Information Extraction from the Long Tail

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

Stuart E. Middleton
University of Southampton, Electronics and Computer Science

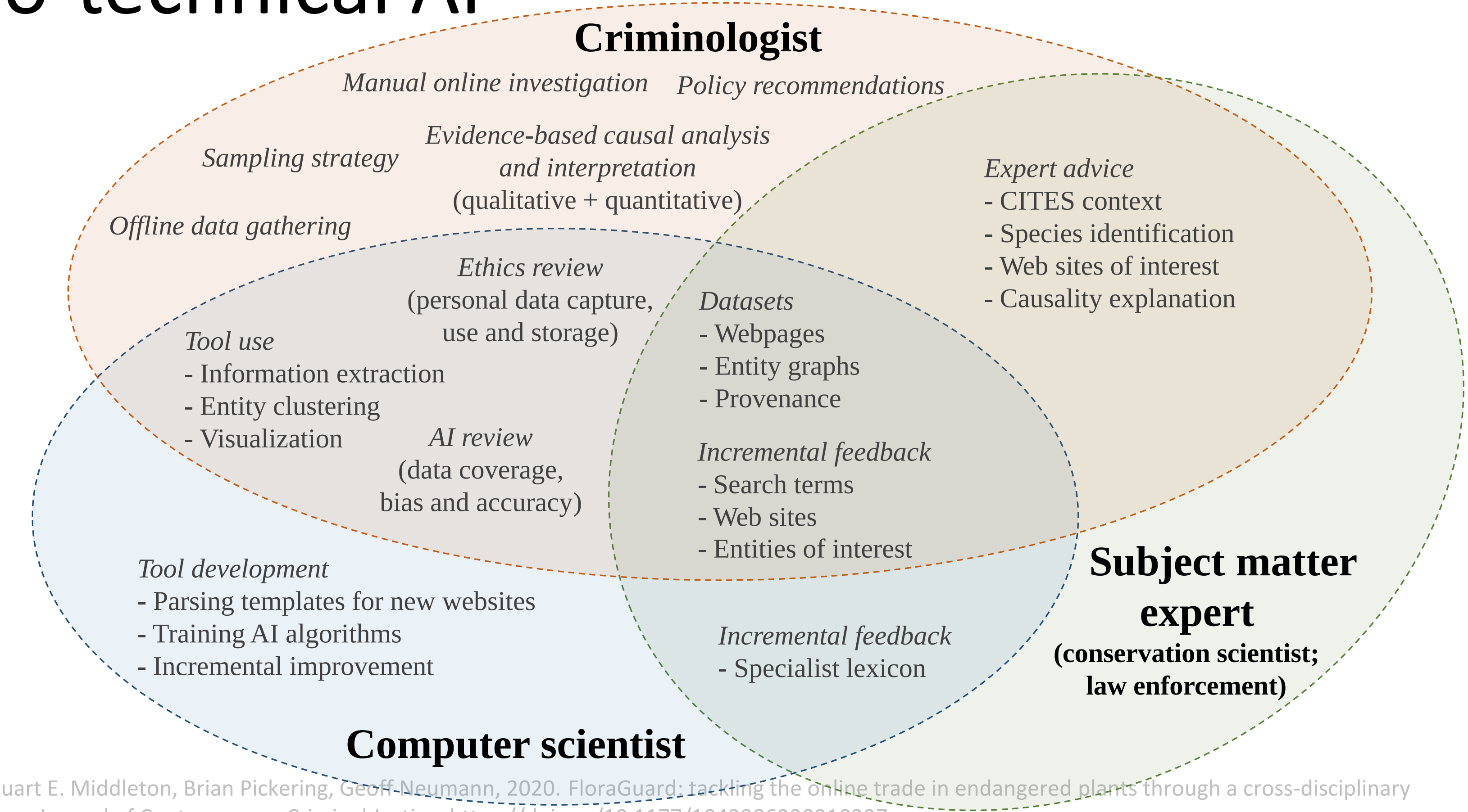
WebSci'20 Workshop: Socio-technical AI systems for defence, cybercrime and cybersecurity (STAIDCC20)
7th July 2020

