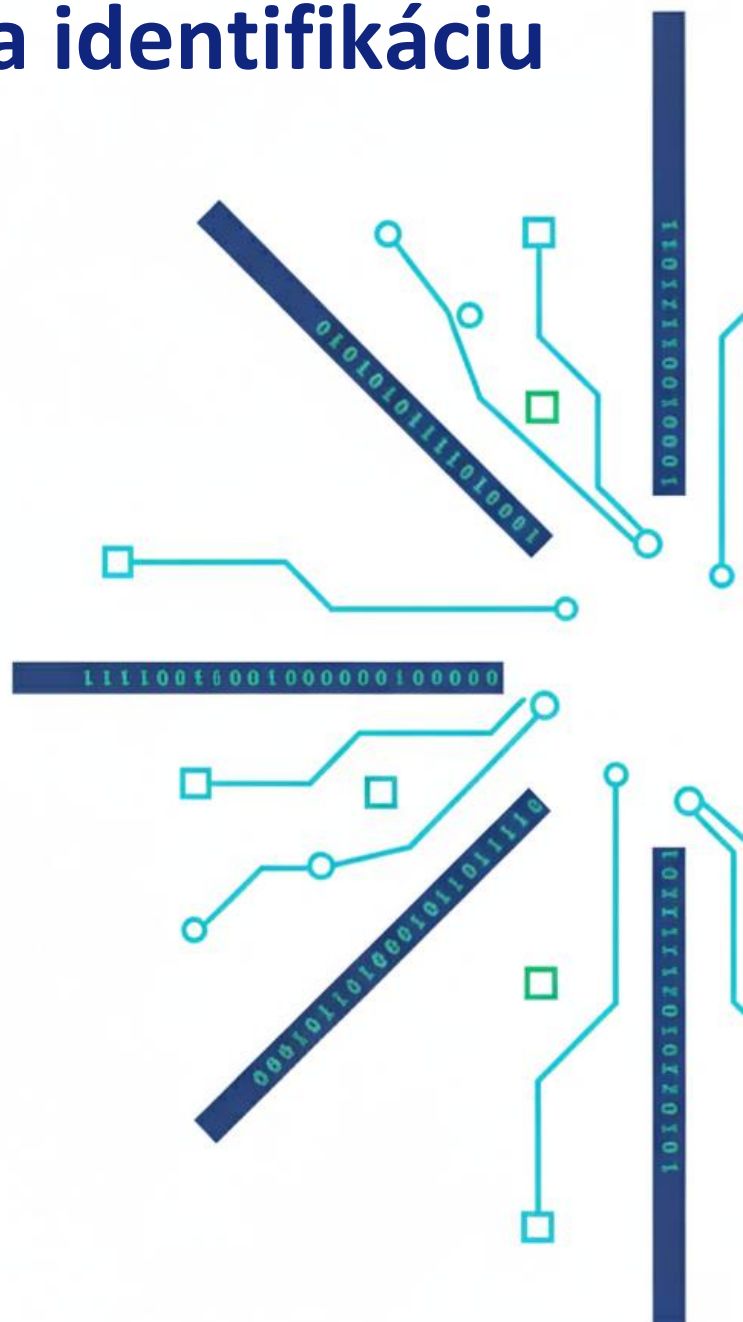


D20 – Metóda na identifikáciu vzťahov (II)



Projekt Automatizácia digitálnej forenznej analýzy a odpovede na incident (ADFIR) financovaný Európskou úniou – Next GenerationEU prostredníctvom Plánu obnovy a odolnosti Slovenskej republiky pod číslom projektu č. 09I05-03-V02-00079.

Obsah

Popis projektu	2
Úvod.....	3
1 Formálna konceptová analýza	4
1.1 Základné pojmy	4
1.2 Atribútové implikácie	5
2 Metóda na identifikáciu významných časových okien	8
2.1 Vstupné dáta a ich predspracovanie	8
2.2 Postup	9
3 Aplikácia metódy na identifikáciu významných časových okien	13
3.1 Predspracovanie dát	13
3.2 Vyhodnotenie výsledkov	14
Záver	18
Bibliografia	19
Prílohy	20

Popis projektu

Projekt **Automatizácia digitálnej forenznej analýzy a odpovede na incident** (ďalej len „ADFIR“) je financovaný Európskou úniou – Next GenerationEU prostredníctvom Plánu obnovy a odolnosti Slovenskej republiky pod číslom projektu č. 09I05-03-V02-00079. Tento projekt sa zaoberá jednou z kľúčových výziev v oblasti kybernetickej bezpečnosti a informačnej bezpečnosti – ako spracovať obrovské množstvo digitálnych dôkazov, ktoré vznikajú počas incidentov kybernetickej bezpečnosti alebo forenznych vyšetrovaní. V súčasnosti je tento proces veľmi náročný z hľadiska ľudských zdrojov a času. Automatizácia pomocou metód strojového učenia môže preto výrazne **zlepšiť kvalitu digitálnej forenznej analýzy** a skrátiť čas potrebný na jej vykonanie. Celkovo to umožňuje bezpečnostným tímom efektívnejšie reagovať na kybernetické hrozby. Hlavné prínosy tohto projektu sú:

