

Įrodymais grįsto tinklo srauto duomenų modeliavimas ir analizė kibernetinių incidentų užkardymui

Ataskaita už 2025-2026 mokslo metų II pusmetį. Doktorantūros pradžios ir pabaigos
metai: 2023-2027

Doktorantas: Virgilijus Krinickij
Darbo vadovas: Prof. Dr. Linas Bukauskas

Vilniaus universitetas
Matematikos ir informatikos fakultetas
Informatikos institutas
Kibernetinio saugumo laboratorija

2026-10-08

- 1 Sutvarkytas straipsnis pagal recenzentų pastabas žurnalui „Baltic Journal of Modern Computing“.
- 2 Straipsnis priimtas publikuoti žurnale „Baltic Journal of Modern Computing“.
- 3 Rengiamas straipsnis žurnalui „Journal of Computer Virology and Hacking Techniques“.

- **Tyrimo objektas** – Duomenų srautų gautu asinchroninio įrašymo metu modeliavimas.
- **Tyrimo tikslas** – Sukurtas naujas algoritminis modelis kompiuterių tinklu perduodamų duomenų efektyviam srautų panašumo vertinimui ir savybių atpažinimui.
- **Uždaviniai:**
 - Sintetiniai atvejai kompiuterių tinklo srauto transliacijai.
 - Duomenų srautų, gautų asinchroninio įrašymo metu simuliacija, modeliavimas ir jų vertinimas.
 - Tinklo srauto parametrų ir duomenų modelio sukūrimas.
 - Algoritmų kūrimas incidentų šablonų atpažinimui.
 - Sukurtų algoritmų efektyvumo vertinimas ir įtaka saugumui.
 - Gautų mokslinių rezultatų taikymas realių duomenų srautų vertinime.

Iš ankstesnės publikacijos rezultatų jau žinome, kad:

- Asinchroninis įrašų lygiavimas **yra įmanomas** nustatant tinklo anomalijas, įspėjimus ir incidentus.
- Asinchroninis įrašų lygiavimas **nesutrumpina laiko** nurodytų problemų sprendimui.
- Asinchroninio įrašų lygiavimo modelio tobulinimas **yra perspektyvus naudojant skirtingus metodus** [KB24].

Definition (Duomenų šaltiniai)

Asinchroniniai duomenų šaltiniai iš heterogeninių taškų pateikiami kaip $\mathcal{D}_t = \{d^1, d^2, \dots, d^n\}_t$. Kur n - tai bet koks duomenų šaltinių, kurie gali būti stebimi sinchronizuotu laiku t , skaičius.

Visus tinklą stebinčius asinchroninius duomenų šaltinius apibrėžiame kaip $\mathcal{D}^*$ su bet koku laiko intervalu. Du skirtingu laiku esantys duomenų šaltiniai apibrėžiami kaip bet kurie du duomenų šaltiniai d_t^1 ir $d_{t'}^2$, kur $d_t^1, d_{t'}^2 \in \mathcal{D}^*$ tenkina sąlygą, kad $t \neq t'$. Laikas t yra laiko žyma, naudojama vietinių duomenų atsiradimui koreliuoti.

Definition (Duomenų srautas)

Apibrėžiame $\mathcal{S}_t$ kaip tinklo stebimų paketų srautą, siunčiamą iš duomenų šaltinio, kuris yra aktorius mašina. Aktorius gali būti užpuoliko arba taikinio mašinos. $\mathcal{S}$ srautas turi paketų įrašus k paketų su tinklo duomenų šaltiniu $d = d_t^i \in \mathcal{D}_t$. Tada $\mathcal{S}_t^d = \langle P_1, P_2, \dots, P_k \rangle_t^d$, kur k yra stebimo duomenų srauto paketų skaičius.

Jei duomenų srautas yra sinchronizuotas ir nenutrūkstamas laike, duomenų srautą žymėsime kaip $\mathcal{S}_{[t;\infty]}^d$ šaltinį $d = d_{[t;\infty]}^i \in \mathcal{D}^*$.

Definition (Paketų filtras)

Paketų filtras iš duomenų srauto $\mathcal{S}_t^d$ išskiria paketus, atspindinčius sudėtingus kibernetinių atakų modelius. Tegul ϕ yra loginis predikatas, apibrėžtas paketo savybėms. Apibrėžiame paketų filtrą $\mathbb{F}$ kaip funkciją $\mathbb{F} : (\mathcal{S}_t^d, \phi) \mapsto \mathcal{S}_t^{d'}$. Čia $\mathcal{S}_t^{d'}$ yra filtruotas duomenų srautas, kuriame yra tik tie paketai, kurių savybių vertės atitinka ϕ .

