Review of Master's Thesis

Student: Surovič Marek, Bc.
Title: Static Behavioral Malware Detection over LLVM IR (id 18603)
Reviewer: Lengál Ondřej, Ing., Ph.D., UITS FIT VUT

1. Assignment complexity **more demanding assignment**

Po studentovi bylo požadováno vytvoření systému pro detekci malware na základě behaviorální analýzy z LLVM mezikódu. Toto na základě studia práce s mezikódem, analýzy kódu a souvisejících behaviorálních analýz.

2. Completeness of assignment requirements **acceptable under serious reservation**

Dle zadání měl student vytvořit systém pro behaviorální detekci malware. Byť to vypadá, že toto bylo splněno, zdá se, že student použil velké množství kódu třetích stran. Toto samo o sobě nepovažuji za problém, jen mi není zcela zřejmé, co vlastně bylo po studentovi požadováno. Jestli vytvoření rozhraní mezi LLVM mezikódem a existujícím nástrojem pro behaviorální analýzu, pak bylo zadání splněno.

3. Length of technical report **within minimum requirements**

Technická zpráva má 36 stran včetně obsahu a referencí, což odpovídá spíše bakalářské práci. V práci citelně chybí obsáhlejší popis provedených experimentů a srovnání a vymezení se vzhledem k souvisejícím přístupům. Větší množství příkladů by také nebylo na škodu.

4. Presentation level of technical report **65 p. (D)**

První část technické zprávy (kapitoly 1-3) je zajímavá a psána velmi hezky a až na drobnosti (lepší popis příkladů by neuškodil, některé obrázky jsou uvedeny bez odkazu a byla by dobrá trocha intuice v sekci 3.2.3) je opravdu radost ji číst. V druhé části jde bohužel kvalita dolů.

V sekci 4.2.1 se začíná objevovat jisté T, u kterého se mi nepodařilo nalézt, co znamená. Definice funkcí Propagate, Create a Sink je dána pouze neformálně a nejsem si úplně jistý, jak bych si je měl představit. Navíc mi není jasné, jak autor řeší situaci kdy instrukce je pro některé proměnné sink a pro některé proměnné source. V posledním odstavci sekce 4.2.1 mi není úplně jasné, jakým způsobem řešení autorem popsaného problému v Algoritmu 1 zapadá do dané abstraktní interpretace a kdy je vůbec volána. V Algoritmu 1 není úplně jasné, jakou funkci mají abstraktní kontexty. Podle definice by to mělo bý mapování proměnných na podmnožiny (oné tajemné domény) T, ale v některých místech se do kontextů přidává mapování instrukcí. Na řádku 11 je asi myšleno sekvenční procházení v rámci základního bloku.

U Algoritmu 2 by bylo dobré doplnit deklarativní popis toho, co algoritmus dělá---pokud se nepletu tak to je sjednocení grafu s jeho primed verzí a nahrazení hrany (u, v) za hranu (u, v'), bez nedosažitelných uzlů---a klidně bych zůstal u tohoto popisu: detailní popis této relativně standardní operace není moc zajímavý (navíc mám pocit, že algoritmus bude v každé iteraci cyklu unfoldovat původní hranu (u,v)).

V sekci 4.3 je zavedena funkce ComputeKey, jejíž smysl není úplně jasný. Pravda, klíče jí vypočítané se používají v Algoritmu 4, ale nejspíš jsou i nějaké požadavky na ně (použití hash funkce může způsobit, že všechny klíče klidně můžou být 0; nejsou nějaké požadavky na hash funkci?). Daná funkce má navíc nikdy nepoužitý vstupní parametr k---měl jsem pocit, že autor původně zamýšlel, aby funkce počítala hashe pro dané k v jednom postorder průchodu, ale později se uchýlil ke k preorder průchodům. Dále u Algoritmu 4 mi není zcela jasné, zda jde o přínos autora, či je to algoritmus z literatury, jaké jsou výstupní podmínky, a proč by vlastně algoritmus měl fungovat.

V sekci 4.4 mi není jasné, jak se bude chovat rovnice 4.1 pokud bude počet stromů přijatých automatem nekonečný.

5. Formal aspects of technical report **87 p. (B)**

Formální úprava technické zprávy je na velmi dobré úrovni. Práce je psána solidní angličtinou, počet překlepů je minimální. Typograficky je až na jednoslovný konec odstavce ve vrchní části strany 12 také na výborné úrovni. Jen seznam příloh opravdu není potřeba je-li jejich počet nula.

6. Literature usage **95 p. (A)**

Nenašel jsem problémy.

7. Implementation results **52 p. (E)**

Z celé práce je dost obtiížné zjistit, co je vlastně přínosem studenta. Pochopil jsem, že vytvořil most mezi LLVM IR a [2] a použil nástroj pro obfuskaci pro vytvoření mutací malwaru, ale nepochopil jsem proč to je vlastně potřeba dělat (co chybí [2] aby se dala použít sama o sobě?)

Pokud to dobře chápu, tak realizačním výstupem jsou:

1. taint analýza implementovaná jako LLVM průchod generující signatury (cca 500 řádků v C++)
2. prostředí pro generování obfuskovaných programů a jejich analýzu využívající analýzu v (1), LLVM obfuskátor a prototyp z [2] (cca 150 řádků v Pythonu).

Není mi jasné, jestli je Algoritmus 4 autorem někde implementovaný.Experimentální vyhodnocení je poskytnuto ve velmi omezené formě. Kromě popisu experimentů obsahuje jen vyhodnocení správné klasifikace malware (tj. true positive). Chybí informace o době trvání experimentů, prostředí, vyhodnocení false positive. Výsledky jsou poskytnuty "as is", chybí jejich diskuze, shrnutí, a srovnání s nějakým jiným přístupem.

Z předchozího to vypadá, že realizační výstup vskutku není příliš obsáhlý.

8. Utilizability of results

Práce by zřejmě šla použít jako základ pro vývoje nástroje pro detekci malware.

9. Questions for defence

1. Prosím vyzdvihněte váš přínos a zdůrazněte které části vyvinutého systému jsou vaše a které jsou použité již existující. (toto by mělo být v těle prezentace, není tedy nutné na otázku odpovídat zvlášť)
2. Prosím, vyjádřete se k výtkám v části 7.
3. Jak se můžete srovnat s jinými způsoby detekce malware používajících formální metody?

10. Total assessment **55 p. sufficient (E)**

Studentem vybrané zadání je obtížnější. Práce je psána zpočátku (kapitoly 1-3) velmi hezky a čtivě, část od kapitoly 4 dále se ale zdá šita horkou jehlou. Množství implementace je dle mého porozumění velmi malé a student by měl v rámci obhajoby vyjasnit, co je jeho přínos. Na základě těchto skutečností, podrobně rozebraných výše, hodnotím práci **známkou E**, pokud však student v rámci obhajoby poskytne přesvědčivé argumenty obhajující jeho přístup, doporučuji komisi hodnocení lepší známkou.

In Brno 8. June 2016

.....
signature