Institution: Lancaster University

Unit of Assessment: UoA 11: Computer Science and Informatics

Title of case study: Investigative toolkit and associated techniques to support online child protection based on digital persona analysis

1. Summary of the impact (indicative maximum 100 words)

Online child protection is one of the key concerns of our time. But there are huge problems in enforcing the law in this area because enforcement agencies have manually to analyse vast quantities of online data, resulting in significant backlogs. In response, we have pioneered the field of digital persona analysis, which significantly automates the process of identifying online sexual predators who masquerade as children. This work is underpinned by internationally-leading research that combines authorship attribution techniques with corpus-based natural language analysis. The results are impressive, yielding highly accurate persona analysis in the face of huge and noisy data sets, and in the face of deliberate attempts to deceive.
The impacts are both wide and significant:

- Impact on law enforcement agencies – we have provided sophisticated analysis tools which are being used by law enforcement agencies nationally and internationally;
- On the economy – we have fostered the creation of a spin-out company, Isis Forensics;
- On education – in close collaboration with the secondary education sector we have developed educational programmes on online protection for children at Key Stages 2-5;
- On public policy wrt Internet governance – we have made significant contributions to the debate on online protection, coupled with public awareness measures.

2. Underpinning research (indicative maximum 500 words)

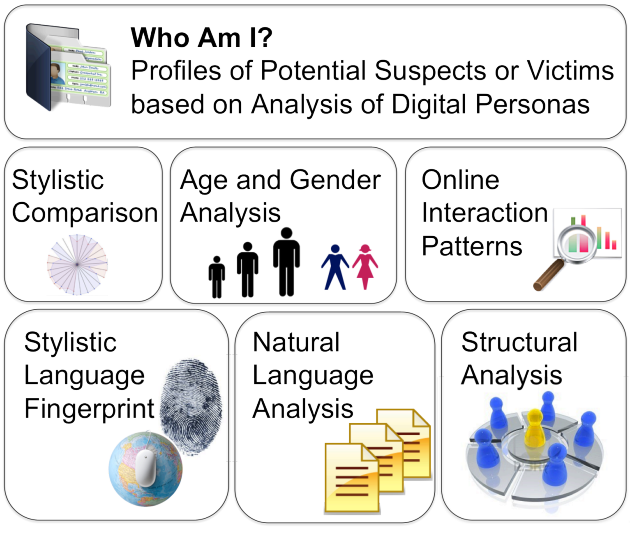
Context and research problem. Online child protection is a key concern of our time. Online social networks are extremely popular with children, and the ubiquity and accessibility of these networks give child sex offenders easy access to potential victims. The scale of the problem is huge:

- 50% of teenagers report having given out personal information online, and 10% have engaged in physical meetings with strangers following online interactions (EU Kids Online project – summarising over 400 studies across Europe);
- 13% of children in London report occasions on which they believe they had been talking online to an adult posing as a child (London Metropolitan Police).

In attempting to combat online sex offenders, the cognitive load on law enforcement investigators is incapacitating because analysis of data from social networks is currently predominantly *manual*, which simply does not scale. This is because existing tools are primitive, typically offering only data extraction and simple keyword search capabilities. There is therefore an urgent need for more sophisticated tools that can efficiently carry out higher-level analyses of vast quantities of online data and report to investigators at the level of personas and behaviours.

Description of the underpinning research and findings. The underpinning research stems from the EPSRC/ESRC Isis Project (2008-11), which involved Rashid, Rayson and Walkerdine; it also builds on earlier research in natural language processing by Rayson and Garside (1997-2003).
The Isis Project focused on the problem of criminals hiding behind multiple identities (including, crucially, adults posing as children) [1, 2, 3]. The key tangible output of the project was an analysis suite called the *Isis Toolkit* (see diagram). This was underpinned by 3 key research contributions:

- *Algorithms to establish a stylistic language fingerprint of potential suspects or victims:*



these fingerprints can be overlaid to determine whether one person is hiding behind a single persona or if multiple persons are sharing a single persona;

- *Algorithms to determine the age and gender of a person behind a digital persona: this is achieved by synthesising the stylistic language fingerprint with additional markers extracted using natural language analysis techniques;*
- *Algorithms to determine online interaction patterns: this involves analysing conversational structures and language patterns (e.g. signature moves when signing off from a conversation, or frequently used words and phrases) to determine a specific persona's identifying characteristics.*

The research builds on internationally-leading work on corpus comparison techniques, using statistical natural language analysis [4, 5, 6]. At the heart of the research is a semantic analysis approach that categorises keywords based on contextual information. In more detail, the approach involves integrating statistically sophisticated but knowledge-poor techniques from authorship attribution work with linguistically-informed methods from corpus-based natural language analysis; and also in combining the macro level (models of language varieties) with the micro level (models of individual's use of language). Our approach operates satisfactorily in the face of noisy language data, and performs well in the face of masquerading or similarly deceptive behaviour that an individual might assume in an attempt to hide his or her identity.