Definition (Paketo savybės vektorius)

Turint duomenų srautą S_t^d , siekiama apibrėžti unikalias paketų savybių vertes kiekvienam paketui P_t^d . Tegul $\mathcal{F} = \langle f_1, f_2, \dots, f_m \rangle$ yra paketo savybės, kur f yra P_t^d savybė. Kiekvienas P_t^d paketas S_t^d yra susijęs su baigtiniu savybių reikšmių rinkiniu. Šias savybių vertes žymime:

$$\mathcal{V}(P_t^d) = \langle p_{t1}^d, p_{t2}^d, \dots, p_{tj}^d \rangle$$

Kur j yra paketų savybių reikšmių skaičius sraute S_t^d .

Tegul $\mathcal{S}_t^d$ yra paketų srautas su savybėmis P_k ir atvykimo laiku $T(P_k)$, o $\mathbb{F} : (\mathcal{S}_t^d, \phi) \mapsto \mathcal{S}_t^{d'}$ sukuriantis filtruotą srautą $\mathcal{S}_t^{d'} = \{ P_k \in \mathcal{S}_t^d \mid \phi(P_k) = \text{true} \}$.

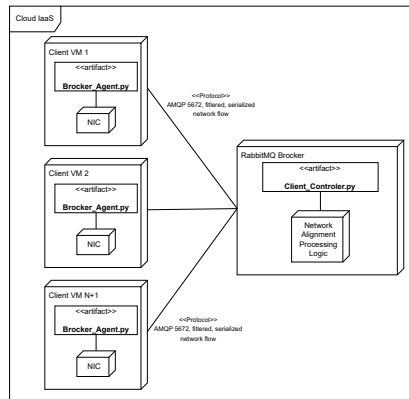
Definition (Išlygiuotas atitikimų laike langas)

Duotas lango dydis ω ir laikai $t_s < t_e$, šį langą apibrėžiame kaip $\mathcal{M}_{[t_s, t_e]}$, kaip filtruotų duomenų rinkinį filtruotame sraute:

$$\mathcal{M}_{[t_s, t_e]} = \langle P_k \in \mathcal{S}_t^{d'} \mid T(P_k) \in [t_s, t_e] \wedge \|\langle P_j \rangle\| \leq \omega \rangle \quad (1)$$

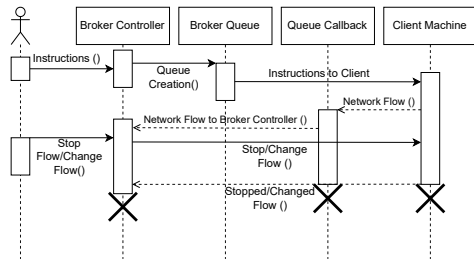
$\mathcal{M}_{[t_s, t_e]}$ yra tiksliai tie paketai, kurie atitinka ϕ kriterijų ir atkeliauja per $[t_s, t_e]$ daugiausia ω išgautų paketų.

- Kontroliuojama, heterogeninė tinklo aplinka.
- Asinchroninis, automatizuotas tinklo srauto įrašymas tarp grėsmės aktoiaus ir taikinio mašinos.
- Sintetiniai atakų scenarijai.



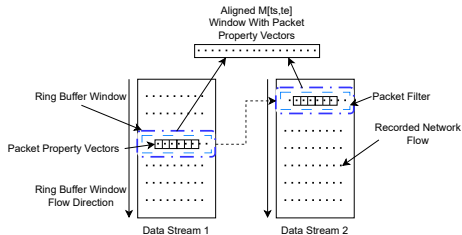
1 pav. Eksperimentinė aplinka

- Sukurta aplinka tenkina tinklo srautui keliamus reikiamivus priklausomai nuo paduodamų parametrų.
- Sukurtoje aplinkoje vykdomi eksperimentai susiję su naujos algoritminės metodikos pritaikymu dideliems tinlo srautams, kibernetinių atakų ir anomalijų detektavimui.



