[32] *The Impact of Adverse Events in Darknet Markets: an Anomaly Detection Approach*

Ziauddin Ursani (University of Bristol), Claudia Peersman (University of Bristol), Matthew Edwards (University of Bristol), Chao Chen (University of Bristol) and Awais Rashid (University of Bristol).

Reviews Phase 1

Review 1

Time: Jun 10, 04:24

Overall evaluation: -1 (weak reject)

This paper proposes to apply anomaly detection on darknet market (DNM) datasets to find adverse events. The authors use an unsupervised evolutionary algorithm to perform such detection on 35 DNM websites and find adverse events corresponding to site shut-downs and stat-ups, and law-enforcement interventions.

=== Strengths ===

I think the idea is interesting.

It did take me a while to understand the goal of the paper (the introduction should be improved). The reason is that most known events such as site shut-downs or law-enforcement actions have clear timestamps. Researchers can simply focus on data following the event timestamps to analyze the impact of the known events.

After reading Section 2, it becomes clear that the paper's goal is to find adverse events that might have gone under the radar, e.g., some previous unknown events or cascade effects caused by known interventions. To this end, the idea makes sense.

The proposed technique is simply to run and is likely to be useful.

=== Weaknesses ===

While I think the idea has potential, the evaluation/analysis part did not fully illustrate the value of the idea. For example, the evaluation has been focused on known events (it is good to know the method can detect them). The authors did not provide much analysis on events that were previously unknown (i.e., detected by this proposed method for the first time).

The true value of anomaly detection is to find such previous-unknown events, perform analysis, and provide new insights. Such analysis is largely lacking.

I believe some of the claims are too strong. For example, “Such cascade impacts can be modelled and utilised by law enforcement actions to bring about maximum disruption to DNMs.” --- from the analysis in the paper, I fail to see how their results inform the law enforcement to change their strategies to maximize the interruption.

Minor issue:

a broken reference: "Therefore, we decided to use all the features described in section ?? and their absolute weighted sum as a model to represent the amount of activity."

Reviewer's confidence: 4 (high)

Review 2

Time: Jun 11, 23:42

Overall evaluation: 2 (accept)

Summary -------

The authors leverage an ML-based method that uses unsupervised learning to find anomalies in the posts that are made on DNM forums. To this end, they combine various features, such as the number of coding/hacking terms, attachments, quotations, etc. The method is the executed on the 35 DNM communities for which data was available. The anomalies are detected for both specific user groups and the entire community and compared to the adverse events that occurred during the time. This shows that mainly the major averse events had an impact on the DNM communities.

Strengths ---------

- I really like the idea of using anomaly detection to study the averse effects in DNMs - The analysis is well-executed, and well-described - A diverse set of features is used, most of which have been shown to be effective in prior work - The anomalies are nicely visualized, and the anomalies detected in several DNMs do coincide with major averse effects

Weaknesses ----------

- The granularity is based on one calendar month; it would've been interesting to see if this approach would also work with a time unit of e.g. one week, which could give more granular insights in the anomalies - It was not entirely clear what extent of the detected anomalies could be attributed to the natural evolution of a DNM

Comments --------

- I would be interesting to have more insights in the weights of the different features that were optimal for the model

Reviewer's confidence: 3 (medium)

Rebuttal Letter

We thank the reviewers for their constructive comments. Below we clarify our responses and changes we will make to address the major comments. We will also make the minor corrections in the final version.

Reviewer 1:

We will bring some of the description from section 2 to section 1, so that the purpose of the paper is clearer in the Introduction.

We will clarify that we did not start with the known major events a-priori. We identified the anomalies and then validated them against known major events. Furthermore, the minor events listed in Table-IV were identified through our anomaly detection approach and we undertook further investigations and analysis to identify the events underpinning these anomalies. We should have marked them clearly with symbols E6-E9 in the table. The graph in figure-2 shows these events but as these were not marked in the table-IV, the reader completely misses them in Figure-2. Therefore, we will make the following modifications.

1. Clarify that events weren’t identified a-priori.
2. Mark the major and minor events in Table-III and Table IV with symbols (E1-E9)

3. Extend analysis of graph in Figure-2 to include minor events to fully illustrate value of the idea.

Regarding claims about use for law enforcement, our discussions with law enforcement have highlighted that there is interest in finding more efficient and effective ways of disrupting DNMs - taking down any market is much more resource and time consuming (and there is not enough evidence that this leads to long term disruption). Table-V shows the extra-ordinary increase in new memberships as compared against long term average of Black Market Reloaded. This caused panic among the administrators who saw this as intrusion from law enforcement, which caused them to shutdown the site voluntarily. This phenomenon can be observed in several DNMs. Emulation of such a phenomenon to trigger voluntary shutdowns can be more cost effective for law enforcement agencies. We will add further discussion to clarify this and add further evidence of the phenomenon across several DNMs.