Research outcomes. The research has transformed the field of online cybercrime and, in doing so, has created a new field of study [1], that of online digital persona analysis. The results from the research are impressive: the Isis Toolkit can detect masquerading tactics with a high degree of accuracy: for example, detecting when an adult is masquerading as a child with an accuracy of 94% (compared to children participating in controlled experiments) [2].

3. References to the research (indicative maximum of six references)

Key references:
[1] Rashid, A., Baron, A., Rayson, P., May-Chahal, C., Greenwood, P., Walkerdine, J. (2013). Who Am I? Analyzing Digital Personas in Cybercrime Investigations. IEEE Computer 46(4): 54-61.
[2] May-Chahal, C., Mason, C., Rashid, A., Greenwood, P., Walkerdine, J., Rayson, P. (2012). Safeguarding Cyborg Childhoods: Incorporating the On/Offline Behaviour of Children into Everyday Social Work Practices. British Journal of Social Work.
[3] Rashid, A., Greenwood, P., Walkerdine, J., Baron, J., Rayson, P. (2012) Technological Solutions to Offending. In: Understanding and Preventing Online Sexual Exploitation of Children. Quayle, E., Ribisl, K. (Eds). Willan. 228-243.

Other references:
[4] Rayson, P. (2003) Matrix: A Statistical Method and Software Tool for Linguistic Analysis through Corpus Comparison, Ph.D. thesis, Lancaster University.
[5] Rayson, P., Leech, G., Hodges, M. (1997). Social Differentiation in the se of English Vocabulary: Some Analyses of the Conversational Component of the British National Corpus, Intl. Journal of Corpus Linguistics. 2(1): 133-152.
[6] Rayson, P., Garside, R. (2000). Comparing Corpora using Frequency Profiling, Proc. of the Workshop on Comparing Corpora, 9: 1-6.

4. Details of the impact (indicative maximum 750 words)

Pathways to impact. The research led to impact via the following three routes:

1. Collaborative research. The underpinning research was collaborative and profoundly cross-disciplinary, covering areas including computer security, linguistic analysis, HCI, law/ethics and social science. The underlying natural language processing results stemmed from a long-running, very successful, cross-disciplinary collaboration between Computer Science and Linguistics at Lancaster (under the auspices of the UCREL Research Centre).

2. Research with user communities. The research involved real-world deployments with law enforcement agencies, and direct engagement with schools. In particular, live trials were held with the Kent, Lancashire and Merseyside Police Forces (2010-2012), with the London Metropolitan Police Force (2013), and in collaboration with the Child Exploitation and Online Protection Centre (CEOP). In addition, teaching sessions were organised at the Queen Elizabeth School, Kirkby

Lonsdale (2009-2010) and at the Lancaster Girls Grammar School (2011) to help children understand the masquerading tactics utilised by online sex offenders.

3. Commercialisation. A spin-out company was created with the support of the KE services that are embedded in the School of Computing and Communications (in InfoLab).

Areas of Impact. The impact arising from this work has been wide and highly significant. The following four areas of impact are identified:

1. Impact on the work of law enforcement agencies (*public service impact*). Our research has pioneered a new approach to online child protection through digital persona analysis. Beneficiaries: The primary beneficiary so far is the *Canadian Royal Mounted Police* who have licensed the Isis Toolkit, which they regard as an "operational necessity" [A], for use across Canada. Other, equally large-scale, agreements are being negotiated but are not yet confirmed as of the REF census date. Secondary beneficiaries are police forces across the UK who, along with CEOP, have participated in highly successful live trials which demonstrate that the Isis Toolkit works accurately on real data sets and significantly reduces manual analysis time: "allows all victims to be identified and therefore a full range of offending to be investigated", "provides the ability to focus analysis on specific information [and] allows investigations to be more focused and therefore potential victims of grooming or contact abuse to be identified more easily", "the only other option is manual analysis which would be much slower" (all quotes from the evaluation of the aforementioned live trials). This work has transformed strategic thinking on child online protection. Reach: National and international. Significance: Hugely significant in delivering a completely new and highly effective approach to online child protection.