- **Rýchlejšie riešenie incidentov v oblasti kybernetickej bezpečnosti.** Projekt ADFIR zavádza automatizované prístupy k zberu, spracovaniu a analýze digitálnych stôp. Vďaka tomu môžu bezpečnostné tímy rýchlejšie identifikovať príčiny incidentov a prijať účinné opatrenia na ich riešenie.
- **Zníženie pracovnej záťaže forenznych analytikov.** Rutinné a časovo náročné úlohy spojené so spracovaním digitálnych stôp budú nahradené automatizovanými metódami. To umožní analytikom sústrediť sa na zložitejšie prípady a strategické rozhodovanie.
- **Vyššia kvalita a konzistentnosť výstupov.** Použitie jednotných metodík a nástrojov zaručuje, že spracované digitálne stopy budú presnejšie, konzistentnejšie a ľahšie overiteľné. To výrazne znižuje riziko chýb spôsobených ľudskými faktormi.
- **Možné využitie v trestnom konaní.** Výstupy projektu budú vyvinuté v súlade s právnymi požiadavkami a normami, čo umožní, aby digitálne stopy boli akceptované ako relevantné dôkazy pre vyšetrovanie a súdne konania.

Úvod

Bezpečnostný analytik často pracuje s veľkým množstvom dát, v ktorom je bezpečnostne relevantná aktivita rozptýlená medzi bežnou aktivitou. Dôležitou úlohou v kybernetickej bezpečnosti je preto identifikovať také časti dát, ktoré sa svojimi vlastnosťami odlišujú od bežnej aktivity a môžu poukazovať na výskyt bezpečnostného incidentu.

Jedným z prístupov vhodných na analýzu vzťahov v dátach je formálna konceptová analýza. Umožňuje identifikovať skupiny objektov so spoločnými vlastnosťami a pomocou atribútových implikácií vyjadriť závislosti medzi týmito vlastnosťami. Atribútové implikácie tak zachytávajú kombinácie atribútov, ktorých výskyt podmieňuje prítomnosť ďalších atribútov.

Cieľom je navrhnúť metódu založenú na formálnej konceptovej analýze, ktorá umožní identifikovať časové okná charakterizované opakujúcimi sa kombináciami atribútov špecifickými pre daný časový úsek. Výstupom metódy sú pravidlá prepájajúce atribútové charakteristiky objektov s konkrétnymi časovými oknami. Takto identifikované úseky môžu poukazovať na výskyt bezpečnostného incidentu alebo inej neštandardnej aktivity a upozorniť analytika na časti dát, ktoré sú vhodné na ďalšiu analýzu.

1 Formálna konceptová analýza

Formálna konceptová analýza (FCA) [1] je matematický rámec založený na teórii zväzov, ktorý slúži na vyhľadávanie konceptov v binárnych dátach a ich hierarchické usporiadanie. Vstupné dáta možno reprezentovať ako binárnu tabuľku, v ktorej riadky zodpovedajú objektom, stĺpce atribútom a hodnota 1 vyjadruje, že daný objekt má príslušný atribút.

1.1 Základné pojmy

Formálne, nech B a A sú neprázdne množiny objektov a atribútov, a nech $R \subseteq B \times A$ je binárna relácia. Potom sa trojica (B, A, R) nazýva **formálny kontext** – príklad uvedený na **Obrázok 1**.

	i	ii	iii
a	×	×	×
b		×	
c			×

Obrázok 1 - Formálny kontext

Na hľadanie spoločných atribútov podmnožiny objektov a spoločných objektov podmnožiny atribútov sa používajú operátory $\uparrow: \mathcal{P}(B) \rightarrow \mathcal{P}(A)$ a $\downarrow: \mathcal{P}(A) \rightarrow \mathcal{P}(B)$, kde $\mathcal{P}(B)$ a $\mathcal{P}(A)$ označujú potenčné množiny množín B a A .

Pre podmnožinu objektov $X \subseteq B$ definujeme

$$X^\uparrow = \{a \in A \mid \forall b \in X: (b, a) \in R\}.$$

Analogicky, pre podmnožinu atribútov $Y \subseteq A$ platí

$$Y^\downarrow = \{b \in B \mid \forall a \in Y: (b, a) \in R\}.$$

Operátory $\uparrow$ a $\downarrow$ tvoria **Galoisovu konexiu**. Ich zložením $X^{\uparrow\downarrow}$ a $Y^{\downarrow\uparrow}$ získavame uzáverové operátory.

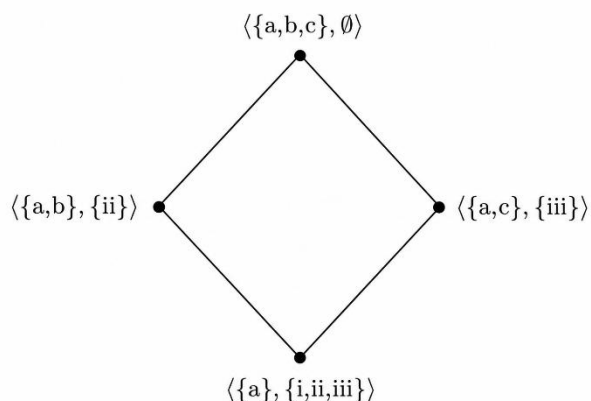
Dvojica (X, Y) , pre ktorú platí $X^\uparrow = Y$ a zároveň $Y^\downarrow = X$, sa nazýva **formálny koncept**. Objektová časť konceptu X sa nazýva extent a atribútová časť konceptu Y sa nazýva intent.