Cite: 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 12th ACM Conference on Web Science (WebSci '20 Companion), July 6–10, 2020, Southampton, United Kingdom. ACM, New York, NY, USA, 7 pages.
https://doi.org/10.1145/3394332.3402838

Problem Statement

- Cybercrime activity within online forums and marketplace
  - Dark Web (TOR-based) - Open criminal marketplaces & forums, 'hard-core' users
  - Surface Web - Criminal and quasi-criminal activity embedded within active discussion forums hiding in plain site in the 'long tail' of discussion thread
  - Long tail = low frequency topics (e.g. niche topics; emergent topics; small communities)
- Popular criminology techniques
  - Manually intensive >> Hard to scale
  - Focus on statistical summaries across websites >> Miss infrequent behaviour patterns
- Our paper
  - Explores information extraction techniques applied to long tail posts, deployed using a socio-technical AI methodology supporting criminology investigations
  - Case study: Online illegal plant trade of CITES listed species

Socio-technical AI



Criminologist

Manual online investigation

Policy recommendations

Sampling strategy

Evidence-based causal analysis and interpretation (qualitative + quantitative)

Offline data gathering

Ethics review (personal data capture, use and storage)

Tool use
- Information extraction
- Entity clustering
- Visualization

AI review (data coverage, bias and accuracy)

Expert advice
- CITES context
- Species identification
- Web sites of interest
- Causality explanation

Datasets
- Webpages
- Entity graphs
- Provenance

Incremental feedback
- Search terms
- Web sites
- Entities of interest

Tool development
- Parsing templates for new websites
- Training AI algorithms
- Incremental improvement

Incremental feedback
- Specialist lexicon

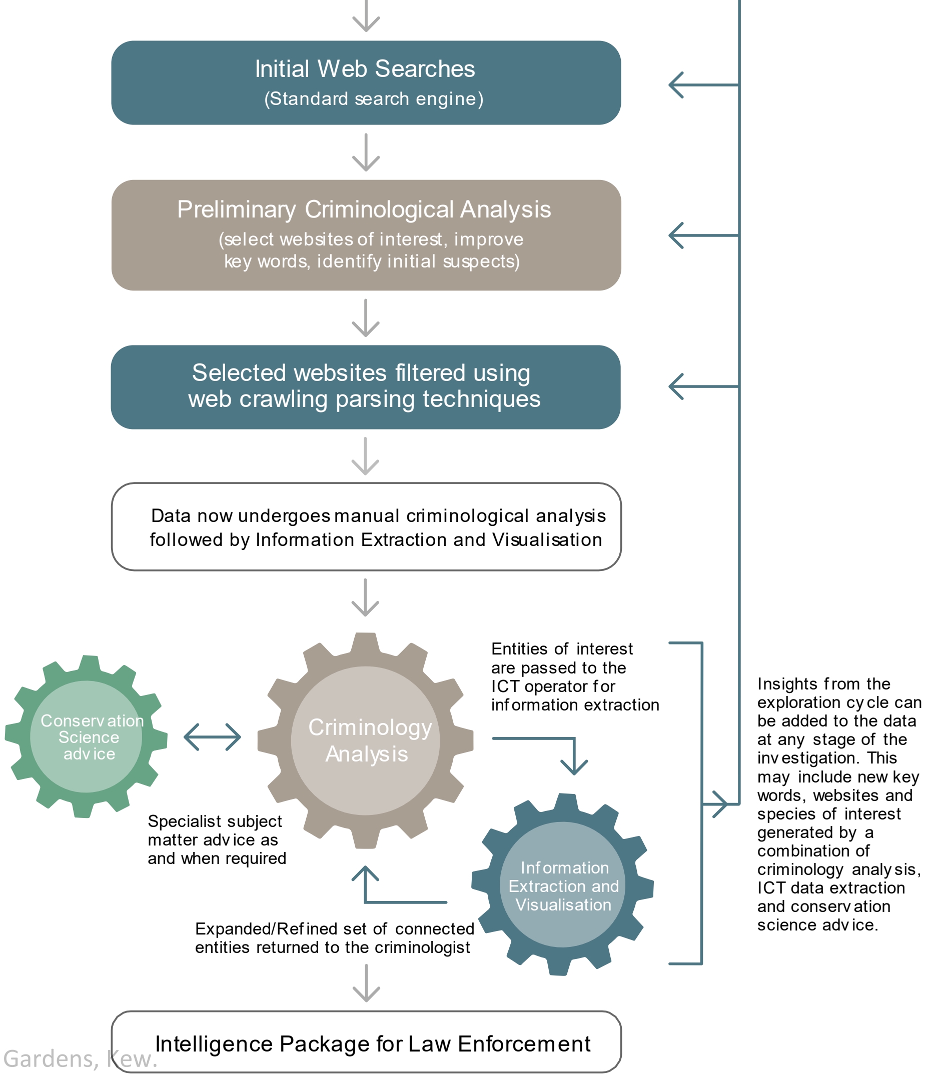
Subject matter expert
(conservation scientist; law enforcement)

