MOKSLAS – LIETUVOS ATEITIS
SCIENCE – FUTURE OF LITHUANIA
ISSN 2029-2341 / eISSN 2029-2252
http://www.mla.vgtu.lt

ELEKTRONIKA IR ELEKTROTECHNIKA
ELECTRONICS AND ELECTRICAL ENGINEERING

2016 8(3): 327–333

http://dx.doi.org/10.3846/mla.2016.928

IŠSKIRČIŲ RADIMO METODŲ TAIKYMAS ANOMALIJOMS KOMPIUTERIŲ TINKLO PAKETŲ SRAUTUOSE APTIKTI

Rimas CIPLINSKAS[1], Nerijus PAULAUSKAS[2]

Vilniaus Gedimino technikos universitetas, Vilnius, Lietuva
El. paštas: [1]rimas.ciplinskas@gmail.com; [2]nerijus.paulauskas@vgtu.lt

Santrauka. Kibernetinių atakų gausa ir įvairovė bei siekis nuo jų apsisaugoti verčia nuolat kurti naujus ir tobulinti jau esamus atakų aptikimo metodus. Kaip rodo praktika, dabartiniai atakų atpažinimo metodai iš esmės veikia pagal antivirusinių programų principą, t.y. sudaromi žinomų atakų šablonai, kuriais remiantis yra aptinkamos atakos, tačiau pagrindinis tokių metodų trūkumas – negalėjimas aptikti naujų, dar nežinomų atakų. Šiai problemai spręsti yra pasitelkiami anomalijų aptikimo metodai, kurie leidžia aptikti nukrypimus nuo normalios tinklo būsenos. Straipsnyje yra pateiktas naujas metodas, leidžiantis aptikti kompiuterių tinklo paketų srauto anomalijas taikant lokalių išskirčių faktorių algoritmą. Atliktas tyrimas leido surasti požymių grupes, kurias taikant anomalūs tinklo srautai yra atpažįstami geriausiai, t. y. pasiekiamos didžiausios tikslumo, atkuriamumo ir F-mato reikšmės.

Reikšminiai žodžiai: anomalijos, anomalijų aptikimo metodai, LOF, tinklo paketų srautas, atakos.

Įvadas

Nepaliaujamai didėjant informacinių grėsmių kibernetinėje erdvėje skaičiui bei kylant naujoms ir modernioms agresijos prieš valstybę formoms, labai svarbu sumažinti grėsmių informacinėms sistemoms keliamą riziką ir galimas pasekmes. Šiam tikslui būtina identifikuoti atakas pradiniame jų realizavimo etape, laiku ir tinkamai į jas reaguoti. Kaip rodo praktika, dabartiniai atakų atpažinimo metodai iš esmės veikia pagal antivirusinių programų principą, t. y. sudaromi žinomų atakų šablonai, kuriais remiantis yra aptinkamos atakos, tačiau pagrindinis tokių metodų trūkumas – negalėjimas aptikti naujų dar nežinomų atakų. Šiai problemai spręsti yra pasitelkiami anomalijų aptikimo metodai, kurie leidžia aptikti nukrypimus nuo normalios tinklo būsenos, pvz., naują ataką, piktnaudžiavimą kompiuterių tinklo resursais ar netinkamai sukonfigūruotą tinklo programinę įrangą.

Išskirčių aptikimas yra naudojamas įvairiose srityse, siekiant identifikuoti neįprastus duomenų šablonus ar anomalijas duomenyse, pvz., aptinkant finansinius sukčiavimus, įsilaužimus kompiuterių sistemose, analizuojant žmonių elgesį ir pan. Išskirtis – tai neįprastas įvykis ar objektas, kurio savybės neatitinka didžiosios dalies duomenų rinkinio savybių. Išskirtys yra laikomos svarbiomis, kadangi gali suteikti ypač reikšmingos informacijos priimant svarbius sprendimus. Įsilaužimai į kompiuterius paprastai yra susiję su kenkėjiško programinio kodo platinimu, konfidencialios informacijos vagystėmis ar užgrobtos sistemos resursų naudojimu vykdant paskirstytas atsisakymo aptarnauti atakas. Būtina pabrėžti, kad atakos sudaro tik labai mažą dalį nukrypimų, lyginant su normaliu viso tinklo ir kompiuterių naudojimu. Ši nedidelė piktavališkos veiklos dalis skiriasi nuo įprastos vartotojo veiklos, todėl gali būti aptikta naudojant išskirčių aptikimo metodus (Mohiuddin *et al*. 2014).

Išskirtys gali būti aptiktos, naudojant statistinius testus, besiremiančius duomenų pasiskirstymo ar tikimybiniais modeliais. Taip pat išskirtys gali būti aptiktos, naudojant įvairius atstumu ar tankiu paremtus metodus, kur sąlyginai nutolę nuo kitų klasterių objektai laikomi išskirtimis (Maimon, Rokach 2010).

Atliekant statistinius testus, dažniausiai išskirtys aptinkamos naudojant tam tikrą statistinį pasiskirstymą, pavyzdžiui, Gauso pasiskirstymą. Turimų duomenų parametrai apskaičiuojami laikantis prielaidos, jog visi duomenys buvo sugeneruoti pagal pasirinktą statistinį pasiskirstymą. Išskirtys yra tie taškai, kurių pasirodymo tikimybė yra maža. Pavyzdžiui, išskirčių nuokrypis skiriasi daugiau kaip tris kartus palyginus su standartiniu nuokrypiu nuo vidurkio (Chandola *et al*. 2009).

