se voisi kasvattaa tuottojaan huomattavasti niin tekemällä. Jos markkinointiyritys rikkoo lupauksensa ja jälleentunnistaa dataa, Amazon tai kukaan muukaan tuskin saisi ikinä tietää. (Ohm, 2010.)

Eräs tärkeä anonymisoinnin käsite on henkilökohtaisesti tunnistava informaatio. Tällainen informaatio voi käsittää nimiä, henkilötunnuksia ja osoitteita, eli mitä tahansa dataa, jota voidaan käyttää yksityishenkilön identiteetin tunnistamiseen tai jäljittämiseen joko sellaisenaan tai yhdistelemällä muuhun informaatioon, joka voidaan liittää tiettyyn yksityishenkilöön (Krishnamurthy & Wills, 2009). Yksi tapa laajentaa anonymisoinnin potentiaalisesti takaamaa yksityisyyttä on monesti ollut laajentaa henkilökohtaisesti tunnistavan informaation määritelmää koskemaan aina vain useampia attribuutteja, mutta kuten on todettu, henkilökohtaisesti tunnistavan informaation määritelmä on aina vain laajentuva kategoria (Ohm, 2010). Viisitoista vuotta sitten tuskin kukaan olisi luokitellut elokuva-arvioita tai hakukyselyitä henkilökohtaisesti tunnistavaksi informaatioksi, eikä niin tehnyt mikään laki tai säädöskään (Ohm, 2010).

Tutkielmassa käsitellyt tunnetut deanonymisointitapaukset ovat informaatioteknologia-alan kehityksen nopeuden huomioon ottaen jo melko ikääntyneitä. Sekä Netflixin järjestämä kilpailu että AOL:n tietovuoto tapahtuivat noin kymmenen vuotta sitten vuonna 2006, ja Sweeney (2000) suoritti Massachusettsin äänestysrekisterin jälleentunnistamistutkimuksensa vuonna 2000. Viime vuosina vastaavia jälleentunnistamistapauksia on vaikuttanut olevan julkisuudessa hyvin vähän, eli joko tietokantojen hallinnoijat ovat olleet varovaisempia julkaisemansa datan suhteen, tai tapaukset eivät vain ole saavuttaneet suurta mediahuomiota. Tietovuotoja on kyllä tapahtunut, kuten esimerkiksi Ashley Madison –seuranhakupalvelun tietomurto (Hackett, 2015), mutta tutkielmassa käsittelemiini tapauksiin verraten Ashley Madisonin kaltaisten tietovuotojen tapauksissa datan hallinnoijat eivät alun perinkään tarkoittaneet julkaista vuodettua dataa, vaan data on päätynyt vääriin käsiin tietomurron kautta.

6 YHTEENVETO

Tutkielman tarkoituksena oli selvittää, millaisia tietojoukkojen anonymisointitekniikoita on yleisesti käytössä, ja että ovatko ne riittäviä yksityishenkilöiden yksityisyydensuojan takaamiseksi henkilön tietoja sisältävää dataa julkaistaessa. Aluksi tarkastelin yleisesti käytössä olevia anonymisointitekniikoita, jonka jälkeen käsittelin deanonymisoinnin eli jälleentunnistamisen keinoja ja menetelmiä. Deanonymisointi tarkoittaa käytännössä anonymisointitoimenpiteiden purkamista ja yksityishenkilöiden identiteettien paljastamista.

Nykyään ihmisistä kerätään paljon dataa niin verkossa kuin sen ulkopuolellakin – yleensä tiedonkeruun kohteena olevien yksityishenkilöiden sitä havaitsematta. Dataa keräävistä tahoista, datan säilytyksen turvallisuudesta tai datan käyttötarkoituksista ei tiedon kohteilla aina ole varmuutta. Esimerkiksi sairaalat keräävät potilaistaan hoitotietoja, joita voidaan julkaista tiettyjen tutkijoiden käyttöön. Ennen julkaisua tiedot on kuitenkin anonymisoitava, jottei yksittäisiä henkilöitä voida tunnistaa datasta. Anonymisoinnin keinoja ja tekniikoita on useita, mutta niillä on heikkoutensa: monet tutkijat (Sweeney, 2000; Narayanan & Shmatikov, 2008; Ohm, 2010) ovat osoittaneet, että anonymisointi on mahdollista purkaa ja saada selville ihmisten henkilökohtaista informaatiota. Henkilökohtaisesti tunnistavaa informaatiota voi nykyään olla käytännössä mikä tahansa, sillä eri aihealueiden dataa ja päällekkäistä dataa sisältävien tietojoukkojen yhdisteleminen voi helpostikin paljastaa hämärrettyjä identiteettejä.