2 pav. Eksperimentinės aplinkos tyrimojo objekto judėjimo kryptis

- Pasinaudojus sukurta aplinka gaunami tinklo srautai kurie įrašomi į PCAP (angl. Packet Capture) failus. Taip pat naudojami srautai iš viešai pasiekiamų bibliotekų internete.
- Gauti objektai pritaikomi modeliui kibernetinių atakų aptikimui mažuose ir dideliuose tinklo srautuose.
- Paraleliai skaitomi PCAP failai iš heterogeninių tinklo taškų. PCAP failų skaičius yra neribotas.



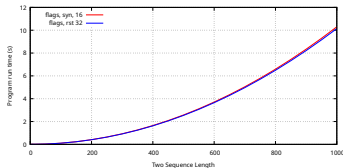
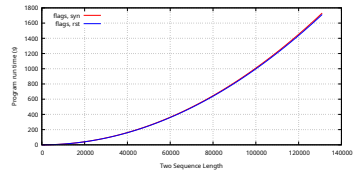
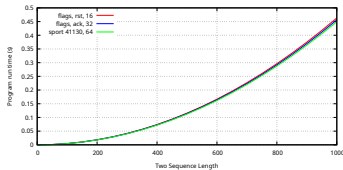
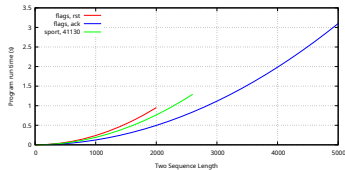
3 pav. Paralelinis PCAP failų apdorojimas su slenkančiais langais ir žiediniu buferiu

- Šablonų požymiai yra nepriklausomi nuo incidento pobūdžio.
- Modelyje naudojami slankūs langai su fiksuotu žiediniu buferiu (RBW) kuriame vyksta paketu filtravimas pritaikant mūsų įvedamus atakų šablonų požymius.
- Modelyje naudojamas grupavimas į intervalus (angl. Binning).
- Visi rasti požymiai yra išlygiuojami į vieną, bendrą $M_{[t_s, t_e]}$ langą tarp visų heterogeninių PCAP srautų įvykių įvertinimui [KB26].

1 Algorithm Duomenų srautų panašumų algoritmas

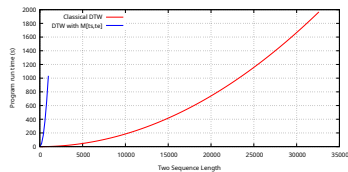
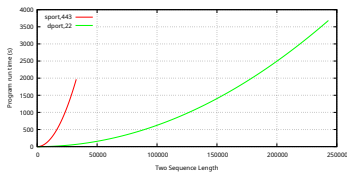
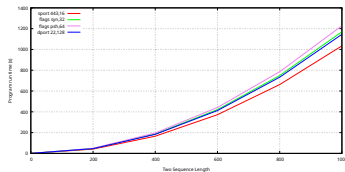
```
1: Input:  $S_t^d$ ,  $f$ ,  $\mathcal{V}(P_t^i)$ , RBW, threshold  $\Omega$ , bins  $\mathcal{K}$ 
2: for all  $S_t^d$  in parallel do
3:   for all packet  $p$  in  $S_t^d$  do
4:      $match(RBW(\mathbb{F}(p_t^d)))$ 
5:     extract  $p_t^d$ 
6:     if  $bytes \leq RBW$  then
7:       enqueue to RBW  $\leftarrow bytes$ 
8:     else
9:       dequeue RBW  $bytes \leftarrow 0$ 
10:    end if
11:  end for
12: end for
13:  $allV \leftarrow \cup match(RBW(\mathbb{F}(p_t^d)))$  if empty then return
14:  $alignedTS \leftarrow ts \leftarrow \min(allV)$ ,  $te \leftarrow \max(allV)$ 
15:  $M_{[ts, te]} \leftarrow \{p_t^d \in alignedTS \mid ts \leq t \leq te\}$ 
16: if  $useBinning$  then
17:    $seq \leftarrow Histogram(M_{[ts, te]}, \Omega, \mathcal{K})$ 
18: end if
19:  $dist \leftarrow DTW(s_1, s_2)$ 
20: return  $dist, t_{E2E}$ 
```

Rezultatai su mažais duomenų srautais generuotais pasinaudojus eksperimentine aplinka. Matuojamas efektyvumas tarp klasikinio DTW ir mūsų kuriamo modelio.



4 pav. PCAP faila su NMAP prievadų ir paketų vėliavų atakomis tarp atakuotojo ir taikinio virtualių mašinų

Efektyvumas tarp didelių srautų pasinaudojus klasikinio DTW ir mūsų siūlomo modelio atveju.



