Tampereen yliopisto

Eveliina Toivanen

# SYBILHYÖKKÄYS SOSIAALISEN MEDIAN ALUSTOILLA

Informaatioteknologian ja viestinnän tiedekunta
Kandidaattitutkielma
Lokakuu 2022

# TIIVISTELMÄ

Eveliina Toivanen: Sybilhyökkäys sosiaalisen median alustoilla
Kandidaattitutkielma
Tampereen yliopisto
Tietojenkäsittelytieteiden tutkinto-ohjelma
Lokakuu 2022

---

Sybilhyökkäys on kansainvälisesti tunnettu tietoturvallisuusuhka, jossa vastustaja luo useita pseudonyymejä identiteettejä, sybileitä, tarkoituksenaan aiheuttaa vahinkoa hyökkäyksen kohteelleen. Tapoja sybilhyökkäyksen käyttämiseen on paljon, mutta sosiaalisen median alustoilla siitä on tullut erityisen suosittu, sillä alustat ovat avoimia ja siten varsin otollisia kohteita hyökkääjälle. Tässä tutkielmassa tutkitaan sybilhyökkäystä. Mikä se on, kuinka siltä voi puolustautua ja mitä käyttökohteita sillä voi olla, muuta laajuuden nimessä keskitytään pääasiassa sybilhyökkäyksiin sosiaalisen median alustoilla.

Tutkielma on toteutettu kirjallisuuskatsauksena, jonka menetelmät on paremmin selitetty luvussa 2. Ensin käsitellään sybilhyökkäystä itseään, sitä miten se toimii ja millä tavalla sitä ilmenee sosiaalisessa mediassa. Sen jälkeen siirrytään esittelemään puolustuskeinoja ja käydään läpi hieman niiden kehityskaarta. Sen jälkeen pohditaan näiden tietojen pohjalta hyökkäystä astetta syvemmältä ja katsotaan sitä, millaisia tulevaisuudennäkymiä hyökkäyksellä ja sitä vastaan puolustautumisessa voisi olla.

Tutkielman pohjalta voidaan sanoa, että sybilhyökkäyksiltä puolustautuminen on haasteellista, mutta sybileistä on löydetty yksi merkittävä avaintekijä. Se, etteivät ne kykene luomaan merkittävää määrää suhteita rehellisten käyttäjien kanssa. Tätä silmällä pitäen on kehitetty lukuisia puolustuskeinoja, joista iso osa nimenomaan sosiaaliseen mediaan kehitetyistä mekanismeista pohjautuvat joko sosiaaliseen graafiin tai koneoppimiseen. Kumpikin näistä on saanut merkittävän suosion tiedeyhteisön keskuudessa. Puolustuskeinojen kehittäminen on kuitenkin kehittänyt myös sybileitä. Osa niistä voi jo imitoida rehellisiä käyttäjiä lähes virheettä. Tämän takia viimeisimpinä vuosina sybilhyökkäyksen tutkimuksessa ollaan alettu nojaamaan sybilien käyttäytymisen tutkimiseen ja kiinnostus hybridimalleihin, joissa käyttäytymistutkimusta ja perinteisempiä puolustuskeinoja valjastetaan yhteistoimintaan toistensa kanssa, on alkanut nostaa päätään tiedeyhteisössä.

Avainsanat: sybilhyökkäys, tietoturvallisuus, kyberturvallisuus, puolustus, sosiaalinen media

# Sisällysluettelo

# 1 Johdanto

Internet tuo ihmiset yhteen globaalisti. Tieto kulkee nopeasti ja helposti maailman toiselle puolelle ja samalla turvallisuus on kohdannut uuden suuren huolenaiheen, sillä vieraiden kanssa kommunikoiminen lisää internetkäyttäjien kohtaamia riskejä (Lobo ym. 2021). Viestinnän mullistuminen on nostattanut sosiaalisen median suosiota huomattavasti. Tämän myötä, ja hyödyistä huolimatta, suosion mukanaan tuomat haitat ovat lisääntyneet ja erityisesti turvallisuusriskit ovat kohoamassa. (Al-Quirishi ym. 2017; Thakur ym. 2019; Alharbi ym. 2021; Lobo ym. 2021.) Kuka tahansa voi luoda helposti profiilin sosiaalisen median palveluun ja levittää sitä kautta tietoa. Alustojen avoimuuden takia tusinakäyttäjästä on siis tulossa suhteellisen helppo saalis kyberrikolliselle. Ja vaikka alustat itsessään eivät välttämättä luo uusia uhkia, ne lisäävät jo olemassa olevien uhkien määrää. (Thakur ym. 2017.)

Tämä tutkielma pyrkii astumaan tähän juuri esiteltyyn maailmaan ja tarkastelemaan yhtä, esimerkiksi Thakurin kollegoineen (2019), perinteiseksi kyberturvauhaksi nimittämää uhkaa, sybilhyökkäystä. Kyberturvallisuus on internetiin liitoksissa oleva turvallisuuden osa-alue, joka muuttuu jatkuvasti. Ei ole tapaa, jolla tieto ja resurssit olisivat täydellisesti turvassa kyberverkossa. (Thakur ym. 2019.) Sybilhyökkäys puolestaan tarkoittaa sitä, että hyökkääjä luo massoittain profiileja väärillä personoiduilla tiedoilla, joita hyökkääjä käyttää iskussaan hyödyksi. Näitä profiileja kutsutaan sybileiksi.

Tässä työssä olisi tarkoitus sukeltaa pintaa syvemmälle ja saada selville, mikä sybilhyökkäys oikein on? Millaisissa tapauksissa sitä käytetään? Ja miten siltä voidaan puolustautua? Tätä kaikkea tarkastellaan sosiaalisen median tulokulmasta. Toisin sanoen, työn tarkoitus on tutkia kuinka sybilhyökkäyksiä käytetään sosiaalisen median alustoilla. Aihe itsessään valikoitui hiljalleen loppuvuoden 2021 aikana. Tietoturvallisuus on kiinnostanut minua jo jonkin aikaa ja halusin ottaa tutkielmaani näkökulman, joka sivuaisi läheltä loppukäyttäjän rajapintaa, liittyi aihe mihin tietojenkäsittelytieteiden osa-alueeseen tahansa. Sosiaalisessa mediassa toteutetuissa sybilhyökkäyksissä sybilit elävät aitojen käyttäjien keskuudessa, joka tuntui menevän sopivan lähelle päämäärää, ja aihe alkoi kiinnostaa, mitä enemmän siihen paneutui.

Aiemmat tutkimukset tuntuvat jakautuvan kahteen hyvin selkään osa-alueeseen. Osa tutkimuksista pyrkii selkeästi tarjoamaan kattavan ja yleisemmän kuvan sybilhyökkäyksestä (Al-Quirishi ym. 2017; Alharbi ym. 2021; Douceur, 2002), ja osa taas keskittyy selkeästi tarjoamaan jonkin puolustautumiskeinon sybileitä vastaan (Gao ym. 2015; Jiang ym. 2015; Shekokar & Kansara, 2016; Wei Wei ym. 2013). Jälkimmäinen osa on aineistohaun perusteella selkeä enemmistö. Näistä Douceurin (2002) artikkelin voisi sanoa olevan jopa historiallisesti merkittävä, sillä se antoi sybilhyökkäykselle pohjan ja Douceur

oli se, joka nimesi sybilhyökkäyksen. Alharbi kumppaneineen (2021) kokoaa sybilhyökkäyksen peruskuvan ja puolustusvaihtoehdot tiiviiseen pakettiin artikkelissaan sosiaalisen median uhkista. Al-Quirishi kollegoineen (2017) taas on laatinut laajan artikkelin sybilhyökkäyksestä itsestään, sen toimintatavoista ja puolustusmekanismeista.

Gao ja kumppanit (2015) esittelevät artikkelissaan kehittelemäänsä koneoppimismenetelmää SybilFramea, ja samoin Wei wei kollegoidensa (2013) kanssa esittelevät kehittämäänsä sosiaalisen graafin menetelmää SybilDefenderiä. Kumpaakin näistä menetelmistä voidaan pitää jonkinlaisena kulmakivenä puolustuksen kehitykselle, sillä lukuisat muut ovat käyttäneet niistä saatuja oppeja hyödykseen omissa tutkimuksissaan. Jiang kumppaneineen (2015) keskittyy artikkelissaan puolustukseen hieman suuripiirteisemmin ja tutkii sybilien joukkokäyttäytymistä kiinan sosiaalisen median palvelussa Ren-Renissä. Shekokar ja Kansara (2016) taas esittelevät kehittämäänsä hybridimallia, jossa sosiaalisen graafin kanssa puolustuksessa käytetään hyväksi tiettyjä havaittuja käyttäytymisaspekteja.

