Posudek vedoucího diplomové práce

Autor: **Ivan Nikolaev**

Název práce: **Network service anomaly detection**

Diplomová práce Ivana Nikolaeva se zabývá detekcí síťových služeb v počítačových sítích a následné detekci anomálních chování uživatelů těchto služeb a to za použití pouze statistické informace o komunikaci na síti ve formě NetFlows.

V teoretické části provedl diplomant krátkou kritickou rešerši existujících technik detekce síťových služeb a představil novou techniku využívající EM algoritmus. Algoritmus je použit bez zastavovací podmínky, takže se neustále přizpůsobuje aktuálnímu stavu sítě, který se během dne rapidně mění. Pro detekci anomálií na identifikovaných síťových službách využívá diplomant dvou přístupů: modelace síťového chování individuálních služeb a modelace chování jednotlivých uživatelů na službách. Pro oba přístupy využil diplomant vhodné modely pro výpočet hodnoty anomálie.

V poslední části práce diplomant navržené matematické modely experimentálně ověřil na datech z reálné sítě. Nechybí ani diskuze o možných využití těchto modelů a budoucí práci.

V průběhu zpracování práce student prokázal schopnost analyzovat problém, vymezit ho vzhledem k existujícím metodám, ty kriticky zhodnotit a vybrat optimální řešení. V průběhu implementace dokázal odhalit slabiny obecného algoritmu a pozměnit ho tak, aby poskytoval optimální výsledky z hlediska kvality závěrů i rychlosti. Rovněž vysoce hodnotím fakt, že práce je zpracována anglicky.

Předloženou diplomovou/bakalářskou práci hodnotím známkou: **A-výborně**.

Ing. Martin Grill
Katedra počítačů
Fakulta elektrotechnická
České vysoké učení technické v Praze
Zikova 1903/4
166 36 Praha 6

Posudek oponenta diplomové práce na téma:

Network Service Anomaly Detection

Autor diplomové práce: **Bc. Ivan Nikolaev**
Rozsah práce: 43 stran, 9 obrázků, 9 tabulek a 31 referencí

Oponent práce: **doc. Ing. David Šišlák, Ph.D.**

Posuzovaná práce je zaměřená na problematiku rozpoznávání nestandardního využívání počítačových síťových služeb založeného na statistickém modelování. Síťová komunikace je v dnešní době obsažená v mnoha každodenních činnostech naší společnosti. Závislost většiny činností na moderních technologiích přináší také rizika zejména v bezpečnostní rovině. Jednotlivci, jež mají velmi dobré znalosti těchto technologií, jsou schopni proniknout k velmi cenným informacím. Proto význam systémů pro automatizovanou identifikaci nestandardního chování při využívání síťových služeb neustále roste. Zvolené téma této diplomové práce je velmi aktuální.

Diplomové práce je přehledná, vhodně strukturována a doplněna řadou obrázků a tabulek. Největší část práce se věnuje práci studenta v oblasti rozpoznávání síťových služeb, modelování chování uživatelů těchto služeb a detekci jejich nestandardního využití. Z pohledu jazykového zpracování, text práce obsahuje přijatelné množství chyb. Zejména se jedná o chybějící čárky a členy. Z pohledu formálního provedení práce lze vytknout chybějící číslování vzorců, jež by umožnilo odkazovat se na jednotlivé definice jak v samotné práci, tak v případě uvedených nesrovnalostí. V textu práce jsem narazil na několik nesrovnalostí týkajících se korektnosti a úplnosti technického textu:

- Popis pod vzorcem na straně 10 definuje x jako poměr pro daný koncový bod. Nicméně není zde úplně jasné, o poměr kterých hodnot se jedná.
- Nejednoznačné použití značení jednotlivých veličin. Například x je jednou použito pro poměr požadavků bez odezvy ke všem požadavkům (strana 10), jindy je x použito pro poměr mezi počtem požadavků vyslaných konkrétním koncovým uzlem k počtu všech zpracovaných spojení stejným koncovým uzlem (strana 12).
- Použití nevhodných značení. Například vzorec na straně 12 používá π jako obecný koeficient a nikoli jako Ludolfovo číslo.
- V kapitole 3.7 je poměrně rozsáhle uvedené odvození výpočtu pro EM algoritmus nicméně zde chybí odvození jedné jeho části, a sice odvození π^{new}.
- Obrázek 6 v popisku zmiňuje logaritmickou škálu vertikální osy grafu, ale osa na obrázku má lineární měřítko a obsahuje také hodnotu 0.
- Poslední věta na straně 22 definuje řešení rovnic jako minimalizaci kvadrátu odchylek modelu, ale uvedený vzorec v této větě neobsahuje mocninu.
- Kapitola 5.2 používá metriku AUC. Tato zkratka je sice zavedená v seznamu zkratek na začátku práce, jež obsahuje její celý název, ale úplně zde chybí popis, vzorec či odkaz na literaturu jak se hodnota AUC spočte.