Vienas iš atstumu pagrįstų išskirčių aptikimo metodų yra *k*-artimiausių kaimynų (angl. *k-nearest neighbour* – toliau „kNN“) metodas. Duomenų objektai yra

klasifikuojami pagal tai, kokią klasę turi artimiausi gretimi duomenų objektai. Dažnai įtraukiamas ne vienas gretimas objektas, o keli gretimi objektai, todėl algoritmas ir vadinasi *k*-artimiausių kaimynų. Algoritmo veikimo metu yra reikalingi mokomieji duomenys, todėl pats algoritmas dar kartais vadinamas atmintimi paremta klasifikacija (Cunningham, Delany 2007).

Kitas atstumu pagrįstas išskirčių aptikimo metodas yra daugiamatės skalės (angl. *multidimensional scaling*, toliau „MDS“). Jos naudojamos daugiamačių duomenų struktūros analizei dvimatėje arba trimatėje erdvėje. Tai vienmačių skalių, kai objektai išdėstomi atkarpoje, apibendrinimas daugiamatėje erdvėje. Dažniausiai naudojamos dvimatės skalės. Jei duomenys pateikiami daugiamačiais vektoriais, jų skirtingumas įvertinamas pagal atstumą daugiamatėje duomenų erdvėje. Pagal daugiamačių skalių metodą sprendžiamas uždavinys, kaip *m* objektų, apibrėžtų artimumo duomenimis, gali būti patikimai pavaizduoti taškais mažo skaičiaus matmenų vaizdo erdvėje (Dzemyda *et al*. 2008).

Vienas iš tankiu pagrįstų metodų yra vietinių išskirčių faktorius (angl. *Local Outlier Factor,* toliau „LOF“). Šis metodas gali būti panaudotas tyrimui, kai duomenys išsidėstę grupėmis ir tos grupės gali būti įvairaus tankio. Vietinių išskirčių faktorius suteikia galimybę gauti įvertį, pagal kurį galima nuspręsti, ar duomenų masyvas yra išskirtis, ar ne. LOF metodo pagrindinė idėja yra parinkto taško tankio palyginimas su jo artimiausiais taškais. Sąlyginis palygintų taškų tankis yra galutinis įvertis (Breuning *et al*. 2000).

Straipsnyje pasiūlytas būdas, leidžiantis aptikti kompiuterių tinklo paketų srauto anomalijas taikant lokalių išskirčių faktorių algoritmą. Straipsnyje taip pat pateikiami rezultatai, parodantys, kaip anomalijų aptikimas priklauso nuo pradinių požymių parinkimo.

Duomenų surinkimas

Informacija apie kompiuterių tinklo veiklą gali būti renkama analizuojant tinklu perduodamus paketus arba tinklo paketų srautą (angl. *flow*). Analizuojant paketus, gaunamas didžiausias tikslumas, kadangi paketais yra perduodama ne tik ryšio seanso tarp galinių tinklo mazgų informacija, bet ir patys duomenys. Tačiau šis privalumas kartu tampa ir trūkumu, kai tenka saugoti ir apdoroti didžiulius duomenų kiekius.

Kompiuterių tinklo paketų srautų duomenys, lyginant juos su atskirų tinklo paketų duomenimis, turi keletą privalumų. Pirmiausia, tinklo srautų įrašams saugoti pakanka kur kas mažiau vietos, tai yra svarbu didelio apkrovimo tinkluose; antra, šie duomenys yra plačiai prieinami, kadangi daugelis šiuolaikinių maršrutų parinktuvų turi galimybę rinkti ir eksportuoti duomenis apie tinklo srautus.

Dažniausiai duomenims apie tinklo srautus registruoti yra naudojamas Cisco pasiūlytas *NetFlow* protokolas (Cisco Systems 2012). Naudojant *NetFlow* duomenis anomalijoms tinkle identifikuoti, susiduriama su įvairiais iššūkiais. *NetFlow* duomenyse nėra informacijos apie paketais perduodamus duomenis. Srauto duomenys apsiriboja duomenimis apie srauto trukmę ir persiųstų duomenų kiekį. Be to, *NetFlow* duomenys yra vienkrypčiai, t.y. vienas srauto įrašas saugo informaciją apie duomenų perdavimą viena kryptimi. Didelio apkrovimo tinkluose *NetFlow* duomenys apie stebimą tinklą neretai yra renkami diskretizuotai, t.y. išsaugomi ne visi srauto įrašai. Šios *NetFlow* savybės trukdo sukurti efektyvią anomalijų aptikimo sistemą. Nepaisant minėtų trūkumų, daugelis jau atliktų tiriamųjų darbų rodo, kad tinklo srauto duomenys yra sėkmingai naudojami identifikuojant tam tikrą piktavališką veiklą (Bilge *et al*. 2012; Lazarevic *et al*. 2003).

