Information Extraction from the Long Tail

A Socio-Technical AI Approach for Criminology Investigations into the Online Illegal Plant Trade

Stuart E. Middleton
Electronics and Computer Science
University of Southampton
sem03@soton.ac.uk

Anita Lavorgna
Department of Sociology, Social Policy & Criminology
University of Southampton
a.lavorgna@soton.ac.uk

Geoff Neumann
Electronics and Computer Science
University of Southampton
g.k.neumann@soton.ac.uk

David Whitehead
Conservation Policy (CITES)
Royal Botanic Gardens, Kew
d.whitehead@kew.org

ABSTRACT

In today's online forums and marketplaces cybercrime activity can often be found lurking in plain sight behind legitimate posts. Most popular criminology techniques are either manually intensive, and so do not scale well, or focus on statistical summaries across websites and can miss infrequent behaviour patterns. We present an inter-disciplinary (computer science, criminology and conservation science) socio-technical artificial intelligence (AI) approach to information extraction from the long tail of online forums around internet-facilitated illegal trades of endangered species. Our methodology is highly iterative, taking entities of interest (e.g. endangered plant species, suspects, locations) identified by a criminologist and using them to direct computer science tools including crawling, searching and information extraction over many steps until an acceptable resulting intelligence package is achieved. We evaluate our approach using two case study experiments, each based on a one-week duration criminology investigation (aided by conservation science experts) and evaluate both named entity (NE) directed graph visualization and Latent Dirichlet Allocation (LDA) topic modelling. NE directed graph visualization consistently outperforms topic modelling for discovering connected entities in the long tail of online forums and marketplaces.

CCS CONCEPTS

•Computing methodologies~Artificial intelligence~Natural language processing~Information extraction •Computing methodologies~Artificial intelligence~Natural language processing •Applied computing~Law, social and behavioral sciences~Sociology

WebSci '20 Companion, July 6–10, 2020, Southampton, United Kingdom

ACM ISBN 978-1-4503-7994-6/20/07…$15.00
https://doi.org/10.1145/3394332.3402838

KEYWORDS

Artificial Intelligence, Information Extraction, Natural Language Processing, Criminology, Socio-technical, Illegal Wildlife Trade, CITES

ACM Reference Format:

Stuart E. Middleton,. Anita Lavorgna, Geoff Neumann and David Whitehead. 2020. Information Extraction from the Long Tail: A Socio-Technical AI Approach for Criminology Investigations into the Online Illegal Plant Trade. In Proceedings of ACM Web Science conference (WebSci 2020). ACM, July 6–10, 2020, Southampton, United Kingdom. 4 pages. https://doi.org/10.1145/3394332.3402838

1 Introduction

Criminologists analysing online forums and marketplaces with cybercrime activity often focus on small subsets of posts. However, the posts often lurk in the long tail, using the bulk of legitimate discussion posts as cover to hide in plain sight. These posts can be time consuming to find, often requiring specialist domain knowledge to identify search terms and wider exploration of connected posts, profiles and external sites to understand the context in which behaviours are taking place. Finding and extracting these posts is important, as they offer important clues needed by both law enforcement and criminologists alike to investigate criminal or deviant behaviours taking place online.

We present in this paper an inter-disciplinary (computer science, criminology and conservation science) socio-technical AI approach to information extraction tailored to helping criminologists explore the long tail of online forums and marketplaces. We focus on case studies around the online trade in wildlife, which is of growing concern to agencies charged with tackling the illegal wildlife trade (IWT) [5]. This persistent form of environmental crime threatens a multitude of species with potential extinction, and erodes economic security and the rule of law in many countries around the world [14].

We explore how two information extraction methods, Latent Dirichlet Allocation (LDA) topic modelling and named entity

(NE) directed graph visualization, perform on the long tail of forum and marketplace posts, and how they can be integrated into a socio-technical system based on an ICT-enabled criminology methodology. We focus on how information extraction can support iterative identification of target suspects and connected entities with low density subgraphs, helping to reduce the volume of posts criminologists need to review and process.

We evaluate our approach using two case study experiments, each involving a one-week duration criminology investigation (aided by conservation science experts) of forums and marketplaces hosting evidence of illegally traded plant species. A standard criminology investigation is first performed, and discovered suspects and connected people, locations, plant species and organisations used as a ground truth. Both LDA topic models and NE directed graph visualizations are then evaluated on their ability to recall this ground truth.