Formálny koncept teda vyjadruje úplnú dvojicu – uzavretú podmnožinu objektov spolu so všetkými atribútmi, ktoré má táto podmnožina spoločné. V tabuľkovej reprezentácii môžeme formálny koncept chápať ako maximálny obdĺžnik vyplnený jednotkami. To znamená, že ho nemôžeme rozšíriť o ďalší riadok alebo stĺpec bez toho, aby obsahoval nulu.

Množinu všetkých formálnych konceptov formálneho kontextu (B, A, R) označujeme $\mathcal{B}(B, A, R)$.

Formálne koncepty možno usporiadať pomocou vzťahu $(X_1, Y_1) \leq (X_2, Y_2)$ práve vtedy, keď $X_1 \subseteq X_2$, resp. $Y_2 \subseteq Y_1$.

Čiastočne usporiadanú množinu formálnych konceptov $(\mathcal{B}(B, A, R), \leq)$ nazývame **konceptový zväz**. Konceptový zväz môžeme vizualizovať pomocou čiarového diagramu (**Obrázok 2**).



Obrázok 2 - Konceptový zväz

Podľa základnej vety o konceptových zväzoch [1] je tento zväz úplný, teda pre každú podmnožinu formálnych konceptov v ňom existuje infimum aj supremum.

Binárne hodnoty atribútov však nie sú vždy postačujúce na zachytenie vlastnosti objektu, keďže vyjadrujú iba jej prítomnosť alebo neprítomnosť. Fuzzy rozšírenie formálnej konceptovej analýzy [3] preto umožňuje pracovať s hodnotami z intervalu $[0,1]$, pričom sa táto hodnota zvyčajne interpretuje ako stupeň príslušnosti objektu k atribútu. Navrhnutá metóda fuzzy formálnu konceptovú analýzu nevyužíva priamo, operácie nad formálnym kontextom sú použité výlučne nad binárnymi atribútmi. Pri predspracovaní časovej pečiatky sme sa však inšpirovali reprezentáciou hodnôt v jednotkovom intervale. Časová pečiatka sa najprv prevedie na číselnú hodnotu epoch time, následne sa normalizuje do intervalu $[0,1]$, a nakoniec sa diskretizuje do konečnej množiny hodnôt reprezentujúcich jednotlivé časové okná. Tieto hodnoty nepredstavujú stupne príslušnosti v zmysle fuzzy formálnej konceptovej analýzy, ale relatívnu časovú polohu objektov.

1.2 Atribútové implikácie

Kým konceptový zväz reprezentuje štruktúru dát prostredníctvom hierarchicky usporiadaných formálnych konceptov, atribútové implikácie [1][2] umožňujú vyjadriť závislosti medzi atribútmi obsiahnutými v tejto štruktúre vo forme logických pravidiel.

Formálne, nech A je neprázdna množina atribútov a nech $Z, W \subseteq A$. **Atribútová implikácia** nad množinou A je formálny výraz

$$Z \Rightarrow W,$$

ktorý vyjadruje, že každý objekt ktorý má všetky atribúty z množiny Z má zároveň všetky atribúty z množiny W . Ľavá strana pravidla sa nazýva **antecedent** a pravá strana **konzekvent**.

Implikácia $Z \Rightarrow W$ je platná na množine atribútov $Y \subseteq A$, ak

$$Z \subseteq Y \Rightarrow W \subseteq Y.$$

Ekvivalentne je implikácia $Z \Rightarrow W$ platná na Y , ak $Z \not\subseteq Y$ alebo $W \subseteq Y$.

Uvedená definícia opisuje platnosť implikácie vzhľadom na jednu množinu atribútov. Pri formálnom kontexte však pracujeme s viacerými objektmi, pričom každému objektu prislúcha vlastná množina atribútov. Zovšeobecňujeme preto platnosť implikácie na systém množín.

Nech $\mathcal{Y} \subseteq \mathcal{P}(A)$ je systém podmnožín množiny atribútov. Atribútová implikácia $Z \Rightarrow W$ je platná na $\mathcal{Y}$, ak je platná na každej množine $Y \in \mathcal{Y}$, teda ak pre každé $Y \in \mathcal{Y}$ platí

$$Z \subseteq Y \Rightarrow W \subseteq Y.$$

Nech (B, A, R) je formálny kontext. Každému objektu $b \in B$ prislúcha množina atribútov

$$\{b\}^\uparrow = \{a \in A \mid (b, a) \in R\}.$$

Formálny kontext (B, A, R) môžeme z hľadiska atribútových implikácií reprezentovať systémom množín

$$\mathcal{Y} = \{\{b\}^\uparrow \mid b \in B\},$$

ktorý obsahuje množiny atribútov prislúchajúce jednotlivým objektom.

Atribútová implikácia $Z \Rightarrow W$ je platná vo formálnom kontexte (B, A, R) , ak je platná na každej množine $\{b\}^\uparrow$. Teda pre každý objekt $b \in B$ platí

$$Z \subseteq \{b\}^\uparrow \Rightarrow W \subseteq \{b\}^\uparrow.$$

Podpora (support) antecedentu Z je počet objektov, ktoré obsahujú všetky atribúty zo Z :

$$\text{supp}(Z) = |Z^\downarrow|.$$

Podpora (support) atribútovej implikácie $Z \Rightarrow W$ je počet objektov, ktoré obsahujú všetky atribúty zo $Z \cup W$:

$$\text{supp}(Z \Rightarrow W) = |(Z \cup W)^\downarrow|.$$

Pravidlo $Z_1 \Rightarrow W$ sa nazýva **a-maximálne** [4], ak neexistuje pravidlo $Z_2 \Rightarrow W$, pre ktoré platí

$$Z_1 \subset Z_2$$

a zároveň

$$\text{supp}(Z_1 \Rightarrow W) = \text{supp}(Z_2 \Rightarrow W).$$

Antecedent Z_1 teda nemožno rozšíriť o ďalší atribút bez toho, aby sa zmenšila množina objektov podporujúcich pravidlo.

