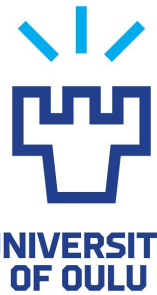


TIETO- JA SÄHKÖTEKNIIKAN TIEDEKUNTA

Ilkka Sovanto

FLOW-TIETOJEN INDEKSOINTI

Kandidaatintyö
Tietotekniikan tutkinto-ohjelma
Kesäkuu 2016

Sovanto I. (2016) Flow-tietojen indeksointi. Oulun yliopisto, tietotekniikan osasto. Tietotekniikan tutkielma, 23 s.

TIIVISTELMÄ

NetFlow sisältää tiivistetyssä muodossa verkkoliikennettä kuvaavaa historiallista tietoa, jota voidaan hyödyntää tietoturvaloukkaustapausten selvittämisessä ja havaitsemisessa. Tässä työssä esitetään katsaus indeksointimenetelmiin, joilla tyypillisimpiä tietoturvaan liittyviä flow-tietojen hakuja voidaan nopeuttaa merkittävästi. Näin mahdollistetaan hakeminen myös huomattavasti laajemmasta aineistosta.

Avainsanat: NetFlow, flow-tiedot, indeksointi, tietoturva

Sovanto I. (2016) Indexing flow records. Department of Computer Science and Engineering, University of Oulu, Oulu, Finland. Bachelor's thesis, 23 p.

ABSTRACT

NetFlow is a compact representation of observed network traffic. It may be leveraged in detecting and investigating computer network breaches. This work gives an overview of indexing methods which allow a significant speed-up of typical security related searches of flow records. Indexing enables searching over a much longer span of time than would otherwise be practical.

Keywords: NetFlow, flow records, indexing, information security

SISÄLLYSLUETTELO

LYHENTEET (ABBREVIATIONS)

DNS	Domain Name System, nimipalvelu
HTTP	Hypertext Transfer Protocol
IANA	Internet Assigned Numbers Authority
ICMP	Internet Control Message Protocol
IETF	Internet Engineering Task Force
IP	Internet Protocol
IPFIX	IP Flow Information Export
LAN	Local Area Network, lähiverkko
NAT	Network Address Translation, osoitteenmuunnos
OSI-malli	Open Systems Interconnection model
SNMP	Simple Network Management Protocol
SPAN	Switched Port Analyzer, peilausportti
TCP	Transmission Control Protocol
UDP	User Datagram Protocol

1. JOHDANTO

Tässä työssä käsitellään Cisco Systemsin kehittämän NetFlow-tiedon indeksointia [1], jotta flow-tietoja voidaan hyödyntää tehokkaasti tietoturvan työkaluna. Tutkielmassa esitetään, että flow-tietovarantojen indeksointi on välttämätöntä, koska muuten yhä suuremmiksi kasvavien tietomäärien kattava hyödyntäminen on käytännössä mahdotonta. Työn tavoitteena on esitellä menetelmiä, joiden avulla NetFlow-tietoihin voidaan kohdistaa tietoturvan kannalta olennaisia hakuja tehokkaasti. Työssä esitettyjä menetelmiä voidaan hyödyntää osana suurtenkin organisaatioiden tietoverkkojen suojaamista, kun kerätyn flow-tiedon määrä kasvaa teratavuluokkaan ja sitä suuremmaksi. Flow-tiedoista on hyötyä etenkin silloin, kun ollaan tekemisissä sellaisten haittaohjelmien kanssa, joita muut tietoturvakontrollit eivät havaitse.

Tutkielman ulkopuolelle on rajattu sellaiset flow-tietoihin perustuvat menetelmät, joiden avulla pyritään tunnistamaan verkkoliikenteestä mahdollisesti haitalliset yhteydet muuten kuin IP-osoitteiden perusteella.

> *"Thou shalt filter no more than a week of traffic at a time. The filter runs for excessive length of time otherwise."*
>
> *— SiLK Commandments [2]*

SiLK-työkalun ohjemateriaalissa kehotetaan rajaamaan haut korkeintaan viikon ajalle, jotta haut valmistuisivat järjellisessä ajassa. Indeksoinnin avulla tavoitellaan usean vuoden ajan kattavia nopeita hakuja.

1.1. NetFlow tietoturvaongelmien havaitsemisessa

NetFlowia voidaan käyttää tietoturvapoikkeamien havainnoinnissa ja selvittämisessä monella tavalla. Esimerkiksi palvelunestohyökkäys voidaan helposti todentaa ja selvittää siihen liittyvät lähdeosoitteet ja hyökkäysliikenteen tyyppi. Flow-tietojen keskeinen etu on siinä, että näistä tiedoista voidaan jälkikäteen tutkia verkkoliikennettä aikaisemmalta ajanhetkeltä. Tyypillisiä kiinnostavia ajankohtia ovat palvelunestohyökkäys tai hetki juuri ennen tietoturvaloukkauksen tapahtumista. [3]

Flow-tiedoista voidaan etsiä poikkeamia ja muita outoja ilmiöitä, jotka voivat kieliä tietoturvaongelmasta. Menetelmät poikkeamien tunnistamiseen voidaan karkeasti ottaen jakaa heuristisiin ja ennalta tiedettyihin tunnisteisiin perustuviin menetelmiin. Heuristiset tunnistusmenetelmät voivat kiinnittää huomiota esimerkiksi poikkeavan suuriin liikennemääriin, jotka saattavat olla merkki käynnissä olevasta laajamittaisesta tietojen anastamisesta [4]. Tiettyyn osoitteeseen kohdistuva säännöllinen mutta harvahko ja määrältään vähäinen liikenne saattaa olla peräisin haittaohjelmasta, joka tarkistaa komentopalvelimeltaan tietyin väliajoin, onko sille annettu tehtäviä. Heuristiset menetelmät haitallisen liikenteen tunnistamiseksi on kuitenkin rajattu tämän tutkielman ulkopuolelle.

Jos kohdeosoite tunnetaan ennakolta haittaohjelman käyttämäksi komentopalvelinosoitteeksi, on siihen kohdistuvan liikenteen havaitseminen flow-tietojen avulla helppoa. Saastuneiden koneiden tunnistaminen onnistuu lähdeosoitteen perusteella riippuen flow-tietojen keräyspisteestä ja mahdollisesti käytetyistä osoitteenmuunnoksista (NAT).

1.2. Flow-tiedot tietoturvapoikkeaman tutkinnassa

Tietoturvapoikkeaman havaitsemisen jälkeen tapausta ryhdytään tutkimaan tarkemmin. Jos kyseessä on haittaohjelman saastuttama työasema ja tiedossa on jo haittaohjelman käyttämän komentopalvelimen osoite, voidaan flow-tiedoista päätellä, milloin tartunta on tapahtunut liikenteen alkuajankohdan perusteella. Lisäksi voidaan takautuvasti tunnistaa muita samaan kohdeosoitteeseen liikennöiviä työasemia, jotka ovat todennäköisesti saman haittaohjelman saastuttamia. Tutkimalla edelleen näiden työasemien liikennöintiä voidaan tunnistaa muita saman haittaohjelman käyttämiä komentopalvelimia. Liikennemääristä voidaan tehdä karkean tason päätelmiä haittaohjelman mahdollisesti varastaman tiedon määrästä.

Vakava tietomurto, jossa yrityksen verkkoon ja palvelimille on tunkeuduttu, havaitaan usein kuukausien tai jopa usean vuoden kuluttua tunkeutumisesta. Tällöin flow-tiedot saattavat olla ainoa jäljellä oleva tietolähde, josta voidaan päätellä, mitä tunkeutumisen alkuhetkillä on verkossa tapahtunut.