5 pav. PCAP objektai iš didelių srautų prieinami interneto šaltiniuose

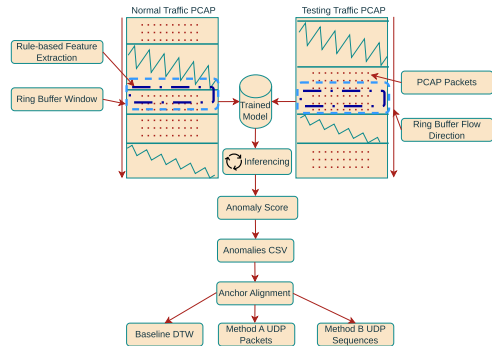
Kokios slenkančio lango savybės padeda DTW veikti greičiau:

- Fiksuota lango trukmė. Ribojamas sekos ilgis, DTW matrica pasiekia tik nustatytas ribas.
- Fiksuotas baitų limitas (RBW) vienam langui. Langai neapkraunami dideliais duomenų kiekiais.
- Žingsniavimas per langus. Gauname tankius, bet valdome kiekio langus, o ne nereguliaras milžiniškas atkarpas.
- Vektorių filtracija. Vietoje $n \times m$ „viskas su viskuo“ matricos DTW vykdoma tik ten kur reali prasmė.
- Lango atnaujinimas. Stumdant langą per srautą pasišalina pasene paketai neperkraunant viso lango iš naujo. Mažesnės sąnaudos ruošiant kandidatuojančius langus DTW.

- DTW su langais pritaikymas yra greitesnis išlygiuojant duomenis tarp skirtingų sekų į prasmingą langą.
- Greitis kyla dėl duomenų sumažinimo per langus kur DTW yra $O(n^2)$ problema mažinama iki $O(n \cdot m) \rightarrow O(n' \cdot m')$, kai $n', m' \ll n, m$
- Technologijos pasirinkimas turės įtakos efektyvumo vertinimui.

- Tolimesniame darbe tiriamė UDP paketo antraštę ir duomenų dalį (angl. *payload*), siekdami nustatyti, ar PCAP failuose užfiksuotų asinchroninių tinklo srautų vertinimas gali tinkamai pakeisti rankinį duomenų įvedimą. Mūsų siūlomi A ir B metodai lyginami su baziniu variantu, kuriame naudojamas tik klasikinis identifikavimas pagal laiko žymą. Metoda A naudoja anomalijos balą ir jo požymius, o ne laiko žymas. Taikant B metodą, užuot vertinę atskirus paketus, juos grupuojame į langus, vadovaudamiesi tuo pačiu RBW principu.
- Eksperimentams atlikti naudojami IEEE „DataPort“ tyrimo objektai PCAP formatu.

- Modeliui apmokyti naudojamas mašininio mokymo algoritmas „Isolation Forest“. Metodas A, kaip ir bazinis metodas, naudoja įvertintą CSV failą, tačiau anomalijos įvertis ir UDP paketo struktūros laukai naudojami kaip požymių vektorius, o ne laiko žymės.
- Taikant B metodą, užuot vertinus atskirus UDP paketus, jie sudedami į langus kaip sekos.
- Visiems metodams taikomas tas pats RBW principas su 64 kB buferiu.



6 pav. Asinchroninio įrašų išlygiavimo bazinio varianto, A metodo ir B metodo modelio architektūros

1 lentelė. Bazinis metodas (angl. baseline)
Anomalijos inkarų analizė pagal UDP paketo
laiko žymes ir išlygiavimas naudojant DTW.

Source A	Source B	Anchors A	Anchors B	DTW Distance	e2e [s]
10.0.0.104	10.0.0.101	2528	2480	414.1739	0.2131
10.0.0.104	10.0.0.102	2528	2985	456.8337	0.2223
10.0.0.104	10.0.0.103	2528	2885	332.0211	0.2109
10.0.0.104	10.0.0.93	2528	1	511.8486	0.2069
10.0.0.104	10.0.0.144	2528	1	511.8584	0.2074
10.0.0.101	10.0.0.102	2480	2985	230.4170	0.2104
10.0.0.101	10.0.0.103	2480	2885	132.1060	0.2108
10.0.0.101	10.0.0.93	2480	1	271.6284	0.2102
10.0.0.101	10.0.0.144	2480	1	271.8051	0.2067
10.0.0.102	10.0.0.103	2985	2885	256.7840	0.2129
10.0.0.102	10.0.0.93	2985	1	473.4797	0.2054
10.0.0.102	10.0.0.144	2985	1	473.4797	0.2069
10.0.0.103	10.0.0.93	2985	1	316.5375	0.2071
10.0.0.103	10.0.0.144	2985	1	316.5754	0.2077
10.0.0.93	10.0.0.144	1	1	0.0000	0.2050

2 lentelė. Metodas A. Anomalijos inkarų analizė
pagal UDP paketus ir išlygiavimas naudojant
DTW.