2. Creation of a spin-out company – Isis Forensics (*economic impact*). A spin-out company called Isis Forensics has been created, its purpose being to license the Isis Toolkit and exploit the associated IPR [B, C]. The company employs 4 FTE, has received venture capital of £400,000, and is seen as having very significant growth potential. It has signed a major deal with the Canadian Royal Mounted Police as discussed above. The company has recently diversified to support 'insight extraction' in areas such as brand management and social media analytics (this is achieved by generalising the core Isis technology and making it available via a cloud-enabled API). Reach: Local. Significance: Part of a series of commercialisation activities coming out of InfoLab and contributing to regional development and regeneration.

3. Impact on educational programmes for children and young people (*public service impact*). Collaboration with local schools has led to significant impact at various levels [D]. Beneficiaries: The main beneficiaries are the two above-mentioned schools with whom we have worked to develop and deliver Internet safety lessons to over 500 students. These lessons include Turing-test like sessions in which children chat electronically with people behind screens, half of whom are children and the other half adults pretending to be children (thus allowing the children to understand the masquerading tactics typically utilised by offenders online). This activity has led to a strong collaboration with teachers at the Queen Elizabeth School to develop comprehensive lesson plans on the broader topic of e-safety for Key Stages 2-5, which we claim are nationally leading [E]. These are now being rolled out across the region through the South Lakes Teaching School Alliance (SLTSA). Isis Forensics has also released a free iTunes app, 'ChildDefence', which empowers children to protect themselves online, thereby extending the reach of the educational work globally. Reach: Local, regional and international for each of the above initiatives resp. Significance: Education is at the core of child protection online.

4. Impact on Internet governance (*public policy impact*). The work has contributed strongly to the national and international debate around Internet governance and online safety. Beneficiaries: Policy makers are the prime beneficiaries in this area. A policy paper was prepared for the BCS and presented to Alun Michael, MP (2009), and subsequently selected as the (single) UK contribution to the 2009 and 2010 Internet Governance Forums (in Sharm-Al-Sheikh and Vilnius respectively) [F, G]. Written evidence was also provided to the Commons Select Committee on Education (2010) [G]. The research also contributed to the *Proposal for a Directive of the European Parliament and of the Council on combating the sexual abuse, sexual exploitation of children and child pornography, repealing Framework Decision 2004/68/JHA* (COM/2010/0094). The research also provided a case study in a report requested by the European Parliament's

Committee on Gender Equality proposing the extension of the Isis Toolkit to assist in the detection and management of cyber coercion and rape of women and girls. Secondary beneficiaries are the general public who benefit from our significant efforts to raise the awareness of online protection issues through the media [H]. Reach: National and international. Significance: It is crucially important in this area to impact both public policy and to inform the general public. As a final accolade, the work was highlighted as one of the 100 big ideas of the future by UUK and RCUK in 2011 [I].

5. Sources to corroborate the impact (indicative maximum of 10 references)

[A] The Technology Manager at the National Child Exploitation Co-ordination Centre of the Royal Canadian Mounted Police can confirm the licensing agreement and the impact on child protection police work across Canada.
[B] http://www.isis-forensics.com has detail on Isis Forensics, and its products and services.
[C] The CEO of Isis Forensics can confirm the business impact of the research on the company.
[D] The Head of School of QeS has provided evidence of the impact of the research on their school and on their online safety curriculum, and can also be contacted to confirm this impact.
[E] The detailed lesson plans for eSafety developed by QeS and Lancaster University are available at: http://scc.lancs.ac.uk/onlinesafetylessons.
[F] http://www.bcs.org/upload/pdf/project-isis.pdf has details of the UK input to the Internet Governance Forum (Protecting Children in Online Social Networks).
[G] http://www.publications.parliament.uk/pa/cm201012/cmselect/cmeduc/writev/1514/cp18.htm details written evidence on 'The Child Protection System in England' stemming from the Isis Project and submitted to the Commons Select Committee on Education.
[H] Media coverage is summarised on these sites: http://www.isis-forensics.com/resources/ (incl. a link to a BBC News item) and http://www.comp.lancs.ac.uk/isis/index_files/Press.htm.
[I] The work was reported in the UUK/ RCUK report on 'Big Ideas of the Future' (p68): http://www.rcuk.ac.uk/Publications/reports/Pages/BigIdeas.aspx.