The contributions of this paper are:

- Proposal of a new socio-technical AI approach incorporating human judgement (criminology, conservation science) into iterative cycles of long tail information extraction.
- Results from a case study evaluating two standard information extraction methods applied to the problem of online illegal wildlife trade investigation.

2 Related Work

2.1 Criminology methods for analysing online forums

Cyberspace provides opportunities both for new criminal markets and for traditional crimes such as internet-facilitated drug trafficking. Criminologists have studied these markets by both adapting traditional social sciences methods (e.g. active or passive virtual ethnographies, content analyses, visual studies) and adopting new strategies, sometimes in partnership with computer or data scientists [9]. Digital data can be collected manually or automatically. Manual collection strategies mostly involve copying and pasting online content. Such a process can be quite time-consuming, but has proven effective [10] [16] in providing in-depth knowledge on the characteristics of online criminal and deviant activities and groups.

Automatic collection involves the use of software to gather digital content, giving the researcher access to larger datasets. For these larger volumes of data, extracting the information relevant to specific research projects can be challenging. As summarised by [3], the three main methods for the automatic collection of digital data are mirroring (e.g. web crawling), active monitoring (e.g. monitoring synchronous and ephemeral online communications) and leaking (e.g. collecting information from online criminal markets willingly made public by someone in an attempt to harm the competition). Examples include [4].

This paper proposes an ICT-enabled criminology analysis that offers a socio-technical approach to help overcome the limits associated with existing combinations of manual criminological analysis and automated collection approaches. We compare methods analysing the important long tail of online forums, and highlight in particular how NE directed graph visualization methods can be used to find a balance between the need for filtering of large forum post volumes whilst still allowing criminologists to perform context analysis and make important subjective assessments on the behaviour patterns discovered.

2.2 Topic model analysis of online criminal forums

Topic models and clustering approaches have received a lot of attention in recent years, being applied to online forums, web pages and social media datasets. For cybercrime and law enforcement applications Latent Dirichlet Allocation (LDA) has been widely explored, with [11] using LDA to generate feature vectors that are then used to prepare training data for supervised classification of crime incidents, and [13] using LDA to generate topics for discourse analysis to explain community level activity within TOR-based online criminal forum dumps such as Silk Road 2.

Other methods exist for topic model analysis of criminal datasets such as the online Hawkes process estimation algorithm [11], working on features generated by a combination of word2vec and the word mover distance function, to compute predictive policing models. Stanford NER has been used to cluster criminal cliques [8], followed by agglomerative clustering applied to posts annotated using parts of speech (POS) tagging and WordNet hypernym classes, to visualize clusters of posts associated with particular criminal activity. Graph-based hierarchical clustering has also been used to classify crime incidents in newspaper articles [2], using NER entity pairs and a word2vec model coupled with a cosine similarity metric to generate the sub-graphs for clustering.

Most topic modelling approaches published today with criminology applications work at a community level, computing topic sets for whole forums and missing potential patterns in the long tail hidden inside threads within these forums. Our work directly addresses the challenge of long tail information extraction, and we compare how the most common topic modelling approach (LDA) performs against a NE directed graph visualization approach.

2.3 Crime data mining and directed graph visualization of online criminal forums

Work on creating graphs of online criminal forum and social media posts has a long history [19], with recent approaches changing from manual visualization towards crime data mining. Crime data mining [1] includes entity extraction, clustering, association rule mining, classification and social network analysis (SNA). The CopLink project [1] is an example, where named entity extraction (of people, addresses, vehicles and drug products) and a mix of hierarchical clustering and string matching based identify disambiguation supports criminal network analysis for law enforcement applications. Work on visualizations within a socio-technical methodology for law enforcement includes [15],

where visualization of open source intelligence (OSINT) is combined with a 'human in the loop' methodology.

Older graph-based approaches [20] have focussed on link analysis, such as page rank, and sentiment analysis coupled with metrics to measure network centrality and group-level activity visualizations. More recent graph-based approaches include GraphExtract [17], which uses sub-graph overlap metrics to perform predictive policing, identifying within manually curated criminal activity graphs suspects engaging in sub-graph behaviour similar to previous offenders. The INSiGHT project [6] used nearest neighbour matching of person-focussed subgraphs to compile suspicious activity reports around radicalization from government intelligence documents. Approaches such as graph-based anomaly detection (GBAD) using minimum descriptive length (MDL) graph beam searches [12] and matrix factorization methods [7] have been successfully applied to online fraud detection within financial transaction datasets.

