Įrodymais grįsto tinklo srauto duomenų modeliavimas ir analizė kibernetinių incidentų užkardymui

Ataskaita už 2024-2025 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

2025-10-02

2024/2025 m. m. II pusmečio atlikti darbai

1. Išlaikytas privalomas dalykas „Šiuolaikinės duomenų bazių sistemos“.
2. Dalyvavimas „Cybersecurity and Privacy (CySeP)“ vasaros mokykloje. Švedija.
3. Pilnai sukurta duomenų srautų gavybos ir eksperimentų aplinka.
4. Modelio kūrimas incidentų atpažinimui ir sudėtingų kibernetinių atakų aptikimui mažuose ir dideliuose tinklo srautuose.
5. Modelio įgyvendinimas atliekant eksperimentus incidentų atpažinimui ir sudėtingų kibernetinių atakų aptikimui masyviuose tinklo srautuose.

„CySep“ vasaros mokykla

1 pav. Bendra dalyvių nuotrauka iš vasaros mokyklos

Tyrimo objektas, tikslas, uždaviniai

- **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.

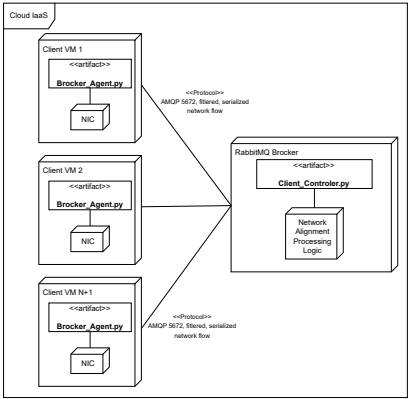
Mokslinių rezultatų pristatymas

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**.

Tyrimo metodas / aplinka

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



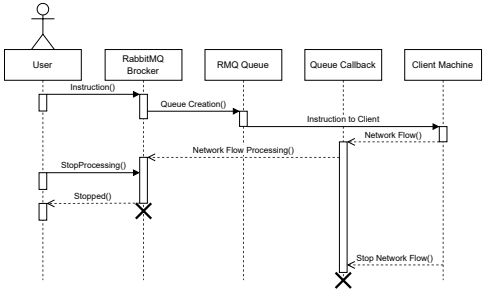
2 pav. Eksperimentinė aplinka

1 lentelė. **„Beacon" signalų parametrai ir jų reikšmės PCAP faile**

Protokolas	**Parametras**	**Reikšmė**
HTTP	http.request.uri	/beacon
HTTP	text	//4A
TLSv 1.3	tls.handshake.ja3	78f0dc5ac5b19daf131a133cfdee9691
TCP	tcp.dstport	5000
TCP	tcp.dstport	8888
TCP	tcp.flags	SYN

Aplinka algoritmams ir eksperimentams

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



3 pav. Eksperimentinės aplinkos tyriamojo objekto judėjimo kryptis

Definition (Duomenų šaltiniai)

Asinchroniniai duomenų šaltiniai iš heterogeninių taškų pateikiami kaip $\mathcal{D}_t = \{d^1, d^2, \ldots, d^n\}_t$. 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 kokiu 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.

Formalus apibrėžimas II

Definition (Duomenų srautas)

Apibrėžiame $\mathcal{S}_t$ kaip tinklo stebimų paketų srautą, siunčiamą iš duomenų šaltinio, kuris yra aktoriaus mašina. Aktorius gali būti užpuoliko arba taikinio mašinos. S srautas turi paketų įrašus k paketų su tinklo duomenų šaltiniu $d = d_t^i \in \mathcal{D}_t$. Tada $\mathcal{S}_t^d = \langle P_1, P_2, \ldots, P_k \rangle_t^d$, 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 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 $\mathbb{F} : \left(\mathcal{S}_t^d,\ \phi \right) \mapsto \mathcal{S}_t^{d\prime}$. Čia $\mathcal{S}_t^{d\prime}$ yra filtruotas duomenų srautas, kuriame yra tik tie paketai, kurių savybių vertės atitinka ϕ.

