Master Thesis Review

Thesis Author: Bc. Vojtěch Létal

Thesis Title: **Discovering of malicious domains using WHOIS database**

Reviewer: RNDr. Petr Somol, Ph.D. (ÚTIA AV ČR, Cisco Systems)

Mr. Letal's thesis addresses a highly interesting and complex problem in an area of top interest. Network security is constantly gaining importance with rapidly increasing proliferation of the Internet, the more so in the age of Internet of Things. It is constantly very difficult to keep security under control, with growing numbers of adversaries who follow military, political, and with growing importance also economic interests (cf. the massive proliferation of click-fraud technology, adware, ransomware and the like).
Massive investments are being made by global corporations into this area in the now standard fields including firewalls, anti-malware scanners, sandboxes etc.. This is not enough. Previously unheard of defense principles are needed to keep up in the race with attackers.
One of the almost unexplored options is to aim at predicting malicious activity before it actually happens. Very limited results have been achieved in this area so far and Mr. Letal's thesis topic has thus been selected very well.
Mr.Letal follows the recent idea of predicting maliciousness of domain addresses by discovering pools of pre-registered domains maintained by malicious actors – where only a portion of addresses has been used so far and the rest kept for replacement of blocked ones. This principle has led security teams to significant discoveries of malicious campains throughout 2014.
The novelty in Mr. Letal's approach is in addressing the problem with utmost rigor and generality. Mr.Letal – having consulted with senior specialists in Bayesian modeling and Machine Learning in Security – proposes a generic framework built on very solid foundations – probably the most solid in industry to date.
The thesis is in fact a joy to read, due to its depth, its serious treatment of the state of the art and overall presentation quality. The text is very focused, non-redundant, notation is well introduced and kept, presented material is well structured. Experimental evaluation is meaningful and convincing.
The proposed methodology is verified on WHOIS database with results applicable in corporate environment. At the same time its generality allows for its application beyond just WHOIS – any source of keys connecting network entities can be used to propagate the probability of maliciousness, starting from a limited number of confirmed cases.

I have found only these non-prohibiting issues:

- Variable identifier b is used in two conflicting ways – as label blocked vs. unblocked, and as one of two parameters a and b of the Beta distribution. I would recommend to use different letters for these two purposes – the current frequent conflicting occurrence throughout the thesis makes reading more difficult
- When introducing equation (3.33) you refer to "hyper parameters a_s and b_s". Is it a typo? The equation is about a_d and b_d

- Description of Algorithm 1 is brief but difficult to follow as it is not self-explanatory. The reader can only assume that convergence is evaluated somehow using $q(m_d)$, $q(a_k)$, $q(b_k)$ but it is difficult to find out how. Also, it would help if each sub-step (a), (b), (c) contained reference to the very equations that serve the updates.
- Table 4.1 shows parameters used for a simple example – how were the parameters chosen?
- When introducing equations (4.11) and (4.12) you state "we have replaced the original vector of parameters... by three new parameters \sigma, *m*, *s* as" but the equations seem to be based on *m*, *s*_\Beta, *s*_\Gamma instead. Is it a mistake ? Also, Table 4.6 shows apparently both sets of interchangeable parameters at once – which were in fact chosen and which were transformed from the chosen ?

To summarise, the material in this thesis is from many perspectives more of Doctoral than Master thesis quality , the difference being mainly just in the extent given by Master thesis requirements. This thesis is certainly among the best Master-level theses I reviewed in recent years. Hence I recommend to accept this thesis as Master thesis with grade A.

25 MAJ 2015

RNDr. Petr Somol, Ph.D.

Head of Research
Cisco Systems R&D
Karlovo namesti 10
120 00 Prague 2
psomol@cisco.com

Research Fellow
Inst. Of Information Theory and Automation
Czech Academy of Sciences
Pod vodarenskou vezi 4
182 08 Prague 8