TRAFICOM
Liikenne- ja viestintävirasto
Kyberturvallisuuskeskus

Tunnisteet ja tietosuoja

Anonymisointi ja sen rajat

Peuhkuri Markus

Traficomin julkaisuja

33/2021

Sisällysluettelo

Taulukot

Kuvat

Tiivistelmä

Jaettaessa tietoa eri organisaatioiden välillä joudutaan tasapainoilemaan tietosuojan ja toisaalta tiedon laadun ja sen jakamisen tehokkuuden välillä. Tunnistetietojen anonymisointiin on kehitetty useita erilaisia teknisiä keinoja. Valitettavasti monet niistä ovat murrettavissa etenkin, jos samalla halutaan säilyttää tiedon käyttöarvo esimerkiksi analysoinnin osalta. Osa tavoista taas on kehityksen alkutaipaleella eikä ole käytännössä järkevästi käytettävissä.

Tässä raportissa tarkastellaan sekä teknisiä että muita keinoja järjestää tiedonvaihtoa siten, että tiedon käyttöarvo ei kärsi liiaksi. Tarkastelussa käydään läpi keskeisimmät aiheeseen liittyvät tieteelliset tutkimukset, anonymisointiin tarjolla olevat työkalut sekä tärkeänä osana tietosuojaan liittyvää lainsäädäntöä. Erityisesti tarkastellaan tietoverkkojen tunnisteita tietoturvatyöhön liittyen mutta periaatteet ovat sovellettavissa muillekin aloille.

Anonymisointi on yksi työkalu tietosuojaperiaatteen toteuttamiseksi ja usein joudutaan tekemään kompromissejä tiedon käyttöarvon kanssa. Tämä ei kuitenkaan ole nollasummapeli tai joko-tai vaan monissa tapauksissa tietosuojaa pystytään parantamaan ilman tiedon arvon olennaista heikentymistä. Päätös anonymisoinnista, käytettävästä menetelmästä ja tietojen jakamisesta tulee tehdä harkiten sekä dokumentoida perusteet tehdylle valinnalle.

Anonymisoinnin kehittäminen ja hyvien käytäntöjen luominen edellyttää yhteistyötä niin tietosuojasta vastaavien, tietoa tuottavien että tietoja hyödyntävien kesken. Hyvää tietosuojaa ei voi olla ilman hyvää tietoturvaa.

1 Johdanto

Tiedonvaihto organisaatioiden välillä helpottaa kokonaiskuvan saamista tietoturvatilanteesta ja -uhkista parantaen kokonaisturvallisuutta. Näissä yhteyksissä joudutaan käsittelemään myös yksityistä tai luottamuksellista tietoa. Yksityinen tieto voidaan muuttaa helpommin käsiteltäväksi, jopa julkiseksi, tiedoksi anonymisoimalla se.

Tiedonvaihdon anonymiteettiä voidaan katsoa kahdelta eri näkökulmalta:

1. Informaatio on anonyymiä: tiedon lähde mahdollisesti tiedetään mutta yksilöitä tai laitetta, joita tieto koskee, ei tunneta.

2. Informaation lähde on anonyymi: ei tiedetä mistä tarkalleen tieto on peräisin.

Yksityisyyden suojan osalta ensimmäinen määritelmä on se, mitä tietosuojalla yleensä tarkoitetaan. Informaation, esimerkiksi verkkoturvallisuuden tilannekuvan, jaossa kuitenkin toisen kohdan arkaluontoisuus voi olla keskeinen este tiedon jakamiselle. Perinteisesti tietoturvapuutteet, -poikkeamat ja haavoittuvuudet on haluttu pitää salassa organisaation maineen ja järjestelmien suojaamisen takia. On helpompaa jakaa arkaluontoista tai mahdollisesti kiusalliseksi koettua tietoa, jos organisaatioitakaan ei voida tunnistaa. Sensorien ja hunaja-ansojen sijainnit verkossa halutaan pitää salassa, jotta hyökkääjät eivät voi kiertää niitä.

Yksi keskeinen työkalu tietosuojan parantamiseksi on jaettavien tietojen anonymisointi eli tiedon muokkaaminen siten, että yksilöitä tai organisaatioita ei voida tunnistaa. Monissa tapauksissa kyse on tarkkaan ottaen pseudonymisoinnista, koska anonymisoidut tunnisteet ovat usein palautettavissa yksilöidyiksi tunnisteiksi. Jäljempänä käytetään termiä **"anonymisointi"** yleisterminä ja termiä **"pseudonymisointi"** vain silloin kun erikseen korostetaan näiden kahden eroa.

Liian tiukka tiedon vaihdon rajoittaminen tai liian voimakas anonymisointi vähentää tai poistaa kokonaan tiedonvaihdon arvon. Liian löysä politiikka taas vaarantaa yksityisyyden ja luo uusia tietoturvariskejä. Yleisesti voidaan todeta, että ei ole yhtä, kaikkiin tapauksiin sovellettavaa anonymisointimenetelmää vaan menetelmiä on tarkasteltava tapauskohtaisesti.

Tässä raportissa käydään läpi käytäntöjä, politiikkoja ja teknisiä ratkaisuja, joiden avuilla saadaan käyttökelpoinen, hyväksyttävä ja lainmukainen kompromissi keskenään ristiriitaisten tavoitteiden ja vaatimusten kesken.

Mahdollistaa mahdollisimman tehokas tiedonvaihto uhista niin pienellä riskillä tietovuodosta kuin mahdollista.

Tämän raportin seuraavaan kappaleeseen on tiivistelmäksi koottu raportin keskeiset aiheet ja havainnot yleiskuvan saamiseksi. Luvussa kaksi käsitellään tiedon jakamisen hyötyjä ja tiedon jakamisesta syntyviä uhkia. Seuraavassa luvussa tarkastellaan yleisesti tiedon anonymisoinnin käsitteitä ja teoreettisia viitekehyksiä. Neljännessä luvussa käydään läpi erilaisia anonymisointimenetelmiä, verkkodatasta löytyviä tunnisteita ja niiden käsittelyä. Viidennessä luvussa esitellään keskeisempiä tieteellisessä kirjallisuudessa esitettyjä työkaluja, joista osa on edelleen aktiivisesti ylläpidettyjä ja sovellettavissa käytäntöön. Kuudennessa luvussa esitetään esimerkkejä ratkaisumalleista, joiden avaulla käytännöllistä anonymisointia voitaisiin toteuttaa niin, että tieto olisi turvassa ja suojattu. Loppupäätelmien jälkeen on kirjallisuusluettelo raportissa viitattuihin artikkeleihin.

1.1 Raportin keskeiset havainnot ja suositukset

Anonymisoinnilla tarkoitetaan tietoaineiston muokkaamista siten, että siitä ei voida tunnistaa henkilöitä tai organisaatioita. Anonymisointi tapahtuu kahdessa vaiheessa: ensin *poistetaan yksilöivät* tunnisteet (esim. nimet, sähköpostiosoitteet) ja tämän jälkeen muuta *tietoa muokataan*, jotta sen perusteella ei voida tunnistaa henkilöitä. Anonymisointi on yksi tärkeä osa tietosuojan työkalupakkia tunnistetietojen minimoinnin toteuttamiseksi.

1.1.1 Verkkodatan anonymisointi

Erityisesti verkkodatan (IP-pakettien, vuotiedon) ongelma on, että yhdestä käyttäjästä syntyy lyhyessä ajassa tuhansia tietueita, joiden perusteella voidaan hyökätä anonymiteettiä vastaan tehokkaasti. Erityisen helposti anonymiteetti rikkoutuu, jos hyökkääjä arvaa, että tiedot julkistetaan ja voi lähettää räätälöityä verkkoliikennettä kohdeverkkoon eli toteuttaa ns. injektiohyökkäyksen. Tämä uhka kohdistuu erityisesti säännöllisiin ja jatkuviin julkaisuihin.

Anonymisoinnin hyvyyttä voidaan arvioida kahden mittarin perusteella:

1. Yksityisyys: kuinka hyvin menetelmä suojaa yksityisyyttä.
2. Käyttökelpoisuus: kuinka vähän tiedon arvo heikkenee anonymisoitaessa.

Nämä kaksi tavoitetta eivät ole aina ristiriidassa keskenään sillä jotkut anonymisointimenetelmät kasvattavat yksityisyyden suojaa siten, että tiedon käyttökelpoisuus heikentyy ainoastaan vähän tai ei ollenkaan. Luvussa Tietoalkioiden anonymisointimenetelmät (sivu 16) on tarkemmin kuvattu erilaisia anonymisointimenetelmiä, jotka voidaan ryhmitellä seuraaviin pääluokkiin:

- poistaminen,
- yleistäminen,
- suora korvaus ja

- häiriöiden lisääminen.

Eri menetelmät soveltuvat eri tietoalkioiden anonymisointiin ja tuottavat yllä mainitulla kahdella kriteerillä – yksityisyys ja käyttökelpoisuus – tuloksia. Käyttökelpoisuuden säilymistä voidaan tutkia sopivilla testijärjestelyillä, esimerkiksi tekemällä sama analyysi alkuperäisellä ja analysoidulla datalla. Mikäli analyysi tuottaa samankaltaisen tuloksen, käyttökelpoisuuden voidaan arvioida säilyvän. Yksityisyyden säilymisen arvioiminen ei ole yhtä suoraviivaista koska säännöllisesti tulee uusia tapoja tunnistaa yksittäisiä koneita tai henkilöitä.

Suomen tietosuojavaltuutetun ohjeen mukaan tieto on anonymisoitua jos:

> *Tunnistamisen täytyy estyä peruuttamattomasti ja siten, että rekisterinpitäjä tai muu ulkopuolinen taho ei voi enää hallussaan olevilla tiedoilla muuttaa tietoja takaisin tunnistettaviksi*.

Tämän ohjeen mukaan pitää huomioida kohtuudella käytettävissä olevat keinot eli ainoastaan teoriassa mahdollisia hyökkäyksiä ei tarvitse huomioida. Teleoperaattori pystyy selvittämään asiakkaistaan kenen käytössä tietty IP-osoite on ollut milläkin hetkellä. Useat operaattorit tietävät myös millä tavoin kukin heidän asiakkaistaan käyttää verkkoa. Tämä tieto ei ole kuitenkaan yleisesti saatavissa ja operaattorin, samoin kuin muiden rekisterinpitäjien, on noudatettava itseään koskevia tietosuojasäädöksiä. Näin ei ole selvää onko esimerkiksi teleoperaattori *muu ulkopuolinen taho*, joka hyötyisi kyvystään purkaa anonymisointi.

Tietosuojakeskustelu tunnisteista kohdistuu pääasiassa selainten käyttämiin evästeisiin ja IP-osoitteisiin. IP-osoitteet voidaan jakaa kolmeen luokkaan lukumäärän mukaisessa järjestyksessä:

1. Dynaamiset käyttäjäosoitteet, jotka ovat yhdellä käyttäjän käytössä tyypillisesti muutamasta tunnista päiviin.

2. Palvelinosoitteet.

3. Staattiset käyttäjäosoitteet, jotka voivat olla jopa vuosia samoja samalla käyttäjällä (kiinteä IP-osoite kotiliittymissä, yritysliittymät).

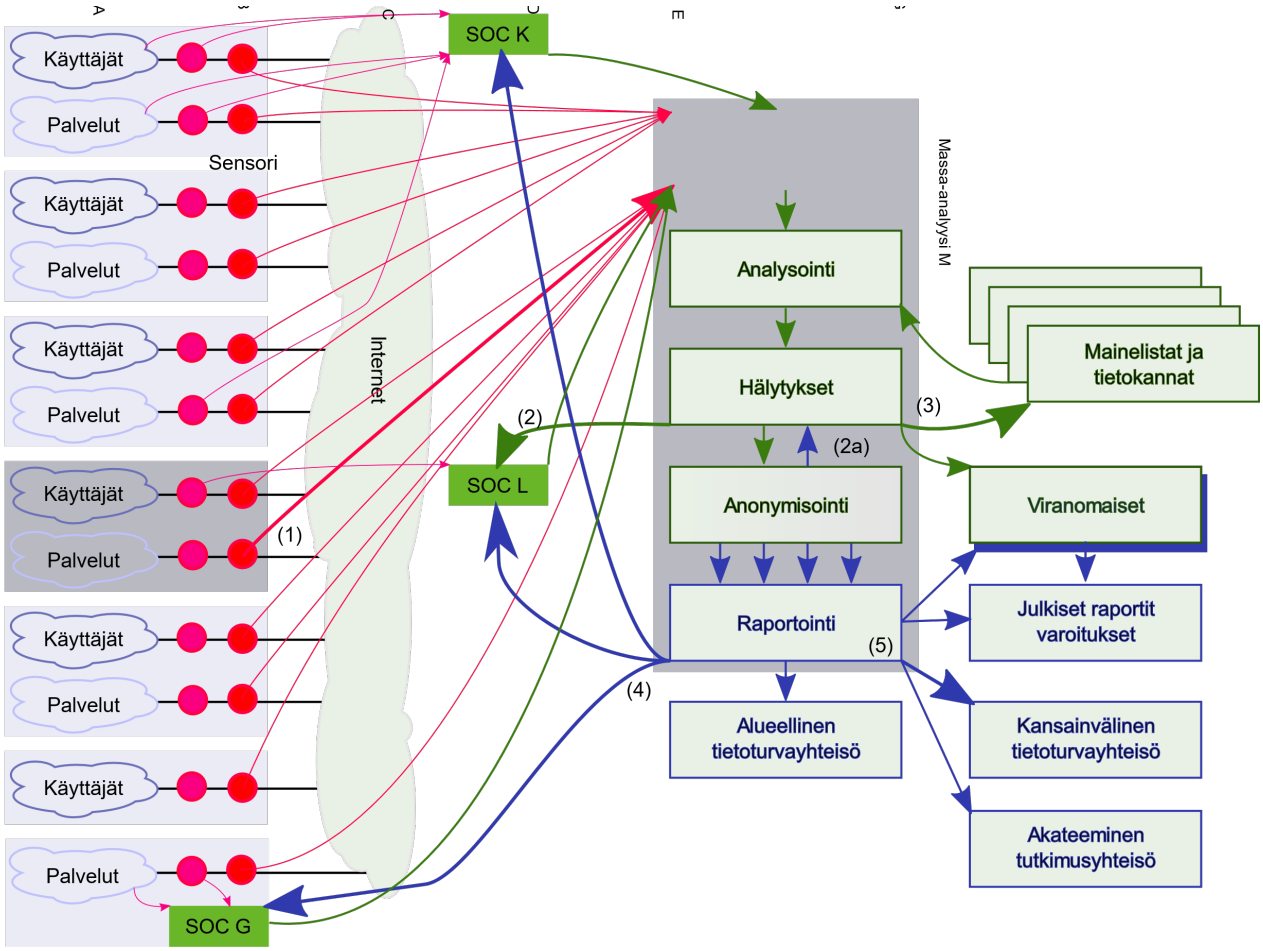
Viimeinen luokka on ongelmallisin, koska käyttäjä voi olla tunnistettavissa vuosikausien ajan, myös jälkikäteen. Myös moni muu tunniste voi paljastaa käyttäjän henkilöllisyyden. Sovellustason tunnisteiden suojausta verkossa on parantanut paljon TLS-salauksen lisääntyminen Internet-liikenteessä. Pitkälle tulevaisuuteen on kuitenkin käytössä protokollia, jotka kuljettavat tunnisteita salaamattomina.

1.1.2 Esimerkki tiedon välittämisestä ja tunnisteiden käsittelystä

Tiedon kulkua ja anonymisointia voidaan havainnollistaa kuvan 1 (sivu 6) mukaisella esimerkillä. Siinä on mukana 7 yritystä tai muuta organisaatiota (A-G). Nämä yritykset ovat asiakkaina jollekin tietoturvavalvomoa (SOC: Security Operations Center) palveluna tarjoajalle yritykselle (K tai L). Suurena yrityksenä *"Globaali G"* toteuttaa SOC-toiminnon yrityksen omana sisäisenä palveluna.

Asiakasyritysten (A-G) verkkoihin on asennettu lisäksi sensori keräämään tietoa verkkoliikenteestä keskitettyyn tietovarastoon, josta sitä analysoidaan. Tietovarastoa operoi tietoturvapalvelua tarjoava *"Massa-analyysi M"*, joka voi olla yritys, yhteisö tai viranomainen. M on myös toimittanut tietoa keräävät sensorit yritysten verkkoon. Yritykset A-G ovat joko asiakassuhteessa tähän M:ään joko suoraan tai tietoturvavalvomoiden kautta.

Erilaisilla punaisen sävyllä kuvataan automaatista tiedonkeruuta, vihreällä tiedon käsittelyä ja sinisellä eri tavoilla anonymisoitua tiedon käsittelyä ja jakamista.



Kuva 1. Tiedon kulku sensoreilta hyödyntämiseen: punainen on automaattisesti tuotettua havaintotietoa, vihreä anonymisoimatonta ja sininen eri tavoilla anonymisoitua tietoa.

Esimerkin tapauksessa keskitetyn tietovaraston M analysoija havaitsee *"Datadiilin"* palvelinverkossa olevalla sensorilla haitallista liikennettä (1). Tästä välitetään tieto *SOC L*:n toimenpiteitä varten (2) sekä soveltuvassa tapauksessa sulkulistoille (3). Edellisissä vaiheissa tietoa ei ole anonymisoitu koska toimien täytyy kohdistua oikeisiin laitteisiin ja palveluihin. Tieto, jolla ei ole merkitystä korjaavien toimien kannalta voidaan anonymisoida (2a). Poistaminen on yksi anonymisointimenetelmistä, joten yksinkertaisesti tarpeeton osa tiedosta jätetään välittämättä.

Tieto poikkeamasta, esimerkiksi liittyen uuteen haavoittuvuuteen, on kiinnostavaa myös muiden SOC-toimijoiden kannalta, mutta heidän ei ole tarpeen tietää tarkkoja tunnistetietoja. Heille voidaan toimittaa anonymisoitu versio ilmoituksesta (4).

Kyseinen uhka on siinä määrin merkittävä, että tietoa siitä halutaan levittää laajasti tietoturvayhteisössä ja julkisesti. Tieto tulee anonymisoida hyvin, koska se tulee olemaan vapaasti kenen tahansa saatavissa. Anonymisointi-kohdan useat nuolet kuvaavat eri anonymisointimenetelmiä, joita käytetään anonymisoinnin vaatimusten mukaan. Tämä tiedon kontrolloitu jakaminen voidaan ajatella rinnasteisena yhteistyöryhmien käyttämälle Traffic Light Protocol (TLP) luokittelulle, jossa jakelua kontrolloidaan värikoodatuilla käsittelyluokilla.

Valitun anonymisointimenetelmän tulee olla harkittu kompromissi eri vaatimusten välillä koska täydellistä anonymisointimenetelmää ei ole olemassa

2 Tietoturvayhteistyö

Yhteistyö on olennainen nykyaikaista tietoturvatyötä ja tiedon jakaminen eri organisaatioiden kesken osa yhteistyötä. Yritykset ja muut organisaatiot hyödyntävät kaupallisten tietoturvaa tarjoavien yritysten palveluja. Nämä samoin kuin yhdistykset, yhteisöt ja viranomaiset jakavat ajantasaista tietoa ajankohtaisista uhista. Tietoturvaongelmiin puututaan valvonnalla, sekä korjaamalla haavoittuvia ohjelmistoja ja järjestelmiä.