Article 29 Working Party (2014) toteaa, että anonymisointitekniikat voivat tarjota takeita yksityisyydestä ja että niitä voidaan käyttää tehokkaiden anonymisointiprosessien luomiseksi, mutta vain jos niiden käyttö suunnitellaan asianmukaisesti. Tämä tarkoittaa sitä, että anonymisointiprosessin esivaatimukset (konteksti) ja päämäärät on selkeästi suunniteltu, jotta haluttu anonymisointi voidaan saavuttaa samanaikaisesti hyödyllistä dataa tuottaen. Ohm (2010) taas on tapahtuneiden tietovuotojen ja jälleentunnistamistapauksien valossa sitä mieltä, että tavanomaisilla ja yleisimmin käytetyillä anonymisointitekniikoilla ja toimenpiteillä ei kyetä

takaamaan riittävää anonymiteettiä yksityishenkilöille. Ohm (2010) ehdottaakin vaihtoehtoisia tai täydentäviä toimintatapoja, jotka nekään eivät kuitenkaan ole ongelmattomia. Ohm (2010) esittääkin, että kaikki informaatio voi olla henkilökohtaisesti tunnistavaa sellaiselle toimijalle, jolla on pääsy sopivaan ulkopuoliseen informaatioon.

Huolimatta siitä, että anonymisointitekniikat eivät ole täydellisiä, ei se tarkoita, ettei niitä tulisi käyttää: selkeästi henkilökohtaisesti tunnistava informaatio kannattaa edelleen karsia tietojoukoista pois, vaikka sen määritelmä onkin hyvin epämääräinen. On kuitenkin ymmärrettävä, että mahdollisuus ja riski datan deanonymisointiin on edelleen olemassa, eikä ihmisille välttämättä kannata siksi tehdä turhia lupauksia täydellisestä datan anonyymiydestä. Kuten katsauksesta käy ilmi, voi olla kohtalaisen helppoa löytää julkisesti tai muuten saatavilla olevia tietojoukkoja ja yhdistellä niitä keskenään tai tehdä hyvinkin tarkkoja päätelmiä ihmisten tunnistamiseksi. Tällaiset ”helpot” jälleentunnistamistapaukset ja ne mahdollistaneiden keinojen julkitulo tarkoittavat toivottavasti muutoksia sekä teknologian että yksityisyyden käsitteiden saralla (Ohm, 2010).

LÄHTEET

Arrington, M. (2006, 6. elokuuta). AOL Proudly Releases Massive Amounts of Private Data. Haettu 13.5.2016 osoitteesta http://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/

Article 29 Data Protection Working Party. (2007). *Opinion 4/2007 on the concept of personal data.*

Article 29 Data Protection Working Party. (2014). *Opinion 05/2014 on Anonymisation Techniques*.

Barbano, M. & Zeller, T. Jr. (2006, 9. elokuuta). A Face is Exposed for AOL Searcher No 4417749. Haettu 19.6.2016 osoitteesta http://www.nytimes.com/2006/08/09/technology/09aol.html

Chen, H., Chiang, R. H. L. & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly, 36(4)*, 1165-1188.

de Montjoye, Y.-A., Hidalgo, C., Verleysen M. & Blondel, V. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, *3(1376)*.

Ding, X., Zhang, L., Wan, Z. & Gu, M. (2010). A Brief Survey on De-anonymization Attacks in Online Social Networks. *2010 International Conference on Computational Aspects of Social Networks* (s. 611-615). Taiyuan: IEEE.

Dwork, C. & Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. Foundations and Trends. *Theoretical Computer Science*, *9(3–4)*, 211–407.

El Emam, K. & Álvarez, C. (2014) A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques. *International Data Privacy Law*.

Fischer-Hübner, S. (2001*). IT-Security and Privacy: Design and Use of Privacy-Enhancing Security Mechanisms*. New York: Springer Berlin Heidelberg.

Frind, M. (2006, 7. elokuuta). AOL Search Data Shows Users Planning to commit Murder. Haettu 17.5.2016 osoitteesta https://plentyoffish.wordpress.com/ 2006/08/07/aol-search-data-shows-users-planning-to-commit-murder/

Ganta, S. R., Kasiviswanathan, S. P. & Smith, A. (2008). Composition Attacks and Auxiliary Information in Data Privacy. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (s. 265-273). New York: ACM.