Požymių aprašymas

Kompiuterių tinklo paketų srautas – tai vienkryptė seka tinklo paketų tarp dviejų tarpusavyje komunikuojančių pusių, t. y. siuntėjo ir gavėjo. Ryšio seansas tarp siuntėjo ir gavėjo yra padalinamas į du srautus: įeinantį ir išeinantį, t.y. kiekvienam ryšio seansui sukuriami du tinklo paketų srauto įrašai. Pavyzdžiui, jeigu iš vieno IP adreso buvo kreiptąsi į tris skirtingus to paties gavėjo prievadus ir į šiuos kreipinius buvo gauti atsakymai, iš viso bus sukurti šeši srauto įrašai. Viename įraše apie tinklo paketų srautą yra pateikiami duomenys apie siuntėjo ir gavėjo IP adresus, siuntėjo ir gavėjo prievadus, srauto trukmę, transporto protokolą, persiųstų duomenų kiekį, tačiau patys persiųsti duomenys nėra saugomi.

Pradiniai požymiai suteikia daug informacijos apie tinklo srautą, tačiau informacija yra daugiau bendro pobūdžio ir ją sudėtinga sukonkretinti. Iš pradinių požymių galima sudaryti išvestinius požymius. Tai atliekama skaičiuojant vidurkį, medianą ir standartinį nuokrypį. Skaičiuojant požymius panaudota programavimo kalba R. R yra skirta statistinei duomenų analizei. Išvestiniai požymiai suskirstomi į grupes pagal jų bendras savybes. Požymių grupių aprašymas pateiktas 1 lentelėje.

Vieną grupę sudaro mažiausiai du požymiai, kadangi reikia įvertinti duomenis abiem srauto kryptimis. Pavyzdžiui, 1-ąją grupę sudaro du požymiai apie bendrą srautų skaičių iš siuntėjo gavėjui ir atvirkščiai. Šių srautų skaičius gali skirtis, kadangi ne į visus kreipinius atsakymai gali būti gauti.

Jeigu išsiųstų ir gautų srautų skaičius stipriai skiriasi, tai gali rodyti bandymą žvalgyti kompiuterio prievadus. Žvalga paprastai vykdoma, pradiniame etape siekiant surasti pažeidžiamas vietas.

Vietinių išskirčių faktorius

Tyrimams atlikti buvo pasirinktas vietinių išskirčių faktoriaus algoritmas. Algoritmas naudojamas anomalių duomenų taškų paieškai, matuojant vietinį nuokrypį nuo artimiausių taškų. LOF algoritme ne tik matuojami atstumai, bet ir skaičiuojamas tankis. Pats algoritmas pagal skirstymą priklauso prie tankiu grįstų algoritmų. Algoritmo idėja yra tokia, jog išskirties vieta yra mažesnio tankio negu artimiausių taškų. Skaičiuojant LOF vertę, gaunamas santykinis tankis: jis yra artimiausių per *k* vertę taškų pasiekiamumo su paties objekto vietiniu pasiekiamumu palyginimas (Breunig *et al*. 2000).

Galutinė LOF vertė naudojama nustatyti, ar duomenys gali būti laikomi išskirtimi, ar ne. Jeigu duomenų LOF vertė yra artima arba lygi vienetui, laikoma, jog duomenys su gretimais taškais yra tokio pat tankio ir išskirčių nėra. Jeigu duomenų LOF vertė yra didesnė nei vienetas, laikoma, kad ta duomenų sritis yra mažesnio tankio, lyginant su artimiausiais taškais. Vadinasi, toji sritis, kurios LOF įvertis yra gerokai didesnis už vienetą, yra išskirčių sritis.

LOF vertės skaičiavimą sudaro 4 žingsniai:

1. Apskaičiuojami atstumai tarp kiekvieno taško. Atstumui apskaičiuoti naudojamas Manheteno atstumas. Jis ypatingas tuo, jog atstumui surasti naudojami tinklelio kvadratai. Turint taškų *p* ir *o* koordinates, atstumas randamas pagal šią išraišką:

$$d(p,o) = \sum_{i=1}^{n} | p_i - o_i | . \quad (1)$$

2. Surandami *k*-artimiausi taškai. *k* reikšmę reikia parinkti. Pavyzdžiui, jeigu $k = 2$, yra parenkamas toks atstumas tarp taškų, jog būtų apimti 2 artimiausi taškai.

