Posudek oponenta diplomové práce

Student: Krůl Michal, Bc.
Téma: Korelace IPFIX záznamů z provozu proxy serverů (id 25050)
Oponent: Jeřábek Kamil, Ing., UIFS FIT VUT

1. Náročnost zadání **průměrně obtížné zadání**

Zadání je obtížnějšího charakteru, pokud se vezme v potaz rozmanitost různých protokolů a variant proxy serverů, které je nutné teoreticky pokrýt a zanalyzovat síťovou komunikaci. Následně navrhnout univerzální a robustní řešení. Což se v této práci však nestalo, značně se tedy omezila a tedy velice zjednodušila (viz dále). Proto snižuji náročnost zadání na průměrně obtížné.

2. Splnění požadavků zadání **zadání téměř splněno s vážnými výhradami**

Jednotlivé body zadání byly splněny takovým způsobem, že jsou v posuzované práci nějak textově obsaženy. Práce však pokrývá pouze minimální požadavky. Například co se týče bodu 5. zadání části testování a bodu 6. vyhodnocení, zde se objevují zmatečné, nepodložené nebo nic neříkající výsledky. Tyto části se tedy nedají považovat za splněné.

3. Rozsah technické zprávy **splňuje pouze minimální požadavky**

Rozsah práce se pohybuje spíše na dolní hranici minimálního počtu normostran. Toto množství je však na pováženou vzhledem k až někdy zbytečným obrázkům (minimálně co se velikosti týče viz str. 15 a 16), vykopírovaným zdrojovým kódům (značná většina zdrojových kódů včetně cca půl strany s funkcí print_help() viz str. 43).

4. Prezentační úroveň předložené práce **40 b. (F)**

Práce je strukturovaná do kapitol, které na sebe navzájem navazují. Práce je však těžce čitelná vzhledem k častému opakování lehce přeformulovaných bloků textu viz například strana 21.

Práce obsahuje popis typů proxy serverů poměrně povrchně spíše na úrovni letmého vysvětlení pojmů, přičemž zde chybí podrobnější popis funkcionality a komunikace těchto serverů což je ztěžejní pro tuto práci. Taktéž zde chybí podrobnější popis protokolů, kterým se tato práce následně věnuje, který je důležitý pro následný návrh, analýzu a realizaci, kde je již autorem zdá se tato znalost předpokládána. Například protokolu HTTP a HTTPS (TLS) je věnována pouze krátká sekce na straně 8, přitom se práce věnuje primárně těmto protokolům a atributům z nich extrahovaným.

Zvláště druhá část práce je pak dost zmatečná (jeví se jako psána ve spěchu). Například na straně 22 se objevuje obrázek (s poznámkou v titulku "vložit aktuální") jevící se jako schéma topologie sítě a následný popis jednotlivých komponentů, popis však neodpovídá schématu. V práci se pak vyskytuje popis vytvořeného algoritmu (sekce 5.4 str 38) a přidruženého vytvořeného pseudokódu, popis se však s pseudokódem rochází. Část testování je velmi zmatečná a není jasné, jakým způsobem byly vytvořeny použité záznamy, které a jak byly upraveny nebo jakým způsobem bylo řešení spouštěno pro ono testování.

5. Formální úprava technické zprávy **50 b. (E)**

Práce obsahuje typografické nedostatky, občasné překlepy, hovorové výrazy. V práci se vyskytují některá schémata zmenšená, obtížně čitelná (viz obr. 3.1 str. 11), a zároveň některá zbytečně veliká (viz již zmiňovaná str. 15 a 16).

6. Práce s literaturou **30 b. (F)**

Student se v práci odkazuje na relevantní zdroje z oblasti řešeného problému. Na některých místech se však jeví, že citace chybí. Zároveň není vždy z umístění citace v textu jasné, co konkrétně je přejato. Například na straně 4 je uvedeno definování standardu IPFIX dle RFC 7011, avšak části textu definice chybí nebo se rozchází s originální definicí, což pak může vést k nepřesnosti.

Problémem však je strana 28 a 29 kde student zmiňuje "inspiraci" diplomovou prací Korelace dat na vstupu a výstupu sítě Tor (Zdeněk Coufal, FIT VUT Brno, 2014) a následují vykopírované pasáže cca 1,5 strany z této práce 1:1 (str. 24-25 v původní práci). V práci zároveň není označeno, co konkrétně autor doslovně převzal. Zde se jedná o porušení citační etiky. Ve zbytku práce nejsou přesné reference na sekce, některé reference na obrázky jsou nedefinované ("??").

7. Realizační výstup **40 b. (F)**

Práce se teoreticky věnuje různým typům proxy serverů, primárně forwarding (v práci uváděno jako dopředné), zřetězené a reverzní proxy. Nicméně se zdá, že se řešením omezuje pouze na forwarding proxy (HTTP, HTTPS) formou mapování hodnot primárně aplikačních protokolů získaných sondou společnosti Flowmon a jakousi "časovou heuristikou" (v práci uváděno časová korelace). Při demonstraci však student přiznal, že spíše než o korelaci se jedná o závěrečné filtrování, zdali se nalezené korelace toků nerozcházejí v čase o napevno nastavené časové rozmezí.

V rámci práce student vytvořil datovou sadu obsahující vytvořený zachycený provoz, který následně využívá pro analýzu, návrh a testování výsledného řešení.

