REVIEWER'S OPINION OF FINAL THESIS

I. IDENTIFICATION DATA

Thesis name:	**Online Advertising Fraud Detection Via Network Traffic Monitoring**
Author's name:	**Lada Ondrackova**
Type of thesis :	Master
Faculty/Institute:	Faculty of Electrical Engineering
Department:	Department of Computer Science
Thesis reviewer:	Sebastian Garcia
Reviewer's department:	Department of Computer Science

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **Extraordinarily challenging**

Evaluation of thesis difficulty of assignment.

The analysis of advertising network in order to detect fraudulent activities is a very hard problem to tackle. This is due to the combined behavior of advertisement that is used both in normal settings and in infected computers. Additionally, Anomaly Detector methods are hard to build and make work in a real environment.

Satisfaction of assignment **fulfilled with minor objections**

Assess that handed thesis meets assignment. Present points of assignment that fell short or were extended. Try to assess importance, impact or cause of each shortcoming.

The handed thesis meets the assignment. The thesis fells short in two accounts: First, obtaining a more accurate understanding of real adware traffic, including labels. Second, obtaining a better anomaly detection method. These shortcoming do not cause a great impact since there is some analysis of labels and captures of adware and also there is some work on anomaly detection that shows how the technique works.

Method of conception **correct**

Assess that student has chosen correct approach or solution methods.

The student has opted for starting to understand first how do the advertisement networks work. This is crucial for understanding the topic. The solution to the problem are built from here.

Technical level	**good**
Assess level of thesis specialty, use of knowledge gained by study and by expert literature, use of sources and data gained by experience.	
The thesis is very specialized since it studies a unique type of malware with a unique type of detector. The student also seem to have understood the inner workings of advertisement networks and the management of large amount of data.	

Formal and language level, scope of thesis	**very good**
Assess correctness of usage of formal notation. Assess typographical and language arrangement of thesis.	
The formal notation was good. The English of the thesis should be improved.	

Selection of sources, citation correctness	**very good**
Present your opinion to student's activity when obtaining and using study materials for thesis creation. Characterize selection of sources. Assess that student used all relevant sources. Verify that all used elements are correctly distinguished from own results and thoughts. Assess that citation ethics has not been breached and that all bibliographic citations are complete and in accordance with citation convention and standards.	
The student cited and analyzed the correct materials for this thesis. The sources on this problem are scarce and the student work with them correctly.	

Additional commentary and evaluation
Present your opinion to achieved primary goals of thesis, e.g. level of theoretical results, level and functionality of technical or software conception, publication performance, experimental dexterity etc.

Section 1: Introduction

The introduction is good. There are some minor corrections to be made:

- Explaining what a *fake impression* is (its only explained later, therefore here is meaningless)
- The problematic distinction of malware and adware should be done better for the reader.
- Better explain that *infection by plugin* is not *infection by popular web pages*.
- Malvertising should be explained if used in this section.
- Some subsections are just too small (e.g. 1.3), in those cases it would be better maybe to merge the text with other section.
- Although I like the separation of the goals of the thesis in the introduction, I think there was a confusion between the real goals to achieve and the different steps involved in the process. A confusion between the 'what' with the 'how'. I would say that the goal seems to be "To improve the detection of fraudulent advertising by using HTTP network logs.". And to fulfil the goal you have to do all the inner steps.
- What I miss in the Introduction are a list of contributions of the thesis.
- The main issue concerning the lacking of a clear goal is that I can not evaluate how each part of the thesis impacts the goal. The steps required are not the same if the goal is to *model* adware traffic,

than if the goal is to *detect* it, or if the detection is to set it apart from *non-advertising* HTTP traffic, or from advertising HTTP traffic. These different goals would give a different answer to the question: Is detecting normal advertising an error?
- The goal is however well defined in Section 6. I would propose to copy it here.

Section 2: State of the Art

The state of the art covered the basis papers that analyze adware and advertising detection. I would suggest to also add some papers on the use of HTTP logs for network detection in order to express how difficult it is to do it, and probably a reference to "Outside the Closed World" (Paxson, Sommers) to emphasise the important need of good datasets and labels.

Section 3: Online-advertisement

This chapter is very good and necessary for understanding how advertisement works. It covers the basic usages and is clear. It shows the reader what we are dealing with.