Yleinen konsensus tiedeyhteisössä tällä hetkellä on se, että sybilhyökkäys on yleinen turvallisuusuhka, jota vastaan palveluiden ja järjestelmien tulisi puolustautua. Sybilhyökkäys on kansainvälisesti tunnettu kyberturvauhka ja sitä on tutkittu mittavasti ja maailmanlaajuisesti. Sosiaalisen median yleistyminen on tuntunut nostavan hyökkääjien mielenkiintoa sybilhyökkäyksille ja sosiaalinen media näyttäisi tällä hetkellä olevan yksi käytetyimmistä alustoista. Sosiaalisessa mediassa sybilit yleensä lähettävät roskapostia ja spämmiviestejä, mutta niitä voi käyttää myös esimerkiksi käyttäjien huijaamiseen ja manipulointiin.

Sybilit keskustelevat pääasiassa lähinnä toistensa kanssa, mutta osa sybileistä voi käyttäytyä hyvin samankaltaisesti aitojen käyttäjien kanssa. Sybileille on tyypillistä luoda omia keskinäisiä ryhmittymiään, ja koska sybilien toiminta on niin monipuolista, puolustusmekanismien olisi hyvä alkaa suosia erilaisia hybriditoteutuksia tai useamman erilaisen puolustusmekanismin käyttöä. Puolustusmekanismeja on kehitetty lukuisia, mutta samalla sybilit ovat pystyneet kehittymään ja osa imitoi todella hyvin rehellistä käyttäjää, mikä on vaikeuttanut sybilien erottamista rehellisistä käyttäjistä.

Toisessa luvussa kuvataan tässä tutkielmassa käytetty tutkimusmenetelmä yksityiskohtaisemmin. Luvussa on kerrottu haun haasteista ja siitä, kuinka ja millä perustein lähteet on koottu. Kolmannessa luvussa käsitellään sitä, mitä sybilhyökkäykset ovat, hyökkäyksen perusteita ja sitä, miten hyökkäystä hyödynnetään sosiaalisen median alustoilla. Neljännessä luvussa käsitellään sitä, millaisia keinoja on kehitetty sybilhyökkäystä vastaan ja käydään hieman läpi sitä, miten tärkeimmät niistä toimivat. Viides luku sisältää keskustelua ja pohdintaa löydettyjen tulosten pohjalta ja viimeinen, kuudes, luku sisältää yhteenvedon tästä tutkielmasta. Loppuun on vielä lisätty lähteet.

# 2 Tutkimusmenetelmä

Tämä tutkielma on toteutettu kirjallisuuskatsauksena. Hauissa on pääsääntöisesti käytetty tietojenkäsittelytieteiden omia alakohtaisia tietokantoja ProQuestia, IEEE:tä ja SpringerLinkiä, mutta myös Tampereen yliopiston omaa hakupalvelua Andoria. Jotkin lähteistä on löydetty Google Scholarista, mutta nämä lähteet on ennen sitä haettu ja löydetty myös aiemmin mainituista tietokannoista. Osa lähteistä on löytynyt puhtaasti jo valituiksi tulleiden lähteiden lähdeluetteloista.

Hakusanat ovat painottuneet englanninkielisiin termeihin nimenomaan suomenkielisen materiaalin puuttuessa oikeastaan täysin. Hakusanavariaatiot pyörivät sybilhyökkäyksen, sosiaalisen median ja puolustuksen, sekä sen osa-alueiden kohdalla. Eniten käytetyt hakusanat ovat nostettu taulukkoon 1 sekä suomeksi että englanniksi niiden perusmuodossa.

Taulukko 1: Yleisimmät hakusanat

| Yleisimmät hakusanat | |
|---|---|
| Sybil | *Sybil* |
| Sybilhyökkäys | *Sybil Attack* |
| Sosiaalinen media | *Social media* |
| Sosiaaliset verkostot | *Social networks* |
| Sosiaalinen graafi | *Social graph* |
| Koneoppiminen | *Machine learning* |
| Turvallisuus | *Security / Safety* |
| Kyberturvallisuus | *Cyber security* |

Lähteiden valintaprosessi on ollut hieman erilainen lähteestä riippuen. Peruslähteet on pyritty valikoimaan huomattavasti tarkemmin kuin esimerkiksi lähteet, jotka on otettu esimerkiksi jostakin tietystä spesifistä toiminnasta ja eivät välttämättä esittele itse aihetta niin syvällisesti. Peruslähteet on ensisijaisesti valittu otsikon, avainsanojen ja tiivistelmän lukemisen kautta. Luotettavuutta on arvioitu sen perusteella, onko kyseisiä samoja artikkeleita löytynyt muista hakupalveluista kuin siitä mistä kyseinen artikkeli on löytynyt. Osa artikkeleista on esimerkiksi tarkastettu julkaisufoorumin kautta. Paljon painoarvoa on tullut valintoihin myös siitä, onko lähdettä käytetty muissa artikkeleissa lähteinä ja kuinka paljon. Erityisen paljon tälle on annettu painoarvoa puolustusmekanismeista kertovien artikkelien suhteen.

Suurin osa sybilhyökkäystä käsittelevistä artikkeleista käsittelee luonnollisista syistä sitä, kuinka hyökkäykseltä voidaan puolustautua, joka on tuottanut hieman vaikeuksia sopivien lähteiden löytymiseen, mutta lähteiksi on pyritty löytämään myös joitakin sellaisia artikkeleja, jotka katselevat sybilhyökkäystä yleisestä näkökulmasta, ja jotka eivät välttämättä keskity johonkin tiettyyn näkökulmaan. Lähteitä on myös verrattu jonkin verran toisiinsa tutkimuksen aikana siitä näkökulmasta, onko kyseisillä lähdeartikkeleilla samankaltaisuutta perusajatuksissaan, jos ne käsittelevät samaa aihetta.

Muut kuin peruslähteet, joista saattaa olla otettu aineistoa tutkielmaan esimerkiksi vain muutaman lauseen verran, ei ole syynätty yhtä perusteellisesti kuin peruslähteitä. Niistä on kuitenkin tarkastettu se, mistä lähteistä tieto on otettu, ja missä yhteydessä asia on nostettu esiin valinnan varmistamiseksi.

Lähteiden analysointi vei huomattavasti enemmän aikaa kuin mitä niiden etsimiseen meni, sillä sybilhyökkäyksestä löytyy huomattavan helposti materiaalia. Materiaalin paljous vaikeutti hieman etsimistä, sillä rajauksista huolimatta materiaalia saattoi löytyä paljon seulottavaksi. Analysointi on tämän vuoksi ollut vahvasti mukana osana valintaprosessia. Ensimmäisen analyysin jälkeen valitut artikkelit on vielä silmäilty läpi ja asetettu sitten omaan kategoriaansa. Kategorioita oli aineistohaun aikana kaksi: sybilhyökkäys yleisellä tasolla ja sybilhyökkäyksen puolustusmekanismit. Artikkelit jaettiin jompaankumpaan sen mukaan, millä tavalla artikkeli oli painottunut ja millaiseen muotoon se oli rakennettu. Luokittelun jälkeen valittu aineisto on luettu läpi ja siitä on otettu haluttuja huomioita ylös.

# 3 Sybilhyökkäys

Selkeyden vuoksi, sybilhyökkäyksen tekijästä käytetään tästä eteenpäin termiä vastustaja ja aitoja käyttäjiä nimitetään rehellisiksi käyttäjiksi.

Sybilhyökkäyksen nimi juontaa juurensa Flora Schreiberin kirjoittamasta kirjasta Sybil, joka kertoo naisesta, jolla on dissosiatiivinen identiteettihäiriö (Al-Quirishi ym. 2017; Douceur, 2002). Ensimmäisenä tätä nimeä tietoturvallisuuden puolella tähän nimenomaiseen hyökkäykseen yhdistettynä käytti John Douceur (Al-Quirishi ym. 2017) ja aineistoa kerätessä vaikutti siltä, että tiedeyhteisössä Douceuria pidetään merkittävänä tekijänä sybilhyökkäysten kehityksessä. Hänen työnsä löytyi monien artikkelien lähdeluetteloon listattuna. Lyhyesti selitettynä sybilhyökkäyksessä vastustaja, joka voi olla yksi tai useampi henkilö, luo useita pseudonyymejä identiteettejä, joilla vastustaja pyrkii tekemään vahinkoa hyökkäyskohteeseensa. Näitä kehitettyjä identiteettejä kutsutaan lyhyesti sybileiksi. Hyökkäyskohde voi olla melkein mitä vain ja sybileistä on moneen, mutta erityisen tehokas työkalu tämä hyökkäys näyttäisi olevan silloin, kuin tavoitteena on hyötyä hyökkäyksestä suuressa skaalassa.