Řešení sestává z programu v jazyce Python, načítajícím csv soubor na vstupu a vytvářejícím csv soubor s mapováním identifikátorů nalezených korelovaných toků.

Vytvořený program je spustitelný a jeví se alespoň částečně jako funkční. Zdrojové kódy jsou nízké kvality. Implementační část textové práce jak již bylo zmíněno obsahuje popis zdrojových kódů včetně rozsáhlých vykopírovaných úryvků kódu.

Validační a verifikační část není dostatečná. Navíc údajně provedené testy nejsou nijak zdokumentované, prezentovány jsou pouze neurčité výsledky nad, jak student sám uvádí, předem upravenými vstupními hodnotami. Což pak z textu vyplývá, že vstupní hodnoty byly upraveny tak, aby program fungoval s následným zhodnocením dobrého výsledku (viz například strana 51 Testování časových korelací). Autor taktéž uvádí, že není možné testování provádět nad větším množství vzorků, jelikož při použití většího množství vzorků, není možné určit korektnost výsledku (viz str. 50 Testování správné korelace), toto tvrzení považuji za zavádějící a nepravdivé.

Při spuštění programu pouze s parametrem -i specifikujícím vstupní csv soubor nad k práci přiloženým souborem Local_server_03_HTTPS.csv s již vyexportovanými vstupními daty, který obsahuje záznamy o 10 TCP tocích (2 páry HTTPS a 3 páry HTTP) byl schopen odhalit 2 páry HTTPS toků. Mapování hodnot na takto malém vzorku HTTPS toků se jeví jako funkční, avšak z návrhu a zdrojových kódů vyplývají značné nedostatky, korektní funkčnost časové korelace nezjištěna. HTTP toky se řešením nad testovaným souborem korelovat nepodařilo.

Funkcionalitu korelace při zřetězených proxy nebylo možné ověřit, student tuto funkcionalitu nebyl schopen demonstrovat ani při osobní schůzce.

8. Využitelnost výsledků

Výsledný program by mohl být využit v rámci firmy Flowmon kdyby neobsahoval značné množství nedostatků.

9. Otázky k obhajobě

1. V části 5.4 > Popis algoritmu v bodě 4.1 uvádíte, že pokud podmnožina vyfiltrovaná podle JA3 hashe obsahuje přesně dva záznamy může jít o korelaci (taktéž obsaženo ve zdrojovém kódu). Proč zrovna přesně hodnota dva, bude toto stále platné i v případě zřetězených proxy serverů? Pokud přijmete tyto dva toky jako korelované, budou skutečně spolu vždy souviset?
2. V sekci 5.3 > Reverzní proxy servery uvádíte, že "pro tento typ proxy serverů nebylo nalezeno žádné řešení". Nezkoušel jste aplikovat korelační metody převzaté z již jmenované diplomové práce "Korelace dat na vstupu a výstupu sítě Tor" a aplikovat je na tento problém? Pokud je nelze použít zdůvodněte konkrétně proč.
3. Není při řešení pouze forwarding proxy pro HTTPS (TLS) jednodušší a dostatečné pro korelaci použít pouze například kombinaci Client Random (32 byte) a Server Random (32 byte), které jsou vždy dostupné dle TLS RFC, spojené dohromady jako klíč? Porovnejte s vašim řešením. Pokud ne, uveďte proč.
4. V sekci 7.2 > Testování robustnosti aplikace uvádíte, že "Z pohledu testování správnosti určování jsou nám velké soubory dat k ničemu". Prosím vysvětlete toto tvrzení.

10. Souhrnné hodnocení

40 b. nevyhovující (F)

Diplomová práce se zaměřuje na korelaci IPFIX záznamů toků, získaných Flowmon sondou, před, za a mezi proxy servery. Student nastudoval a seznámil se s principy proxy serverů a technologii NetFlow/IPFIX. Práce se však omezuje pouze na forwarding proxy servery a protokoly HTTP a HTTPS a mapování hodnot extrahovaných z toků, čímž se práce velmi značně zjednodušuje. Ani v takto zjednodušené práci není využit potenciál informací z toků extrahovaných. V rámci práce student nastavil virtuální prostředí, v němž vytvořil datovou sadu skládající se ze zachyceného síťového provozu, kterou následně používá pro analýzu, návrh a testování výsledného řešení. Textová část práce je těžkopádná s opakujícími se bloky textu a často velmi těžce pochopitelná. Práce zároveň

obsahuje cca 1,5 strany vykopírovaného textu z jiné diplomové práce 1:1 (je možné považovat za plagiát). Zejména druhá polovina práce se jeví jako psaná a dodělávaná ve velkém spěchu, což student sám při demonstraci přiznal. Nasvědčují tomu taktéž některé soubory z přiložené datové sady, které obsahují časové značky z doby blízké finálnímu termínu odevzdání práce. Pro hodnocení diplomové práce navrhuji známku F vzhledem k výše zmíněným nedostatkům. Student nezvládl práci řádně dokončit, textová část sice pokrývá každý bod zadání, avšak bod 4. zadání je splněn částečně, bod 5. část testování funkčnosti není dostatečné a bod 6. nelze možné považovat za splněné. Zároveň tedy navrhuji práci zadat k dopracování.

Prohlášení: Uděluji VUT v Brně souhlas ke zveřejnění tohoto posudku v listinné i elektronické formě.

V Brně dne: 3. června 2022

Jeřábek Kamil, Ing.
oponent