Posudek vedoucího diplomové práce

Autor: **Tomáš Komárek**

Název práce: **Passive NAT detection using HTTP logs**

Diplomová práce Tomáše Komárka se zabývá pasivní detekcí síťových zařízení využívajících systém pro překlad síťových adres (Network Address Translation neboli NAT) v počítačových sítích za použití pouze statistické informace o síťové komunikaci ve formě HTTP logů.

V teoretické části provedl diplomant krátkou kritickou rešerši existujících technik detekce NAT zařízení a představil novou metodu využívající techniky strojového učení. Navržený systém využívá metody učení s učitelem, která používá trénovací data. V oblasti síťové bezpečnosti je velice obtížné takováto data získat, proto navržené řešení využívá simulace ke generování síťového provozu uměle vytvořených NAT zařízení. Tento způsob umožňuje analyzovat daný problém s různými scénáři a extrémními případy. V experimentální části diplomant ukázal, že použitím SVM klasifikátoru, natrénovaném na takto vytvořených datech, lze dosáhnout vysoké přesnosti detekce, převyšující schopnosti existujících technik. Dále ukázal, že navržený postup má velice nízkou generalizační chybu, což ukazuje na možné použití již natrénovaného klasifikátoru v neznámém prostředí.

V průběhu zpracování práce student prokázal schopnost analyzovat problém, vymezit ho vzhledem k existujícím metodám, ty kriticky zhodnotit a navrhnout řešení poskytující výrazně lepší výsledky z hlediska kvality závěrů i rychlosti. Samotná práce je sepsána v anglickém jazyce a v některých místech působí text poněkud těžkopádně, zatížený teorií ne zcela nutnou k vysvětlení navrhovaného řešení. I přes tuto výtku splňuje práce všechny požadavky kladené na diplomovou práci a doporučuji ji k obhajobě.

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

Ing. Martin Grill
Katedra počítačů
ČVUT v Praze

Martin Rehak, Ph.D.
Cisco Systems (Czech Republic) s.r.o.
rehak@cisco.com

Assessment of the Diploma Thesis:

Tomas Komarek: Passive NAT Detection using HTTP access logs

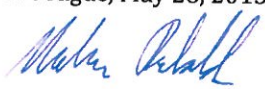
Summary and recommendation: I had a pleasure of reviewing the diploma thesis of Bc. **Tomas Komarek**, entitled: "**Passive NAT Detection using HTTP access logs**". The thesis presents a significant contribution to the state-of-the-art approaches in the area. The student has designed, implemented and evaluated a method that can reliably distinguish between the HTTP requests generated by a single device and the HTTP requests generated by a set of devices hidden behind a single IP address. *The thesis fulfils the requirements commonly associated with the diploma thesis and I recommend its acceptance and grade it as* ***Excellent*** *(A).* I also recommend the student for further study at Ph.D. level.
On the other hand, the thesis is not without downsides. I *do not* recommend distinguishing it with Dean's prize or any other award due to the eclectic and highly unusual style of writing. This makes the contribution of the student harder to discern and does not show the ability to independently document his research.

Questions:

Could you explain whether the distribution of NAT devices in the generated mix represented the real-world distribution of observed NAT devices in terms of size, activity and other characteristics?

In description of pre-processing on page 29, I'm not entirely clear over which domain is the normalization performed. Is it over values of a single feature over time (reader's assumption), across users, or across other domain?

In Prague, May 28, 2015



Martin Rehak

Detailed remarks for the student:

Source port discussion should have been presented as a trade-off and a deliberate choice made by the author, rather than a necessity or a property of the domain. It is simply one of the downsides of the proposed training schema.

Some places of Section 4.5 read like a lab notebook of a student reading through ML for the first time. Not a big problem, shows learning and right reading, but bit unusual. I'd have preferred this discussion in a dedicated Related work or Methodology section.

The thesis discusses subjects that are only tangentially relevant to the contribution: discussion of VC dimension, discriminative / generative learning, and other topics.

Removal of large negative outliers based on assumption of them being proxies is not unreasonable, but should have been measured, verified and documented.

Fig. 4.11. and relative importance of features may be biased due to the randomize construction of the synthetic training set, where the devices are merged randomly. In some environments, the devices in behind a single proxy can be more standardized/uniform. A typical example would be a call center with enforced standardized configuration of HW and SW, compared to personal laptops issued to regular employees.