TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS
INFORMACIJOS IR INFORMACINIŲ TECHNOLOGIJŲ SAUGA

BARTAS ALEKSANDRAVIČIUS

SNMP MONITORINGO ĮRAŠŲ ANONIMIZAVIMO MODELIO SUDARYMAS IR TYRIMAS

Magistro darbas

Vadovas
dr. D. Matulis

KAUNAS, 2013

KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS
INFORMACIJOS IR INFORMACINIŲ TECHNOLOGIJŲ SAUGA

BARTAS ALEKSANDRAVIČIUS

SNMP MONITORINGO ĮRAŠŲ ANONIMIZAVIMO MODELIO SUDARYMAS IR TYRIMAS

Magistro darbas

Darbo vadovas
dr. D. Matulis

Recenzentas
doc.dr. S. Maciulevičius

KAUNAS, 2013

AUTORIŲ GARANTINIS RAŠTAS
DĖL PATEIKIAMO KŪRINIO

2013 m. gegužės 24 d.
Kaunas

Autorius, **Bartas Aleksandravičius**,
(vardas, pavardė)

patvirtina, kad Kauno technologijos universitetui pateiktas baigiamasis magistro darbas (toliau vadinama – Kūrinys) SNMP MONITORINGO ĮRAŠŲ ANONIMIZAVIMO MODELIO SUDARYMAS IR TYRIMAS
(kūrinio pavadinimas)

pagal Lietuvos Respublikos autorių ir gretutinių teisių įstatymą yra originalus ir užtikrina, kad

1) jį sukūrė ir parašė Kūrinyje įvardyti autoriai;
2) Kūrinys nėra ir nebus įteiktas kitoms institucijoms (universitetams) (tiek lietuvių, tiek užsienio kalba);
3) Kūrinyje nėra teiginių, neatitinkančių tikrovės, ar medžiagos, kuri galėtų pažeisti kito fizinio ar juridinio asmens intelektinės nuosavybės teises, leidėjų bei finansuotojų reikalavimus ir sąlygas;
4) visi Kūrinyje naudojami šaltiniai yra cituojami (su nuoroda į pirminį šaltinį ir autorių);
5) neprieštarauja dėl Kūrinio platinimo visomis oficialiomis sklaidos priemonėmis.
6) atlygins Kauno technologijos universitetui ir tretiesiems asmenims žalą ir nuostolius, atsiradusius dėl pažeidimų, susijusių su aukščiau išvardintų Autorių garantijų nesilaikymu;
7) Autoriai už šiame rašte pateiktos informacijos teisingumą atsako Lietuvos Respublikos įstatymų nustatyta tvarka.

Autorius

Bartas Aleksandravičius _____
(vardas, pavardė) (parašas)

SANTRAUKA

Kompiuterinių tinklų valdytojai plačiai naudoja paprastąjį tinklų valdymo protokolą (toliau – SNMP), kuris padeda aptikti tinklų anomalijas ir įvairius gedimus. Norint patobulinti patį protokolą ir/ar stebėti specifinės tinklinės įrangos elgseną, būtina disponuoti šio protokolo monitoringo įrašais.

Problema iškyla tada, kai norima gauti SNMP monitoringo įrašus iš realius kompiuterių tinklus valdančių organizacijų ir atlikti su jais įvairius tyrimus. SNMP monitoringo įrašai apima daug organizacijai jautrios informacijos saugumo požiūriu, todėl jų valdytojai neviešina. Bet koks šios informacijos atskleidimas sukelia realias grėsmes organizacijos informacijos saugumui.

Siūlomas modelis leidžia atrinkti dominančią informaciją iš SNMP monitoringo įrašų visumos, anonimizuoti juos pagal pasirenkamus anonimiškumo kriterijus (priklausomai kokio formato duomenys anonimizuojami), išanalizuoti jų anonimiškumo laipsnį, gauti rekomendacijas ir jeigu anonimizuota informacija pakankamai saugi, perduoti trečioms šalims be grėsmės pakenkti disponuojamo tinklo saugumui.

Pasinaudojus šiuo modeliu galima pasiekti tokį jautrios informacijos apsaugos lygį, kuris leistų apsaugoti informacijos tiekėją nuo jo tinklo silpnų vietų analizės ir galimų atakų. Kita vertus, būtų galima išsaugoti pakankamą kiekį naudingos informacijos, kuri būtų naudinga moksliniams bei analitiniams tyrimams. Naudojant skirtingus anonimizavimo metodus, priklausomai nuo informacijos atributų formato, būtų galima sustiprinti ar sumažinti saugumo lygius, siekiant padidinti ar sumažinti anonimizuotuose monitoringo įrašuose saugomos informacijos išliekamąją vertę. Tokiu būdu suteikiama galimybė patiems tinklų administratoriams ar jų valdytojams spręsti kurią ir iki kokio lygio informaciją reikia anonimizuoti, prieš pateikiant ją naudojimui trečiose šalyse.

SUMMARY

Computer networks owners widely use the simple network management protocol (SNMP), as it helps to identify varies anomalies and failures of computer networks. In the case of improvement of that method, or/and controlling the status or behaviour of specific network equipment, we must have logs of protocol.

The problem arises when you want get them out from organizations, that manage computer networks. The SNMP monitoring records are not published, as any disclosure of real data would cause real threats to the organization and its security of information.

The proposed model allows to anonymise, analize the level of anonymization, valuate it and if the anonymised data enough secure, forward it for public use without the threat to damage own computer network.

It can be achieved a level of protection for sensitive information, by using this model, it will protect the suppliers and their networks from the real vulnerabilities analysis and possible attacks through them. On the other hand, it still could be saved a sufficient amount of information that would be useful for scientific and analytical research. Using the different anonymization techniques, depending on the attributes of the information format, it is possible to enhance or reduce the levels of security in order to increase or decrease the data lasting values. It gives the opportunity to network administrators or managers to decide, which data should be anonymized and how deep level of security must be reach before being released for the public use.

Turinys

LENTELIŲ SĄRAŠAS

PAVEIKSLŲ SĄRAŠAS

Įvadas

Realiame pasaulyje interneto paslaugų tiekėjai ir tinklinių sistemų administratoriai disponuoja dideliais informacijos kiekiais, susijusiais su tinklo inventoriaus veikimo būsenomis, incidentais, įvairiausio pobūdžio informaciniais pranešimais, kurie indikuoja ir ateina į pagalbą sistemų administratoriams, kai sistema nukrypsta nuo natūralios būsenos. Yra prikurta įvairiausių programų ir įrankių sistemų administratoriams, bet jais galima pasinaudoti tik su savo konkrečia uždara sistema ir pritaikyti savam tinklui.

Problema iškyla tada, kai norima tą informaciją perduoti kitiems. Kitų sistemų administratoriai, analitikai, inžinieriai ar mokslininkai norėdami geriau suprasti, išanalizuoti ir patobulinti informavimo protokolą, metodą ar tinklo įrangą, neturi tam pilnos informacijos disponavimo laisvės. Natūraliai kyla klausimas, kodėl taip yra?

Sistemų administratoriai, interneto paslaugų tiekėjai nelinkę teikti informacijos viešam naudojimui apie savo valdomų tinklų įrenginių būsenų, adresų, siuntėjų, gavėjų, gedimų, klaidų ir kitos informacijos iš esmės tik todėl, kad nebūtų galima analizuoti jų tinklų informacijos ir nebūtų kuriamos priemonės nukreiptos prieš jų pačių tinklus bei jų silpnąsias vietas. Kitaip sakant, pagrindinis argumentas yra saugumo grėsmės.

Jeigu sistemų administratoriai ar interneto tiekėjai turėtų galimybę anonimizuoti savo disponuojamų tinklų informaciją iki tokio lygio, kad nebūtų galima atsekti realių įrenginių, jų adresų ir būsenų, datų ir kitos informacijos, kuri galėtų pakenkti interneto tiekėjams, tikrai daug daugiau reikiamos informacijos atsirastų viešose terpėse ir būtų laisvai pasiekiama moksliniams tyrimams.

Darbo tikslas – sukurti modelį, leidžiantį pagal kelis atributus anonimizuoti SNMP monitoringo įrašus ir juos ištirti, įvertinant anonimiškumą bei naudingų duomenų išliekamumo lygmenį palyginant su pradiniais neanonimizuotais duomenimis.

Darbo uždaviniai:

- Sukurtas modelis turi saugiai ir patikimai anonimizuoti SNMP monitoringo įrašus, nes organizacijos tinkle saugomi organizacijų privatūs duomenys, tinklų būsenos bei struktūros;
- Turi būti ištirta ir modelyje įvertinta SNMP monitoringo įrašų anonimizavimo proceso seka, kurios pagalba būtų galima nuosekliai anonimizuoti įrašus pagal pageidaujamą naudingų duomenų išliekamumo lygmenį;
- Anonimiškumo lygmuo turi apsaugoti organizacijų teikiamą informaciją iki tokio lygio, kad nebūtų galimybės jai pakenkti piktavališkais tikslais;
- Anonimizuota informacija turi turėti tam tikrą tiriamąją vertę moksliniu ir analitiniu požiūriu;
- Turi būti sudaryta rekomendacinė kriterijų lentelė, pagal kurią būtų galima nustatyti anonimizavimo saugumo lygį.

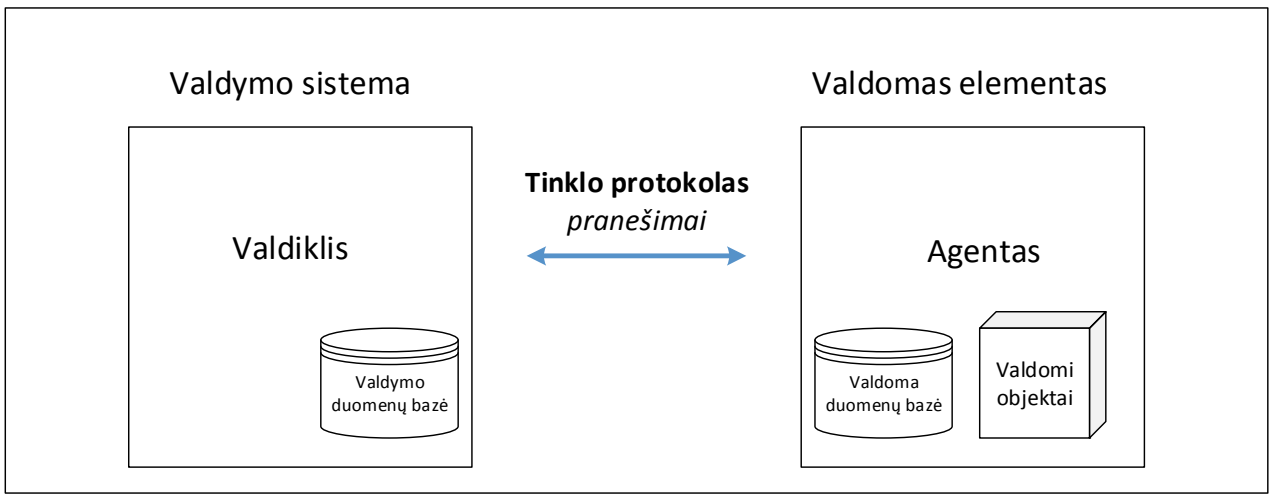
I. SNMP PROTOKOLO APŽVALGA IR ANALIZĖ

1.1. Įvadas

Nuo savo sukūrimo datos 1988-aisiais Paprastas Tinklo Valdymo Protokolas (angl. *Simple Network Management Protocol* - toliau SNMP) tapo standartu "de facto" tinklo valdymo srityje. Dė1 SNMP paprastumo ir mažo kodo apimties, tinklo įrangos gamintojai gali lengvai montuoti agentus į savo produkciją. SNMP yra lankstus ir leidžia gamintojams papildyti savo gaminius tinklo valdymo funkcijomis SNMP bazėje. Taip pat, SNMP atskiria valdymo architektūrą nuo aparatinės architektūros, kas plečia skirtingų gamintojų produktų palaikymo bazę. Ir svarbiausiai, SNMP skirtingai nuo kitų taip vadinamų "protokolų" yra ne popierinė specifikacija, o šiuo metu plačiausiai naudojamas tinklo valdymo protokolas.

1.2. Tinklo valdymo architektūra

Tinklo valdymo sistema susideda iš dviejų pagrindinių elementų: valdiklio *(angl. Manager)* ir agentų *(angl. Agents)*. Valdiklis yra konsolė, per kurią tinklų administratorius vykdo valdymo funkcijas. Agentai yra esybės, kurie atstovauja realų įrenginį ir tarpininkauja valdyme. Komutatoriai, maršrutizatoriai, kartotuvai, bridžai, tinklo serveriai - įrenginių, turinčių valdymo objektų, pavyzdžiai. Valdymo objektais gali būti aparatinė įranga, konfigūracijos parametrai, darbo našumo statistika ir t.t., visa tai kas sudaro informaciją apie momentinę įrenginio darbo būseną. Šie objektai surūšiuoti virtualioje duomenų bazėje, taip vadinamoje valdymo informacijos bazėje *(angl. Management Information Base* - toliau MIB). SNMP leidžia valdikliams ir agentams bendrauti tarpusavyje, aprūpinant priėjimą prie valdomų objektų (1 pav.)



1 pav. SNMP veikimo principas

SNMP protokolo savybės. Realizuotas kaip tinklo valdymo stotis (Network Managment Station - toliau NMS) ir pilnai realizuoja SNMP protokolą. Tinklo valdiklio savybės:

- Siųsti užklausas agentams;
- Keisti kintamųjų reikšmes agentuose;
- Priiminėti asinchroniškus signalus apie įvykius iš agentų.

Agento savybės:

- Pilnai realizuoja SNMP protokolą;
- Saugoja ir ima informaciją iš valdymo informacijos bazės (MIB);
- Gali asinchroniškai signalizuoti vadybininkui apie įvykį;
- Gali tarpininkauti, suteikiant SNMP sąsają įrenginiui, nepalaikančiam SNMP.

1.3. SNMP protokolo aprašymas

SNMP (Paprastas Tinklo Valdymo Protokolas) veikia pagal valdiklio/agento modelį. Protokolas yra "paprastas", nes agentas minimaliai reikalauja programinio aprūpinimo. Pagrindines valdymo funkcijas atlieka valdymo sistema, o ją papildo valdomųjų sistemų funkcijų rinkiniai. Kad būtų galima tikslą pasiekti paprastai, SNMP sugeba apdoroti ribotą valdymo funkcijų bei atsakymų skaičių. Valdanti sistema (valdiklis) naudoja pranešimus, kad gautų atitinkamai vieno ar kelių objektui kintamųjų reikšmes bei nustatytų tam tikro kintamojo reikšmę. Yra penki pranešimų tipai:

- *Get*. Vieno valdymo objekto išrinkimas.
- *GetNext*. Kelių valdymo objektų ar duomenų lentelės išrinkimas.
- *Set*. Valdymo objekto keitimo žinutė.
- *Get Response.* Agentas atsako į Get, GetNext ir Set pranešimus.
- *Trap*. Žinutė informuojanti apie nenumatytus įvykius. Agentai patys be užklausimo siunčia šias žinutes apie įvykius (traps) valdikliui, kai atsitinka tam tikra nenumatyta situacija, pvz., kintamojo reikšmė viršija nustatytą ribą.

SNMP pranešimų kūrimas. Bendras SNMP pranešimo formatas

- Versijos numeris;
- Bendrijos identifikatorius (angl. community string) - vietoj slaptažodžio;
- Vienas ar keletas SNMP PDU (angl. protocol data unit) - protokolo duomenų vienetas.

SNMP Trap PDU formatas:

- enterprise - identifikuoja signalizuojantį objektą;
- agent address - signalizuojančio agento IP adresas;
- generic trap id - standartinės signalizacijos identifikatorius;
- specific trap id - specifinis signalizacijos identifikatorius;
- time stamp - signalizacijos suveikimo laikas impulsuose;
- OIDs ir jų reikšmės – OIDs, kurie turi būti perduoti NMS.

SNMP protokolo aprašas:

- Kiekvienas SNMP valdomas objektas priklauso bendrijai;
- NMS stotys gali priklausyti keletui bendrijų;
- Bendrija identifikuojama pagal vardą, kuris yra oktetų seka (maks. 255 oktetų);
- Kiekvienas SNMP pranešimas susideda iš trijų komponentų:
    a. Versijos numerio;
    b. Bendrijos pavadinimo;
    c. Duomenų - PDU sekos, susijusios su užklausa.

1.4. Standartinės SNMP saugumo priemonės

Autentifikacija. Triviali autentifikacija realizuota per bendrijos pavadinimo (community string) siuntimą atviru tekstu SNMP pranešimuose. Autentifikacija numano, kad pranešimai nėra koreguojami, pakeičiami bei padirbinėjami.

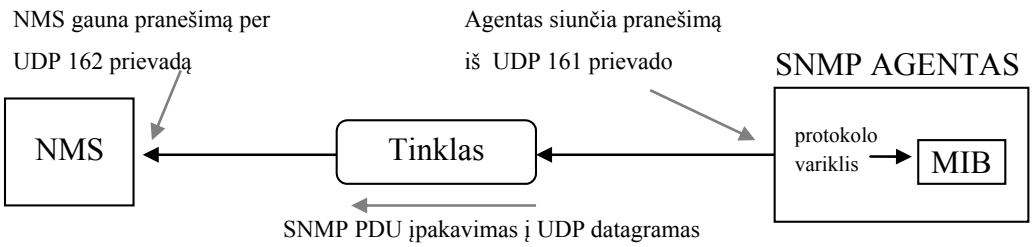
Autorizacija. Kai bendrijos pavadinimas patikrintas, agentas ir vadybininkas tikrina, ar pranešimo adresas yra leistinas ir turi pakankamai teisių operacijos vykdymui.

Siuntimo adreso ir bendrijos pavadinimo patikrinimas sudaro pagrindą užklausos patvirtinimui arba atmetimui.

SNMP protokolu pasiekiami duomenys. SNMP pasiekia apibrėžtus valdymo informacijos bazėje (MIB) objektus:

- Visi objektai saugomi MIB;
- MIB pomedžio žemesnio lygio mazguose;
- SNMP protokolas pasiekia objektus pagal jų įdentifikatorius: *x,y*.

Kur *x* yra "tikras" objekto OID MIB, o *y* yra raktas, kuris identifikuoja apibrėžtą valdymo objektą (pvz. kai operuojama su lentele). Jei objektas visame MIB yra vienintelis, pvz. nėra sukurtas kiekvienam interfeisui, tada *y=0.* Pvz: sysDescr (OID: 1.3.6.1.2.1.1.1) bus pasiektas SNMP protokolu kaip 1.3.6.1.2.1.1.1.0 (sysDescr.0)



2 pav. TRAP signalizacijos apie įvykį schema

1.5. SNMP protokolo monitoringo įrašas.

SNMP protokolo monitoringo įrašai gali būti standartiniai (angl. generic trap), kurie siunčia 6 tipų informaciją apie standartines įrenginių būsenas (coldStart, WarmStart, LinkUp, LinkDown, authenticationFailure, egpNeighborLoss) (žr. 3pav.). Septintas tipas yra vadinamas specifiniu (angl. enterpriseSpecific), nes šiame pranešime, tinklo įrangą valdančios ar gaminančios organizacijos, gali kurti specializuotus pranešimus, būtinus stebėti konkrečius tinklo įrangos parametrus (žr. 4pav.).

```
Nov 21 07:44:17: %LINK-3-UPDOWN: Interface Loopback1, changed state to up
4d23h: SNMP: Queuing packet to 172.17.246.162
4d23h: SNMP: V1 Trap, ent products.45, addr 172.17.246.9, gentrap 3, spectrap 0
 ifEntry.1.23 = 23
 ifEntry.2.23 = Loopback1
 ifEntry.3.23 = 24
 lifEntry.20.23 = up
```

3 pav. Standartinio tipo TRAP pranešimas

3 pav. pavaizduotame SNMP TRAP pranešime sakoma, kad tinklo įrenginio, kurio IP adresas yra 172.17.246.9 tinklo plokštė atstatė ryšį su tinklu, todėl siunčiamas pranešimas į tinklo valdiklį IP adresu 172.17.246.162.