Toinen tyypillinen tilanne on se, että ulkopuolisesta lähteestä saadaan tieto yhdestä tai useammasta IP-osoittesta, joita on käytetty haittaohjelman komentopalvelinosoitteina. Siltä varalta, että verkko on jo saastunut, on järkevää tutkia, onko omasta verkosta otettu yhteyttä kyseisiin osoitteisiin. Jos flow-aineistoa ei ole indeksoitu, joudutaan koko aineisto lukemaan lävitse. Tällainen haku voi kestää useita tunteja tai jopa päiviä, jos varastoitujen flow-tietojen määrä on suuri. Jos etsittyihin osoitteisiin todetaan olleen liikennettä, on arvioitava, voiko liikenne olla haittaohjelman aiheuttamaa ja tarvittaessa ryhdyttävä lisätutkimuksiin. On pidettävä mielessä, että sama IP-osoite on voinut olla myös muussa käytössä.

2. NETFLOW

NetFlow on Cisco Systemsin kehittämä tapa kuvata verkkoliikennettä [5]. NetFlow kuvaa verkkoliikenteestä tiivistetysti yhteenvetotietoja OSI-mallin kerroksilta 3 ja 4. Flow-tiedoista ilmenee käytetystä versiosta riippuen ainakin lähde- ja kohdeosoite, käytetyt portit ja protokolla, yhteyden kesto ja siirretty paketti- ja tavumäärä. Taulukossa 1 on esitetty flow-tietueessa esiintyvät keskeisimmät kentät [6]. Flow-tietoa voidaan lisäksi rikastaa yhdistämällä siihen muista lähteistä kerättyjä tietoja, kuten esimerkiksi nimipalvelutietoja tai ylempien verkkoprotokollien kontekstitietoja [7].

NetFlow kehitettiin alun perin verkon ylläpitäjien työkaluksi, jotta he saisivat paremman käsityksen verkossa kulkevan liikenteen luonteesta. Flow-tietojen avulla voi esimerkiksi saada nopeasti selville verkkoa kuormittavan liikenteen lähteen. Käyttökohteita on nykyisin runsaasti. Teleoperaattorit käyttävät flow-tietoja usein liikenteen määrään perustuvaan laskutukseen, verkon kuormituksen seuraamiseen ja vikatilanteiden selvittelyyn. NetFlow on havaittu arvokkaaksi työkaluksi verkon tietoturvan tukena.

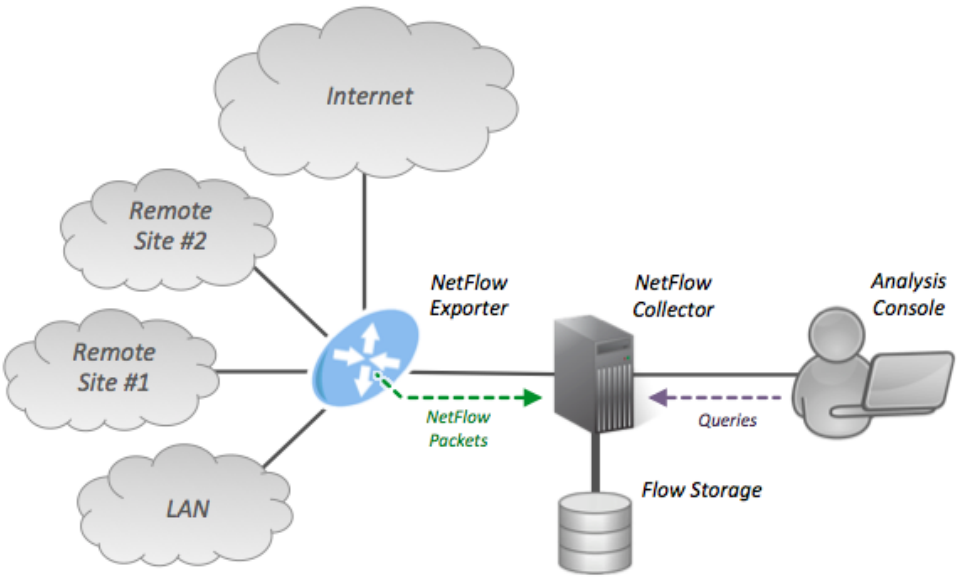
Vastaavat tiedot on suurimmaksi osaksi mahdollista kerätä myös muilla menetelmillä, kuten palomuurilokeista sekä reititinten ja kytkinporttien liikennemäärälaskureista tai viime kädessä verkkoliikenteen täydellisellä kaappauksella. Flow-tietoja vastaavaan tarkkuuteen tai järjestelyn keveyteen on kuitenkin vaikea päästä muilla menetelmillä. Hyvin käyttökelpoinen vaihtoehto on pakotetun välityspalvelimen (proxy) lokitiedot, jotka sisältävät enemmän tietoja kuin flow-tietueet, mutta rajoitteena tietysti on kattavuus ainoastaan HTTP-liikenteeseen.

Taulukko 1. NetFlow v5 -tietueen keskeisimmät kentät

Tavut	Nimi	Kuvaus
0-3	srcaddr	Lähteen IP-osoite
4-7	dstaddr	Kohteen IP-osoite
12-13	input	Sisääntuloverkkoliitännän SNMP-index
14-15	output	Lähtöverkkoliitännän SNMP-index
16-19	dPkts	Pakettien määrä flow:ssa
20-23	dOctets	Tavujen määrä flow:ssa
24-27	First	Flow:n alkuajankohta
28-31	Last	Viimeisimmän havaitun paketin ajankohta
32-33	srcport	TCP/UDP-lähdeportti
34-35	dstport	TCP/UDP-kohdeportti
37	tcp_flags	TCP-yhteyden havaitut liput
38	prot	IP-protokolla (esim. TCP, UDP, ICMP)

Flow-tietoa kerätään tyypillisesti verkon reunareitittimeltä, jolloin saadaan tietoa organisaation verkosta lähtevästä ja sinne saapuvasta liikenteestä. Kerätessään flow-tietoja reititin pitää kirjaa aktiivisista yhteyksistä välitettyjen IP-pakettien otsaketietojen perusteella. Pakettien sisältöä ei tutkita. Tietyin väliajoin reitin lähettää flow-tiedot koosteena NetFlow-keräimelle, joka tallentaa tiedot myöhempää käyttöä varten. Tyypillinen keräysjärjestely on esitetty kuvassa 1 [8]. Flow-tietoja voidaan kerätä verkon eri pisteistä käyttötarkoituksesta riippuen. Reitittimen sijaan flow-tieto voidaan tuottaa erillisellä komponentilla, jolle välitetään kopio halutusta liikenteestä (kuva 2 [9]).

Flow-tiedot voidaan käsittää eräänlaisena havaitun verkkoliikenteen häviöllisenä pakkauksena, jossa tallennetaan vain keskeisimmät liikennettä kuvaavat tiedot. Olennaisesti pienenty-



Kuva 1. Tyypillinen NetFlow-keräysjärjestely.

neen tietomäärän ansiosta flow-tietoja on mahdollista tallentaa myöhempää käyttöä varten jopa usean vuoden ajalta.

Verkkoliikennettä on mahdollista profiloida hyvin monella tavalla flow-tietojen avulla. Perinteisiä käyttötarkoituksia ovat erilaiset tilastointitarkoitukset, verkon suorituskyvyn seuranta, vianetsintä ja liikenteen määrään perustuva laskutus.

Tyypillisimmät käytetyt NetFlow-versiot ovat 5, 9 ja IPFIX [10], joka on Ciscon NetFlow-versioon 9 pohjautuva IETF-standardi. NetFlow-versio 9 ja IPFIX tukevat IPv4:n lisäksi IPv6-osoitteita sekä organisaatiokohtaisten laajennuskenttien käytön mallineen (record template) avulla. Mallineessa määritellään flow-tietueisiin kerättävät tietokentät, jotka voidaan valikoida käyttötarkoitukseen nähden sopivalla tavalla.