Graph-based approaches are often visualized using dense sub-graph graphical representations for users such as law enforcement to see discovered connections, again mostly focussing on the whole dataset and not individual target-focussed graph analysis. Our work visualizes named entity sub-graphs, starting with a root node of a specific suspect, with a variable graph depth to control the node density of our visual graphs, and thus how easy it is for criminologists or law enforcement officers to quickly identify useful information within them.

3 Method

Our experimental method was to first perform a standard criminology analysis to create a ground truth intelligence package of suspects likely to be engaging in illegal behaviour, along with their connected entities of interest including people, locations, species being traded and organisations discussed such as plant nurseries illegally selling species. These entity types are similar to the UK law enforcement and Home Office standard POLE format (People, Object, Location and Events). We then ran a forum crawling pipeline for automated data collection and performed both a topic model analysis and NE directed graph visualization. Details of these approaches can be found in this section.

Our case studies focus on a limited number of selected genera and species, which were deemed by our conservation experts at Royal Botanic Gardens, Kew, (RBG Kew) to be particularly relevant for study as they are both threatened in the wild and protected by The Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES). Our final shortlist of target species were as follows: Seven species (and associated subspecies) of Ariocarpus, a type of cacti native to Mexico/Southern USA; thirteen species of succulent Euphorbia, whose wild populations are endemic to a number of islands including Madagascar; Saussurea costus, a perennial herb (thistle) native to the Himalaya region. More details of the conservation context behind this work can be found in [18].

Our case studies examined in total five plant-focussed forums for Ariocarpus, one forum for Euphorbia and for Saussurea costus we looked at eBay, Alibaba and Esty. In total we had nine websites crawled, providing 13,697 posts by 4,009 authors in 1,826 forum threads, posted on dates ranging from 2006 to 2019. The full dataset breakdown can be seen in table III and more details found in section 4.

3.1 Criminology Analysis

For the criminological analysis, data was organised into separate Excel files (one per forum and one grouping together data from different marketplaces in the case of the *Saussurea* group). The researcher went manually through all forums to identify a subset of relevant forum threads and individual posts that could be conceptualised into three main groups of interest. First, posts with information regarding (potentially) illegal trades; in particular, the researcher was interested in the specific cases reported, the countries involved, the central actors involved, and how these trades were perceived by the forums' users. Second, the researcher looked for information regarding discussions about the need of CITES permits, and how these were framed by forums' users into broader debates of species conservation. Third, a selection of posts highlighting and accounting for relevant subcultural elements of forums' users were identified.

The subset of relevant posts (N=543 for *Ariocarpus*; N=768 for *Euphorbia*; and N= 947 for *Saussurea*, of which 20 on Esty, 695 on eBay and 232 on Alibaba) was then qualitatively analysed through NVivo (a data analysis software package that allows one to manage and arrange unstructured information).

The fact that most posts from the forums were discarded is not surprising: the forums analysed are mostly intended for the sharing of legitimate information (on cultivation, on plant-related events, etc.) and photos. Hence, most posts were discussions about how to better grow cacti (in terms of soil, pot, watering, and so on), or to exchange and comment pictures taken from plant shows from around the world. Overall, these posts suggested that the online communities considered are generally very friendly environments, open to new members (only one habitual forum user was accused once to be a potential scammer, and two new users accused of trying to sell potentially illegal specimens). Most users genuinely enjoy the slow (but rewarding) process of cacti cultivation, and do not seem interested in pursuing illegal activities.

We use a subset of users from this analysis, who are exhibiting behaviour or links associated with the illegal plant trade, in the evaluation (section 4) as a ground truth set of target suspects.

