PUBLIC
FINAL VERSION

Web Science Institute Research Collaboration Stimulus Fund: The Digital Police Officer Final Report

Project Outline

The Digital Police Officer (DPO) project investigated the use of linguistic analysis to profile cybercriminals in carding forums, online venues for buying and selling stolen credit card data. When one carding forum closes, another rises: some criminals with a sound reputation on a defunct site may then 'port' their reputations to another without necessarily keeping the same username or other identifying features.

Bringing linguistic analysis into the context of online security, DPO examined how we can profile carding forum users by the way in which they communicate, analysing characteristics (such as vocabulary and grammar) to build a linguistic fingerprint. We examined Natural Language Processing (NLP) technologies and particular linguistic features such as forum-specific argot. By examining criminal user behaviour within these communities DPO was able to profile both community users and active discussion threads. These profiles would be useful for law enforcement agencies when perusing a focussed investigation (e.g. looking at activity and associates involved in a criminal carding network).

Project Outcomes

At the midpoint of the project, the DPO team had obtained ethics approval for the study, completed literature reviews on three topics, gathered two appropriate datasets, developed a protocol for dataset analysis, produced scripts to extract usernames and posts from one dataset, begun qualitative analysis regarding use of pseudonyms, and presented a poster at a cybercrime symposium. Please see the midterm report for more detail on these activities.

Since the midterm report, the DPO team has:

1. Created a Python-based TOR criminal marketplace forum crawler able to handle forum logins, challenge pages and navigate forum structures (e.g. follow links to next pages for large discussion posts). The crawler generates an HTML dump of a forum.
2. Crawled the live TOR-based AlphaBay forum, generating a substantial and up-to-date HTML dump (4,100 pages of discussion each with up to 25 posts, 31,000 posts, 12,000 profile pages of which most are inactive).
3. Created a Python-based HTML parser to create from HTML dumps CSV-formatted datasets suitable for NLP work. This work included discussion thread pre-processing such as HTML parsing, signature extraction, reply history extraction and support for the discussion thread dialogue model forums use.
4. Analysed AlphaBay and ShadowCrew dataset criminal vocabulary (e.g. criminal jargon for products on sale) and classified sets of posts to identify criminal activity types. This social scientific analysis has been encoded into computer scientific regex patterns for use in criminal named entity recognition matching, using both token text & parts of speech tag information.
5. Created a Python-based user and discussion profiler to extract users, URIs, criminal vocabulary named entities and other such information from parsed

CSV datasets. This profiler creates a detailed forum report containing a list of user & discussion pairings ranked according to similarity computed using a variety metrics (e.g. common NE mentions, common user mentions, common URI's). This report is useful as it helps law enforcement agencies to identify known associates, users with suspected multiple accounts (i.e. pseudonyms) and similar discussion threads to those hosted by criminals under investigation.

6. Profiled the entire AlphaBay forum and created a report detailing user and discussion thread similarity. Performed manual inspection and validation of the results.
7. Acquired an HTML dump of SilkRoad2 from Don Hobson (ECS). This dataset arrived too late to analyse but has been useful in helping us demonstrate credibility when getting into proposals.
8. Created a web and social media presence with the targeted aim to build credibility for getting into proposals. This includes creating a website, linking to conference posters, creating a Twitter account and plans to make a web observatory to advertise the forum datasets we now have access to. See the Impact Review section for more details.

Collaboration with External Stakeholders

The DPO team has drawn on its contacts in Law Enforcement and Academia to provide advice and support throughout the duration of the project. Many of these contacts are also interested in seeing the outcome of the project and to work with the team as we seek to extend and deepen the research.

Impact Review

The DPO project advanced the state of the art by combining datasets crawled from modern-day criminal forums (i.e. over the dark web), insights from social science regarding criminal vocabulary and activity patterns and natural language processing approaches to entity profiling, disambiguation and text stylometry. The results seen are allowing us to move beyond domain unaware text stylometry (e.g. on clear web forums unrelated to criminal activity) and experiment with alternative features that have a greater potential information value.

DPO has had impact beyond the technical also. Besides engaging with contacts as described above we have generated a targeted web and social media presence, engaging with web observatories and attending key domain events. The targeted web and social media presence in particular has been instrumental in convincing proposal consortia of our track record in this area and securing us a significant place in these proposals.

The DPO website and Twitter account have been launched and maintained at http://wordpress.it-innovation.soton.ac.uk/dpo/ and https://twitter.com/DPOProject respectively. The website provides an accessible introduction to the project as well as information about related organisations, resources and events. The Twitter account has been used to publicise DPO activities, particularly around the time of events.

An IT Innovation web observatory is planned to be setup in 2015. This will advertise datasets (both labelled and unlabelled) to encourage future collaborative research and benchmarking type activities. The DPO datasets will be listed here, with access restricted to organizations working in a collaborative project containing UoS as a partner and as such under an established legal contract with ethical approval. We will work with WSI to find the right way to advertise this type of material. IT Innovation will also list datasets from the REVEAL FP7 project and look to expand this in the years to come.

Regarding events, DPO took an abstract to EUROCRIM2015 (September 2015, Porto, Portugal), and posters to the Second Cybercrime Symposium (March 2015 Winchester) as well as the Web Science Institute “One Year On” event (June 2015, London, UK). These events were key for generating contacts, maintaining existing relationships, and raising the profile of the project.