1 lentelė. Išvestinių požymių apibūdinimas
Table 1. Derivative signs description

Požymių grupės Nr.	Išvestinių požymių bendras tipas	Aprašymas
1	Srautų skaičius	Skaičiuojamas IP adresų poroje išsiųstų ir gautų paketų srautų skaičius.
2	Unikalūs prievadų numeriai	Skaičiuojamas agreguotų paketų srautų unikalių siuntėjo ir gavėjo, į kuriuos ir iš kurių buvo kreiptąsi, prievadų skaičius.
3	Paketai	Skaičiuojama paketų suma, vidurkis, mediana, standartinis nuokrypis.
4	Baitai	Skaičiuojama baitų suma, vidurkis, mediana, standartinis nuokrypis.
5	Trukmė	Skaičiuojama paketų srautų trukmių suma, vidurkis, mediana, standartinis nuokrypis.
6	Paketų srautų pasirodymai	Skaičiuojamas laiko tarp paketų srautų pasirodymo vidurkis, mediana, standartinis nuokrypis.
7	Paketų dydis	Skaičiuojama vidutinis paketo dydis, paketo dydžio mediana, standartinis nuokrypis.
8	IP adresų kiekis	Skaičiuojama, kiek kartų siuntėjo IP adresas pasirodė kaip siuntėjo IP adresas; gavėjo IP adresas, kaip gavėjo, taip pat ir priešingos krypties paketų sraute.
9	IP adresų kreipimosi skaičius	Skaičiuojama, į kiek skirtingų gavėjo IP adresų kreipėsi siuntėjo IP adresas; kiek skirtingų siuntėjo IP adresų kreipėsi į gavėjo IP adresą, taip pat ir priešingos krypties paketų sraute.
10	IP adresų kreipimosi skaičiaus vidurkis pagal prievadą	Skaičiuojami skirtingi siuntėjo IP adresai, kurie kreipėsi iš to paties siuntėjo prievado; skirtingi gavėjo IP adresai, kurie kreipėsi į tą patį siuntėjo prievadą, taip pat priešingos krypties paketų sraute.
11	Skirtingas prievadų skaičius pagal IP adresą	Skaičiuojami skirtingi gavėjo prievadai į kuriuos kreipėsi siuntėjo IP adresas; skirtingi siuntėjo prievadai, iš kurių buvo kreiptasi į gavėjo IP adresą, taip pat priešingos krypties paketų sraute.
12	Kreipimosi iš prievadų vidurkis	Skaičiuojamas suagreguotų paketų srautų siuntėjo prievadų skaičiaus vidurkis, gavėjo prievadų skaičiaus vidurkis, taip pat priešingos krypties paketų sraute.
13	Kreipimosi iš prievadų vidurkis neatskiriant prievadų	Skaičiuojamas kreipimosi iš prievado ir į prievadą skaičius, neatskiriant siuntėjo ir gavėjo prievadų, taip pat priešingos krypties paketų sraute.
14	Prievadų panaudojimas	Skaičiuojamas skirtingų gavėjų IP adresų skaičiaus vidurkis, į kuriuos kreipėsi siuntėjas iš to paties prievado, taip pat skirtingas siuntėjų IP adresų skaičius, kurie kreipėsi į gavėjo IP adresą tą patį prievadą, vidurkis, taip pat priešingos krypties paketų sraute.
15	Prievadų panaudojimo vidurkis	Skaičiuojamas siuntėjo IP adreso kreipimosi į tą patį gavėjo prievadą vidurkis, siuntėjo prievado kreipimosi į gavėjo IP adresą vidurkis, taip pat priešingos krypties paketų sraute.

3. Surandama vietinio pasiekiamumo vertė lrd_k. Ji nusako vietinius gretimų taškų pasiekiamumus. Vietinio pasiekiamumo vertė randama pagal šią formulę:

$$lrd_k(p) = 1 / \left(\frac{1}{|N|} \sum_{o \in N} reach\text{-}dist_k(p \leftarrow o) \right). \quad (2)$$

Formulėje 2 dydis $reach\text{-}dist_k$ yra atstumas tarp stebimo taško ir jo artimiausių k taškų. $|N|$ – skaičius taškų, patenkančių į atstumo iki k-tojo artimiausio taško ribas.

4. Surandama LOF vertė. Ji nusako sąlyginį tankį. Kuo vietinio išskirčių faktoriaus vertė didesnė, tuo vietinės srities tankis mažesnis. LOF vertė apskaičiuojama pagal šią formulę:

$$LOF(p) = \frac{\sum_{o \in N} \frac{lrd_k(o)}{lrd_k(p)}}{|N|} = \frac{\sum_{o \in N} lrd_k(o)}{|N|} / lrd_k(p). \quad (3)$$

Remiantis LOF anomalijų aptikimo metodu, taikomas slenkstinis įvertinimas. Slenkstis nustatomas per bandymus. Visi paketų srautai, kurių įvertis gautas didesnis už nustatytą slenkstinę reikšmę, laikomi anomaliais. LOF įverčių paaiškinimai pateikti 2 lentelėje.

2 lentelė. LOF reikšmių paaiškinimai
Table 2. LOF values explanations

LOF reikšmė	Paaiškinimas
~ 1	Normalus paketų srautas
< 1	Paketų srautas yra didelėje sankaupoje su kitais paketų srautais.
> 1	Išskirtis – tokį paketų srautą galima laikyti anomaliu.
Inf	Gretimo taško pasiekiamumo tankis yra begalinis, paketų srautas yra sankaupoje su kitais paketų srautais.
NaN	Pats paketų srautas yra begalinio pasiekiamumo tankyje, vadinasi, paketų srautai dubliuojasi.