Definition (Paketo savybės vektorius)

Turint duomenų srautą $\mathcal{S}_t^d$, siekiama apibrėžti unikalias paketų savybių vertes kiekvienam paketui P_t^d. Tegul $\mathcal{F} = \langle f_1, f_2, \ldots, f_m \rangle$ 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:

$$\mathcal{V}(P_t^i) = \langle p_{t\,1}^d, p_{t\,2}^d, \ldots, p_{t\,j}^d \rangle$$

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 $\mathbb{F} : \left(\mathcal{S}_t^d, \phi\right) \mapsto \mathcal{S}_t^{d\prime}$ sukuriantis filtruotą srautą $\mathcal{S}_t^{d\prime} = \{\, P_k \in \mathcal{S}_t^d \mid \phi(P_k) = \texttt{true} \,\}$.

Definition (Išlygiuotas atitikimų laike langas)

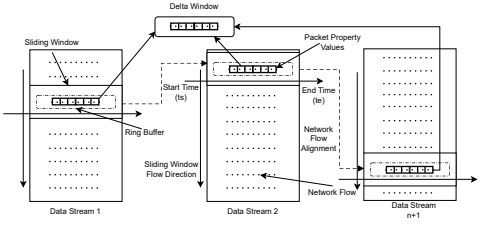
Duotas lango dydis $\omega \in \mathbb{N}$ ir laikai $t_s < t_e$, šį langą apibrėžiame kaip $\mathcal{M}_{[t_s,t_e]}$, kaip filtruotų duomenų rinkinį filtruotame sraute:

$$\mathcal{M}_{[t_s,t_e]} = \left\langle\, P_k \in \mathcal{S}_t^{d\prime} \;\middle|\; T(P_k) \in [t_s, t_e] \;\wedge\; \|\langle P_j \rangle\| \leq \omega \,\right\rangle \tag{1}$$

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

Duomenų srautų analizė

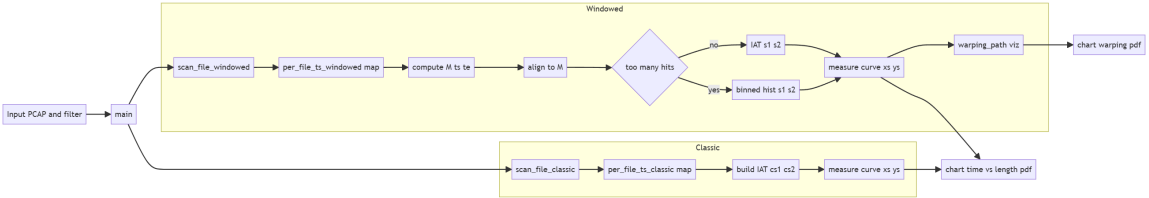
- Pasinaudojus sukurta aplinka gaunami tinklo srautai kurie įrašomi į PCAP (angl. Packet Capture) failus. Taip pat naudojami srautai iš viešai pasiekiamų bibliotekų internete.
- Gauti objektai pritaikomi modeliui kibernetinių atakų aptikimui mažuose ir dideliuose tinklo srautuose.
- Modelyje naudojami slankūs langai su fiksuotu žiediniu buferiu kuriame vyksta paketu filtravimas pritaikant mūsų įvedamus atakų šablonų požymius.