Source A	Source B	Anchors A	Anchors B	DTW Distance	e2e [s]
10.0.0.104	10.0.0.101	2528	2480	49.3802	0.0292
10.0.0.104	10.0.0.102	2528	2985	59.3606	0.0341
10.0.0.104	10.0.0.103	2528	2885	84.3868	0.0337
10.0.0.104	10.0.0.93	2528	1	50.8029	0.0001
10.0.0.104	10.0.0.144	2528	1	110.2392	0.0000
10.0.0.101	10.0.0.102	2480	2985	109.3421	0.0341
10.0.0.101	10.0.0.103	2480	2885	33.3725	0.0339
10.0.0.101	10.0.0.93	2480	1	67.8264	0.0001
10.0.0.101	10.0.0.144	2480	1	87.0493	0.0000
10.0.0.102	10.0.0.103	2985	2885	144.9284	0.0393
10.0.0.102	10.0.0.93	2985	1	93.6546	0.0001
10.0.0.102	10.0.0.144	2985	1	171.9167	0.0000
10.0.0.103	10.0.0.93	2985	1	95.4549	0.0000
10.0.0.103	10.0.0.144	2985	1	88.6677	0.0000
10.0.0.93	10.0.0.144	1	1	1.6038	0.0000

3 lentelė. Metodas B. Anomalijus inkarų analizė pagal UDP paketų sekas languose ir išlygiavimas naudojant DTW.

Source A	Source B	Sequences A	Sequences B	DTW Distance	e2e [s]
10.0.0.144	10.0.0.101	86	106	23.6899	0.0002
10.0.0.144	10.0.0.103	86	117	24.3152	0.0001
10.0.0.144	10.0.0.104	86	54	26.6407	0.0000
10.0.0.144	10.0.0.102	86	35	36.7009	0.0000
10.0.0.101	10.0.0.103	106	117	8.7211	0.0001
10.0.0.101	10.0.0.104	106	54	12.1685	0.0001
10.0.0.101	10.0.0.102	106	35	26.5110	0.0000
10.0.0.103	10.0.0.104	117	54	21.5010	0.0000
10.0.0.103	10.0.0.102	117	35	36.5061	0.0000
10.0.0.104	10.0.0.102	54	35	10.4956	0.0000

- 1 Konferencinio straipsnio rengimas į vieną iš preleminarių kandidatų.
- 2 Disertacijos rengimas.

4 lentelė. Egzaminai pagal studijų planą

Studijų metai	Egzaminai	
	Planas	Įvykdyta
I (2023/2024)	1	1
II (2024/2025)	2	2
III (2025/2026)	1	1
IV (2026/2027)		

Studijų metai	Dalyvavimas konferencijose				Publikacijos					
	Tarptautinėse		Nacionalinėse		Su citav. rodikliu			Be citavimo rodiklio		
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Būklė	Planas	Įvykdyta	Būklė
I (2023/2024)	1	1						1	1	Publikuota
II (2024/2025)					1	1	Publikuota			
III (2025/2026)					1		Įteikta			
IV (2026/2027)	1							1		
Iš viso:	2	1			2	1		2	1	

5 lentelė. Konferencijų ir publikacijų planas

- [KB24] Virgilijus Krinickij ir Linas Bukauskas. Asynchronous record alignment of network flows for incident detection and reconstruction. Iš *European Conference on Cyber Warfare and Security: Proceedings of the 23rd European Conference on Cyber Warfare and Security*. T. 23 iš numeris 1, p. 249–256. Academic Conferences International Limited, 2024.
- [KB26] Virgilijus Krinickij ir Linas Bukauskas. Modelling and assessment of asynchronous evidence-based network flows for cyber-attack detection. *Baltic journal of modern computing.*, 14(2):376–401, 2026.