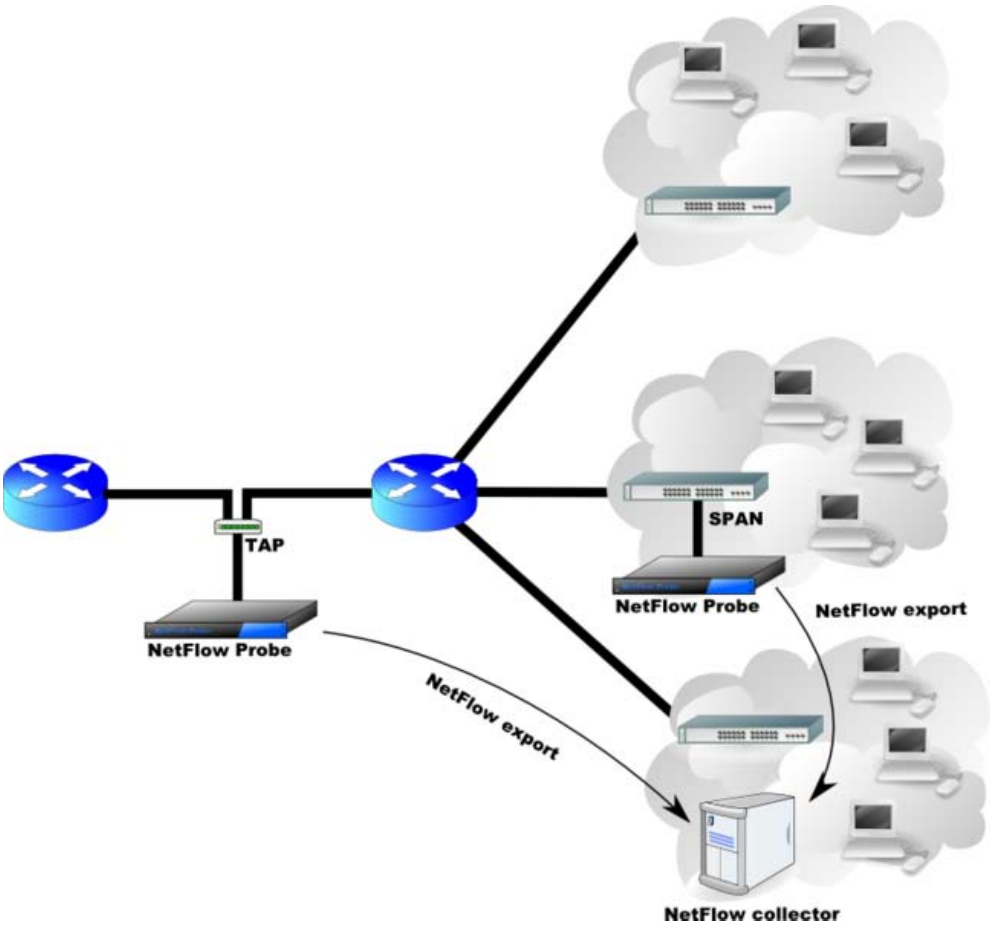
Flow-tietojen käytön kannalta valitulla protokollaversiolla ei ole kovinkaan suurta merkitystä, koska flow-keräin yleensä muuntaa ja tallentaa tiedot omassa muodossaan myöhempää analysointia varten.

2.1. Työkalut

Flow-tietojen keräykseen ja käsittelyyn on olemassa useita työkalukokoelmia, joista tunnetuimpia on esitelty alla. Työkalut sisältävät yleensä varsin kattavat mahdollisuudet flow-tietojen käsittelyyn, analysointiin ja raportointiin, mutta indeksoinnin puuttuessa haettavan aineiston määrää usein joudutaan rajoittamaan.

2.1.1. Flow-tools

Flow-tools on vanhimmasta päästä oleva työkalukokoelma ja -kirjasto flow-tietojen keräämiseen, käsittelyyn ja raportointiin [11]. Flow-toolsin suosio on vähentynyt uudempien tulokkaiden myötä.



Kuva 2. NetFlow-keräysjärjestely erillislaitteiden avulla.

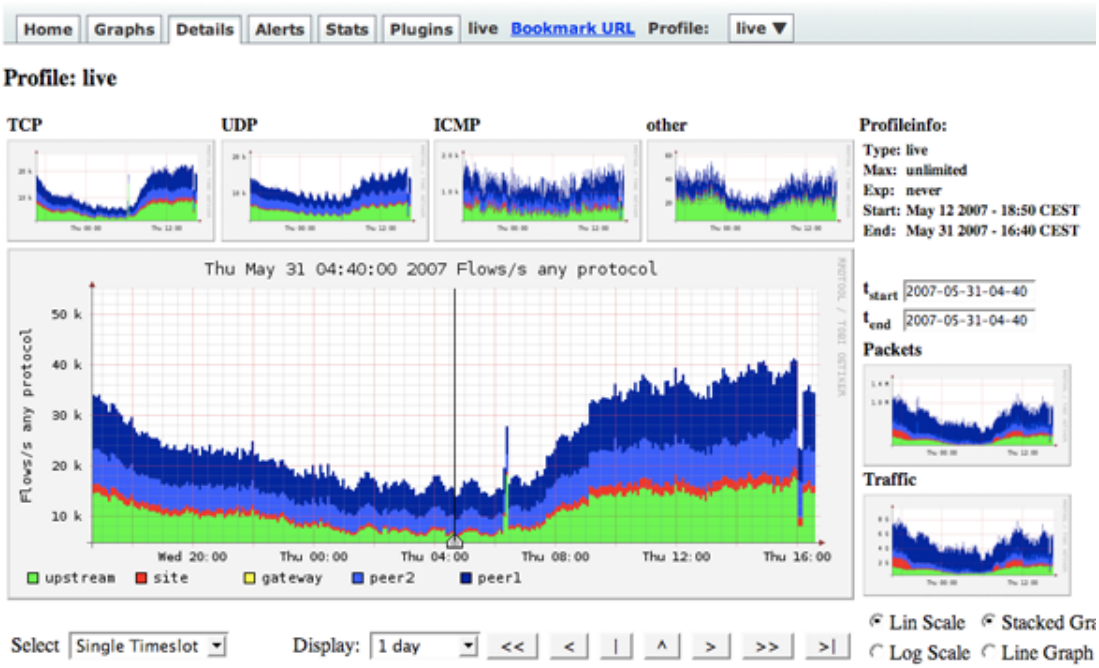
2.1.2. Nfdump ja Nfsen

Nfdump ja Nfsen ovat monipuolinen kokoelma työkaluja, jotka vastaanottavat flow-tietoja keräimeltä ja tallentavat ne levylle. Nfsen piirtää flow-tiedoista havainnollistavia kuvaajia (kuva 3). Nfsenin avulla on myös mahdollista tehdä selainkäyttöliittymästä hakuja ja tilastoja tallennetuista flow-tiedoista, mutta käytännössä nämä on rajoitettava lyhyehkölle aikavälille, koska indeksointia ei käytetä.

2.1.3. SiLK

SiLK on Carnegie Mellon -yliopiston Software Engineering Instituten tuottama työkalukokoelma, joka vastaanottaa ja tallentaa flow-tietoja hyvin tiiviissä muodossa [12]. Monipuoliset työkalut tarjoavat runsaasti vaihtoehtoja flow-tietojen analysointiin, suodattamiseen, etsimiseen ja yhteenvetojen tekemiseen. Komentorivityökalut ja ohjelmalliset kirjastorajapinnat tarjoavat hyvät mahdollisuudet ohjelmalliseen käsittelyyn ja skriptien tekemiseen.

SiLK ei itsessään sisällä menetelmiä flow-tietojen indeksointiin, ja sen koulutusmateriaalissa kehotetaankin rajoittamaan haut siten, että haettava aikaväli on korkeintaan viikko kerrallaan, jotta hakuajat pysyisivät kohtuullisina.