3.2 Crawling forums

The information extraction methods defined in sections 3.3 and 3.4 require forum datasets to work on. To collect posts from online forums we employed a basic search, crawl and parse data ingest pipeline using off the shelf tools. First, we used a set of search keywords, outlined in table I, to run a set of Microsoft Bing [1] queries to identify candidate forums containing posts

[1] www.bing.com

relevant to each case study. We then created an HTML dump of each candidate forum using the open source DARPA MEMEX Undercrawler[2]. Because Undercrawler downloads all posts in all forum threads, a relevance filter is then applied to store only those threads that contain mentions of the phrases in the case study search keyword list. Each forum's HTML dump is then parsed using a combination of Python's standard HTML parser library and a set of HTML tag names, pre-defined in forum specific configuration file templates. The final output for each parsed forum is a JSON file containing relevant forums threads and posts, with each post containing metadata, such as author name and timestamp, and the sentences of the post.

3.3 Topic Modelling Analysis

The idea of the topic modelling analysis is to generate a small set of phrase groups around a target suspect, to allow a criminologist to identify other connected people, locations, species and organisations that can then become new suspects for investigation in the iterative ICT methodology (see section 3.5).

Ariocarpus search terms
Ariocarpus agavoides; Ariocarpus bravoanus; Ariocarpus bravoanus bravoanus; Ariocarpus bravoanus hintonii; Ariocarpus kotschoubeyanus; Ariocarpus kotschoubeyanus albiflorus; Ariocarpus retusus; Ariocarpus retusus scapharostroides; Ariocarpus confuses; Ariocarpus scaphitostris; Ariocarpus trigonus; tamaulipas living-rock cactus; pezuña de venado; nuevo leon living-rock cactus; biznaga maguey pequeño; biznaga peyotillo; biznaga maguey pata de venado; biznaga maguey peyote cimarrón; biznaga maguey chautle

AND Buy/ for sale
Exclude: seeds

Euphorbia search terms
Euphorbia decaryi; Euphorbia ampanihyensis; Euphorbia robinsonii; Euphorbia sprirosticha; Euphorbia quartziticola; Euphorbia tulearensis; Euphorbia capsaintemariensis var. tulearensis; Euphorbia francoisii; Euphorbia parvicyathophora; Euphorbia handiensis; Cardón de Jandía; Euphorbia lambii; Tabaiba Amarilla de Tenerife; Euphorbia bourgeana; Euphorbia stygiana; Euphorbia stygiana subsp. santamariae; spurge

AND Buy/ for sale

Saussurea search terms
Saussurea costus; S. lappa; Auklandia lappa; A. costus; Kuth; Aucklandia; Saussurea Root; Costus Root; Kustha; Kut; Postkhai; Kur; Kot: Kostum; Sepuddy; Koshta; Kotu; Aplotaxis lappa; Theodorea costus

AND variants with root, oil, or root oil together

Table I: Search terms used for case studies. Keywords created by criminology and conservation experts, then used in Bing searches and later to perform relevance filtering.

We first filter each forum to remove any irrelevant threads that do not contain a single post with the target suspect mentioned. A sentence corpus is then created by aggregating all relevant thread posts, and a bag of words collated from sentence unigram and bigram phrases. A document frequency filter is applied to remove very common phrases, removing phrases with >95% document frequency. Very uncommon phrases are also removed with a document frequency <2 occurrences. A scikit-learn[3] CountVector is computed using a maximum of 1,000 features, and a scikit-learn LatentDirichletAllocation (LDA) model. The LDA model hyper-parameters values (i.e. feature set size, number of topics and LDA learning offset) were found empirically using a grid search method, values for which can be seen in table II. The top 20 entity phrases were taken for each topic, since larger phrase sets (e.g. top 100 phrases per topic) are not practical for a criminologist to manually view quickly.

We also tested topic modelling using named entities labelled using Stanford NER, but the results were too poor to use as the topic models degraded in quality substantially with the lower occurrence frequency of named entities, as opposed to the full set of unigram and bigram phrases in each post.

3.4 Named Entity Recognition and Directed Graph Visualization Analysis

The graph visualization aims to generate small directed graphs of connected entities (i.e. people, locations, species and organisations) with the target suspect as the root node. This automates to some degree the approach used in criminological analysis, where users are first identified and then posts analysed to see who is connected and what behaviours are being exhibited. The NE directed graph model hyper-parameters values (i.e. NE filter settings and depth of graph) were found empirically using a grid search method, values for which can be seen in table II.