Organizacijos, kuriančios tinklų valdymo ir stebėjimo įrangą, kuria ir specifinės paskirties TRAP pranešimus, skirtus specifiniams įvykiams pranešinėti. Todėl pranešimo turinys ir sąlygos, pagal kurias siunčiamas šio tipo TRAP pranešimas, gali būti labai įvairūs (4 pav.). Pagrindiniai panaudojimo atvejai gali būti įsiskirstyti į šias grupes:

- Aplinkos aliarmai (temperatūra, drėgmė, fizinė sauga ir kt.);
- Transporto įranga (mikrobangos, mobilūs jungikliai ir kt.);

- Maitinimo sistemos (generatoriai, baterijos lygintuvai ir kt.);
- IT infrastruktūra (komutatoriai, maršrutizatoriai, serveriai, duomenų masyvai ir kt.).

```
4d23h: SNMP: Queuing packet to 172.17.246.162
4d23h: SNMP: V2 Trap, reqid 2, errstat 0, erridx 0
 sysUpTime.0 = 43053404
 snmpTrapOID.0 =
 clogHistoryEntry.2.958 = SYS
 clogHistoryEntry.3.958 = 6
 clogHistoryEntry.4.958 = CONFIG_I
 clogHistoryEntry.5.958 = Configured from console by vty0 (10.10.10.10)
 clogHistoryEntry.6.958 = 43053403
```

4 pav. Specifinio tipo TRAP pranešimas

Bet kuriuo atveju visų tipų TRAP pranešimuose galima išrinkti svarbią ir tą pačią informaciją (pranešimo data, siuntėjas, gavėjas, objekto identifikatorius ir jų būsenas).

1.6. Išvados

SNMP nėra visiškai saugus, nes protokolo 1 ir 2 versijose nėra numatytas joks duomenų šifravimas, o prieiga prie informacijos apsaugoma tik atviro teksto tekstine eilute (angl. community string). Papildomos saugumo priemonės įmanomos su dabartiniu protokolu (v3): DES ir MD5, kurios nesumažina SNMP galingumo tinklo stebėjime bei valdyme.

SNMP TRAP pranešimų informacijoje yra daug jautrios informacijos, kurią atitinkamai apdorojus ir išanalizavus, galima susidaryti pakankamai aiškią organizacijos vidaus tinko struktūrą, gedimų pobūdžius, jų tendencijas. Todėl, tai suprasdami, organizacijų saugumo politiką prižiūrintys specialistai griežtai draudžia platinti tokio pobūdžio monitoringo įrašus.

II. ANONIMIŠKUMO APŽVALGA IR ANALIZĖ

2.1. Anonimiškumo savybės

Šiame skyriuje mes pirmiausia apibrėšime, ką vadiname "anonimiškumu" ir "sistemos anonimiškumu“, taip pat aptarsime kai kuriuos dažniausiai naudojamus terminus literatūroje. Taip pat pateiksime įvairių tipų, anonimiškumui priešiškų modelių klasifikacijas, kurias dažniausiai siekiama anonimizuoti.

Anonimiškumas ir pseudoanonimiškumas. Siekdami standartizuoti anonimizavimo terminologiją naudojamą literatūroje, Pfitzmann ir Hansen [4] apjungė ir pasiūlė savo terminologiją. Šiame darbe mes taip pat griežtai laikysimės jų siūlomos terminologijos. Laikysime, kad bet kuri sistema yra veikėjų visuma, kuri turi klientus, serverius ar jiems analogiškus įrenginius viename komunikacijos tinkle. Šie aktoriai apsikeitinėja informacija tarpusavyje per viešuosius bendravimo kanalus, kaip pavyzdžiui Internetas. Pagal grafų teoriją kartais aktorius tinkle vadiname mazgais (angl. node), o komunikacijos kanalus tarp jų – jungtimis (angl. links).

Neformaliai Pfitzmann ir Hansen [4] pasiūlė anonimiškumo apibrėžimą “Anonimiškumas – tai objekto neidentifikuota būsena objektų visumoje“. Kitaip žodžiais tariant, anonimiškumas yra tokia būsena, kai visi sistemoje esantys dalyviai gali būti konkrečios žinutės ir siuntėjai, ir gavėjai. Plačiau tikslinant sąvokas galima jas išskirti pagal poreikį, kad būtų galima apibrėžti žinutės anoniminių siuntėjų visumą ir atitinkamai anoniminių gavėjų visumą. Taip pat, esant poreikiui, galima nustatyti ir informacijos siuntimo datos ir laiko anoniminių intervalų sąvoką, kur būtų anonimizuojama data ir laikas nustatytame datos ir laiko intervale.

Iš principo, anoniminių sistemų tikslas - teikti ***ryšio nenutraukiamumą*** (angl. ***unlinkable***) tarp žinutę siuntusio anoniminio siuntėjo ir jos tikrojo gavėjo, bei tarp anoniminio gavėjo ir tikrojo siuntėjo. Taip pat Pfitzmann ir Hansen [4] nustatė ir santykių anonimiškumą tarp siuntėjo ir gavėjo. Santykių anonimiškumas sukuriamas tam, kad neteisėtas tyrimas neleistų nustatyti kas su kuo bendrauja.

Siunčiamos ar gaunamos žinutės savybė, būti neišskirtai iš kitų atsitiktinių žinučių sistemoje, yra vadinama **nepastebimumu** (angl. ***unobservability***), kuri yra dar didesnio saugumo lygmens savybė nei ***ryšio nenutraukiamumas***.

Labai svarbu atskirti anonimiškumą ir pseudo anonimiškumą. Jeigu anonimiškumas reiškia sistemos dalyvių neidentifikavimą anonimiškumo visumoje, tai pseudo anonimiškumas tiesiog reiškia, kad aktorius, ar siunčiantis ar gaunantis žinutes, yra susijęs su konkrečiu identifikatoriumi, bet ne jis pats yra identifikuojamas. Pfitzmann ir Hansen [4] taip pat pažymėjo, kad vienas pseudonimas turi būti priskiriamas tik vienam turėtojui, o ne grupė skirtingų subjektų dalintųsi vienu

pseudonimu, žinomu kaip grupinis pseudonimas. Jei yra naudojamasi grupiniu pseudonimu, tai yra beveik visiškai tas pats, kas naudoti anonimiškumo būseną.

Kadangi anonimiškumas ir pseudo anonimiškumas yra skirtingos sąvokos, bet iš esmės labai susijusios, Goldberg [9] savo darbe pademonstravo "Nymity Slider" koncepciją, kaip įmanoma sukurti pseudo anonimiškumą ant anonimiškumo. Aktorius iš anoniminės sistemos gali tiesiog prisiskirti konkretų pseudonimą sau, skirtą atlikti eilę operacijų (angl. transaction) per anoniminę sistemą su tuo identifikatoriumi, kuris labiau toleruoja pseudo anonimiškumą, nei anonimiškumą.

Anonimiškumui priešiški modeliai. Visuose, su saugumu susijusiuose darbuose ir tyrimuose, anonimiškumui priešingos sąvokos yra dažnai naudojamos apibūdinti jų tikslams ir saugumo stiprumui nusakyti. M. Edman ir B. Yener apibendrino ir sukvalifikavo šias savybes:

- Gebėjimas (angl. capability);
- Matomumas (angl. visibility);
- Judrumas (angl. mobility);
- Dalyvavimas (angl. participation).

Gebėjimas. Tai svarbi priešiškai nusiteikusio subjekto (toliau – piktavalis) savybė, kuris gali būti tiek pasyvus, tiek ir aktyvus. Pasyvus piktavalis laikomas toks subjektas, kuris gali stebėti ir įrašinėti tinklo srauto nuorodas, kuriomis įeina ir išeina klientai bei serveriai anonimizuotame tinkle. Pasyvus piktavalis taip pat gali įrašinėti metaduomenis apie tinklo srautą, tokius kaip paketo ilgis ir jo atvykimo laikas. Net paprastas pasyvus srauto stebėjimas gali leisti galingas statistines atakas.

Aktyvus piktavalis turi visas tokias pačias stebėjimo galimybes kaip ir pasyvus piktavalis. Jis taip pat turi ir galimybę ir manipuliuoti tinklo srautu, galbūt net kontroliuoti vieną ar daugiau tinklo ryšius ar valdyti anoniminio tinklo mazgą. Jis ar ji gali keisti, nutraukti ar panaikinti dalį tinklo srauto, įterpti savo valdomą srautą ar perleisti teisėtą tinklo srautą, kurį anksčiau buvo įrašęs.

Matomumas. Tai tokia savybė, kuri literatūroje tankiai naudojama, kai norima apibūdinti grėsmės modelį. Piktavalio matomumas nurodo, kiek tinklų piktavalis gali pasyviai stebėti ir aktyviai valdyti. Globalus piktavalis, kaip ir pavadinimas sufleruoja, yra galingas stebėtojas, kuris turi prieigą prie visų tinklo ryšių tarp klientų ir serverių anonimiškame tinkle.

Ir priešingai, dalinis piktavalis gali stebėti tik ryšius viename iš tinklo mazgų tarp klientų ir serverių ar tik mažą potinklio dalį susijusio su konkrečiu mazgu iš viso tinklo. Dalinio piktavalio pavyzdys būtų kas nors tame pačiame vietiniame tinkle, kaip mazgas anoniminiame tinkle.

Mobilumas. Net jei piktavalis neturi didelio mąsto infrastruktūros, reikalingos kontroliuoti visus klientus ir serverius anonimiškumo tinkle tuo pačiu laiku, jis ar ji gali sugebėti išrinkti, kuriuos tinklo poaibius jis nori kontroliuoti, įvertinant anksčiau gautą informaciją.

Tokį piktavalį vadiname adaptyviu piktavaliu. Pavyzdžiui, vyriausybė gali šaukti į teismą pasirinkimą interneto paslaugų tiekėją pagal jo teisėtą jurisdikciją, kad galėtų kontroliuoti tinklo eismą kai kurių serverių anonimiškumo tinkle. Tada adaptyvus piktavalis gali nuspręsti kontroliuoti kitokį, galbūt analogišką serverių poaibį, pasinaudodamas ankstesne (teisėtai surinkta) informacija.

Ir priešingai, statiškas piktavalis sugeba sukelti grėsmę ar kontroliuoti kažkokį tinklo poaibį, bet nesugeba parinkti, kokį poaibį ji ar jis gali stebėti. Globalus piktavalis yra, beveik iš esmės, statiškas piktavalis, kadangi jis turi globalinį anonimiško tinklo vaizdą, bet jis negali atrankos būdu kontroliuoti mažesnių tinklo dalių.

Dalyvavimas. Raymond [8] taip pat patarė skirstyti piktavalius į vidaus ir išorinius. Vidaus piktavalis yra toks, kuris dalyvauja anonimiško tinklo protokole kaip klientas, ar, galbūt, valdo tinklo infrastruktūros dalį, valdydamas serverį tinkle. Pažymėtina, kad vidaus piktavalis nėra iš esmės taip pat aktyvus piktavalis. Vietoj to, jis gali tiesiog kontroliuoti tinklo srautą, kuris praeina per jo valdomą serverį, aktyviai nekeisdamas jo.

Išorinis piktavalis, nedalyvauja anonimiškumo tinkle ar jo protokole. Vietoj to, jis stato į pavojų komunikacijas, panaudodamas klientų (t.y., jų tinklo sąsajas), kad kontroliuotų ar valdytų jų tinklo srautą.

2.2. Anonimizavimo laipsnio apskaičiavimas

Kadangi mūsų darbas remiasi anonimizavimo principais, tai iš pradžių būtina apžvelgti ir išanalizuoti duomenų anonimizavimo specifiką. Pagrindiniai parametrai, kuriais mes remsimės savo tyrime ir taikysime mūsų modelyje, bus k-anonimiškumas ir entropija su įvairiomis jos formomis.

K-anonimiškumas. K-anonimiškumo koncepcija sako, kad norint išsaugoti objekto privatumą ir išlaikyti jį anonimišku, būtina, kad pateikiami duomenys būtų taikytini daugiau nei vienam respondentui. Pateikiamų privačių duomenų lentelės stulpeliuose surašomi objektą identifikuojantys atributai, arba kitaip vadinami kvazi-identifikatoriai (toliau – QI) {Data, laikas, IP adresas, OID, Būsena}, o eilutėse pateikiamos objekto QI reikšmės (žr. 1 lentelę).

Lentelė 1. Pirminiai SNMP trap pranešimo duomenys.

Data laikas	**IP adresas**	**OID**	**Būsena**
2011-01-12 20:41:11	172.30.10.34	a3Com	linkDown
2011-01-12 04:08:18	172.30.20.39	compaq	linkUp
2011-01-11 07:15:37	172.30.10.33	microsoft.1.1.3.1.1	AuthenticationFailure
2011-01-09 06:37:11	172.30.10.12	dell.server3.baseboardGroup	alertAmperageWarning
2011-01-01 03:02:10	172.30.20.18	cio2	ColdStart
2011-01-04 07:46:54	172.30.10.19	dell.storage	arrayDiskRemoved

Apibrėžimas 1 (k-anonimiškumo sąlygos) [21]. Kiekvienas paviešinamas duomenų masyvas turi būti toks, kad kiekviena QI reikšmių kombinacija būtų taikytina ne mažiau kaip k objektams.

Iš pirmo žvilgsnio atrodo, kad įgyvendinti tokias sąlygas prieš pateikiant viešam naudojimui jautrius duomenis yra sudėtingas uždavinys, bet atlikus duomenų apibendrinimą (angl. generalization) pagal jų atributus, tai galima pasiekti. Garantuoti k-anonimiškumo sąlygų įvykdymą, kad kiekviena QI reikšmė turėtų ne mažiau kaip k atvejų, taip kaip aprašyta sekančiame apibrėžime.

Apibrėžimas 2 (k-anonimiškumas) [21]. Tarkime, kad T (A_1,...,A_m) yra lentelė, o QI yra su ja susiję. Tada T tenkina k-anonimiškumą su QI, jeigu kiekviena reikšmė iš T[QI] taikytina ne mažiau kaip *k* respondentams T lentelėje pagal tuos pačius QI.

Šis apibrėžimas yra pakankama sąlyga k-anonimiškumo reikalavimams tenkinti. Bet kokia lentelė, tenkinanti 2 apibrėžimą duotajam k, tikrai tenkins k-anonimiškumo reikalavimus duotajam k. Jeigu išorinės lentelės atributų aibė bus sutapatinama su jautrių duomenų lentele, kuri tenkina 2 apibrėžimą, tai paviešintų duomenų deriniai su išoriniais duomenimis niekada neleis identifikuoti objektų tiksliau nei k galimų objektų tikslumu.

Pavyzdžiui, stebint 1 lentelę ir jos QI {Data laikas, IP adresas, OID, Būsena} lengva pamatyti, kad lentelė tenkina k-anonimiškumą tik k=1, nes kiekvienas IQ reikšmių derinys yra unikalus (pvz., 2011-01-04 07:46:54 /172.30.10.19 / dell.storage / arrayDiskRemoved). Duomenų su tokiu k-anonimiškumu jokiu būdu negalima viešinti, todėl reikia labai atidžiai įvertinti ir nuspręsti kuriuos IQ pateikti viešam naudojimui. K-anonimiškumo didinimui naudojami įvairūs metodai, tokie kaip apibendrinimas (angl. generalization), slopinimas (angl. suppression), netikrų duomenų pridėjimas ir/ar specifinių anonimizavimo algoritmų naudojimas.

Entropijos anonimiškumo laipsnis nusako visos matuojamos sistemos neapibrėžtumo lygį. Šis matas leidžia mums įvertinti kaip, patikimai ar ne, yra anonimizuotas mūsų tiriamas duomenų masyvas. Kitaip sakant, tai teorinis informacinis matas, kuris nusako, koks yra laukiamos vertės pasiskirstymas visoje anonimiškoje sistemoje.

Entropija H (X) apima atskirų tikimybių Pr_i visumą:

$$H(X) = -\sum_{i \in X} p(x_i) * log(p(x_i)) \ , \quad (1)$$

Sąlyginės entropijos H (X|C) apskaičiavimas, kai yra žinoma dalinė informacija C:

$$H(X|C) = -\sum_{i \in X, j \in C} p(x_i, y_j) * log(p(x_i|y_j)) \ , \quad (2)$$

kur C yra perimtos informacijos visuma ar informacija, gauta iš paviešintos anonimizuotos lentelės; $Pr_{i,j}$ yra jungtinė dalyvio *i* ir perimtos informacijos *j* tikimybė; $Pr_{i|j}$ yra sąlyginė tikimybė, kad objektas *i* yra atpažintas pagal duotą informaciją *j*.

Sandūros (angl. joint) entropijos H (X,Y) apskaičiavimas, naudojamas kai reikia įvertinti dviejų kvazi-identifikatorių galimų reikšmių derinius sistemoje.

$$H(X,Y) = -\sum_{i=n} \ \sum_{j=n} p\left(x_i, y_j\right) * \log\left(p\left(x_i, y_j\right)\right) \ , \quad (3)$$

Dalinės entropijos apskaičiavimo metodas. Vertinant anonimiškumą yra labai svarbu įvertinti kokių veiksmų gali imtis piktavalis, norėdamas identifikuoti realius objektus su tikslu įsibrauti į sistemą. Kaip teigiama [20], anonimiškumo lygį vertinti būtina atsižvelgiant į šias nuostatas:

a) Anonimiškumo puolimas yra visada efektyvus ir po kiekvienos atakos anonimiškumo laipsnis kinta;
b) Kuo sistemoje yra didesnis objektų pasiskirstymas, tuo piktvaliui sunkiau apsispręsti, todėl anonimiškumas geresnis;
c) Bendruoju atveju, piktavalio nustatyti objektų pasikartojimo dažniai, yra jo statistinio stebėjimo rezultatas.

Dėl šių priežasčių, puolėjas prieš pradėdamas įsibrovimo ar identifikavimo veiksmus, iš pradžių juos statistiškai įvertins ir, be abejo, tuos elementus, kurių pasikartojimų tikimybė mažiausia, tiesiog atmes, o pradės analizuoti tuos, kurie sudaro didžiąją visų duomenų pasikartojimo dažnio dalį. Dėl šios priežasties labai svarbu įvertinti ne tik bendrąją sistemos entropiją, bet ir dalinę entropiją.

Dalinės entropijos metodas [20] apskaičiuoja tik tų skirtingų anonimiškų objektų neapibrėžtumą, kurių pasikartojimo dažnių pasiskirstymas sistemoje daro esminę įtaką bendram sistemos anonimiškumui (toliau - dominuojantys objektai).

Apibrėžimas. Surūšiuokime mažėjimo tvarka visų sistemos objektų pasikartojimų dažnius ir, atrinkus tik dominuojančių objektų pasikartojimų dažnius, gaunama nauja sistema:

$$D = \frac{\sigma}{\beta/_m}, \tag{4}$$

kur

D - bendras sistemos anonimiškumas, vertinant tik dominuojančius objektus sistemoje;

σ – dominuojančių objektų entropija, $\sigma = -\sum_{i \in m} p(x_i) * log(p(x_i))$;

β - dominuojančių objektų pasikartojimo dažnių suma, $\beta = \sum_{i \in m} p(x_i)$;

m – dominuojančių objektų kiekis sistemoje.

Iš apibrėžimo tampa aišku, kad sistemos anonimiškumas tiesiogiai priklauso nuo σ, β ir m. Kuo didesnis entropijos laipsnis σ, tuo geresnis anonimiškumas D. Kuo didesnis β, tuo prastesnis anonimiškumas ir kuo didesnis kiekis m, tuo geresnis anonimiškumas.

Taikymas. Pavyzdžiui, turime anonimizuotų duomenų aibę, kurios skirtingų objektų pasikartojimo dažniai yra apskaičiuoti ir surūšiuoti mažėjančia tvarka A={0.24, 0.24, 0.24, 0,24, 0.01, 0.01, 0.005, 0.005, 0.005, 0.005}. Lengva nustatyti, kad dominuojantys objektai yra m = 4, o jų pasikartojimų dažnių suma β = 0.96. Apskaičiavus šių objektų entropiją σ ir rezultatus sustačius į 4 formulę, gauname sistemos neapibrėžtumą D=2.48, kuris atitinka tik minimalius saugumo reikalavimus anonimizuotai sistemai. Palyginimui pateikiama keletas kitų dominuojančių objektų pasiskirstymas sistemoje ir jų vertės:

Lentelė 2. Dominuojančių objektų pasiskirstymas anonimizuotoje sistemoje.