Kuva 3. Nfsen piirtää kuvaajia flow-tietojen perusteella liikennemäärästä tavujen, pakettien tai yhteysmäärän (flows) perusteella. Liikennettä voidaan lisäksi eritellä protokollan (TCP / UDP / ICMP / muut) tai havainnointipisteen mukaan. Kuvaajista on helppo havaita liikennepiikkejä tai ongelmatilantaista kieliviä pudotuksia.

3. FLOW-TIETOJEN INDEKSOINTI

Flow-tietoja on järkevää indeksoida, jotta niihin voidaan tehdä nopeita hakuja. Tarkoituksena on minimoida haun toteuttamiseksi luettavan datan määrä. Indeksoinnin avulla voidaan välttää lukemasta aineistoa, jossa etsittävä hakutermi ei esiinny. Varsin usein etsittävää IP-osoitetta ei esiinny aineistossa lainkaan, jolloin sopivalla indeksoinnilla kielteinen vastaus voidaan saada erittäin nopeasti.

Indeksin avulla tehtävässä haussa annetaan hakuparametreina yksi tai useampi etsittävä IP-osoite, haluttu aikaväli ja flow-keräimet, joiden tietoihin haku kohdistetaan. Tuloksia voidaan rajata tarvittaessa myös käytetyn protokollan ja portin perusteella, tai määritellä tietty kynnysarvo havaitulle liikennemäärälle.

Indeksointimenetelmille on yhteistä, että ne ottavat syötteekseen joukon flow-tietoja sisältäviä tiedostoja, joista kerätään niissä esiintyvät IP-osoitteet hakuindeksiin. Tiedot voidaan jaotella useisiin eri indekseihin aikavälin tai tiedot tuottaneen keräimen perusteella. Indeksiin liitetään usein tehokkaalla tietorakenteella toteutettu hakutermilista, jonka perusteella voidaan nopeasti todeta, esiintyykö haettava IP-osoite indeksin viittaamassa aineistossa lainkaan. Tätä voidaan entisestään tehostaa käyttämällä bittikarttaa tai Bloom Filteriä [13], jonka avulla on mahdollista todeta erittäin nopeasti, että hakutermi ei esiinny indeksissä [14]. Lopulliseen indeksiin tallennetaan yleensä viittaus alkuperäiseen flow-aineistoon, jolloin hakutuloksena saadut flow-tiedot saadaan esitettyä kokonaisuudessaan alkuperäisessä muodossa.

Indeksit voidaan muodostaa esimerkiksi päiväkohtaisesti ja kullekin keräimelle erikseen. Näin ollen haut voidaan kohdistaa halutulle aikavälille ja keräimille valitsemalla haettavaksi vain näihin liittyvät indeksit. Indeksejä on mahdollista järjestellä myös muilla tavoin riippuen muun muassa siitä, kuinka alkuperäinen flow-aineisto on jaoteltu.

Jos indeksien määrä tällä tavalla muodostettuna kasvaa suureksi ja on tarve tehostaa pitkän aikavälin ja suuren määrän keräimiä kattavia hakuja, voidaan indeksejä yhdistämällä muodostaa yhdistelmäindeksejä, jotka kattavat esimerkiksi kuukauden tai vuoden kerrallaan. Tällaisista yhdistelmäindekseistä on hyötyä varsinkin, jos halutaan tutkia, esiintyykö tiettyihin osoitteisiin liittyvää liikennöintiä missään kohtaa indeksoitua flow-aineistoa. Yhdistelmäindeksejä on mahdollista edelleen yhdistämällä muodostaa hakuhierarkia, jonka avulla voidaan minimoida haussa tarvittavien indeksien määrä ja siten haun tarvitsema kokonaisaika.

Indeksit voidaan käsittää häviöllisenä pakkauksena, jossa alkuperäisestä aineistosta säilytetään ainoastaan tarvittavat osat. Pienimmillään indeksin voi muodostaa syöteaineistossa esiintyneiden IP-osoitteiden joukko tai jopa pelkkä verkko-osoite vaikkapa /24-verkkoalueen tarkkuudella.

Toisaalta indeksiin voidaan sisällyttää enemmänkin tietoa yhdistämällä havaintoihin liittyviä tunnuslukuja. Tällöin pelkän hakuindeksin lisäksi saadaan valmiiksi kerättyä tilastotietoa osoitteeseen liittyvistä havainnoista. Käyttötarkoituksesta riippuen indeksin yhteyteen voidaan kerätä ensimmäisen ja viimeisen havainnon aikaleimat, havaittujen yhteyksien määrä ja kokonaisliikennemäärä tai muuta tietoa alkuperäisistä flow-tietueista.

3.1. Netflow-indexer

Netflow-indexer on avoimen lähdekoodin projekti [15], joka pohjautuu avoimeen Xapian-indeksiin. Netflow-indexer muodostaa kullekin päivälle oman indeksinsä, josta voidaan nopeasti tarkistaa etsityn osoitteen esiintyminen kyseisenä päivänä. Löytyneen viitetiedon perusteella voidaan myös hakea alkuperäiset flow-tiedot ja tulostaa ne näytölle. Jos käytössä on useampia Netflow-keräimiä, on tietojen hallinnoinnin kannalta järkevää säilyttää kunkin keräimen tuottamia tietoja omissa hakemistoissaan. Vastaavasti näiden pohjalta luodut indeksit

voidaan pitää toisistaan erillään, jolloin hakuja voidaan tarvittaessa kohdistaa vain tiettyihin keräimiin.

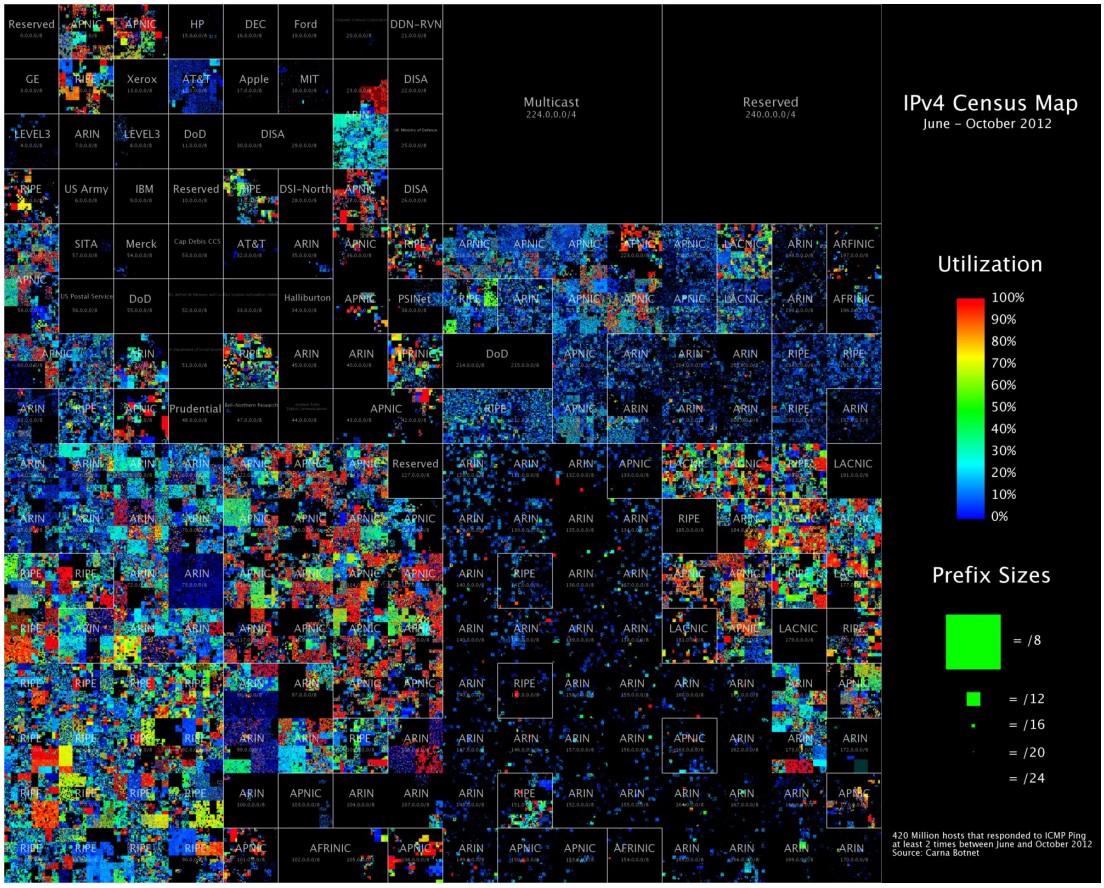
Netflow-indexerin tuottamien indeksien koko on tyypillisesti noin 5-10% alkuperäisestä aineistosta [16]. Pitkän ajanjakson yli hakeminen voi kestää useita minuutteja, kun haettavia päiväkohtaisia indeksejä on runsaasti. Xapian-kirjasto sisältää kuitenkin xapian-compact -työkalun, jolla on mahdollista yhdistää useita indeksejä yhdistelmäindeksiksi. Netflow-indexerin hakutyökaluja voidaan tämän jälkeen käyttää myös näiden yhdistelmäindeksien kanssa.