Computer scientist

Cite: 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, https://doi.org/10.1177/1043986220910297

Socio-technical AI

- Cyclic methodology
  - Human in the loop with AI tools
  - Cycles of crim analysis & info extract
  - Refine final intelligence package



Initial Web Searches
(Standard search engine)

Preliminary Criminological Analysis
(select websites of interest, improve key words, identify initial suspects)

Selected websites filtered using web crawling parsing techniques

Data now undergoes manual criminological analysis followed by Information Extraction and Visualisation

Conservation Science advice

Criminology Analysis

Information Extraction and Visualisation

Entities of interest are passed to the ICT operator for information extraction

Specialist subject matter advice as and when required

Expanded/Refined set of connected entities returned to the criminologist

Insights from the exploration cycle can be added to the data at any stage of the investigation. This may include new key words, websites and species of interest generated by a combination of criminology analysis, ICT data extraction and conservation science advice.

Intelligence Package for Law Enforcement

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

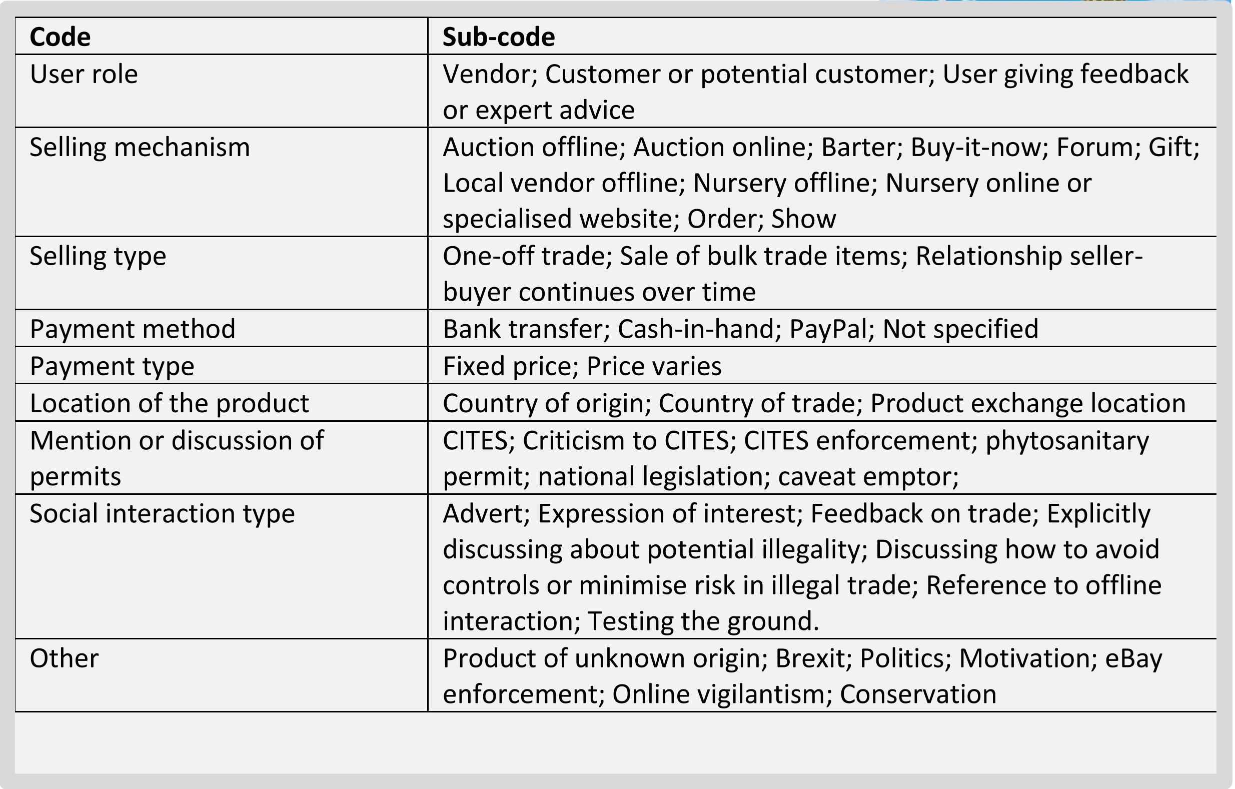
Approach - Criminology Analysis

- Subject matter expert >> Lexicon (species, trade jargon)
  - Target species (Ariocarpus, Euphorbia, Saussurea)
  - Search keywords, Plant-based lexicon
- Criminologist >> Manual browsing and behaviour coding
  - Posts about illegal trades and CITES permits
  - Subcultural examples of relevant forum user behaviour
  - Coding of posts using NVivo >> Update target suspect lists
- Criminologist >> Analysis of information extraction results
  - Refinement of lists of relevant POLE entities
  - POLE (People, Object, Location and Events) >> UK law enforcement & Home Office

Images: Copyright Royal Botanic Gardens, Kew

Approach - Criminology Analysis

- Subject matter exper
  - Target species (Arioca
  - Search keywords, Plai
- Criminologist >> Mar
  - Posts about illegal tra
  - Subcultural examples
  - Coding of posts using
- Criminologist >> Anal
  - Refinement of lists of
  - POLE (People, Object,



Code	Sub-code
User role	Vendor; Customer or potential customer; User giving feedback or expert advice
Selling mechanism	Auction offline; Auction online; Barter; Buy-it-now; Forum; Gift; Local vendor offline; Nursery offline; Nursery online or specialised website; Order; Show
Selling type	One-off trade; Sale of bulk trade items; Relationship seller-buyer continues over time
Payment method	Bank transfer; Cash-in-hand; PayPal; Not specified
Payment type	Fixed price; Price varies
Location of the product	Country of origin; Country of trade; Product exchange location
Mention or discussion of permits	CITES; Criticism to CITES; CITES enforcement; phytosanitary permit; national legislation; caveat emptor;
Social interaction type	Advert; Expression of interest; Feedback on trade; Explicitly discussing about potential illegality; Discussing how to avoid controls or minimise risk in illegal trade; Reference to offline interaction; Testing the ground.
Other	Product of unknown origin; Brexit; Politics; Motivation; eBay enforcement; Online vigilantism; Conservation

Images: Copyright Royal Botanic Gardens, Kew

Information Extraction

- Search for relevant forums
  - Microsoft Bing Search >> Discover forums/marketplaces trading species
- Crawl forum threads
  - DARPA MEMEX Undercrawler >> Crawl HTML pages from forum
- Parse posts
  - Python HTML Parser (templates for HTML tags) >> Dataset [thread, author, text, timestamp]
  - Stanford CoreNLP >> Tokenized Text, Named Entity (NE) annotation to n-gram phrases
  - Author, thread and post >> Directed Acyclic Graph (DAG) of conversations
- Information extraction model
  - Posts (Text) >> Scikit-learn LDA Topic Model >> Topics (each containing target suspect)
  - Posts (NE, DAG) >> Graph Walk >> NetworkX & Matplotlib >> Viz (target suspect as root)