Objektų pasiskirstymo dažnis sistemoje	Dominuojančių objektų pasiskirstymas	m	β	D
0.42,0.42,0.03,0.03,0.03,0.02, 0.02,0.01,0.01,0.01	0.42,0.42	2	0,84	1.2
0.24,0.24,0.24,0.04,0.04,0.04, 0.04,0.04,0.04,0.04	0.24,0.24,0.24	3	0.72	1.8
0.24,0.24,0.24,0,24,0.01,0.01, 0.005,0.005,0.005,0.005	0.24,0.24,0.24,0,24	4	0.96	2.48
0.30,0.23,0.22,0.21,0.01,0.006, 0.006,0.006,0.006,0.006	0.30,0.23,0.22,0.21	4	0.96	2.46

2.3. Anonimizavimo algoritmai ir metodai

Norint, kad anonimizuoti SNMP žurnalai būtų naudingi ir tolimesniems tyrimams bei nebūtų atskleista jautri interneto tiekėjų informacija, būtina anonimizuoti ne tik siuntėjo ir gavėjo IP adresus, bet taip pat ir kitus jautrius žurnalų laukus, tokius kaip siuntimo data ir laikas, įrenginio ar objekto identifikatorius bei jų būsenas. Apžvelkime keletą anonimizavimo algoritmų, kurie panaudoti panašiems informacijos anonimizavimo tikslams.

Ryšių nutraukimo algoritmas. Weijia Yang ir Sanzheng Qiao [12] pasiūlė anonimizavimo algoritmą (3 pav.), kurio tikslas yra nutraukti ryšius tarp skirtingų QI ir reikšmių. Tai nereikalauja anonimizuoti visų QI reikšmių. Pagrindinė idėja yra ta, kad atsitiktiniu būdu (angl. random) pakeisti QI reikšmes naudojant originalių verčių pasiskirstymą. Tokiu būdu, jokia nauja informacija nėra pridėta prie anonimizuotų duomenų, ryšiai tarp QI ir jautrių duomenų yra nutraukti, o originalus duomenų pasiskirstymas yra išsaugotas. Nors ir ryšiai atskiruose įrašuose yra nutraukti, bet ryšių statistika visame duomenų rinkinyje yra išsaugota.

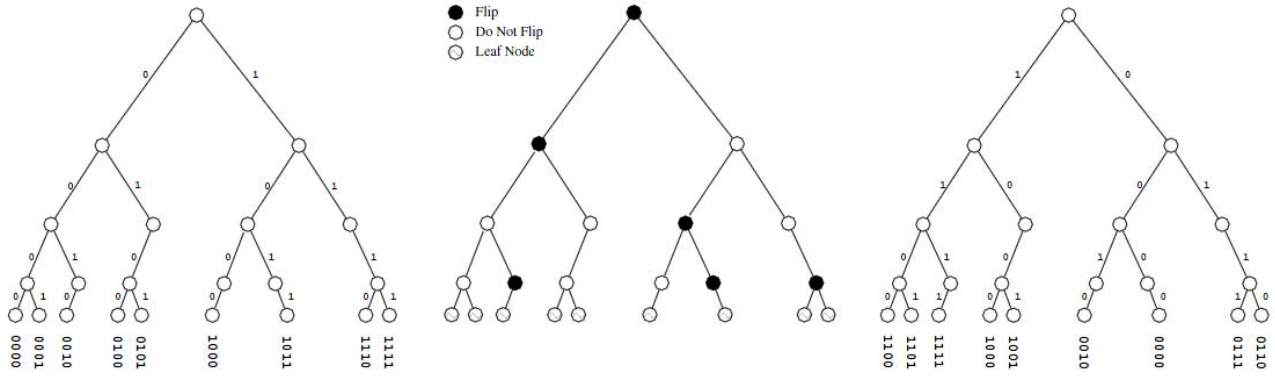
1: **Input** : the original data set D, the Q-I attributes Q, and the probability distribution $\{p_1, \ldots, p_m\}$, where $m = |Q|$, the number of attributes in Q-I.
2: **Output**: the original data set D is overwritten by an anonymized one
3: **begin**
4: $n := |D|$;
5: $Dist := \emptyset$;
6: **for** i: = 1 **to** m **do**
7: **begin**
8: $Dist_i$:= the distribution of the values of Q_i;
9: **end**
10: **for** j: = 1 **to** n **do**
11: **begin**
12: Randomly select an attribute Q_k in Q-I of the jth record with probability p_k;
13: Randomly generate a new value for Q_k based on $Dist_k$;
14: Replace the value of Q_k with the new value;
15: **end**
16: **end**

5 pav. Ryšių nutraukimo algoritmas [12].

IP adreso anonimizavimo algoritmas, išsaugantis prefiksą ir leksikografinį eiliškumą. Jun Xu, Jinliang Fan ir Mostafa H. Ammar [18] pasiūlė naują IP adreso anonimizavimo algoritmą, kuris išsaugo tokį patį IP adreso prefiksą, jei yra anonimizuojami du to paties potinklio adresai. Šio anonimizavimo tipo nauda yra ta, kad tinklų ir potinklių struktūra yra visiškai išsaugota, kai tuo pačiu metu ir nežinoma. Piktavalis bandydamas skanuoti tokį anonimizuotą IP adresą jokios apčiuopiamos naudos negaus, nes nei realus potinklis, nei tikras IP adreso savininkas nebus žinomas.

Apibrėžimas. Du IP adresai $a = a_1a_2 \ldots a_n$ ir $b = b_1b_2 \ldots b_n$ turi *k* bitų ilgio prefiksą ($0 \leq k \leq n$), jei $a_1a_2 \ldots a_k = b_1b_2 \ldots b_k$ ir $a_{k+1} \neq b_{k+1}$, tada $k < n$. Anonimizavimo funkcija F yra apibrėžta kaip vienas vienam funkcija nuo $\{0, 1\}^n$ į $\{0, 1\}^n$. Anonimizavimo funkcija F yra priešdėlio išsaugojimas, jei duotiems dviem IP adresams a ir b, kurie turi k – bitų ilgio priešdėlį, tada F (a) ir F (b) turi k – bitų ilgio priešdėlį taip pat.

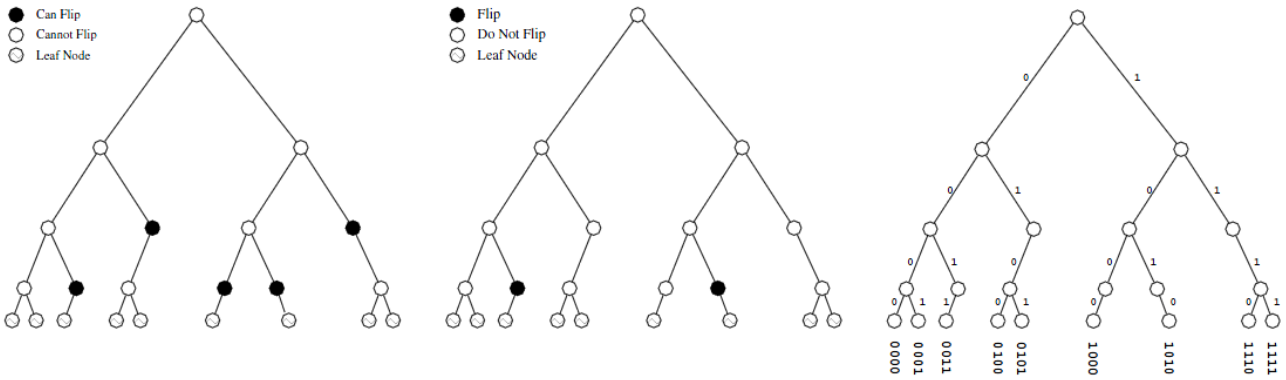
Įsivaizduokime anonimizavimo metodo, išsaugant prefiksą, geometrinę interpretaciją (4 pav.). Atkreipkime dėmesį, kad pilnai IP adreso erdvei gali atstovauti užbaigtas dvejetainis medis. Kadangi IPv4 adresai, šis medis turėtų 32 mazgus, tuo metu, kai IPv6 turės128 mazgus. Kiekvienam IP adresui atstovauja mazgas. Be to, kiekvienas mazgas atitinka bito poziciją (išreikštas pagal mazgo aukštį) ir vertes (išreikštas pagal jos tėvinio mazgo šakos kryptį).



(a) Originalus adresų medis (b) anonimizavimo funkciją f_i (c) Anonimizuotas adresų medis

6 pav. Prefikso išsaugojimo anonimizavimo funkcijos geometrinė interpretacija.

Prefikso išsaugojimo anonimizavimo funkcijos geometrinė interpretacija kaip apibrėžta [18] (6 pav.). Kairė dalis (a) atstovauja devyniems adresams, paimtiems iš 4 bitų adresų erdvės kaip dvejetainis medis. Vidurinė dalis (b) rodo atsitiktinai pasirinktą anonimizavimo funkciją, t.y., komplektas mazgų dvejetainiame medyje, kuriais pakeičiama bito reikšmė, kad gauti anonimizuotus adresus. Anonimizuoto adreso geometrinis vaizdavimas (c) rodo 4 bitų adresus, gautus pritaikius anonimizavimo funkciją f_i, naudojant (b) reikšmes.



(a) Bitai, kurie gali būti pakeisti, (b) anonimizavimo funkciją f'_i (c) Anonimizuotas adresų medis

7 pav. Prefikso ir leksikografinio eiliškumo išsaugojimo anonimizavimo funkcijos geometrinė interpretacija.

Prefikso ir leksikografinio eiliškumo išsaugojimo anonimizavimo funkcijos interpretacija, kurią pasiūlė M. Harvan ir J. Schonwalder [2].

7 paveikslėlio kairė dalis (a) rodo mazgus, kuriems galima pakeisti bito reikšmę. Vidurinė dalis (b) rodo kombinaciją su mazgais, kuriuos būtina pakeisti pagal priešdėlio išsaugojimo anonimizavimo funkciją, parodytą 7 paveikslėlio (b) dalyje. Anonimizuoto adreso geometrinis vaizdavimas (c) rodo 4 bitų adresus, gautus pritaikius anonimizavimo funkciją f'_i naudojant 5 paveikslėlio (b) reikšmes.

PAVYZDYS

Originalūs IP adresai:

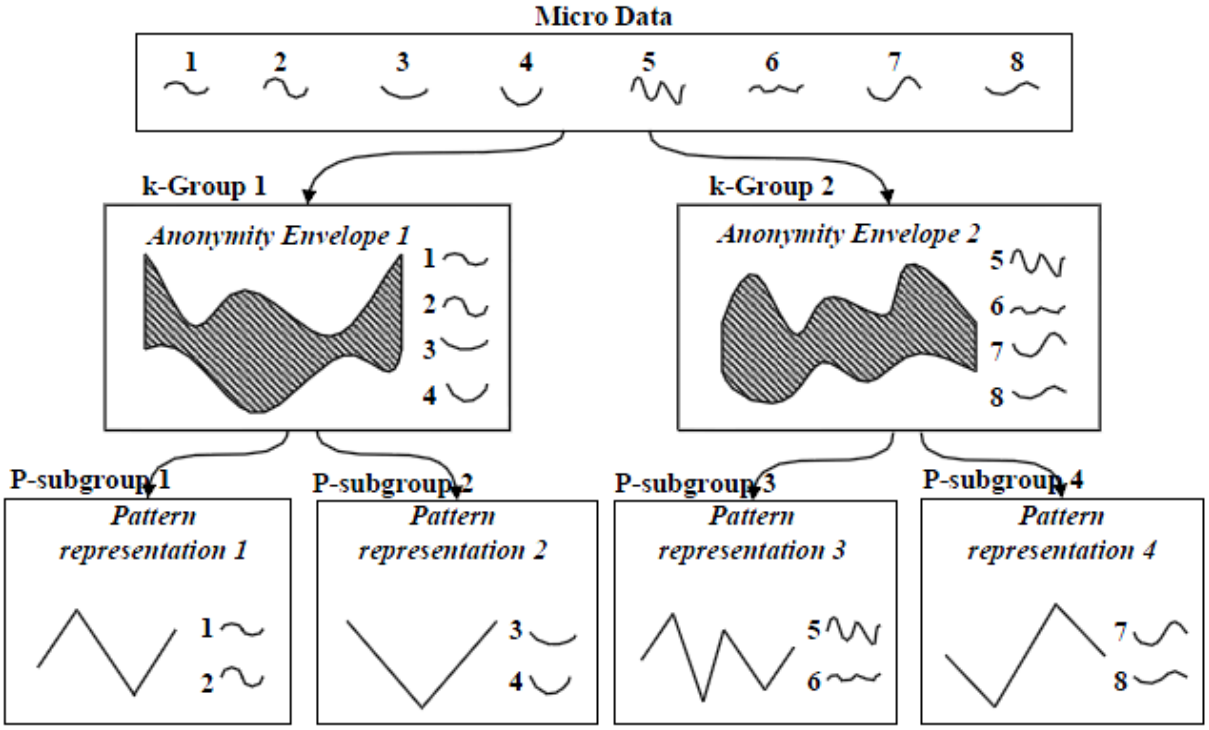
IP1:	172.17.32.223	(10101100.00010001.00100000.11011111)
IP2:	172.21.9.223	(10101100.00010101.00001001.00000101)

Anonimizuoti IP adresai

F(IP1):	113.6.154.129	(01110001.00000110. 10011010. 10000001)
F(IP2):	113.29.202.62	(01110001.00011101. 11001010.00111110)

(k,P)- anonimiškumo modelis: laiko duomenų struktūrą saugantis anonimizavimas. Xuan Shang [14] pasiūlė algoritmą, skirtą laiko duomenims anonimizuoti. Jo metode yra numatyta, kad kiekviena laiko eilutė yra publikuota trijuose komponentuose: apibendrinti QI požymiai, QI struktūros pavaizdavimas ir jautri informacija. Patikslinant, (k, P) - anonimiškumo modelis gali būti apibūdintas kaip konceptualus tradicinio k-anonimiškumo išplėtimas. Vis dėlto, (k, P) - anonimiškumas nepriklauso tradiciniam k-anonimiškumo algoritmui.

Kaip iliustruoja 8 pav., jų modelis garantuoja anonimiškumą dviejuose lygmenyse. Pirmame lygmenyje QI požymiai yra apibendrinti, kad įvykdytų tradicinį k-anonimiškumą, nepriklausomai nuo QI struktūros pavaizdavimo. Apibendrinimo rezultatai turi savyje daugiau padalijimų, žinomų kaip k-grupės. Pažymėtina, kad apibendrinti QI požymiai panėšėja į tuos tradiciniame k-anonimiškume. Antro lygmens įrašų anonimiškumas laikomas kiekvienoje k-grupėje. Bet kokiam įrašui *r* k-grupėje, jei ten egzistuoja bent jau P − 1 kiti įrašai, kurie turi tą patį struktūros pavaizdavimą kaip *r,* teigiama, kad P-anonimiškumas yra priverstinai sukuriamas visai šiai k-grupei. Todėl, galima padalinti k-grupę toliau į subgrupes, iš kurių kiekviena turi savyje mažiausiai P įrašų, turinčių identišką struktūros pavaizdavimą.



8 pav. (k,P)- anonimiškumo modelis

2.4. Išvados

Piktavalio tipas, kuris dažniausiai svarstomas literatūroje apie anonimines komunikacijos sistemas, yra galingas pasyvus globalinis piktavalis. Bendra saugumo praktika taiko priemones tokias, kad būtų bandoma gintis prieš blogiausią įvykių sekos variantą, manant, kad pasyvaus globalinio piktavalio buvimas atrodo pats apdairiausias ir išmintingiausias.

Pasyvaus globalinio piktavalio egzistavimas yra praktiškai nerealus dideliems tinklams, išdėstytiems internete. Geografinė ir tinklų pavaldumo įvairovė sukuria tokį bendrą tinklą, kad beveik neįmanoma, kad viena valstybė, ar net maža organizacija, galėtų kontroliuoti visus serverius plačiai paskirstytame tinkle. Ir netgi, jei piktavalis būtų gana galingas kontroliuoti visą tinklą, tai tada turėtų ir aktyviai valdyti tinklo srautą, kas yra nerealu.

Taigi, dėl šių argumentų labiau tikėtinas realiojo pasaulio piktavalis, galintis būti aktyviu piktavaliu su tiktai daliniu viso tinklo vaizdu.

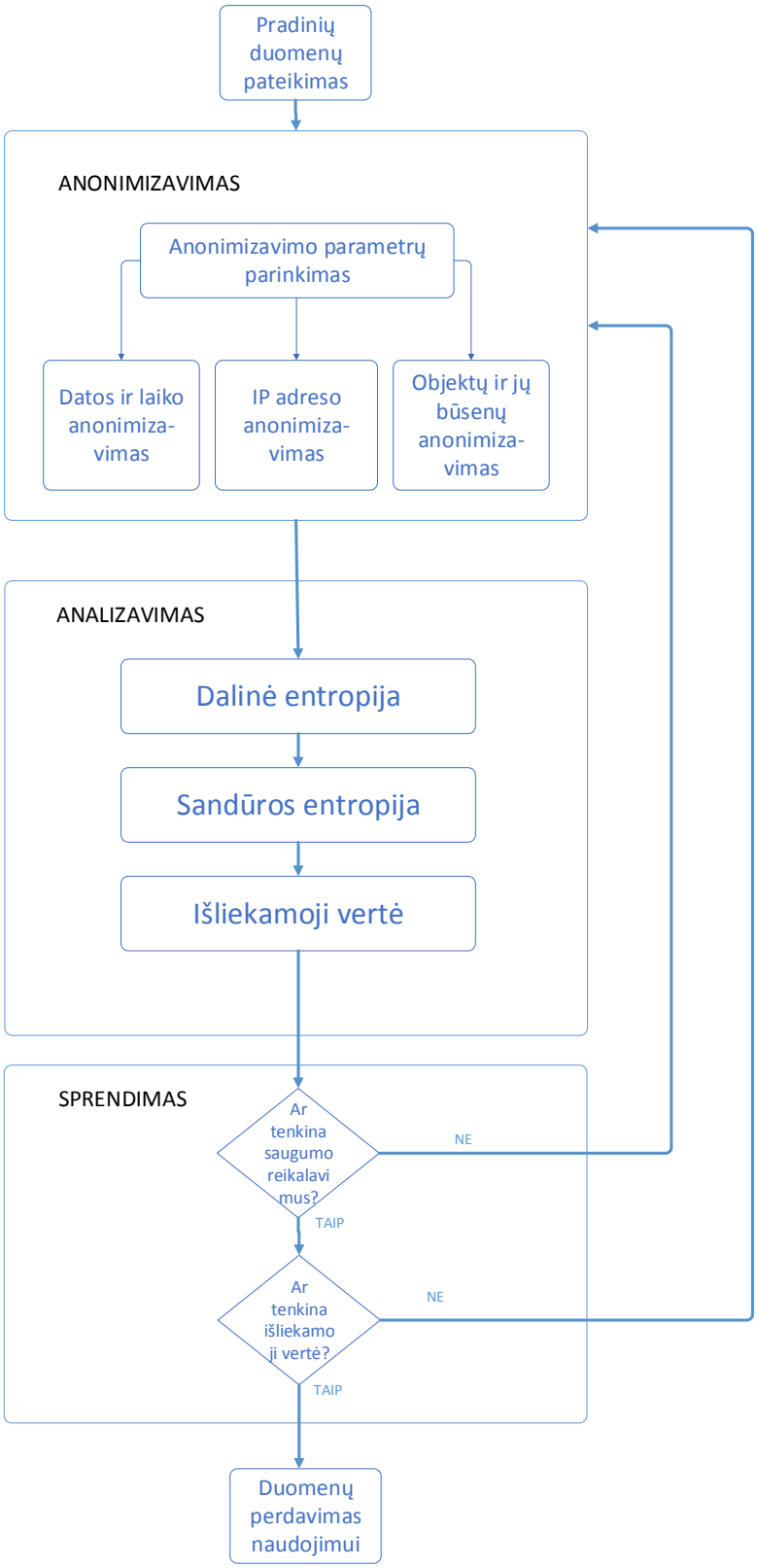
Apibendrinant apžvelgtus anonimizavimo algoritmus ir metodus juos galima pritaikyti skirtingai SNMP monitoringo įrašų informacijai. Priklausomai nuo jos formato, taikyti vieną ar kitą anonimizavimo būdą :

- **Datos ir laiko** (angl. time stamp) informacijos anonimizavimui tikslinga metodą "*Laiko duomenų struktūrą saugantis anonimizavimas"*, nes juo galima anonimizuoti laiko duomenis intervalais, kuriuos taip pat galima dar papildomai skaidyti į mažesnius pogrupius, jeigu reikia padidinti anonimiškumą;
- **Agento IP adreso** anonimizavimui tikslinga apžvelgtą IP adreso anonimizavimo algoritmą, kuris leidžia išsaugoti prefiksą ir leksikografinį jo eiliškumą;
- **Objektų identifikatorių** (toliau – OID) **ir jų būsenų** anonimizavimui tikslinga panaudoti ryšių nutraukimo algoritmą, kuris leidžia nutraukti ryšius tarp realių identifikatorių ir jautriosios informacijos.