Netflow-indexer tukee Nfdumpin ja flow-toolsin tallentamien flow-tietojen indeksointia ja hakemista, mutta tämän tutkielman yhteydessä on toteutettu myös alustava tuki SiLK:n tallentamien tietojen indeksointiin. Modulaarisen rakenteen ansiosta tuen rakentaminen oli suoraviivaista.

3.2. Bittikarttasuodatus

Kielteinen hakutulos saadaan nopeimmin, jos heti alkuun voidaan todeta, ettei hakutuloksia ole luvassa. Työn yhteydessä toteutettiin yksinkertainen bittikarttasuodatin, jossa koko 32-bittinen IPv4-osoiteavaruus kuvataan lineaarisena bittitaulukkona. Taulukon alkion numero vastaa IP-osoitetta ja bitin arvo sitä, esiintyykö siihen liittyvä IPv4-osoite indeksoidussa aineistossa kertaakaan. Yhteen tavuun voidaan pakata kahdeksan bittiä, jolloin taulukon kooksi tulee puoli gigatavua. Lineaarisen tietorakenteen ansiosta haut onnistuvat O(1)-ajassa. Useita hakutermejä etsittäessä haku voidaan bittikarttasuodatuksen avulla rajata ainoastaan niihin termeihin, joilla on odotettavissa osumia.

Bittikartta ei sellaisenaan sovellu IPv6-osoiteavaruuden kuvaamiseen sen merkittävästi suuremman koon vuoksi. Kompromissina voisi toimia IPv6-osoitteiden kuvaaminen Bloom Filterillä tai /48-verkkoalueen tarkkuuteen rajoittuminen. Näiden verkkoalueiden kuvaaminen olisi mahdollista muutamalla 2^{32}-bittisellä bittikartalla, sillä tällä hetkellä suuri osa käytetyistä IPv6-osoitteista on 2001:- tai 2a00:-alkuisia [17]. Myös IPv4-avaruuteen jää alueita, jotka eivät ole käytössä (kuva 4 [18]).



Kuva 4. IPv4-osoiteavaruuden käyttöaste jakautuu epätasaisesti.

4. ESIMERKKIHAKU JA SUORITUSKYKYMITTAUKSET

Indeksoinnin tuomaa suorituskykyhyötyä mitattiin vertaamalla hakuaikoja ilman indeksointia tehtyyn hakuun. Koeaineistona käytettiin palvelinympäristöstä kerättyä 27 gigatavun kokoista flow-aineistoa reilun kuukauden ajalta. Hakuaikojen muutosten selvittämiseksi haku toistettiin eripituisilla ajanjaksoilla. Toisena koeaineistona käytettiin laajempaa flow-aineistoa, josta ilmenee myös indeksoitujen hakuaikojen piteneminen kuvan 7 mukaisesti, kun haettavia päiväindeksejä on runsaasti.

Hakuesimerkissä etsittiin tietoturvayhtiö F-Securen blogissa [19] julkaistuja osoitteita, joita on käytetty kehittyneen haittaohjelmakampanjan yhteydessä. Blogissa mainitut DNS-nimet selvitettiin taulukkoon 2 IP-osoitteiksi hakua varten. Haun avulla selvitetään, onko verkosta ollut liikennettä etsittäviin osoitteisiin. Positiivinen tulos viittaisi mahdolliseen haittaohjelmatartuntaan ja antaisi aihetta jatkotutkimuksiin. Saatu hakutulos osoittaa, että liikennettä ei ole ollut aineiston kattamalla ajanjaksolla.

Haut toistettiin kolmella eri menetelmällä. Ensiksi haku tehtiin nfdump-työkalulla, jolloin indeksointia ei hyödynnetty lainkaan ja koko haettavan aikavälin aineisto oli käsiteltävä. Seuraavaksi haku toistettiin netflow-indexerin avulla, jolloin hakua varten riitti tutkia aikavälin kattavat päiväindeksit. Indeksoinnin tuoma hyöty näkyy selkeästi kuvassa 5. Viimeisenä hakua tehostettiin ylempänä kuvatulla bittikarttasuodattimella, jonka avulla voitiin vakioajassa todeta, ettei hakutuloksia tulisi ja varsinainen haku voitiin välttää kokonaan.

Hakujen ensimmäisellä toistolla käyttöjärjestelmän tiedostovälimuistin vaikutuksen vuoksi hakuaika on pidempi kuin sitä seuraavilla toistoilla. Kuvassa 6 näkyvä välimuistin vaikutus on huomioitu mittauksissa toistamalla haut useampaan kertaan. Indeksoimattomalla haulla välimuistin vaikutus on vähäisempi, koska luettava aineisto ei mahdu kokonaisuudessaan välimuistiin.

Kuvassa 8 verrataan hakutermien määrän vaikutusta hakuaikoihin sekä netflow-indexerillä että bittikarttasuodatusta käyttämällä. Hakutermeinä käytettiin satunnaisesti muodostettuja IPv4-osoitteita. Koska sattumanvaraisia osoitteita ei joitakin poikkeuksia lukuun ottamatta esiintynyt haettavassa aineistossa, bittikarttasuodatus poisti lähes kokonaan tarpeen käydä läpi netflow-indexerin muodostamat indeksit. Niissäkin tapauksissa, joissa hakutermeissä oli bittikarttasuodatuksen läpäiseviä osoitteita, tarvitsi indekseistä etsiä ainoastaan yksittäisiä osoitteita kaikkien hakutermien sijaan.

Taulukko 2. Tarkistettavat IP-osoitteet

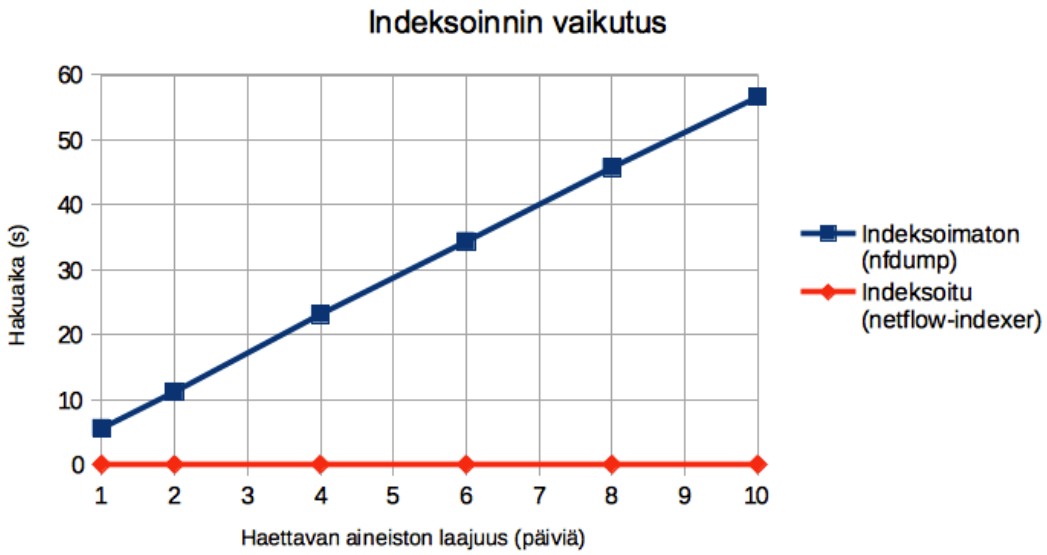
DNS-nimi	IP-osoite
cognimuse.cs.ntua[.]gr	147.102.10.1
portal.sbn.co[.]th	115.178.58.19
flockfilmseries[.]com	198.246.200.97
www.recordsmanagementservices[.]com	12.180.240.37
files.counseling[.]org	38.111.140.188
-	58.80.109.59
-	97.75.120.45

Esimerkkihaku nfdumpin avulla:

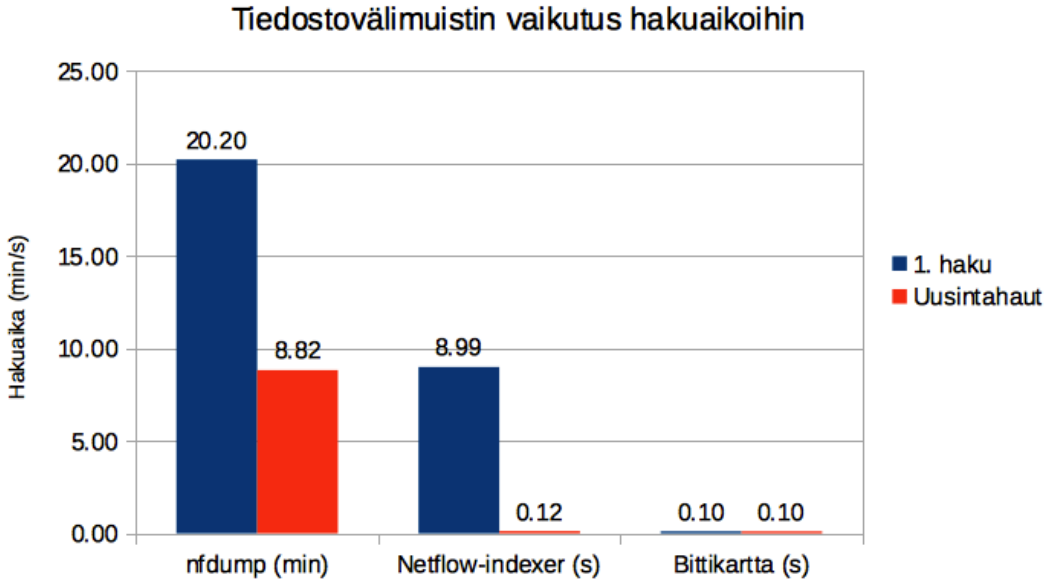
```
nfdump -R . -B "host 97.75.120.45 or host 58.80.109.59 or \
 host 147.102.10.1 or host 115.178.58.19 or host 198.246.200.97 \
 or host 12.180.240.37 or host 38.111.140.188"
```

Esimerkkihaku netflow-indexerin avulla:

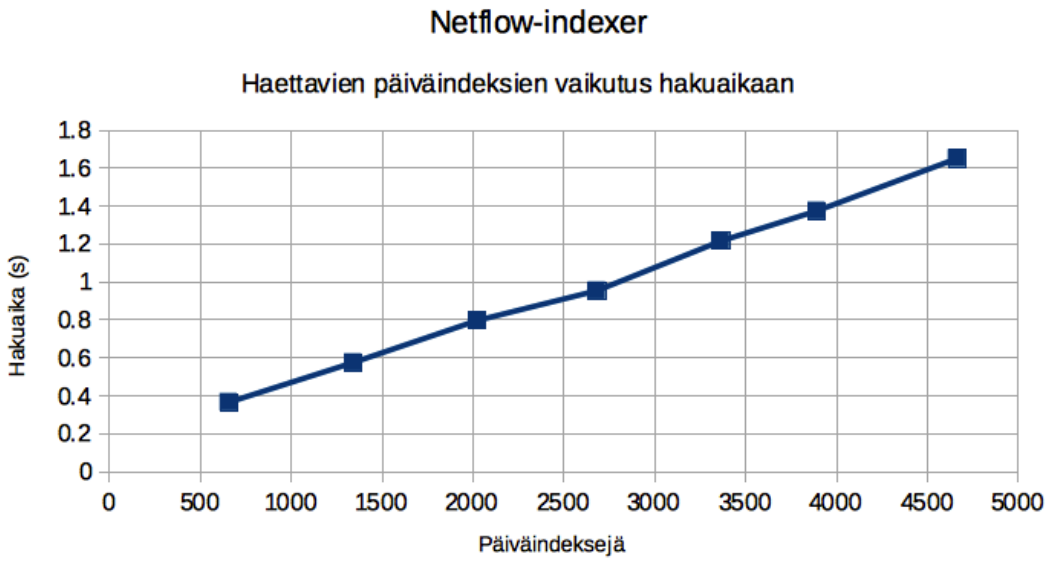
```
netflow-index-search-all -d nfdump.ini 97.75.120.45 58.80.109.59 \
 147.102.10.1 115.178.58.19 198.246.200.97 12.180.240.37 \
 38.111.140.188
```



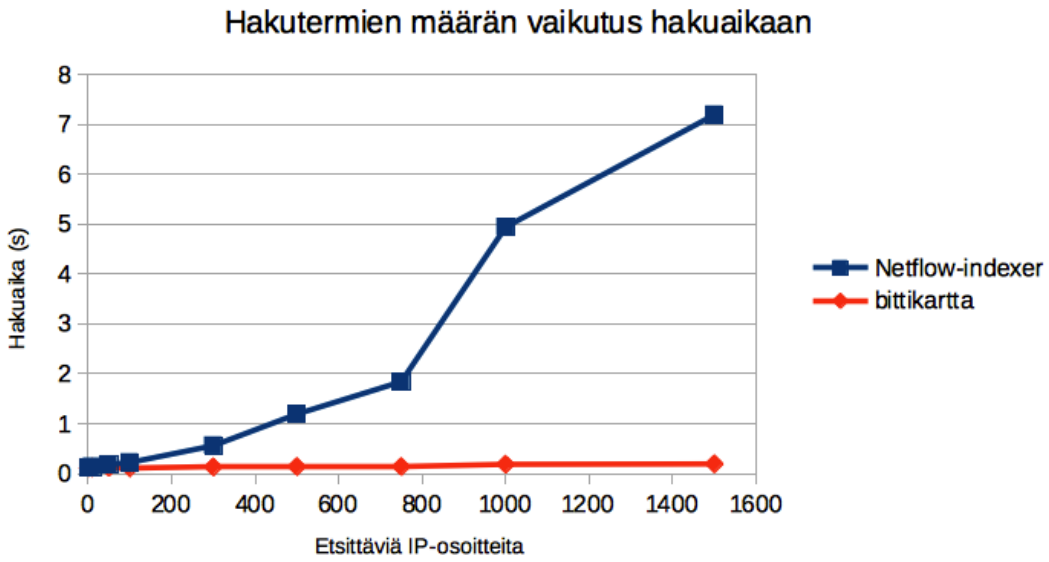
Kuva 5. Indeksoinnin avulla hakuaika pienenee ratkaisevasti. Indeksihakuaika on noin 0,09 sekuntia.



