Posudek vedoucího diplomové práce

Autor: **Ondřej Fikar**

Název práce: **P2P and Anonymity Network Detection**

Práca sa zaoberá problematikou detekcie TOR-u a anonymizačných sietí všeobecne za využitia agregovaných informácii o komunikácii v sieti vo forme proxy logov. Na túto úlohu študent zvolil kombináciu techník umelej inteligencie, s aj bez učiteľa.

V teoretickej časti diplomant podáva detailný popis využitých techník, ktore zároveň porovnáva s inými v praxi používanými metódami. Keďže vo svojom prístupe používa kombináciu techník strojového učenia, taktiež v teoretickej časti predkladá základy rôznych tém v danej oblasti.

V praktickej časti študent aplikuje vybrané metódy na problematiku detekciu užívateľov anonymizačnej siete. Obzvlášť praktické sa javí transformovanie tohto problému na problém ekvivalentný - detekciu TOR relay nodes, ktorý poskytuje lepšiu separáciu sieťovej aktivity TORu od zvyšku. Z experimentov vyplýva, že kombinácia učenia s učiteľom a bez učiteľa bola dobrá voľba.

Za istý nedostatok práce pokladám, že nediskutuje a neanalyzuje false positives použitého detečného algoritmu.

V priebehu spracovávania práce študent preukázal schopnosť analyzovať problém a vybrať efektívne riešenie. Vysoko hodnotím aj fakt, že práca je spracovaná v angličtine.

Predloženú diplomovú prácu hodnotím známkou: **A - výborně**.

16.1.2015

Ing. Jan Jusko
Fakulta elektrotechnická
České vysoké učení technické v Praze
Zikova 1903/4
166 36 Praha 6

Master Thesis Review

Thesis Author: Bc. Ondřej Fikar

Thesis Title: **Detection of P2P and anonymity networks**

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

The topic of the reviewed thesis aligns well with an important research area of the large-scale computer network security, as well as with industrial interest in the area of cloud administration. Network security currently offers important open problems of applied research type due to the rapidly growing size and importance of computer networks, especially with respect to the upcoming phenomena like Internet of Things and Internet of Everything, that are inevitably going to change our lives in the next couple of years.

Network security as a field is well established, yet seems to be generally delayed behind the latest developments listed above. The ever greater complexity of network traffic, growing variety of networking, application and security protocols, as well as the growing danger of network attacks puts strain on traditional security solutions, especially those relying on regular signature updates, often maintained manually by large teams of specialists. Relying on manpower never scales, especially in context of high-end computer infrastructure. Automated methods, often based on machine learning approaches, are generally considered the only feasible way to the future.

The author chose as topic of his research the detection of Tor traffic and/or peer-to-peer infrastructure, based solely on observations of proxy logs. The Tor service, though not malicious in itself, is typically considered a security threat at least in business environment (though on individual user level it may have a positive role if used, e.g., for safe communication among people in strongly oppressive countries where freedom of thought is punished). In business environment Tor usage can with high confidence be expected to mark intellectual property theft or other breach of security.

In order to build a robust detector the author correctly chose the hard way of relying on raw proxy logs only, without external information on the state of Tor infrastructure that could be taken into account, but at the cost of dependence on unreliable external information source. Building a detection system capable of transparent operation in arbitrary business network is the right choice.

As part of introduction the thesis provides good overview over network traffic anonymisation systems beyond Tor, discusses several examples of general classification tools and graph clustering methods. A two-layer Tor traffic detection system is then proposed, followed by experimental evaluation, discussion of related work and conclusion.

The quality of English is high, despite occasional typos and minor mistakes. In this respect the thesis is more mature than many other that I reviewed in last two years.

The structure of the thesis is generally good. It would be better though to make it easier for the reader to grasp the overall context sooner; when reading sections 3 and 4 one wonders what their purpose is. It takes to read through to section 5 where the purpose of the preceding chapters gets clarified.

I have a couple of comments from scientific practice point of view; some of them address found weaknesses. Nevertheless, I should stress right away that I am listing these more like suggestions for further scientific work than as critique of this text given its Master Thesis level:

- Opting for a two-layer classification system with the first classifier used for pre-filtering and the second for classification on the pre-filtered data is a good solution. Regarding the particular choice of tools for each layer there is certainly space for other options; a discussion on this issue would be welcome. If the text were to be formulated as journal paper, the reviewers would probably ask for more thorough reasoning behind the choice of classification tools. A deeper evaluation and comparison of the properties of logistic regression, threshold-shifting SVMs as well as other tools usable in this context would lead to a more defendable choice of the final setup
- In page 28 the sentence "After the reduction we ended up with histograms for 1.4 million servers." assumes the reader knows what histograms are meant. The meaning of histograms, however, becomes clear only later in text.
- Some of the preliminary methods described in sections 3 and 4 seem not to be needed later
- Last paragraph in section 5.2.1 mentions that more features have been considered beyond those in final use. The explanation which extra features proved to be useless would be worth extending to give more detail, including discussion what were the probably reasons of the problems.
- When reporting classifier performance, it is usual to describe the performance estimation procedure in detail. It is not clear here what was the training and testing set; more specifically, whether testing was done on the same data as training. If so, discussing the implications as well as the estimation bias incurred by that would improve the work
- The studied problem falls into the large group of contemporary problems being inherently hard due to the difficult properties of data available for learning. In this case the number of labeled samples is very small when compared to the size of data, with no feasible possibility to get large enough number of reliable samples. The available labels are not necessarily precise in all cases, and they may vary in time. The problem is also heavily imbalanced. I understand that deeper discussion of these issues would be out of scope of this thesis. Yet should the author intend to continue this research, it would be good to investigate more in this direction.
- Would it make sense to include FPs in Figure 5?
- The threshold value in Edge pruning (section 5.3.2) is obtained in ad-hoc manner. This leads to a good working result. Better way, however, may exist and can be worth investigating. Is it possible to describe formally what would be the optimum here ?

- It is not clear how sensitive the method is to a different setting of parameters summarized in Table 3. If these parameters are set manually, a deeper discussion of recommended and/or unusable values would solidify the reader's understanding of the method.
- It would be interesting to have a discussion on whether there are options to evade the proposed detection system. A related question is: what are the properties of the misclassified cases, what makes them difficult to classify. This is partly discussed in the experimental section but can be followed deeper.
- The author compares the resulting proposed method to a baseline in form of standard general SVM. Comparison to prior art on the same data is good practice, which can, however, be made better by including also comparison to the best known prior-art method defined to solve the same specific problem. "Best known" might mean "best found in external sources with reasonable effort".
- Related work is more common to be discussed before the description of the novel material, so that the novel solution can be reflected in context of the related work throughout the text as well as in the experiments.

Regardless the comments above I should stress that the text shows the author's good understanding of every discussed concept. I could not discover any mistake in formulas nor in other non-trivial descriptions. Large part of the material discusses prior art and sources are properly and frequently cited, this is also positive. There is space for improvement in terms of the overall command of context and reasoning presentation – consecutive sections should ideally build up the reader's understanding and gradually add more ground for the claims; the thesis as it is opens some minor additional questions that are left unanswered.

The key substance, however, makes good sense, is non-trivial, contains author's own innovation and is correct in all key claims. The extent of the thesis is adequate for a Master Thesis.

I have no doubt that this thesis should be accepted as Master Thesis. I give the thesis the overall rating of Very Good. I believe the author has good potential to become proficient in research work, especially as he gains wider experience in finding the right way through the endless landscapes of various prior art. I recommend the author to follow his studies as a doctoral student.

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