Anonimizavus duomenis skirtingais metodais su galimybe keisti saugumo stiprumo lygį, galima pasiekti tokį jautrios informacijos apsaugos lygį, kuris apsaugos informacijos tiekėją nuo realių jo tinklo silpnų vietų analizės ir galimų atakų per jas. Kita vertus, būtų galima išsaugoti dar vis pakankamą kiekį informacijos, kuri neleistų atpažinti tikrųjų įrenginių ir jų būsenų, susieti su konkrečiais informacijos siuntėjais ar/ir gavėjais su realiais tinklo vartotojais ir/ar jų įrenginiais ir jų identifikaciniais numeriais, bet moksliniams bei analitiniams tyrimams vis dar naudingos informacijos. Naudojant skirtingus anonimizavimo modelius skirtingiems informacijos stulpeliams, priklausomai nuo jų formato, ir tai numatant programinės įrangos konfigūraciniame išpildyme, pagal poreikį būtų galima sustiprinti ar sumažinti saugumo lygius atitinkamuose SNMP žurnalų duomenų anonimizavimo etapuose, siekiant padidinti ar sumažinti duomenų išliekamąsias vertes. Tokiu būdu būtų suteikta galimybė patiems tinklų administratoriams ar jų valdytojams spręsti, kurią ir iki kokio lygio informaciją reikia anonimizuoti, prieš pateikiant ją viešam naudojimui. Atlikus naujojo sudėtinio algoritmo modeliavimą, visus jo galimus anonimizavimo lygius būtina patikrinti pagal apskaičiuojant individualias, sandūros bei dalines entropijas, kad būtų galima įvertinti kokio lygio yra tikėtinos grėsmės.

III. SNMP MONITORINGO ĮRAŠŲ ANONIMIZAVIMO MODELIS

Mūsų sukurtas modelis (9 pav.) sudarytas iš trijų pagrindinių dalių: anonimizavimas, analizavimas ir sprendimo priėmimas.



9 pav. SNMP monitoringo įrašų anonimizavimo modelis

Organizacija, kuri valdo SNMP monitoringo įrašus, nusprendusi perduoti viešam naudojimui savo valdomo tinklo SNMP TRAP pranešimų informaciją ir dėl to nenukentėti saugumo požiūriu, pagal šį modelį privalo atlikti šiuos duomenų apdorojimo etapus:

- Pradinių duomenų pateikimas;
- Anonimizavimas;
- Analizavimas;
- Sprendimas.

Pradinių duomenų pateikimas. Labai svarbu atrinkti tik pareikalautą informaciją, išfiltruojant tik konkrečius SNMP TRAP pranešimus. (pvz., vieno gamintojo tinklo įrenginių būsenas per nustatytą laikotarpį). Jeigu bus pateikta visas SNMP monitoringo žurnalas, tai didėja grėsmė saugumu, nes vykdant įvarius duomenų rūšiavimo ir lyginimo metodus, galima susieti realius įrenginius ir jų būsenas bei datas;

Anonimizavimas. Anonimizavimo procesas atliekamas prieš tai nustačius, kurią informaciją anonimizuoti ir iki kokio lygio didinti anonimiškumą. Galima anonimizuoti su keičiamais parametrais šią informaciją:

- *Datos žymė*. Jos anonimizavimo diapazonas gali būti keičiamas 2, 10, 30, 90, 180 dienų tikslumu (pvz., jeigu reali data yra 2013-05-10, tai anonimizavus 10 dienų tikslumu, data gali būti atsitiktinai generuojama nuo 2013-05-05 iki 2013-05-15). Naudojamas k,P-anonimiškumo modelis: laiko duomenų struktūrą saugantis anonimizavimo algoritmas;
- *Laiko žymė*. Jos anonimizavimo diapazonas gali būti keičiamas 2, 10, 30 minučių, 1, 3, 24 valandų tikslumu (pvz., jeigu realus laikas yra 10:00, tai anonimizavus 30 minučių tikslumu, laikas gali būti atsitiktinai generuojamas nuo 9:45 iki 10:15). Naudojamas k,P-anonimiškumo modelis: laiko duomenų struktūrą saugantis anonimizavimo algoritmas;
- *IP adreso prefiksas*. Jis gali būti anonimizuojamas oktetais nuo 0 iki 3. Pvz., anonimizavus IP adreso prefiksą 3 oktetais, bus visi realūs tinklo adresai pakeisti, tačiau bus išlaikyta realaus tinklo struktūra. Naudojamas IP adreso anonimizavimo algoritmas, kuris leidžia išsaugoti IP adreso prefiksą ir leksikografinį jo eiliškumą;
- *IP adresas*. Jis gali būti anonimizuojamas pasirenkant nuo 2 iki 253 skirtingais adresais. Kuo daugiau IP adresų naudojama, tuo didesnė įvairovė ir labiau didėja anonimiškumas. Naudojamas IP adreso anonimizavimo algoritmas, išsaugantis IP adreso prefiksą ir leksikografinį eiliškumą;
- *Objektų identifikatoriai ir jų būsenos*. Juos galima anonimizuoti naudojant ryšių nutraukimo algoritmą. Prireikus padidinti anonimiškumą, papildomai galima įtraukti

netikros informacijos nuo 5 iki 100 % esamos informacijos (pvz., naudinga, kai norima paslėpti tikruosius įvykius). Naudojamas ryšių nutraukimo algoritmas.

Analizavimas. Atlikus anonimizavimą, privalomas jau iš dalies pakeistos informacijos įvertinimas, kuris atliekamas pasitelkiant entropijos (objektų neapibrėžtumui sistemoje vertinimo) metodus. Apskaičiuojant entropijas, būtina patikrinti, ar nėra galimybės susieti realių tinklo įrenginių (pvz., IP adresų su jų objektais OID), todėl būtina apskaičiuoti kiekvieno duomenų stulpelio (pvz, data, laikas, IP adresas ir objektų reikšmės) individualias ir sandūros entropijas.

*Individualios entropijos apskaičiavimas***.** Individuali entropija apskaičiuoja anonimizuotų duomenų vieno stulpelio (pvz. IP adreso) neapibrėžtumą visoje anonimizuotoje sistemoje, vertinant tik tuos argumentus, kurie daro didžiausią įtaką anonimizuotai sistemai.

Pvz., IP adresas 23.239.43.192 pasikartojo 16 kartų iš 86 įrašų, taigi jo pasikartojimo dažnis visoje anonimizuotų įrašų aibėje yra 16/86 = 0.186. Naudojant dalinės entropijos metodą, atrenkami dominuojantys objektai, kurie daro didžiausią įtaką anomizuotai sistemai, apskaičiuojamos jų entopijos σ, sumuojami jų pasikartojimų dažniai β, bei kiekiai m ir gauti rezultatai sustatomi į (4) formulę. Joje:

σ – dominuojančių objektų entropija, $\sigma = -\sum_{i \in m} p(x_i) * log(p(x_i))$;

β - dominuojančių objektų pasikartojimo dažnių suma, $\beta = \sum_{i \in m} p(x_i)$;

m – dominuojančių objektų kiekis sistemoje.

σ = -1 * [(0.186 * log(0186)) + (0.1744* log(0.1744)) + (0.1395* log(0.1395)) + (0.1279* log(0.1279)) + (0.1163* log(0.1163))] = 0.6

β = 0,742

m=5

$$D = \frac{\sigma}{\beta/_m} = \frac{0.61}{0.742/_5} = 4.1$$

Tokiu pačiu būdu apskaičiuojama dalinė entropija visiems anonimizuotiems duomenų argumentams (pvz., data, laikas, IP adresas, OID ir jų būsenos).

Sandūros entropijos apskaičiavimas. Sandūros entropija leidžia apskaičiuoti įrašų porų neapibrėžtumą anonimizuotose duomenyse, kas yra labai svarbu, norint įvertinti anonimizuotų duomenų neapibrėžtumą lyginant kelis anonimizuotus stulpelius.

Pvz.: IP adresas 23.239.43.192 su OID microsoft.1.1.3.1.1 pasikartojo 3 kartus iš 86 įrašų, taigi jo pasikartojimo dažnis sandūroje, tarp IP ir OID anomizuotų aibių yra 0.0349. Tokiu būdu surandamos visos porų kombinacijos, sudaroma pasikartojimų dažnių matrica:

$p(x_i,y_i)$ =[0.0814, 0.0698, 0.0698, 0.0698, 0.0698, 0.0581, 0.0581, 0.0581, 0.0465, 0.0465, 0.0349, 0.0349, 0.0349, 0.0349, 0.0349, 0.0349, 0.0233, 0.0233, 0.0233, 0.0233, 0.0233, 0.0233, 0.0116, 0.0116].

Gautos reikšmės $p(x_i,y_i)$ sustatomos į (3) formulę ir gaunamas rezultatas yra H= 4.41407.

Sprendimas. Vadovaujantis atliktais tyrimais ir eksperimentais buvo sudaryta rekomendacinio pobūdžio anonimizuotų duomenų vertinimo skalė (žr. 3 lentelę), pagal kurią galima nustatyti esamą sistemos saugumo lygmenį, bei anonimizuotuose duomenyse išlikusią naudingą informaciją. Pagal šią vertinimo skalę būtina įvertinti visus anominizuotų duomenų stulpelius, po ko tikrinami individualios bei sandūros entropijų rezultatai.

Lentelė 3. Anonimizuotų duomenų išliekamos vertės ir saugumo lygio rekomendacijų skalė.

Sistemos neapibrėžtumas, D	**Išliekamoji duomenų vertė**	**Grėsmės lygis**
< 2.5	Daug realios informacijos, labai mažai dominuojančių objektų	Nesaugu
2.5 - 5	Optimalus realios informacijos kiekis	Patenkinama
5 - 8	Minimalus ryšys su realia informacija, labai daug dominuojančių objektų	Saugu
> 8	Nėra jokio ryšio su realia informacija, dominuojančių objektų kiekis m=n, kai $n>10^8$	Labai saugu

Jei bent vienas anonimizuoto stulpelio neapibrėžtumas gali sukelti grėsmę informacijos saugumui, rekomenduojama duomenis anonimizuoti iš naujo, prieš tai sustiprinus anonimizavimo kriterijus. Jeigu anonimizuoto stulpelio neapibrėžtumas pakankamas, bet netenkina anonimizuotų duomenų išlikusi vertė, rekomenduojama duomenis anonimizuoti iš naujo, prieš tai susilpninus anonimizavimo kriterijus. Tačiau jei anonimiškumas pakankamas ir duomenyse dar yra naudingos informacijos tyrimams, informacija gali būti perduodama trečioms šalims.

IV. EKSPERIMENTINĖ MODELIO REALIZACIJA IR REZULTATŲ ĮVERTINIMAS

Vadovaujantis nustatytais funkciniais ir nefunkciniais techniniais reikalavimais, buvo sukurtas programinės įrangos prototipas (toliau – PĮP), kuris gali atlikti iškeltus uždavinius ir patikrinti mokslinius tyrimus, bei greitai bei tiksliai pateikti rezultatus vartotojui suprantama kalba.

4.1. Realizuoti anonimizavimo metodai

Šis PĮP yra pritaikytas anonimizuoti SNMP žurnalų informaciją, bet gali būti lengvai perorientuotas ir prie kito tipo įvykių ar būsenų žurnalų (angl. logs), nes iš esmės juose visuose yra ta pati informacija:

- *Data ir laikas.* Šių laukų anonimizavimui taikomas "*Laiko duomenų struktūrą saugantis anonimizavimas"*, nes juo galima anonimizuoti laiko duomenis intervalais, kuriuos taip pat galima dar papildomai skaidyti į mažesnius pogrupius, jeigu reikia padidinti anonimiškumą.
- *IP adresas***.** Taikomas IP adreso anonimizavimo algoritmas, išsaugantis pasirenkamo ilgio prefiksą oktetais ir leksikografinį jo eiliškumą, t.y. gali būti išsaugota kompiuterinio tinklo topologija.
- *Objekto identifikatorius.* Anonimizuojama taikant algoritmą, kuris leidžia nutraukti ryšius tarp realių identifikatorių, jų būsenų ir IP adresų bei esant poreikiui įterpiant papildomo triukšmo (nerealių objektų).
- *Objekto būsena***.** Anonimizuojama taikant algoritmą, kuris leidžia nutraukti ryšius tarp realių identifikatorių, jų būsenų ir IP adresų bei esant poreikiui įterpiant papildomo triukšmo (nerealių objekto būsenų).

4.2. Realizuoti analizavimo metodai

Šis PĮP analizuoja SNMP žurnalų anonimizuotą informaciją, beto, pagal anonimizavimo savybes pateikiama likutinė informacija, t.y. kiek naudingos arba vis dar atitinkančios realybę informacijos yra išlikę šiuose anonimizuotose duomenyse. Analizei buvo pasitelktas anonimiškumo matavimo parametras – entropija. Mūsų anonimizuotos informacijos neapibrėžtumui įvertinti naudojama individuali, dalinė ir sandūros entropijos.

Individuali ir dalinė entropija apskaičiuojama kiekvienam anonimizuotam stulpeliui (žr. 2 lentelę), o ten, kur egzistuoja tamprus ryšys tarp dviejų stulpelių (kaip šiuo atveju IP-OID), vertinama ir sandūros (angl. joint) entropija [15].

Remiantis naujuoju neapibrėžtumo matavimu vadovaujantis daline entropija [16], išvadose pateikiama saugumo lygio skalė nurodanti rizikos grėsmę, priklausomai nuo esamo anonimiškumo lygio, kur yra vertinamas neapibrėžtumo laipsnis (D), skirtingų QI kiekis ir jų pasikartojimo dažnio pasiskirstymas visoje duomenų aibėje:

- NESAUGU. Labai mažas neapibrėžtumas (D< 2.5) leidžia lengvai atstatyti ryšius tarp tinklo įrenginių ir jo objektų, nes objektų pasikartojimo dažnis pasiskirstęs labai nevienodai, dominuojančių objektų yra labai mažai ir jie sudaro didelę dalį visos sistemos.

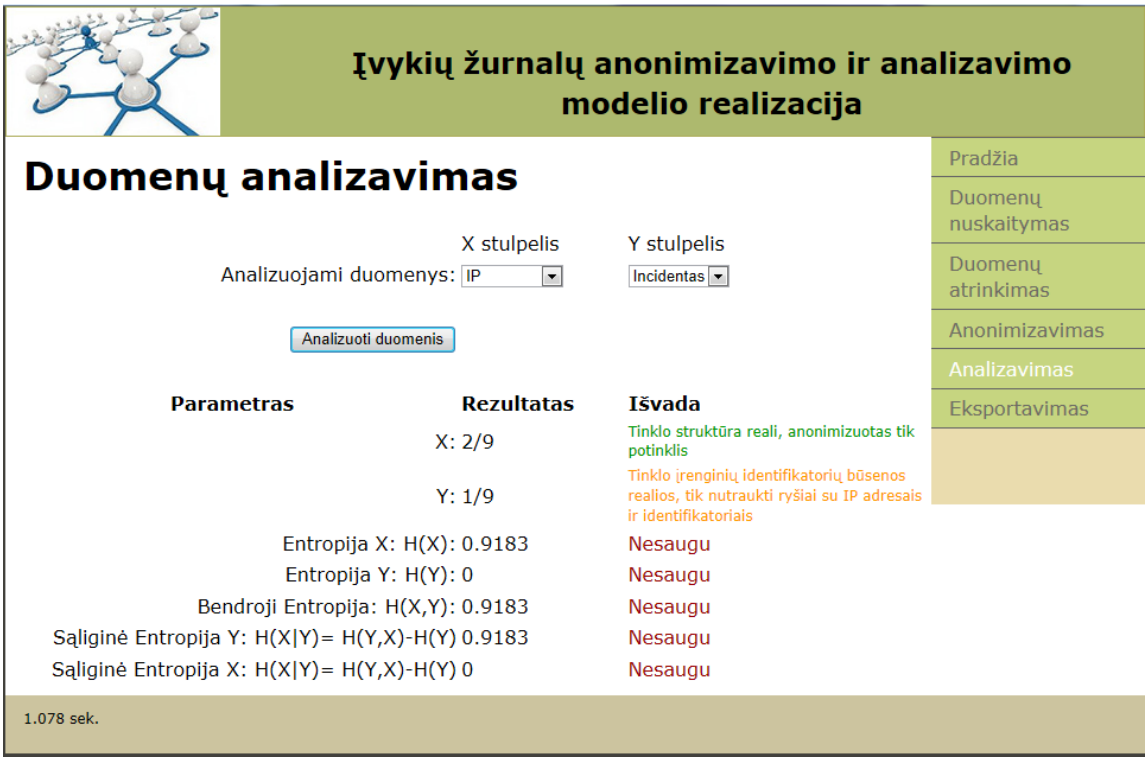
- NEPATIKIMA. Minimalus neapibrėžtumas (D=2.5-5) leidžia lengvai atstatyti ryšius tarp tinklo įrenginių ir jo objektų, nes pasikartojimo dažnis pasiskirstęs daugiau nei keliuose objektuose vienodai, bet jų yra mažai ir sudaro didelę dalį visos sistemos.

- SAUGU. Neapibrėžtumas pakankamas (D=5-8), kuris neleidžia lengvai atstatyti ryšius tarp tinklo įrenginių ir jo objektų, nes pasikartojimo dažnis pasiskirstęs daugiau nei keliuose objektuose vienodai, o dominuojančių pasikartojimų daugiau nei 3.

- LABAI SAUGU. Neapibrėžtumas labai didelis (D>8) neleidžia atstatyti ryšius tarp tinklo įrenginių ir jo objektų, nes objektų pasikartojimo dažnis pasiskirstęs labai tolygiai ir jų yra labai daug (artima ar lygu n).

4.3. Vartotojo sąsaja

Įrankis su kuriuo bus atliekami šiame aprašomo modelio naudingumo ir teisingo veikimo eksperimentai bei analizė, turi pakankamai patogią vartotojo sąsają ir visas reikalingas realizuotas funkcijas šiems tyrimams atlikti (10 pav.).



10 pav. Vartotojo sąsajos vienas iš langų (anonimizuotų duomenų analizavimo metodų išvados).

4.4. Eksperimento aprašymas

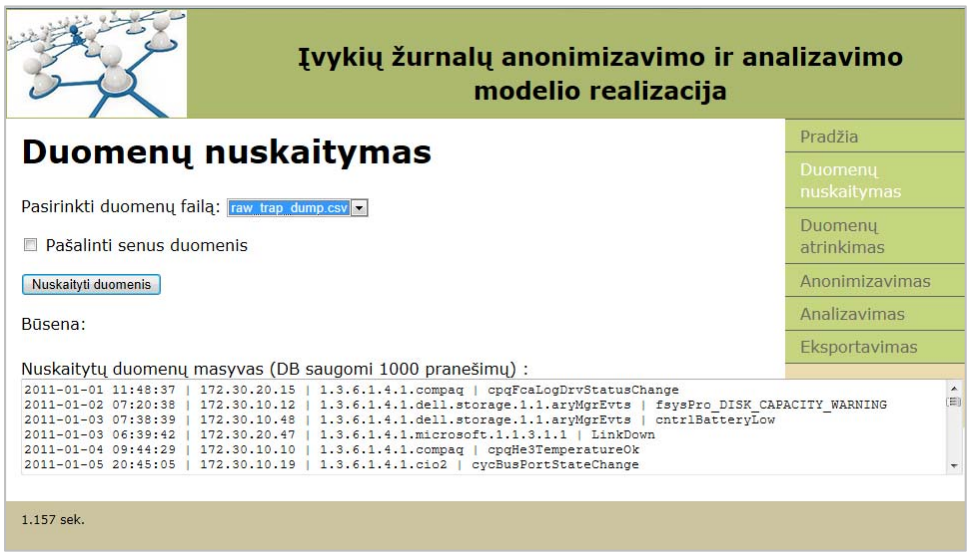
Eksperimento tikslas – įvertinti modelio anonimizavimo metodus ir išanalizuoti anominiškumo lygį. Mūsų sukurtas programinės įrangos prototipas (toliau – prototipas), leidžiantis sparčiai atlikti visus pristatomo modelio veiksmus nuo duomenų nuskaitymo iki išanalizuotų duomenų perdavimo trečioms šalims, įskaitant duomenų atrinkimą, anonimizavimą ir jų analizę, leidžia efektyviai įvertinti modelio pagrįstumą. Mūsų prototipas turi funkcijas, leidžiančias įtakoti anonimizavimo procesą, priklausomai kiek ir kokios realios informacijos norima palikti anonimizuotose duomenyse.

4.5. Eksperimento eiga

Pasirenkame užduotį “Kaip tankiai „lūžinėja“ organizacijos vietinis tinklas?“ ir bandome jį ištirti. Mūsų pagrindiniai įrankiai bus PĮP, testiniai duomenys. Skaitinių reikšmių vizualiam atvaizdavimui bus galima pasinaudoti bet kuriuo įrankiu, kad būtų galima nubraižyti kreives.

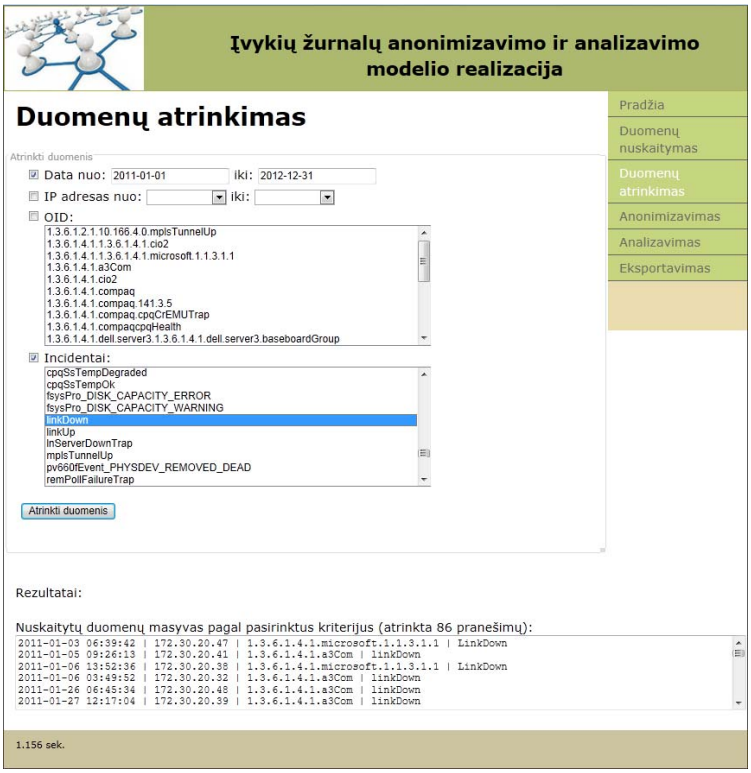
Testiniams žurnalų duomenims mes naudosime 1000 realių SNMP Trap pranešimų, kuriuos sugeneravo du potinkliai po 50 tinklinių įrenginių, turinčių savo unikalius IPv4 adresus.

Duomenų nuskaitymas. Programos paleidimas ir duomenų nuskaitymas vykdomas iš duomenų failo. Jeigu dėl kokios nors priežasties reikia pakartoti duomenų įkėlimą, pasirinkus failą raw_trap_dump.csv, galima pažymėti varnelę, „pašalinti senus duomenis“ ir viskas bus įkelta iš naujo. Jeigu reikia daugiau duomenų nei 1000 pranešimų, varnelės nededame ir kartojame veiksmą tiek kartų, kiek norime kad būtų duomenų, nes duomenų failas turi tik 500 įrašų. Taigi, kad įkelti 1000 įrašų pagal mūsų pirminius reikalavimus, šį veiksmą atlikome du kartus (11 pav.).



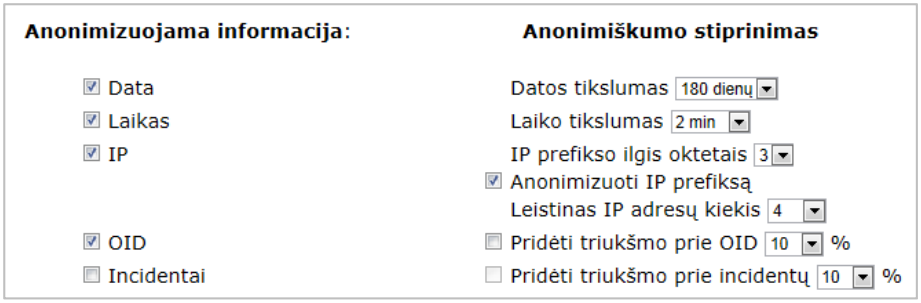
11 pav. PĮP vartotojo sąsajos duomenų nuskaitymo langas.

Duomenų atrinkimas. Šiame etape atrenkami tie duomenys, kurie yra mums reikalingi, t.y. visi incidentai LinkDown. Duomenų atrinkimas užtruko apie 1,5 sekundės, tad darbo našumui įtakos neturi. Duomenys atrinkti ir buvo rasti 86 pranešimai, kurie atitiko mūsų užsibrėžtą tikslą (12 pav.).

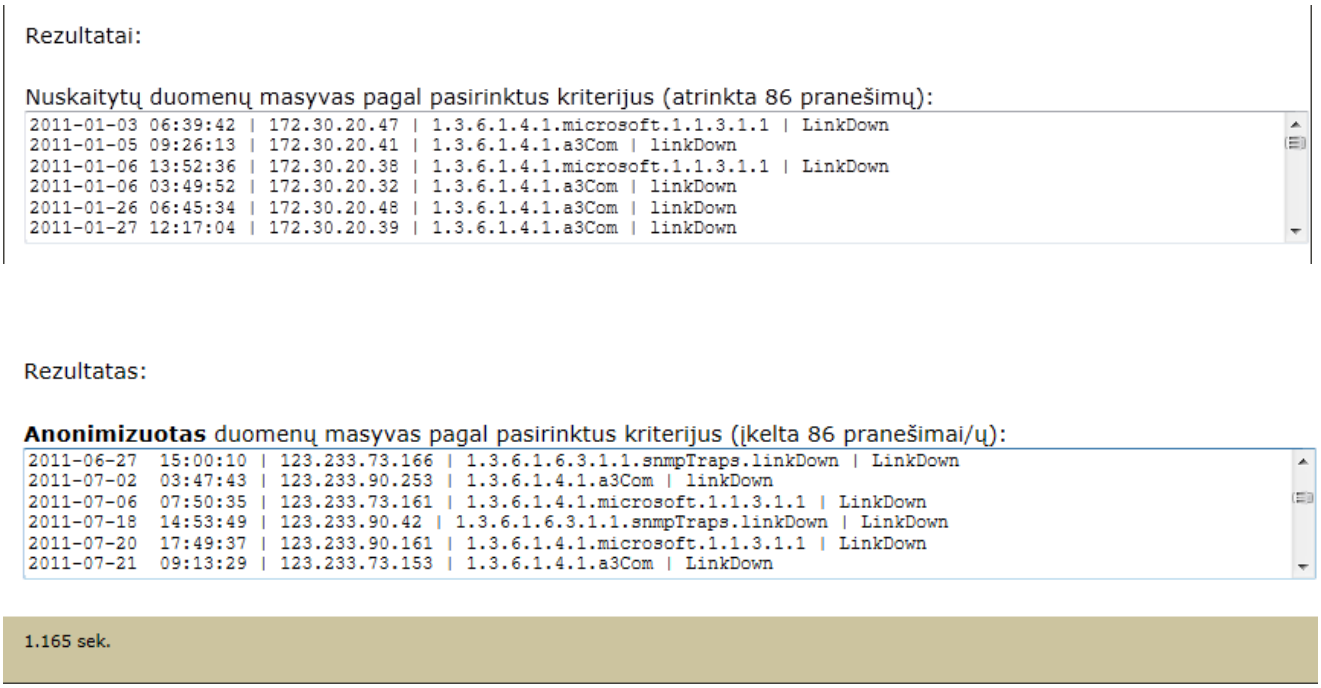


12 pav. PĮP vartotojo sąsajos duomenų atrinkimo langas.

Duomenų anonimizavimas. Šiame etape vertinama, kokių duomenų mes pateikti negalime ir kokius privalome anonimizuoti. Reikia nepamiršti ir šių duomenų galinio naudotojo, kad išliktų tyrimams reikalinga informacija. Šiuo atveju svarbu anonimizuoti realius IP adresus ir ryšį su objektų identifikatoriais (OID), bei reiktų anonimizuoti datos bei laiko informaciją, kad nebūtų galimybės atrinkti realių gedimų tendencijų ir jų pasikartojimo dažnių laiko atžvilgiu (13 pav.).



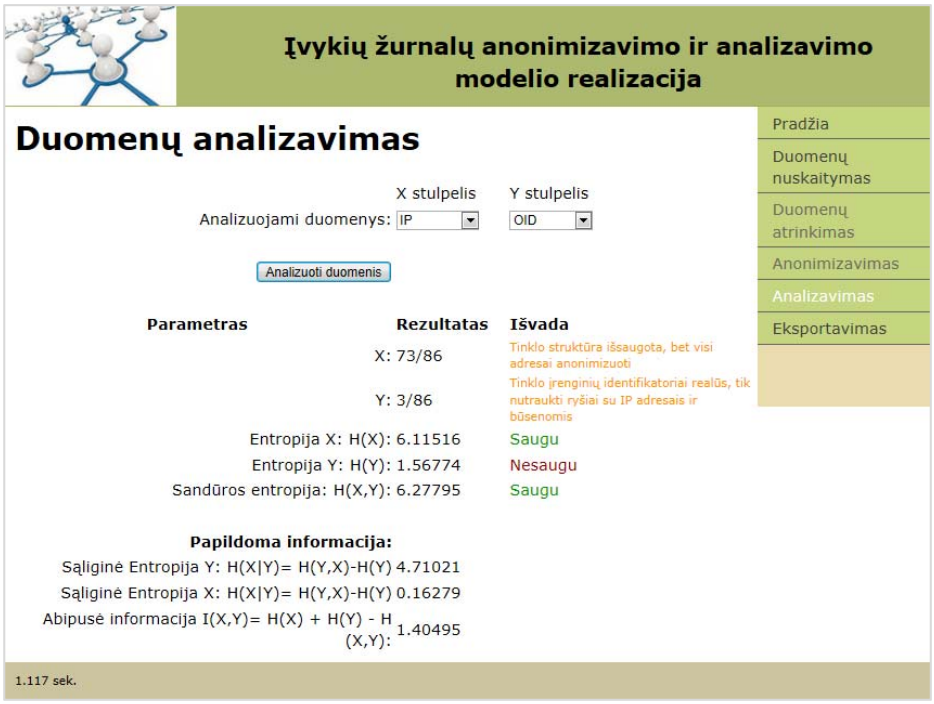
13 pav. PĮP vartotojo anonimizavimo parametrų nustatymo langas.



14 pav. PĮP vartotojo sąsajos duomenys prieš ir po anonimizavimo įvykdymo.

Iš rezultatų (14 pav.) matome, kad identiški duomenys vizualiai neturi jokio ryšio su prieš tai buvusiais duomenimis, išskyrus tai, kad liko paskutinis stulpelis “Incidentas“ nepakitęs, kurio mes ir nenorėjome anonimizuoti, nes klausimas buvo suformuluotas, kiek ir iš kokių tinklo įrenginių ir kokių jų objektų (OID) gaunami tokio tipo “LinkDown“ pranešimai.

Anonimizuotų duomenų analizavimas. Analizės etape reikia atlikti kiekvienos anonimizuotos poros vertinimą ir įsitikinti, kad entropijos laipsniai neperžengia kritinių verčių. O taip pat įvertinti, ar anonimizuotas stulpelis vis dar turi apčiuopiamos naudos vykdant numatytus tyrimus. PĮP informuos apie kiekvieną stulpelį atskirai iki kokio lygio jis buvo anonimizuotas (15 pav.).



15 pav. PĮP vartotojo sąsajos anonimizuotų duomenų analizės langas.

Papildoma informacija pateikiama per išvestines entropijos formules, kurios pateikiamos tik kaip papildoma informacija reikiamam sprendimui priimti.

Sąlyginė entropija X reiškia, koks bus entropijos laipsnis, jei būtų atskleistas stulpelis X. Šiuo atveju atskleidus IP adresus, anonimiškumo sistemoje beveik nelieka ir atvirkščiai, atskleidus OID reikšmes (Sąlyginė entropija Y), mes matome, kad entropijos laipsnis beveik nepakito ar bent jau priklauso kategorijai saugu.

Abipusė informacijos (angl. Mutual information) žema vertė reiškia, kad naudojant vieną reikšmę tarp šių dviejų stulpelių, kitos bet kurios reikšmės neapibrėžtumas mažėja ir atvirkščiai, kuo didesnė vertė, tuo didesnis neapibrėžtumo atsparumas vienos vertės atskleidimo kitai.

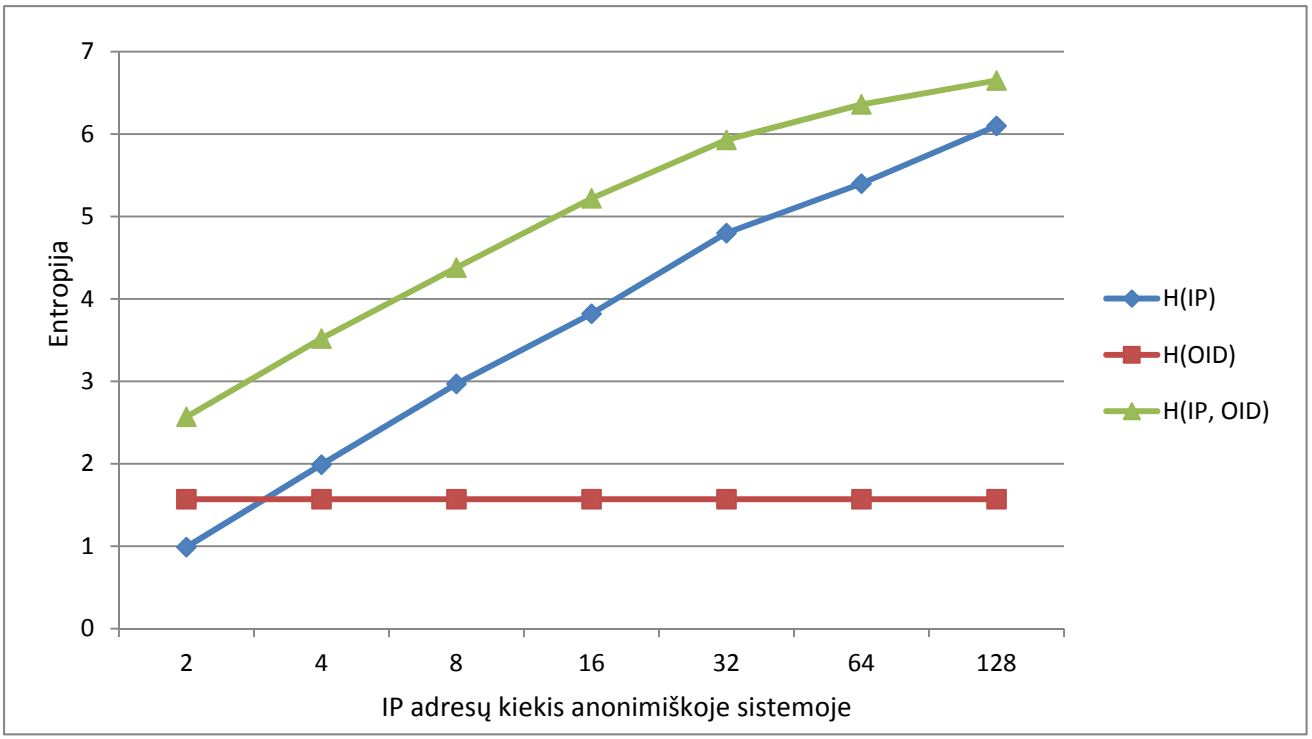
Anonimizuotų duomenų eksportavimas. Galimybė parsisiųsti anonimizuotus duomenis ir perduoti juos prašiusiam vartotojui (16 pav.)

16 pav. PĮP vartotojo sąsajos anonimizuotų duomenų analizės langas.

4.6. Analizė

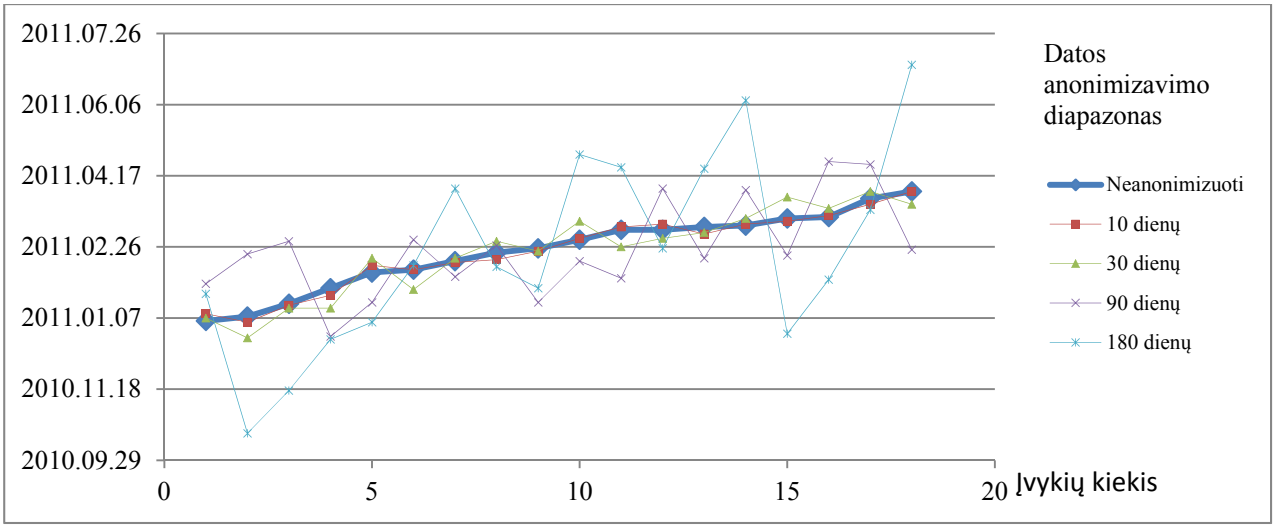
Pabandykime tuos pačius duomenis šiek tiek permodeliuoti ir atlikti palyginimus tarp tuos pačius kriterijus turinčių duomenų, keičiant tik vieno parametro anonimizavimo lygį, pvz. keičiant IP adreso anonimizavimą panaudojant 2, 8, 16, 32, 64, 128 galimus adresus. Dabar patikrinkime, kokias vertes gausime (17 pav.).

Atlikus mūsų pasirinktų anonimizuotų duomenų analizę, galime lengvai nustatyti, kad pagal mūsų sudarytą vertinimo skalę IP adresų neapibrėžtumo laipsnis atitinka kriterijui SAUGU, o OID skirtingų objektų kiekis tik 3, tad ir neapibrėžtumo laipsnis tik NESAUGU, bet sandūros entropija atitinka kriterijų SAUGU, nes ryšiai tarp tikrų tinklų įrenginių ir jų objektų pasiskirstę pakankamai tolygiai, o jų dominuojančių objektų daugiau nei 3 (0.0814, 0.0698, 0.0698, 0.0698, 0.0698).