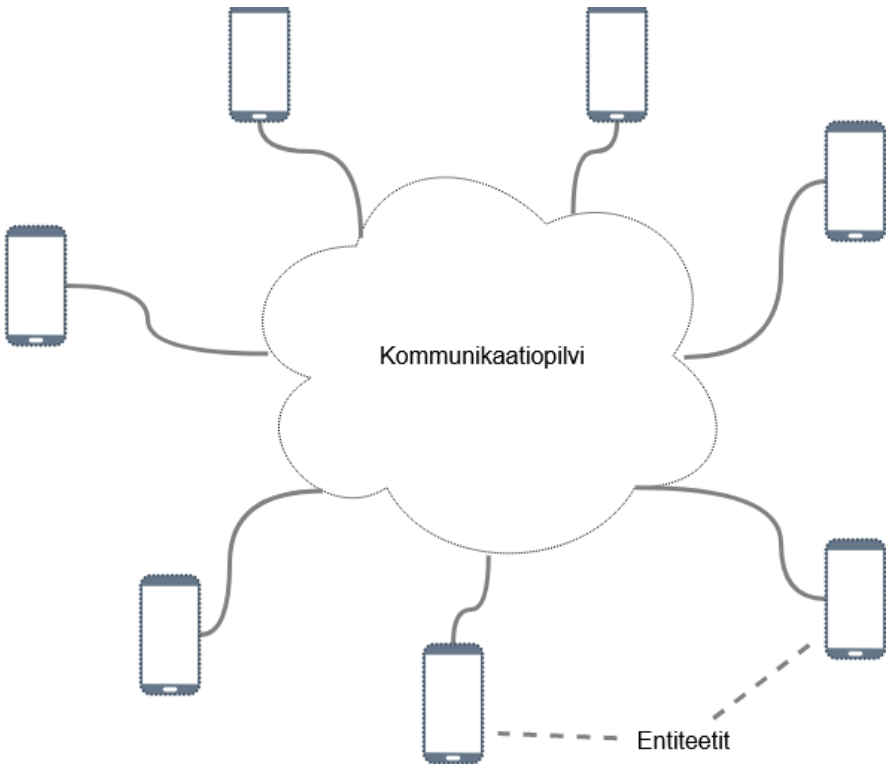
## 3.1 Hyökkäyksen toimintatapa

Koska sybilhyökkäyksiä on mahdollista käyttää monipuolisesti eri kohteisiin, on helpompaa käydä hyökkäystä ensin läpi hieman abstraktimmassa ja yleismaailmallisemmassa muodossa. Pohjana tässä kuvailussa on käytetty Douceurin (2002) mallinnusta sybilhyökkäyksistä. Vaikka malli on parikymmentä vuotta vanha, katson sen yhä antavan hyvän perustan sille, kuinka sybilhyökkäys oikein toimii. Toimintatavan perusteet pysyvät muuttumattomina kohteesta riippumatta.

Ensinnäkin on laaja joukko entiteettejä, joka jakautuu kahteen osaan: rehellisiin entiteetteihin ja petollisiin entiteetteihin. Rehelliset entiteetit noudattavat luotuja sääntöjä ja laadittuja protokollia. Petollisten entiteettien joukko puolestaan pystyy suorittamaan mitä tahansa mielivaltaista toimintaa, mikäli resurssirajoitukset eivät vain tule vastaan. Entiteettien lisäksi tarvitaan alusta, pilvi, kommunikaatiota varten. Entiteetit kommunikoivat toistensa kanssa tämän kommunikaatiopilven kautta, johon jokainen entiteetti on linkittyneenä. Tämä kaikki on koottu kuvaan 1.

Kuva 1 on mukailtu versio Doucerin (2002) kuvasta. Alkuperäinen kuva on hivenen vanhentunut nykyajan maailmaan verrattuna. Entiteetit on esitetty kuvassa älypuhelinten joukkona, jotka ovat yhteydessä kommunikaatiopilveen. Tämän pilven kautta entiteettien välinen kommunikaatio tapahtuu. Sybilhyökkäykset monesti tapahtuvat nykypäivänä jonkin laitteen kautta, ja kommunikaatiopilvi on se alusta, johon laite on yhteydessä. Oli

Kuva 1: Sybilhyökkäyksen perusta




se sitten sosiaalinen media tai jokin muu vastaava palvelu tai järjestelmä. Sosiaalisen median kohdalla jokaisen entiteetin voi ajatella vastaavan yhtä käyttäjäprofiilia ja kommunikaatiopilven vastaavan sitä sosiaalisen median alustaa, jonne käyttäjät on rekisteröity.

Entiteetit kommunikoivat toistensa kanssa viesteillä, jotka ovat käytännössä katkeamaton, äärellisen pituinen bittijono, jonka merkitys määräytyy joko eksplisiittisen protokollan tai entiteettijoukon välisen implisiittisen sopimuksen perusteella. Viestintä entiteettien välillä toimii hieman samalla tavalla kuin bysanttilainen yhteisymmärrys. Eli lähetetyt viestit välittyvät kaikille muille entiteeteille tietyn rajallisen ajan sisällä. Kaikki entiteetit eivät kuitenkaan välttämättä saa viestiä samassa järjestyksessä kuin mitä viesti on lähetetty. (Douceur, 2002.)

Toimintatapa ja sybilien muoto voi hyökkäyksissä vaihdella sen mukaan, millä alustalla ja mitä varten sybilhyökkäystä käytetään. Tässä kandidaatintyössä käsitellään sybileitä niiden ehkä yhdessä näkyvimmässä muodossa, sosiaalisen median palveluissa, mutta ne voivat toimia myös paljon huomaamattomammin, ikään kuin taustalla. Esimerkiksi Abbas (2019) esittää, kuinka sybilit voivat häiritä esineiden internetin, IoT:n, reititysprosesseja. Sybilit asetetaan reitityspoluille antamaan väärää kuvaa solmukohdista eri sijainneissa. Abbaksen (2019) mukaan tällainen toimintatapa on omiaan esimerkiksi useissa ilkeämielisissä toiminnoissa, kuten palvelunestohyökkäyksissä.

Sybileitä voi myös olla eri laatuisia. Trifunovic ja Hossmann-Picu (2016) kertovat artikkelissaan, kuinka sybileitä voidaan luoda kahdella eri tavalla: käyttämällä todellista identiteettiä tai vaihtoehtoisesti luomalla niitä virtuaalisen identiteetin kautta. Todellinen

identiteetti tarkoittaa sitä, että vastustaja on luonut sybileitä käyttämällä yhtä fyysistä laitetta tekemään useita todellisia identiteettejä, joista jokainen ilmoittaa olemassaolostaan muille solmuille joko eri aikoina, tai samanaikaisesti. Tämä tapahtuu samalla tavalla kuin rehellisten identiteettien kommunikointi keskenään. Ne vaativat enemmän resursseja ja erikoistaitoja, jonka takia niitä on vaikeampaa luoda. Todelliset identiteetit ovat myös virtuaalisia identiteettejä paljon tehokkaampia, sillä ne tunnistetaan virtuaalisia identiteettejä todennäköisemmin rehellisiksi solmuiksi.

Virtuaalinen identiteetti puolestaan on helpompi tapa luoda sybileitä. Vastustaja voi luoda virtuaalisen identiteetin väärentämällä rehellisille identiteeteille lähetetyt yhteyslokit. Tämä käytännössä merkitsee sitä, että virtuaalinen identiteetti on väärennyksen myötä vuorovaikutuksessa ainoastaan vastustajan todellisten identiteettien kanssa. Virtuaalinen identiteetti ei osaa lainkaan kommunikoida rehellisten identiteettien kanssa, sillä virtuaaliset identiteetit eivät välitä tietoa olemassaolostaan muulle maailmalle. Tämä tapa luoda sybileitä pätee ainakin opportunistissa verkoissa (Trifunovic & Hossmann-Picu, 2016), mutta viitteitä muihin verkkoihin ei vaikuta olevan ainakaan tällä hetkellä.

### 3.2 Sybilhyökkäys ja sosiaalinen media

Sosiaalisesta mediasta on tullut viimeisten parin vuosikymmenen aikana merkittävä ja tärkeä osa monen ihmisen elämää. Ne tarjoavat mutkattoman alustan yhteydenpitoon ja muiden kanssa kommunikointiin ilman, että siihen tarvitse muuta kuin oman tietokoneensa tai älylaitteensa. Sosiaalinen media näkyy kattavasti jokaisella arjen osa-alueella. Siihen luotetaan työelämässä ja jakamalla tietoa eri alustoilla, sosiaalisen median avulla voidaan myös esimerkiksi pyrkiä vaikuttamaan ihmisten maailmankuvaan.