Atribútová implikácia $Z \Rightarrow W$ sa nazýva triviálna, ak $W \subseteq Z$.

Formálny kontext z **Obrázok 1** obsahuje 6 netriviálnych atribútových implikácií:

$$\begin{aligned}\{i\} &\Rightarrow \{ii\}, \\ \{i\} &\Rightarrow \{iii\}, \\ \{i\} &\Rightarrow \{ii, iii\}, \\ \{i, ii\} &\Rightarrow \{iii\}, \\ \{i, iii\} &\Rightarrow \{ii\}, \\ \{ii, iii\} &\Rightarrow \{i\}.\end{aligned}$$

A-maximálne sú posledné štyri atribútové implikácie.

2 Metóda na identifikáciu významných časových okien

Cieľom metódy je identifikovať časové okná charakterizované opakujúcimi sa kombináciami atribútov, ktoré sa v ostatných častiach datasetu nevyskytujú. Takto vymedzené okná môžu indikovať výskyt bezpečnostného incidentu alebo inej neštandardnej aktivity a upozorniť bezpečnostných analytikov na úseky dát vhodné na ďalšiu analýzu.

Výstupom metódy je množina atribútových implikácií, v ktorých ľavú stranu pravidla tvorí kombinácia binárnych atribútov a pravú stranu časové okno. Pre množinu objektov $E \subseteq B$, ktoré patria do rovnakého časového okna τ , má výsledné pravidlo tvar

$$E^\uparrow \Rightarrow \text{time_window}(\tau),$$

kde

$$E^\uparrow = \{a \in A \mid \forall b \in E: (b, a) \in R\}$$

predstavuje množinu všetkých binárnych atribútov spoločných pre objekty z E a $\text{time_window}(\tau)$ je predikát, ktorý hovorí, že objekt patrí do časového okna τ . Slovné môžeme pravidlo interpretovať takto: ak objekt obsahuje všetky atribúty z $E^\uparrow$, potom patrí do časového okna τ .

Pravidlo je a-maximálne vzhľadom na danú množinu objektov E , pretože jeho antecedent obsahuje všetky atribúty, ktoré sú spoločné pre každý objekt z E . Nemožno k nej pridať ďalší atribút bez toho, aby sa zmenila množina objektov spĺňajúcich pravidlo.

Parameter *minsupp* určuje minimálny počet objektov, na ktorých sa musí daný atribútový vzor v jednom časovom okne vyskytovať:

$$|E| \geq \text{minsupp}.$$

Hodnota *minsupp* = 1 vo všeobecnosti nie je vhodná, pretože umožňuje vytvárať pravidlá založené na jedinom objekte. Takéto pravidlá môžu opisovať iba individuálne alebo náhodné vlastnosti konkrétneho záznamu, nie opakujúci sa vzor charakteristický pre daný časový úsek. Vyššia minimálna podpora zabezpečuje, že identifikované charakteristiky sú potvrdené opakovaným výskytom pri viacerých objektoch.

Metóda tak ku každému identifikovanému časovému oknu priraduje uzavreté kombinácie atribútov, ktoré sa v ňom opakovane vyskytujú a v rovnakej kombinácii sa nevyskytujú mimo neho. Výsledkom sú časové okná spolu s atribútovými charakteristikami, ktoré sú pre nich špecifické.

2.1 Vstupné dáta a ich predspracovanie

Navrhnutá metóda predpokladá vstupný dataset obsahujúci binárne atribúty a jeden atribút reprezentujúci časovú pečiatku každého objektu. Binárne atribúty nadobúdajú hodnoty 1 a 0,

pričom vyjadrujú prítomnosť resp. neprítomnosť príslušnej vlastnosti. Časová pečiatka je určená dátumom a časom zaznamenania udalosti.

Časové pečiatky je potrebné previesť na hodnoty z jednotkového intervalu $[0,1]$, ktorý predstavuje štandardný rozsah hodnôt používaný vo fuzzy formálnej konceptovej analýze.

Časová pečiatka sa najskôr prevedie na číselnú reprezentáciu epoch time, ktorá vyjadruje počet sekúnd uplynutých od pevne stanoveného referenčného okamihu. Pre objekt b_i označme túto hodnotu ako e_i . Vysoké hodnoty epoch time sa následne normalizujú do jednotkového intervalu. Normalizovaná časová hodnota objektu b_i sa určí vzťahom

$$t_i = \frac{e_i - e_{\min}}{e_{\max} - e_{\min}},$$

kde $e_{\min}$ a $e_{\max}$ predstavujú najmenšiu a najväčšiu hodnotu epoch time v datasete. Hodnota $t_i = 0$ zodpovedá začiatku sledovaného obdobia a hodnota $t_i = 1$ jeho koncu.

