

Tracking Discourse Influence in Darknet Forums

Submission of Team SamSepi0l to the AMoC Hackathon

Christopher Akiki, Lukas Gienapp, and Martin Potthast

Text Mining and Retrieval Group

Leipzig University

Abstract—This technical report documents our efforts in addressing the tasks set forth by the 2021 AMoC (Advanced Modelling of Cyber Criminal Careers) Hackathon. Our main contribution is a joint visualisation of semantic and temporal features, generating insight into the supplied data on darknet cybercrime through the aspects of novelty, transience, and resonance, which describe the potential impact a message might have on the overall discourse in darknet communities. All code and data produced by us as part of this hackathon is publicly available.¹

I. INTRODUCTION

The hackathon encompassed two separate tasks. The goal of the first task was to create an innovative approach to visualising the temporal nature of the dataset to allow for a longitudinal analysis of how significant events might affect the nature of messages exchanged on the forums. The second task seeks to perform authorship attribution to re-identify individuals' accounts across different forums.

Both tasks aim at gaining novel insight into financially-motivated cybercrime on darknet markets. The hackathon makes use of a subset of two datasets: the Darknet Market Archives [1] and the hacker forums of AZsecure.org [2]. The final dataset consists of 40 fora of the dark web. These fora typically serve as escrow spaces where buyers and sellers of illicit goods and services converge to conduct transactions.

II. METHODOLOGICAL APPROACH

This section details the methodological approaches and design decisions influencing our solutions to both of proposed tasks.

A. First Task

The underlying goal of the visualisation approach we chose for Task 1 is to make the relation between the content of messages posted on dark web forums and the time messages were posted there both visible and explorable to an end user. Therefore, the visualisation dashboard we created (see Figure 1) includes three distinct modes of visualisation. The first one is the temporal nature, represented by a simple timeline at the top, allowing users to browse the data by making time-based selections. The second is content, represented by the semantic space embedding to the left of the visualisations; here, posts are plotted by their position in the semantic space of their respective community. The third

component visualises the interaction effect between time and semantic space, plotting the three features novelty, transience, and resonance. For all three, we plot both the distribution, as well as the x-y interaction plot between them.

To calculate the position of messages in their communities' semantic space, we rely on the transformer-encoder-based variant of the Universal Sentence Encoder (USE) [3] to calculate phrase-level embeddings for the body text of all posts. The USE model consists of a transformer-encoder architecture very similar to BERT [4], albeit trained with two key differences: first through the use of the rule-based PBT tokenizer, and second through a more downstream-aware multi-task supervised pretraining regime.

The resulting vector space spanned by the 512-dimensional embeddings USE produces can be used to calculate the semantic similarity of texts. To make these high-dimensional semantic relations visible to the end user, we resort to manifold learning whereby we try to learn a 2-dimensional non-linear topological space that best approximates the data in low-dimensions. To that end, we experimented both with t-distributed Stochastic Neighbor Embedding (t-SNE) [5] and Uniform Manifold Approximation and Projection (UMAP) [6], [7], and ultimately chose t-SNE as it provided for a better visual result upon manual inspection. Manifold learning was performed separately per community as the final visualisation is centred around community-specific views.

Furthermore, we performed density-based clustering using the DBSCAN [8] algorithm to make different semantic groupings in the data more easily visible.

To estimate the interaction effect between time and semantic space, we expand upon the methods developed by Barron, Huang, Spang, *et al.* [9], originally meant to study a political body—that of the national assembly of revolutionary France—as a heteroglossic system that evolves through time within a bounded political context. We find the parallel between a longitudinal corpus of political speeches and a longitudinal corpus of forum posts structurally similar enough to warrant an adaptation of their methods. This approach boils down to computing three longitudinal vectors using a sliding window approach: novelty, which is quantified by the divergence of a document to its local past; transience, which is quantified by the divergence of a document to its local future; and resonance, which quantifies the difference of these two dynamic quantities, measuring their interplay. We calculate the three features novelty, transience and resonance to model the influence to