17 pav. Entropijos pokytis, keičiant anonimizavimo procese naudojamų IP adresų kiekį.

Išliekamoji vertė anonimizuotame duomenų masyve iš dalies išsaugota, nes IP adresų potinklių struktūra išsaugota su jų leksikografiniu eiliškumu, OID visi realūs, tik nutraukti ryšiai su realiais tinklo įrenginiais.



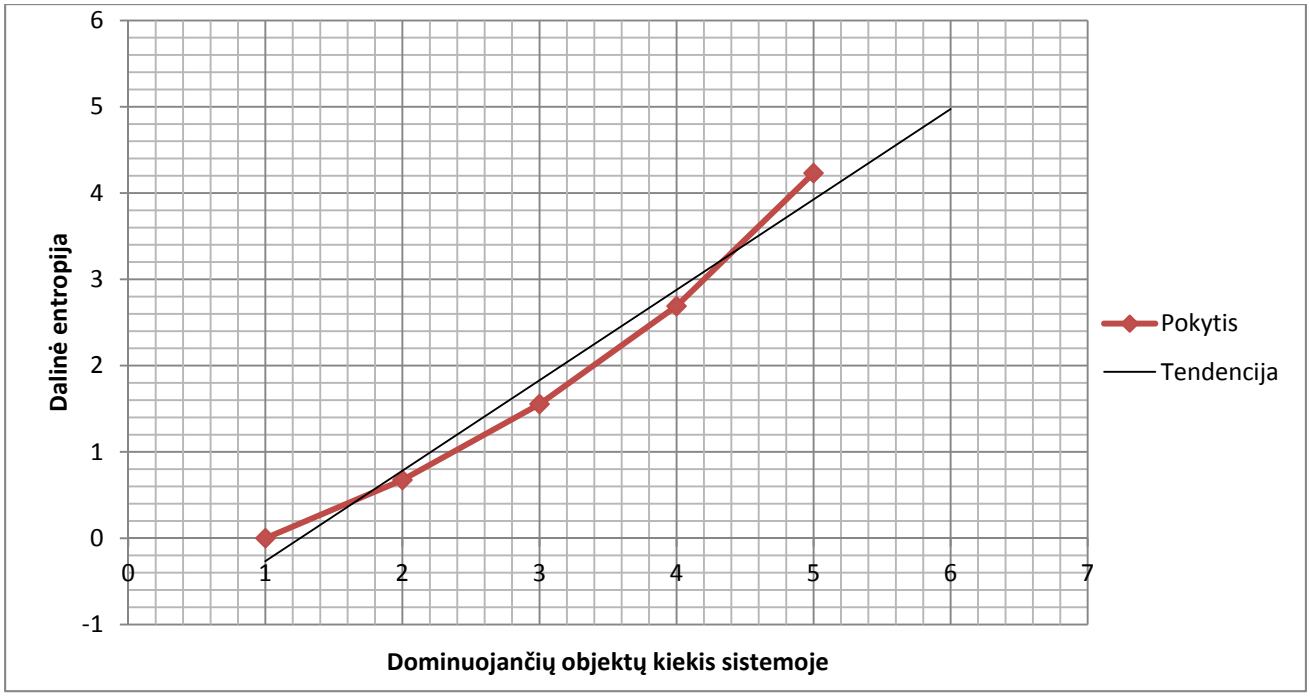
18 pav. Priklausomai, koks pasirinktas datos anonimizavimo diapazonas, gautų reikšmių lyginimas su realiais duomenimis.

TRAP pranešimų data ir laikas gali būti anonimizuojamas minimaliai (pvz., 2 dienų tikslumu). Tokiu būdu reali įvykio data nežinoma, bet tendencijas ir/ar įvykių pasikartojimus galima identifikuoti. Esant poreikiui, naudojantis prototipo galimybėmis, galima anonimizuoti net 180 dienų tikslumu (neapibrėžtumo laipsnis daugiau nei 8, t.y. LABAI SAUGU), bet tada visiškai prarandamas ryšys su įvykio realia data (18 pav.).

Panaudojant dalinės entropijos metodą, galime stebėti, kokia įtaka daroma sistemos bendram anonimiškumui, kintant dominuojančių objektų kiekiui m ir jų suminiam pasikartojimo dažniui β. Kaip matome, jie yra tiesiogiai susiję ir tendencijos tiesė rodo, kad kuo didesnis dominuojančių objektų kiekis anonimizuotoje sistemoje, tuo didesnis yra bendras sistemos anonimiškumas (19 pav.).

Lentelė 4. Įtaka anonimiškai sistemai, kintant dominuojančių objektų kiekiui ir jų pasikartojimo dažniui joje

Objektų pasiskirstymas sistemoje	**Dominuojančių objektų pasiskirstymas**	**m**	**β**	**D**
0.46, 0.46, 0.02, 0.02, 0.01, 0.01, 0.005, 0.005, 0.005, 0.005	0.46, 0.46	2	0,92	0.7
0.31,0.3,0.3,0.04,0.02,0.01,0.005,0.005,0.005,0.005	0.31,0.3,0.3	3	0.91	1.6
0.22, 0.21, 0.21, 0.21, 0.079, 0.05, 0.006, 0.005, 0.005, 0.005	0.22, 0.21, 0.21, 0.21	4	0.85	2.7
0.19, 0.19, 0.19, 0.19, 0.18, 0.022, 0.01, 0.01, 0.01, 0.008	0.19, 0.19, 0.19, 0.19, 0.18	5	0.94	4.2



19 pav. Anonimiškos sistemos neapibrėžtumo kitimas, kintant dominuojančių objektų kiekiui

V. IŠVADOS

1. Analizės etape buvo išnagrinėta kas yra anonimiškumas, kvazi-anonimiškumas ir jiems priešiški modeliai. Taip pat ištirta, kaip veikia ir kokią informaciją saugo SNMP protokolas, kokio tipo bei formato informacija saugoma monitoringo įrašuose ir protokolo pranešimuose. Pagal tai buvo išanalizuoti ir parinkti anonimizavimo algoritmai, kurie labiausiai tinkami, kai reikia įtakoti monitoringo įrašuose saugomų duomenų anonimizavimo procesą, norint pasiekti lankstumą ir išsaugoti tą informaciją, kuri reikalinga viešam naudojimui ir slėpti pagal pasirenkamą saugumo lygį tą informaciją, kurios piktavališkas panaudojimas gali pridaryti žalos informaciją pateikusiai organizacijai.
2. Anonimizuotai informacijai įvertinti buvo ištirti anonimiškos sistemos vertinimo metodai: dalinė, sandūros ir sąlyginė entropijos ir sudaryta rekomendacinio pobūdžio vertinimo skalė, leidžianti IT saugos specialistui daryti išvadas apie anonimizuotų duomenų saugumą ir išlikusios naudingos informacijos būseną.
3. Sukurtas SNMP monitoringo įrašų anonimizavimo modelis, kurio pagalba galima atrinkti reikiamus duomenis, juos anonimizuoti pagal pasirenkamus anonimiškumo didinimo ar mažinimo kriterijus, vėliau patikrinti anonimizuotos informacijos lygį keliais būdais ir pagal gautas vertes ir modelio teikiamas rekomendacijas ir esant tenkinamam anonimiškumo rezultatui perduoti informaciją trečioms šalims naudojimui, arba grįžti į anonimizavimo fazę ir keisti anonimizavimo parametrus, kad būtų pasiektas balansas tarp saugumo ir išlikusios naudingos informacijos anonimizuotoje sistemoje.
4. Pasinaudojus šiuo modeliu galima pasiekti tokį jautrios informacijos apsaugos lygį, kuris leistų apsaugoti informacijos tiekėją nuo jo tinklo silpnų vietų analizės ir galimų atakų. Kita vertus, dar būtų galima išsaugoti pakankamą kiekį informacijos, kuri būtų naudinga moksliniams bei analitiniams tyrimams. Eksperimentų ir analizės metu buvo nustatyta, kad anonmizuojant IP adresą, būtina naudoti bent 32 skirtingus adresus, kad būtų pasiektas saugą užtikrinantis lygis (neapibrėžtumo laipsnis lygus 5). Datos ir laiko stulpelių anonimizavimas turi vis dar išliekamosios vertės (pranešimų tendencija, pasikartojimo dažnis), kai anonimizavimui naudojama ne daugiau kaip 30 dienų anonimizavimo diapazonas.
5. Informacijos išliekamoji vertė anonimizuotame monitoringo įraše iš dalies išsaugota ir dar tinkama naudojimui, jei sistemos neapibrėžtumo laipsnis D ne daugiau kaip 8. Sistemos neapibrėžtumo laipsnis D tiesiogiai priklauso nuo dominuojančių objektų entropijos laipsnio σ, jų kiekio m ir dominuojančių objektų pasikartojimų dažnio sumos β. Kuo didesnis σ ar m, tuo geresnis sistemos anonimiškumas, o kuo didesnė dominuojančių objektų pasikartojimų dažnio suma β, tuo prastesnis sistemos anonimiškumas.

VI. PANAUDOTA LITERATŪRA

1. J. Schonwalder, A. Pras, M. Harvan, J. Schippers, and R. van de Meent, " SNMP Traffic Analysis: Approaches, Tools, and First Results", Proc. 10th IFIP/IEEE International Symposium on Integrated Network Management, 2007.

2. M. Harvan, J. Schonwalder, " Prefix- and Lexicographical-order-preserving IP Address Anonymization", 10th IEEE/IFIP Network Operations and Management Symposium p.519–526, 2006.

3. O'CONNOR L., „ On blending attacks for mixes with memory", In Proceedings of Information Hiding Workshop (IH), p. 39–52, 2005.

4. Andreas Pfitzmann, Marit Hansen, "Anonymity, Unlinkability, Unobservability, Pseudonymity, and Identity Management – A Consolidated Proposal for Terminology", 2005.

5. Latanya Sweeney, "K-anonymity: a Model for Protecting Privacy", 2002.

6. M. F. Chi-Yin Chow, X. Liu, "A Peer-to-Peer Spatial Cloaking Algorithm for Anonymous Location-Based Services", Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems. Arlington, Virginia, USA, 2006.
7. B. Gedik, Ling Liu, "Protecting Location Privacy with Personalized k-Anonymity: Architecture and Algorithms.", IEEE Transactions on Mobile Computing. University of Illinois at Urbana-Champaign P.1-18, 2008.
8. Raymond, J.-F., "Traffic analysis: Protocols, attacks, design issues, and open problems. In Proceedings of Designing Privacy Enhancing Technologies: Workshop on Design Issues in Anonymity and Unobservability", 2000.
9. Goldberg I, "A pseudonymous communications infrastructure for the internet. Ph.D. dissertation.", University, of California, Berkeley, Berkeley, CA, 2000.
10. A.Serjantov, R. E. Newman, "On the anonymity of timed pool mixes", In Proceedings of the Workshop on Privacy and Anonymity Issues in Networked and Distributed Systems, Kluwer, Athens, Greece, p. 427–434., 2003.
11. M. Reiter , A. Rubin, "Crowds: Anonymity forWeb transactions. ACMTrans." Inform. Syst. Sec. 1, 1, 66–92. , 1998.
12. Weijia Yang, Sanzheng Qiao, "A novel anonymization algorithm: Privacy protection and knowledge preservation", 2009

13. Aggarwal, C. C., Yu, P. S., “A general survey of privacy-preserving data mining models and algorithms. Privacy-preserving data mining“ p. 11–52, 2008.
14. Xuan Shang, Ke Chen, Lidan Shou, Gang Chen, Tianlei Hu, „(k,P)-Anonymity: Towards Pattern-Preserving Anonymity of Time-Series Data“, CIKM’10, October 26–30, Toronto, Ontario, Canada, 2010
15. RFC 3411 – 3418 standartų apie SNMP protokolą aprašymas [prieiga per internetą] http://www.rfc-editor.org/rfc/std/std62.txt, žiūrėta 2012-01-20
16. M. Edman, B. Yener, “On Anonymity in an Electronic Society: A Survey of Anonymous Communication Systems“, ACM Computing Surveys, Vol. 42, No. 1, Article 5, 2009.
17. M. Bezzi, “An Entropy based method for Measuring Anonymity“, Proceedings of the third international conference on security and privacy in communication networks. Nice, France. p. 28 – 32., 2007
18. J. Xu, J. Fan, M. H. Ammar, “Prefix-Preserving IP Address Anonymization: Measurement-based Security Evaluation and a New Cryptography-based Scheme”, Proceedings of the 10th IEEE International Conference on Network Protocols (ICNP’02), 2002.
19. Paul Meagher,, Calculating Entropy for Data Miners“, [prieiga per internetą] http://onlamp.com/pub/a/php/2005/03/24/joint_entropy.html?page=1, žiūrėta 2013-01-15
20. Guihua Duan, Weiping Wang+, Jianxin Wang, Luming Yang, “A new anonymity measure based on partial entropy“, IEEE Communications Society publication in the ICC, 2008.
21. V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, “k-anoonymity“, Springer US, Advances in Information Security, 2007.

VII. PRIEDAI

1. Programinės įrangos prototipo techninė specifikacija;
2. Programinės įrangos prototipo informacinė posistemė;
3. Bartas Aleksandravičius, Darius Matulis,“ SNMP žurnalų anonimizavimo tyrimas“, Informacinės Technologijos, XVIII tarpuniversitetinė magistrantų ir doktorantų konferencija, Konferencijos pranešimų medžiaga Psl. 8-12; ISSN 2029-249X, 2013-04-25.

1 priedas

PROGRAMINĖS ĮRANGOS TECHNINĖ SPECIFIKACIJA

Paskirtis

Kuriamas programinės įrangos prototipas (toliau PĮP) skirtas anonimizuoti SNMP monitoringo įrašus ir apskaičiuoti jų anonimiškumo lygmenį pagal nustatytus anonimiškumo matavimo vienetus. PĮP turi galimybę keisti anonimizavimo lygį vartotojui pasirenkant kuriuos informacijos laukus anonimizuoti, o kuriuos palikti nepakeistus. Po duomenų anonimizavimo PĮP automatiškai įvertins anonimiškumo lygį, pateiks jį skaitine išraiška. Jei anonimizavimo lygis bus per žemas, PĮP informuos apie tai vartotoją.

PĮP funkcijos

Duomenų nuskaitymas ir saugojimas. PĮP nuskaito SNMP žurnalų informaciją iš tekstinių failų arba iš duomenų bazės.

Reikiamos informacijos atrinkimas. priklausomai kokios informacijos reikia, išrenkama tik reikiama iš visų monitoringo įrašų aibės.

Anonimizavimo parinkimas. priklausomai koks anonimizavimo algoritmas buvo pasirinktas, galima papildomai pasirinkti kuriuos SNMP žurnalo žinutės laukus anonimizuoti. Nuo to priklauso anonimizavimo lygio stiprumas ir jo vykdymo laikas.

Anonimizavimo lygio matavimas. pagal nustatytus matavimo vienetus įvertinami anonimizuoti duomenys, pateikiamos vartotojui jų vertės ir rekomendacijos, jei anonimizavimo lygmuo yra per mažas ar per didelis.

PĮP reikalavimai

Funkciniai reikalavimai

- PĮP privalo nuskaityti nurodytus SNMP žurnalus iš tekstinių laikmenų arba iš duomenų bazės;
- PĮP privalo anonimizuoti SNMP žurnalų informaciją pagal pasirenkamą lygmenį;
- PĮP privalo apskaičiuoti SNMP žurnalų informacijos anonimizavimo lygmenį pagal nustatytus matavimo vienetus;
- PĮP privalo teikti vartotojui išvadas ir rekomendacijas apie atlikto SNMP žurnalų informacijos anonimizavimo saugos lygmenį;

- PĮP privalo leisti eksportuoti anonimizuotus duomenis.

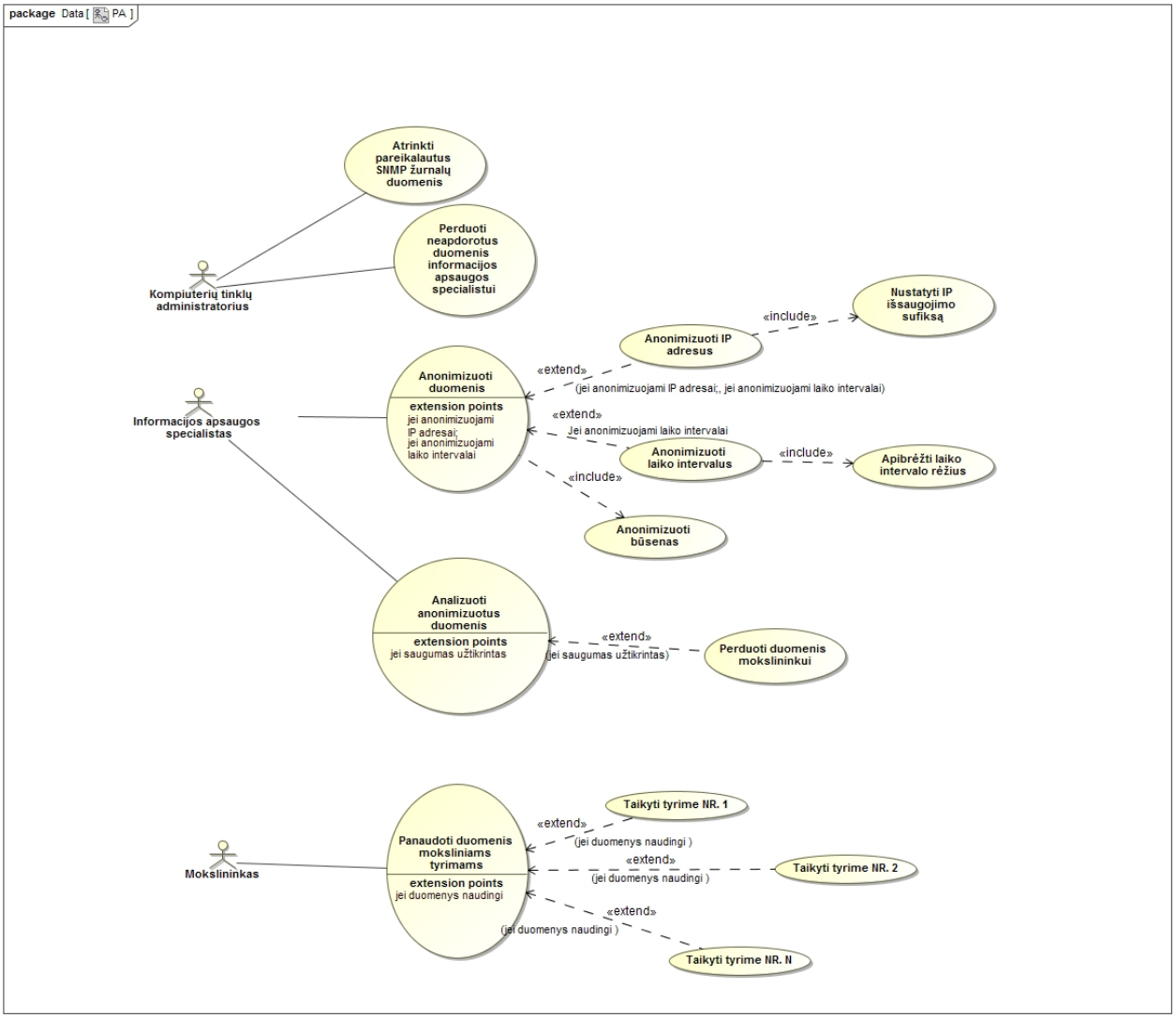
Nefunkciniai reikalavimai

- PĮP privalo būti saugus ir patikimas, nes sistema saugo organizacinių privačius duomenis, tinklų būsenas bei struktūras;
- Anonimizavimo lygmuo turi apsaugoti organizacijų teikiamą informaciją iki tokio lygio, kad nebūtų galimybės jai pačiai pakenkti piktavališkais tikslais;
- Anonimizavimo lygmuo turi apsaugoti organizacijų teikiamą informaciją iki tokio lygio, kad anonimizuota informacija vis dar turėtų tiriamąją vertę moksliniu ir analitiniu aspektu;
- PĮP turi užtikrinti, kad duomenis būtų korektiškai anonimizuoti, kad būtų išvengta duomenų praradimo, dėl anonimizavimo;
- PĮP turi dirbti patikimai ir optimaliai;
- dėl saugumo reikalavimų PĮP gali būti prieinama tik iš organizacijos vidaus tinklo;
- PĮP vartotojo sąsaja turi būti intuityvi, paprasta, aiški;
- PĮP saugumo lygio nustatymo laukai turi būti patogiai išdėstyti;
- PĮP klaidų pranešimai turi būti informatyvūs ir neapsunkinti technine informacija.