Presné časové pečiatky sú však spravidla pre jednotlivé objekty jedinečné, zatiaľ čo cieľom metódy je analyzovať aktivitu v rámci širších časových úsekov. Z tohto dôvodu sa normalizovaný časový interval rozdelí na q rovnako širokých časových okien. Šírka jedného okna je

$$\Delta = \frac{1}{q}.$$

Každý objekt sa podľa svojej normalizovanej časovej hodnoty priradí do príslušného okna, pričom hraničná hodnota 1 patrí do posledného okna. Následne sa objektu priradí diskretná časová hodnota zodpovedajúca začiatku daného časového okna.

Objekty s blízkymi časovými pečiatkami sú tak zaradené do rovnakého časového okna. Diskretizácia umožňuje analyzovať opakujúce sa kombinácie atribútov v rámci spoločných časových úsekov namiesto posudzovania každej časovej pečiatky samostatne.

2.2 Postup

Krok 1 – Generovanie kandidátnych množín atribútov

Z množiny binárnych atribútov A sa generujú neprázdne kandidátne množiny atribútov

$$C \subseteq A,$$

pričom ich veľkosť je obmedzená parametrom $k_{\max}$, teda

$$1 \leq |C| \leq k_{\max}.$$

Množina všetkých preverovaných kandidátov je

$$C_{k_{\max}} = \{C \subseteq A \mid 1 \leq |C| \leq k_{\max}\}.$$

Každý kandidát C predstavuje potenciálnu kombináciu atribútov spoločnú pre určitú skupinu objektov.

Krok 2 – Výpočet extentu kandidáta

Pre každého kandidáta $C \in \mathcal{C}_{k_{max}}$ sa pomocou operátora $\downarrow$ určí množina objektov, ktoré majú všetky atribúty z C :

$$C^\downarrow = \{b \in B \mid \forall a \in C: (b, a) \in R\}.$$

Množina $C^\downarrow$ je extent kandidáta C .

Podpora kandidáta je definovaná ako počet objektov v jeho extente:

$$supp(C) = |C^\downarrow|.$$

Do ďalšieho spracovania sa ponechajú iba kandidáti spĺňajúci minimálnu podporu určenú hodnotou $minsupp$:

$$supp(C) \geq minsupp.$$

Množina ponechaných kandidátov je

$$\mathcal{C}^* = \{C \in \mathcal{C}_{k_{max}} \mid |C^\downarrow| \geq minsupp\}.$$

Krok 3 – Zoskupenie kandidátov podľa rovnakého extentu

Rôzne kandidátne množiny môžu určovať rovnakú skupinu objektov. Pre kandidátov $C_1, C_2 \in \mathcal{C}^*$ sa preto definuje ekvivalencia

$$C_1 \sim C_2 \iff C_1^\downarrow = C_2^\downarrow.$$

Pre každý unikátny extent $E \subseteq B$ sa vytvorí trieda kandidátov

$$\mathcal{E}(E) = \{C \in \mathcal{C}^* \mid C^\downarrow = E\}.$$

Každý unikátny extent E sa v ďalších krokoch spracuje iba raz. Kandidáti patriaci do tej istej triedy $\mathcal{E}(E)$ opisujú rovnakú skupinu objektov, hoci obsahujú rôzne kombinácie atribútov.

Krok 4 – Výpočet intentu

Pre každý unikátny extent $E \subseteq B$ sa pomocou operátora $\uparrow$ vypočíta jeho intent – množina atribútov spoločných pre všetky objekty z E :

$$E^\uparrow = \{a \in A \mid \forall b \in E: (b, a) \in R\}.$$

Množina $E^\uparrow$ predstavuje úplný spoločný atribútový opis objektov patriacich do extentu E . Keďže E vznikol ako extent niektorej kandidátnej množiny atribútov, platí

$$E^{\uparrow\downarrow} = E.$$

Množina $E^{\uparrow}$ je teda uzavretou množinou atribútov. Obsahuje všetky atribúty, ktoré možno zahrnúť do opisu bez zmeny príslušnej množiny objektov. V tomto zmysle predstavuje maximálny atribútový opis vzhľadom na extent E .

Krok 5 – Overenie časovej homogenity extentu

Pre každý extent E sa určí množina časových okien jeho objektov:

$$T(E) = \{\tau(b) \mid b \in E\}.$$

Extent E je časovo homogénny, ak všetky jeho objekty patria do rovnakého časového okna:

$$|T(E)| = 1.$$

Ak extent obsahuje objekty z viacerých časových okien, časové pravidlo sa preň nevytvorí.

Krok 6 – Vytvorenie uzavretého časového pravidla

Pre každý časovo homogénny extent E , ktorého objekty patria do časového okna τ , sa vytvorí pravidlo

$$E^{\uparrow} \Rightarrow \text{time_window}(\tau).$$

Antecedent $E^{\uparrow}$ obsahuje všetky atribúty spoločné objektom z extentu E a predstavuje maximálny atribútový opis vzhľadom na túto skupinu objektov.

Podpora pravidla je

$$\text{supp}(E^{\uparrow} \Rightarrow \text{time_window}(\tau)) = |E|.$$

Výsledné pravidlo opisuje úplnú podmnožinu atribútov spoločnú pre najmenej *minsupp* objektov, ktoré patria do rovnakého časového okna. Zároveň platí, že žiadny objekt mimo tohto časového okna nespĺňa celú ľavú stranu pravidla.