Tästä sosiaalisen median yleisyydestä, avoimuudesta ja helppoudesta johtuen, sosiaalisen median palvelut ovat jatkuvasti kyberturvallisuusuhan alla. Tämän jatkuvan uhan seurauksena sosiaalisessa mediassa esiintyvä haitallinen ja ilkeämielinen toiminta voi vaikuttaa vakavasti niihin sosiaalisiin aktiviteetteihin, joihin käyttäjät osallistuvat verkossa ollessaan. Näitä aktiviteetteja ovat esimerkiksi sisällönjulkaisu, ystävyyssuhteiden luominen, viestittely, profiilien selaaminen ja julkaisujen kommentointi. Näin ollen siis suosion kasvaessa, kasvavat myös kannustimet hyökätä, sillä vahinko kohdistuu nyt laajemmalle alueelle yhteisöä. (Al-Quirishi ym. 2017.)

Tämä yleisyys ja avoimuus ovat nimenomaan ne tekijät, jotka tekevät sosiaalisesta mediasta kiinnostavan vastustajan silmissä. Avoimuuden takia, alustoilla monesti vallitsee tietynlainen sokea luottamus täysin uusiinkin kontakteihin. Sen lisäksi sosiaalisessa mediassa käytössä olevien suosittelualgoritmin rajoitukset ja mainejärjestelmän haavoittuvuus ovat myös vastustajien mieleen, kuten ovat myös mahdollisuudet laatia vääriä arvioita ja arvosteluja. (Al-Quirishi ym. 2017.)

Sosiaalisessa mediassa sybilit ovat väärennettyjä käyttäjäprofiileja, jotka vastustaja pyrkii rekisteröimään alustalle, ja joilla vastustaja pyrkii ystävystymään mahdollisimman monen oikean käyttäjän kanssa (Alharbi ym. 2021). Aiempaa mallinnusta kuvassa 1 katsottuna käyttäjäprofiilit ovat siis nyt entiteettejä ja sosiaalisen median alusta on kommunikaatiopilvi, jossa entiteetit vaikuttavat ja lähettävät viestejä toisilleen.

Sybileille yksi tyypillisimmistä ominaisuuksista näiden käyttäjäprofiilien kohdalla on se, että niiden jakamat käyttäjätiedot ovat epätarkkoja. Toisin sanoen ne ovat epätäydellisiä tai huonolaatuisia. (Al-Quirishi ym. 2017; Jiang ym. 2015.) Jiang kollegoineen (2015) antaa tästä esimerkkinä sen, kuinka esimerkiksi profiilien sähköpostit ovat harvoin vahvistettuja, profiilikuvat voivat olla samankaltaisia ja kuinka profiilit harvoin ovat erityisen aktiivisia. Yleinen oletus on, etteivät sybilit kykene juuri luomaan ystävyyssuhteita rehellisten käyttäjien kanssa.

Sybilhyökkäyksillä vastustajat pyrkivät sosiaalisessa mediassa saavuttamaan mahdollisimman monta rehellistä käyttäjää ja yksi keino tämän mahdollistamiseksi on luoda lukuisia ja lukuisia sybileitä palvelemaan tätä tarkoitusta. Sosiaalisessa mediassa suoritetuilla sybilhyökkäyksillä on helppo vaikuttaa alustan käyttäjien kokemuksiin alustan käytöstä. Käyttäjäkokemuksen lisäksi vastustajat voivat kuitenkin hyökkäyksillään aiheuttaa vahinkoa myös alustojen markkinointiin ja mainostukseen, sekä vaikuttaa negatiivisesti suorituskykyyn. Eli toisin sanon, vastustajien tarkoituksena on hyökkäystavasta riippumatta pyrkiä vaikuttamaan ja horjuttamaan merkittävällä tavalla hyökkäyskohdettaan. (Al-Quirishi ym. 2017; Alharbi ym. 2021; Lobo ym. 2021.)

Tärkeää on kuitenkin huomata, että sosiaalisen median alustoilla liikkuu useampi erilainen tapa identiteettien väärentämisen. Sybilhyökkäys ei tarkoita profiilien kloonausta, ja kaikki valeprofiilit eivät saman tien ole sybileitä (Hamid ym. 2020).

Sosiaalisessa mediassakin on useita erilaisia käyttökohteita sybileille. Yksi perinteisimmistä käyttökohteista sybilien kohdalla ovat roskaposti ja spämmäys. Toisaalta sybileitä voidaan käyttää myös vaikkapa käyttäjien manipulointiin ja huijaamiseen. Esimerkiksi Nitin kumppaneineen (2012) tuo esille, kuinka sybileitä voidaan käyttää sosiaalisen median äänestyksissä eli toisin sanoen, sybilhyökkäys voidaan kohdentaa esimerkiksi YouTube-videon tykkäyksiin. Collins kollegoineen (2011) mainitsee sybileitä voivan hyödyntää valeuutisten levittämiseen sosiaalisen median alustoilla. Sybileillä voidaan myös pyrkiä vaikuttamaan ihmisten mielipiteisiin. Esimerkiksi Lobo kollegoineen (2021) kertoo, kuinka vuonna 2016 pidettyjen Yhdysvaltojen presidentinvaalien aikaan sosiaalisen median palvelu Twitterissä 19 prosenttia kaikista vaaleja käsittelevistä twiiteistä oli sybilien luomia. Samoin 1/3 Donald Trumpia kannattaneista twiiteistä ja 1/5 Hillary Clintonia kannattaneista twiiteistä olivat sybilien luomia.

Sybilit voivat myös tehdä yhteistyötä toistensa kanssa ja Haifeng kumppaneineen (2010) mainitsee, että on havaittu tapauksia, joissa sybilit ovat käytännössä äänestäneet rehellisiä käyttäjiä ulos käyttämällä hyödykseen Bysantin konsensusta.

Toisaalta, sybilhyökkäys voi tuottaa myös hyviä asioita. Tang kumppaneineen (2019) tuo esiin sen, kuinka sybileitä varten kehitettyjä puolustusmekanismeja on voitu käyttää potilaille tarkoitetusta terveyteen keskittyvästä sosiaalisesta mediasta tunnistamaan ja löytämään valheelliset tulokset oikeiden tulosten joukosta.

# 4 Puolustuskeinot

Koska puolustuskeinoista puhuttaessa entiteetit usein nimetään solmuiksi, tehdään niin myös tässä luvussa.

Sybilhyökkäystä vastaan luotujen puolustusmekanismien päämäärä on tunnistaa täsmällisesti kaikki sybilit. Eli toisin sanoen puolustus tunnistaa rehelliset solmut ja hyväksyy ne, mutta pyrkii samalla olemaan hyväksymättä väärennettyjä solmuja eli sybileitä. (Alharbi ym. 2021.)

Sybilhyökkäyksiä vastaan on esitetty monia puolustusvaihtoehtoja, mutta yksikään ei tietenkään ole aivan aukoton. Shekokar ja Kansara (2016) mainitsevat, että sybileitä ei aina pystytä huomaamaan, sillä ne voivat käyttäytyä hyvin samankaltaisesti kuin rehelliset solmut. Tämä on yksi perimmäisistä syistä, jotka vaikeuttavat sybilien löytämistä huomattavasti. Jiang kumppaneineen (2015) myös lisää, että koska sybilien tekemät haitat ovat niin monipuolisia, yksi puolustusmekanismi ei voi mitenkään riittää kaikkien niiden tarkistamiseen. He painottavat sitä, kuinka vakavien hyökkäysten välttämiseksi olisi erityisen tärkeää tunnistaa sybilit ja niiden muodostamat ryhmät etukäteen ja seurata näiden toimintaa koko ajan. Usein sosiaalisen median kohteissa puolustuskeinoissa jaotellaan entiteetit rehellisiin käyttäjiin ja sybileihin, mutta esimerkiksi Lobo ja kumppanit (2021) käyttävät tutkimuksessaan vielä lisäluokkaa kyseenalaiset käyttäjät, joilla on sybilien kaltaista toimintaa, mutta eivät aivan täytä sybilien varsinaista kuvausta.