Section 4: Analysis of advertisement loading process
This chapter introduces the methodology used by the student to obtain the data necessary for the experimentations. In particular were the data should be collected and how it should be analyzed to obtain features. The first analysis was done with real advertising sites. My only concern here is about the fact that no real adware traffic was used for this analysis. Although its true that adware traffic is expected to behave very similarly to advertising, we won't know if adware implements new techniques until we take a loot at the traffic. One of the conclusions of this section is that the URLs of advertising can not be generalized. However, since the final goal is not completely clear to me, I can not say if this is good or bad, or it doesn't affect the experiments.
I would add in this section that the ads servers used by real advertisement are supposed to be the same as the adware.

Section 5: Data Explanation

This is an explanation of the datasets used. My main concern is that its not clear which label was assigned to the Cisco's users traffic. Although the traffic is good and vast, it would be good to be clear that this traffic is unknown and that without labels the use of the dataset is limited. It would be also good to say that some labels are already present in the dataset by means of the reputation tools used by Cisco. This is a very good advantage since most real Advertising Networks are known and Cisco may be capable of detecting them.
Something more that I missed in this section is a better explanation of the features of the HTTP and HTTPS traffic generated by Cisco. This would give me a better idea of what is, for example, the *status* of HTTP, since HTTP does not have a status normally. The same with the blocked tag: can we use this somehow?
The most important information lacking in this section is whether the Man-in-the-middle HTTPS traffic was included or not in the dataset. This is crucial.
Other information that is lacking are a summary of the amounts in the dataset. Not only users, but HTTPs flows, how many were blocked, etc.

The other concern with this section is that Figure 11 is not accurate. The Inspection of HTTPs traffic does not just *look* at the traffic going through. The inspecting device creates two different channels with different encryption keys, impersonating the destination web server. The process is more complex and is not well summarized in the Figure.

Section 6: Advertisement traffic detection

This section deals with the analysis of how real advertising looks like and how to detect it.
My main concern is that its not clear if the student is now trying to separate advertising from normal non-advertising traffic, or if its trying to separate advertising from adware.
The problem is that all the adblockers lists to block advertising may also detect adware. Because adware is not about using new servers or new URLs, its about *how* they are used.
So I would consider any blacklist or whitelist to separate non-advertising related traffic, with advertising AND adware.
The problem of not having a complete set of labels means that some conclusions are difficult to support. Like that the recall of ad-blockers is low. In here its probably a good idea to separate ad-servers of the publisher from the agency, etc. Following the terminology of Section 5.
The section says that YoYo rules can work well on HTTPs traffic. However, it should be better analyzed because not all HTTPs requests use the SNI for the hostname.

The motivation for 6.2 is unclear. Although its true that an algorithm that can *expand* the knowledge is good, its not stated that this is done because of the lack of labels.
In 6.2.1 there is a threshold that is being used for the amount of hostnames, but its not clear how the value of 100 was defined. Maybe expert knowledge. The same happened in 6.3 with the 15 domain names taken from the back propagation algorithm. The thresholds should be more clearly stated.
Its also not clear how the original set of patterns were found for the first round of the propagation algorithm. Was it from the adblock list? Was it by expert knowledge? The use of googlesyndication is expert knowledge.

The idea of the backpropagation algorithms is good, but I'm still concern about how much coverage it will have. Again the recall is very difficult or impossible to measure without labels. So actually we don't know how much the recall is compared to the simple block lists.

To summarise:

- The propagation algorithm finds hostnames in URLs that are referred from a list of known advertisers URLs. The original list of advertisement URLs come from a known list of ads URLs.
- The back-propagation algorithm finds new domains from URLs that refer to the same hostnames found by the propagation algorithm.

The whitelisting algorithm is a good idea and seems to be able to help in recognising adservers. However it uses a threshold of 10 visited times that is not defined. We assume expert knowledge.

Why this algorithms is not run multiple times until convergence? Due to the error?
It should be also be noted that this method may have some bias toward closed-networks of ads servers, it its possible that it can not find other unrelated ad networks.
Also, by not setting or training any threshold for the AD feature (the ratio) and by assuming that adware users would have a score of 1, you are effectively setting the anomaly score threshold to 1.

Section 7: Evaluation of the advertisement traffic detection

This section evaluates the results of the algorithm in section 6 and compares it with the adblockers lists.
One of the concerns of this section is the manual evaluation of the results. Although its a good idea to use experts in the evaluation, I would expect some numerical results of the evaluation from the expert. How many of the hostnames found in the first round were not ad-servers? Can we estimate the error of the algorithm?

The evaluation says that the generated adserver hostnames are 'exclusively related to advertisement', but then it says that some non-ad servers may appear very rarely. This means that the error is not 0, and we need to know the exact value.