Reviewer 2:

Regarding anomalies at the level of small-scale samples, our experience is that distinguishing the anomalies which are caused by adverse events and those that are mere false positives depends on their persistence over several weeks. This is the reason we chose one calendar month sample. It would be interesting to explore how the fidelity of the approach is impacted by smaller-scale time periods. We will note this as part of future work.

Natural evolution of DNM was modelled and anomalies were only considered which were beyond certain threshold limit away from that model (Section III.C, model learning). The linear graph in Figure 3, represents the model (natural evolution) while the data points can be seen how far they are from the model. The bar chart in figure-4 shows the distance of those data points from the model line in terms of standard error units. Any data point which crosses the limit of two standard units is considered anomalous. We will add a brief explanation to clarify this further.

We currently use a default set of parameters to undertake anomaly detection and do not tailor weights to specific DNMs. We can conduct analysis with respect to optimal weights on some of the major darknets such as Silk Road 1, Silk Road 2, Black Market Reloaded.

Reviews Phase 2

Review 3

Time: Jun 23, 12:43

Overall evaluation: -2 (reject)

The paper analyses 35 darknet markets using an anomaly detection approach to automatically identify anomalous behaviors in DNM users and in DNM as communities. The identified anomalies are then compared with external adverse events to find correlations. Overall the approach is novel, and further research in this area could potentially assist organizations to disrupt DNMs.

--General Comments-- The structure of the paper should be improved. More emphasis should be put on the dataset, the features, and how they were extracted, and the models used for the anomaly detection. Readers are left to guess important aspects of these stages, which are key to understand the work in full.

Many areas of the paper are not clearly explained, concepts and ideas are mixed with events, making it hard to read and follow.

--Detailed Comments--

1. Abstract. From reading the abstract is not clear in full what was done. Most of the abstract discusses the results, but not the approach.

2. Section III. The authors claim that there are no reference models for DNM because they are short-lived and volatile. The criteria of being 'short lived' is subjective, and some DNMs have remained active for 2-4 years. Many of these markets have many thousands of messages per day, so it's hard to understand this claim.

3. Section III C: This subsection is the core of the work and is very weak. The authors do not explain in detail the exact model used, the configurations, the data processing, and the algorithms used. Additionally, - The selection of the threshold limit is not explained. - The authors do not explain the criteria used to select a 'one month' time unit, when other finer time units could have worked maybe better. - It is not clear if all selected DNMs have English as their main forum language. - Authors do not explain how the features were extracted (techniques used). - Explanations of some features are unclear or not explained in detail (e.g.: Number of Active Users, Number of membership) - Configuration parameters are not explained (Fig 1). - The input information (Fig.1) is not explained. - The weights for the feature set are trained, however, no details on how it was done are presented.

4. Section IV: - Fig2, it is not clear what the colors of the dots represent. Would have been better to use different symbols instead of sizes of the circles to represent differences in the type of anomaly. - It is difficult to evaluate the results presented due to the shortcomings of Section III.

The idea presented in the paper seems promising and novel, however, the research lacks enough clarity to be fully understood and possibly reproduced.

Reviewer's confidence: 3 (medium)

Review 4

Time: Jun 27, 13:42

Overall evaluation: 1 (weak accept)

This paper presents a measurement study in which the authors apply anomaly detection to identify changes in activity in darknet markets (DNMs) happening in proximity with disruptive events like takedowns. The authors analyze a dataset from 35 DNMs and find several changes in activity, many corresponding to well known adverse events.

The effect of takedowns and other adverse events on DNMs is not well understood, and in particular their cascading effect on other markets. This paper presents an interesting approach to better understand these, and I think that it would be a nice addition to the workshop.

The authors use external lists to identify coding and hacking terms. How comprehensive are these? Are all the DNMs in the dataset mainly using English? If not, what are the potential biases that might arise here?

It would be interesting to analyze the anomalies that correspond to known adverse reactions in more detail. What is it that changed in the studied DNMs? Was it users migrating to them in reaction to a takedown happening on another market, or was it users stopping posting for fear of getting caught? This would add a lot to the results presented in the paper.

The figures are of very low quality. The authors should either significantly increase their resolution or plot them in pdf format to improve readibility.

Reviewer's confidence: 4 (high)

Review 5

Time: Jul 02, 16:50

Overall evaluation: 0 (borderline paper)

=== META REVIEW: REQUESTS FOR THE SHEPHERDING PROCESS ===

The reviewers agreed that the paper has potential for publication, but the following issues must be addressed before it is ready for publication:

1. Provide further clarifications on the methodology (Section III). Refer to detailed comments by Reviewer 3. 2. To show what the actual contribution is regarding anomaly detection, analyse some of the anom alous events found that were previously unknown (R1), and confirm these are not always the natural evolution of the DNM (R2). Clearly mark the minor events from Table IV on Figure 2, and explain how your method helped to spot these. Also, indicate what triggered the anomalies for some of the known adverse events (R4). 3. Improve the quality of the Figures. Use preferably PDF or PS formats (R4). Explain the colours of Figure 2 (R3).

Reviewer's confidence: 5 (expert)