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

Tinklo srauto apdorojimas

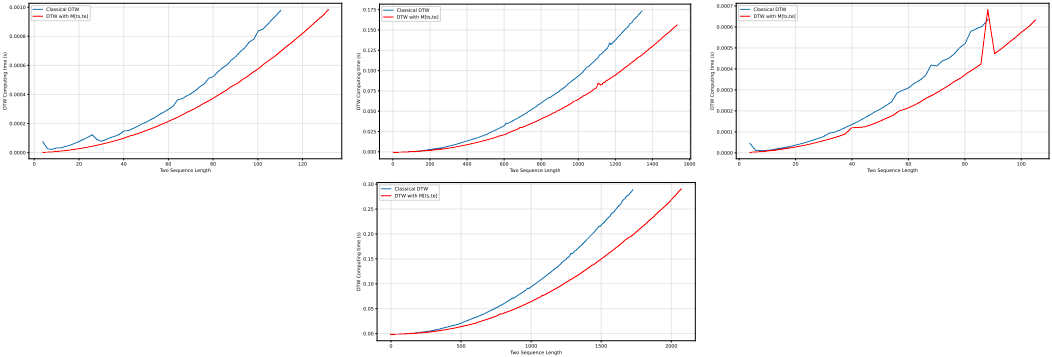
- Eksperimentu metu užkraunami PCAP failų duomenys išskiriami tarp klasikinio ir langų varianto.
- Toliau duomenys perduodami DTW skaičiavimams efektyvumo matavimams.



5 pav. Tinklo srauto aprodojimas iš PCAP failų panaudojant skirtingas DTW implementacijas

Tyrimų rezultatai mažiems srautams

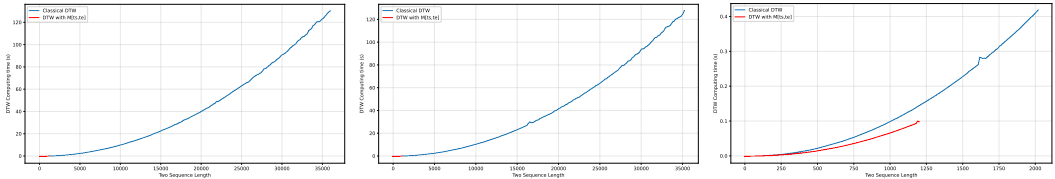
Rezultatai su mažais duomenų srautais generuotais pasinaudojus eksperimentine aplinka. Matuojamas efektyvumas tarp klasikinio DTW ir varianto su langais.



6 pav. Eksperimentinės aplinkos generuoti PCAP objektai su C2 tipo atakomis tarp atakuotojo ir taikinio virtualių mašinų

Tyrimų rezultatai dideliems srautams

Efektyvumas tarp didelių srautų pasinaudojus klasikinio DTW ir varianto su langais atveju.



7 pav. PCAP objektai iš didelių srautų prieinami interneto šaltiniuose

Slenkančio lango savybės

Kokios slenkančio lango savybės padeda DTW veikti greičiau:

- Fiksuota lango trukmė. Ribojamas sekos ilgis, DTW matrica pasiekia tik nustatytas ribas.
- Fiksuotas baitų limitas vienam langui. Langai neapkraunami dideliais duomenų kiekiais.
- Žingsniavimas per langus. Gauname tankius, bet valdome kiekio langus, o ne nereguliarias 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.

Išvados

- 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 $n', m' << n, m$
- Technologijos pasirinkimas turės įtakos efektyvumo vertinimui.

2025/2026 m. m. I pusmečio suplanuoti ir vykdomi darbai

1. Straipsnio pateikimas į „IEEE Access“ žurnalą.
2. Išlaikyti pasirenkamą dalyką „Tinklų modeliai ir algoritmai“.
3. Antro žurnalinio straipnio rengimas į vieną iš preliminarių kandidatų.
4. Galima trumpalaikė stažuotė viename iš preliminarių kandidatų.
5. Dalyvavimas „16th Data Analysis Methods for Software Systems“ tarptautinėje konferencijoje.

Doktorantūros studijų planas

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

2 lentelė. Egzaminai pagal studijų planą

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	Priimta
II (2024/2025)					1					
III (2025/2026)					1					
IV (2026/2027)	1							1		
Iš viso:	2	1			2			2	1	

3 lentelė. Konferencijų ir publikacijų planas

Egzaminai 2024-2025 I pusmetis		
Planas	**Įvykdyta**	**Būklė**
Šiuolaikinės duomenų bazių sistemos (2025-06-18)	Šiuolaikinės duomenų bazių sistemos (2025-06-18)	Išlaikytas

4 lentelė. Išlaikyti egzaminai