Jeigu LOF reikšmė yra artima 1, tai normalu. Jeigu mažesnė už 1, vadinasi, taškas yra apsuptas daugelio greta esančių kitų taškų. Jeigu LOF reikšmė gerokai viršija 1, tai galima laikyti išskirtimi. Galimos situacijos, kai LOF reikšmė yra lygi begalybei (*Inf*) arba skaičiui, kurio nėra (*NaN*), tai yra, įvyko negalima matematinė operacija. Tokias reikšmes galima gauti skaičiuojant pasiekiamumo tankį. Jį skaičiuojant begalinės reikšmės gaunamos, kai artimiausių taškų pasiekiamumas yra 0 – tada visa suma gaunama 0, ir pasiekiamumo tankis yra begalinis. Nulinio pasiekiamumo prasmė yra ta, kad taškai sutampa, tai yra, begalinis tankis reiškia besidubliuojančius taškus. Begalinis tankis reiškia, kad yra visa taškų sankaupa. Pats pasiekiamumas reiškia ne tai, kad stebimas taškas gali pasiekti gretimus taškus, o tai, kad tokiu atstumu artimiausi taškai gali pasiekti stebimą tašką. Jeigu gautas begalinis vietinio pasiekiamumo tankis, skaičiuojant LOF gaunama *NaN* reikšmė, nes pats taškas yra begalinio pasiekiamumo srityje. LOF įvertis *Inf* reiškia, jog stebimas taškas turi baigtinį vietinį pasiekiamumo tankį, tačiau kažkuris gretimas taškas turi begalinį pasiekiamumo tankį, todėl visa pasiekiamumų suma lygi begalybei. Galima teigti, kad LOF reikšmės *NaN* ir *Inf* atsiranda dėl taškų sankaupų, vadinasi, tie taškai atitinka bendrą grupę ir nėra išskirtys.

Metodo sudarymas

Metodas yra sudaromas per kelis etapus. Pirmiausia jis yra apmokomas, tada gerinamas ir galiausiai pritaikomas realiems duomenims.

Apmokymo metu naudojami žinomi duomenų pavyzdžiai. Panaudojami iš anksto suklasifikuoti savarankiškai sugeneruoti paketų srautai, imituojantys prievadų žvalgos ir užtvindymo SYN paketais atakas.

Gautiems rezultatams įvertinti naudojamos duomenų klasifikavimo metodams taikomos šios efektyvumo vertinimo metrikos: tikslumas (angl. *Precision*, P), atkuriamumas (angl. *Recall*, R) ir F-matas (angl. *F-measure*, F). Metrikos apskaičiuojamos remiantis šiais vertinimo kriterijais: teisingi rezultatai (angl. *True Positive*, TP), netikėti rezultatai (angl. *False Positive*, FP), trūkstami rezultatai (angl. *False Negative*, FN), teisingai atmesti rezultatai (angl. *True Negative*, TN). Tikslumas parodo teisingų rezultatų ir visų atpažinimų santykį (TP/(TP+FP)). Atkuriamumas parodo teisingų rezultatų ir visų srautų, kurie turi būti atpažinti kaip anomalūs, santykį (TP/(TP + FN)). F-matas parodo pusiausvyrą tarp tikslumo ir atkuriamumo (2 × P × R/(P + R)).

LOF metodas priklauso nuo gretimų taškų k skaičiaus parinkimo ir slenkstinės reikšmės. Gerinimo metu parenkamos įvairios k ir slenkstinės reikšmės ir tikrinama, su kuriomis iš jų gaunami geriausi rezultatai. Įvertinimui naudojami apmokymo metu geriausius rezultatus pasiekę požymių rinkiniai. Apmokymo metu gauti požymių rinkiniai, gerinimo metu gautos k gretimų taškų skaičiaus ir slenkstinių ribų reikšmės naudojamos realių paketų srautų analizei ir anomalių paketų srautų paieškai.

Metodo tyrimas

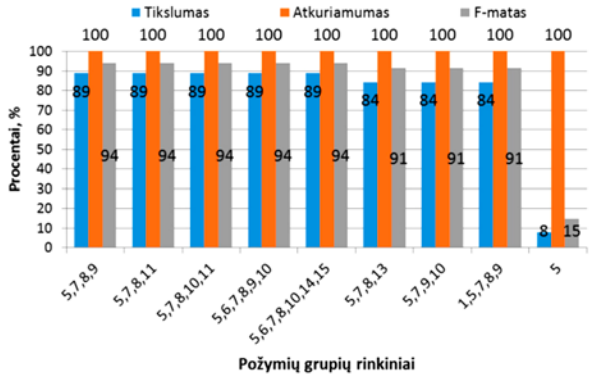
Metodo tyrimui buvo naudojami savarankiškai sugeneruoti tinklo paketų srautai, imituojantys prievadų žvalgą ir užtvindymo SYN paketais ataką. Generavimui buvo naudota „Kali Linux“ operacinė sistema, turinti nemažai įrankių

atakoms vykdyti. Tinklo paketų srautams registruoti buvo pasirinkti *fprobe* ir *nfcapd* įrankiai. Registruojant paketų srautus, yra sukuriami du atskiri įrašai apie kiekvieną sesiją tarp dviejų galinių tinklo mazgų. Sugeneruoti tinklo srauto duomenys buvo išanalizuoti, suagreguoti pagal siuntėjo ir gavėjo IP adresus bei sužymėti. Iš viso buvo gauti 3334 paketų srautai, iš jų 342 laikomi anomaliais. Agreguotų paketų srautų skaičius yra 612, o anomalių – 16. Taigi, agreguoti anomalūs paketų srautai sudaro apie 2,6 % visų agreguotų paketų srautų. Pagrindinis žvalgos atakos požymis yra įvairių prievadų naudojimas, o SYN paketais užtvindymo atakos pagrindinis požymis yra vienintelio 80 prievado naudojimas.