Puolustusmekanismit yleensä perustuvat joko *Random Walk* (RW) -tekniikoihin tai *Loopy Belief Propagation* (LBP) -tekniikoihin (Alharbi ym. 2021). RW-tekniikoissa käytännössä liikutaan solmujen välillä satunnaisessa järjestyksessä, kuten nimestä hieman voi päätellä. RW-tekniikat perustuvat siihen, että alkuperäisestä solmusta navigoidaan satunnaisesti yhteen naapurisolmuista, jolla on yhdistetty reuna alkuperäisen solmun kanssa. LBP-tekniikat puolestaan ovat RW-tekniikoita paljon systemaattisempia. Käytännössä LBP pitää sisällään dynaamisen ohjelmointitavan, jolla ratkaistaan ehdollisen todennäköisyyden kyselyitä graafien sisällä. LBP siis laskee jokaiselle havaitsemattomalle solmulle marginaalijakauman, joka on ehdollinen havaittujen solmujen suhteen. LBP-tekniikat levittävät iteratiivisesti tunnistetietoja sosiaalisissa graafeissa. Tämän johtuu siitä, että solmun tunnisteita pystyttäisiin ennustamaan. (Alharbi ym. 2021.)

Puolustuskeinojen kehittämisessä on yksi huomattava heikko kohta, kun puhutaan sosiaalisen median alustoista. Nimittäin se, että tutkijoilla ei ole kovin helppoa pääsyä suuren skaalan dataryhmiin oikeista sybileistä. Tällaiset suuren skaalan dataryhmät olisivat merkittävä apu puolustuskeinojen kehitystyössä, sillä oikean datan syväluotaus voisi tuottaa tärkeää tietoa sybilien piirteistä, sekä hyödyllistä numeerista tietoa. Sosiaalisen median palveluntarjoajilta tällaista tietoa esimerkiksi on hyvin vaikeaa, sillä jakaminen usein voi rikkoa palveluntarjoajien luottamuskäytänteitä. (Jiang ym. 2015.)

Alharbi kumppaneineen (2021) esittää, kuinka puolustuskeinoja sybilhyökkäystä vastaan on kolme eri kategoriaa niissä käytettyjen teknologioiden mukaan: sosiaaliseen graafiin perustuvat tekniikat, koneoppimiseen perustuvat tekniikat ja avaintenhallintaan perustuvat tekniikat. Myös Al-Quirishi kollegoineen (2017) jaottelee puolustusmekanismit kategorioihin. Heidän kategorioitaan ovat sosiaaliseen graafiin perustuvat metodit, koneoppimisen mallit, manuaalinen verifiointi, käyttäjäpalautteen käyttö ja ehkäisemiseen perustuvat menetelmät.

Näissä metodeissa, kuten esimerkiksi Jiang ja kumppanit (2015) tuo esille, on perusoletuksena se, että sybilit harvoin kykenevät kommunikoimaan rehellisten solmujen kanssa. Näistä menetelmistä sosiaalisen graafin menetelmät ja koneoppimisen mallit dominoivat sybilhyökkäyksiltä puolustautuessa. Näin on myös sosiaaliseen mediaan kohdistuvien sybilhyökkäysten kohdalla.

Puolustusmekanismien kehittämisen takia on hyvä tutkia sybileitä myös yleisemmällä tasolla. Jiang kollegoineen (2015) esimerkiksi suosittelee sybilien ryhmäkäyttäytymisen tutkimista ja sen pohjalta puolustusmekanismien parantamista. He tutkivat sybilien ryhmäkäyttäytymistä yhdessä Kiinan suurimmista sosiaalisen median palveluista Renrenissä, ja tulokset puhuivat sen puolesta, että sybilit tapaavat muodostaa ryhmiä keskenään. Näissä ryhmissä olevilla sybileillä on vahvat suhteet toistensa kanssa ja sybilit voivat luoda toistensa kanssa peräti kokonaisia sybilyhteisöjä, jotka sisältävät suuren määrän sybileitä. Jiang kollegoineen (2015) myös havaitsi, että näillä sybilryhmillä näyttäisi olevan myös omia erityisiä kehityskuvioita niiden välisissä yhteysrakenteissa. He kommentoivat uskovansa, että tämä löydös saattaisi muuttaa puolustusmekanismeja tulevaisuudessa.

### 4.1 Sosiaalisen graafin menetelmät

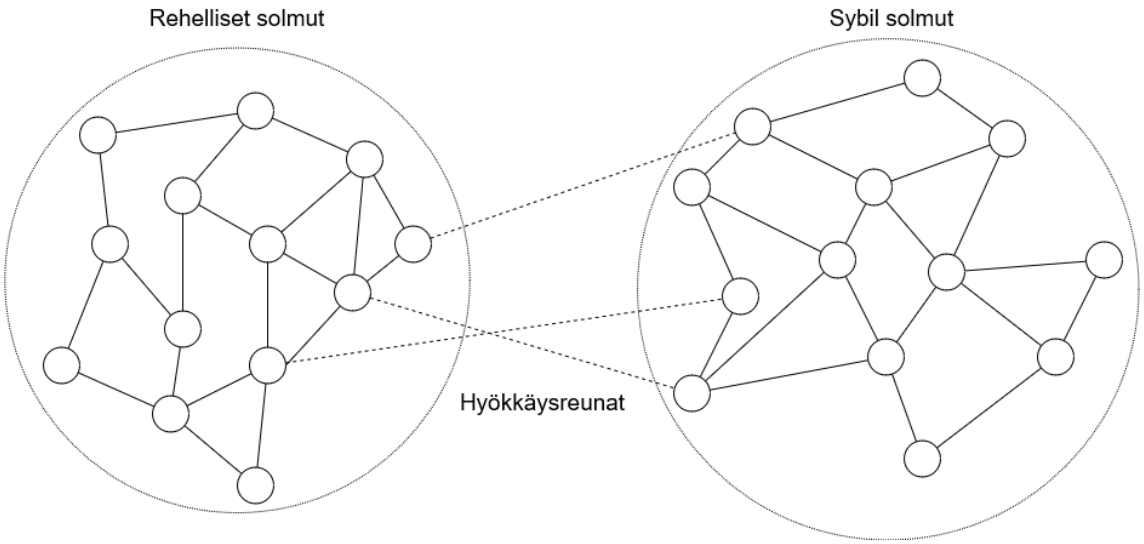
Sosiaalinen graafi on viimevuosina ollut vallitseva trendi sybilhyökkäykseltä puolustautumisessa ja aineistohaun perusteella se näyttäisi olevan ehdottomasti suosituin puolustautumiskeino.

Sosiaalisen graafin menetelmät pohjautuvat sosiaalisista rakenteista muodostettuihin sosiaalisiin graafeihin, eli nämä menetelmät pohjimmiltaan käyttävät solmuista keräämäänsä dataa sybilhyökkäyksiltä puolustautumiseen (Alharbi ym. 2021). Sosiaalisen median kohdalla tämä tarkoittaa siis sitä, että sosiaalinen graafi muodostetaan käyttäjäprofiilista kerättyjen tietojen perusteella. Nämä sosiaalisen graafin sosiaaliset rakenteet yleensä perustuvat useisiin oletuksiin siitä, millä tavalla sybilit toimivat ympäristössään verrattuna rehellisiin solmuihin. Tärkeimpänä osa-alueena voidaan pitää ja tarkastella sybilien rajallisia sosiaalisia yhteyksiä rehellisten solmujen kanssa, joita kutsutaan myös hyökkäysreunoiksi. (Shekokar & Kansara, 2016; Al-Quirishi ym. 2017.)

Sosiaalinen graafi muodostetaan siten, että katsotaan millaisia suhteita ja linkkejä, eli eräänlaisia luottamussiteitä, yksi solmu luo muiden solmujen kanssa ja sen lisäksi tarkastellaan kyseisen solmun toimintaa. Nämä kaksi yhdistetään ja sen perusteella pyritään löytämään hyökkäysreunat. Lopulta kokonaisuudesta syntyy sosiaalinen graafi. Havainnollistava esimerkki sosiaalisesta graafista on esitelty kuvassa 2. Kuvan 2 mallinnus sosiaalisesta graafista on hyvin perinteinen. Rehelliset käyttäjät ja sybilit on eroteltu toisistaan omiin ryhmiinsä. Solmut on merkattu kuvassa ympyröinä ja viivat solmujen välillä ovat niiden luomia siteitä muiden solmujen välillä. Rehellisten solmujen ja sybilien väliset siteet on esitetty katkoviivoina ryhmien välillä. Nämä ovat hyökkäysreunat. Joissakin tapauksissa tähän malliin on lisätty leikkausviiva sybilien ja rehellisten solmujen väliin hyökkäysreunojen päälle (esimerkiksi Shekokar & Kansara 2016), mutta se ei vaikuta olevan millään muotoa standardi ja kuvan 2 malli vaikuttaa yleisemmältä tavalta.