Krok 7 – Spätný prevod časového okna na pôvodný časový interval

Pre každé identifikované časové okno τ sa určí zodpovedajúci interval z pôvodnej časovej osi datasetu. Ak má časové okno v normalizovanom intervale $[0,1]$ tvar $[\tau, \tau + \Delta)$, jeho začiatok a koniec sa vypočítajú ako

$$e_{\text{start}} = e_{\min} + \tau(e_{\max} - e_{\min})$$

a

$$e_{\text{end}} = e_{\min} + (\tau + \Delta)(e_{\max} - e_{\min}),$$

kde $e_{\min}$ a $e_{\max}$ predstavujú najmenšiu a najväčšiu hodnotu epoch time v datasete a Δ je šírka časového okna. Pri poslednom časovom okne sa horná hranica intervalu nastaví na $e_{\max}$. Získané hodnoty epoch time sa prevedú späť na pôvodnú reprezentáciu dátumu a času.

Spätný prevod umožňuje priradiť identifikované pravidlá ku konkrétnemu obdobiu na pôvodnej časovej osi. Následne je možné vyhľadať objekty podporujúce jednotlivé pravidlá a analyzovať ich spolu s pôvodnými forenznými údajmi. Bezpečnostný analytik tak môže overiť, aké udalosti alebo aktivity sa v danom časovom úseku vyskytovali a či súvisia s bezpečnostným incidentom.

Identifikované časové okná možno ďalej analyzovať opakovanou aplikáciou metódy. Objekty patriace do vybraného okna sa vyčlenia, ich časové pečiatky sa opätovne normalizujú vzhľadom na časový rozsah tohto okna a následne sa rozdelia do menších okien. Opakovaná aplikácia metódy umožňuje presnejšie lokalizovať atribútové vzory a určiť kratšie časové úseky, v ktorých sa špecifická aktivita vyskytuje.

3 Aplikácia metódy na identifikáciu významných časových okien

Na aplikáciu metódy sme použili dataset z dátovej sady „Magnet CTF 2019“, ktorý bol agregovaný s použitím časového okna 900 sekúnd. Podrobné predspracovanie datasetu je popísané vo výstupe KPB2-02. Jednotlivé riadky datasetu reprezentujú časové zhľuky udalostí vytvorené v rámci tohto intervalu. Celý výstup z aplikácie metódy uvádzame v prílohe Magnet_CTF_2019_rules_900s.xlsx, kde sa v riadkoch nachádzajú id pravidiel, podpora pravidiel, ich dĺžka a samotné pravidlo.

3.1 Predspracovanie dát

Dataset obsahuje 5 830 objektov a 223 atribútov. V prvom kroku boli odstránené nebinárne atribúty, s výnimkou atribútu *time_cluster*, ktorý obsahuje dátum a čas príslušného objektu. Z pôvodných atribútov boli ako nebinárne identifikované a vyradené atribúty *version*, *container_id*, *id* a *cluster_length*. Po tejto úprave zostalo 218 binárnych atribútov a jeden časový atribút.

Časová pečiatka bola následne prevedená na číselnú reprezentáciu epoch time a normalizovaná do intervalu [0,1]. Normalizovaný interval bol rozdelený na 20 rovnako širokých časových okien so šírkou

$$\Delta = \frac{1}{20} = 0,05.$$

Každému objektu bola podľa jeho normalizovanej časovej hodnoty priradená diskretná hodnota reprezentujúca začiatok príslušného časového okna. Počty objektov v jednotlivých oknách sú uvedené v **Tabuľka 1**.

Začiatok časového okna	Počet objektov
0	35
0,05	24
0,1	29
0,15	27
0,2	28
0,25	11
0,3	23
0,35	31
0,4	54
0,45	23
0,5	10
0,55	36
0,6	16
0,65	583
0,7	602

0,75	615
0,8	856
0,85	901
0,9	993
0,95	933

Tabuľka 1 - Počet objektov v časových oknách

V ďalšom kroku boli medzi binárnymi atribútmi identifikované konštantné atribúty, ktoré nadobúdali hodnotu 0 pri všetkých objektoch. Takéto atribúty neumožňujú rozlišovať medzi objektmi a nemajú vplyv na výsledné pravidlá. Ich odstránenie preto zachováva informačný obsah relevantným pre navrhnutú metódu a zároveň znižuje dimenziu vstupných dát. Celkovo bolo odstránených 56 konštantných nulových atribútov. Konštantné atribúty s hodnotou 1 sa v datasete nevyskytovali.

Výsledný dataset použitý pri aplikácii metódy obsahuje 5 830 objektov, 162 binárnych atribútov a 1 diskretizovaný časový atribút. Časový atribút pritom neslúži ako súčasť kandidátnych množín na ľavej strane pravidiel, ale určuje časové okno uvedené na ich pravej strane.

3.2 Vyhodnotenie výsledkov

Metóda bola na dataset aplikovaná s parametrami $k_{\max} = 3$ a $\text{minsupp} = 3$.

Výstupom aplikácie metódy sú pravidlá pre **časové okná 0,9 a 0,95**. Identifikované časové okná zodpovedajú približne týmto intervalom z pôvodného datasetu: od 25. 2. 2019 08:00 do 9. 3. 2019 02:30, a od 9. 3. 2019 02:30 do 20. 3. 2019 21:15.