¹<https://github.com/webis-de/AMOC-21>

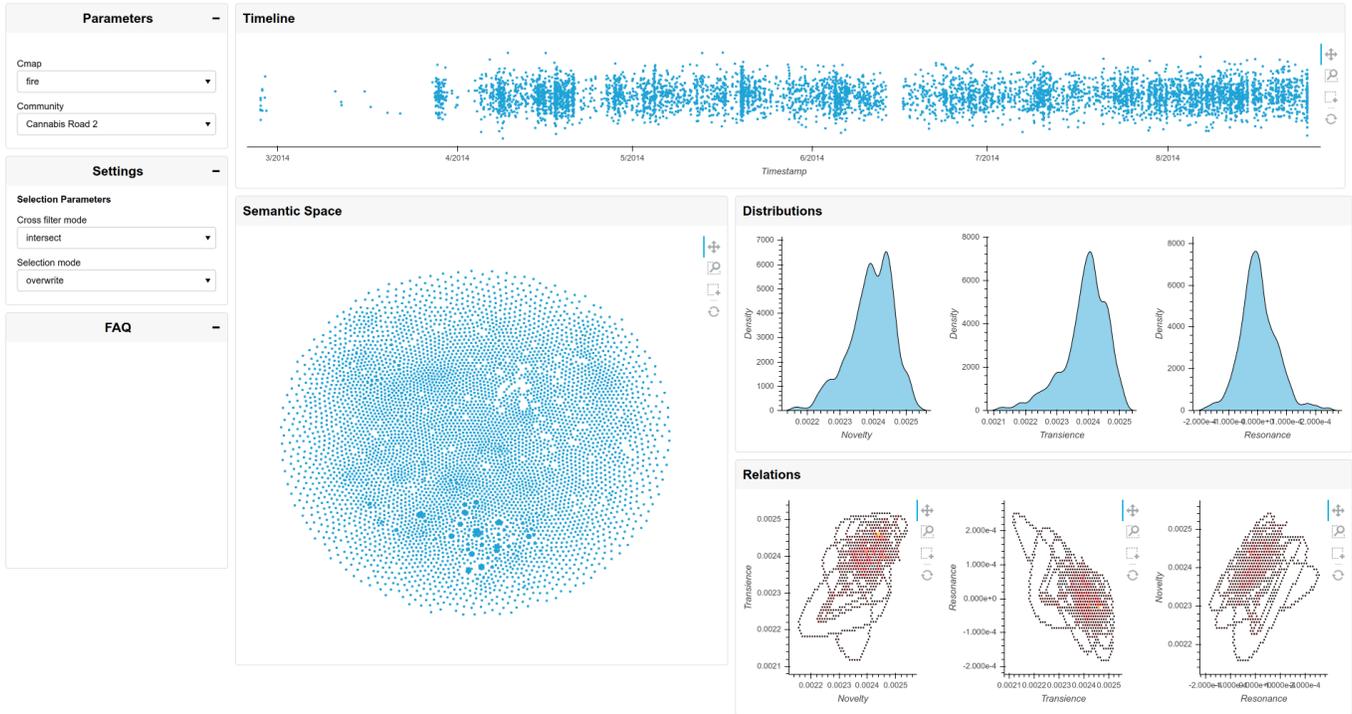


Fig. 1. Final Visualisation Dashboard

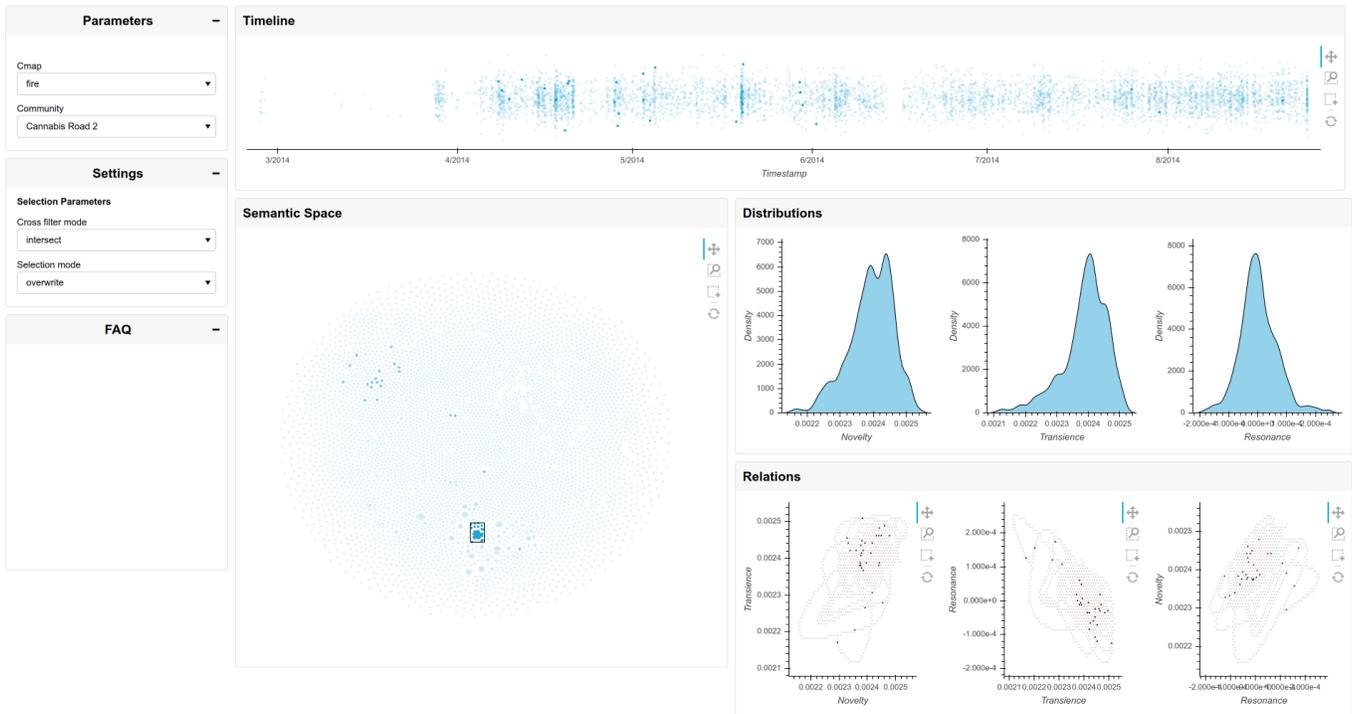


Fig. 2. Final Visualisation Dashboard with active selection

the communities discourse a single message has.

In the context of this task, a high novelty value could, for example, be used to identify messages that introduce a new product to a darknet market, while an additional low transience value might help identifying members that are highly influential on the overall discourse of platform, and are key community leaders. However, we refrain from making any further assumptions in this direction because we lack the domain-specific knowledge required to make a useful interpretation of the collected data.

The features are calculated in a sliding-window manner: here, for each post p_t at point in time t , its novelty is measured as the Kullback-Leibler divergence of a semantic probability distribution over p_t to the average distribution of all previous posts p_{t-1}, \dots, p_{t-n} in window of size n . For transience, the same method is applied, but comparing to all following posts p_{t+1}, \dots, p_{t+n} . Finally, resonance is measured as the asymmetric difference between novelty and transience. We infer a semantic probability distribution for each post p by applying a softmax function to the semantic embedding vector as produced by the USE (see previous section).

B. Second Task

We set about solving this task using the novel approach introduced by Sun, Schuster, and Shmatikov [10] and went as far as implementing it using TensorFlow and the Huggingface library [11]. This approach leverages the generation dynamics of causal language models, GPT-2 [12] in this instance, to compute a fingerprint for a given text. Upon deeper examination, it became clear to us that this method of fingerprinting text would be better suited for a side-channel scenario where one does not have access to the original text, but merely to the smart device upon which said text is generated.

As a fallback implementation, we started to apply the unmasking algorithm originally developed by Koppel and Schler [13] and refined for the domain of short texts by Bevendorff, Stein, Hagen, *et al.* [14]. However, due to the short time frame of the hackathon and the time “lost” on the first approach, we did not finish this part of the task and can therefore not present meaningful results.

III. RESULTS

Our main contribution to the hackathon is the visualisation dashboard pictured in Figure 1. While we initially planned to include cluster information of semantic space, the clustering results are not displayed on the final visualisation as computations did not finish within time. However, cluster information is available in the individual visualisations as produced by UMAP (Figure 3).

Furthermore, we implement an interactive component, such that if a user highlights a data point, or area of data in one of the plots, the corresponding data in other plots is highlighted as well (Figure 2).

Results are displayed per individual community. In the demo version, not all communities included in the original dataset

are available, since most of them are too large to be interactively displayed in-browser on commonly available computer systems. However, the visualised features were computed for all communities for possible downstream analysis applications.

REFERENCES

- [1] G. Branwen, N. Christin, D. Décary-Héту, R. M. Andersen, StExo, E. Presidente, Anonymous, D. Lau, D. K. Sohlz, V. Cakic, V. Buskirk, Whom, M. McKenna, and S. Goode, *Dark net market archives, 2011-2015*, <https://www.gwern.net/DNM-archives>, dataset, Accessed: 2021-02-11, Jul. 2015. [Online]. Available: <https://www.gwern.net/DNM-archives>.
- [2] Alsayra (web forum), 2011–2012. [Online]. Available: <http://azsecure-data.org/other-forums.html>.
- [3] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y. Sung, B. Strope, and R. Kurzweil, “Universal sentence encoder,” *CoRR*, vol. abs/1803.11175, 2018.
- [4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of NAACL-HLT 2019*, J. Burstein, C. Doran, and T. Solorio, Eds., Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/n19-1423.
- [5] G. E. Hinton and S. T. Roweis, “Stochastic neighbor embedding,” in *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]*, S. Becker, S. Thrun, and K. Obermayer, Eds., MIT Press, 2002, pp. 833–840.
- [6] L. McInnes and J. Healy, “UMAP: uniform manifold approximation and projection for dimension reduction,” *CoRR*, vol. abs/1802.03426, 2018. arXiv: 1802.03426.
- [7] C. J. Nolet, V. Lafargue, E. Raff, T. Nanditale, T. Oates, J. Zedlewski, and J. Patterson, “Bringing UMAP closer to the speed of light with GPU acceleration,” *CoRR*, vol. abs/2008.00325, 2020. arXiv: 2008.00325.
- [8] M. Ester, H. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*, E. Simoudis, J. Han, and U. M. Fayyad, Eds., AAAI Press, 1996, pp. 226–231.
- [9] A. T. J. Barron, J. Huang, R. L. Spang, and S. DeDeo, “Individuals, institutions, and innovation in the debates of the french revolution,” vol. 115, no. 18, pp. 4607–4612, 2018, ISSN: 0027-8424. DOI: 10.1073/pnas.1717729115.
- [10] Z. Sun, R. Schuster, and V. Shmatikov, “De-anonymizing text by fingerprinting language generation,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December*