Yksi aineistohaun perusteella tiedeyhteisön arvostetuimmista ja vahvasti viitatuimmista sosiaalisen graafin puolustusmekanismeista on Wei Wein ja kollegoiden SybilDefender vuodelta 2013. Se on sosiaaliseen mediaan kehitetty, verkostotopologioita hyödyntävä, puolustusjärjestelmä. SybilDefender perustuu RW-tekniikkaan, jossa solmujen välillä tapahtuvien satunnaisesti valittujen siirtymien ”kävelyiden” määrä on rajoitettu sosiaalisten graafien sisällä. Menetelmä on erittäin skaalautuva käyttökohteen koon mukaan ja sen erikoisuus lienee siinä, että se kykenee havaitsemaan oikeat sybilit sellaisissakin tilanteissa, kun jokaisen hyökkäysreunan tuomien sybilien määrä lähestyy teoreettisesti havaittavaa alarajaa.

Kuva 2: Sosiaalisen graafin malli




SybilDefender koostuu kahdesta erilaisesta algoritmista: sybilintunnistusalgoritmista ja sybilyhteisönhavaitsemisalgoritmista. Jälkimmäinen on kehitetty siitä syystä, että kaik-

kien solmujen tutkiminen yksitellen sybilien varalta on varsin tehotonta suuressa mittakaavassa. On helpompaa ryhmitellä ensin isompi määrä solmuja kerralla. Näiden algoritmien lisäksi puolustusmekanismille on kehitetty myös kaksi erilaista lähestymistapaa hyökkäysreunojen rajoittamiseksi. Toinen näistä antaa käyttäjälle itselleen mahdollisuuden arvostella suhdettaan kaverilistansa henkilöihin joko arvioimalla heidät kavereiksi tai tuntemattomiksi. Toinen lähestymistapa taas perustuu siihen miten aktiivisesti käyttäjät kommunikoivat toistensa kanssa. (Wei Wei ym. 2013.)

Puolustusmekanismien kohdalla on hyvä huomata, että tutkijat saattavat tehdä vanhoista mekanismeista uudempia, tehokkaampia, versioita käyttämällä pohjalla samaa puolustusideologiaa ja toimintatapaa. Tunnettu tällainen pari on esimerkiksi Haifengin kumppaneineen kehittämä SybilGuard (2006) ja sen päivitetty versio SybilLimit (2010), joita yhä usein käytetään uudemmissa artikkeleissa lähdeteoksina.

Suurin huomio näiden kahden välillä, puolustustehon lisääntymisen lisäksi, on uuden teorian todistaminen aidossa ympäristössä. SybilGuard (Haifeng ym. 2006) puolustusmekanismina olettaa, että sosiaaliset verkostot sekoittuvat nopeasti, eivätkä pysy staattisina kovin pitkään. Tällaisia merkkejä oli esiintynyt mallidatassa. Myöhemmin Haifeng kumppaneineen totesi, että todellisessa elämässä, tälle ei ole vahvistettua näyttöä. He päättivät testata teoriaa käytännössä SybilLimitin avulla, ja tulivat todisteiden valossa positiiviseen tulokseen teorian paikkaansa pitävyydestä. (Haifeng ym. 2010.) Varsinkin tämän jälkeen tästä teoriasta on pidetty kiinni tiedeyhteisössä ja esimerkiksi Wei Wein ja kumppaneiden SybilDefender (2013) toimii myös saman oletuksen pohjalta.

Kolmantena esimerkkinä on hieman uudempi, Jian kollegoineen vuonna 2017 julkaisema, RW-tekniikkaan perustuva SybilWalk. He kehittivät SybilWalkin tukemaan RW-tekniikoiden hyviä puolia; teoreettisesti taattua suorituskykyä verkkoyhteisössä, tarkkuutta ja skaalautuvuutta, mutta vähentäen RW-tekniikoiden kärsimiä rajoituksia. Tällaisiksi ominaisuuksiksi he listasivat kyvyttömyyden hyödyntää samaan aikaan sekä merkittyjä rehellisiä solmuja ja merkittyjä sybileitä, rajallisen havaitsemistarkkuuden heikon homofilian sosiaalisissa verkostoissa ja heikon kyvyn melun leimaamiseen koulutusdatassa. (Jia ym. 2017.)

SybilWalk toimii niin, että se pyrkii tavoittamaan rehellisten solmujen ja sybilien välillä olevan rakenteellisen kuilun RW-tekniikkaa hyödyntämällä sosiaalisessa verkostossa, joka on lokeroitu. Tämä tarkoittaa sitä, että käyttäjät ovat eri tasoilla toisiinsa nähden. (Jia ym. 2017).

Usein sosiaalinen graafi yksinään ei itsessään ole puolustuskeino, vaan sen ympärille rakennetaan erilaisia algoritmeja ja muita välineitä, joilla sybileitä etsitään rehellisten solmujen joukosta. Hiljattain sosiaalisen graafin rinnalle on alettu nostamaan käyttäytymis-

tieteen osasia tukemaan graafin löydöksiä. Esimerkiksi Shekokar ja Kansara (2016) kehittivät puolustuskeinon SNI (*sybil node identification*) ja sen lisäosan SNI-B:n (*sybil node identification – behavioural aspects*).

SNI itsessään on perinteinen sosiaaliseen graafiin nojaava menetelmä, joka pyrkii sosiaalisen graafin avulla löytämään sybilit rehellisten solmujen joukosta. Se laskee millä tavalla solmujen väliset luottamussiteet ja toiminta solmujen välillä toimii, ja pyrkii sitä kautta löytämään hyökkäysreunat. SNI perustuu täysin solmujen luomiin siteisiin toistensa kanssa, jonka takia tarvitaan SNI-B:tä. SNI-B tarkastaa SNI-menetelmän tuottamat tulokset käyttämällä avukseen kahta käyttäytymiseen perustuvaa näkökulmaa: solmun vastaanottamien ystäväpyyntöjen hylkäykset ja solmun vastaanottamien profiilikäyntien määrä, eli solmun piilevät vuorovaikutukset muiden kanssa. Nämä aspektit on valittu siksi, että sybileillä niitä on vähemmän kuin rehellisillä solmuilla. SNI-B:n tarkoituksena on varmistaa, ettei SNI ole tuottanut valheellisia negatiivisia tuloksia sybileistä. (Shekokar & Kansara, 2016.)

### 4.2 Koneoppimisen menetelmät

Toinen suosituista menetelmistä sosiaalisen median alustoilla on puolustusmekanismit, joissa on käytetty hyödyksi tekoälyn yhtä osa-aluetta, koneoppimista. Koneoppiminen on itsestään toimiva tekniikka, joka hankkii ja integroi itsenäisesti tietoa, jota se on hankkinut kokemuksen tai analyyttisten havaintojen kautta. Se on kehitetty erityisesti ratkaisemaan ongelmia, joihin liittyy valtavia tietomääriä, joilla on monia eri muuttujia. (Al-Quirishi ym. 2017). Nämä menetelmät perustuvat sybilhyökkäyksen kohdalla siihen, että solmujen toimintahistorian perusteella pyritään muodostamaan käyttäytymismalleja. Toisin sanoen solmujen toimintaa siis seurataan sosiaalisen median alustalla tietyn rekisteröintijakson ajan. Alharbin ja kumppaneiden mukaan (2021) tämä seurattu toiminta käsittää kaiken sen, mitä kuuluu käyttäjäprofiilin ja sosiaalisen median alustan muiden elementtien välisiin yhteyksiin ja kommunikaatioon. Yksittäisen solmun kohdalla tällaista toimintaa on helppo seurata, sillä se on hyvin läpinäkyvää ja julkista.

Koneoppimisen algoritmit yrittävät sybileitä etsiessään löytää yleisiä laadittuja käyttäytymissääntöjä rikkovia solmuja. Mallien koulutusdata koostuu rehellisistä solmuista sekä sybileistä, ja yhtä solmua edustaa joukko solmujen tiedoista kerättyjä ominaisuuksia, sekä solmuista ja niiden toiminnasta muodostettu sosiaalinen graafi. (Alharbi ym. 2021.) Koneoppimisen menetelmät voidaan karkeasti jakaa kahteen päätyyppiin: valvotut menetelmät ja valvomattomat menetelmät (Al-Quirishi ym. 2017).

Valvotut menetelmät pitävät usein sisällään erilaisia regressiomalleja tai päätöspuumalleja, sekä sellaisia luokitinmalleja kuin naiivi Bayesin luokitin (*naive bayes classifier*) ja tukivektorikone (*SVM, support vector machine*). Valvotuissa menetelmissä hyvin olen-

naista on se, kuinka datan tietoaluetta käytetään ominaisuuksien rakentamisessa. Eli toisin sanoen tärkeää on käyttää tunnettua dataa. Tämä aiheuttaa ongelmia sosiaalisen median kohdalla, sillä kuten jo aiemmin tuli ilmi, oikeita sybildataryhmiä sosiaalisen median alustoilta on vaikea saada haltuun. (Al-Quirishi ym. 2017; Jiang ym. 2015.) Tästä johtuen valvottuja menetelmiä on kehitetty vain muutamia.