Zadání práce je splněno studentem ve všech bodech beze zbytku. Vlastní přínos studenta je patrný v jednotlivých částech práce. K obsahu práce mám několik specifických a také obecnějších otázek:

- Kapitola 3.2 popisuje algoritmus identifikace dvojic požadavek-odezva, jež předpokládám, probíhá v rámci jedné dávky toků (tedy toků z jednoho 5 minutového intervalu). V části popisu vzniků toků se zmiňuje, že informace o toku je zaznamenána při detekci RST nebo FIN příznaku v paketu. Pro každý směr stejného TCP/IP spojení je vytvořen samostatný tok. V rámci TCP/IP protokolu je zcela normální a běžné, že se obousměrné spojení se uzavírá v každém směru v jiném čase. Například uživatel služby odešle svůj požadavek a uzavře svůj směr spojení, ale čeká dále na odpověď od služby. Takto se snadno stane, že informace o jednotlivých směrech stejného spojení jsou zaznamenány v jiných dávkách toků. Další problém vidím v dlouhodobých spojeních, jež se nikdy neuzavřou, a tedy se nikdy nezaznamenají informace o tomto spojení. Jak je řešen tento problém?
- V kapitole 3.7 je EM algoritmus použit bez konvergenčního kritéria pro identifikaci koeficientů funkce $r_{pdf}(x)$, jež slouží jako základ pro určení prahové hodnoty pro rozhodnutí, zda koncový uzel je uzel nabízející službu či se jedná o uživatele služby. Na začátku jsou parametry funkce nastaveny na počáteční hodnoty. Při zpracování každé dávky toků se provede jen jedna iterace EM algoritmu a ta pak

slouží k identifikaci služeb. Další krok se provede až zpracováním další dávky toků. Kolik dávek toků je potřeba aby se identifikované koeficienty $\lambda_1, \lambda_2, \pi$ stabilizovaly pro danou síť?

- Pro globální modely použití služeb v kapitole 4.4 jsou uvedeny pouze dva modely. Proč se pro globální modely nepoužijí také ostatní modely jako Holt-Winter či auto-regresní?
- Experimenty v kapitole 5.1 porovnávají jedno-vláknový a paralelizovaný algoritmus identifikace dvojic požadavek-odpověď. Paralelizovaná varianta vykazuje přibližně dvojnásobné zrychlení. Nicméně celý experiment je proveden jen se čtyřmi výpočetními vlákny. Jaká je závislost zrychlení algoritmu na počtu výpočetních vláken? Má smysl přidávat stále více výpočetních vláken?
- V kapitole 5.4.1 se vyhodnocují kvalitativně jednotlivé metody predikce za použití dvou metrik. Nicméně obě metriky se počítají pouze z případů, kdy skutečná hodnota není nulová – omezení na vzorky T_{nz}. Takovéto omezení ovšem nezohledňuje situace, kdy predikce predikuje nenulové hodnoty. Proč je zde toto omezení, vyhodnocujeme-li přesnost predikce?
- Tabulky 4, 5 a 6 porovnávají chyby jednotlivých prediktorů v různých scénářích a na různých skupinách uživatelů. V textu je napsáno, jak se spočtou tyto metriky pro jednotlivé uživatele. Jak se ale agregují hodnoty pro skupiny uživatelů prezentované v uvedených tabulkách?
- Nelze všechny tyto popsané metody detekce služeb a jejich anomálního použití obejít snadno tím, že budou služby měnit porty, jež využívají v čase? Uživatelé budou schopni v každém čase jednoznačně určit jaký port nyní použít či použijí nějaký typ adresářové služby pro zjištění aktuálního portu služby. V případě protokolu IPv6 jsou možnosti v kombinaci s velkým rozsahem IPv6 adres daného uzlu obrovské a tak se ani v delším čase nebudou opakovat a model nepůjde ani vytvořit. Další možnost obejití IDS škodlivým kódem spatřuji v použití řady vlastních IP protokolu z rozsahu nedefinovaných protokolů a taky volitelných parametrů v hlavičce IP protokolu pro přenos dat a znemožnit tak detekci jednotlivých toků.

I přes několik výtek k formálnímu provedení práce zmíněných výše, řadím, na základě zhodnocení celkové úrovně diplomové práce, tuto diplomovou práci k těm výtečným. Proto, navrhuji práci ohodnotit stupněm **A výborně**.

V Praze dne 23. května 2014

doc. Ing. David Šišlák, Ph.D.