Topic model hyper-parameters

num features	num topics	learning_offset	entities to view	mean recall
10	10	50	200	< 0.1
100	10	50	200	< 0.1
1000	10	50	200	0.14
5000	**10**	**50**	**200**	**0.27**
10000	10	50	200	0.23
20000	10	50	200	0.23
5000	20	50	400	0.22
5000	50	50	1000	< 0.1
5000	10	1	200	0.14
5000	10	10	200	0.18
5000	10	100	200	0.23
5000	10	200	200	0.14

NE graph hyper-parameters

NER filter	graph depth	entities to view	mean recall
any	1	49	0.20
any	2	872	1.00
any	3	4285	1.00
plant, person, loc, nationality, org email, money, criminal charge	2	462	1.00
plant, person, loc, nationality, org	**2**	**426**	**1.00**

Table II: Grid of model hyper-parameters for dataset (euphorbia) and connection type (people), optimising mean recall of ground truth entities. Entities to view are the number of entries in a visualization (too big and it's hard to read). Optimal model hyper-parameters in bold.

[2] https://github.com/TeamHG-Memex/undercrawler

[3] www.scikit-learn.org

We first take all the forum posts (relevant or not) and label named entities within sentences using the Stanford CoreNLP toolkit[4]. Optionally named entities are filtered based on type, allowing less useful types (e.g. religion) to be removed. An inverted index is then built up for all named entity phrases, listing the threads and posts which mention them. Authors, threads and posts are added in as entities as well. No entity disambiguation is performed, so entity names are suffixed by their post identifier to avoid jumps from one post to a completely unconnected post which happens to also mention the entity phrase.

A breadth first directed graph walk is then performed using the inverted index, starting from the target suspect's named entity and going 1, 2 or 3 levels deep. Connected entities are added to the graph at each level, and used as new root nodes for the next level's graph walk. Walks are allowed to find connections in the forward direction (e.g. thread A contains post B, post B mentions entity C) and backward direction (e.g. entity C was mentioned by post B). Because entities are suffixed by post identifiers these graphs represent sequences of entities in the context of a conversation. Once graphs are generated all entity nodes with the same base name can be safely merged, discarding the post identifier suffixes. This allows the user to see sets of posts talking about the same thing in the context of a conversational sequence.

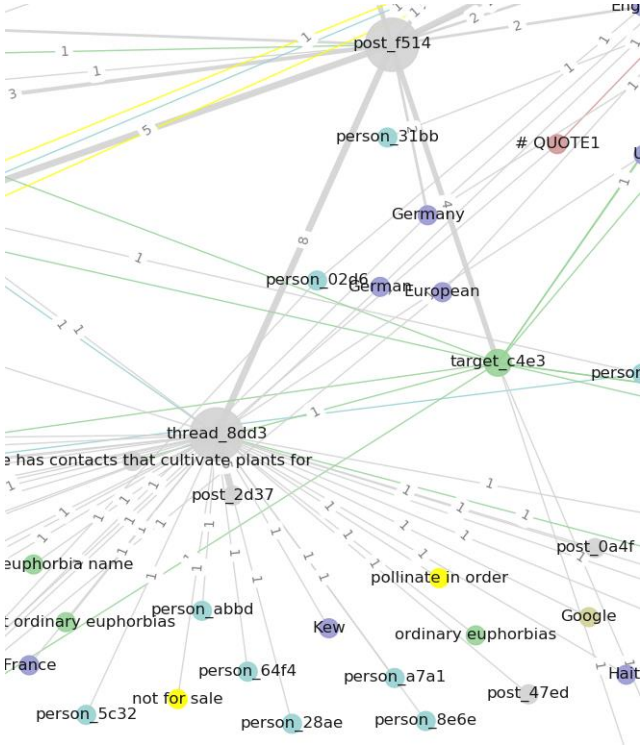

Figure 1: NE directed graph visualization for species Ariocarpus with a depth of two connections. Target suspects are green, posts and threads grey, predicates yellow, locations purple and people blue. Names hashed to pseudonymize.

[4] stanfordnlp.github.io/CoreNLP

Once a target suspect's graph is generated it is visualized using a matplotlib [5] and networkx [6] visualization. Pseudonymized examples of graph visualizations can be seen in figure 1. To make it easier for a criminologist to process 100's of connected entities on a graph we colour code nodes by entity type. The visualization is interactive, and can be zoomed in and panned around to explore dense data more easily. Our directed graph visualization code is released as open source[7].