Pre časové okno 0,9 bolo identifikovaných 44 časových pravidiel, zatiaľ čo pre okno 0,95 až 710 pravidiel. Výsledky teda naznačujú, že najmä v poslednom časovom okne sa nachádza veľké množstvo opakujúcich sa atribútových vzorov, ktoré sa v rovnakej podobe nevyskytujú v žiadnom inom časovom úseku datasetu. Toto okno je preto z pohľadu navrhnutej metódy najvýraznejšie a môže poukazovať na špecifickú alebo neštandardnú časovo sústredenú aktivitu.

V **Tabuľka 2** uvádzame porovnanie vybraných časových okien a ich pravidiel.

Začiatok časového okna	Počet objektov	Počet pravidiel	Maximálna podpora pravidla	Maximálny počet atribútov v pravidle
0,9	993	44	9	47
0,95	933	710	379	53

Tabuľka 2 - Vybrané časové okná a ich pravidlá

Absencia pravidiel v ostatných časových oknách neznamena, že v nich neprebíhala žiadna významná aktivita. Pri zvolených parametroch však nebola nájdená dostatočne podporená kombinácia atribútov, ktorá by sa vyskytovala výhradne v jednom z týchto okien.

Podpora určuje, koľko objektov spĺňa dané pravidlo. Pravidlá s vyššou podporou opisujú opakujúcu sa aktivitu väčšej skupiny objektov. Naopak, pravidlá s nižšou podporou zachytávajú užšie, ale potenciálne bezpečnostne významnejšie kombinácie atribútov.

Ľavá strana pravidla obsahuje úplnú podmnožinu atribútov spoločnú pre objekty podporujúce pravidlo. Veľký počet atribútov automaticky nenaznačuje významnejšie pravidlo. Pravidlo často obsahuje všeobecné technické atribúty, ktoré sa spoločne vyskytujú pri danom type forenzného záznamu. Tieto atribúty sa nachádzajú vo veľkom počte pravidiel a môžu predstavovať skôr spoločný technický základ objektov než charakteristiku konkrétnej udalosti. Pri interpretácii sú preto zaujímavejšie atribúty, ktoré tento všeobecný základ dopĺňajú, napríklad atribúty súvisiace so skriptami, podozrivými doménami, spustiteľnými súbormi, prihlasovaním, či zmenou oprávnení.

Po zoradení pravidiel podľa podpory sú najviac podporované pravidlá sústredené v časovom okne 0,95. Ich bližšie preskúmanie ukazuje, že nejde o úplne odlišné pravidlá, ale prevažne o jeden spoločný základný vzor dopĺňaný ďalšími atribútmi. Spoločný základ reprezentuje súborové a NTFS artefakty z používateľského adresára a priečinku AppData, ktoré sú spojené s Chrome cache, udalosťami Windows a podozrivou doménou. V ďalších pravidlách bol tento základný vzor doplnený najmä o atribúty súvisiace s HTTPS komunikáciou a históriou prehliadača, obnovenými záznamami, zmenami oprávnení, udalosťami prihlásenia, záznamami Prefetch a opakovaným spúšťaním programov.

Prvých 10 pravidiel s najvyššou podporou obsahuje 18 spoločných atribútov:

- has_macb_accessed,
- has_macb_modified,
- has_modified_flag,
- has_sha256,
- is_accessed,
- is_file,
- is_file_entry,
- is_filename_attr,
- is_from_users_dir,
- is_in_appdata,
- is_ntfs,
- is_ntfs_event,
- is_standard_info,
- is_suspicious_domain,
- mask_Chrome_Cache,
- mask_File_stat,
- mask_NTFS_file_stat,

- mask_WinEVTX.

V

Tabuľka 3 sú k týmto pravidlám uvedené iba atribúty, ktorými sa ľavá strana pravidla od spoločného základu líši.

Číslo pravidla	Podpora	Dodatočné atribúty oproti spoločnému základu
45	379	-
46	173	has_transition_type, is_https, mask_Chrome_History
47	144	is_log_recovered
48	113	is_privilege_change
49	99	has_transition_type, is_https, is_log_recovered, mask_Chrome_History
50	65	is_login_event
51	61	mask_WinPrefetch
52	59	is_run_often, mask_WinPrefetch
53	56	mask_Registry_Key
54	55	has_transition_type, is_https, is_privilege_change, mask_Chrome_History

Tabuľka 3 - Vybrané pravidlá a ich dodatočné atribúty

Z pravidiel s najnižšou podporou boli vybrané tri potenciálne zaujímavé pravidlá, ktoré obsahujú atribúty súvisiace s bezpečnostne relevantnou aktivitou. Ich prehľad je uvedený v **Tabuľka 4**.

Číslo pravidla	Začiatok časového okna	Počet objektov	Počet atribútov	Atribúty
22	0,9	3	45	IsDynamicLibrary, IsInAppData, IsUserHive, MissingMinMACB, has_macb_accessed, has_macb_modified, has_modified_flag, has_sha256, is_accessed, is_executable, is_file, is_file_entry, is_filename_attr, is_from_system32, is_from_users_dir, is_in_appdata, is_in_system32, is_in_temp_or_spool, is_in_winsxs, is_link_file, is_log_recovered, is_login_event, is_ntfs, is_policy_change, is_privilege_change, is_rare_app, is_script_file, is_standard_info, is_suspicious_domain, is_system_path, is_unallocated, is_user_profile_shortcut, is_xml_file, key_new_registry_branch, key_suspicious_value, macb_is_only_access_birth, mask_Chrome_Cache, mask_File_entry_shell_item, mask_File_stat, mask_MSIE_WebCache_container_record, mask_NTFS_file_stat, mask_Registry_Key, mask_WinEVTX, mask_WinPrefetch, mask_Windows_Shortcut
726	0,95	3	28	accessed_local_file, accessed_sensitive_file, contains_secret_path, has_macb_accessed, has_macb_modified, has_modified_flag, has_sha256, has_transition_type, is_accessed, is_file, is_file_entry, is_filename_attr, is_from_users_dir, is_https, is_in_appdata, is_log_recovered, is_modified_after_access, is_ntfs, is_ntfs_event, is_standard_info, is_suspicious_domain, is_suspicious_keyword, mask_Chrome_Cache, mask_Chrome_History,

