ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	**Použití Certificate Transparency pro detekci malwaru ze síťového provozu**
Jméno autora:	**Jan Karsch**
Typ práce:	diplomová
Fakulta/ústav:	Fakulta elektrotechnická (FEL)
Katedra/ústav:	Katedra počítačů
Oponent práce:	Ing. Martin Svatoš
Pracoviště oponenta práce:	Katedra počítačů

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání **průměrně náročné**

Hodnocení náročnosti zadání závěrečné práce.

K úspěšnému vypracování zadání se musel student seznámit s *Certificiate Transparency*, částí strojového učení pro detekci malware a práci s velkými daty v produkčním prostředí.

Splnění zadání **splněno**

Posuďte, zda předložená závěrečná práce splňuje zadání. V komentáři případně uveďte body zadání, které nebyly zcela splněny, nebo zda je práce oproti zadání rozšířena. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.

Všechny body zadání byly splněny.

Zvolený postup řešení **správný**

Posuďte, zda student zvolil správný postup nebo metody řešení.

Zvolený postup je správný.

Odborná úroveň **A - výborně**

Posuďte úroveň odbornosti závěrečné práce, využití znalostí získaných studiem a z odborné literatury, využití podkladů a dat získaných z praxe.

Autor prokázal schopnost aplikování školou získaných vědomostí (včetně nastudování témat mimo svůj obor) na realném problému (terabajty dat).
K menšímu pokulhání došlo v grafech (např. 6.2), kde chybí popis osy *y*; současně nejsou osy *y* dvou grafů v jednom obrázku naškálovány na stejné rozpětí. Práce také neobsahuje žádné zmínky o potřebách či limitaci HW (pamět, počet procesorů, výpočetní čas experimentů, apod.); pouze zmiňuje 11TB dat.

Formální a jazyková úroveň, rozsah práce **A - výborně**

Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku.

Práce je napsána velice čtivou angličtinou a ve svém velikém rozsahu obsahuje jen minimum chyb (např. přehozený slovosled). Jediným cizorodým prvkem je používání zkrácených tvarů („n't", „let's"), které se pro formální text nehodí. Některé obrázky a rovnice nejsou z textu přímo odkazovány (*see in Fig….*), což je poměrně netypické, ale na přehlednosti práce to není nijak znát.

Výběr zdrojů, korektnost citací **B - velmi dobře**

Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení závěrečné práce. Charakterizujte výběr pramenů. Posuďte, zda student využil všechny relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.

Formální použití citací je správné, avšak k rozsahu relevantních zdrojů mám následující poznámky: i) tvrzení zdůvodňující výběr metody náhodného lesa (kap. 7, první odstavec) by mělo být podpořeno citaci na relevantní zdroj (či soubor metod pro detekci malware založených na metodě náhodného lesa); ii) práci z mého pohledu chybí rešeršní paragraf o ostatních

přístupech pro detekci malware a to i ve spojení *Certificate Transparency* (jakýsi „*prior art*“ v tomto poli). V případě, že sekce 4.3.1 shrnuje veškeré vědomosti v tomto poli, je práce až moc skromná ve svém novátorství.

Další komentáře a hodnocení

Vyjádřete se k úrovni dosažených hlavních výsledků závěrečné práce, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, publikačním výstupům, experimentální zručnosti apod.

Práce předkládá slibné výsledky v oblasti detekce malwaru.

III. CELKOVÉ HODNOCENÍ, OTÁZKY K OBHAJOBĚ, NÁVRH KLASIFIKACE

Shrňte aspekty závěrečné práce, které nejvíce ovlivnily Vaše celkové hodnocení. Uveďte případné otázky, které by měl student zodpovědět při obhajobě závěrečné práce před komisí.

Předložená práce vypadá velice pěkně nejenom rozsahem, se kterým se autor hladce vypořádal, ale i díky dosaženým výsledkům, které napovídají, že v praxi použití *feature* založených na *Certificate Transparency* zlepšuje detekci malwaru. Studentovi předkládám následující otázky:

- Bylo provedeno více experimentů s rozdílným nastavením hloubky stromů nebo jejich počtu (hyperparametrů), aby byla vidět přidaná hodnota nových *feature* (z certifikátů)? V případě, že běžel pouze jeden experiment: jsme schopni nějak odvodit důležitost nových *feature* i v lesu se stromy menší nebo větší hloubky?
- Jak často by se měl model učit (týdně, měsíčně,…) v ideálním případě pro produkční pipeline? Je nějaký výrazný rozdíl v evaluačním čase základního a rozšířeného modelu?
- V sekci 5.2.5 se píše, že se práce nakonec věnovala přístupu *proof-of-concept* kvůli problému se zpracováním velkého objemu dat v jednom okamžiku. Dá se tento problém *velkého objemu* nějak kvantifikovat (např. počtem *hostnames*, certifikátů, terabajty)?

Předloženou závěrečnou práci hodnotím klasifikačním stupněm **A - výborně.**

Datum: 20.1.2021 Podpis: