The Digital Police Officer

Using Linguistic Analysis to Identify Cybercriminals

Clare J. Hooper[1], Craig Webber[2], Stuart E. Middleton[1], Gert Jan van Hardeveld[3], Mike Surridge[1]
1 IT Innovation Centre 2 School for Social Sciences 3 Web Science Centre for Doctoral Training

Contact
http://wordpress.it-innovation.soton.ac.uk/dpo/

https://twitter.com/DPOProject

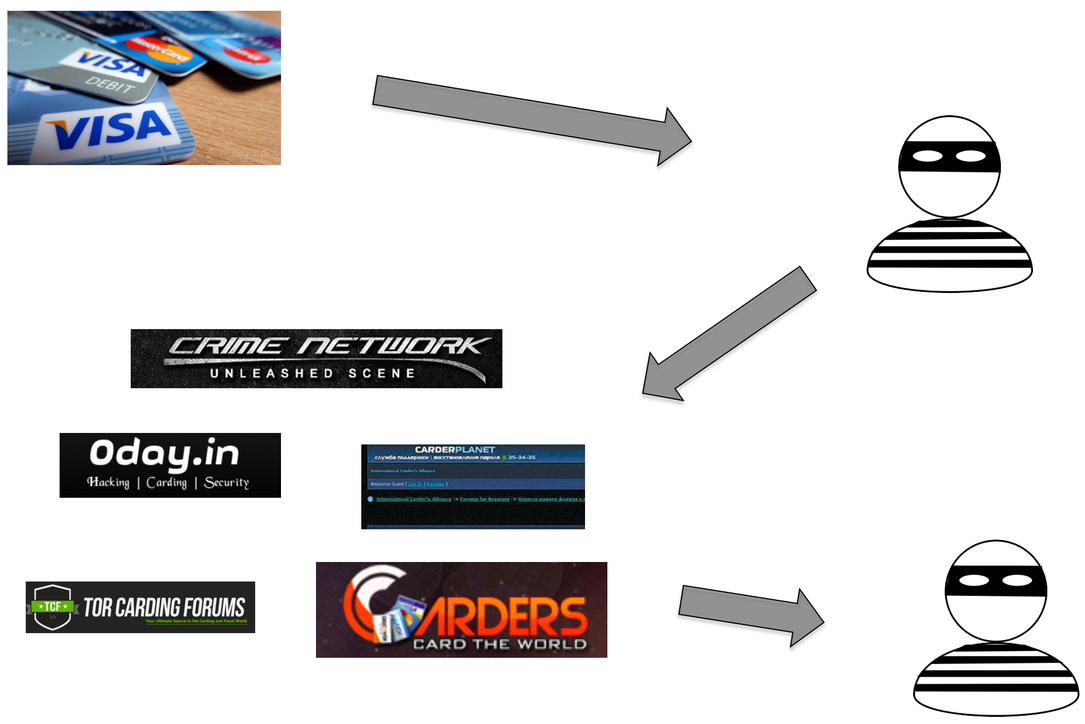
NLP: Stuart E. Middleton (sem@it-innovation.soton.ac.uk, @stuart_e_middle, @ITInno),
Criminology: Craig Webber (c.webber@soton.ac.uk, @cwebber01),
Gert Jan van Hardeveld (gjvh1g13@soton.ac.uk, @gertjangj)

Identifying cybercriminals

The aim of the Digital Police Officer project (DPO) is to identify cybercriminals based on their writing style. When a criminal underground forum is closed down, cybercriminals move to another one to further their illicit business. These users do not necessarily return with the same username. We are producing a demo that can still identify such cybercriminals. We look at the way they communicate, analysing the characteristics of forum users (i.e. based on their vocabulary and grammar) to build a linguistic fingerprint.

Interdisciplinary approach

The demo will be based on Natural Language Processing (NLP) technologies and will give us insight into how online (criminal) communities work. To build the demo, we have to use an interdisciplinary approach. In addition to the technical aspects, approaches from linguistics and criminology are needed to make the demo accurate in the context of criminal forums, where a specific jargon is used.



Carding forums

In our demo we will specifically focus on carding forums. These forums are used by cybercriminals to buy and sell stolen credit card data. Members of such forums use pseudonyms to communicate with one another to make sure they stay out of hands of law enforcement. Still, members interact by posting in threads on the forum and therefore leave a big trail of text that can be analysed. These large amounts of text lend themselves well to linguistic analysis.

Collaboration

This cross-faculty, cross-disciplinary collaboration combines technical and social expertise. The IT Innovation Centre brings expertise in cybersecurity and linguistic analysis, while the Faculty of Social Sciences and the Web Science Centre for Doctoral Training bring expertise in applying criminological research to a Web context. Together we will contribute to an interdisciplinary state of the art, i.e. produce a technical demonstrator and improved understanding of how online communities work.

What we have done so far

- Received ethics approval to conduct our research
- Conducted literature reviews on NLP and criminal forums, machine learning and language analysis, pseudonymity on the Web
- Gathered appropriate datasets
- Developed a protocol for analysing the datasets
- Produced scripts to extract usernames and posts from datasets (HTML and SQL)
- Tagged parts of dataset manually to obtain more accurate features for demo
- Presented poster at the Second Southampton Cybercrime Symposium

Future work

In future efforts we hope to improve our demo by expanding the domain. Linguistic analysis can also help combat cybercriminal activity on forums relating to extremism, weapons, drugs, child pornography, counterfeit money and counterfeit documents. Therefore, we want to gather data from broader sources and create a robust tool that will help law enforcement agencies in identifying a wide variety of cybercriminals. In doing this, we will contribute to research on criminal communities.

iT innovation

UNIVERSITY OF Southampton | Web Science DTC