Kuva 6. Käyttöjärjestelmän tiedostovälimuistin ansiosta hitaat levyoperaatiot näkyvät vähemmän uusintahauissa.



Kuva 7. Indeksoidut haut hidastuvat etsittävien indeksien määrän kasvaessa.



Kuva 8. Hakutermien määrän lisääntyminen hidastaa hakua indeksistä.

5. POHDINTAA

Indeksoinnin avulla IP-osoitteen perusteella tehtävät haut nopeutuvat ratkaisevasti. Indekseistä on erityistä hyötyä silloin, kun haun kohteena oleva ajanjakso on esimerkiksi tuntia pidempi, jopa kuukausia tai vuosia. IP-osoitteeseen perustuvat indeksit soveltuvat erityisesti tilanteisiin, joissa kiinnostuksen kohteena olevat osoitteet ovat tiedossa, mikä on yleinen tapaus tietoturvaloukkauksia tutkittaessa tai niitä etsittäessä. Muun tyyppiset haut eivät välttämättä hyödy tällaisista indekseistä, mikäli hakutermeinä ei voida käyttää IP-osoitteita. Sellaisiakin hakuja voi kuitenkin olla mahdollista nopeuttaa merkittävästi järjestämällä toisenlainen indeksi tarkoitukseen sopivalla tavalla tai esikäsittelemällä flow-tiedot tarkoitusta varten siten, ettei alkuperäisiin tietueisiin ole tarvetta palata. Monissa tapauksissa taas indeksiä ei välttämättä tarvita lainkaan, jos esimerkiksi tarvittava aikaväli on lyhyt ja hyvin tiedossa.

5.1. Näytteistys

Flow-tietoja on mahdollista kerätä myös näytteistämällä tilastollisesti esimerkiksi joka kymmenes tai tuhannes paketti, jolloin jokaista pakettia ei huomioida flow-tietoja muodostettaessa [20]. Tämä vähentää kuormitusta flow-tietoja muodostettaessa sekä niiden vaatimaa tallennustilaa. Näytteistys voi olla laitteiston asettamien rajoitusten vuoksi välttämätöntä, jos flow-tiedot muodostetaan reitittimellä, minkä vuoksi erillislaite voi olla parempi vaihtoehto. Tilansäästö syntyy lähinnä hyvin lyhyiden tai vähän paketteja sisältävien yhteyksien jäädessä näytteistyksen satunnaisotannan ulkopuolelle. Esimerkiksi suurin osa porttiskannauksen aiheuttamista yhteyksistä jää näytteistyksen vuoksi tallentamatta.

Näytteistetyn keräämisen huonona puolena on se, että myös tietoturvan kannalta merkittäviä tapahtumia voi jäädä flow-aineiston ulkopuolelle. Haittaohjelmat voivat liikennöidä komentopalvelimen kanssa hyvin harvoin tai lähettää vain yksittäisiä paketteja yhteydenottoa kohden. Näytteistäminen voi haitata tällaisten heikkojen signaalien havaitsemista puutteellisen tiedon vuoksi. Näytteistämällä menetetään myös mahdollisuus todeta varmuudella, onko jokin yhteydenotto tapahtunut vai ei.

5.2. Flow-tietojen suodatus ja taustakohina

Turhan flow-tiedon määrää on mahdollista kuitenkin vähentää sijoittamalla keräyspiste esimerkiksi ulkoreunan palomuurin sisäpuolelle, jolloin suurin osa porttiskannauksista ja muusta internetin taustakohinasta pysähtyy palomuuriin eikä päädy flow-tietoihin.

Flow-tietoja voi olla muutenkin hyödyllistä kerätä eri pisteistä kattavamman näkyvyyden saamiseksi esimerkiksi verkon sisäiseen liikenteeseen. Tavallisesta poikkeava liikennöinti, vaikkapa kahden palvelimen välillä, joiden ei normaalisti pitäisi keskustella keskenään, voi olla merkki tietoturvapoikkeamasta. Toisaalta sisäverkon osoitteiden näkyminen voi huomattavasti helpottaa saastuneen työaseman tunnistamista, jos ulkoreunalta kerättynä flow-tiedoissa näkyisi vain yhdyskäytävän tai välityspalvelimen osoite. Tämä on huomioitava myös arvioitaessa flow-tietojen keräämisen ja käsittelyn vaikutuksia verkon käyttäjien tietosuojaan ja yksityisyyteen.

Palomuurin ja keräyspisteen sijoittamisen ohella flow-aineiston ja indeksien kokoa voidaan jossain määrin rajoittaa myös suodattamalla pois liikennettä, jolla ei todennäköisesti ole merkitystä tietoturvan kannalta. Tällaista liikennettä voi olla esimerkiksi porttiskannaukset tai muualla internetissä tapahtuvista palvelunestohyökkäyksistä syntyvä takaisinsirontaliikenne, jota

syntyy hyökkäyksen kohteen vastaillessa väärennetyillä lähdeosoitteilla lähetettyihin paketteihin.

Suodatus voidaan tehdä esimerkiksi rajoittumalla ainoastaan onnistuneesti muodostuneisiin TCP-yhteyksiin ja muiden protokollien osalta tarkastelemalla ainoastaan sellaista liikennettä, joka sisältää myös verkosta ulospäin suuntautuvaa liikennettä. Yksinkertaisimmillaan tähän voidaan indeksoinnin osalta päästä sisällyttämällä indeksiin ainoastaan flow-tietueiden kohdeosoitteet sekä tarkastelemalla TCP-yhteyksien lippuja esimerkiksi yksittäisestä TCP reset -vastauksesta koostuvan tietueen poissulkemiseksi.

5.3. Tietojen säilytys

Flow-tietoja on järkevää pyrkiä säilyttämään usean vuoden ajalta, koska vakavan haittaohjelmatartunnan ja sen havaitsemisen välillä on usein kuukausia tai jopa vuosia[21]. Jos liikennettä on runsaasti ja tallennustilaa ei voida järjestää riittävästi, voi vanhempaa aineistoa siirtää edullisemmalle tallennusvälineelle. Pelkistä indekseistäkin voi olla merkittävää hyötyä esimerkiksi tiettyyn osoitteeseen alkaneen liikennöinnin alkuajankohdan määrittämisessä, vaikka alkuperäinen flow-aineisto olisikin jo hävitetty.

Flow-tiedot ovat liikennettä kuvaavaa metatietoa, jonka käsittelyssä on huomioitava yksityisyydensuojaan liittyvät seikat. Flow-tiedot ovat välitystietoja, joiden käsittelyä Suomessa säännellään tietoyhteiskuntakaaressa [22]. Flow-tietojen sopivaa säilytysaikaa on harkittava sekä tietoturvapoikkeamien tutkinnan turvaamisen että yksityisyydensuojan kannalta.

5.4. Kehitysajatuksia

Tutkielman edetessä syntyi useita ajatuksia siitä, kuinka indeksointia voitaisiin kehittää edelleen nopeampien ja monipuolisempien hakujen mahdollistamiseksi. Joitakin näistä ehdittiin jo toteuttaa tai ainakin kokeilla.

Bittikarttojen laajempi hyödyntäminen vaikuttaa järkevältä seuraavalta kehitysaskeleelta. Bittikarttasuodattimia voidaan järjestää hierarkkisesti siten, että kielteisen hakutuloksen optimoinnin lisäksi saadaan tarkempaa tietoa siitä, missä osassa aineistoa hakutuloksia olisi saatavilla. Hierarkiatasoiksi sopisi esimerkiksi keräimen ja vuoden perusteella jaottelu, mutta myös esimerkiksi kuukausitasolle on mahdollista mennä. Ratkaisuna tämä on samantyyppinen kuin päiväkohtaisten indeksien kokoaminen yhdistelmäindekseiksi, ja näitä kannattaisikin vertailla tilankäytön ja suorituskyvyn suhteen.