Geriausius rezultatus pasiekę požymių rinkiniai ir jų F-mato rezultatai, tiriant tinklo paketų srautus imituojančius prievadų žvalgą, pateikti 1 pav. Geriausius rezultatus parodė keletas įvairių požymių rinkinių. Visų jų atkuriamumas siekia 100 %, tikslumas 89 %, o F-matas 94 %. Galima teigti, jog visuose požymių rinkiniuose yra 5, 7, 8 požymių grupės. Penkta požymių grupė susijusi su paketų srautų trukme. Ši požymių grupė išskiria anomalius paketų srautus, nes žvalgos atakos metu generuojami paketų srautai trunka trumpą laiko tarpą. Septinta požymių grupė susijusi su paketų dydžiu. Žvalgos atakos metu siunčiami nedideli paketai tik su SYN požymiais ir kita tarnybine informacija, todėl palyginus su kitų paketų srautų dydžiu, juos galima laikyti išskirtimis. Aštunta požymių grupė susijusi su IP adresų kiekiu. Neslepiant tikrojo IP adreso, žvalgos atakos metu dominuojantis požymis yra IP adresas, iš kurio kreipiamasi į įvairius prievadus. Palyginimui su sujungtais požymių rinkiniais, 1 pav. pateikti rezultatai, kai naudojama tik viena požymių grupė. Nors ji ir tinkama, atpažino 100 % anomalių paketų srautų, tačiau jos tikslumas labai prastas – siekia vos 8 %. Vadinasi, viena požymių grupė be papildomų požymių grupių negali pakankamai tiksliai atpažinti anomalių paketų srautų.

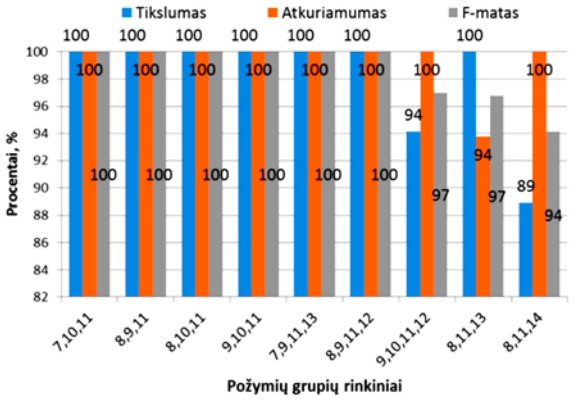
Geriausius rezultatus pasiekę požymių rinkiniai ir jų F-mato rezultatai tiriant tinklo paketų srautus imituojančius užtvindymą SYN paketais pateikti 2 pav.

Labai gerus rezultatus parodė aštuoni požymių grupių rinkiniai, iš jų keturi požymių grupių rinkiniai naudoja tik po tris požymių grupes, o jos yra iš 7, 8, 9, 10, 11 grupių sąrašo. Jos atpažino 100 % anomalių paketų srautų su 100 % tikslumu. Septinta požymių grupė susijusi su paketų dydžiu. SYN užtvindymo atakos metu, kaip ir žvalgos atakos metu, siunčiami tik paketai su SYN požymiais ir kita tarnybine informacija, dėl to, palyginus su visuma, tokius paketų srautus galima laikyti išskirtimis. Kita naudojama požymių grupė yra 8 – ji susijusi su IP adresų kiekiu. Skaičiuojami IP adresų kiekiai. Tiesioginės SYN atakos



1 pav. Geriausi F-mato rezultatai, tiriant tinklo paketų srautus, imituojančius prievadų žvalgą

Fig. 1. The highest F-measure values obtained by analyzing flows with port scanning activity



2 pav. Geriausi F-mato rezultatai, tiriant tinklo paketų srautus, imituojančius užtvindymą SYN paketais

Fig. 2. The highest F-measure values obtained by analyzing flows with SYN flood activity

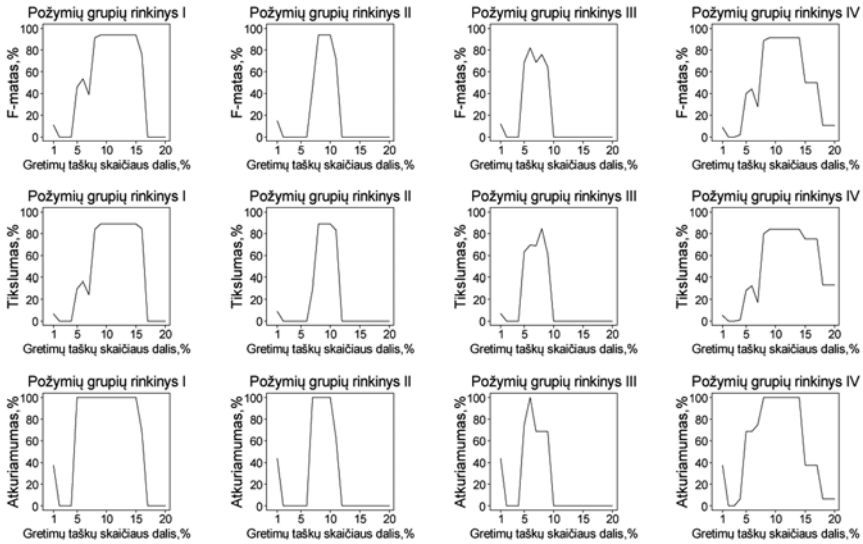
metu yra dominuojantis siuntėjo IP adresas. Jeigu atliekama SYN užtvindymo ataka su netikrais IP adresais, dominuoja keletas IP adresų. Taip pat vienas iš šios atakos požymių yra dominuojantis gavėjo IP adresas. Devinta požymių grupė susijusi su IP adresų kreipimosi skaičiumi. Tai irgi leidžia aptikti išskirtis, nes siuntėjo IP adresų kreipimosi skaičius nėra didelis. Jis kreipiasi tik į dominuojantį gavėjo IP adresą. Požymių grupės 10 ir 11 susijusios su prievadų skaičiumi. Tai labai ryškus požymis, nes SYN užtvindymo atakose išsiskiria vienas dominuojantis prievadas.