3.5 ICT-Enabled Criminology Analysis

We observed from our analysis of the criminology method in section 3.1 that the exploration of forums was often iterative. Posts from a set of target suspects would be labelled, new connected people identified and then they would be the source of posts for the next iteration. Coverage of forums was limited by the bandwidth of how many posts the criminologists could process at any one time.

Building on these observations we created an iterative ICT-enabled methodology where criminologists and computer scientists work together in exploration cycles to develop intelligence packages of suspects and evidence of illegal behaviour. Each exploration cycle starts with a criminologist manually analysing sets of posts and identifying target suspects. Posts relating to these suspects are then crawled and modelled by the computer scientist, using a topic model or directed graph visualization, and an expanded set of connected entities identified. The criminologist reviews these new connected entities, examines the original posts mentioning them to understand context and uses a filtered subset as targets for the next cycle.

This type of cyclic methodology can easily integrate expertise from other disciplines, especially at the end of each cycle, such as in our study conservation science (RBG Kew to identify species in posted images) and law enforcement (UK Border Force to advise on illegal behaviour types and priority areas). Key to the success of this type of cyclic exploration is the ability of computer science methods at each cycle to identify relevant new suspects within large forum datasets. We evaluate in section 4 how well two computer science methods perform doing just that.

[5] matplotlib.org
[6] networkx.github.io
[7] https://github.com/stuartemiddleton/intel_viz_entity_graph

Dataset	# websites	# authors	# threads	# posts
ariocarpus	5	3308	1281	9676
euphorbia	1	545	545	3733
saussurea	3	156		288
all	9	4009	1826	13697

		Mean recall			
Connection type	Model type	ariocarpus	euphorbia	saussurea	all
people	topic model	0.00	0.27		0.14
	NE graph	0.34	0.78		0.56
location	topic model	0.00	0.00	0.00	0.00
	NE graph	0.56	1.00	1.00	0.85
plant species	topic model	0.05	0.24	0.00	0.10
	NE graph	0.20	0.40	0.14	0.25
organisation	topic model	0.00	0.00		0.00
	NE graph	0.33	0.14		0.24

Table III: Dataset and model result breakdown. Saussurea dataset was from auction sites with no threads. NE directed graph model outperforms the topic model in all cases.

4 Experimental Results

We ran two experiments, the first focussing on species of Ariocarpus (cacti) and the second on species of Euphorbia (succulents) and Saussurea costus (thistle). Each experiment ran over a one-week period, with a criminologist using Bing searches to identify forums and marketplaces where posts contained mentions of species of interest. They then applied the criminology analysis method detailed in section 3.1, manually browsing posts from 100's of authors to identify 4 or 5 ground truth suspects who were probably engaging in illegal trade activity and their connected entities (i.e. people, locations, species and organisations). We examined in five plant-focussed forums for Ariocarpus, one forum for Euphorbia and for Saussurea costus we looked at eBay, Alibaba and Esty. In total we had nine websites crawled, providing 13,697 posts from 4,009 authors. The full dataset breakdown can be seen in table III. These data volumes give you an idea of how many posts the criminologist needs to manually browse to check for connected entities around sets of potential suspects; the goal for our models is to automatically extract ground truth connected entities from the dataset, which would help speed up the criminology analysis process around potential suspects.

After the full criminology analysis generated a ground truth we ran the crawling pipeline in section 3.2 to download datasets of posts from these forums and marketplace for automated analysis. These datasets were processed using the topic modelling (section 3.3) and NE directed graph visualization (section 3.4) methods, with optimal parameters from table II, and a set of likely connected entities computed for each ground truth suspect.

We then calculated the mean recall of ground truth entities per species, averaged across the 4 or 5 suspects, which can be seen in table III. We defined recall as the number of ground truth entities reported by a model as a fraction of the total ground truth entities.

5 Discussion and Conclusions

Finding our set of target suspects took a relatively long time, with the criminology analysis of each case study taking around 25 hours of effort browsing 1,000's of posts, curating and annotating example conversations and exploring the social networks and context around potential suspects. This highlights how much time could be saved with more automated tools.