Valvomattomat menetelmät yleensä tarkoittavat erilaisia klusterointialgoritmeja. Klusterointi näissä algoritmeissa suoritetaan siten, että objektit, joilla on samankaltaisia ominaisuuksia, ryhmitellään yhdeksi joukoksi. Klusteroinnin takia nämä menetelmät eivät tarvitse valvottujen menetelmien tavoin jo ennalta tiedossa olevaa dataa ja nämä algoritmit voivat jopa paljastaa piilotettuja tai piilossa olevia rakenteita merkitsemättömästä tiedosta. (Al-Quirishi ym. 2017.)

Tiedeyhteisössä nämä mallit eivät kuitenkaan tunnu herättävän erityisen suurta luottoa, vaan tunnetuimmat koneoppimismallit sybilhyökkäyksiä vastaan ainakin sosiaalisen median suhteen ovat puolivalvotut mallit, joissa käytetään sekä valvottujen menetelmien merkittyä ja tiedettyä dataa kuin myös valvomattomien mallien merkitsemätöntä dataa. Tämä johtuu siitä, että on huomattu, että näiden kahden mallin datan käyttö yhdessä on nostanut huomattavasti koneoppimisen laatua. (Al-Quirishi ym. 2017.)

SybilFrame (Gao ym. 2015) ja SybilBelief (Gong ym. 2014) ovat tunnettuja puolivalvottuja menetelmiä, joihin tunnutaan tiedeyhteisössä nojaavan paljon. SybilBeliefin (Gong ym. 2014) koulutusdata perustuu pieneen määrään sybileitä ja rehellisiä solmuja, joiden sosiaalista verkostoa menetelmä vertailee ja luokittelee tarkoituksenaan tunnistaa rehelliset solmut sybileistä. SybileFrame (Gao ym. 2015) taas on menetelmä, jossa pyritään monivaiheiseen luokittelumekanismin avulla analysoimaan heterogeenisiä lähteitä ja tietotyyppejä sosiaalisen median profiileista. Keskeinen osa SybilFramen toimintaa on se, että sosiaalisessa graafissa, jossa reunat edustavat vahvoja luottamussuhteita solmujen välillä, on vastustajien vaikea luoda näitä samanlaisia linkkejä rehellisiin solmuihin.

Kumpikin, niin SybilFrame kuin SybilBelief, on suunniteltu toimimaan nimenomaan sosiaalisen median alustoilla. Kolmas tällainen puolivalvottu ja sosiaaliseen mediaan suunniteltu, hieman kahta aiempaa uudempi, koneoppimismenetelmä on SybilTrap. SybilTrap on sosiaalista graafia hyväkseen käyttävä menetelmä, joka hyödyntää RW-tekniikkaa ja sen absorbointitiloja levittääkseen tunnisteita sosiaaliseen graafin sisällä. (Al-Qurishi ym. 2018.)

Sekä SybilFrame (Gao ym. 2015) että SybilTrap (Al-Qurishi ym. 2018) on suunniteltu toimimaan ympäristössä, jossa hyökkäysreunoja on todella paljon. SybilBelifin (Gong ym. 2014) tarkkuus puolestaan tuntui kärsivän, jos hyökkäysreunojen määrä kasvoi erittäin suureksi. Toisaalta Gao ja kollegat (2015) noteerasivat, että koska suurin osa

sybileistä on eristettyjä käyttäjiä, on niiden havaitseminen vaikeaa SybilFramen tekniikkaa käyttämällä. Voisi siis sanoa, että koneoppimisen metodeissa on vielä paljon kehitettävää.

# 5 Keskustelu

Sybilhyökkäys on suhteellisen vanha ja monimuotoinen kyberturvauhka, jonka yksi suosittu käyttökohde on sosiaalinen media. Sosiaalinen media itsessään on varsin hyvä kohde vastustajille, sillä se on alustana avoin ja näin ollen rehelliset käyttäjät ovat helposti aivan vastustajan käden ulottuvilla. Aineistohaun perusteella sybilhyökkäysten lisääntyminen näyttäisi olevan tapahtunut melko lailla käsikädessä sosiaalisen median yleistymisen kanssa 2010-luvulta lähtien. Sosiaalista mediaa koskevien aineistojen määrä sybilhyökkäyksestä näyttäisi lisääntyneen huomattavasti tuon rajapyykin jälkeen.

Sybilit esiintyvät sosiaalisessa mediassa usein toisten kyberturvahaittojen muodossa, kuten spämminä tai roskapostin lähettämisenä. Yksi selkeä nykyaikainen suosikki vastustajien keskuudessa näyttäisi olevan ihmisiin vaikuttaminen väärän informaation ja valeuutisten turvin. Vastustajat luovat sybileitä, jotka sitten ujuttautuvat ympäristöönsä levittämään valeuutisia. Sybilit voivat myös levittää esimerkiksi poliittista propagandaa pyrkien sitä väylää pitkin vaikuttamaan rehellisten käyttäjien ajatuksiin ja mielipiteisiin. Tämä olisi hyvä ottaa huomioon jo ihan lähitulevaisuudessa. Mielenkiintoista dataa tällaisesta toiminnasta voisi tulla esimerkiksi tutkimalla sybilien mahdollista käyttöä esimerkiksi koronapandemian ajan viestinnässä.

Koska sybilhyökkäystä on tutkittu niin monen vuoden ajan, on puolustuskeinojakin syntynyt vuosien saatossa runsaasti. Tämän myötä monet puolustusmenetelmät muistuttavat paljon toisiaan ja monet keinot ovatkin kehitetty yhdestä tai useammasta toisesta, jo kehitetystä, puolustuskeinosta. Tällä hetkellä tiedeyhteisö on vahvasti nojautunut käyttämään sosiaalista graafia sybilhyökkäyksiltä suojautumisessa. Sitä käytetään monissa tilanteissa yksinään, mutta myös muiden keinojen pohjatietona tai perustana. Sosiaalinen graafi on oletettavasti ainakin osasyy sille, että sybilien on huomattu usein toimivan omissa ryhmissään ja muodostavan kokonaisia yhteisöjä keskenään. Ylipäänsä käyttäytymistutkimuksen kohdistaminen sybileihin ja sen lisääminen aktiivisesti puolustuskeinojen kehitykseen, voisi tuottaa lisää tehoja sybilhyökkäysten torjuntaan. Tällainen tutkimus tekisi sybileistä enemmän ennalta-arvattavia ja sitä kautta hyökkäysten ennaltaehkäisy ja sybilien löytäminen nopeasti rekisteröinnin jälkeen voisi tuottaa arvoa sosiaalisen median alustoille.

Sosiaalisessa mediassa sybilhyökkäyksiltä tavanomaisesti tunnutaan puolustautuvan kahdella tavalla: sosiaalisen graafin turvin ja koneoppimisen menetelmin. Koneoppimisen menetelmät tuntuvat olevan jo suhteellisen vanhoja ja tuoreimmat tutkimukset, jotka löytyivät pyörivät jossakin vuoden 2017 tienoilla. Jos uudempia on, niitä on todella hankala löytää. Koneoppimisen malleissa näkyy selkeä potentiaali hyökkäysten torjuntaan, ja tuskin sen kaikkea tehokkuutta on vielä käytetty. Voisi siis olettaa, että joko sitä tutkitaan tällä hetkellä lisää tai uusia menetelmiä on mahdollisesti tulossa jossain lähitulevaisuudessa.

Toinen todennäköinen seuraus tulevaisuudessa on, että käyttäytymistutkimuksen käyttö sybileitä vastaan kasvaa ja kasvaa vielä nimenomaan sosiaalisen median kohdalla, jossa sybilien käytös on usein avainasemassa.

Yksi selkeä olennainen osa tulevaisuutta ja tulevaisuuden puolustusta ajatellen on, että tutkijoilla olisi mahdollisuus päästä käsiksi suuren skaalan dataryhmiin aidoista sybileistä sosiaalisen median alustoilta. Esimerkiksi koneoppimisen mallit hyötyisivät tästä mahdollisuudesta todennäköisemmin huomattavasti enemmän kuin sybileistä mallinnetuista datasta. Alustat ovat kuitenkin olleet nihkeitä tämän suhteen luottamuskäytänteiden ja tietosuojan takia. Tulevaisuudessa sosiaalisen median alustat voisivat yrittää luomaan tutkijoiden kanssa jonkinlaisia yhteistyökuvioita aiheen tiimoilta, sillä paremmin toimivat puolustusmenetelmät hyödyttävät myös alustoja itseään.