				mask_File_stat, mask_MSIE_WebCache_container_record, mask_NTFS_file_stat, mask_WinEVTX
735	0,95	3	27	has_macb_accessed, has_macb_modified, has_modified_flag, has_sha256, is_accessed, is_admin_activity, is_error_event, is_failed_login, is_file, is_file_entry, is_filename_attr, is_frequent_execution, is_in_appdata, is_log_recovered, is_login_event, is_ntfs, is_policy_change, is_privilege_change, is_run_often, is_standard_info, is_system_process, is_xml_file, mask_File_stat, mask_NTFS_file_stat, mask_Registry_Key, mask_WinEVTX, mask_WinPrefetch

Tabuľka 4 - Vybrané potenciálne bezpečnostne relevantné pravidlá

Pravidlo 22 obsahuje atribúty *key_suspicious_value*, *is_script_file*, *is_suspicious_domain*, *is_rare_app*, *is_system_path*, *is_in_temp_or_spool*, *is_unallocated*, *is_privilege_change* a záznamy registra. Atribút *key_suspicious_value* priamo označuje potenciálne podozrivú hodnotu registra. V kombinácii so skriptom, podozrivou doménou a systémovými alebo dočasnými cestami môže ísť o konfiguráciu súvisiacu so spustením neštandardnej aplikácie.

Pravidlo 726 obsahuje atribúty *accessed_local_file*, *accessed_sensitive_file*, *contains_secret_path*, *is_https*, *is_suspicious_domain*, *is_suspicious_keyword*, *mask_Chrome_History*, *mask_Chrome_Cache* a *is_modified_after_access*. Takáto kombinácia môže naznačovať prácu s citlivými údajmi v súvislosti s webovou aktivitou.

Pravidlo 735 obsahuje atribúty *is_failed_login*, *is_login_event*, *is_admin_activity*, *is_privilege_change*, *is_frequent_execution*, *is_run_often*, *is_system_process* a *is_error_event*. Táto kombinácia môže poukazovať na neúspešné pokusy o prihlásenie v súvislosti s administrátorskou aktivitou. Môže ísť o bežnú chybu používateľa, ale aj o pokusy o získanie prístupu.

Záver

Navrhnutá metóda umožňuje identifikáciu významných časových okien vo forenzných dátach, ktoré môžu poukazovať na výskyt neštandardnej alebo bezpečnostne relevantnej aktivity. Navrhnutý postup založený na metódach formálnej konceptovej analýzy využíva opakujúce sa kombinácie atribútov, ktoré sú špecifické pre vybrané časti datasetu. Výstupom sú časové pravidlá prepájajúce uzavretý atribútový opis skupiny objektov s konkrétnym časovým oknom. Metóda tak okrem lokalizácie časového úseku poskytuje aj jeho charakteristiky, ktoré možno ďalej využiť pri forenznej analýze.

Navrhnutá metóda negeneruje všetky platné atribútové implikácie s časovým atribútom v konzekvente. Pre každý identifikovaný časovo homogénny extant vytvára iba jedno pravidlo s maximálnym antecedentom, čím redukuje redundanciu výsledných pravidiel.

Výstup zároveň závisí od obmedzenia $k_{\max}$. Hodnota parametra $k_{\max}$ predstavuje kompromis medzi rozsahom prehľadávania a výpočtovou náročnosťou. Vyššie hodnoty môžu viesť k identifikácii ďalších časovo špecifických extantov, počet preverovaných kandidátnych množín však s rastúcim počtom atribútov prudko narastá. V uvedenom datasete umožnila už hodnota $k_{\max} = 3$ identifikovať časové okná obsahujúce špecifické atribútové vzory. Výsledky tak naznačujú, že aj nízka hodnota $k_{\max}$ môže postačovať na lokalizáciu relevantných časových úsekov.

Bibliografia

- [1] Ganter, B., Wille, R. (2024). *Formal Concept Analysis: Mathematical Foundations*. Springer.
<https://doi.org/10.1007/978-3-031-63422-2>
- [2] Carpineto, C., Romano, G. (2004). *Concept Data Analysis: Theory and Applications*. John Wiley & Sons.
<https://doi.org/10.1002/0470011297>
- [3] Bělohlávek, R. (2002). *Fuzzy Relational Systems: Foundations and Principles*. Kluwer Academic/Plenum Publishers.
<https://doi.org/10.1007/978-1-4615-0633-1>
- [4] Bayardo, R. J., Agrawal, R. (1999). *Mining the Most Interesting Rules*. ACM.
<https://doi.org/10.1145/312129.312219>

Prílohy

ADFIR-D20-KPB4-02-Magnet_CFT_2019_rules_900s.xlsx