Perustellusti voidaan sanoa, että mitä enemmän havaintotietoa pystytään jakamaan "hyvien" kesken, sitä parempi näkemys on erilaisista tietoturvaan kohdistuvista uhista.

2.1 Politiikka ja käytännöt

Suoraviivaisin ratkaisu on löytää menetelmät ja toimintatavat, jossa tietoa jaetaan samalla tavalla kaikille toimijoille. Eri toimijoiden kyky hyödyntää ja toisaalta suojata tietoa voi olla erilainen. Voikin olla järkevää välittää eri toimijoille eri tasoilla anonymisoitua tietoa riippuen keskinäisestä luottamussuhteesta.

Luottamus voidaan käsitellä ehdottomana, jolloin taho on joko täysin luotettu tai ei ollenkaan luotettu. Tämä ei kuitenkaan toimi reaalimaailmassa. Toinen ääripää on pitää luottamusta vaikkapa jatkuvana bayeslaisittain laskettuna arvona, jonka perusteella tehdään päätöksiä tiedon jakamisesta arvioiden sekä tietoa että vastaanottajaa (Vasilomanolakis *ym.*, 2015). Mikäli ihminen arvioi luottamuksen määrän, tällöin rajallinen määrä tasoja on paremmin hallittavissa. Esimerkkinä luottamusta organisaatioiden välillä voidaan luokitella PGP:n Web of Trust:in mukaisella kuudella tasolla avaimen luotettavuudelle:

1. tuntematon,
2. riittämätön tieto,
3. ei koskaan luotettu,
4. marginaalisesti luotettu,
5. täysin luotettu ja
6. ehdottomasti luotettu.

Avoimia mainepohjaisia järjestelmiä vastaan voidaan hyökätä ns. Sybil-hyökkäyksellä mikäli on mahdollista luoda suuri määrä identiteettejä, jotka keskinäisesti lisäävät mainepisteitään (Sirivianos, Kim ja Yang, 2011). Ongelmaa voidaan torjua, jos toimijoiden liittyminen maineverkostoon on kontrolloitua ja perustuu todellisiin olemassa oleviin organisaatioihin tai henkilöihin. Ongelmat voivat esiintyä käytännössä esimerkiksi erilaisissa avoimissa tietoturva- ja roskapostiraportointijärjestelmissä. Näiden hyökkäysten havaitseminen, tunnistaminen ja tiedon siivoaminen virheellisestä tiedosta jälkikäteen on usein paremmin toteutettavissa kuin pyrkiä ennakolta estämään kaikki valheellinen tieto.

Tiedon jakamisessa käytetyt anonymisointialgoritmit olettavat uhkamalliksi tyypillisesti *puolirehellinen* tai *rehellinen, mutta utelias* toimijan, esimerkiksi Nguyen ja Roughan (2013) ja Ricciato ja Burkhart (2011). Tällainen toimija

1. toteuttaa protokollan ja laskutoimitukset oikein, mutta

2. yrittää selvittää identiteettejä joko yksin tai yhdessä muiden kanssa käyttäen siihen enintään realistisen määrän resursseja.

Tiedon suojaamisen järjestelyt toimivat, jos valtaosa osallistujista toimii oikein. Esimerkiksi monenkeskisessä laskennassa tyypillisesti oletetaan, että yli puolet protokollaan osallistuvista ei tee yhteistyötä identiteettien paljastamiseksi. Identiteettien selvittäminen voi olla myös tahatonta: tietoa julkaistaan erehdyksessä esimerkiksi raportissa tai muulla tavoin välitetään kolmannelle osapuolelle. Tiedon anonymisointi suojaa myös oikeutettua tiedon käsittelijää: selväkielisestä tunnisteesta voi tunnistaa ystävän, tutun tai perheenjäsenen mikä taas voi johtaa kiusallisiin tilanteisiin vaikka noudattaisikin ammatillista vaitiolovelvollisuutta.

2.2 Tiedon keräämiseen ja tallentamiseen liittyviä riskejä

Tallennettaessa tietoa joudutaan ottamaan myös kantaa erilaisiin yksityisyyttä vaarantaviin riskeihin. Seuraavassa on eritelty eräitä yksityisyyteen kohdistuvia uhkia (Claffy ja Kenneally, 2010):

- *Julkistaminen*, jossa tietoa annetaan saataviksi raportteina, lokitiedostoina, lokitietoina, sulkulistoina tai muuten.

- *Tahaton tai pahantahtoinen tiedon julkistaminen* voi tapahtua kun tietoa, jonka ei ymmärretä olevan yksityistä tai arkaluontoista, julkistetaan. Tunnettu esimerkki on AOL:n vuonna 2006 julkistama lista verkkohauista. Tietoa voi paljastua myös esimerkiksi tietomurron yhteydessä ja tiedon paljastamisella voidaan kiristää uhria.

- *Painostettu julkistaminen* kun kerättyä tietoa vaaditaan luovutettavaksi oikeuden päätöksellä esimerkiksi epäillyn tekijänoikeusloukkausten perustella. Tämän ongelman välttääkseen moni organisaatio ei alunperinkään tallenna tietoa. Tässä kuitenkin voidaan menettää arvokasta tietoturvallisuus- tai tutkimustietoa.

- *Valtiollinen paljastaminen* vastaa edellistä mutta tiedon vaatijana on valtiollinen toimija. Näiden tapausten läpinäkyvyys tietopyynnön kohteena olevalle henkilölle tai ylipäänsä yhteiskunnalle voi olla heikko.

- *Käyttäjä- tai verkkoprofiilien väärinkäyttö*, jossa toimijoihin voidaan kohdistaa joko hyökkäyksiä tai muuta kohdennettua vaikuttamista, mikäli esimerkiksi henkilön käyttäytyminen tai yrityksen liiketoimintasuunnitelmat paljastuvat.

- *Päätelmien väärinkäyttö* kun luodaan virheellisiä 1. tai 2. tason tunnisteita (sivu 13) käytöksen tai identiteetin perusteella.

- *Uudelleentunnistaminen tai anonymisoinnin purku* jossa yksityisiä henkilöitä tunnistetaan tai toissijaisia tunnisteita käytetään toisten tietolähteiden avulla paljastamaan henkilötietoja. Riittämätön tiedon aggregointi ei suojaa uudelleentunnistamiselta.

2.3 Hajautettu tietoturvatoiminta

Tietoturvaa voidaan toteuttaa tarkastelemalla yhtä pistettä tai sitten laajemmin yhdistelemällä eri pisteistä saatua informaatiota, jota yleensä kutsutaan sensoritiedoksi. Yleinen konsensus on, että jälkimmäinen parantaa hyökkäysten havainnointitodennäköisyyttä. Siitä huolimatta monet organisaatiot arkailevat tietojen jakoa muun muassa yksityisyyteen liittyvien huolien takia. Yrityksellä huolena voi olla myös merkittävien liiketoimintatietojen paljastuminen,

esimerkiksi tiedot suunnitteilla olevista yrityskaupoista, muut liiketoimintaa koskevat tiedot tai järjestelmissä olevien haavoittuvuuksien paljastuminen.

Hajautetun tietoturvan järjestelmän tulee olla (Vasilomanolakis *ym.*, 2015; Meng *ym.*, 2015):

- Skaalautuva suunnitellulle määrälle osallistujia.
- Sietää sekä ulko- että sisäpuolelta tulevia hyökkäyksiä.
- Välttää keskeistä vikapistettä.
- Ei saa vaarantaa yksityisyyttä
- Ei saa aiheuttaa uusia haavoittuvuuksia, esimerkiksi käyttää monitorointia vakoiluun tai oikeudettomaan tietojen hankkimiseen.
- Uuden osapuolen liittyminen ei saa heikentää jo mukana olevien turvallisuutta

Yllä olevan ideaalilistan suhteen todennäköisesti joudutaan tekemään kompromisseja ja tarkastelemaan menetelmiä, joilla turvallisuutta voidaan parantaa.

Tiedonvaihdossa on huomioitava kriteerit millä tietoa välitetään. Mikäli välitettävä tieto on vain sitä, mikä normaalissa IDS-järjestelmässä aiheuttaa hälytyksen, ei järjestelmästä saada täyttä hyötyä. Laajan sensoriverkon etuna on havaita esimerkiksi liikenteen korreloinnin avulla hyökkäyksiä, jotka muuten jäisivät näkymättömiksi havaintorajojen alle.

Tietoa voidaan välittää järjestelmien välillä:

- Tallennettuna tietona verkkoliikenteestä (paketti- tai vuoinformaatiota, ns. raakadataa), jolloin tietoa on mahdollisimman paljon mutta toisaalta tiedon välittäminen ja käsittely vaatii eniten resursseja, koska samanlainen analyysi tehdään useissa eri paikoissa. On myös mahdollista, että ainoastaan tiedon keruu on hajautettu ja analyysi tehdään keskitetysti ja tulokset välitetään takaisin hajautetulle organisaatioille.
- Osittain prosessoituna datana, jota on esimerkiksi suodatettu ja johon on lisätty metadataa. Yksi tiedonvaihtomuoto on IDMEF[1] (Debar, Curry ja Feinstein, 2007).
- Prosessoituna ilmoitusdatana, jossa ei ole välttämättä enää mukana alkuperäistä informaatiota, ainoastaan siitä tehtyjä päätelmiä.

Havaintotiedon lisäksi voidaan välittää myös muuta havaintojen todistetietoa, analysointituloksia sekä myös turvallisuuteen liittyviä päätöksiä kuten liikenteen rajoittamista, vaikka ne olisivatkin paikallisia.

Edellä kuvattun tiedon keruun ohjalta tietoturvatoiminnan järjestelyä voidaan mallintaa IDS-järjestelmän mahdollisilla malleilla (Vasilomanolakis *ym.*, 2015):

1. Keskitetty arkkitehtuuri, tähtimäinen topologia, jossa (syvällisempi) analyysi tapahtuu keskitetysti. Keskipisteen on oltava luotettu taho, mutta se voi olla myös mahdollinen kriittinen vikapiste.

[1] Intrusion Detection Message Exchange Format

2. Hierarkkinen arkkitehtuuri, puumainen topologia, jossa tietoa käsitellään eri portaissa ja tieto jalostuu puun juurta kohti. Juuresta välitetään tietoa tarvittavista toimenpiteistä takaisin keruupisteisiin.

3. Hajautettu arkkitehtuuri, jossa rakenne on suoraan vertaisverkkomainen vapaasti tai organisatorisesti muotoutunut rakenne. Tietoa jalostetaan ja analysoidaan kaikissa solmuissa. Osa solmuista voi olla myös "supersolmuja", joihin tiedon fuusio on keskittynyt.

Nämä mallit vastaavat eri tavoin vaatimuksiin:

- **Skaalautuvuus** suurelle määrälle organisaatioita ja sensoreita.
- **Sietokyky** sekä ulkoisille että sisäpuolisille uhille. Kriittisten vikapisteiden välttäminen.
- **Yksityisyys** edellyttää joko poikkeamatiedon vaihdon rajoittamista tai tämän tiedon anonymisointia. Tieto voi olla herkkää käyttäjien, käyttäjäorganisaatioiden tai palveluntarjoajien kannalta. Myös sensorien sijainnit ja kyvyt ovat suojattavaa tietoa.
- **Vasteaika** havainnosta ilmoitukseen ja korjaaviin toimenpiteisiin.

Tiedon jakamisen arkkitehtuureilla on puolensa. Mikäli on yksi luotettava ja hyvin resussoitu taho, tällöin keskitetty arkkitehtuuri on usein tehokkain valinta. Siinäkin tapauksessa tulee valita tarkasti mitä tietoa välitetään keskitettyyn pisteeseen koska kaikki turha tiedon välittäminen on mahdollinen tietosuojariski.

3 Tiedon anonymisointi

Tiedon käsittely on muuttunut tiedon keräämisen kustannusten pienentyessä. Jos ennen oli tarpeen rajata tiedon keruuta kustannuksien takia – kerätä ainoastaan tietoa mistä on hyötyä – nykyään tulee pohtia mitä tietoja kannattaa kerätä ja tallentaa, jotta niistä ei tule tarpeettomia vastuita (Domingo-Ferrer ja Soria-Comas, 2016). Kaiken mahdollisen tiedon kerääminen voi olla järkevää vain kun ei ennakolta tiedetä mitä tietoa lopulta tarvitaan. Tässä luvussa käsitellään anonymisoinnin käsitettä ja eräitä teoreettisia ja käytännöllisiä menetelmiä.

Tiedon anonymisointi tapahtuu kahdessa vaiheessa:

1. Yksilöivien tunnisteiden poisto.
2. Kvasitunnisteiden peittäminen.

Anonymisointia suunnitellessa keskeinen kysymys on se, milloin tieto on riittävästi anonymisoitu. Kaksi merkittävää tietosuojakehystä lähestyy näitä eri tavoin (Domingo-Ferrer ja Soria-Comas, 2016):

- HIPAA[2] määrittää tiedon olevan anonymisoitua kun joko
  1. riittävän pätevä asiantuntija määrittää riskin olevan hyvin pieni, tai
  2. useita määreitä on poistettu tai yleistetty määrättyyn tasoon.
- GDPR[3] tulkitsee tiedon olevan anonymisoitu kun siitä ei voi tunnistaa osapuolia. GDPR tuntee myös pseudonymisoinnin käsitteen, jossa

[2] Health Insurance Portability and Accountability Act, yhdysvaltalainen terveystietojen käsittelyä säätelevä asetus

[3] General Data Protection Regulation, EU:n yleinen tietosuoja-asetus.

> tunnistetiedot voidaan palauttaa käyttäen muuta tietoa. (Euroopan Unioni, 2016)

GDPR:n 4. artiklan määritelmässä 5 pseudonymisoinnilla tarkoitetaan (korostus kirjoittajan):

> henkilötietojen käsittelemistä siten, että henkilötietoja ei voida enää yhdistää tiettyyn rekisteröityyn käyttämättä lisätietoja, edellyttäen että tällaiset lisätiedot **säilytetään erillään** ja niihin **sovelletaan teknisiä ja organisatorisia toimenpiteitä,** joilla varmistetaan, ettei henkilötietojen yhdistämistä tunnistettuun tai tunnistettavissa olevaan luonnolliseen henkilöön tapahdu.

Määrittely käsittelee perinteistä tietojenkäsittelyn tilannetta, jossa pseudonymisoinnin purkamiseen tarvittava tieto on ainoastaan tiedon kerääjän tai käsittelijän hallinnassa. Pseudonymisoinnin purkamiseen tarvittavaa tietoa voi olla myös muilla tahoilla joko julkisesta, rajoitetuista tai salaisista lähteistä saatavissa. Julkista informaatiota voivat kaikki käyttää hyväksi. Lisäksi on huomioitava mahdollisesti laittomasti tai toisen maan lakeja noudattaen[4] tietoja hankkineet tahot, jotka näiden avulla voivat identifioida aineistossa olevia henkilöitä. Esimerkiksi Euroopan tietosuojaneuvosto (2020) tunnistaa, että erityisesti tietojen vastaanottomaan[5] viranomaisilla voi olla pääsy tietoihin joko muista järjestelmistä tai avoimista lähteissä. Tässä tapauksessa vaaditaan perusteellinen analyysi pseudoanonymisoinnin toimivuudesta huomioiden kaikki a.o. maan viranomaisilla olevat tiedot.

Useimmat tiedon käyttökelpoiseksi jättävät tekniikat voidaan murtaa ilman, että siihen tarvitaan vain tiedon kerääjään hallussa olevaa tietoa. Tieto paljastuu riippumatta siitä kuinka hyvä tiedon kerääjän ja käsittelijän tietoturva on. Luvussa Hyökkäykset anonymisointia vastaan (sivu 27) on käsitelty eräitä anonymisointi- ja pseudonymisointitekniikoita vastaan tapahtuvia hyökkäyksiä.

3.1 Anonymiteetin käsitteitä

Seuraavassa on selitetty lyhyesti anonymiteettiin liittyiä käsitteitä (Bianchi, Bracciale ja Loreti, 2012).

Tietojulkaisujen yhteydessä keskeinen termi on *k-anonymiteetti*: aineistossa on vähintään *k* yksilöä, joilla on tietty yhdistelmä tunnisteita julkaistavassa tiedossa. Yksilöä ei pysty tunnistamaan kuin korkeintaan yhtenä tiettyyn, vähintään *k* yksilön, ryhmään kuuluvana.

Edellistä voidaan tarkentaa joko *k^m-anonymiteetillä* (Gkountouna *ym.*, 2014) tai *(k, l)-anonymiteettillä* (Stokes, 2012), joissa hyökkääjän oletetaan tuntevan korkeintaan *m* (*l*) attribuuttia yksilöstä. Tämän jälkeen on edelleen vähintään *k* yksilön ryhmä pienin tunnistettava ryhmä.

Mikäli anonymiteetin ehdot eivät täyty, voidaan yhdistellä arvoja (*l-diversiteetti*), jotta ryhmäkoko saadaan riittäväksi. Yhdistelyssä voidaan huomioida myös muuttujien jakauma (*t-läheisyys*), jolla voidaan paremmin suojautua tilastolliseen analyysiin perustuvilta hyökkäyksiltä.

Differentiaalinen yksityisyys mittaa kuinka yksilön yksityisyys heikkenee, kun hänen tietoja käytetään tuottamaan tilastotietoja. Termi on noussut esille tilanteessa, jossa aineistoon on mahdollista tehdä vapaasti rakennettuja

[4] Useissa maissa on hyvin puutteellinen tietosuojalainsäädäntö.

[5] Maat, joita Euroopan komissio ei GDPR 45 artiklan mukaan ole tulkinnut tarjoavan riittävää tietosuojaa. Erityisesti on huomioitava maat, joissa ihmisoikeuksien suoja on heikko.

tietokantakyselyjä. Poimimalla sopivat hakuyhdistelmät, jotka kaikki palauttavat vähintään *k* tulosta, voidaan yksittäinen henkilö tunnistaa näiden tulosten yhdistelmästä. Tiedon suojaamiseen voidaan käyttää mekanismeja, joilla tuloksiin lisätään satunnaisuutta mikä vaikeuttaa luottamuksellisen tiedon uudelleen rakentamista toistuvien kyselyiden avulla (Dwork, 2008). Alkuperäisten tietojen sijasta voidaan mahdollisesti julkaista esimerkiksi tilastollisesti syntetisoituja tietoja. Myös syntetisoitu tieto voi vuotaa informaatiota, jos se tehdään huolimattomasti.

Anonymiteetti voidaan jakaa kahteen luokkaan: ehdottomaan eli teoreettiseen ja ehdolliseen eli laskennalliseen anonymiteettiin. Jälkimmäinen olettaa, että hyökkääjällä on käytettävissä rajallisesti sekä laskentakapasiteettia että aikaa ja tuntee taustainformaatiota rajallisesti. Kaikissa tapauksissa pätee, että **anonymiteettiä ei ole ilman moninaisuutta**. (Stokes, 2012)

3.2 Kvasitunnisteet

Tilastotietoa julkistetaan yleensä taulukkomuodossa, jossa yksittäistä tietuetta (riviä) kohtaan on useita attribuutteja. Sarakkeissa oleva tieto on joko julkista tai luottamuksellista. Luottamuksellista tietoa, esimerkiksi henkilöiden nimiä, ei julkisteta. (Gkountouna *ym.*, 2014; Domingo-Ferrer ja Soria-Comas, 2016)

Kvasitunnisteiksi (QI[6]) kutsutaan niitä tietoja, joiden perusteella voidaan yksilöidä tietue esimerkiksi tiettyyn henkilöön tai rajata mahdollisten henkilöiden joukkoa. Tällöin paljastuu luottamuksellista tietoa. Kvasitunnisteet muodostuvat yhdestä tai useammasta attribuutista. Teoreettisen anonymiteetin kannalta kaikkia attribuutteja tulee käsitellä kvasitunnisteattribuutteina.

