UNIVERSITY OF CAMBRIDGE

Cambridge Cybercrime Centre
COVID Briefing Paper # 17

Bursty and Emerging Topic Detection: Tracking Terms on Underground Forums

Jack Hughes

27 October 2020

Executive summary

CCC COVID Briefing Papers are an ongoing series of short-form, open access reports aimed at academics, policymakers, and practitioners, which aim to provide an accessible summary of our ongoing research into the effects which the coronavirus pandemic (and government responses) are having on cybercrime.

Underground cybercrime forums contain discussions of a range of topics, adverts for services, hacking tutorials, sales of items on marketplaces, and more general chatter. Off-the-shelf natural language processing techniques may struggle with 'noisy' text data from communities such as these, where community-specific slang and jargon evolve over time and topics come and go very quickly. We developed a tool [1] for detecting these so-called 'bursty' trending topics (using a Bayesian log-odds approach [2]) which detects change in the forum vocabulary and filters out consistently used jargon and slang. We apply this to look at emerging coronavirus-related topics within the news and marketplace sections of two underground forums. Forum 1 is a large, well-known, hacking forum, and Forum 2 is a smaller forum for sharing gaming-related hacking techniques. These forums showed clear differences in their discussions of the pandemic, suggesting that its effect on the cybercrime underground may well not be homogeneous.

Modelling hot topics

The tool works by comparing two time windows, using a 'prior' window and a 'target' window, the length of which can be changed to detect long or short-term trends. The prior window acts as a reference for the model, allowing the emergence of new topics to be detected. The **log odds score** for each term is similar to odds ratios; if a particular word has higher odds of appearing in the target (current) window than the prior window, we have established an emerging topic. When visualised, we see the emergence of these 'bursty' trending topics, with peaks appearing where a term becomes significant, and disappearing when it ceases to be the 'hot new topic'. For this analysis, we use a month-long prior window starting on 1 January 2020, followed by a gap of ten days, and then a two-week target period. We then slide these windows forward day-by-day to track emerging terms over time. The gap of ten days between the reference and target windows is important as this means that slowly-emerging trends are likely to be present in both windows, and so would not be as significant; this modelling approach is therefore best suited to detect *rapidly-emerging* topics.

Discussions of the pandemic in the cybercrime underground

Figure 1 shows COVID-related emerging terms for Forum 1's news-based subforums. Initially, we find "corona" trends during the middle of February, and later in February, "covid" also emerges. Following this are discussions of vaccines and mentions of a lockdown. "China" and "Chinese" emerge around April 2020, at the same time as "Zoom" (this is the time Zoom-bombing became a concern). There is a second rise of "vaccine" towards the end of April, followed by a significant increase in discussions relating to "Trump".

While Forum 1 has emerging terms spread across time (with new pandemic topics becoming 'hot' over time, indicated by coloured lines in Figure 1), Figure 2 shows a different picture for COVID-related topics

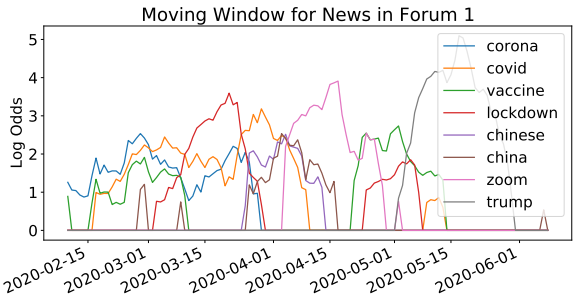

Figure 1: News subforums on Forum 1

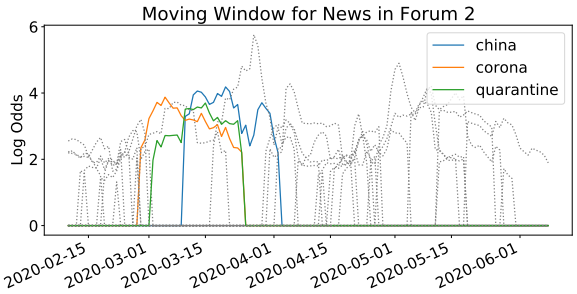

Figure 2: News subforums on Forum 2

emerging in Forum 2. While the pandemic continues to be discussed, we do not see the emergence of a series of pandemic 'hot topics' in Forum 2 in the same way, with new COVID-19 topics ceasing after the first month. The dotted lines represent non-COVID terms, provided for context across all trending terms.

COVID-19 and underground cybercrime markets

Forum 1 also contains a marketplace subforum. As shown in Figure 3, the only pandemic-related emerging term occurs during March 2020, when "covid" trends in relation to discussions around the disruption of global supply chains. Other research has found the pandemic resulted in increased sales on this market, with a sharp, but short-lived, increase in trading volumes [3]. Our evidence suggests that this increase was due to already established products and services being traded at higher volumes rather than novel, COVID-related products.

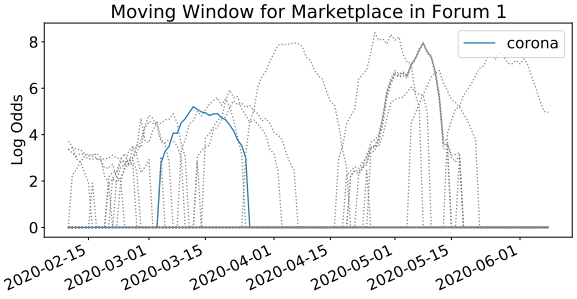

Figure 3: Market subforums on Forum 1

Conclusions

We showcase in this briefing paper the use of 'bursty' topic modelling to analyse cybercrime forums. This method highlights how discussions within a topic (such as the pandemic) change over time. Forum 1 contained a varied set of COVID-related terms emerging over time, keeping the overall topic of the pandemic significant, whereas discussions of coronavirus in Forum 2 did not develop into a series of novel emerging topics. This suggests that the effects of the pandemic on discussions in the underground community are not homogeneous, with larger, multi-purpose sites engaging in what are clearly ongoing social discussions as the pandemic progresses, and others lacking a sustained interest in COVID-19 as a 'hot topic'.

[1] J. Hughes, S. Aycock, A. Caines, P. Buttery, & A. Hutchings, *Detecting trending terms in cybersecurity forum discussions*. Workshop on Noisy User-generated Text (W-NUT), 2020.
[2] B. Monroe, M. Colaresi, & M. Quinn, *Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict*. Political Analysis, 16(4), pp.372-403, 2008.
[3] A. V. Vu, J. Hughes, I. Pete, B. Collier, Y. T. Chua, I. Shumailov, & A. Hutchings, *Turning up the dial: The evolution of a cybercrime market through set-up, stable, and COVID-19 eras*. Proceedings of the ACM Internet Measurement Conference, 2020.

At the Cambridge Cybercrime Centre we make our research data available to other academics, sometimes before we have looked at it ourselves! Researchers can be provided access to our 'CrimeBB' dataset of (26 and counting) underground cybercrime forums, our extensive collections of chat channel data, and our new collections of forums relating to online right-wing extremism and radicalisation. We can also share email spam and sensor data related to DDoS and IoT malware. All these collections are regularly updated and can be rapidly provided under licence – for full details see: `https://cambridgecybercrime.uk`

The full set of CCC COVID Briefing Papers can be found at: `https://cambridgecybercrime.uk/COVID`