Our experiments show that the NE directed graph visualization method delivered a much better recall of connected ground truth entities compared to the topic model. Many connected entities had a low term frequency and were buried in the long tail of forum discussions, reducing their importance in the topic model. The numbers of entities shown in our directed graphs were not too large, and it was easy to identify entity types with colour coding. A few targets had a lot of connected entities (>400) and these created dense and hard to read graphs, but in these cases the topic model also produced dense hard to read entity phrases lists.

It is hoped that by generating evidence around the performance of ICT-enabled criminology analysis methodologies we can start to build the case for wider adoption of socio-technical AI into law enforcement practices, helping to more efficiently counter and disrupt the illegal wildlife trade through the use of online intelligence, while minimising some biases intrinsic to monodisciplinary data-driven technologies. Of equal value, indeed, is also a deeper understanding of the socio-economic and socio-behavioural aspects of these crimes made possible by this interdisciplinary approach. These insights have the potential to inform a range of alternative interventions, `softer' than traditional enforcement but, for the conservation of endangered species, potentially more effective.

Our socio-technical AI approach generates a series of incremental intelligence packages relating to target suspects and their connected entities. In its raw form this type of intelligence package does not reach the evidential standard required to be presented in a court of law. However, because we maintain provenance links from all entities back to the original online posts that mention them, we think a target-focussed evidence package could be generated in principle. This might contain an archive of all posts found mentioning a target suspect, set of connected entities mentioned in these posts with snippets of context approved by the criminologist, and a summary sheet again edited and approved by the criminologist. Optionally links to relevant out of band corroborating evidence, obtained from other means such as human intelligence resources could be added also. This type of provenance-based intelligence package might meet the evidential standard required in a court of law.

Recent big data approaches by companies such as Amazon AWS have seen the development of massive 'data lakes', centralizing vast sets of structured and unstructured data into a single multi-purpose searchable repository. Our intelligence packages are well suited for this type of collective intelligence sharing, and this would allow application of the advanced information retrieval methods being offered today by these cloud service providers. It should also be noted that for forums that hide on the dark web our Bing search approach is unlikely to be an

effective strategy to discover forum sites. In this case the criminologist would need to engage in an expert analysis of the community, maybe working to gain a level to trust from community members in order to discover relevant forums which can be later crawled. This highlights the strength of a socio-technical AI approach, as there is no way to automate this type of interactive community search without a human in the loop.

With regards ethics of our data collection [9] we did not engage in participant observation of virtual communities, but rather in their passive monitoring and download of data created by online community users. Therefore, we did not engage in any entrapment activity or encourage the illegal trade. Where we needed to login to forums, and therefore to agree to terms and conditions around crawling data, we honoured any robot crawling policy. For sites that do not allow crawling in their terms and conditions, we have only used public pages discovered by search engines without using a forum login. This is in line with what law enforcement working on illegal wildlife trades can currently do, as they do not have legal permission to consider in an intelligence package what is not available to them through the open webpages unless they obtain separate authority such as a search warrant.

For next steps the authors are starting to apply these techniques to other online cybercrime areas, such as TOR-based forums discussing ransomware, hacking services, denial of service attacks and cryptocurrencies, and to explore analysis of the 'crime as a service' space in general. For example, the CYShadowWatch project (DSTL, ACC2005442) is combining statistical machine translation and information extraction to explore socio-technical approaches to analyse Russian cybercrime posts. There are also opportunities to experiment with sub-graph classification and partial graph matching algorithms to try and automatically classify the criminal behaviour patterns associated with each suspect's connected entity graph.

ACKNOWLEDGMENTS

This work was supported by the Economic and Social Research Council (ES/R003254/1) and UK Defence and Security Accelerator, a part of the Ministry of Defence (ACC2005442).

REFERENCES

[1] Hsinchun Chen, Wingyan Chung, Jennifer J. Xu, Gang Wang, Yi Qin and Michael Chau, 2004. Crime data mining: a general framework and some examples. *Computer*, vol. 37, no. 4, pp. 50-56

[2] Priyanka Das, Asit K. Das, Janmenjoy Nayak, Danilo Pelusi and Weiping Ding, 2019. A Graph Based Clustering Approach for Relation Extraction From Crime Data. *IEEE Access*, vol. 7, pp. 101269-101282.

[3] David Décary-Hétu and Judith Aldridge, 2015. Sifting through the net: monitoring of online offenders by researchers. *The European Review of Organised Crime*, 2(2):122-141.