Bittikarttasuodattimia kasattaessa hierarkiaksi niiden lukumäärä kasvaa, jolloin ne vaativat myös enemmän tallennustilaa. Tilaa voi säästää käyttämättä pakattuja bittikarttoja, kuten Roaring Bitmap -tekniikkaa [23], jolloin vältytään tallentamasta bittikartan käyttämättöämiä osia, jotka sisältävät vain nollabittejä.

Bloom filterit muistuttavat läheisesti bittikarttoja, mutta ne eivät ole samalla tavalla täsmällisiä kuin bittikartat. Toisaalta niiden avulla IPv6-osoitteita voidaan käsitellä samoin kuin IPv4-osoitteita.

Indeksoinnin rajoittaminen vain yhteyksien kohdeosoitteisiin vähentäisi vähemmän kiinnostavia hakutuloksia sekä pienentäisi indeksien vaatimaa tilaa, mahdollisesti myös nopeuttaen hakuja.

Flow-tietojen tapaan verkosta on mahdollista kerätä DNS-kyselyhistoriaa, josta selviää mitä DNS-nimiä kulloinkin on pyydetty ja mihin IP-osoitteisiin ne viittasivat. Historiatietoa voidaan hyödntää flow-tietojen tapaan, kun selvitetään onko johonkin haitalliseksi tiedettyyn DNS-nimeen liittyen tehty kyselyitä. Toisaalta pelkkä DNS-kysely ei tarkoita, että selvitettyyn

osoitteeseen olisi välttämätt myös liikennetty, minkä taas voi tarkistaa flow-tiedoista. DNS-kyselyhistorian yhdistäminen flow-tietoihin toisi lukuisia kiintoisia mahdollisuuksia etenkin tilanteisiin, joissa samaan IP-osoitteeseen liittyy lukuisia DNS-nimiä tai dynaamisten DNS-nimien vaihdellessa niihin liittyvää IP-osoitetta. Flow-tietohakujen tekeminen DNS-nimen perusteella tulisi myös mahdolliseksi. Historiatiedon automaattisella ristiinvertaamisella ja yhdistämisellä voitaisiin flow-haun tulosten yhdeytessä näyttää osoitteeseen havaintohetkellä liittynyt DNS-nimi.

6. YHTEENVETO

Tutkielmassa tarkasteltiin NetFlow-aineiston indeksointia, kun flow-tietoja hyödynnetään tietoturvan työkaluna. Tavoitteena oli osoittaa indeksoinnin tarpeellisuus sekä tarjota käytännöllisiä tapoja sen toteuttamiseen. Samalla tuotiin esille, miksi flow-tietojen kerääminen on järkevää organisaation tietoverkon suojaamiseksi.

Työssä luotiin katsaus erilaisiin indeksointimenetelmiin ja kokeiltiin käytännössä netflow-indexeriä, johon myös toteutettiin pieniä parannuksia. Mittausten perusteella indeksoinnin hyödyt ovat kiistattomat.

Laajoista flow-aineistoista etsittäessä indeksit ovat välttämättömiä, koska muuten haut eivät valmistu järkevässä ajassa. Mittausten perusteella on ilmeistä, että suoraviivaisillakin indeksointimenetelmillä on mahdollista päästä ratkaisevaan nopeutukseen. Hakutermien esisuodattamisella bittikartan tai bloom filterin avulla voidaan saada merkittävää lisänopeutta niissä tilanteissa, jossa haettava aineisto on laaja ja osa tai mikään hakutermeistä ei tuota hakutuloksia. Hierarkkisella indeksien indeksi -tietorakenteella voidaan edelleen nopeuttaa hakuja suuresta joukosta yksittäisiä päiväindeksejä.

7. LÄHDELUETTELO

[1] Fusco F., Stoecklin M.P. & Vlachos M. (2010) Net-fli: On-the-fly compression, archiving and indexing of streaming network traffic. Proceedings of the VLDB Endowment 3, s. 1382–1393.

[2] Bandes R. (2013) Network analysis with silk. In: FloCon 2013 Proceedings, s. 109.

[3] Scheck M. (2009), Netflow for incident detection. URL: http://www.first.org/global/practices/Netflow.pdf.

[4] Rashid A., Ramdhany R., Edwards M., Kibirige S.M., Babar A., Hutchison D. & Chitchyan R. (2014), Detecting and preventing data exfiltration. URL: https://www.cpni.gov.uk/Documents/Publications/2014/2014-04-11-de_lancaster_technical_report.pdf.

[5] Cisco Systems (2012), Introduction to cisco ios netflow - a technical overview. URL: http://www.cisco.com/c/en/us/products/collateral/ios-nx-os-software/ios-netflow/prod_white_paper0900aecd80406232.html.

[6] Cisco Systems, Netflow export datagram format. URL: http://www.cisco.com/c/en/us/td/docs/net_mgmt/netflow_collection_engine/3-6/user/guide/format.html#wp1006186.

[7] Ray T. (2012), Augmented netflow: Using layer 7 metadata to enhance netflow analysis. URL: http://www.cert.org/flocon/2012/presentations/ray-using-layer-7-metadata-augment-flow-analysis.pdf.

[8] Powers A. (2012), Netflow architecture. CC-BY-SA-3.0, URL: https://cdn.plixer.com/blog/wp-content/uploads/2012/11/NetFlow-Diagram2.png.

[9] Pazder (2008), Netflow architecture using standalone probes. Public domain.

[10] IETF (2004), Ip flow information export (ipfix), ietf rfc 3917. URL: http://tools.ietf.org/html/rfc3917.

[11] Fullmer M. & Romig S. (2000) The osu flow-tools package and cisco netflow logs. In: 14th Systems Administration Conference, s. 291–303.

[12] Carnegie Mellon University (2015), Silk faq. URL: https://tools.netsa.cert.org/silk/faq.html.

[13] Bloom B.H. (1970) Space/time trade-offs in hash coding with allowable errors. Commun. ACM 13, s. 422–426. URL: http://doi.acm.org/10.1145/362686.362692.

[14] Roblee C. (2008), Bloomdex: Analyst workflow integration. URL: http://www.cert.org/flocon/2008/presentations/roblee_bloomdex-flocon2008.pdf.

[15] Azoff J. (2014), Netflow-indexer. URL: http://justinazoff.github.io/netflow-indexer/index.html.

[16] Azoff J. (2012), Sites using netflow-indexer. URL: http://justinazoff.github.io/netflow-indexer/sites.html.

[17] IANA (2016), Ipv6 global unicast address assignments. URL: http://www.iana.org/assignments/ipv6-unicast-address-assignments/ipv6-unicast-address-assignments.xhtml.

[18] Internet Census (2012), 2012 ipv4 census map. URL: http://internetcensus2012.bitbucket.org/paper.html.

[19] Lehtiö A. (2015), Duke apt group's latest tools: cloud services and linux support. URL: https://www.f-secure.com/weblog/archives/00002822.html.

[20] Cisco Systems (2003), Sampled netflow. URL: http://www.cisco.com/c/en/us/td/docs/ios/12_0s/feature/guide/12s_sanf.html.

[21] Mandiant (2015), M-trends 2015: A view from the front lines. URL: https://www2.fireeye.com/rs/fireye/images/rpt-m-trends-2015.pdf.

[22] Liikenne- ja viestintäministeriö (2014), Tietoyhteiskuntakaari (917/2014). URL: http://www.finlex.fi/fi/laki/ajantasa/2014/20140917.

[23] Chambi S., Lemire D., Kaser O. & Godin R. (2014) Better bitmap performance with roaring bitmaps. CoRR abs/1402.6407. URL: http://arxiv.org/abs/1402.6407.