Approach - Information Extraction

- Search for rele
  - Microsoft Bi
- Crawl forum t
  - DARPA Unde
- Parse posts
  - Python HTM
  - Stanford Co
  - Author, threa
- Information ex
  - Posts (Text) >> Scikit-learn LDA Topic Model >> Topics (each containing target suspect)
  - Posts (NE, DAG) >> Graph Walk >> NetworkX & Matplotlib >> Viz (target suspect as root)

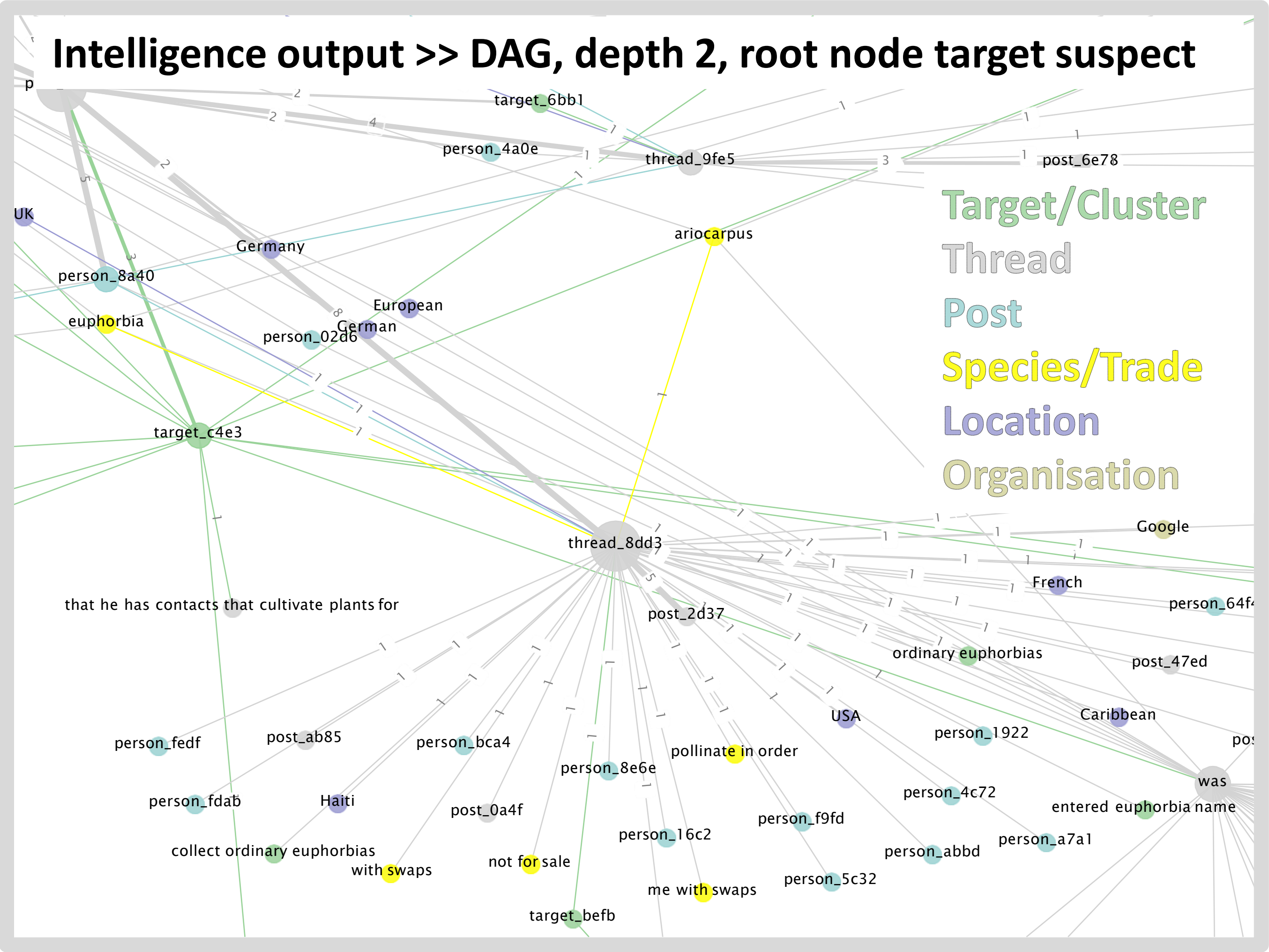
Intelligence output >> list of 10 topics, 20 phrases per topic

