Analysing the Forum Activities Inside the Dark Web

S. Hürol Türen, Rafiqul Islam , Kenneth Eustace
School of Computing, Mathematics and Engineering - Charles Sturt University Australia

Project Overview

The dark web is an encrypted subset of the deep web, whose content cannot be indexed by search engines. Dark web pages can be accessed from private networks such as TOR (The Onion Routing), I2P (Invisible Internet Project), Freenet. TOR is widely used by the dark web users in a domain defined by .onion extension. Dark web users can communicate with each other without using their identifications. However, the anonymity of the users encourages them to perform illegal activities.

Objectives

OBJECTIVES

1. To investigate and identify imminent criminal threats and the algorithms, techniques and tools used to protect from data loss and mitigate the risks from attacks inside the Dark Web;
2. To use patterns in the data with (WEKA or Python) Machine Learning algorithms and other Data Science, Machine Learning (ML) and Model Building strategies for training and development of a model.

Problems/Hypothesis

PROBLEM 1. The timely and pre-emptive detection of dark web forum activities before the dark web actor(s) can put their threats into action.

PROBLEM 2. The accuracy of the attacks on Tor network and use of IoT and Streaming technologies, requiring new algorithms to monitor the forums and to attack data on Tor networks.

HYPOTHESIS

Can the findings of this research on detecting illegal forum activities, be used to develop an effective Cyber Security Forensics Maturity Model for continuous monitoring and prevention?

Key Research Question

How is it possible to detect illegal forum activities and attack from a Dark Web TOR user, before the threat is put into action?

Methodology

- This project uses quantitative research methods such as content analysis with an experimental research design.
- A group the techniques and methods which are applied for monitoring the dark web forums and which are used for the attacks on Tor network.
- Then it will be possible to define the gaps of these methods according to their usage. After the gap analysis of the previous research methods, it is possible to try to extend these algorithms or the data which the techniques are applied to, to see if the gaps are closed.
- The experimental research methodology leads to a Model evaluation in machine learning testing from the data patterns.

Procedure

Step 1

Review of existing techniques

Step 2

Define case studies

Step 3

Data collection for the case study

Step 4

Compare results of different techniques on the same case study

Data / Observations

The observations made from the Data Collection and from existing research elsewhere follows a sequence as described below:

- Other external data resources for the selected case studies such as Kegel
- Apply the exiting Data Science and ML techniques for the recent research case studies
- Perform Gap analysis after the literature review
- Testing or Creating algorithm(s) depending on the gap analysis
- Train and test data with the algorithm e.g. random forest
- Analyse the results and iterate this process for final model building

Gap Analysis

- **(Nicolas Ferry, 2019)** has applied his semantic analysis with an external open-source framework (Alyze Framework) That's why it is better to analyse his research through other frameworks to comply his research. Another gap is, that this research group tried to do the semantic analysis locally. They have grabbed an .onion site once, saved locally and then they have run their semantic analysis modelling. I have not read any approach with streaming data.
- **(Hussein Alnabulsi, 2018)** had only Tor Network in his research. The other networks such as ISP, Freenet should also be applied for the related work. The developed algorithm (Vector Space Model) is based on TF / IDF. The most difficult problem is the text frequency. Mostly the people cyphered their text and voice communication and avoid using general words, which can be caught easily by adversaries through text and voice recognition. For example, they can use beans instead of bullets or munition. That could be a bit misleading.
- **(Matthias Schäfer, 2019)** has developed BlackWidow with a half-automated system. The first step of the lifecycle (planning and requirements) is manual. The language translation is made by Google translator. I would use rather other online translator APIs such as deepl. It is always better to use different APIs for the same purpose to see the accuracy of the work. This research is also done with local data structure. Streaming was out of interest.

Conclusion

With the large number of illegal forum activities happening inside the dark web, the main aim of this topic is to make it easier and faster to extend the previous machine learning algorithms which are applied to detect and monitor forum activities or gather the web fingerprints of the illegal web sites which are visited frequently by dark web users.

The algorithms which are included in this paper are the most updated and relevant ones to detect and extract data from dark web forums. Unfortunately, the hackers are also trying their best to not to be caught by the governments. That is why research is ongoing beyond this project scope.

Works Cited

- Nicolas Ferry, T. H., Francine Herrmann, Alexandre Tourette. (2019). *Methodology of dark web monitoring* 11th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), Pitesti, Romania.
- *Hussein Alnabulsi, R. I. (2018). Identification of Illegal Forum Activities Inside the Dark Net International Conference on Machine Learning and Data Engineering (iCMLDE), Sydney, Australia.*
- Matthias Schäfer, M. F., Martin Strohmeier, Markus Engel, Marc Liechti, Vincent Lenders. (2019). BlackWidow: Monitoring the Dark Web for Cyber Security Information 11th International Conference on Cyber Conflict (CyCon), Tallinn, Estonia