Kvasitunnisteisiin liittyy edellä mainittu *k*-anonymiteetin käsite: kaikkia QI-yhdistelmiä on vähintään *k* kappaletta.

Päällekkäiset datajulkaisut heikentävät *k*:n tehollista arvoa, joten anonymisointi on tehtävä esimerkiksi kahdessa vaiheessa (Domingo-Ferrer ja Soria-Comas, 2016) mitä on sovellettu esimerkiksi Abt ja Baier (2016) ja Riboni *ym.* (2015). Pseudotunnisteita on uudistettava säännöllisesti, jotta vältetään toistuvista julkaisuista seuraava tunnisteiden korrelaatio (Burkhart ja Schatzmann *ym.*, 2010). Yhden identiteetin paljastuminen heikentää myös niiden muiden alkioiden anonymiteettiä, jotka jakavat samoja kvasitunnisteita: *k*-anonymiteetistä tulee *(k-1)*-anonymiteetti.

Verkkoliikenteestä saatavat tiedot, vuodata tai pakettikaappaus, eroavat tyypillisestä tietotaulukkojulkistuksista, kuten väestö- tai yritystilastoista, siinä, että samaa identiteettiä vastaava pseudotunniste esiintyy tyypillisesti useita kertoja yhdistettynä joukkoon muita identiteettejä.

QI-luokkien rakentamiseen voidaan käyttää luvussa Anonymisointi- ja pseudonymisointitekniikat (sivu 15) mainittuja tapoja.

3.2.1 Ulottuvuuksien kirous

Mikäli QI-määritteitä on vain pieni määrä, on mahdollista löytää sopivat anonymisointimenetelmät, joilla estetään uudelleentunnistaminen. Tilanne muuttuu vaikeaksi mikäli eri luokkia ja ulottuvuuksia on paljon. Esimerkiksi Soininvaara, Oinonen ja Nissinen (2014) tutkivat voisiko maantieteellisiin alueisiin ja toimialoihin perustuvaa tilastotietoa jakaa osin tarkemmalla jaottelulla ilman, että yksittäisten yritysten tietoja paljastuisi. Tulos oli päätelmä, että mitä

[6] Quasi-Identifier

enemmän ominaisuuksia kirjataan sitä karkeampia ja laajempia lokeroiden tulee olla.

Ulottuvuksieen määrään liittyy tapahtumaketjujen ainutlaatuisuus. Esimerkiksi Montjoye *ym.* (2013) tutkivat matkapuhelinten sijaintietojen perusteella kuinka hyvin näistä pystyttiin tunnistamaan henkilöitä. Neljä aikaan sidottua sijaintia matkapuhelinverkon solukoon tarkkuudella riitti tunnistamaan 95 % käyttäjistä.

3.2.2 Ensi- ja toissijaiset tunnisteet

Tunnisteet voidaan jakaa ne ensi- ja toissijaisiin tunnisteihin (Claffy ja Kenneally, 2010).

Ensisijainen tunniste yksilöi henkilön, perheen tai kotitalouden. Näitä ovat esimerkiksi nimi, henkilötunnus, katuosoite, sähköpostiosoite ja eräät biometriset tunnisteet.

Toissijainen tunniste voi olla IP- tai MAC-osoite, syntymäaika, sukupuoli, taloudellinen, terveydellinen tai maantieteellinen tieto. Samaan luokkaan voi kuulua myös käyttäytymistä koskeva tietoa: missä henkilö fyysisesti liikkuu, millä verkkosivuilla vierailee tai mitä sovelluksia käyttää.

GDPR mainitsee IP-osoitteet evästeiden tai radiotaajuustunnisteiden ohella yhtenä protokollien verkkotunnistetietona, jota voidaan käyttää luonnollisten henkilöiden tunnistamiseen ja profilointiin *yhdistettäessä ne yksilöiviin tunnisteisiin ja muihin palvelimelle toimitettuihin tietoihin* (2016, johdanto-osan 30. perustelukappale). Tämä vastaa yllä mainittua toissijaisuuden määritelmää: IP-osoitteen toimiminen henkilötunnisteena edellyttää, että on pääsy samanaikaiseen tietoon, jolla liittäminen tietoon voidaan tehdä. IP-osoite on yhdellä henkilöllä käytössä tyypillisesti vajaasta tunnista useisiin päiviin (Kuva 2, sivu 20).

3.3 Salatut tietokannat

Käytännössä suurin osa salatuista tietokannoista on toteutettu siten, että valitut tauluissa olevat sarakkeet ovat symmetrisesti salattuja ja operaatioiden yhteydessä palvelimelle välitetään avausavain. Tämä suojaa tietokannassa olevaa dataa levossa kun operaatioita ei ole käynnissä. Eräissä tapauksissa tämä voi olla hyödyllinen ominaisuus, mutta usein on yksinkertaisempaa salata tallennusmedia koska järjestelmän pääkäyttäjä pystyy usein selvittämään käytetyn salausavaimen. Tyypillisä tietokannoissa käytettäviä salaustapoja on kuvattu esimerkiksi suositun PostgreSQL tietokannan dokumentaatiossa.

Tieteellisessä kirjallisuudessa salatuilla tietokannoilla tarkoitetaan tietokantoja, joiden sisältö on salattu tietokantaa palveluna tarjoavalta. Yhtenä tavoitteena on, että voidaan hyödyntää esimerkiksi pilvilaskentapalveluita ilman, että palveluntarjoajalla on mahdollisuus vakoilla tallennettua tietoa. Näitä on tutkittu paljon, mutta käytännön toteutuksia ei juuri ole (Dowsley *ym.*, 2017). Niiden vaihtoehtoina on käyttää mm. homomorfisia salauksia tai monenkeskistä laskentaa[7] mutta esimerkiksi avainsanahaku salatusta tekstistä ei ole mahdollista. Käytännön suorituskyky ei myöskään ole riittävä.

Useissa ehdotetuissa tietokannoissa avainjakelu on jätetty tarkastelun ulkopuolelle tai oletetaan, että tietokantaa käyttää ainoastaan yksi asiakas. Useiden avainten käsittely on vaikeaa. Yksikään ratkaisu ei edes teoriassa vastaa kaikkiin tarpeisiin. Tyypillisiä rajoituksia on, että tietueiden lisääminen ja poisto on mahdotonta tai raskasta, tietoa vuotaa sitä päivitettäessä tai haettaessa, mahdollisia avainsanoja on rajallinen määrä tai tieto on tallennettava erikseen

[7] MPC: Multi-Party Computation

jokaiselle halutulle operaatiolle. Yleensä tämä tarkoittaa moninkertaista resurssien käyttöä. Esimerkiksi Dowsley *ym.* (2017) Openstack-virtualisointiympäristössä tehdyssä kokeilussa lukuisten optimointienkin jälkeen levytilan käyttö oli 130-kertainen normaaliin tietokantaan nähden ja loppupäätelmänä oli:

> Valitettavasti huolimatta kaikesta tämän mallin teorian kehitymisestä, malli ei siltikään ole toteuttamiskelpoinen useimpiin tosielämän sovelluksiin.

Ala on aktiivisen akateemisen tutkimuksen kohteena, mutta on epävarmaa milloin on olemassa nykyisiä tietokantoja ominaisuuksiltaan vastaavia järjestelmiä.

3.4 Homomorfiset salaukset ja suojattu laskenta

Homomorfinen salaus mahdollistaa, että puoliluotettu taho voi tehdä tiedolle operaatioita saamatta selville lähtöarvoja tai lopputulosta. Tavoite on sama kuin salattujen tietokantojen tapauksessa. Nämä perustuvat esimerkiksi RSA-salauksen kaltaisen julkisen avaimen algoritmien käyttöön. Helpoimpia operaatioita ovat arvojen yhteenlasku tai kertominen, mitkä voidaan toteuttaa vastaavasti salattujen arvojen kertolaskulla tai potenssiin korottamisella. Salauksessa tarvitaan vastaavia menetelmiä kuin julkiseen avaimeen perustuvissa salauksissa, joten operaatiot ovat väistämättä hitaita (Nguyen ja Roughan, 2013).

Suojattu monikeskeinen laskenta olisi hyödyllinen monissa verkkoihin liittyvässä tiedon vaihdossa. Esimerkkejä voisi olla kokonaisliikenteen määrän laskeminen paljastamatta operaattorikohtaisia tietoja, haittaliikenteen määrän laskenta tai IP-osoitteiden vertailu esimerkiksi IDS-hälytysten korrelointia varten, jossa havainnon tehneet saavat selville ovatko muut toimijat havainneet k.o. osoitteesta poikkeamia. Yksi tapa on jakaa tietoa joistain arvoista siten, että voidaan nähdä kuinka monella on sama havainto ilman, että paljastuu kenellä kaikilla (Huang, Wang ja Borisov, 2005).

Suorituskyvyltä käyttökelpoisempia ovat turvalliseen monikeskeiseen laskentaan[8] perustuvat menetelmät. Shamir's Secret Sharing (SSS) on muuten käyttökelpoinen mutta monet operaatiot vaativat esimerkiksi useita kommunikaatiokierroksia ratkaisua varten, koska tulos näissä on myös salattu. Tuloksena on suuri määrä viestejä ja sitä myötä paljon verkkoliikennettä. Esimerkiksi kahden IPv4 osoitteen yhtäsuuruusvertailu vaatii eräällä algoritmilla 2592 hajautettua kertolaskua, joista jokainen tuottaa m^2 viestiä verkkoon (m on laskentaan osallistujien määrä) (Burkhart ja Strasser *ym.*, 2010). Tämä ei ole riittävän tehokas ollakseen käyttökelpoinen. Toisaalta lopputulokseen riittää ennalta määritetty määrä osallistujia. Kaikkien ei tarvitse osallistua, joten satunnaiset vikatilanteet eivät aiheuta ongelmia.

Näitä menetelmiä saadaan optimoitua rajaamalla operaatioita ja luopumalla joistakin vaatimuksista. Yllä mainitussa IP-osoitteen vertailussa saivat Burkhart ja Strasser *ym.* (2010) vähennettyä tarvittavien operaatioiden määrän 34 kertolaskuun eli 1/76 osaan alkuperäisestä. Kommunikaatioviestien määrä on tässäkin pullonkaula, vaikka kokeellisessa ympäristössä pystyttiinkin saavuttamaan lähes reaaliaikainen toiminta ts. edeltävän viiden minuutin aikana tehtyjen havaintojen vaatimat laskelmat saatiin valmiiksi viidessä minuutissa.

Viestien määrää voidaan vähentää laskemalla salauksessa tarvittavat satunnaisjoukkovektorit etukäteen suuremmissa erissä. Tätä on käytetty Ricciato ja Burkhart (2011) esittämässä GCR-menetelmässä[9], joka pohjautuu SMC:n

[8] SMC: Secure Multiparty Computing, likimain sama kuin MPC.

[9] Globally-Constrained Randomization

yksinkertaistetuun versioon. Haittapuolena on, että mikäli jokin ryhmän jäsenistä ei osallistu laskentaan, ei tulosta voida ratkaista. Tämä asettaa korkean luotettavuusvaatimuksen osallistuville solmuille, koska jokainen on mahdollinen kriittinen vikapiste. Muuten GCR toimii tehokkaasti tukien useita tyypillisesti tarvittavia operaatiota:

- yhteenlasku ja kertolasku,
- ehdollinen laskenta: pelaajat (kyllä/ei) tai tapahtumien määrä (0–x),
- histogrammien muodostaminen ennalta asetetuilla raja-arvoilla yhdellä kerroksella sekä minimi ja maksimiarvojen löytäminen alle $log_2\ m$ kierroksella,
- osa joukko-operaatioista bloom-suotimilla sekä
- anonyymi julkaiseminen (aloha-tyyppisellä uudelleenlähetyksellä) ja aikatauluttaminen.

Prototyyppejä lukuun ottamatta sovelluksia tai järjestelmiä, jotka käyttäisivät tätä (tai muuta) mainittua menetelmää, ei kuitenkaan ole tiettävästi toteutettu.

3.5 Bloom-suotimet

Bloom-suotimet (Bloom, 1970) ovat moniin tarkoituksiin hyviä tilatehokkaita tietorakenteita vastaamaan kysymykseen kuuluuko tietty alkio todennäköisesti joukkoon. Nämä tarjoavat luontaisesti tietosuojaa eli monissa tapauksissa ei voida osoittaa, että tietty alkio kuuluu joukkoon. Tarkat raja-arvot tiedon optimaaliseen suojaukseen riippuvat mahdollisten alkioiden määrästä, bittien määrästä ja montako arvoa on tallennettu. Väärien positiivisten määrä riippuu tietorakenteen täyttöasteesta.

Niissä tapauksissa, joissa mahdollisia arvoja on hyvin rajallinen määrä, esimerkiksi IPv4-osoitteita tai vieläpä sen alijoukko, on mahdollista saada selville, että tietty arvo on todennäköisesti tallennettu (Parekh, Wang ja Stolfo, 2006; Bianchi, Bracciale ja Loreti, 2012). Tietosuojaa voidaan parantaa lisäämällä valikoidusti "valebittejä" kasvattamaan väärien positiivisten esiintymistä alkioille, jotka voisivat muuten olla tunnistettavissa. Seurauksena väärien positiivisten määrä lisääntyy jonkin verran.

Mahdollisia bloom-suodinten käyttömahdollisuuksia ovat esimerkiksi:

- Epäilyttävien IP-osotteiden lista: esimerkiksi tiettynä aikaikkunana havaittujen hyökkäysten lähdeosoiteet.
- DNS-kyselyiden tarkkailulistat.

Bloom-suodinten yhteydessä voidaan myös käyttää salausta, mutta kuten yleensäkin, avainhallinta muodostuu helposti ongelmaksi. Toinen bloom-suotimiin liittyvä rajoite on niiden inkrementaalinen rakenne ts. tietoalkioita voidaan vain lisätä, ei poistaa. Ratkaisuna voidaan julkistaa kokonaan uusi bloom-suodin määrävälein.

4 Anonymisointi- ja pseudonymisointitekniikat

Anonymisoinnin hyvyyttä voidaan arvioida kahden mittarin perusteella:

1. Yksityisyys: kuinka hyvin menetelmä suojaa yksityisyyttä.
2. Käyttökelpoisuus: kuinka paljon (vähän) tiedon hyöty heikkenee.

Nämä kaksi tekijää eivät ole vakiosumma, vaan jotkut menetelmät kasvattavat yksityisyyden suojaa tiedon analysointitehon juurikaan kärsimättä. Toiset taas eivät paranna yksityisyyttä merkittävästi, mutta tiedot tulevat hyödyttömiksi tarkoitukseensa. Vaikutus riippuu myös siitä, millaista analyysiä tietojen perusteella tehdään. Mittareina on käytetty esimerkiksi liikenteen tilastollisia samankaltaisuutta Kolmogorov-Smirnov testillä (Farah ja Trajković, 2013), liikennetilastoja, ponnahduslautahyökkäyksen analyysiä (Riboni *ym.*, 2015) ja IDS:n havainnointikykyä (Yurcik *ym.*, 2007, 2008; Lakkaraju ja Slagell, 2008) sekä IDS:n koneoppimista (Chew *ym.*, 2019). Näissä verrataan tulosta, joka on saatu anonymisoimattomalla aineistolla, tuloksiin, jotka saadaan kullakin eri tavalla anonymisoidulla aineistolla.

Pelkästään IDS-hälytysten määrä ei kerro kuinka vähän anonymisointi vaikuttaa tiedon käyttökelpoisuuteen. Näitä on katsottava tarkemmin kolmen ryhmän määrien suhteessa alkuperäisellä ja anonymisoidulla datalla (Lakkaraju ja Slagell, 2008):

- **Oikeat positiiviset:** hälytykset, jotka ovat samat molemmissa ryhmissä.
- **Väärät positiiviset:** hälytykset, jotka ovat seurausta anonymisoinnista.
- **Väärät negatiiviset:** tapahtumat, jotka jäävät havaitsematta hälytyksinä anonymisoinnin takia.

Väärien hälytysten osuuden tulee olla mahdollisimman pieni. Käyttötarkoituksesta riippuuu ovatko väärät positiiviset (turhien hälytysten määrä kasvaa) vai negatiiviset (havaintoja jää huomaamatta) haitallisempia.

Sisäänrakennettu ja oletusarvoinen tietosuoja on EU:n tietosuoja-asetuksen (Euroopan Unioni, 2016) 25. artiklassa mainittu periaate. Siinä mainitaan pseudonymisointi yhtenä teknisenä ja organisatorisena menetelmänä, jolla tietoja voidaan suojata. Toinen tietosuojaperiaate on tietojen minimointi: tarkastellaan kriittisesti tallennettavan tiedon määrää sekä tallennusaikaa.

4.1 Tietoalkioiden anonymisointimenetelmät

Anonymisointi- ja pseudonymisointimentelmiä voidaan luokitella eri tavoilla. Alla on koottu yhdistelmä eri julkaisuissa (Boschi ja Trammell, 2011; Farah ja Trajković, 2013; Gkountouna *ym.*, 2014; Muralidhar ja Domingo-Ferrer, 2016; Lin *ym.*, 2016) esitetyjä menetelmiä. Näiden käyttöä kuvataan tarkemmin eri tietoalkioita koskevissa kappaleissa. Käytettävä anonymisointi voi olla yhdistelmä erilaisista menetelmistä. Esimerkkejä alla olevista menetelmistä on liitteessä Esimerkkejä tunnisteiden anonymisoinnista (sivu 41).

- Poistaminen
    - Päällekirjoitus/pyyhkiminen: tieto korvataan tyhjällä arvolla. Ns. musta tussi, voi olla myös muu arvo kuin tyhjä tai nolla, esimerkiksi tietueen oletus- tai tyyppiarvo.
    - Suodatus: tietoalkioita jätetään kokonaan pois.
- Yleistäminen
    - Katkaisu: esimerkiksi vähiten merkitsevä tavu osoitteesta tai sekunnin murto-osat aikaleimasta nollataan.
    - Käänteinen katkaisu: eniten merkitsevä arvo poistetaan.
    - Tarkkuuden heikentäminen, pyöristäminen: vastaava kuin katkaisu.

    - Ryhmittely: useita arvoja asetetaan samaan arvoon ryhmän sisällä; voidaan varmistaa, että vähintään *k* alkiota saa saman arvon.
    - Mikroyhdistely: arvo korvataan *k* lähimmän arvon keskiarvolla.