The comparison done in subsection 7.2 is misleading since the mining algorithm works with hostnames, but the rest of the lists work with domain names. Therefore the comparison should have been done only with domains. The huge difference in hostnames is not indicative of a real difference in coverage.

In every comparison it should be noted that the adblockers are not supposed to work with network flows. So their real coverage can be much more when working inside the browser. Without explicitly saying this, a bias is introduced.

Section 8: Network intrusion detection techniques

This section discusses network intrusion methods. One concern is that the description of the characteristics and problems is not applied to the problem being solved. For example, an analysis of the balance is not done in the current dataset.
I would suggest not to use subsections for paragraphs (8.3.2, 8.3.3, etc.)

Section 9: Advertisement fraud detection

This section describes the current CTA system and the new advertisement fraud AD system.
In 9.2 the system states that in a normal case the non-ad flows come before the ad-flows. This makes sense. However it later states that the **deviant** behavior is when a 'significant portion of ad flows appears after non-ad flows'. There is probably an error here, since the two phrases say the same: That non-ad flows come before ad-flows.
Assuming that there was an error and that the first deviant characteristic was 'that non-ad flows come after ad-flows', I would say that this doesn't make much sense since in a normal computer non-ad flows also come after ad-flows. The sequence of non-ad flows and ad flows is almost continuous.
The second characteristic of adware is stated as 'non-ad flows are missing between ad flows'. But how can this be measured? If there are missing non-ad flows, wouldn't all the ad flows look like a continuous stream of requests? So what would the anomaly be? That the stream of ad flows is too long?

The problem here is the none of these 2 features are later used as part of the AD system. Which uses a ratio of ads to non-ads flows. So I'm not sure why they are here.

I would have liked that these characteristics of anomalous traffic would be based on one of two things:
1) Or only modelling the normal traffic and searching for anomalies
2) Or modelling the AD system based on real observed adware

It seems that the 2) approach is used, but there is no explanation of why these specific characteristics.

The goal of the system as described in 9.2 is confusing. It says that the goal is to detect infected users. However, the previous goal of the system was to detect fraud ads in HTTP logs. These goals are not the same.

Later on a ratio between total flows and ad flows is proposed. This is a good ratio to work in AD. However, its sated that no more than 50% of the traffic should be ads. Why? What is the rationale behind that no more than 50% should be ads? I can image several situations where a normal, non-infected user, generates more ads than normal traffic. And actually Figure 20 and Figure 21 show that there are a large amount of users, supposedly normal, that have more than 50% of this ratio.

Ia agree that labels would help in evaluating the results, however, labels are not required to study the behavior of the ratio on every user. The assumption that a normal user would not have more than 50% and that adware would have close to 100% can be easily verified by looking at all the users. Although not perfect, its better.

Another idea to solve the problem of HTTPS would be to resolve the hostnames to IP addresses per day, and then filter the HTTPs flows per IP address. This would not be perfect but it would give a very very good estimation of ad hostnames being used in HTTPs for the purpose of training the AD algorithm. Once the AD algorithm is trained you don't need so many of them.

More importantly, you need to go and see how the adware looks like to understand the properties and propose features. All these assumptions can not hold in reality since the user is still doing normal actions while being infected.

Also the AD algorithm needs at least 4 days of data and at least 10k flows to detect something, which is an important restriction for most users.

There is a small confusion between histograms and distributions. The ones shown in Table 20 and 21 do not sum up to 1, so they are histograms probably. But the principle is similar.

The normalisation process is rather confusing also. First a linear transformation is proposed, that consists of the classic normalisation process of taking out the min value and dividing by the range. This is ok, but it turns out that this normalisation is not used later. Then a Gaussian Scaling Method is proposed and used. However this method is not explained.

The proposed AD algorithm is difficult to grasp. The summary is that for each user (all their traffic) a ratio of ad flows to non-ad flows is computed. This is the real anomaly score of feature used.
Later some transformations are applied, but the AD feature is the same. The problem is that this is a feature, but its not a *method* yet.

Some analysis after the transformation is also confusing. The thesis states that the majority of the users have a score of 0. However this is not true, since the majority of users are in the middle of the distribution in Figure 23. If you sum up the scores there are more users with score >0 and <0.5 than with score 0. Also there are more users with score >0.5 and <1 than with score 0.

My main concern for this section is that no AD system is proposed. Only a feature. But there is no training or adjustment of what is an anomaly. This could have been approximated even without a dataset. Some users could have been analyzed by hand.