PĮP INFORMACINĖ POSISTEMĖ

Panaudojimo atvejų diagrama

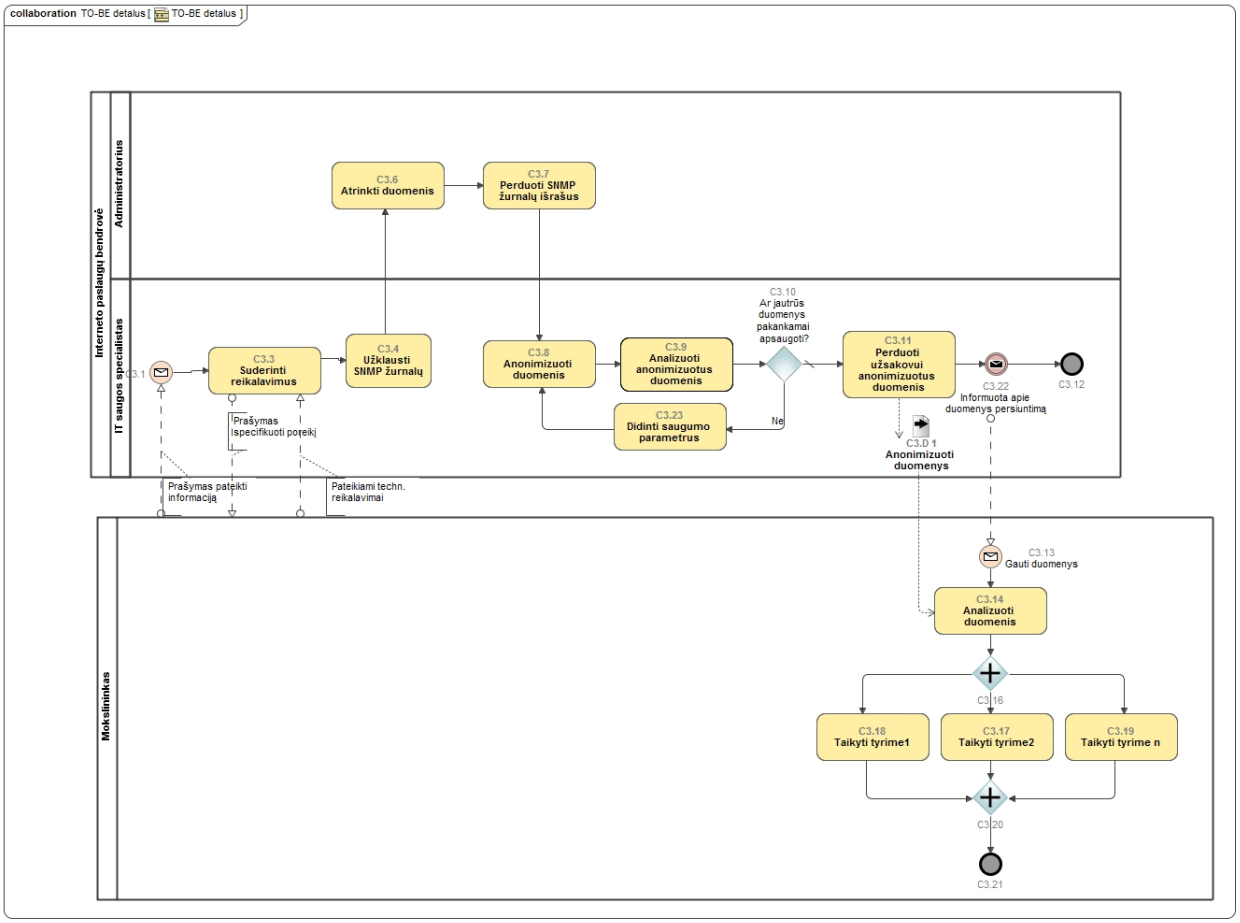
Iš PĮP panaudojimo diagramos matome, kad programa iš esmės yra skirta tik organizacijos kompiuterinio tinklo SNMP žurnalų anonimizavimo algoritmų tyrimui ir gautų rezultatų analizavimui. Matome, kad pagrindiniai programos vartotojai yra kompiuterių tinklų administratorius ir IT saugos specialistas, kurie tiesiogiai naudoja programos teikiamas funkcijas (duomenų atrinkimas, anonimizavimo parametrų parinkimas, anonimizavimas, analizavimas).



20 pav. PĮP panaudojimo atvejų diagrama, demonstruojanti visus programos naudotojus ir galimas funkcijas.

Veiklos diagrama

Veiklos diagramoje (21pav.) pavaizduotos galimos vartotojo veiklos programoje ir kaip yra susiję gauti rezultatai bei kada ir kur yra saugomi gauti visų stadijų duomenys.



21 pav. Veiklos diagrama.

Kaip matome iš veiklos diagramos, egzistuoja 3 pagrindinės veiklos (duomenų atrinkimas, anonimizavimas ir analizavimas), kitos yra šalutinės, kurios neturi įtakos modelio veikimui ir jo rezultatams, bet pateikus visą bendrą vaizdą, galima geriau įsivaizduoti, kurioje vietoje pradedamas ir baigiamas PĮP veikimas ir naudojimas.

Klasių diagrama

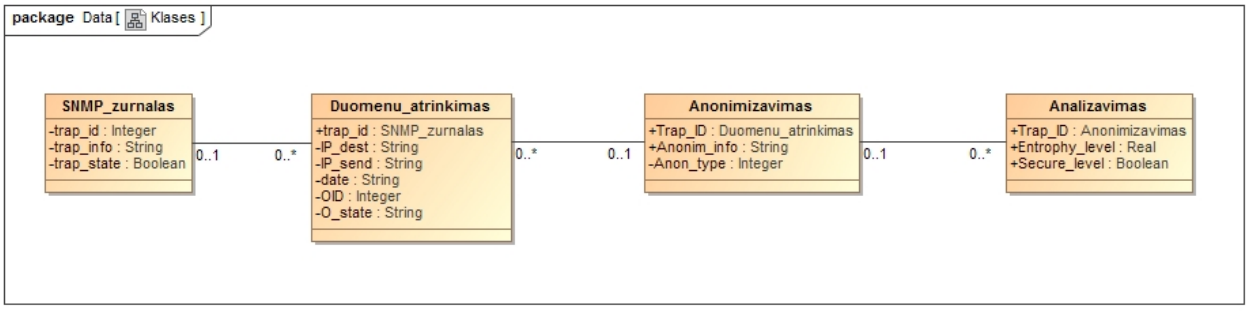
Klasių diagrama (22 pav.) susideda iš 5 tarpusavyje susijusių klasių su savo metodais ir kintamaisiais:

- Klasė **SNMP_zurnalas** skirta pirminės informacijos nuskaitymui iš SNMP žurnalo failo, reikiamų anonimizuoti laukų aptikimo jame ir sukėlimo į duomenų bazės reikiamas lenteles, kad vėliau būtų galima atrinkti reikiamus laukus anonimizavimo procesams.

- Klasė **Duomenu_atrinkimas** skirta reikiamų duomenų išfiltravimui iš visų monitoringo įrašų žurnalo.

- Klasė **Anonimizavimas** skirta atrinktos informacijos anonimizavimui pagal vartotojo pasirinktą konfigūraciją bei kriterijus. Priklausomai kokia informacija turi būti anonimizuota, naudojamas tam tikras anonimizavimo algoritmas.

- Klasė **Analizavimas** atlieka visus matematinius skaičiavimus, skirtus įvertinti anonimizuotų duomenų anonimiškumo lygį (apskaičiuojami verčių pasikartojimo dažniai, vienetinės, sandūros, dalinės entropijos) ir pateikiamas gautas rezultatas.



22 pav. PĮP klasių diagrama.

SNMP žurnalų anonimizavimo tyrimas

Bartas Aleksandravičius

KTU Informatikos fakultetas, Studentų g. 50, Kaunas, Lietuva
bartas.aleksandravicius@stud.ktu.lt

Darius Matulis

KTU Informatikos fakultetas, Studentų g. 50, Kaunas, Lietuva
darius.matulis@ktu.lt

Santrauka. Kompiuterinių tinklų valdytojai plačiai naudoja paprastąjį tinklų valdymo protokolą, kuris padeda aptikti tinklų anomalijas ir įvairius gedimus. Norint tobulinti patį protokolą ir/ar stebėti specifinės tinklinės įrangos elgseną, būtina disponuoti šio protokolo registracijos žurnalais. Siūlomas modelis leidžia anonimizuoti šiuos žurnalus, išanalizuoti jų anoniniškumo laipsnį ir perduoti trečiosios šalis be grėsmės pakenkti disponuojamo tinklo saugumui.

Reikšminiai žodžiai: SNMP įvykių žurnalai; anoniniškumas; entropija.

I. ĮVADAS

Realiame pasaulyje interneto paslaugų tiekėjai ir tinklinių sistemų administratoriai disponuoja dideliais informacijos kiekiais, susijusiais su tinklo inventoriaus veikimo būsenomis, incidentais, įvairiausio pobūdžio informaciniais pranešimais, kurie parodo ir padeda sistemų administratoriams, kai sistema nukrypsta nuo natūralios būsenos.

Problema iškyla, kai norima tą informaciją perduoti kitiems, ar kitų sistemų administratoriai, analitikai, inžinieriai ar mokslininkai norėdami geriau suprasti, išanalizuoti ir patobulinti informavimo protokolą ar metodą, tam neturi visos informacijos disponavimo laisvės. Natūraliai kyla klausimas, kodėl taip yra?

Sistemų administratoriai, interneto paslaugų tiekėjai nelinkę skleisti informaciją viešam naudojimui apie savo valdomų tinklų įrenginių būsenų, adresų, siuntėjų, gavėjų, gedimų, klaidų ir kitos informacijos iš esmės tik todėl, kad nebūtų galima analizuoti jų tinklų informacijos ir nebūtų kuriamos priemonės, nukreiptos prieš jų pačių tinklus bei jų silpnąsias vietas. Kitaip sakant, pagrindinis argumentas yra saugumo grėsmės.

Jeigu sistemų administratoriai ar interneto tiekėjai turėtų galimybę anonimizuoti savo disponuojamų tinklų informaciją iki tokio lygio, kad nebūtų galima atsekti realių įrenginių adresų, datų ir kitos informacijos, kuri galėtų kaip nors pakenkti kompiuterių tinklų valdytojams, tikrai daug daugiau reikiamos informacijos moksliniams tyrimams atsirastų viešuose terpėse ir būtų laisvai pasiekiama.

Problemos sprendimas. Sukurtas modelis, leidžiantis pagal kelis atributus anonimizuoti SNMP žurnalus, pasirenkant naudingų duomenų išliekamumo žurnale lygmenį. Taip pat jis leidžia įvertinti SNMP žurnalų anonimizuotų atributų įtaką bendram anonimizavimo lygmeniui ir naudingos informacijos išliekamumui.

Esminiai reikalavimai modeliui

- Anonimizavimo lygmuo turi apsaugoti organizacijų teikiamą informaciją iki tokio lygio, kad nebūtų galimybės jai pačiai pakenkti piktavališkais tikslais.
- Anonimizavimo lygmuo turi būti optimizuotas iki tokio lygio, kad nepažeidžiant pirmo reikalavimo, anonimizuota informacija vis dar turėtų tiriamąją vertę moksliniu ir analitiniu aspektu.
- Modelis turi turėti įrankį, leidžiantį įvertinti anonimizuotų duomenų anonimiškumo lygį.

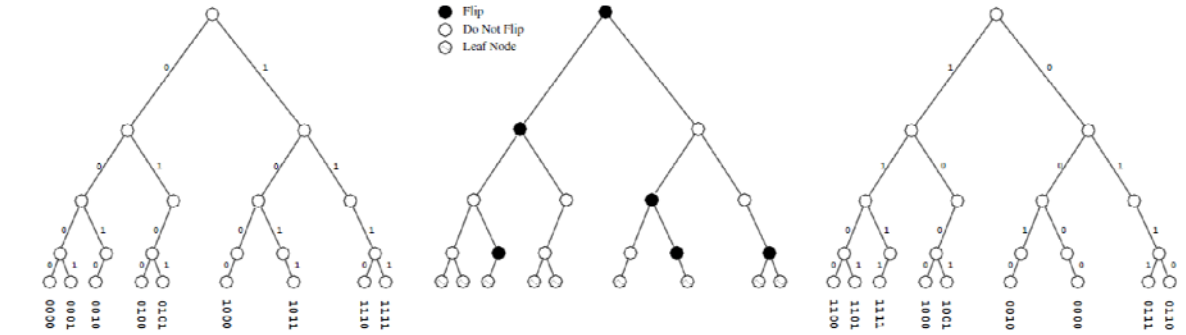
II. SNMP APŽVALGA IR ANALIZĖ

Paprastas tinklo valdymo protokolas (angl. Simple Network Management Protocol - SNMP) veikia pagal serverio/agento modelį [11]. Agentai siunčia signalus apie įvykius (angl. Trap) serveriui, kai atsitinka tam tikra situacija, pvz., kintamojo reikšmė viršija nustatytąją ar pasiekia kritinę vertę.

Kadangi pastarieji pranešimai (Trap) turi didžiausią kiekį su tinklų sauga susijusios jautrios informacijos, mes ir pasirinkome analizuoti būtent šiuos prancšimus, ncs iš jų galima susidaryti organizacijos tinklo topologiją, anomalijų ir gedimų bendrą vaizdą, galimus atakos taškus bandant įsiskverbti į organizacijos tinklą. Būtent dėl šių priežasčių Trap pranešimai yra griežtai saugomi organizacijų vidaus saugumo politikų, nes pasinaudodamas šiais duomenimis piktavalis gali bandyti pakenkti organizacijai per kompiuterinį tinklą pasinaudodamas šia jautria informacija.

III. ANONIMIZAVIMO METODŲ APŽVALGA

Kadangi mūsų darbas remiasi anonimizavimo principais, tai iš pradžių būtina apžvelgti ir išanalizuoti duomenų anonimizavimo specifiką. Privatumas yra atskleistas, kai specifinis kenkėjas sugeba teisingai sujungti asmenis su jų jautriomis vertėmis. Metrinis anonimiškumas nustato lygį, kaip gerai anonimizavimo technika slepia atstovo tapatybes ar santykius prieš specifinį kenkėją. Norint, kad anonimizuoti SNMP žurnalai būtų naudingi ir tolimesniems tyrimams bei nebūtų atskleista jautri interneto tiekėjų informacija, būtina anonimizuoti ne tik siuntėjo ir gavėjo IP adresus, bet kitus jautrius žurnalų laukus, tokius kaip siuntimo data ir laikas, įrenginio ar objekto identifikatorius bei jų būsenas ar įvykius. Toliau mes juos vadinsime kvazi-identifikatoriais (netiesiogiai objektą identifikuojančių atributų aibė, toliau – QI). Apžvelkime keletą anonimizavimo algoritmų, kurie panaudoti panašiems informacijos anonimizavimo tikslams.



(a) Originalus adresų medis (b) Anonimizavimo funkciją f_1 (c) Anonimizuotas adresų medis

1 pav. Prefikso išsaugojimo anonimizavimo funkcijos geometrinė interpretacija [14]

A. RYŠIŲ NUTRAUKIMO ALGORITMAS

Weijia Yang ir Sanzheng Qiao [8] pasiūlė anonimizavimo algoritmą, kurio tikslas yra nutraukti ryšius tarp k-identifikatorių ir su jais susijusia jautria informacija. Tai nereikalauja anonimizuoti visų QI reikšmių. Pagrindinė idėja yra ta, kad atsitiktiniu būdu (angl. random) pakeisti dalį QI reikšmių kiekvienam įrašui, vartojant originalių verčių požymių pasiskirstymą. Tokiu būdu jokia nauja informacija nėra pridėta prie anonimizuotų duomenų, ryšiai tarp QI ir jautrių duomenų yra nutraukti, o originalus duomenų pasiskirstymas yra išsaugotas.

B. IP ADRESO ANONIMIZAVIMAS IŠSAUGANT PREFIKSĄ IR LEKSIKOGRAFINĮ ĮILIŠKUMĄ

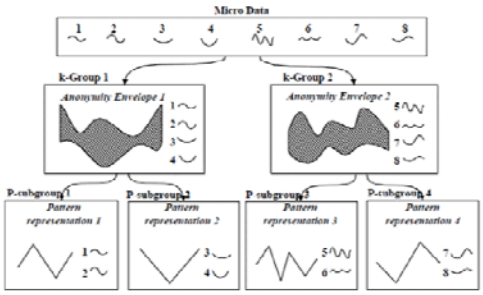
Jun Xu, Jinliang Fan ir Mostafa H. Ammar [14] pasiūlė naują IP adreso anonimizavimo algoritmą, kuris išsaugo tokį patį IP adreso prefiksą, jei yra anonimizuojami du to paties potinklio adresai. Šio anonimizavimo tipo nauda yra ta, kad tinklų ir potinklių struktūra yra visiškai išsaugota, kai tuo pačiu metu ir nežinoma. Kenkėjas bandydamas skenuoti tokį anonimizuotą IP adresą jokios apčiuopiamos naudos negaus, nes nei realus potinklis, nei tikras IP adreso savininkas nebus žinomas.

Apibrėžimas. Du IP adresai $a = a_1a_2 \ldots a_n$ ir $b = b_1b_2 \ldots b_n$ turi k bitų ilgio prefiksą ($0 \le k \le n$), jei $a_1a_2 \ldots a_k = b_1b_2 \ldots b_k$ ir $a_{k+1} \neq b_{k+1}$, tada $k < n$. Anonimizavimo funkcija F yra apibrėžta kaip vienas vienam funkcija nuo $\{0, 1\}^a$ į $\{0, 1\}^a$. Anonimizavimo funkcija F yra prieš dėlio išsaugojimas: jei duotiems dviems IP adresams a ir b, kurie turi k – bitų ilgio prieš dėlį, tada F (a) ir F (b) turi k – bitų ilgio prieš dėlį taip pat.

Įsivaizduokime anonimizavimo metodo išsaugant prefiksą geometrinę interpretaciją. Atkreipkime dėmesį, kad visai IP adreso erdvei gali atstovauti užbaigtas dvejetainis medis. Kadangi IPv4 adresai, šis medis turėtų 32 mazgus tuo metu, kai IPv6 turės 128 mazgus. Kiekvienam IP adresui atstovauja mazgas. Be to, kiekvienas mazgas atitinka bito poziciją (išreikštas pagal mazgo aukštį) ir vertes (išreikštas pagal jos tėvinio mazgo šakos kryptį).

C. (K,P)- ANONIMIŠKUMO MODELIS: LAIKO DUOMENŲ STRUKTŪRĄ SAUGANTIS ANONIMIZAVIMAS

Xuan Shang [10] pasiūlė algoritmą, skirtą laiko duomenims anonimizuoti. Jo metode yra numatyta, kad kiekviena laiko eilutė yra publikuota trijuose komponentuose: apibendrinti QI atributai, QI struktūros pavaizdavimas ir jautri informacija. Patikslinant (k, P) - anonimiškumo modelis gali būti apibūdintas kaip konceptualus tradicinio k-anonimiškumo išplėtimas. Vis dėlto (k, P) - anonimiškumas nepriklauso tradiciniam k-anonimiškumo algoritmui.