Hackett, R. (2015, 26. elokuuta). What to know about the Ashley Madison hack. Haettu 16.6.2016 osoitteesta http://fortune.com/2015/08/26/ashley-madison-hack/

Hafner, K. (2006, 23. elokuuta). Researchers Yearn to Use AOL Logs, but They Hesitate. Haettu 13.5.2016 osoitteesta http://www.nytimes.com/

2006/08/23/technology/23search.html?fta=y&_r=2&

Information Commissioner's Office (ICO). (2016, 11. toukokuuta). The Guide to Data Protection. Haettu 15.6.2016 osoitteesta https://ico.org.uk/media/for-organisations/guide-to-data-protection-2-4.pdf

Krishnamurthy, B. & Wills, C. E. (2009). On the Leakage of Personally Identifiable Information Via Online Social Networks. *WOSN '09 Proceedings of the 2nd ACM workshop on Online social networks* (s. 7-12). August 17, 2009, Barcelona, Spain. New York: ACM.

Lessig, L. (2006). *Code: Version 2.0*. New York: Basic Books.

Lundin, E. & Jonsson, E. (2000). Anomaly-based intrusion detection: privacy concerns and other problems. *Computer Networks, 34(4)*, 623-640.

Machanavajjhala, A., Gehrke, J., Kifer, D. & Venkitasubramaniam, M. (2007). l-Diversity: Privacy Beyond k-Anonymity. *ACM Transactions on Knowledge Discovery from Data, 1(1)*, Article No. 3.

Martin, D. J., Kifer, D., Machanavajjhala, A., Gehrke, J. & Halpern, J. (2007). Worst Case Background Knowledge for Privacy-Preserving Data Publishing. *Proc. 23rd International Conference on Data Engineering* (s. 126-135). Istanbul: IEEE.

Mivule, K. (2013). Utilizing Noise Addition for Data Privacy, an Overview. *Proceedings of the International Conference on Information and Knowledge Engineering (IKE 2012)*, (s. 65-71). Las Vegas.

Narayanan, A. & Shmatikov, V. (2008). Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset). The University of Texas at Austin. *arXiv preprint cs/0610105*.

Narayanan, A. & Shmatikov, V. (2009). De-Anonymizing Social Networks. *2009 30th IEEE Symposium on Security and Privacy* (s. 173-187). Berkeley, CA: IEEE.

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

Ohm, P. (2010). Broken promises of privacy: responding to the surprising failure of anonymization. *UCLA Law Review, 57,* 1701-1077.

Oikeusministeriö. (1999). Henkilötietolaki, 32§, Tietojen suojaaminen. Haettu 19.6.2016 osoitteesta http://www.finlex.fi/fi/laki/ajantasa/1999/19990523#L7P32

Pfitzmann, A. & Koehntopp, M. (2001). Anonymity, Unobservability and Pseudonymity - a Proposal for Terminology. Teoksessa H. Federrath, *Designing Privacy Enhancing Technologies* (s. 1-9). Berkeley: Springer Berlin Heidelberg.

Riedl, B., Grascher, V., Fenz, S. & Neubauer, T. (2008). Pseudonymization for improving the Privacy in E-health Applications. *Proceedings of the 41st Hawaii International Conference on System Sciences* (s. 255-264). Waikoloa, HI: IEEE.

Skandia. (2011). Haettu 14.3.2016 osoitteesta https://www.oldmutualwealth.co.uk/Media-Centre/2011-press-releases/May-2011/SKANDIA-TAKES-THE-TERMINAL-OUT-OF-TERMS-AND-CONDITIONS/

Smolan, R. & Erwitt, J. (2012). The Human Face of Big Data. Sausalito, CA: Against All Odds Productions.

Sweeney, L. (1997). Weaving Technology and Policy Together to Maintain Confidentiality. *The Journal of Law, Medicine & Ethics, 25(2-3)*, 98-110.

Sweeney, L. (2000). *Simple Demographics Often Identify People Uniquely*. Carnegie Mellon University, School of Computer Science, Data Privacy Laboratory.

Sweeney, L. (2002). K-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, *10(5)*, 557–570.

Tene, O. & Polonetsky, J. (2012). Privacy in the Age of Big Data: A Time for Big Decision. *64 Stanford Law Review Online 63*.

Netflix. (2006) The Netflix Prize Rules. Haettu 24.1.2016 osoitteesta http://www.netflixprize.com/rules

Tietosuojalautakunta. (2006, 4. huhtikuuta). Finlex, Henkilötieto – Arkaluonteinen henkilötieto. Haettu 13.4.2016 osoitteesta http://www.finlex.fi/fi/viranomaiset/ftie/2006/20060001

Tinabo, R., Mtenzi, F. & O'Shea, B. (2009). Anonymisation vs. Pseudonymisation: Which one is most useful for both privacy protection and usefulness of e-healthcare data. *International Conference for Internet Technology and Secured Transactions* (s. 1-6). London: IEEE.