Vietinių išskirčių faktoriaus metodas priklauso nuo parinkto artimiausių taškų skaičiaus. Taip pat vietinių išskirčių faktoriaus metodu gauti rezultatai priklauso ir nuo parinktos slenkstinės reikšmės. Šie du parametrai daro įtaką aptikimo rezultatams. Apmokymo metu buvo rasti geriausi požymių grupių rinkiniai. Su šiais rinkiniais buvo tirta, kaip F-matas, tikslumas ir atkuriamumas priklauso nuo parinktų gretimų taškų skaičiaus ir slenkstinės vertės.

Žvalgos atakos atpažinimo gerinimo metu panaudoti tokie požymių grupių rinkiniai: 5, 7, 8, 9 (I); 5, 7, 8, 9, 10 (II); 5, 6, 7, 8, 9, 10, 14 (III); 1, 5, 6, 8, 9, 10, 11, 15 (IV).

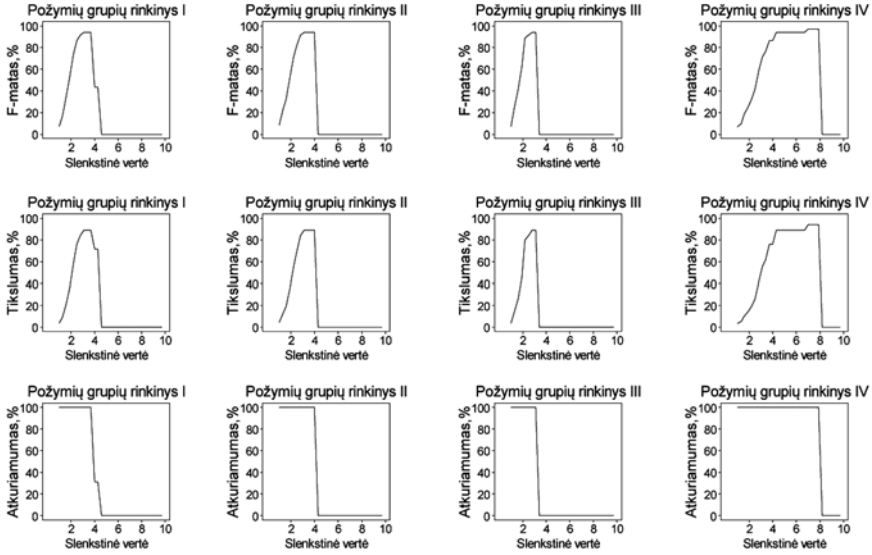
Žvalgos atakos atpažinimo rezultatų priklausomybės nuo gretimų taškų skaičiaus pateiktos 3 pav. Požymių I ir IV grupių rinkinių priklausomybės apima platesnes parenkamų gretimų taškų skaičiaus ribas. Rinkiniai II ir III apima siauresnį ruožą, jų atkuriamumas ir tikslumas siekia apie 90 %, kai gretimų taškų skaičiaus dalis yra tarp 5 % ir 13 %. Vadinasi, tam, kad būtų gautos mažo tankio išskirčių sritys, turi būti parinktas tikslus gretimų taškų skaičius, priešingu atveju, parinkus daugiau taškų, anomalūs paketų srautai bus priskiriami prie bendros duomenų srities. Palyginus I ir IV požymių grupių rinkinius, kai naudojamas skirtingas požymių skaičius, galima teigti, jog, parinkus daugiau kaip 17 % gretimų taškų, I rinkinio tikslumas ir atkuriamumas krenta iki 0 %, o IV rinkinio rezultatai taip staigiai nemažėja. Vadinasi, parinkus daugiau požymių grupių, aptikimo rezultatai ne taip stipriai priklauso nuo parinktų gretimų taškų skaičiaus. Įvertinus visus rinkinius, geriausi F-mato rezultatai gauti, kai gretimų taškų skaičius yra 9 % nuo visų duomenų skaičiaus.

4 pav. pateiktos žvalgos atakos atpažinimo rezultatų priklausomybės nuo parinktos slenkstinės vertės. Visos priklausomybės panašios tuo, jog, pasiekus atitinkamą ribą, aptikimo rezultatai staigiai sumažėja iki 0 %. Požymių I,



3 pav. Žvalgos atakos atpažinimo rezultatų priklausomybės nuo gretimų taškų skaičiaus pagal atskiras požymių grupes

Fig. 3. Port scanning detection dependencies on different parameter values



4 pav. Žvalgos atakos atpažinimo rezultatų priklausomybės nuo parinktos slenkstinės vertės pagal atskiras požymių grupes

Fig. 4. Port scanning detection dependencies on different threshold values

II ir III grupių rinkiniuose slenkstinė riba yra artima 5. Parinkus daugiau požymių grupių, slenkstinė riba yra apie 9. Geras tikslumas, siekiantis apie 90 %, pasiekiamas labai siaurose ribose – tada, kai slenkstinė reikšmė yra 2,8–3,5. Šiame ruože atkuriamumas siekia 100 %. Vadinasi, visi anomalūs paketų srautai yra sutelkti į vieną sritį, kuriai aptikti reikia tikslių parametrų reikšmių. Naudojant daugiau požymių grupių, kaip IV rinkinio atveju, gaunamas daugiau kaip 80 % tikslumas, pasiekiamas esant įvairesnėms slenkstinėms reikšmėms. Jos siekia 4,3–8. Atkuriamumas šiose ribose siekia 100 %. Vadinasi, išskirčių sritis pagal tankio įvertinimą yra pasiskirsčiusi šiek tiek plačiau, negu parinkus mažiau požymių grupių.