[4] Benoit Dupont, Anne-Marie Côté, Jean-ian Boutin and José Fernandez, 2017. Darkode: recruitment patterns and transactional features of 'the most dangerous cybercrime forum in the world'. *American Behavioral Scientist*, 61(11).

[5] Teresa Fajardo del Castillo, 2016. The European Union's approach in the fight against wildlife trafficking: challenges ahead. *Journal of International Wildlife Law and Policy*, 19(1), 1-21.

[6] Benjamin W.K. Hung, Anura P. Jayasumana and Vidarshana W. Bandara, 2017. INSiGHT: A system for detecting radicalization trajectories in large heterogeneous graphs. In *IEEE International Symposium on Technologies for Homeland Security (HST)*, Waltham, MA, pp. 1-7.

[7] Dongxu Huang, Dejun Mu, Libin Yang and Xiaoyan Cai, 2018. CoDetect: Financial Fraud Detection With Anomaly Feature Detection. *IEEE Access*, vol. 6, pp. 19161-19174.

[8] Farkhund Iqbal, Benjamin C.M. Fung, Mourad Debbabi, Rabia Batool and Andrew Marrington, 2019. Wordnet-Based Criminal Networks Mining for Cybercrime Investigation. *IEEE Access*, vol. 7, pp. 22740-22755.

[9] Anita Lavorgna, Stuart E. Middleton, Brian Pickering, Geoff Neumann, 2020. FloraGuard: tackling the online trade in endangered plants through a cross-disciplinary ICT-enabled methodology. *Journal of Contemporary Criminal Justice* (forthcoming).

[10] James Martin, 2014. Lost on the Silk Road. Online drug distribution and the 'cryptomarket'. *Criminology and Criminal Justice*, 14(3):351-367.

[11] George Mohler and P.J. Brantingham, 2018. Privacy Preserving, Crowd Sourced Crime Hawkes Processes. In *International Workshop on Social Sensing (SocialSens)*, Orlando, FL, pp. 14-19.

[12] Lenin Mookiah, William Eberle and Lawrence Holder, 2014. Detecting suspicious behavior using a graph-based approach. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, Paris, 2014, pp. 357-358.

[13] Rasmus Munksgaard and Jakob Demant, 2016. Mixing politics and crime – The prevalence and decline of political discourse on the cryptomarket. *International Journal of Drug Policy*, Volume 35, Pages 77-83, ISSN 0955-3959.

[14] Christian Nellemann, Rune Henriksen, Arnold Kreilhuber, Davyth Stewart, Maria Kotsovou, Patricia Raxter, Elizabeth Mrema and Sam Barrat, 2016. The Rise of Environmental Crime – A Growing Threat To Natural Resources Peace, Development And Security. *UNEPINTERPOL Rapid Response Assessment, United Nations Environment Programme and RHIPTO Rapid Response–Norwegian Center for Global Analyses*.

[15] Mariam Nouh, Jason R.C. Nurse, Helena Webb and Michael Goldsmith, 2019. Cybercrime Investigators are Users Too! Understanding the SocioTehnical Challenges Faced by Law Enforcement, In *Proceedings of the 2019 Workshop on Usable Security (USEC) at the Network and Distributed System Security Symposium (NDSS)*, 24-27 February 2019, San Diego, CA, USA.

[16] Meltem Odabaş Thomas J. Holt, Ronald L. Breiger, 2017. Markets as governance environments for organizations at the edge of illegality: insights from social network analysis. *American Behavioural Scientist*, 61(11).

[17] David Robinson and Chris Scogings, 2018. The detection of criminal groups in real-world fused data: using the graph-mining algorithm "GraphExtract". *Security Informatics*, 7 (2).

[18] Anita Lavorgna, Stuart E. Middleton, David Whitehead, Carly Cowell, 2020. FloraGuard, Tackling the illegal trade in endangered plants, Project report. *Royal Botanic Gardens, Kew*.

[19] Jennifer Xu and Hsinchun Chen, 2005. Criminal network analysis and visualization. *Communications of the ACM*, 48 (6), pp 100–107.

[20] Christopher C. Yang and Tobun D. Ng, 2007. Terrorism and Crime Related Weblog Social Network: Link, Content Analysis and Information Visualization. *IEEE Intelligence and Security Informatics*, New Brunswick, NJ, pp. 55-58.