topic_1: ["**greenfingers123**", "**Ariocarpus**", "seeds", "Here", "http www", "www", "old", "picture", "http", "years", "**markthegardener**", "**Dino54**", "retusus", "list", "This", "10", "**eBay**", "**Ariocarpus retusus**", "**greenfingers123 Here**", "Smith"]

topic_2: ["cacti", "And", "satin", "crash satin", "crash", "like", "Cactisaurus", "looks", "probably", "**greenfingers123**", " **Plantnursery**", "cacti cacti", "sell", "legal", "good", "looks like", "seeds", "hybrid", "live", "bought"]

Approach

- Search for relev
  - Microsoft Bing
- Crawl forum th
  - DARPA Under
- Parse posts
  - Python HTML
    timestamp]
  - Stanford Core
  - Author, threa
- Information ext
  - Posts (Text) >>
  - Posts (NE, DA



Intelligence output >> DAG, depth 2, root node target suspect

Experiments

- Research question
  - Given a known target suspect, can information extraction methods discover connected POLE entities useful for a criminology investigation without information overload
- Experiment setup
  - Participants: 1 criminologist, 1 computer scientist, 1 subject matter expert
  - Experiment 1 - Ariocarpis (forums), 1 week of analysis
  - Experiment 2 - Euphorbia and Saussurea (forums & marketplaces), 1 week of analysis
  - Intelligence outputs focus on POLE (People, Object, Location and Events)
- Dataset from experiments
  - 9 websites crawled
  - 13,697 posts by 4,009 authors in 1,826 forum threads
  - Posts were aged from 2006 to 2019

Experiments

- Ground truth
  - 25 hours of criminology analysis >> 4 or 5 target suspects per species + POLE entities
- Evaluation
  - For each target suspect execute (a) LDA topic model, and (b) NE directed graph viz
  - Limit size of intelligence outputs to something a criminologist can easily review (400 entries)
  - Evaluate recall of ground truth connected entities per target suspect

		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

NE graph
clearly outperforms
Topic models

Discussion

- Target suspects were mostly found in the long tail
  - # posts in dataset >> over 13,000
  - # posts with target suspect >> typically about 100 posts (lowest 5, highest 2,000)
  - Connected POLE entities >> max 10 people, max 5 species/locations/organisations
- Named Entity directed graph visualization clearly better than topic models
  - Not perfect (recall ranged from 0.24 to 0.85)
  - NE graph recall was best for people and locations
- Socio-technical AI methodology
  - Scale up analysis >> NE graphs to discover potential suspects and POLE entities
  - Human context checking >> Criminology analysis to check automated results, using manual browse and human judgement to look at the context behind connections

Discussion

- NE graph evidence ready for a court of law?
  - Not on its own
  - It could be used to support target-focussed evidence packages though, with input from criminologist and maybe some relevant out of band corroborating evidence
- Next steps for NE graph analysis
  - Sub-graph classification, Partial graph matching >> Automate behaviour classification
- Collective intelligence sharing?
  - Data lakes are trending with cloud providers like Amazon
  - Centralizing vast sets of structured and unstructured data into a single multi-purpose searchable repository for various intelligence analysis tasks
  - NE graphs, rich with metadata and connected entities, are well suited to feed a data lake

Questions?

Dr Stuart E. Middleton

University of Southampton, Electronics and Computer Science

email: sem03@soton.ac.uk

web: www.ecs.soton.ac.uk/people/sem

twitter:@stuart_e_middle

CYShadowWatch: https://www.ecs.soton.ac.uk/research/projects/1019

FloraGuard: http://floraguard.org/

UNIVERSITY OF Southampton

Royal Botanic Gardens Kew

Border Force

FloraGuard
Tackling the Illegal Trade
in Endangered Plants

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

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).

Defence and Security Accelerator

NCA National Crime Agency