# 6 Yhteenveto

Sybilhyökkäys on kansainvälisesti suhteellisen tunnettu ja monesti perinteiseksi uhaksi luokiteltu kyberturvallisuusuhka, jossa vastustaja luo useita toistensa kanssa samankaltaisia identiteettejä tarkoituksenaan vahingoittaa hyökkäyksen kohdetta tavalla tai toisella. Näitä pseudonyymejä identiteettejä kutsutaan sybileiksi. Sybilhyökkäystä on tutkittu jo parivuosikymmentä, jos aika lasketaan lähteväksi siitä, kun Douceur (2002) ensimmäistä kertaa kutsui hyökkäystä sybilhyökkäykseksi. Suomessa sybilhyökkäyksen käsitettä ei tunnuta juuri tunnettavan, vaikka konsepti ehkä on tuttu tavalla tai toisella, ja esimerkiksi sosiaalisessa mediassa moni on saattanut jopa sybileitä kohdata.

Sosiaalisessa mediassa sybilhyökkäykset ovat olleet läsnä valtaosan sen historiasta ja hyökkäystä käytetään sosiaalisen median alustoilla hyvin monipuolisesti. Oli tavoite mikä tahansa, yleensä vastustajan tavoitteena on soluttautua rehellisten käyttäjien kontaktilistoilla ja sitä kautta levittäytyä sosiaalisen median ympäristössä. Sybilit usein lähettävät rehellisille käyttäjille roskapostia tai spämmiviestejä, mutta sybilien käyttö ei rajaudu vain näihin kahteen vaan sybilien käyttömahdollisuudet ovat hyvin monipuoliset. Viime vuosien aikana yksi käyttökohde, jossa sybilien käyttö on yleistynyt, on yritykset vaikuttaa rehellisiin käyttäjiin. Sybilit voivat esimerkiksi jakaa valeuutisia, tai vaikkapa vaalipropagandaa, ja yrittää sitä kautta vaikuttaa ihmisten mielipiteisiin ja maailmankuvaan.

Sybilhyökkäykseltä suojautumista varten on kehitetty lukuisia erilaisia puolustusmekanismeja, myös sosiaalisen median näkökulmasta. Monet sosiaalisen median alustoilla käytetyistä puolustusmekanismeista perustuvat joko sosiaaliseen graafiin tai sitten koneoppimisen malleihin. Nämä kaksi ovat selkeästi tiedeyhteisön suosiossa, kun puhutaan sybilien torjunnasta. Sosiaalisen graafin ja koneoppimisen menetelmissä, kuten monissa muissa puolustusmekanismeissa, pohja usein luodaan sen oletuksen perusteella, etteivät sybilit kykene luomaan merkittävää määrää suhteita rehellisten käyttäjien kanssa. Jotkin sybileistä voivat kuitenkin jäljitellä rehellisen käyttäjän toimintaa sosiaalisessa mediassa todella pitkälle, joka luo omat haasteensa puolustuksen kehittämiselle.

Sybilien kehittymisen myötä tutkimus on viime vuosina alkanut kääntymään käyttäytymistutkimuksen puoleen ja nimenomaan sybilien käytöksen analysointiin. Tutkijat ovat löytäneet viitteitä samankaltaisista käyttäytymismalleista sybilien kesken sekä siitä, että sybilit tapaavat ryhmäytyä keskenään luoden yhdessä jopa omia keskinäisiä sybilyhteisöjään. Pelkän käyttäytymisen tutkimuksen lisäksi tiedeyhteisössä vaikuttaa olevan kiinnostusta myös hybridimalleille, joissa perinteiset puolustusmekanismit, kuten sosiaalinen graafi, valjastetaan yhteistyöhön käyttäytymistutkimuksen kautta luotujen menetelmien kanssa. Tuloksien perusteella tällä yhteistyöllä näyttäisi olevan positiivista vaikutusta puolustuksen tehoon nähden.

Suuret haasteet sybilhyökkäyksiltä puolustautumisessa sosiaalisen median alustoilla luo se, ettei palveluntarjoajat voi tai ole halukkaita erinäisistä syistä jakamaan tietoa sybilkäyttäjistään tutkijoiden kesken. Nyt mallidatana käytetään valtaosin sybilien käytöstä simuloivaa dataa, varsinkin suuren skaalan dataryhmien kohdalla. Aidoista sybileistä koostetun suuren skaalan dataryhmien saaminen mallidataksi auttaisi puolustuksen kehitystä, kun menetelmää voisi testata aidoilla sybileillä, eikä luottaa siihen, että ihmisen tekemä mallinnus on tarpeeksi tarkka.

## Lähdeluettelo

Abbas, Sohail. (2019). An Efficient Sybil Attack Detection for Internet of Things. *New Knowledge in Information Systems and Technologies*. Springer International Publishing. 339–349.

Al-Qurishi, M. et al. (2017). Sybil Defense Techniques in Online Social Networks: A Survey. *IEEE access*. [Online] 51200–1219.

Al-Qurishi, M. et al. (2018). SybilTrap: A graph-based semi-supervised Sybil defense scheme for online social networks. *Concurrency and computation.* [Online] 30 (5), e4276–n/a.

Alharbi, A. et al. (2021). Social Media Identity Deception Detection: A Survey. *ACM computing surveys*. [Online] 54 (3), 1–35.

Collins, B., Dinh, T. H., Ngoc, T. N., & Hwang, D. (2021). Trends in combating fake news on social media – a survey. *Journal of Information and Telecommunication, 5*(2), 247-266. *http://dx.doi.org/10.1080/24751839.2020.1847379*

Douceur, J. R. (2002). The Sybil Attack. *Peer-to-Peer Systems*. [Online]. 2002 Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 251–260.

Gao P., Gong N. Z., Kulkarni S., Thomas K., and Mittal P. (2015). Sybilframe: A defense-in-depth framework for structure-based Sybil detection. *arXiv preprint*. [Online]. Available: https://arxiv.org/abs/1503.02985

Gong N. Z., Frank M. and Mittal P. (2014). SybilBelief: A Semi-Supervised Learning Approach for Structure-Based Sybil Detection. *IEEE Transactions on Information Forensics and Security, vol. 9, no. 6*. pp. 976-987. doi: 10.1109/TIFS.2014.2316975.

Yu, H. et al. (2006). SybilGuard: Defending against sybil attacks via social networks. *Computer Communication Review*. [Online]. 2006 pp. 267–278.

Haifeng Yu et al. (2010). SybilLimit: A Near-Optimal Social Network Defense Against Sybil Attacks. *IEEE/ACM transactions on networking*. [Online] 18 (3), 885–898.

Hamid, A., Alam, M., Sheherin, H., & Pathan, A. K. (2020). Cyber security concerns in social networking service. *International Journal of Communication Networks and Information Security, 12*(2), 198-212.

Jia, J., Wang B. and Gong N. Z. (2017). Random Walk Based Fake Account Detection in Online Social Networks, *2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pp. 273-284, doi: 10.1109/DSN.2017.55.

Jiang, J., Shan, Z., Wang, X., Zhang, L. & Dai, Y. (2015). Understanding Sybil Groups in the Wild. *Journal of Computer Science and Technology,* vol. 30, no. 6, pp. 1344-1357.

Lobo, A. et al. (2021). Detection of Sybil Attacks in Social Networks. *Computational Data and Social Networks.* [Online]. Cham: Springer International Publishing. pp. 366–377.

Nitin Chiluka, Nazareno Andrade, Johan Pouwelse, and Henk Sips. (2012). Leveraging trust and distrust for sybil-tolerant voting in online social media. *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media* (*PSOSM '12*). Association for Computing Machinery, New York, NY, USA, Article 1, 1–8.

Shekokar N. M. & Kansara K. B. (2016). Security against sybil attack in social network. *2016 International Conference on Information Communication and Embedded Systems (ICICES)*, pp. 1-5. doi: 10.1109/ICICES.2016.7518887.

Tang, W., Ren, J., & Zhang, Y. (2019). Enabling Trusted and Privacy-Preserving Healthcare Services in Social Media Health Networks. *IEEE Transactions on Multimedia, 21*(3), 579-590. *http://dx.doi.org/10.1109/TMM.2018.2889934*

Thakur, K., Hayajneh, T., & Tseng, J. (2019). Cyber security in social media: Challenges and the way forward. *IT Professional Magazine.* 41-49. doi: *http://dx.doi.org/10.1109/MITP.2018.2881373*

Trifunovic, S. & Hossmann-Picu, A. (2016). Stalk and lie—The cost of Sybil attacks in opportunistic networks. *Computer communications*. [Online] 7366–79.

Wei Wei et al. (2013). SybilDefender: A Defense Mechanism for Sybil Attacks in Large Social Networks. *IEEE transactions on parallel and distributed systems*. [Online] 24 (12), 2492–2502