2 pav. (k,P)- anonimiškumo modelis [10]

Kaip iliustruoja 2 pav., jų modelis garantuoja anonimiškumą dviejuose lygmenyse. Pirmame lygmenyje QI atributai yra apibendrinti, kad įvykdytų tradicinį k-anonimiškumą, nepriklausomai nuo QI struktūros pavaizdavimo. Apibendrinimo rezultatai turi daugiau padalijimų, žinomų kaip k-grupės. Pažymėtina, kad apibendrinti QI atributai panašėja į tuos tradiciniame k-anonimiškume. Antro lygmens įrašų anonimiškumas laikomas kiekvienoje k-grupėje. Bet kokiam įrašui r k-grupėje, jei ten egzistuoja bent jau P – 1 kiti įrašai, kurie turi tą patį struktūros pavaizdavimą kaip r, teigiama, kad P-anonimiškumas yra priverstinai sukuriamas visai šiai k-grupei. Todėl galima padalinti k-grupę toliau į pogrupius, iš kurių kiekvienas talpina mažiausiai P įrašų, turinčių identišką struktūros pavaizdavimą.

Apibendrinant apžvelgtus anonimizavimo algoritmus ir metodus galima juos pritaikyti skirtingiems SNMP žurnalų laukams; priklausomai nuo kintamųjų reikšmių formato, taikyti vieną ar kitą anonimizavimo būdą:

- **Registracijos datos** (angl. time stamp) laukų anonimizavimui tikslinga taikyti 3.3 punkte apžvelgtą metodą „Laiko duomenų struktūrą saugantis anonimizavimas“, nes juo galima anonimizuoti laiko duomenis intervalais, kuriuos taip pat galima dar papildomai skaidyti į mažesnius pogrupius, jeigo reikia padidinti anonimiškumą.
- **Objekto IP adresui** (angl. agent address) tikslinga būtų panaudoti 3.2 punkte apžvelgtą IP adreso anonimizavimą, išsaugant prefiksą ir leksikografinį jo eiliškumą.
- **Objekto identifikatoriams** (toliau – OID) **ir jų reikšmėms** tikslinga būtų panaudoti 3.1 punkte apžvelgtą algoritmą, kuris leidžia nutraukti ryšius tarp realių identifikatorių ir jautrios informacijos.

IV. ANONIMIZAVIMO LYGIO ANALIZĖ

Prieš pateikiant trečioms šalims anonimizuotus SNMP žurnalus, juos būtina ištirti, kiek naudingos arba vis dar atitinkančios realybę informacijos yra išlikę šiuose anonimizuotose duomenyse bei kokia tikimybė yra susieti realius duomenis su jų būsenomis ar kitais realiais atributais (pvz., kada įvyko įvykis, iš kurio tinklo įrenginio su kuriuo jo objektu?). Analizei buvo pasitelktas anonimiškumo matavimo parametras – entropija. Norint visapusiškai įvertinti ar anonimizuoti, duomenys neturi galimybės atskleisti realios informacijos ten, kur nenorima; reikia įvertinti visus informacijos laukus, patikrinti jų individualias, sandūros (angl. joint) ir sąlygines (angl. conditional) entropijas, o taip pat duomenų stulpelių priklausomybes vienam nuo kito (pvz., žinant ar pašalinant vieną vieno stulpelio reikšmę, kokia įtaka daroma kitam stulpeliui individualiai ir sandūros entropijai).

Naudojamos formulės vertėms gauti:

$$H(X) = -\sum_{i=n} p(x_i) * \log\left(p(x_i)\right) \quad (1)$$

$$H(Y) = -\sum_{i=n} p(y_i) * \log\left(p(y_i)\right) \quad (2)$$

$$H(X,Y) = -\sum_{i=n} \sum_{j=n} p(x_i, y_j) * \log\left(p(x_i, y_j)\right), \quad (3)$$

kur

$p(x_i)$ yra kiekvieno skirtingo X aibės argumento pasikartojimo tikimybė x_i,

$p(y_i)$ yra kiekvieno skirtingo Y aibės argumento pasikartojimo tikimybė y_i,

$p(x_j, y_i)$ yra X ir Y skirtingų pasikartojimų tikimybė x_j, y_i,

n yra suminis įvykių kiekis.

Entropijos apskaičiuojamos kiekvienam anonimizuotam stulpeliui (žiūrėti 2 lentelę), o ten, kur egzistuoja tamprus ryšys tarp dviejų stulpelių (kaip šiuo atveju IP-OID), vertinama ir sandūros (angl. joint) entropija [15].

Remiantis naujuoju anonimiškumo matavimo vadovaujantis daline entropija [16], išvadose pateikiama saugumo lygio skalė nurodanti rizikos grėsmę, priklausomai nuo esamo anonimiškumo lygio, kur yra vertinama entropijos laipsnis (H), skirtingų QI kiekis ir pasikartojimo dažnio pasiskirstymas visoje duomenų aibėje:

- NESAUGU. Labai mažas neapibrėžtumas (H< 1.5) leidžia lengvai atstatyti ryšius tarp tinklo įrenginių ir jo objektų, nes pasikartojimo dažnis pasiskirstęs labai nevienodai ir jų yra labai mažai.
- NEPATIKIMA. Neapibrėžtumas minimalus (H=1.5-2.5) leidžia lengvai atstatyti ryšius tarp tinklo įrenginių ir jo objektų, nes pasikartojimo dažnis pasiskirstęs daugiau nei keliuose objektuose vienodai, bet jų yra mažai.
- SAUGU. Neapibrėžtumas pakankamas (H=2.5-5), kuris neleidžia lengvai atstatyti ryšius tarp tinklo įrenginių ir jo objektų, nes pasikartojimo dažnis pasiskirstęs daugiau nei keliuose objektuose vienodai, o dominuojančių pasikartojimų daugiau nei 2.
- LABAI SAUGU. Neapibrėžtumas labai didelis (H>5) neleidžia atstatyti ryšius tarp tinklo įrenginių ir jo objektų, nes pasikartojimo dažnis pasiskirstęs labai tolygiai ir jų yra labai daug (artima ar lygu n).

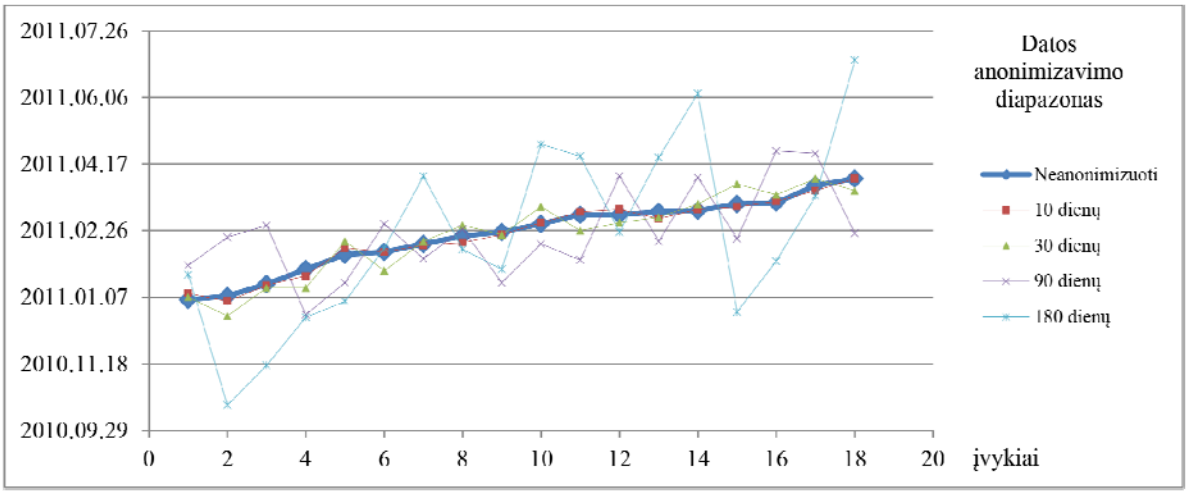
V. EKSPERIMENTAS IR JO REZULTATAI

Eksperimento tikslas – įvertinti modelio anonimizavimo metodus ir išanalizuoti anonimiškumo lygį.

Mūsų sukurtas programinės įrangos prototipas (toliau – prototipas), leidžiantis sparčiai atlikti visus pristatomo modelio veiksmus nuo duomenų nuskaitymo iki išanalizuotų duomenų perdavimo trečioms šalims, įskaitant duomenų atrinkimą, anonimizavimą ir jų analizę, mums leidžia efektyviai įvertinti modelio pagrįstumą. Mūsų prototipas turi funkcijas, leidžiančias įtakoti anonimizavimo procesą, priklausomai kiek ir kokios realios informacijos norima palikti anonimizuotose duomenyse.

- Anonimizuojant datą ir laiką (naudojamas 3.3 punkte aprašymas algoritmas) galima nustatyti, kokiame diapazone generuoti datos nukrypimą nuo realios (pvz, paslėpti tikslias įvykusio įvykio datas ir/ar laikus, arba keičiant minimaliai, kad būtų galima įvertinti įvykių tendencijas laiko atžvilgiu).
- Anonimizuojant IP adresą (naudojamas 3.2 punkte aprašymas algoritmas) galima nustatyti prefikso ilgį (jei reikia, išsaugoma organizacijos tinklo struktūra).

Anonimizuojant OID ir jų reikšmes (naudojamas 3.1 punkte aprašymas algoritmas) yra galimybė sumaišyti esamas reikšmes, nutraukiant realius gedimų ar įvykių ryšius su realiais, juos generuojančiais, objektais. Be to, esant poreikiui, galima įterpti papildomų neegzistuojančių objektų ar atsitiktinai sugeneruotų ne tikrų įvykių (naudinga, kai norima paslėpti tikruosius įvykius).



3 pav. Priklausomai, koks pasirinktas datos anonimizavimo diapazonas, gautų reikšmių lyginimas su realiais duomenimis

Testiniams žurnalų duomenims mes naudosime 1000 realių SNMP Trap pranešimų, kuriame yra du potinkliai po 50 tinklinių įrenginių, turinčių savo unikalius IPv4 adresus. Ištyrėme, kaip tankiai „lūžinėja" organizacijos vietinis tinklas ir kurie įrenginiai ir/ar moduliai tai įtakoja. Būtina pateikti visus įrašus, kur yra pranešta apie SNMP standartinį aliarmą (angl. generic trap) su jo būsena LinkDown (GT=3). Išfiltravę informaciją iš visų duomenų, gavome 86 įrašus, kurie atitiko mūsų užsibrėžtą tikslą.

Lentelė I. Pirminiai SNMP pranešimo duomenys

Data laikas	IP adresas	OID	Įvykis
2011-01-12 20:41:11	172.30.10.34	a3Com	linkDown
2011-01-12 04:08:18	172.30.20.39	a3Com	linkDown
2011-01-11 07:15:37	172.30.10.33	microsoft.1.1.3.1.1	linkDown
2011-01-09 06:37:11	172.30.10.12	a3Com	linkDown
2011-01-01 03:02:10	172.30.20.18	a3Com	linkDown
2011-01-04 07:46:54	172.30.10.19	snmpTraps	linkDown

Duomenų anonimizavimui buvo pasirinkta anonimizuoti datą/laiką, IP adresą ir OID reikšmes, o įvykis paliktas nekeistas. Dabar tas pats duomenų masyvas atrodo kitaip (rodomas tik fragmentas):

Lentelė II. Anonimizuoti SNMP pranešimo duomenys

Data laikas	IP adresas	OID	Įvykis
2011-01-01 03:04:10	44.83.61.18	snmpTraps	linkDown
2011-01-05 07:48:54	44.83.10.204	a3Com	linkDown
2011-01-10 06:39:11	44.83.10.222	microsoft.1.1.3.1.1	linkDown
2011-01-11 07:13:37	44.83.10.139	a3Com	linkDown
2011-01-12 20:39:11	44.83.10.196	snmpTraps	linkDown
2011-01-12 04:06:18	44.83.61.238	microsoft.1.1.3.1.1	linkDown

Apskaičiuojant entropijas būtina patikrinti, ar nėra galimybės susieti realių tinklo įrenginių (IP adresų) su jų objektais (OID), todėl buvo skaičiuojamos entropijos H(IP), H(OID) ir sandūros entropija H(IP, OID). Naudojamų algoritmų anonimizavimo sparta nebuvo vertinama, nes sparta labai priklauso nuo programinio išpildymo ir techninės įrangos galimybių. Norime pabrėžti, kad mūsų metodas skirtas saugumo lygiui įvertinimui, o ne anoninimizavimo spartai tirti, nors atliekant tyrimus su dideliais (daugiau nei 100 000 SNMP žurnalo įrašų) duomenų kiekiais, sparta tenkino poreikius. Sustačius reikšmes į formules, gauti rezultatai.

X aibės argumento pasikartojimo dažnis x_i apskaičiuojamas suskaičiuojant, kiek tokių pačių elementų yra aibėje ir surandama jo dalis visoje aibėje X. Pvz.: IP adresas 23.239.43.192 pasikartojo 7 kartus iš 86 įrašų, taigi jo pasikartojimo dažnis aibėje X yra 7/86 = 0.0814. Tokiu būdu surandami visų egzistuojančių skirtingų IP adresų pasikartojimo dažniai ir sustatomi į 1 formulę:

H(IP) = -1 * [(0.0814* log(0.0814)) + (0.1279* log(0.1279)) + (0.1395* log(0.1395)) + (0.0814 * log(0.0814)) + (0.1163* log(0.1163)) + (0.093* log(0.093)) + (0.1744* log(0.1744)) + (0.186* log(0186))]
H(IP) = 2.93558

Tokiu pačiu metodu apskaičiuojami visų OID pasikartojimo dažniai (naudojama 2 formulė):

H(OID) = -1 * [(0.3256 * log(0.3256)) + (0.3256 * log(0.3256)) + (0.3488 * log(0.3488]
H(OID) = 1.58419

Sandūros entropijos (H(IP,OID)) įvykių porų pasikartojimų dažnių x_j, y_i apskaičiavimas atliekamas suskaičiuojant, kiek tokių pačių elementų porų yra aibėje ir surandama jų dalis sandūros aibėje X,Y.

Pvz.: IP adresas 23.239.43.192 su OID microsoft.1.1.3.1.1 pasikartojo 3 kartus iš 86 įrašų, taigi jo pasikartojimo dažnis sandūros aibėse X,Y yra 0.0349. Tokiu būdu surandamos visos porų kombinacijos, apskaičiuojami jų pasikartojimų dažniai.

$p(x_i,y_i)$ =[0.0814, 0.0698, 0.0698, 0.0698, 0.0698, 0.0581, 0.0581, 0.0581, 0.0465, 0.0465, 0.0349, 0.0349, 0.0349, 0.0349, 0.0349, 0.0349, 0.0233, 0.0233, 0.0233, 0.0233, 0.0233, 0.0233, 0.0116, 0.0116]

Gautos reikšmės $p(x_i,y_i)$ sustatomos į 3 formulę ir gaunamas rezultatas yra H(IP, OID) = 4.41407.

VII. IŠVADOS IR TOLIMESNI DARBAI

Atlikus mūsų pasirinktų anonimizuotų duomenų analizę, galime lengvai nustatyti, kad pagal mūsų sudarytą vertinimo skalę IP adresų neapibrėžtumo laipsnis atitinka kriterijui SAUGU, o OID skirtingų objektų kiekis tik 3, tad ir neapibrėžtumo laipsnis tik NEPATIKIMA, bet sandūros entropija atitinka kriterijų SAUGU, nes ryšiai tarp tikrų tinklų įrenginių ir jų objektų pasiskirstę pakankamai tolygiai ir jų dominuojančių daugiau nei 3 (0.0814, 0.0698, 0.0698, 0.0698, 0.0698).

Išliekamoji vertė anonimizuotame duomenų masyve iš dalies išsaugota, nes IP adresų potinklių struktūra išsaugota su jų leksikografiniu eiliškumu, OID visi realūs, tik nutraukti ryšiai su realiais tinklo įrenginiais.

Trap pranešimų data ir laikas gali būti anonimizuojamas minimaliai (pvz., 2 dienų tikslumu). Tokiu būdu reali įvykio data nežinoma, bet tendencijas ir/ar įvykių pasikartojimus galima identifikuoti. Esant poreikiui, naudojantis prototipo galimybėmis, galima anonimizuoti net 180 dienų tikslumu (neapibrėžtumo laipsnis daugiau nei 6, t.y. LABAI SAUGU), bet tada visiškai prarandamas ryšys su įvykio realia data (3 pav.).

Atlikta modelio realizacija ir jos patikrinimas pirminiais duomenimis užsibrėžtus tikslus pateisino. Kitame modelio vystymo etape paaiškės, ar visais atvejais yra patikima taikyti šį modelį, ar gali reikėti atlikti pataisymus, ar įtraukti papildomų išimčių. Jau dabar galima pastebėti, kad būtina atlikti visapusišką anonimizuotos informacijos analizę. Būtina įvertinti ir kitus, kai kuriais atvejais nereikšmingus, parametrus, tokius kaip sąlyginė entropija ar abipusė informacija, kuri atrodo nereikšminga, bet pasikeitus situacijai ar atsiradus galimybei atskleisti vieną iš anonimizuotų duomenų aibės, galima padaryti nepataisomą žalą organizacijai ar jos įvaizdžiui su visomis blogomis pasekmėmis.

LITERATŪROS SĄRAŠAS

[1] J. Schonwalder, A. Pras, M. Harvan, J. Schippers, and R. van de Meent, "SNMP Traffic Analysis: Approaches, Tools, and First Results," Proc. 10th IFIP/IEEE International Symposium on Integrated Network Management, May 2007.

[2] M. Harvan ir J. Schonwalder, „Prefix- and Lexicographical-order-preserving IP Address Anonymization," 10th IEEE/IFIP Network Operations and Management Symposium p.519–526, 2006.

[3] Latanya Sweeney, "K-anonymity: a Model for Protecting Privacy," 2002.

[4] M. F. Chi-Yin Chow and X. Liu, "A Peer-to-Peer Spatial Cloaking Algorithm for Anonymous Location-Based Services.", Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems. Arlington, Virginia, USA, 2006.

[5] Gedik B and Ling Liu, "Protecting Location Privacy with Personalized k-Anonymity: Architecture and Algorithms", IEEE Transactions on Mobile Computing. University of Illinois at Urbana-Champaign. P.1-18, 2008.

[6] Raymond, J.-F., "Traffic analysis: Protocols, attacks, design issues, and open problems. In Proceedings of Designing Privacy Enhancing Technologies: Workshop on Design Issues in Anonymity and Unobservability," 2000.

[7] Serjantov, A. and Newman, "On the anonymity of timed pool mixes. In Proceedings of the Workshop on Privacy and Anonymity Issues in Networked and Distributed Systems. Kluwer, Athens, Greece," p. 427–434, 2003.

[8] Weijia Yang ir Sanzheng Qiao, "A novel anonymization algorithm: Privacy protection and knowledge preservation," 2009.

[9] Aggarwal, C. C., & Yu, P. S., "A general survey of privacy-preserving data mining models and algorithms. Privacy-preserving data mining," p. 11–52, 2008.

[10] Xuan Shang, Ke Chen, Lidan Shou, Gang Chen, Tianlei Hu, "(k,P)-Anonymity: Towards Pattern-Preserving Anonymity of Time-Series Data," CIKM'10, October 26–30, Toronto, Ontario, Canada, 2010.

[11] RFC 3411 – 3418 SNMP protokolo standartų aprašymas [prieiga per internetą] http://www.rfc-editor.org/rfc/std/std62.txt, žiūrėta 2012-01-20

[12] M. Edman, B. Yener, "On Anonymity in an Electronic Society: A Survey of Anonymous Communication Systems," ACM Computing Surveys, Vol. 42, No. 1, Article 5, 2009.

[13] M. Bezzi, "An Entropy based method for Measuring Anonymity," Proceedings of the third international conference on security and privacy in communication networks. Nice, France. p. 28 – 32, 2007.

[14] J. Xu, J. Fan, and M. H. Ammar, "Prefix-Preserving IP Address Anonymization: Measurement-based Security Evaluation and a New Cryptography-based Scheme," Proceedings of the 10th IEEE International Conference on Network Protocols (ICNP'02) , 2002.

[15] Paul Meagher, "Calculating Entropy for Data Miners," [prieiga per internetą] http://onlamp.com/pub/a/php/2005/03/24/joint_entropy.html?page=1, žiūrėta 2013-01-15

[16] Guihua Duan, Weiping Wang+, Jianxin Wang, Luming Yang, "A new anonymity measure based on partial entropy," IEEE Communications Society publication in the ICC, 2008.