- Suora korvaus
    - Sekoittaminen: vaihdetaan tai korvataan arvoja.
    - Etuliitteen säilyttävä anonymisointi: alunperin lähellä toisiaan olevat arvot ovat lähellä toisiaan myös anonymisoituna.
    - Rakenteellinen anonymisointi: arvot sotketaan tietyn arvoluokan tai ryhmän sisällä.
    - Luettelointi: arvot korvataan luettelon mukaisilla arvoilla, esimerkiksi aiemmin esiintymätön arvo korvataan seuraavalla vapaalla arvolla.
    - Luettelointi järjestys säilyttäen: arvojen keskinäinen suuruusjärjestys säilyy.
    - Satunnaiset siirrot: arvoihin lisätään satunnainen vakio; paikka siirtyy mutta keskinäiset suhteet säilyvät.
    - Tuhoaminen: aikaleimasta poistetaan esimerkiksi päivä- ja kuukausiosat eli aikaleima `2018-12-06T15:45:00` muuttuu arvoksi `2018-01-01T15:45:00`. Saman tyyppinen kuin satunnainen siirto.
    - Pituuden säilyttävä anonymisointi: tarpeen esimerkiksi esitettäessä IP-osoite tekstimuodossa. Arvon tilalle voidaan sijoittaa sopivan mittainen osa esimerkiksi HMAC-tiivisteestä.

- Häiriöiden lisääminen
    - Riippumaton summattu kohina: arvoihin lisätään satunnaista kohinaa.
    - Korreloitu summattu kohina: kohinan määrä riippuu alkuperäisestä arvosta.
    - Kerrottu kohina: arvo kerrotaan kohinalla.
    - Sijoitusvaihto: arvot järjestetään suuruusjärjestykseen ja arvo vaihdetaan $\pm p$ askeleen päästä. Askeleen suuruus voi riippua myös arvojen jakaumasta.

Osa menetelmistä tuottaa 1:1 suhteen alkuperäisen ja anonymisoidun tiedon välillä. Tällöin käytetään myös termiä pseudonymisointi, mikäli on olemassa menetelmä, jolla alkuperäiset tunnisteet voidaan palauttaa. Ero anonymisoinnin ja pseudonymisoinnin välillä on usein vaikeasti tulkittava. GDPR:n osalta ero on merkittävä tietojen käsittelyssä, mutta vielä ei ole oikeusasteiden käytännön tapauksesta tekemää tulkintaa siitä, mikä on riittävä ja mikä riittämätön anonymisointi.

Suomen tietosuojavaltuutettu määrittelee anonymisoinnin sellaiseksi, että

> Tunnistamisen täytyy estyä peruuttamattomasti ja siten, että rekisterinpitäjä tai muu ulkopuolinen taho ei voi enää hallussaan olevilla tiedoilla muuttaa tietoja takaisin tunnistettaviksi.

Tietosuojavaltuutetun ohjeen mukaan pitää huomioida kohtuudella toteutettavat keinot. Esimerkiksi hyökkääjällä käytössä olevat resurssit tulee huomioida, mutta miten pitää huomioida pääsy erilaisiin tietokantoihin? Esimerkiksi mobiilioperaattori tai pankki voi tietää tiettynä ajanhetkenä IP-osoitetta käyttävän henkilöllisyyden melko luotettavasti. Mahdollisesti myös keskustelufoorumin ylläpitäjä tai sähköpostipalveluntarjoaja voi tietää tähän IP-osoitteeseen liittyvän identiteetin, joka voi olla sidoksissa todelliseen identiteettiin (nimeen, sähköpostiosoitteeseen) tai ei. Onko tämä tietoa, johon hyökkääjällä voidaan olettaa olevan pääsy ja mahdollisuus käyttää sitä väärin? Valtiollisella toimijalla voi olla hyvät mahdollisuudet selvittää yhdistäviä tietoja (European Data Protection Board, 2020).

Sanotaan, että "mitään Internetiin julkaistua ei saa sieltä pois". Anonymisoinnissa on huomioitava myös se, että käytetty anonymisointi voi heikentyä ajan myötä. Esimerkiksi Pang *ym.* (2006) mainitsee TCP:n aikaleimoihin perustuvan hyökkäyksen tulleen tunnetuksi samaan aikaan kun he määrittivät anonymisointipolitiikkaansa. Mikäli politiikka olisi määritelty aikaisemmin, kyseistä ongelmaa ei olisi huomioitu mitenkään. Laskentatehon kasvaminen pitää luonnollisesti myös huomioida.

Osa menetelmistä tekee M:N kohdennuksia. Useampi alkuperäisarvo muutetaan yhdeksi arvoksi (M>N) tai yksi alkuperäisarvo voi muuttua useammaksi kohdearvoksi (M<N). Tässä yhteydessä käytetään termejä *palautettavuus* (Recoverability) ja *numeroitavuus* (Countability) (Boschi ja Trammell, 2011).

Pelkästään häiriöiden lisääminen ei tuo muodollista luottamuksellisuutta ja lisäksi se voi altistaa toisenlaisille hyökkäyksille (Burkhart ja Schatzmann *ym.*, 2010). Anonymisoinnin tulee toimia oikein myös virheellisen datan tapauksessa. Menetelmät yleensä olettavat, että tieto on määrittelyn mukaista, mutta näin ei aina ole. Virheellinen tieto voi olla tarkoituksellista tai vikatilanteen seurausta. Tiedon virheellisen tulkinnan seurauksena anonymisointi voi epäonnistua piilottamaan tunnisteet.

Mikään yksittäinen anonymisointitekniikka ei ole paras kaikkiin olosuhteisiin, vaan samaakin tietoa joudutaan anonymisoimaan eri tavoin riippuen käyttötarkoituksesta (Yurcik *ym.*, 2008). **Täydellistä anonymisointia ei ole olemassa** (Xu *ym.*, 2002).

4.1.1 Tunnisteiden käsittely

Keskeinen kysymys tunnisteiden käsittelyssä on voiko tunnisteen muoto muuttua. Mikäli anonymisoitua tietoa on tarkoitus käyttää samoilla verkkoprotokollia tulkitsevilla työkaluilla kuin anonymisoimatonta tietoa, tällöin arvojen on pysyttävä saman muotoisina jopa semantiikkaa myöten.

Esimerkiksi IP-osoite (versio 4) on binaarimuodossa neljä tavua mutta sen pituus on 7-15 merkkiä esitettynä normaalissa ASCII-muodossa desimaalisena. Mikäli osoite esiintyy sovellusprotokollassa ASCII-muodossa, tämä rajoittaa IP-osoiteen anonymisointia myös binaarisessa muodossa. Myös IP-osoitteen muuttuminen jakelulähetysosoitteeksi (*224.0.0.0 – 239.255.255.255*) tai ns. marsilaiseksi (*240.0.0.0* →) tuo ongelmia, mikäli analysointiohjelma käsittelee näitä eri tavalla kuin normaaleja kohdelähetysosoitteita. Osoitteiden muokkaus edellyttää aineiston läpikäymistä kahteen kertaan. Ensimmäisellä kerralla selvitettään onko mitään rajoitteita tietyn osoitteen anonymisoinnille (Lin *ym.*, 2016). Muut osoitteet voivat rajoittaa miten yksittäinen osoite voidaan anonymisoida.

Yksinkertainen tapa tunnisteen tuottamiseksi on laskea tiivistefunktio arvosta (Lincoln, Porras ja Shmatikov, 2004). Mikäli mahdollisia lähtöarvoja on

kohtuullinen[10] määrä, tämä suojaus on helposti murrettavissa väsytyshyökkäyksellä. Haun estämiseksi voidaan käyttää avainnettua tiivistefunktiota (HMAC) (Krawczyk, Bellare ja Canetti, 1997).

Toinen vaihtoehto on salata arvo menetelmällä, joka tuottaa samalla avaimella saman tuloksen samalle arvolle, esimerkiksi lohkosalain ECB-moodissa[11]. Avaimen myötä tulevat myös avainhallinan haasteet, mikäli halutaan useamman osapuolen salaavan tunnisteita yhtäpitävästi. Yksinkertainen ratkaisu on julkisen avaimen käyttö mutta suorituskyvyn hinnalla.

Tunnisteiden käsittelyyn yllä mainitun binaarisen ja desimaalisen esitystavan huomioinnin lisäksi samoja tai saman näköisiä tunnisteita voi esiintyä eri rooleissa. Sama merkkijono voi olla käyttäjän nimi, tiedoston nimi, DNS-nimi tai WLAN-verkon nimi. Jos nämä anonymisoidaan samassa tunnusavaruudessa, nämä voivat paljastua niiden keskinäisten suhteiden takia (Pang ja Paxson, 2003).

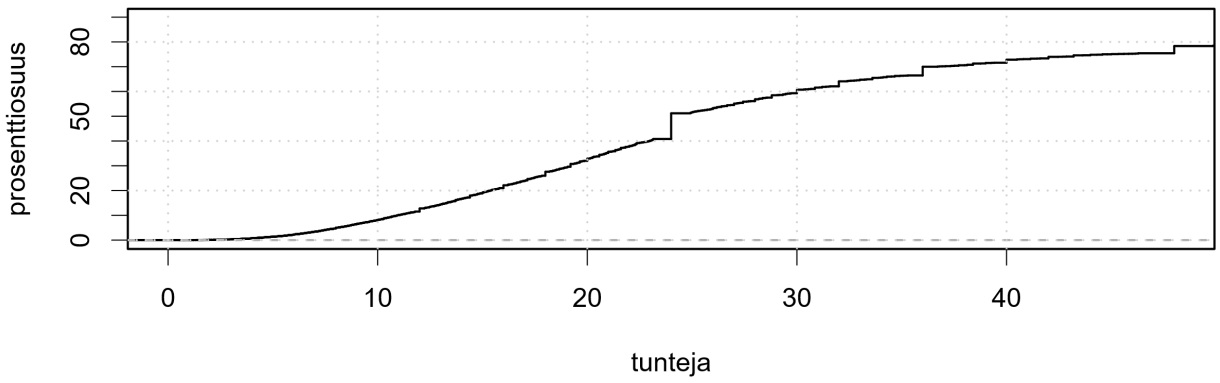
4.2 IP-osoitteet

IP-osoitteet, mahdollisesti evästeiden lisäksi, ovat eniten keskusteltu yksityisyyteen liittyvä tunniste verkossa. Usein tämä rinnastetaan puhelinnumeroon tunnisteena, mutta tosiasiassa suurella osalla verkon käyttäjistä IP-osoite on vain hetkellinen – tunteja tai päiviä hallussa oleva – tunniste. Sen liittäminen henkilöön edellyttää pääsyä esimerkiksi operaattorin AAA-tietokantaan tai johonkin palveluun, jota käyttäjä on käyttänyt ja sitä kautta tunnistettu samoilla ajanhetkillä. Kumulatiivinen jakauma osoitteiden elinajasta mobiiliverkossa on kuvassa 2.

Merkittävällä määrällä käyttäjiä IP-osoite säilyy samana kuukausia tai vuosia ja heidät voidaan liittää sen perusteella vähintäänkin tiettyyn talouteen. Yksittäisestä IP-osoitteesta ilman taustatietoa ei voida tietää kumpaan luokkaan se kuuluu vai onko se esimerkiksi palvelimen osoite, joka ei yksilöi henkilökäyttäjää.

[10] Esimerkiksi käytettäviä IPv4 osoitteita on alle neljä miljardia kappaletta, mikä on kohtuullisilla resursseilla läpi käytävä määrä algoritmista riippumatta.

[11] Electronic CodeBook



Kuva 2. Kumulatiivinen jakauma kuinka pitkään käyttäjällä on ollut sama IP-osoite mobiiliverkoissa vähintään X tuntia. Lähde: Netradar/Jukka Manner, otos (N=53000, t>2 päivää) tammikuu 2019

IP-osoite voidaan jakaa kahteen osaan: organisaation yksilöivään verkko-osaan ja käyttäjän organisaatiossa yksilöivään laiteosaan[12]. Likimääräinen maantieteellinen sijainti voidaan arvioida organisaation perusteella. Usein tietosuojassa ollaan kiinnostuneita pääasiassa yksilönsuojasta, mutta tieto siitä, mihin liikenne yrityksen verkosta suuntautuu, voi paljastaa yrityksen toiminnasta esimerkiksi tietoturvaongelmia, joiden ei haluta paljastuvan.

Yksinkertaisimmillaan IP-osoitteita voidaan anonymisoida korvaamalla osoitteet järjestelmällisesti eli ensimmäisenä havaittu osoite `192.0.2.55` korvataan arvolla `1.0.0.1`, seuraava `203.0.133.107` arvolla `1.0.0.2` ja niin edelleen. Tämä muunnos on uniikki jokaiselle ajolle, ellei muunnostaulukkoa tallenneta. Tässä kadotetaan tieto laitteiden verkkotopologisista suhteista.

Mikäli verkon rakenne halutaan säilyttää, joudutaan käyttämään menetelmää, joka säilyttää osoitteiden suhteet. Peuhkuri (2001) esitteli pakettidatan pakkauksen yhteydessä menetelmän, jolla alimmat 8 bittiä (tai verkkokohtaisesti enemmän) salataan, mutta joka voidaan palauttaa yksikäsitteiseksi anonymisoiduiksi osoitteiksi. Samassa artikkelissa esiteltiin myös injektointihyökkäys, jonka kuvasi tarkemmin Burkhart ja Schatzmann *ym.* (2010) osoittaen sen tehokkuuden. Päätelmänä on, että mikä tahansa IP-osoitteiden suoraan korvaukseen perustuva anonymisointi voidaan ratkaista käyttäen kohtuullisesti aikaa ja resursseja.

Useita algoritmeja, joilla topologiatieto voidaan säilyttää on esitetty. Näistä eniten käytetty ja tunnetuin on Crypto-PAn (Xu *ym.*, 2001), lisäksi esimerkiksi Lakshmanan, Ng ja Ramesh (2005) ja Pang ja Paxson (2003). Topologian säilyttäminen ei tarkoita ylempien bittien säilyttämistä alkuperäisinä vaan sitä, että jos alkuperäisillä osoitteilla on *k* eniten merkitsevää bittiä yhteisiä myös muutetuilla osoitteilla ovat samat *k* bittiä yhteisiä. Menetelmä vaatii paljon laskentaa, mutta sitä voidaan vähentää tallentamalla osatuloksia osoitteita anonymisoitaessa. Tämä luonnollisesti vie taas enemmän muistikapasiteettia.

Osoitteet voidaan salata myös joko kokonaan tai osittain laskemalla osoitteella ja salaisella avaimella HMAC-arvo (Krawczyk, Bellare ja Canetti, 1997), josta otetaan tarvittava määrä bittejä korvaamaan piilotettava arvo. Tässä on pieni

[12] Laiteosa voi pitää sisällään organisaation sisäisen aliverkotuksen eli poikkeaa laitteen itsensä näkemästä jaosta

todennäköisyys törmäyksiin eli tilanteeseen jossa kaksi osoitetta voi tuottaa saman arvon (Yurcik *ym.*, 2008). Anonymisoinnin kannalta mahdollisuus tähän M>N on hyvä ominaisuus.

IP-osoitteita voidaan myös ryhmitellä yhdistämällä laitteet, joilla on samanlainen liikenneprofiili (liikennesormenjälki), samalle "ryhmä-IP:lle". Tavoitteena on saada riittävä määrä yhteyksiä *j* jokaiselle IP-osoitteelle eli *k,j*-anonymiteetti. (Riboni *ym.*, 2015)

IP-osoitteiden anonymisointiin ja niiden julkistamiseen liittyy edellä mainitun injektiohyökkäyksen lisäksi myös monia muita näkökulmia. Pang *ym.* (2006) ryhmitteli organisaation verkosta havaitut IP-osoitteet useampaan luokkaan. Jokainen luokka käsiteltiin erikseen kyseiseen luokkaan parhaiten soveltuvalla tavalla.

- **Ulkoiset osoitteet** anonymisoitiin verkkotopologian säilyttävällä anonymisointimenetelmällä.
- **Sisäiset osoitteet** käsiteltiin erikseen edellisestä anonymisoimalla erikseen aliverkko-osa ja laiteosa: sisäinen topologia piilotettiin ja tuloksesta pystyi saamaan selville, että kaksi laitetta ovat samassa aliverkossa mutta ei aliverkkojen suhteita. Osoitteet muokattiin siten, että ne olivat alueella, joka oli vapaa ulkoisista anonymisoiduista osoitteista.
- **Jakelulähetysosoitteet** tallennettiin sellaisenaan, koska ne eivät yksilöi laitteita ellei ole jotain erityistä sovellusta käytössä.
- **Yksityiset osoitteet** eli osoitteet, jotka eivät reitity Internetissä. Näitä ei tässä tapauksessa anonymisoitu, mutta tunnistettiin, että joissain verkoissa näidenkin anonymisointi samalla tavalla kuin sisäiset osoitteet on tarpeen.
- **Verkkoskannaukset** tuovat ongelmia anonymisoinnille, koska yleensä nämä käyttävät luonnollista järjestystä laitteiden etsimiseen. Eli aliverkossa *192.0.2.0* ensimmäinen skannaus osuu verkko-osoitteeseen (*.0*), seuraavat *192.0.2.1*, *.2*, *.3* ja niin edelleen. Ratkaisuna oli heuristiikan avulla tunnistaa skannaukset – tietty kone ottaa peräkkäin yhteyttä useisiin (tässä tapauksessa vähinään 16:sta) osoitteisiin numerojärjestyksessä. Vaikka tällä tavalla voidaan paljastaa useimmat skannaukset, satunnaista järjestystä käyttävä tai hidas skannaus voi jäädä huomaamatta.

  Liikenne, joka tunnistetaan skannaukseksi, anonymisoidaan erikseen sisäisistä osoitteista. Tämä edellyttää aineiston käymistä läpi kahteen kertaan.
- **Virheelliset osoitteet**, esimerkiksi väärästä verkosta havaittu tai käyttämätön aliverkko, anonymisoidaan myöskin erikseen.

Näiden osoiteryhmien lisäksi verkko-operaattori voi tunnistaa omasta verkostaan eri tarkoituksiin olevat aliverkot: osa verkoista on dynaamisia asiakasverkkoja, osa kiinteitä osoitteita käyttäviä. Tämä tieto ei kuitenkaan ole yhtenäisesti saatavissa kaikilta verkko-operaattoreilta.

4.3 Linkkikerroksen osoitetiedot

MAC-osoitteet muodostuvat kahdesta osasta: 3 tavua laitteen valmistajan organisaatiotunnistetta (pääsääntöisesti) ja loppu laitetunnistetta. Vastaavasti kuin IP-osoitteen yhteydessä, nämä voidaan käsitellä yhtenä anonymisoitavana tunnisteena tai anonymisoida erikseen.

MAC-osoitteen kohdalla sen arkaluontoisuus riippuu havainnointipisteestä. Jos liikenne tallennetaan runkoverkossa kahden reitittimen väliseltä linkiltä, näiden MAC-osoitetta ei yleensä pidetä arkaluontoisena: sehän ei yksilöi kenenkään yksittäisen henkilön käyttämiä laitteita. Laitteen valmistaja, kenties malli ja sarjanumero voidaan päätellä niistä. Jos tallennus tapahtuu esimerkiksi WLAN-verkossa, laitteet voidaan yksilöidä MAC-osoitteen perusteella ja tunnistaa laitteiden valmistajat. Osa käyttöjärjestelmistä käyttää nykyään satunnaista MAC-osoitetta WLAN-liitännöissä käyttäjien seurannan vaikeuttamiseksi. Mikäli DHCP-pyyntöjä hallinnoidaan keskitetysti (käyttäen DHCP-relay toiminnallisuutta) tällöin voi myös organisaation runkoverkossa kulkea MAC-osoitteita.