Section 10: Evaluation of advertisement fraud anomaly detector

This section describes the evaluation of the AD
The evaluation was done in some of the datasets of Stratosphere. The main problem of using these datasets is that they do not have normal traffic at the same time, making them useful only for certain actions. So the datasets could be used to learn about how aware really works and define features, or they can be used mixed with only normal captures.
However, the error shown in the thesis for these captures are that the score is very low. This doesn't not make sense if we are evaluating in an all-ad capture. We would expect to have a score of 1. I believe that the problem may be that the hostnames found by the mining algorithm do **not** appear in the Stratosphere capture. Looking at one of the captures (177-1) I can see that most requests are related to Chinese ad servers. Maybe the bias of the locality of the users make the to ad networks separate.
However the thesis says that the mining algorithm **do** detect the ads hostnames in the stratosphere captures, which then raises the question of why the score is not close to 1. If indeed the mining algorithm finds most of the ad server hostnames in the datasets and if the score is not one, then we must conclude that the assumption that the score would be close to 1 in adware was not correct, at least for these adware in the capture.
In this case it may be worth analyzing by hand real captures of different types of adware.

Also the assumption that the datasets do not contain enough ad flows because there was no user interacting is hard to maintain, because if a user is interacting there will be also more non-ad flows, moving the ratio down.

Some definitions are missing, such as what are blacklisted users in subsection 10.1.2.

Later on there is an evaluation with CTA users. This is a good comparison since CTA has a list of blacklisted domains and adwares. Although it may not be complete it should sample part of the adware behavior.
Even when the evaluation analysis was done with the CTA users it was found that infected users do not communicate much with popular ad-servers. This is a good finding in the thesis, but it marks that its hard to maintain the original assumption that adware-infected users generate more ad flows than normal. It also marks the need of a better analysis, since CTA is not focusing on adware so much, meaning that

Subsection 10.2 presents a manual evaluation of the top 0.1% of the most anomalous users. But again there is no indication of any training of the thresholds, so why 0.1%?

The methodology for manually analysis is not clear.
First, any analytic method introduces automation, which means that is based on features and therefore in a mental model of what you are looking for. When doing manual analysis, any automation should be avoided because it adds more errors and bias. Now we should evaluate if the analytic method was good enough, with good coverage, etc.
Second, the description of what its being looked is confusing (subsection 10.2.1). For each user, each hostname is evaluated, then the distribution of hostnames is evaluated if it contains normal user behavior. However, there is no description how this is done, by hand? Using the Internet?. Finally the methodology extracts a C&C server, but there is no description of what is a C&C server, how to recognize it or even why it was not in the original list of features for the detector.

There was also a confusion about if the method is evaluating flows or users. The thesis goes from one to the other.

Section 11: Future Work

I believe more future work can be done. Starting from studying more real adwares captures, training the AD feature to crate some method, etc.

Section 12: Conclusion
I would make it a conclusion of what happened in the thesis as a concept, not of the thesis as a paper. In summary we want to happen with the proposal, not a summary of the Sections (that is for the introduction).

III. OVERALL EVALUATION, QUESTIONS FOR DEFENSE, CLASSIFICATION SUGGESTION

Summarize thesis aspects that swayed your final evaluation. Please present apt questions which student should answer during defense.

This thesis studies a very complex and difficult problem: the use of advertising networks and their use in fraudulent activities. This is a hard problem that is usually covered in larger research. Therefore the research done by the student is good for a master thesis, presenting good results and a good base for future work. The main comments are that: (1) real adware malware should have been analyzed more in order to better understand the real word of fraud, and (2) that the anomaly score proposed should have been adjusted in a better way.

I believe that the work was good. The main issue may be that the expectations were set too high. The evaluation and comparison should have been done according to the standards in the CTA system, which knows that its not possible to find all the anomalous flows completely. I would have put the requirements lower in order to contribute to CTA without trying to completely solve the AD problem.

The main contribution of the thesis is a novel and very useful algorithm for finding all the advertising networks, which can have a huge impact in the security community. Furthermore, I would suggest to keep improving the AD method.

Questions for the student:

1. How does adware differentiate from normal advertising in the network? In terms of structure of the communications, sequence of connections and referrers used.
2. Which can be the 3 most useful applications in the security community of your mining algorithm considering the implications of such attacks as Exploit kits and web cryptocurrency mining?
3. How would you train your anomaly score based on incomplete labels as those in CTA? Consider that you don't want a perfect detection and that other algorithms may already catch some related anomalies from the same users.
4. If we define an ad-network *bubble* as a group of ad-related infrastructure servers that does not have any connection to other ad-networks (separated for example, by language). How would the existence of these bubbles affect your mining algorithm?

I evaluate handed thesis with classification grade **B**

Date: **2018, June 9th.** Signature: