Visualizing Cyber-Threats in Underground Forums

DECISION: accept

AUTHORS: James Burroughs, Michal Tereszkowski-Kaminski and Guillermo Suarez-Tangil

Summary of Reviews

- Review 1: 2 (4)
- Review 2: 0 (5)
- Review 3: 0 (3)

Reviews

Review 1

TOTAL SCORE: 2 (accept)
Overall evaluation: 2 (accept)
Reviewer's confidence: 4 (high)

The authors present an alternative approach to classifying posts through visualizations that convey semantic metadata of the posts. The key features represented include keywords/key phrases (red), contact info (cyan), URL/domains/IP addresses (blue), prices (pink), questions (yellow), CVE numbers (grey), code (green), and keywords based on custom dictionaries (dark red). They use data from a large Russian cybercrime forum and demonstrate how such a visualization can expedite the identification of a post in four categories: 1) debugging, 2) know-how, 3) trading, and 4) miscellaneous.

The technique showcased in the paper is innovative, and the paper is well-written and well-structured. I believe this has great merit and should be published in WACCO's proceedings.

As a reviewer, it is difficult to not provide further suggestions; thus, here are a few comments that the authors can consider enhancing the quality of their paper:

- The background and related work section does not adequately credit previous work on the matter. The authors could better situate their work within the literature and explain why their approach is innovative. Many previous studies (see below) have developed techniques to classify posts, and they should not be overlooked.

- The authors might consider applying the same method to an actor's entire corpus. Through such visualization, one could quickly determine if the actor is an info seeker, a seller, etc. This could be added to future research and considered by the authors as a subsequent research step.

- Who are the analysts used to classify the posts and test the efficiency of the methods? Do they have domain knowledge on the topic? The authors could provide more information about them.
Why were the four categories chosen (debugging, know-how, trading, and miscellaneous)? How are these categories helpful in understanding the relevance of a threat? The authors could better explain the rationale behind these categories.

- Sections D and F are missing the "threat level" score.

Examples of studies:
Caines, A., Pastrana, S., Hutchings, A., & Buttery, P. J. (2018). Automatically identifying the function and intent of posts in underground forums. Crime Science, 7(1), 19.

Huang, C., Guo, Y., Guo, W., & Li, Y. (2021). HackerRank: Identifying key hackers in underground forums. International Journal of Distributed Sensor Networks, 17(5), 15501477211015145.

Sun, Z., Rubio-Medrano, C. E., Zhao, Z., Bao, T., Doupé, A., & Ahn, G. J. (2019, March). Understanding and predicting private interactions in underground forums. In Proceedings of the Ninth ACM Conference on Data and Application Security and Privacy (pp. 303-314).

Review 2

TOTAL SCORE: 0 (borderline paper)
Overall evaluation: 0 (borderline paper)
Reviewer's confidence: 5 (expert)

The paper proposes a visualization method to represent posts in a darkweb forum. It creates multi-colored pictograms corresponding to keywords, questions, code etc., while the density represents post length. The authors claim that it improves analyst's processing speed.

I do like an idea of addition a visual dimension to threat analysis. Going beyond text could be helpful in a variety of applications. However, I do have two groups of concerns. First, I think the authors didn't provide a sufficient argument for the usability of their method in attacker detection space. Second, I do not think there is sufficient novelty in the underlying data processing. Specifically,

1. I think the authors over-sell a bit regarding the effectiveness of the method. They claim speed improvement with "tolerable decrease in accuracy" - while in fact, accuracy drops from 100% to 81%-72% for both analysts. This is a huge drop in performance - now we are missing one in five, or even one in four threats!

2. Is speed that relevant in an underground forum detection? It is by definition an asynchronous (non real-life) resource. Shaving a few seconds of an analyst time is not as critical as in live detection.

3. Keywords that the system visualizes need to be provided in advance. The system has no capability to detect evolving threats.

4. There is no novelty in text pre-processing; it is relying on existing algorithms. Further, it is not clear how multi-lingual nature of posts is addressed here.

5. The classification applied (links, code, prices, keywords, questions...) is too generic. Why is it relevant to cybersecurity? On the upside, I see that this method can be useful in other text-visualizing applications.

6. The classification task presented to the analysts is also generic and not directly related to cybersecurity (debugging, know-how, trading, misc.) Might as well represent any IT-related discussion board.

Review 3

TOTAL SCORE: 0 (borderline paper)
Overall evaluation: 0 (borderline paper)
Reviewer's confidence: 3 (medium)

This paper presents an analysis tool visualising cyber threats in posts on an Russian underground forum, and is tested on experts for validation.

Strengths

+ Welcome contribution to the body-of-work analysing underground forums
+ Well-structured and readable paper, that can add value to manual analysis of underground forums by analysts

Weaknesses

+ Unclear methodology to validate the visualisation of post on expert community
+ Little scientific novelty

Comments for the authors

The paper in its currents forms lacks a solid empirical basis. Both parts of the work - the visualisation design, and the testing thereof on the intended population - are shallow. In essence, the authors present a visualisation tool based on existing box-counting dimension techniques, which has little novelty in itself.
My most prominent concern however is, the lack of an elaborate evaluation section. How is the evaluation performed? With which conditions? How many experts were recruited and how? What were the parameters that were used to evaluate the tool against manual analysis? And if more than one analyst was included in the evaluation, how comparable are their respective tasks and internal procedures?