4.4 Muut otsikkotiedot

Erinäisiä otsikkotietoja voidaan myös tallentaa joko vuotiedossa tai pakettikaappauksissa. Osalla näistä on myös yksityisyysvaikutuksia (Yurcik *ym.*, 2008).

Aikaleimoja voidaan käyttää apuna tunnistamaan tai kuvaamaan laitteita. Tiedon korreloinnissa tarkka ja oikea aika on tarpeen mutta mikäli tietoja julkistetaan esimerkkinä, voi tietoihin lisätä satunnaisen siirtymisen, esimerkiksi tapahtuman ensimmäinen aikaleima voi olla tasan keskiyöllä. Tämän jälkeen voidaan yksittäisten pakettien aikaleimoihin lisätä kohinaa ellei tämä aiheuta ongelmia itse analyysin kannalta. Monissa analyyseissä kohinan lisääminen ei aiheuta ongelmia. On myös mahdollista asetetaa kaikki ajat samaan arvoon tai vakioaikavälein järjestys säilyttäen.

Aikaleimat voivat paljastaa verkon ominaisuuksia. Esimerkiksi kiertoaikaviiveestä voidaan päätellä laitteiden keskinäinen maantieteellinen etäisyys. Mikäli kiertoaikaviive on normaalisti lyhyt (< 20 ms), voidaan viiveen kasvun perusteella päätellä, onko asunnossa parhaillaan verkkoa käyttäviä (Trammell ja Kühlewind, 2018) samalla tavalla kuin sähkön kulutuksesta voidaan päätellä ollaanko kotona vai ei. Metatiedon lisäksi aikaleimoja on myös protokollakentissä[13] ja hyötykuormassa[14]. Aikaleimoja muokattaessa nämä on otettava huomioon.

Paketin pituus (joko metadatasta tai paketin kentistä) voi myös paljastaa tietoa. Esimerkiksi voidaan erottaa, saatiinko nimipalvelukyselyyn vastaus vai ei. Tai onko kysymys äänipuhelusta vai videopuhelusta. Vuon koko on vastaava suure. Yhteyden suurimman käytetyn pakettikoon perusteella voidaan tehdä arvaus, että on käytetty jotain VPN-yhteyttä, jos pakettien maksimikoko on pienempi kuin 1500 tavua. Yksi tapa on valita esimerkiksi viisi pituusluokkaa (< 64, 64 – 127, 128 – 511, 512 – 1023, > 1024) ja merkitä kunkin paketin pituudeksi kyseisen luokan suurin arvo. Paketin pituuden muokkauksella ei ole merkitystä, jos kuitenkin koko paketin sisältö on tallennettu.

Palveluluokkakenttä (*ToS* tai *DS-tavu*) on yleensä vakioarvossa (0). Eräät sovellukset asettavat sen määrättyyn arvoon vaikkei sitä yleensä hyödynnetä verkossa muuten kuin operaattori-VPN yhteyksien yhteydessä. Näissäkään ei yleensä noudateta sovelluksen asettamaa arvoa. Palveluluokkakentän arvon avulla voi siis tunnistaa joitain sovelluksia mutta sillä ei yleensä ole merkitystä analyysin kannalta. Arvo voidaan yleensä nollata.

Otsikossa oleva *TTL* tai *Hop Count* arvo taas kertoo kuinka monen reitittimen kautta paketti on kulkenut (verkkotopologia) ja tehdä arvaus laitteen käyttöjärjestelmästä. Tällä arvolla on kuitenkin käyttöä esimerkiksi DDoS-hyökkäysten yhteydessä tunnistamaan käytetäänkö vääriä lähettäjätietoja. Arvo

[13] Esimerkiksi NTP ja RTP aikaleimat sekä TCP aikaleimaoptiot

[14] Esimerkiksi sähköpostiviestien ja HTTP:n otsakkeet sisältävät kellonaikoja tekstimuodossa

voidaan joko nollata tai jakaa se esimerkiksi 32:lla ja tallentaa jakojäännös, jolloin eri pakettien välillä säilyy tieto niiden kulkemien polkujen pituuksien eroista.

Vastaavasti IPv4:n *fragmentointikenttää* voidaan hyödyntää käyttöjärjestelmien tai sovellusten tunnistamiseen. Sen vaikutus yksityisyyteen on samankaltainen kuin *paketin pituus*- ja *TTL*-kentillä. IPv6:n *laajennusotsakkeissa* voi olla tunnistavia tietoja, esimerkiksi reitityslaajennuksessa.

Kuljetuskerroksen protokolla kentän arvo on käytännössä jokin arvoista 1, 6, 17 tai 50.[15] Tämän arvon piilottamisella ei juurikaan ole merkitystä, mutta sen arvon tunteminen toisten otsikoiden ja hyötykuorman analysoinnissa on tarpeen. ESP liikenteen perusteella voidaan esimerkiksi tunnistaa VPN-yhdyskäytävät.

TCP- ja UDP-protokollien *porttinumerot* antavat viitteitä käytettävistä sovelluksista ja mitä palveluita tietty laite tarjoaa. Tämä paljastaa mahdollisia haavoittuvuuksia. Vastaavasti liikennettä voidaan yrittää piilottaa käyttämällä "vääriä" porttinumeroita. Laitteessa avoimina olevat portit tai portit, joihin laite ottaa yhteyttä, kertovat mitä palveluita ja sovelluksia käytetään ja luo näin mahdollisesti yksilöivän sormenjäljen. Tämä mahdollistaa laitteen tunnistamisen (Pang *ym.*, 2006). Muokkaus voidaan tehdä rakenteellisena anonymisointina, jolloin käsitellään erikseen porttinumerot 0–1023, 1024–49151 ja 49152–65535. Porttinumeroiden käsittely, etenkin niiden nollaus, tuottaa helposti suuren määrän vääriä positiivisia IDS-hälytyksiä (Lakkaraju ja Slagell, 2008).

TCP-protokollan *sarja- ja kuittausnumeroja* voidaan käyttää tunnistamaan laitteiden käyttöjärjestelmiä ja siirretyn tiedon määrä. *Ikkunakoko*, *TCP:n lipputiedot* ja *TCP:n optiot* voivat myös määrittää käyttöjärjestelmää tai sisältää tietoja kellonajasta.

4.5 Nimipalvelutiedot

Nimipalvelutiedot voivat sisältääö enemmän yksilöivää tietoa kuin IP-osoite mutta osa nimipalvelutiedoista ei taas yksilöi käyttäjää ollenkaan. Mikäli henkilö on vieraillut esimerkiksi sivulla https://yle.fi, se ei Suomessa paljasta käyttäjästä juuri mitään. Ulkomaisessa yrityksessä se voi tunnistaa ainoan suomalaistaustaisen henkilön. Toisaalta, mikäli sähköpostiohjelma on ottanut yhteyttä `smtp.timovirtanen.example` osoitteeseen, saattaa se yksilöidä henkilön yhdeksi 233:sta tai heidän perheenjäsenestään.

DNS-kyselyjen yhteydessä voidaan käyttää esimerkiksi katkaisua (tallennetaan vain 1. ja 2. tason nimet), ryhmittelyä tai näitä yhdessä. Voidaan esimerkiksi määritellä, että vain nimet, joita on hakenut yli *k* laitetta *m* kertaa ajan *t* sisällä, muuten kirjataan vain ylemmän tason nimi, mikäli sitä on haettu riittävästi (Favale *ym.*, 2021). Toisaalta haittaohjelmat tekevät ainoastaan muutamia nimipalvelukyselyjä, mutta ne voidaan tunnistaa hakemiensa nimien perusteella (Fejrskov, Pedersen ja Vasilomanolakis, 2020). Haittaohjelmien käyttämien verkkonimien listaa joudutaan päivittämään säännöllisesti.

DNS-tietueita voidaan tunnistaa myös tietueen elinajan perusteella, joten anonymisoidessa ne voidaan pyöristää vakioarvoihin (esim. 1, 100, 300, 900 sekunttia) (Fejrskov, Pedersen ja Vasilomanolakis, 2020).

Vastaavasti kuin DNS-kyselyissä, myös haettavat URL:t voivat paljastaa tietoa kuten myös sähköpostiosoitteet ja muut viestitunnisteet. Näistä tarkemmin seuraavassa luvussa.

[15] ICMP, TCP, UDP tai IPsec ESP.

4.6 Tunnisteet sovellus- ja käyttäjädatassa

Sovellusprotokollissa tai hyötykuormassa voi olla tunnisteita tai muuta arkaluontoista sisältöä kuten sähköpostiosoitteita ja nimiä. URI:t voivat myös sisältää käyttäjää tunnistavaa tietoa kuten myöskin HTTP:n evästeet sekä muut protokollakentät. Varmenteista voidaan tunnistaa mihin palvelimiin on otettu yhteyttä, vaikka niissä usein onkin useita palvelinnimiä määritettynä.

Nämä tunnisteet voidaan tallentaa esimerkiksi korvaamalla tunniste avainnetulla tiivisteellä (HMAC), josta otetaan sopiva osa. Esimerkiksi merkkijono `/js/jquery-3.3.1.min.js` voitaisiin korvata saman mittaisella merkkijonolla `/92880635170456b8c3c.js`, jossa on myös tiedoston tyyppiä kuvaava pääte. Merkkijonoon on valittu tarvittava määrä HMAC-arvon heksadesimaalisia merkkejä. Tarkoituksesta riippuu halutaanko koneen nimi tai sähköpostin domain-osa käsitellä erikseen vai käsitelläänkö koko tunniste yhtenä tunnisteena. Edellinen säilyttää enemmän informaatiota sallien esimerkiksi domaineihin liittyvää analyysiä. Anonymiteetti on vastaavasti pienempi.

Valtaosa Internet-liikenteestä on nykyään salattua. Salatun liikenteen sisältöä ei käytännössä pystytä tulkitsemaan. Siitä voidaan kuitenkin hakea esimerkiksi varmenteisiin tai palvelimen nimeen liittyviä tunnisteita. Jälkimmäiseltä voidaan suojautua Encrypted Client Hello (ECH) -menetelmällä, joka on parhaillaan Internet Draft-vaiheessa (Rescorla *ym.*, 2020). Myös liikenteen tyypistä voidaan tehdä päätelmiä TLS-protokollan perusteella vastaavasti kuin sarjanumeroista ja pakettien koosta.

Eräät haittaohjelmat käyttävät itsepurkautuvaa salausta piilottamaan hyökkäyskoodia yksinkertaiselta merkkijonopohjaiselta havaitsemiselta. Salaus voidaan purkaa emuloimalla suoritusta, jolloin viestissä mahdollisesti olevat tunnistetiedot paljastuvat. (Foukarakis, Antoniades ja Polychronakis, 2009)

Siirrettävän tiedon määrä voi olla myös yksilöivä tieto. Palvelin tai sovellus voidaan yksilöidä siirrettyjen tietoalkioiden koon tai kokojakauman mukaan (Pang *ym.*, 2006). Satunnaisuuden lisääminen tiedostojen kokoon edellyttää vastaavasti TCP-yhteyksien muokkausta.

4.7 Hyötykuorman ja tiedostojen yksilöiminen

Esimerkiksi haittaohjelmien tapauksessa voi olla hyödyllistä jakaa tietoa viestin tai tiedoston sisällöstä. Yksinkertaista on jakaa *koko viesti sellaisenaan*, mutta tämä vie paljon kapasiteettia ja tuo yksityisyysongelmia. Kahta viestiä tai tiedostoa voidaan vertailla suoran yhtäläisyyden lisäksi hakemalla yhtenäisiä jaksoja, jaksojen sarjoja, Levenshtein- tai kosinisamanlaisuutta (Parekh, Wang ja Stolfo, 2006; Dara, Zargar ja Muralidhara, 2018).

Kapasiteetin käytön kannalta tehokkaampaa on laskea *tarkistesumma* (esimerkiksi SHA-256) ja jakaa tämä. Yksikin muuttunut bitti muuttaa kuitenkin kryptografisesti vahvan tarkisteen arvon täysin toiseksi. Vaihtoehtona on käyttää *sumeita tarkistussummia*, jotka pystyvät tunnistamaan lähes identtiset tiedostot, esimerkkeinä `ssdeep` ja Lempel-Ziv Jaccard Distance (LZJD) (Kornblum, 2006; Raff ja Nicholas, 2018).

Yksi tapa jakaa tietoa on jakaa viesti *N-gram* tietueisiin eli määrätyn mittaisiin jaksoihin tai merkkijonoihin. Näiden jakelua voidaan optimoida. Haittapuolena on se, että nämä eivät tarjoa juurikaan yksityisyyden suojan parannusta, koska 5 merkin mittaiseen merkkijonoon voi mahtua vaikka tunnistetieto tai salasana. Sen sijaan käyttämällä bloom-suodinta ilmaisemaan mitä N-grammeja on nähty, parantaa yksityisyyttä merkittävästi. Todennäköisyys, että onnistuneesti arvaa suotimeen tallennetun 5-merkkisen arvon 4096-bittisestä bloom-suotimesta on yksi neljännesmiljardista (Parekh, Wang ja Stolfo, 2006).

Anonymiteetiltään hyvä ja nopea tapa laskea tiedostoa kuvaava arvo on käyttää *tavujen esiintyvyystaulukkoa*, johon kirjataan kullekin tavuarvolle havaintojen määrä. Mikäli käytetään tavuja, saadaan 256 alkion mittainen taulukko. Anonymiteettiä voidaan edelleen parantaa järjestämällä arvot lukumäärän mukaiseen järjestykseen ja jättämällä myös lukumäärät pois. Tuloksena on tällöin *Ziph-merkkijono*[16], jonka koko on 256 tavua (Parekh, Wang ja Stolfo, 2006).

5 Anonymisointityökalut

Kirjallisuudessa on esitetty useita menetelmiä tiedon anonymisointia varten. Edellisessä luvussa käytiin läpi menetelmiä ja seuraavissa kappaleissa esitetään eräitä työkaluja, joista osaa ei ole kehitetty prototyyppiä pidemmälle kun taas osa on yhä aktiivisesti ylläpidettyjä projekteja. Suurin osa keskittyy pakettikaappaustiedostojen (PCAP) anonymisointiin, mutta ratkaisuja on myös loki- ja vuotietueiden (NetFlow) anonymisointiin. Tuoreet melko kattavat katsaukset työkaluista löytyvät esimerkiksi Dijkhuizen ja Ham (2018) sekä CAIDAn verkkosivuilta arkistoitu versio. Seuraavassa on käyty läpi lajiensa ensimmäisiä ja keskeisempiä työkaluja.

Suurin osa työkaluista on UNIX-tyyppisissä järjestelmissä toimivia komentoriviohjelmia. Komentorivipohjaiset työkalut sopivat yleensä hyvin osaksi automatisoitua ja jatkuvaa tietojen käsittelyä. Joukossa on myös käyttöliittymän tarjoavia sovelluksia esimerkiksi Matlabiin toteutettu Anonym (Farah ja Trajković, 2013) ja Windows-sovellus TraceWrangler (”TraceWrangler”, 2018), jotka molemmat tarjoavat erilaisia anonymisointimenetelmiä protokollien eri kentille.

Anonymisointityökalut muodostuvat käytännössä kahdesta komponentista: tiedon tulkitsevasta jäsentäjästä ja tunnistetut tietoalkiot anonymisoivasta osasta. Yksinkertainen anonymisointi tapahtuu suoraan korvaamalla ja poistamalla, mutta edistyneemmissä menetelmissä myös eri protokollatasot ja niiden riippuvuudet huomioidaan.

5.1 Verkko- ja kuljetuskerroksen anonymisointi

Tcpdpriv (Minshall, 2005) oli ensimmäisiä pakettikaappaustiedostojen anonymisointiin tarkoitettuja työkaluja. Ensimmäiset versiot on julkistettu vuonna 1996, jolloin se tuki neljää eri menetelmää IP-osoitteiden anonymisointiin. Sitä oli myös mahdollista käyttää ilman IP-osoitteiden anonymisointia mikäli muita osia haluttiin anonymisoida tai esimerkiksi poistaa käyttäjädataa (TCP- tai UDP-liikenteen hyötykuorma). Ensimmäinen anonymisointimenetelmä (“0”) korvasi IP-osoitteet ensimmäisen esiintymisen mukaisella järjestysnumerolla. Toinen (“1”) anonymisoi ylimmät ja alimmat 16 bittiä erikseen ja kolmas (“2”) vastaavasti kukin tavu erikseen. Neljäs tapa (“50”) toteutti etuliitteen säilyttävän salauksen. Tähän – ja yleensäkin IP-osoitteiden anonymisointiin – kohdistuvaa hyökkäystä spekuloi Ylönen (1996).

Nykyinen versio[17] tcpdpriv:stä käyttää, kuten usea muukin, Crypto-PAn:ia (Xu *ym.*, 2002) yhtenä mahdollisista anonymisointimenetelmistä.

Useimmat työkalut keskittyvät ainoastaan IP-osoitteiden anonymisointiin. Tämän lisäksi eräät työkalut anonymisoivat TCP- ja UDP-porttinumeroita sekä käsittelevät joitakin muita otsikkokenttiä ja poistavat sovellusprotokollat hyötykuormineen. Esimerkiksi (”pktanon”, 2011) tarjoaa vapaasti konfiguroitavan anonymisoinnin kullekin linkki-, verkko- ja kuljetuskerroksen protokollille. Eräissä

[16] Z-string

[17] Tcpdpriv-nimellä on useita eri henkilöiden muokkaamia versioita. Tässä tarkoitetaan Minshallin itsensä julkaisemia versiota.

työkaluissa on tuki joillekin sovellusprotokollille kuten (”tcpanon”, 2009) tukee sähköpostiprotokollia HTTP:n ja FTP:n lisäksi mutta muut sovelluskerroksen protokollat hyötykuormineen poistetaan. Järjestelmien haasteena on rajallinen laajennettavuus sekä uusien anonymisointitapojen, niihin kohdistuvien uhkien, paikallisten vaatimusten ja eri tiedostomuotojen soveltaminen.

Tcpmkpub (Pang *ym.*, 2006) tavoitteena oli tehdä joustava sovellus liikenteen anonymisointiin erilaisten vaatimusten perusteella. Samana vuonna Slagell, Lakkaraju ja Luo (2006) kehittivät modulaarisen FLAIM:n (Framework for Log Anonymization and Information Management), jossa halutut anonymisointitarpeet voidaan toteuttaa yhdistelemällä haluttu politiikka tai uusi protokolla olemassa oleviin tai uusiin moduuleihin ilman, että työkalua tarvitsee tehdä alusta uudelleen. Usein halutaan julkaista esimerkiksi sekä palomuurilokeja että pakettikaappauksia. Nämä täytyy anonymisoida yhtenäisesti. FLAIM tukee mm. PCAP-tiedostoja sekä NetFlow- ja Netfilter-lokitietoja. Kehitystyön taustalla oli aikaisempi työkalu CANINE (Converter and ANonymizer for Investigating Netflow Events) NetFlow-tiedoille (Slagell, Li ja Luo, 2005).

FLAIM:a vastaava työkalu on Anontool (Koukis *ym.*, 2006), joissa mm. BPF-sääntöjen[18] perusteella voidaan valita erilainen anonymisointi. Sääntönä voi olla myös, että hyökkäykseksi tunnistettua liikennettä ei anonymisoida hyökkäyksen lähdeosoitteen osalta, ainoastaan kohdeosoite. SCRUB-tcpdump (Yurcik *ym.*, 2007) tarjoaa useita eri anonymisointimenetelmiä mutta sekään ei esitetyssä muodossaan tukenut sovellusprotokollia.

5.2 Sovellusprotokollien anonymisointi

Protokollien ja hyötykuorman kenttien muokkaus tuottaa usein muutoksia itse hyötykuorman pituuteen. Bro (nykyisin Zeek) IDS-järjestelmän yhteyteen toteutettu anonymisointi- ja muokkausympäristö toimii seuraavasti (Pang ja Paxson, 2003):

1. TCP-yhteyden hyötykuorma tallennetaan.
2. Sovellusprotokolla tulkitaan.
3. Tarvittavat sovellusprotokollan kentät muokataan. Yksi mahdollisuus on korvata varsinainen hyötykuorma tiedolla datan pituudesta, datasta lasketusta tiivisteestä sekä muusta meta-tiedosta, esimerkiksi mediatyypistä. Tällä voidaan säästää merkittävästi levytilaa. Myös Anontool (Koukis *ym.*, 2006) sisältää tämän option.
4. Anonymisointia vaativat kentät anonymisoidaan
5. Tieto tallennetaan anonymisoituihin TCP-segmentteihin ja IP-paketteihin. Näitä muokataan tarpeen mukaan, jotta muokattu tieto on oikean mittainen.

Tämä menetelmä tarjoaa useita mahdollisuuksia optimointiin. Anonymisoinnin haittapuolena on, että esimerkiksi tieto paketeista vääristyy ja ns. evasio-tyyppiset hyökkäykset (Niemi, Levomäki ja Manner, 2012) eivät tallennu oikein. Vaihtoehtona on lisätä tieto protokollan “virheellisyyksistä” tai poikkeavuuksista metatietoihin.

Uusien protokollien lisääminen anonymisointiin on työlästä. Ensin täytyy rakentaa k.o. protokollaa tukeva tulkki, jotta erilaiset anonymisointia vaativat kentät tunnistetaan. Mikäli protokollaa ei osata tulkita, turvallinen ratkaisu on poistaa se kokonaan. Esimerkiksi edellä mainittu FLAIM tukee anonymisointityökalujen

[18] BPF: Berkeley Packet Filter. Yleinen käytetty kuvauskieli verkkosuodatusäännöille.

joukossa harvinaisen suurta määrää protokollien kenttiä, 152 kappaletta (Lakkaraju ja Slagell, 2008). Suuri osa protokollista jää kuitenkin näiden ulkopuolille. Tässä voidaan menettää tietoa, josta olisi apua tietoturvaongelman analysoinnissa.

PCAPAnon (osa PCAPLib-työkalua) tekijät keksivät hyödyntää Wireshark verkkoliikenneanalysaattoriohjelman protokollajäsentäjiä, jotka tuntevat satoja tuhansia kenttiä tuhansien protokollien dekoodaukseen (Lin *ym.*, 2016). Anonymisoitavat kentät valitaan asiantuntijan päättäminä, heuristisesti (Lin ja Lin, 2012) tai hakulauseilla kenttien nimistä. Esimerkiksi `addr|host|user` valitsee kentät, joiden nimen osana on `addr`, `host` tai `user`. Toinen tietojen jatkokäyttöä helpottava ominaisuus on, että tunnisteiden pituudet säilyvät samoina olivat ne sitten binaarisia tai tekstimuotoisia. Näin anonymisoitua tietoa voidaan käsitellä edelleen samoilla tavallisilla työkaluilla kuin alkuperäistä tietoa.

6 Tunnisteiden suojaaminen

Tunnisteiden suojaamiseen voidaan käyttää edellä mainittuja menetelmiä. Näillä on rajoitteensa, joita vastaan voidaan suojautua yhdistelemällä teknisiä, toiminnallisia, organisatoorisia ja sopimuksellisia keinoja.

6.1 Hyökkäykset anonymisointia vastaan

Suurin osa anonymisointitekniikoista ja niiden turvallisuuden anonymisointi perustuu staattisen historiallisen datan, esimerkiksi väestötilastoinnin, suojaamiseen. Näissä ei oleteta aktiivista hyökkäystä, jossa hyökkääjä voi syöttää järjestelmään haluamaansa tietoa (Burkhart ja Schatzmann *ym.*, 2010). Tätä voi verrata vähintään valitun selväkielisen tekstin hyökkäykseen mutta useimmiten hyökkäyksen havaitseminen vaatisi piilokanavan tunnistamista, mikä on erittäin vaikea ongelma.

Oletetaan esimerkiksi, että hyökkääjä haluaisi tunnistaa tietyistä verkoista missä osoitteissa on hälyyttäviä hunaja-ansoja tai haittaliikennettä tunnistavia sensoreita. Hyökkääjällä on myös käytössä tuhansia kaapattuja koneita. Tällöin hyökkääjä voisi pyrkiä lähettämään tunnettua haittaliikennettä tiettyihin IP-osoitteeseen ainoastaan tietyiltä koneilta ja tarkkailla koska osoite päätyy mainelistalle. Hyökkääjän kannalta epätietoisuutta aiheuttaa epätietoisuus reagointinopeudesta, joten ei ole varmaa koska joku osoite on todettu sokeaksi haittaliikenteelle ja voidaan siirtyä seuraavan osoitteen pommittamiseen. Tälläinen hidas analyysi on kustannus hyökkääjälle koska kyseistä kaapattua konetta ei voi käyttää muuhun.

Edellä oli esimerkki lähes reaaliaikaisesta tiedon julkistamisesta missä yksittäiset IP-osoitteet ovat mielenkiintoisia. Toinen tyypillinen tapa julkaista tietoa on verkkoliikennekaappauksen julkaisu, jossa on varmasti ongelmaan liittyvää kiinnostavaa tietoa, siihen mahdollisesti liittyvää tietoa sekä taustaliikennettä. Tämä halutaan tarjota suuremmalle joukolle analysoitavaksi poikkeamien löytämiseksi.

Mikäli hyökkääjällä on arvaus käytetyistä anonymisointimenetelmistä, voi hyökkääjä konstruoida injektoimansa tiedon rakenteen sellaiseksi, jonka pystyy tunnistamaan vielä anonymisoinnin jälkeenkin. Riittävän pitkän ajan kuluessa tehty hyökkäys naamioituna normaaliksi liikenteeksi on käytännössä mahdoton havaita. Anonymisointimenetelmästä riippuu millaista injektointia voidaan tunnistaa mutta usein toimiva menetelmä on lähettää liikennettä tietyn mallin mukaan, jossa kohdeosoitetta ja -porttia sekä paketin pituutta vaihdetaan sopivassa järjestyksessä.

Injektion lisäksi esimerkiksi liikennemääristä tai liikenteen tyypistä voidaan tunnistaa joitakin verkossa olevia koneita. Mikäli IP-osoitteiden anonymisointiin

on käytetty etuliitteen säilyttävää menetelmää, voidaan tunnistaa näiden "naapurustossa" olevia koneita.

Edellä mainittut tavat hyökätä koskevat pitkäaikaista, jatkuvaa julkaisua. Mikäli tietojen julkaisut ovat kertaluontoisia tai lyhyeltä ajalta, mukaan otettavilta tiedoiltaan rajoitettuja sekä tunnisteiden vaihto tapahtuu usein, tietovuotoriski on selvästi pienempi.

6.2 Tunnisteiden käsittely raporteissa

Ongelmallisen koneen IP-osoitetta ei ole järkevää anonymisoida raportoitaessa tietoturvaongelmista verkon tai järjestelmän ylläpitäjälle. Mikäli tämä tieto on anonymisoitu, oikean koneen löytyminen on tarpeettoman vaikeaa. Jos kyseisestä koneesta on hyökätty muihin koneisiin, näiden kohdeosoitteiden anonymisointia voidaan harkita. Anonymisointia puoltaa organisaation havainnointikyvyn ja sensorien piilottaminen, päinvastaista taas ylläpitäjän mahdollisuus löytää tapahtumat omista lokitiedoista.

6.3 Sulkulistat ja anonymisointi

Tieto siitä, että tietty IP-osoite on ollut sulkulistalla tai muulla huonomainesten osoitteiden listalla, voi olla tärkeää analysoinnille. Jos osoite on anonymisoitu, tätä liitosta ei voida tehdä jälkikäteen. Anonymisoituun tietoon kannattaa liittää tieto siitä, millä sulkulistalla osoite oli havaintohetkellä tai pian sen jälkeen.

Yksi mahdollisuus olisi tuottaa sulku- ja mainelistoista anonymisoidut versiot, mutta tämä vuotaisi helposti liikaa informaatiota anonymisoinnista, koska monet listat ovat joko julkisia tai kohtuullisilla sopimuksilla kenen tahansa saatavissa. Lisäksi kaikki listat eivät ole sellaisenaan ladattavissa vaan osoitteiden maine voidaan selvittää vain kysymällä osoite kerrallaan.

6.4 Politiikka ja käytännöt tiedon suojaamisessa

Tekniset keinot tarjoavat suojan parhaimmillaankin vain tiettyyn rajaan asti. On järkevää tarkastella keinoja, joilla teknisten ratkaisujen jäännösriskit voidaan hyväksyttävästi käsitellä. Myös tiedon käytettävyyttä voidaan parantaa mikäli "riittävän hyvä" tiedon anonymisointi heikentäisi tiedon käyttöarvoa liikaa. (Claffy ja Kenneally, 2010)

Tiedon jakamiselle tulee olla *valtuutus*. Mikäli verkon tunnistetietoja käsittelee organisaation ulkopuolinen taho tai näitä jaetaan ryhmittymän sisällä, oikeat toimintatavat on tarpeen käsitellä YT-menettelyllä, mikä edellyttää toiminnan *läpinäkyvyyttä*. Tutkimuslaitoksissa kuten yliopistoissa on usein tutkimusryhmän ulkopuolinen tutkimuseettinen lautakunta, joka voi *valvoa* tiedon jakamista ja julkaisua. Tiedon *luovuttamisessa kolmannelle osapuolelle* tulee mukana seurata vähintään saman tasoiset suojamekanismit.

Tiedon käsittelyssä tulee luonnollisesti noudattaa *asiaankuuluvia lakeja*, tietoa käsittelevät vain *siihen oikeutetut henkilöt* ja käsittelyn tulee olla sovitun *tarkoituksen mukaista*. *Tietojen yhdistelyssä* pitää esimerkiksi huolehtia, että uusia riskejä yksityisyydelle ei synny. Kaksi julkaisua samasta aineistosta voi mahdollisesti paljastaa arkaluonteista tietoa, jos näiden tietoja yhdistellään. Organisaation sisällä samasta aineistosta tehtävät julkaisut on koordinoitava, mutta myös organisaatioiden välillä on huolehdittava, etteivät julkaisut aiheuta tiedon paljastumista.

On arvioitava säännöllisesti *tiedon keruun laajuutta*, jotta tarpeetonta tietoa ei kerätä, mutta toisaalta hyödyllistä tietoa ei jää keräämättä.

Käsittelystä pitää jäädä riittävät *lokitiedot*, jotta tietovuotoepäilyn yhteydessä syyttömyys voidaan selkeästi osoittaa. Osana *riskiarviota* pitää myös ennakkoon selvittää miten mahdollinen tiedon – oikean tai väärän – paljastuminen *korvataan tai hyvitetään*. Väärä tieto tai väärät päätelmät tiedosta voivat aiheuttaa ongelmia Tätä voidaan torjua huolehtimalla sekä *tiedon että analyysin laadusta*.

Tietoturva, käyttäjien *kouluttaminen* ja tapahtumien *kirjaaminen* ovat itsestäänselvyyksiä.

6.5 Yhteenveto

Anonymisointityökaluja löytyy IP- ja TCP-tason informaation suojaamiseen hvvin. Satunnaisen PCAP-tiedoston anonymisointi onnistuu yleensä hyvin, mutta käytössä on huomioitava Anonymisointi- ja pseudonymisointitekniikat luvussa mainitut hyökkäykset (sivu 15).

Tilanne ei ole niin hyvä mikäli halutaan säilyttää osa sovellusprotokollista. Näiden anonymisointiin on vähemmän työkaluja. Yksi sovellusprotokolliin liittyvä ongelma on niiden joustavuus ja monimuotoisuus. IP-otsake on tarkkaan määritetty, koska reitittimien pitää pystyä käsittelemään niitä nopeasti suuria määriä. Sovellusprotokollaan, esimerkiksi HTTP:hen, voidaan sen sijaan lisätä mielivaltaisia otsakkeita ilman, että niitä on mihinkään dokumentoitu. Tällainen epästandardi otsake voi vaikkapa sisältää mobiiliselainkäyttäjän puhelinnumeron. Otsake on lisätty, jotta palveluntarjoajat voivat kohdistaa palveluja haluamilleen käyttäjälle. Yksi vaihtoehto on käsitellä ainoastaan oikeamuotoiset tunnetut otsakkeet ja poistaa tai ylikirjoittaa tuntemattomat ja virheelliset tietoalkiot.

Tietoturvaan liittyen erityisesti hyökkääjä voi käyttää epästandardeja tai rikkinäisiä otsakkeita, arvoja tai muotoiluja. Anonymisointi ei välttämättä osaa käsitellä näitä. Tietojen korjaus ja normalisointi taas voi hukata tietoturvaongelman kannalta tärkeää tietoa.

Tietoa julkistettaessa on hyvä tarkistaa, että tiedostoon ei jää vahingossa anonymisoimatonta tietoa. Esimerkiksi Pang *ym.* (2006) etsi tiedostosta mm. organisaation IP-osoitteita sekä binaari- että tekstimuodossa.

7 Tiedonvaihdon anonymisointi tilannekuvan jakamisessa

Seuraavassa tarkastellaan esimerkkiä tilannekuvatoiminnasta erityisesti tietosuojan näkökulmasta. Johdantoluvussa esitetyn kuvan 1 (sivu 6) mukaisesti tietoa kuljetetaan varsinaisesti käsittelemättömänä, käsiteltynä mutta anonymisoimattomana ja anonymisoituna tietona.

Yleisesti yksityisten tietojen vuotaminen tapahtuu seuraavien ongelmien takia (Silva, Monteiro ja Simões, 2021):

1. tietojen yhdistely,
2. tietojen haravointi,
3. tietomurrot,
4. tietoturvaongelmat ja
5. kolmannen osapuolen tietoturvaongelmat.

Kahdessa ensimmäisessä tapauksessa suojautuminen onnistuu, mikäli tietojen jakelun hallinta ja tarkoitukseen sopiva anonymisointi on valittu oikein. Seuraavissa kahdessa oman tietoturvan taso ja tietoturvakulttuuri ovat ratkaisevat. Viimeisessä kohdassa työkaluina ovat sopimukset ja tarkastukset organisaatioiden välillä.

7.1 Koneoppimisen mahdollisuudet

Tietojen käsittely automaattisilla menetelmillä tarjoaa mahdollisuuden parantaa tietosuojaa. Vaikka tunnistetietojen suhteen myös automaattisessa käsittelyssä on noudatettava asiaankuuluvia tietosuojasäännöksiä, voidaan anonymisoinnilla vähentää riskiä siihen, että analysoija tahattomasti näkee tunnistetietoja. Osa käsittelystä voidaan tehdä hyvin aikaisessa vaiheessa ja olla tallentamatta tietoa, jonka todennäköisesti voidaan arvioida olevan tarpeetonta tietoturvan kannalta.

Mikäli koneoppimisen opetusaineisto on osittain tai kokonaan anonymisoitua, mahdolliset anonymisoinnin aiheuttamat virheellisyydet pitää huomioida itse datan käsittelyssä. Vaikutusta voidaan arvioida tekemällä opetus erikseen vain anonymisoimattomalla aineistolla sekä erikseen aineistolla, joka on anonymisoitu. Tämän jälkeen verrataan näiden kahdella tavalla opetetun kykyä luokitella tietoa kuten yleensäkin anonymisointia arvioidessa (sivu 16).

Anonymisoidun tiedon mahdollisesti aiheuttama vinouma opetusaineistossa tulee käsitellä samoin kuin muukin mahdollinen vinouma, joka voi johtaa virheellisiin luokitteluihin. Joissain tapauksissa anonymisointi voi vähentää vinoumaa yleistämällä opetusaineistoa (Chew *ym.*, 2019).

Federoitu koneoppiminen, missä kukin osallistuja kehittää yhteistä mallia omalla aineistolla, ratkaisisi osin tietojen jakoon liittyvät ongelmat. Tämä ottaa kuitenkin vasta ensimmäisiä askeleitaan. (Silva, Monteiro ja Simões, 2021). Koneoppimisen yhtenä riskinä on luottamuksellisen tiedon vuotaminen kuten on havaittu ongelmaksi suurien kielimallien tapauksessa (Carlini *ym.*, 2021).

7.2 Tilannekuvan tuottaminen

Tietoturva-alan toimijat tuottavat säännöllisiä raportteja, joissa käsitellään ajankohtaisia ongelmia. Tietoturvakenttä ei ole yhtenäinen vaan yleisten uhkien lisäksi eri toimialoihin kohdistuu erilaisia uhkia ja tavat, joilla esimerkiksi kalasteluhyökkäyksiä kohdennetaan, vaihtelevat. Yritysten on myös helpompi arvioida omaa toimintaansa suhteessa oman alan yrityksiin kuin koko organisaatiokenttään.

Yksi ratkaisu anonymisointiongelmaan voi olla syntetisoidun esimerkkidatan tuottaminen. Tällöin aineistossa on kaikki olennainen tieto uhan tunnistamiseksi muttei mitään tunnistetietoja. Synteettisen datan tuottamisessakin pitää olla huolellinen, koska tilastollisen käsittelyn seurauksena siihen voi päätyä oikeita tunnistetietoja kuten yllä mainittiin.

Mikäli tiedon julkistaminen tapahtuu yksittäisiin tapauksiin liittyen esimerkinomaisesti ja sisältää vain kohtuullisen määrän IP-osoitteita, anonymisoidut osoitteet voi valita varattujen verkkojen joukosta (Cotton ja Vegoda, 2010). Osoitteilla ei ole keskinäistä riippuvuutta peräkkäisissä julkaisuissa. Muiden tunnisteiden osalta pitää huolehtia niiden anonymisoinnista.

Monilla aloilla on melko vähän toimijoita, joten organisaatiot voivat olla helposti tunnistettavissa alakohtaisista raporteista, vertaa esimerkiksi Soininvaara, Oinonen ja Nissinen (2014) tekemä tilastotietojen arviointi. Näissä tapauksissa myös organisaation anonymisoinnilla on suuri merkitys.

7.3 Tiedon jakaminen yhteisössä

Tietoa tietoturvaongelmista, -loukkauksista tai niiden epäilyistä välitetään organisaatioiden välissä. Tiedot voivat sisältää IP-osoitteita tai muita tietoja, jotka voivat olla henkilötietoja. Tietoturvatoiminnan voidaan tulkita olevan Euroopan unionin tietosuoja-asetuksen 23. artiklassa mainittua yleisen turvallisuuden sekä rikosten ennalta ehkäisemisen ja paljastamisen mukaista toimintaa, jolloin

tietosuoja ei estä tietojen käsittelyä ja jakamista. Myös muille kuuluvat oikeudet ja vapaudet eli esimerkiksi tietoverkkojen pitäminen turvallisena käyttäjille ovat hyväksyttäviä perusteita tietojen käsittelylle.

Tämän käsittelyn pitää olla kuitenkin suhteessa saatavaan hyötyyn eli jos yllä mainitut tarkoitukset täyttyvät, vaikka tunnistetietoja anonymisoimatta tai poistettaisiin, tällöin ei ole perustetta olla tehdä tätä.

Tietoturvatoiminta on kansainvälistä, joten henkilötietoja mahdollisesti sisältäviä ilmoituksia ja raportteja välitetään myös EU:n ulkopuolisiin maihin. Eräiden maiden, alueiden tai sektoreiden voidaan todeta tarjoavan riittävän tietosuojan (Euroopan Unioni, 2016, s. 45 artikla, 3. kohta), jolloin tietosuoja on samalla tasolla kuin EU:n sisällä. Vuoden 2021 lopussa näitä maita oli 13 (osa rajoituksin).

Jäljelle jää suuri joukko maita, jotka eivät täytä näitä vaatimuksia. Tällöin voidaan toimia 46 artiklan asianmukaisia suojatoimia soveltaen mikäli tiedonvaihto on säännöllistä. Prosessi (1. kohdan e ja f) on kuitenkin hidas ja raskas esimerkiksi tiedon välittämiseen yksittäiselle operaattorille, jonka kanssa kommunikoidaan yksittäisen akuutin tietoturvaongelman ratkaisemisen. Käytäntö soveltuu lähinnä tiiviimmän yhteistyön toteuttamiseen.

Yksittäisissä tapauksissa voitaneen soveltaa 49. artiklan erityistilanteita koskevaa poikkeusta ja määrittää tiedonsiirto tarpeelliseksi tärkeän yleisen edun vuoksi. Tämä ei voi kuitenkaan olla tavanomaista toimintaa vaan sen tulee tapahtua sattumanvaraisesti epäsäännöllisesti kohtuullisin pitkin aikavälein. Tietojen pitää joka tapauksessa olla tarkasti rajotettuja sekä yleisen edun pitää kohdistua myös EU:n kansalaisiin tai organisaatioihin.

7.3.1 Sopimukset ja tutkimustyö

Tiedonjakoa voidaan säädellä useilla eri tasoilla. Osa tiedoista ja tuloksista halutaan jakaa julkisesti esimerkiksi uutisoinnin ja koulutuksen yhteydessä, jolloin mitään sopimusta, käytön rajoittamista tai käyttölupaa ei voida soveltaa. Tällaisessa tiedossa anonymisointia voidaan tehdä varsin voimakkaasti.

Toinen taso on tiedon käyttö tutkimuksen, tuotekehityksen ja opetuksen tarpeisiin. Tällöin voidaan käyttöluvalla tai -säännöillä ohjata tiedon käsittelyä. Esimerkiksi käyttäjä sitoutuu siihen, että ei julkista alkuperäistä tietoa, suojaa sen riittävästi sekä ei pyri uudelleentunnistamaan tai muuten purkamaan anonymisointia. Tiedon tulee olla sopivalla tasolla anonymisoitua, jotta tunnisteita ei paljastu tietoa käsittelevälle vahingossa.

GDPR määrittää 89 artiklassa tietojen käsittelyn periaatteet muun muassa tieteellisiä tutkimustarkoituksia varten. Nämä ovat samoja kuin yleisesti eli tietojen minimointi mm. pseudonymisoimalla.

Tietosuojavaltuutetun kannanoton mukaan rekisteröidyn oikeuksista voidaan tutkimustyössä tapauskohtaisen harkinnan perusteella tarvittaessa poiketa, jos

1. käsittely perustuu asianmukaiseen tutkimussuunnitelmaan;
2. tutkimuksella on vastuuhenkilö tai siitä vastaava ryhmä; ja
3. henkilötietoja käytetään ja luovutetaan vain historiallista tai tieteellistä tutkimusta taikka muuta yhteensopivaa tarkoitusta varten sekä muutoinkin toimitaan niin, että tiettyä henkilöä koskevat tiedot eivät paljastu ulkopuolisille.

Aineiston käsittely mahdollisimman täydellisesti eristetyssä hiekkalaatikossa on yksi tapa välttää aineiston vuotoa. Eräs tapa toteuttaa ratkaisu on seuraava:

1. Tutkija kehittää analyysiohjelman mallidatalla ja dokumentoi sen tuottamat analyysit ja tulosteet.
2. Datan omistaja tarkistaa, että ohjelma ei tuota muuta kuin sovitut tulosteet.
3. Ohjelma ajetaan hiekkalaatikossa datalle.
4. Tulokset tarkistetaan ja ajoympäristö tyhjennetään.
5. Tulokset toimitetaan tutkijalle.

Mallisopimukset voivat olla hyvä työkalu tiedon luovuttamiseen, jossa tietoa luovutetaan toiselle organisaatiolle hyödynnettäväksi, jos ei aivan rutiininomaisesti mutta kuitenkin usein. Näissä tapauksissa luovuttaminen on edelleen yksisuuntaista tai kahdenkeskistä.

Laajemmassa joukossa kahdenkeskiset sopimukset ovat työläitä, joten näissä puitesopimuksen tai verkostosopimuksen tapainen ratkaisu voi olla käyttökelpoisin tapa toimia. Tiedonvaihdon sääntöjä voidaan täsmentää vielä yhteisymmärryspöytäkirjalla.

8 Päätelmät: anonymisointi yhtenä työkaluna

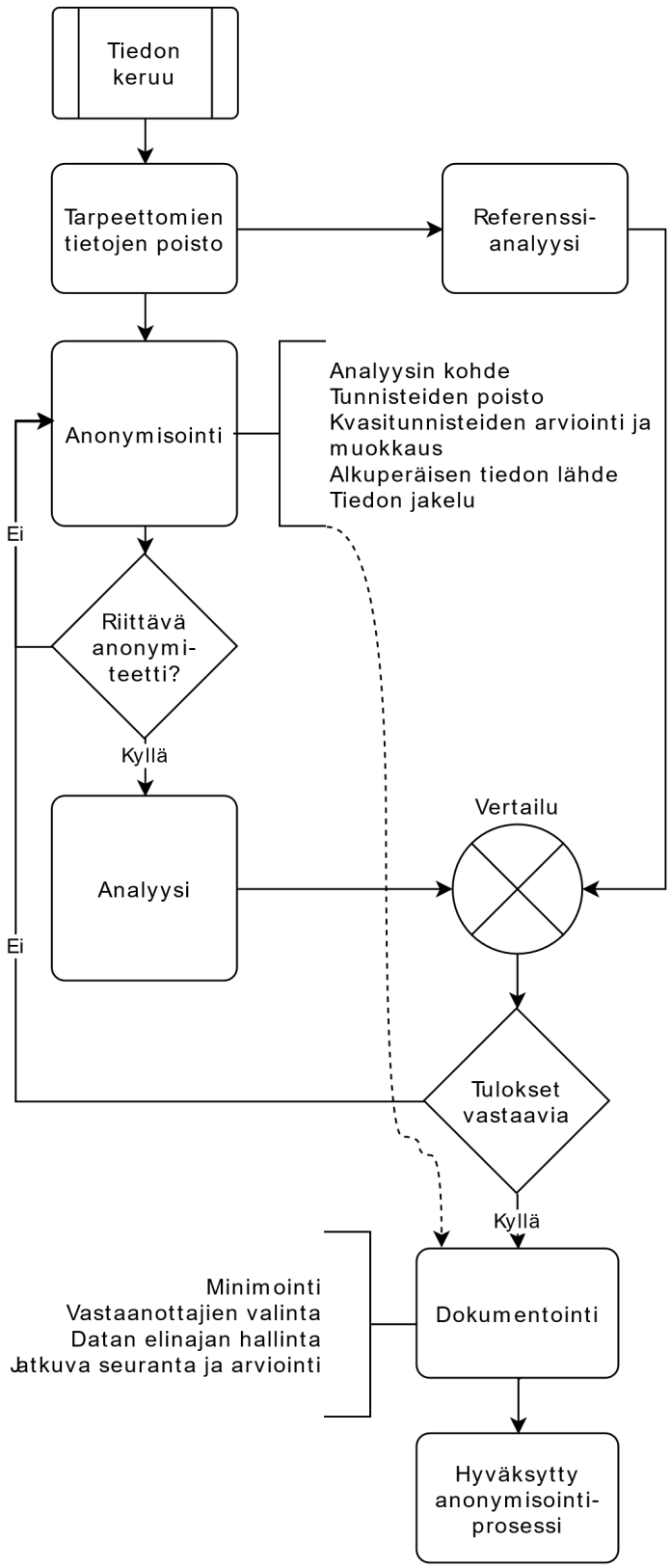
Anonymisointi on keskeinen työkalu parannettaessa tietosuojaa. Se ei kuitenkaan ole yksinkertaisesti toteutettava kaikkeen sopiva *"plug-and-play"* ratkaisu vaan vaatii aina harkintaa tapauskohtaisesti. Aikojen saatossa on ollut useita tapauksia, joissa tieto on kuviteltu anonymisoiduksi, mutta tietoa analysoimalla ja yhdistämällä muuhun tietoon on kyetty tunnistamaan henkilöitä. Tunnettuja tapauksia ovat esimerkiksi AOL hakuhistoria 2006 ja Netflix Prize 2007 reilun vuosikymmenen takaa.

Tietojen minimointi on yksi tärkeimmistä tietosuojaperiaatteista. Monissa tapauksissa tietoja analysoitaessa ei ole tarvetta tietää, käsitellä ja tallentaa yksilöiviä tietoja vaan nämä voidaan joko poistaa tai anonymisoida jo aikaisessa tietojenkäsittelyn vaiheessa. Anonymisointi ja pseudonymisointi tarjoavat suojaa ja rajoittavat vahinkoa mahdollisen tietovuodon sattuessa toteuttaen sisäänrakennetun ja oletusarvoisen tietosuojan periaatetta.

Anonymisointia tulee tarkastella jo suunniteltaessa tietojen keruuta, analysointia ja jakamista. Hyvä käytäntö on anonymisoida tieto ja tämän jälkeen vertailla sille tehtyä analyysiä anonymisoimattomalla tiedolla tehtyyn analyysiin. Mikäli anonymisoitu tieto tuottaa laadultaan samaa tasoa olevia tuloksia, sitä voidaan käyttää. Mikäli käyttökelpoisia tuloksia ei saada anonymisoidulla tiedolla, voidaan sen perustella tehdä päätös olla anonymisoimatta tietoa tässä tapauksessa.

Päätös anonymisoinnista, käytettävästä menetelmästä ja tietojen jakamisesta tulee tehdä harkiten sekä dokumentoida perusteet tehdylle valinnalle. Mahdollisen tietosuojavahingon sattuessa on parempi, että vääräksi osoittautunut päätös on tehty kuitenkin perustellusti ja parhaan silloisen käsityksen mukaisesti.

Anonymisoinnin kehittäminen ja hyvien käytäntöjen luominen edellyttää yhteistyötä niin tietosuojasta vastaavien, tietoa tuottavien että tietoja hyödyntävien kesken. Hyvää tietosuojaa ei voi olla ilman hyvää tietoturvaa.



Kuva 3. Oikean anonymisoinnin valitseminen.

9 Lähdeluettelo

Abt, S. ja Baier, H. (2016) ”Correlating network events and transferring labels in the presence of IP address anonymisation”, teoksessa *2016 12th International Conference on Network and Service Management (CNSM)*, ss. 64–72. doi: 10.1109/CNSM.2016.7818401.

Bianchi, G., Bracciale, L. ja Loreti, P. (2012) ””Better Than Nothing” Privacy with Bloom Filters: To What Extent?”, teoksessa Domingo-Ferrer, J. ja Tinnirello, I. (toim.) *Privacy in Statistical Databases*. Berlin, Heidelberg: Springer Berlin Heidelberg, ss. 348–363.

Bloom, B. H. (1970) ”Space/Time Trade-offs in Hash Coding with Allowable Errors”, *Commun. ACM*. New York, NY, USA: ACM, 13(7), ss. 422–426. doi: 10.1145/362686.362692.

Boschi, E. ja Trammell, B. (2011) ”IP Flow Anonymization Support”. Fremont, CA, USA: RFC Editor; RFC Editor; RFC 6235 (Experimental); RFC Editor (Internet Request for Comments). doi: 10.17487/RFC6235.

Burkhart, M. *ym.* (2010) ”The Role of Network Trace Anonymization Under Attack”, *SIGCOMM Comput. Commun. Rev.* New York, NY, USA: ACM, 40(1), ss. 5–11. doi: 10.1145/1672308.1672310.

Burkhart, M. *ym.* (2010) ”SEPIA: Privacy-preserving Aggregation of Multi-domain Network Events and Statistics”, teoksessa *Proceedings of the 19th USENIX Conference on Security*. Berkeley, CA, USA: USENIX Association (USENIX Security’10), ss. 15–15. Saatavissa: http://dl.acm.org/citation.cfm?id=1929820.1929840.

Carlini, N. *ym.* (2021) ”Extracting Training Data from Large Language Models”. Saatavissa: https://arxiv.org/abs/2012.07805.

Chew, Y. J. *ym.* (2019) ”Privacy Preserving of IP Address through Truncation Method in Network-Based Intrusion Detection System”, teoksessa *Proceedings of the 2019 8th International Conference on Software and Computer Applications*. New York, NY, USA: Association for Computing Machinery (ICSCA ’19), ss. 569–573. doi: 10.1145/3316615.3316626.

Claffy, K. ja Kenneally, E. (2010) ”Dialing Privacy and Utility: A Proposed Data-Sharing Framework to Advance Internet Research2”, *IEEE Security Privacy*, 8(4), ss. 31–39. doi: 10.1109/MSP.2010.57.

Cotton, M. ja Vegoda, L. (2010) ”Special Use IPv4 Addresses”. Fremont, CA, USA: RFC Editor; RFC Editor; RFC 5735 (Best Current Practice); RFC Editor (Internet Request for Comments). doi: 10.17487/RFC5735.

Dara, S., Zargar, S. T. ja Muralidhara, V. (2018) ”Towards privacy preserving threat intelligence”, *Journal of Information Security and Applications*, 38, ss. 28–39. doi: 10.1016/j.jisa.2017.11.006.

Debar, H., Curry, D. ja Feinstein, B. (2007) ”The Intrusion Detection Message Exchange Format (IDMEF)”. Fremont, CA, USA: RFC Editor; RFC Editor; RFC 4765 (Experimental); RFC Editor (Internet Request for Comments). doi: 10.17487/RFC4765.

Dijkhuizen, N. V. ja Ham, J. V. D. (2018) ”A Survey of Network Traffic Anonymisation Techniques and Implementations”, *ACM Comput. Surv.* New York, NY, USA: ACM, 51(3), ss. 52:1–52:27. doi: 10.1145/3182660.

Domingo-Ferrer, J. ja Soria-Comas, J. (2016) ”Anonymization in the Time of Big Data”, teoksessa Domingo-Ferrer, J. ja Pejić-Bach, M. (toim.) *Privacy in Statistical Databases*. Cham: Springer International Publishing, ss. 57–68.

Dowsley, R. *ym.* (2017) ”A survey on design and implementation of protected searchable data in the cloud”, *Computer Science Review*, 26, ss. 17–30. doi: 10.1016/j.cosrev.2017.08.001.

Dwork, C. (2008) ”Differential Privacy: A Survey of Results”, teoksessa Agrawal, M. ym. (toim.) *Theory and Applications of Models of Computation*. Berlin, Heidelberg: Springer Berlin Heidelberg, ss. 1–19.

Euroopan Unioni (2016) ”Euroopan parlamentin ja neuvoston asetus (EU) 2016/679, annettu 27 päivänä huhtikuuta 2016, luonnollisten henkilöiden suojelusta henkilötietojen käsittelyssä sekä näiden tietojen vapaasta liikkuvuudesta ja direktiivin 95/46/EY kumoamisesta (yleinen tietosuoja-asetus)”, *Euroopan unionin virallinen lehti*, L119, ss. 1–88. Saatavissa: https://eur-lex.europa.eu/legal-content/FI/TXT/HTML/?uri=CELEX:32016R0679&from=EN.

European Data Protection Board (2020) *Recommendations 01/2020 on measures that supplement transfer tools to ensure compliance with the EU level of protection of personal data*. Recommendations 01. European Data Protection Board. Saatavissa: https://edpb.europa.eu/our-work-tools/our-documents/recommendations/recommendations-012020-measures-supplement-transfer_en.

Farah, T. ja Trajković, L. (2013) ”Anonym: A tool for anonymization of the Internet traffic”, teoksessa *2013 IEEE International Conference on Cybernetics (CYBCO)*, ss. 261–266. doi: 10.1109/CYBConf.2013.6617434.

Favale, T. *ym.* (2021) ”α-MON: Traffic Anonymizer for Passive Monitoring”, *IEEE Transactions on Network and Service Management*, 18(2), ss. 1233–1245. doi: 10.1109/TNSM.2021.3057927.

Fejrskov, M., Pedersen, J. M. ja Vasilomanolakis, E. (2020) ”Cyber-security research by ISPs: A NetFlow and DNS Anonymization Policy”, teoksessa *2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)*, ss. 1–8. doi: 10.1109/CyberSecurity49315.2020.9138869.

Foukarakis, M., Antoniades, D. ja Polychronakis, M. (2009) ”Deep Packet Anonymization”, teoksessa *Proceedings of the Second European Workshop on System Security*. New York, NY, USA: ACM (EUROSEC ’09), ss. 16–21. doi: 10.1145/1519144.1519147.

Gkountouna, O. *ym.* (2014) ”km-Anonymity for Continuous Data Using Dynamic Hierarchies”, teoksessa Domingo-Ferrer, J. (toim.) *Privacy in Statistical Databases*. Cham: Springer International Publishing, ss. 156–169.

Huang, Q., Wang, H. J. ja Borisov, N. (2005) ”Privacy-Preserving Friends Troubleshooting Network.”, teoksessa *NDSS*.

Kornblum, J. (2006) ”Identifying almost identical files using context triggered piecewise hashing”, *Digital Investigation*, 3, ss. 91–97. doi: 10.1016/j.diin.2006.06.015.

Koukis, D. *ym.* (2006) ”A Generic Anonymization Framework for Network Traffic”, teoksessa *2006 IEEE International Conference on Communications*, ss. 2302–2309. doi: 10.1109/ICC.2006.255113.

Krawczyk, H., Bellare, M. ja Canetti, R. (1997) ”HMAC: Keyed-Hashing for Message Authentication”. Fremont, CA, USA: RFC Editor; RFC Editor; RFC 2104

(Informational); RFC Editor (Internet Request for Comments). doi: 10.17487/RFC2104.

Lakkaraju, K. ja Slagell, A. (2008) ”Evaluating the Utility of Anonymized Network Traces for Intrusion Detection”, teoksessa *Proceedings of the 4th International Conference on Security and Privacy in Communication Netowrks*. New York, NY, USA: ACM (SecureComm ’08), ss. 17:1–17:8. doi: 10.1145/1460877.1460899.

Lakshmanan, L. V. S., Ng, R. T. ja Ramesh, G. (2005) ”To do or not to do: the dilemma of disclosing anonymized data”, teoksessa *SIGMOD ’05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM Press, ss. 61–72. doi: 10.1145/1066157.1066165.

Lin, P. ja Lin, Y.-W. (2012) ”Towards packet anonymization by automatically inferring sensitive application fields”, teoksessa *2012 14th International Conference on Advanced Communication Technology (ICACT)*, ss. 87–92.

Lin, Y. *ym.* (2016) ”PCAPLib: A System of Extracting, Classifying, and Anonymizing Real Packet Traces”, *IEEE Systems Journal*, 10(2), ss. 520–531. doi: 10.1109/JSYST.2014.2301464.

Lincoln, P., Porras, P. A. ja Shmatikov, V. (2004) ”Privacy-Preserving Sharing and Correlation of Security Alerts.”, teoksessa *USENIX Security Symposium*, ss. 239–254.

Meng, G. *ym.* (2015) ”Collaborative Security: A Survey and Taxonomy”, *ACM Comput. Surv.* New York, NY, USA: ACM, 48(1), ss. 1:1–1:42. doi: 10.1145/2785733.

Minshall, G. (2005) ”TCPDPRIV”. on web http://ita.ee.lbl.gov/html/contrib/tcpdpriv.html.

Montjoye, Y.-A. de *ym.* (2013) ”Unique in the Crowd: The privacy bounds of human mobility”, *Scientific Reports*, 3(1), s. 1376. doi: 10.1038/srep01376.

Muralidhar, K. ja Domingo-Ferrer, J. (2016) ”Rank-Based Record Linkage for Re-Identification Risk Assessment”, teoksessa Domingo-Ferrer, J. ja Pejić-Bach, M. (toim.) *Privacy in Statistical Databases*. Cham: Springer International Publishing, ss. 225–236.

Nguyen, H. X. ja Roughan, M. (2013) ”Multi-Observer Privacy-Preserving Hidden Markov Models”, *IEEE Transactions on Signal Processing*, 61(23), ss. 6010–6019. doi: 10.1109/TSP.2013.2282911.

Niemi, O.-P., Levomäki, A. ja Manner, J. (2012) ”Dismantling Intrusion Prevention Systems”, teoksessa *Proceedings of the ACM SIGCOMM 2012 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*. New York, NY, USA: ACM (SIGCOMM ’12), ss. 285–286. doi: 10.1145/2342356.2342412.

Pang, R. *ym.* (2006) ”The Devil and Packet Trace Anonymization”, *SIGCOMM Comput. Commun. Rev.* New York, NY, USA: ACM, 36(1), ss. 29–38. doi: 10.1145/1111322.1111330.

Pang, R. ja Paxson, V. (2003) ”A high-level programming environment for packet trace anonymization and transformation”, teoksessa *SIGCOMM ’03: Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*. New York, NY, USA: ACM Press, ss. 339–351. doi: 10.1145/863955.863994.

Parekh, J. J., Wang, K. ja Stolfo, S. J. (2006) ”Privacy-preserving Payload-based Correlation for Accurate Malicious Traffic Detection”, teoksessa *Proceedings of the 2006 SIGCOMM Workshop on Large-scale Attack Defense*. New York, NY, USA: ACM (LSAD ’06), ss. 99–106. doi: 10.1145/1162666.1162667.

Peuhkuri, M. (2001) ”A Method to Compress and Anonymize Packet Traces”, teoksessa *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*. New York, NY, USA: ACM (IMW ’01), ss. 257–261. doi: 10.1145/505202.505233.

”pktanon” (2011). http://www.tm.uka.de/software/pktanon/.

Raff, E. ja Nicholas, C. (2018) ”Lempel-Ziv Jaccard Distance, an effective alternative to ssdeep and sdhash”, *Digital Investigation*, 24, ss. 34–49. doi: 10.1016/j.diin.2017.12.004.

Rescorla, E. *ym.* (2020) *TLS Encrypted Client Hello*. Internet-Draft draft-ietf-tls-esni-13. Internet Engineering Task Force; Internet Engineering Task Force. Saatavissa: https://datatracker.ietf.org/doc/html/draft-ietf-tls-esni-02.

Riboni, D. *ym.* (2015) ”Obfuscation of Sensitive Data for Incremental Release of Network Flows”, *IEEE/ACM Transactions on Networking*, 23(2), ss. 672–686. doi: 10.1109/TNET.2014.2309011.

Ricciato, F. ja Burkhart, M. (2011) ”Reduce to the Max: A Simple Approach for Massive-Scale Privacy-Preserving Collaborative Network Measurements (Extended Version)”, *CoRR*, abs/1101.5509. Saatavissa: http://arxiv.org/abs/1101.5509.

Silva, P., Monteiro, E. ja Simões, P. (2021) ”Privacy in the Cloud: A Survey of Existing Solutions and Research Challenges”, *IEEE Access*, 9, ss. 10473–10497. doi: 10.1109/ACCESS.2021.3049599.

Sirivianos, M., Kim, K. ja Yang, X. (2011) ”SocialFilter: Introducing social trust to collaborative spam mitigation”, teoksessa *2011 Proceedings IEEE INFOCOM*, ss. 2300–2308. doi: 10.1109/INFCOM.2011.5935047.

Slagell, A. J., Lakkaraju, K. ja Luo, K. (2006) ”FLAIM: A Multi-level Anonymization Framework for Computer and Network Logs.”, teoksessa *Proceedings of the 20th USENIX Large Installation System Administration Conference*, ss. 63–77.

Slagell, A. J., Li, Y. ja Luo, K. (2005) ”Sharing network logs for computer forensics: a new tool for the anonymization of netflow records”, teoksessa *Workshop of the 1st International Conference on Security and Privacy for Emerging Areas in Communication Networks, 2005.*, ss. 37–42. doi: 10.1109/SECCMW.2005.1588293.

Soininvaara, K., Oinonen, T. ja Nissinen, A. (2014) ”Balancing Confidentiality and Usability”, teoksessa Domingo-Ferrer, J. (toim.) *Privacy in Statistical Databases*. Cham: Springer International Publishing, ss. 338–349.

Stokes, K. (2012) ”On Computational Anonymity”, teoksessa Domingo-Ferrer, J. ja Tinnirello, I. (toim.) *Privacy in Statistical Databases*. Berlin, Heidelberg: Springer Berlin Heidelberg, ss. 336–347.

”tcpanon” (2009). http://netweb.ing.unibs.it/~ntw/tools/tcpanon/.

”TraceWrangler” (2018). https://www.tracewrangler.com/.

Trammell, B. ja Kühlewind, M. (2018) ”Revisiting the Privacy Implications of Two-Way Internet Latency Data”, teoksessa Beverly, R., Smaragdakis, G., ja Feldmann, A. (toim.) *Passive and Active Measurement*. Cham: Springer International Publishing, ss. 73–84.

Vasilomanolakis, E. *ym.* (2015) ”Taxonomy and Survey of Collaborative Intrusion Detection”, *ACM Comput. Surv.* New York, NY, USA: ACM, 47(4), ss. 55:1–55:33. doi: 10.1145/2716260.

Xu, J. *ym.* (2002) ”Prefix-preserving IP address anonymization: measurement-based security evaluation and a new cryptography-based scheme”, teoksessa *10th IEEE International Conference on Network Protocols, 2002. Proceedings.*, ss. 280–289. doi: 10.1109/ICNP.2002.1181415.

Xu, J. *ym.* (2001) ”On the Design and Performance of Prefix-preserving IP Traffic Trace Anonymization”, teoksessa *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*. New York, NY, USA: ACM (IMW ’01), ss. 263–266. doi: 10.1145/505202.505234.

Ylönen, T. (1996) ”Thoughts on How to Mount an Attack on tcpdpriv’s ’-A50’ Option…”. on web http://ita.ee.lbl.gov/html/contrib/attack50/attack50.html.

Yurcik, W. *ym.* (2007) ”SCRUB-tcpdump: A multi-level packet anonymizer demonstrating privacy/analysis tradeoffs”, teoksessa *2007 Third International Conference on Security and Privacy in Communications Networks and the Workshops - SecureComm 2007*, ss. 49–56. doi: 10.1109/SECCOM.2007.4550306.

Yurcik, W. *ym.* (2008) ”Privacy/Analysis Tradeoffs in Sharing Anonymized Packet Traces: Single-Field Case”, teoksessa *2008 Third International Conference on Availability, Reliability and Security*, ss. 237–244. doi: 10.1109/ARES.2008.189.

Lyhenteet ja termit

Lyhenne	Termi
AAA	Authentication, Authorization and Accounting. Käyttäjätiedot ja oikeudet sisältävä tietokanta.
AOL	American Online. Internet-palveluntarjoaja.
ASCII	American Standard Code for Information Interchange. Merkistöstandardi, joka sisältää kirjaimet A-Z, numerot, väli- ja kontrollimerkkejä
BPF	Berkeley Packet Filter. Kieli pakettisuotimen määrittelyyn.
DDoS	Distributed Denial of Service. Hajautettu palvelunestohyökkäys.
Crypto-PAn	Etuliitteen yhteneväisyyden säilyttävä IP-osoitteiden anonymisointimenetelmä.
DHCP	Dynamic Host Configuration Protocol. Menetelmä laitteen verkkoasetusten määrittämiseen.
DNS	Domain Name Service. Nimipalvelu, jolla mm. verkkonimet (www.example.com) muutetaan IP-osoitteiksi.
ECB	Electronic Code Book
ECH	Encrypted Client Hello. TLS 1.3 laajennus, joka estää ulkopuolista tarkkailemasta minkä nimiseen palveluun otetaan yhteyttä.
ESP	Encapsulating Security Payload. IP-pakettityyppi, jolla IPsec-salattu liikenne kuljetetaan verkossa
GCR	Globally-Constrained Randomization. Menetelmä monenkeskisessä laskennassa.
GDPR	General Data Protection Regulation. EU:n yleinen tietosuoja-asetus.
HIPAA	Health Insurance Portability and Accountability Act. Yhdysvaltalainen terveystietojen käsittelyä säätelevä asetus.
HMAC	Hash-based Message Authentication Code. Viestitiiviste, jonka laskemiseen käytetään salaista avainta.
HTTP	HyperText Transfer Protocol. Protokolla, jolla web-sivut yleensä siirretään.
ICMP	Internet Control Message Protocol. Vikaraportointiin ja diagnostiikkaan IP-protokollan yhteydessä käytetty protokolla.
IDMEF	Intrusion Detection Message Exchange Format. Tietomuoto, jolla voidaan välittää mm. poikkeamatietoja IDS-järjestelmien välillä.
IDS	Intrusion Detection System. Tunkeilijan havaitsemisjärjestelmä.
IP	Internet Protocol. Protokolla, jonka päällä mm. kaikki Internet-liikenne kulkee. Käytössä kaksi versiota (IPv4 ja IPv6).
IPsec	Internet Protocol Security. Protokollakokoilma, joka mahdollistaa tiedon suojaamisen IP-verkoissa.
LZJD	Lempel-Ziv Jaccard Distance. Algoritmi karakterisoimaan tiedostoja siten, että samankaltaiset tiedostot voidaan tunnistaa.
MAC	Media Access Control. MAC-osoitetta käytetään mm. Ethernet-verkoissa yksilöimään laite. Tyypillisesti jokaisella laitteella on uniikki MAC-osoite mutta voi käyttää myös dynaamisia osoitteita.
MPC	Multi-Party Computation. Moninkeskinen laskenta.
NTP	Network Time Protocol. Käytetään seinäkelloajan välittämiseen IP-verkoissa.
PCAP	Packet Capture. Ohjelmointirajapinta ja tiedostomuoto pakettidatan tallentamiseen.
PGP	Pretty Good Privacy. Sala- ja allekirjoitusstandardi.
QI	Quasi-identifier. Kvasitunniste: ei välttämättä suoraan yksilöi henkilöä mutta voidaan käyttää tunnistamisen apuna.
RSA	Rivest–Shamir–Adleman. Julkisen avaimen salausjärjestelmä.
RTP	Real-Time Protocol. Protokolla esimerkiksi audion, videon tai reaaliaikaliikenteen välittämiseksi IP-verkoissa.
SHA	Secure Hash Algorithms. Joukko kryptografisia tiivistefunktioita.
SMC	Secure Multiparty Computing. Turvallinen moninkeskinen laskenta.
SOC	Security Operations Center. Tietoturvavalvomo.
SSS	Shamir's Secret Sharing. Tapa jakaa salaisuus useaan osaan siten, että tarvitaan vähintään valittu määrä osia salaisuuden paljastamiseksi.

TCP	Transmission Control Protocol. Tiedonsiirtoprotokolla, joka pystyy uudelleenlähettämään mahdollisesti hukkuneet paketit.
TLP	Traffic Light Protocol. Arkaluontoisen tiedon jakamiseen tarkoitettu tiedon luokittelu. Käytössä erityisesti tiedon jakamiseen tietoturvaongelmista.
TLS	Transport Layer Security. Tiedon salaustapa, jota käytetään mm. HTTP-yhteyksien suojaamiseen.
TTL	Time To Live. IP-paketissa ilmoittaa kuinka monen reitittimen kautta tieto voidaan vielä välittää. DNS-tietueissa kertoo kuinka pitkään siinä oleva tieto on vielä voimassa.
UDP	User Datagram Protocol. TCP:n ohella toinen IP-verkoissa käytetty kuljetusprotokolla. UDP ei huolehdi kadonneiden pakettien uudelleenlähetyksestä.
UNIX	Alunperin AT&T:n kehittämä monen käyttäjän käyttöjärjestelmä. Linux on nykyään yleisin UNIX-tyyppisistä käyttöjärjestelmistä.
URI	Uniform Resource Identifier. Yksiselitteinen tunniste Web-palveluissa fyysiselle tai loogiselle resussille, esim. URL.
URL	Uniform Resource Locator. URI:n muoto mikä sisältää verkko-osoitteen ja käytetyn protokollan resussin löytämiseksi.
VPN	Virtual Private Network. Verkkoliikenteen välittäminen turvallisesti tai eriytetysti toisessa verkossa. Yleensä perustuu joko IPsec- tai TLS-suojaukseen.
WLAN	Wireless Local Area Network. Langaton lähiverkko.

Esimerkkejä tunnisteiden anonymisoinnista

Alla oleviin taulukoihin on koottu esimerkkejä erilaista verkossa esiintyviä tunnisteista ja kvasitunnisteista sekä esitetty miten ne muuttuvat eri anonymisointimenetelmissä. Oletetaan, että alkuperäiset esimerkit tulevat tiedostossa vastaan vasemalta oikealle.

Tauukko 1. IP-osoitteen anonymisointiesimerkkejä

Menetelmä				
Ei muutosta	192.0.2.1	192.168.1.8	10.20.30.40	192.0.2.50
Poisto	0.0.0.0	0.0.0.0	0.0.0.0	0.0.0.0.0
Katkaisu	192.0.2.0	192.168.1.0	10.20.30.0	192.0.2.0
Katkaisu (2)	192.0.0.0	192.168.0.0	10.20.0.0	192.0.0.0
K-katkaisu	0.0.0.1	0.0.0.8	0.0.0.40	0.0.0.50
Etuliite	67.140.31.8	67.68.4.243	140.9.185.230	67.140.31.55
Luettelointi	1.0.0.1	1.0.0.2	1.0.0.3	1.0.0.4

Ensimmäinen katkaisu jättää 24 bittiä IP-osoitteesta muuttamatta, toinen katkaisu 16 bittiä. Käänteinen katkaisu (K-katkaisu) vastaavasti nollaa 24 ylintä bittiä.

Taulukko 2. DNS-nimien anononymisointiesimerkkejä

Menetelmä					
Ei muutosta	a.example.com	b.example.com	f.z.fi	l.g.z.fi	c.example.org
Poisto	.	.	.		.
Katkaisu	example.com	example.com	z.fi	z.fi	example.org
Katkaisu (2)	com	com	fi	fi	org
K-katkaisu	a	b	f	l.g	c
Etuliite	k5t.s4b.gmp	fy0.s4b.gmp	nb1.lj0.ih2	od9.q3u.lj0.ih2	p8q.6ay.x9x
Luettelointi	a	b	c	d	f

Ensimäinen katkaisu säilyttää ylimmän tason toimialueen ja sen alapuolella olevan toimialueen, toinen katkaisu ainoastaan päätason. Käytännössä katkaisussa voisi olla hyvä hyödyntää Public Suffix-listaa, jotta normaalista poikkeavat organisaatiorajat otettaisiin huomioon. Esimerkiksi Suomesta *iki.fi* kuuluu tuolle listalle siten, että kolmannen tason nimet ovat eri henkilöiden hallussa.

Taulukko 3. IP-paketin TTL-arvon anonymisointiesimerkkejä

Menetelmä										
Ei muutosta	60	255	45	37	64	55	12	12	13	54
Poistaminen	0	0	0	0	0	0	0	0	0	0
Katkaisu	38	31	13	5	0	23	12	12	13	22
Pyöristäminen	60	250	40	40	60	60	10	10	10	50
Ryhmittely	58	255	42	42	58	58	12	12	12	58
Luettelointi	7	9	4	3	8	6	1	1	2	5

Katkaisussa TTL arvo on todellinen arvo jaettuna 32 jakojäännös. Pyöristäminen tehtiin lähimpään kymmeneen. Ryhmittelyssä valittiin lähimpänä toisiaan olevat arvot (klusterointi) ja ryhmän alkioiden arvoksi valittiin lähinnä keskiarvoa oleva kokonaisluku. Luetteloinnissa arvot on asetettu suuruusjärjestykseen ja luetteloitu pienimmästä alkaen.

Taulukko 4. Kiertoaikaviiveen (RTT) anonymisointi

Menetelmä										
Ei muutosta	4.5	235.7	19.0	13.5	163.4	19.9	4.9	3.1	167.8	29.1
Poistaminen	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Katkaisu	4.0	235.0	19.0	13.0	163.0	19.0	4.0	3.0	167.0	29.0
Pyöristäminen	0.0	230.0	10.0	10.0	160.0	20.0	0.0	0.0	160.0	20.0
Yhdistely	4.2	189.0	27.2	27.2	189.0	27.2	4.2	4.2	189.0	27.2
Kohina	5.0	245.8	25.9	18.4	163.7	22.9	14.7	3.7	187.9	32.0
Luettelointi	2	10	5	4	8	6	3	1	9	7

Kiertoaikaviiveen anonymisoinnissa käytettäessä katkaisumenetelmää, sekunnin murto-osat jätettiin pois. Pyöristäminen on tehty 10 sekunnin tarkkuudella. Yhdistelyssä lähinnä toisiaan olevat arvoista on laskettu keskiarvo ja tämä on kaikkien k.o. ryhmän arvona. Kohinassa arvoihin on lisätty tasajakautunut satunnainen arvo väliltä 0.0–20.0 sekunttia. Luetteloinnissa arvot ovat suuruusjärjestyksessä kuten TTL-arvon tapauksessa.

Liikenne- ja viestintävirasto Traficom
Kyberturvallisuuskeskus

PL 320, 00059 TRAFICOM
p. 029 534 5000

kyberturvallisuuskeskus.fi

ISBN 978-952-311-778-5
ISSN 2669-8757 (verkkojulkaisu)

Liikenne- ja viestintävirasto
Kyberturvallisuuskeskus