Analogiškai buvo tiriama ir užtvindymo SYN paketais atpažinimo rezultatų priklausomybė nuo gretimų taškų skaičiaus bei slenkstinės reikšmės. Nustatyta, kad ir šiuo atveju geriausi rezultatai pasiekiami pasirinkus gretimų taškų skaičių, lygų 9 % nuo visų duomenų skaičiaus. Slenkstinė reikšmė SYN užtvindymo atakos atveju yra šiek tiek mažesnė ir svyruoja 3,5–4,5 ribose. Vadinasi, anomalūs paketų srautai yra išsidėstę artimesnio tankio srityse.

Išvados

1. Anomalijų aptikimo metodo rezultatai priklauso nuo naudojamų požymių kiekio. Naudojant daugiau požymių grupių, aptinkama mažiau teisingų rezultatų, tačiau pasiekiamas geresnis tikslumas. Nustatyta, kad, naudojant daugiau požymių, išskirčių LOF įvertis yra mažesnis ir jos labiau sutelktos į vieną sritį. Naudojant mažiau požymių grupių, gaunami bendresni rezultatai, todėl tarp užfiksuotų anomalių paketų srautų patenka ne tik iš tiesų anomalūs, tačiau ir normalūs paketų srautai.
2. Įvertinus visus požymių grupių rinkinius, nustatyta, kad geriausi rezultatai pasiekiami, kai artimiausių taškų skaičius lygus 9 % nuo visų duomenų taškų.
3. Slenkstinės reikšmės parinkimas priklauso nuo naudojamo požymių grupių rinkinio. Rinkinio su daugiau požymių LOF įverčiai gaunami mažesni, todėl parenkama mažesnė slenkstinė vertė. Išanalizavus gautus rezultatus nustatyta, kad geriausiai anomalūs tinklo paketų srautai atpažįstami, kai slenkstinė vertė lygi 3,5.

Literatūra

Bilge, L.; Balzarotti, D.; Robertson, W.; Kirda, E.; Kruegel, C. 2012. Disclosure: detecting botnet command and control servers through large-scale NetFlow analysis, in *Proceedings of the 28th Annual Computer Security Applications Conference,* 3–7 December 2012, Orlando, Florida, 129–138.

Breuning, M. M.; Kriegel, H. P.; Ng, R. T.; Sander, J. 2000. LOF: Identyfying Density-Based Local Outliers, in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 15–18 May 2000, Dallas, USA, 93–104.

Cisco Systems. 2012. *Introduction to Cisco IOS NetFlow – a technical overview*. White paper, 16 2012.

Chandola, V.; Banerjee, A.; Kumar, V. 2009. Anomaly detection: a survey, *ACM Computing Surveys (CSUR)* 41(3): 15–58.

Cunningham, P.; Delany, S. J. 2007. *K-Nearest Neighbour Classifiers*. Technical Report UCD-CSI-2007-4. 17 p.

Dzemyda, G.; Kurasava, O.; Žilinskas, J. 2008. *Daugiamačių duomenų vizualizavimo metodai*. Vilnius: Mokslo aidai. 206 p.

Lazarevic, A.; Ertoz, L.; Kumar, V.; Ozgur, A.; Srivastava, J. 2003. A comparative study of anomaly detection schemes in network intrusion detection, in *Proceedings of the Third SIAM International Conference on Data Mining*, 1–3 May 2003, San Francisco, USA, 25–36.

Maimon, O.; Rokach L. 2010. *Data mining and knowledge discovery handbook.* New York: Springer. 1285 p.

Mohiuddin, A.; Abdun, N. M.; Jiankun, H. 2014. Outlier detection, in A-S. K. Pathan (Ed.). *The state of the art in intrusion prevention and detection.* New York: CRC Press.

OUTLIER DETECTION METHOD USE FOR THE NETWORK FLOW ANOMALY DETECTION

R. Ciplinskas, N. Paulauskas

Abstract

New and existing methods of cyber-attack detection are constantly being developed and improved because there is a great number of attacks and the demand to protect from them. In practice, current methods of attack detection operate like antivirus programs, i. e. known attacks signatures are created and attacks are detected by using them. These methods have a drawback – they cannot detect new attacks. As a solution, anomaly detection methods are used. They allow to detect deviations from normal network behaviour that may show a new type of attack. This article introduces a new method that allows to detect network flow anomalies by using local outlier factor algorithm. Accomplished research allowed to identify groups of features which showed the best results of anomaly flow detection according the highest values of precision, recall and F-measure.

Keywords: anomaly, anomaly detection methods, LOF, network flow, network attack.