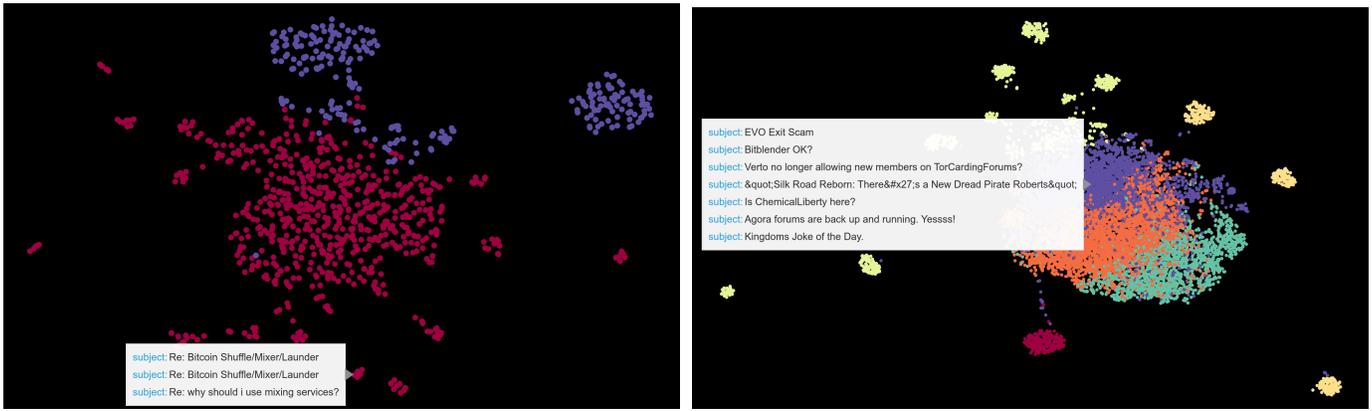


Fig. 3. Example plot for semantic space with clusters highlighted as produced by UMAP for the Hydra Forums (left) and for the Kingdom Forums (right).

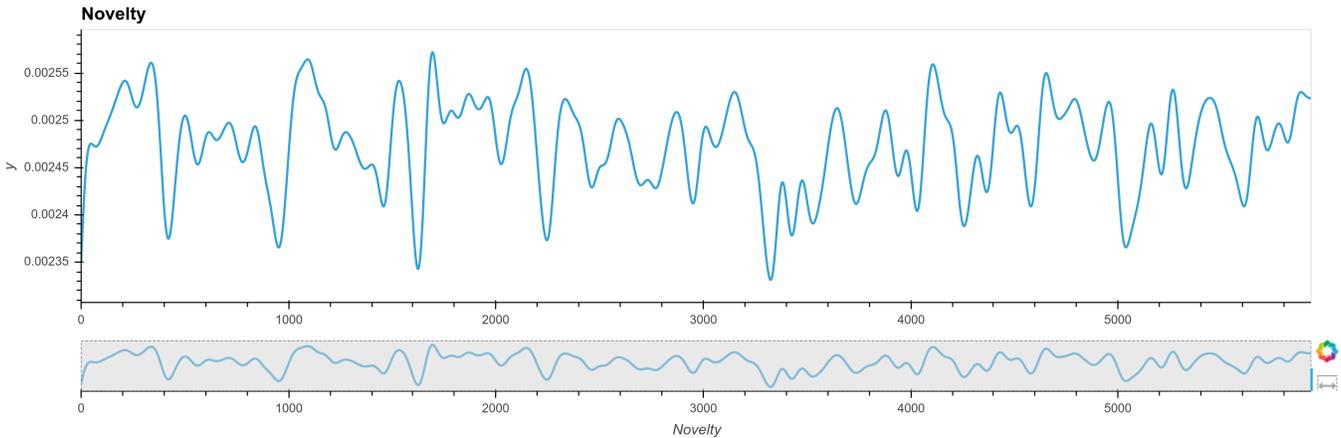


Fig. 4. Example plot for novelty over time

6-12, 2020, virtual, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.

[11] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online: ACL, Oct. 2020, pp. 38–45.

[12] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” 2019.

[13] M. Koppel and J. Schler, “Authorship verification as a one-class classification problem,” in *Machine Learning, Proceedings of (ICML 2004*, C. E. Brodley, Ed., ser. ACM International Conference Proceeding Series, vol. 69, ACM, 2004. DOI: 10.1145/1015330.1015448.

[14] J. Bevendorff, B. Stein, M. Hagen, and M. Potthast, “Generalizing Unmasking for Short Texts,” in *14th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019)*, J. Burstein, C. Doran, and T. Solorio, Eds., ACL, Jun. 2019, pp. 654–659. [Online]. Available: <https://www.aclweb.org/anthology/N19-1068>.