Who am I? Analysing Digital Personas in Cybercrime Investigations

January 23, 2013

Awais Rashid, Alistair Baron, Paul Rayson, Corinne May-Chahal *(Lancaster University, UK)*
Phil Greenwood, James Walkerdine *(Isis Forensics Ltd)*

Abstract

Online cybercrime activities often involve criminals hiding behind multiple identities (so-called digital personas). Unraveling these multiple digital personas is a non-trivial problem owing to the large amounts of text communicated in online social media and the large numbers of digital personas involved. The cognitive load for cybercrime investigators is immense – existing tools lack the sophisticated capabilities required to analyse digital personas in order to provide investigators with clues to the identity of the individual or group hiding behind one or more personas. In this article, we present the Isis toolkit which addresses this very problem.

1 Introduction

Digital communities not only bring people closer together but also, inadvertently, provide criminals with new ways to access potential vicitms online. Digital personas play a key role in criminal tactics in online social media. One criminal may hide behind multiple digital personas or a single persona may be shared by a criminal group when engaging with potential victims. Furthermore, the fluid nature of identity on online social media means that criminals can disguise themselves with relative ease to gain the trust of potential victims. Examples of such criminal exploitation of digital personas include:

- *Child sex offenders* masquerading as young persons to gain the trust of their victims. An offender may use multiple personas over the course of an interaction (introducing himself/herself as a young person and then introducing another persona, e.g., that of an older relative). Alternatively, a single persona may be shared by an offender group so that a victim is groomed by multiple people over a period of time [1].
- *Romance scam operators* using digital personas with appropriate age and gender to engage with multiple victims in online dating sites, gaining their trust and exploiting them for financial gain [2].
- *Radicalisation of youth* in online forums through persuasive messaging [3]. Multiple digital personas are used as a tactic at times. For instance, one persona is used to vigorously support a radical cause, followed by silence for a few days and then a different persona is used to claim that the original protagonist has left to fight for the cause.

Effective policing of such environments is, however, extremely challenging – a vast amount of information is communicated within online social media making its manual analysis difficult nigh impossible. Consequently, law enforcement agencies face huge backlogs of online communication data analysis during cybercrime investigations – backlogs of 6-9 months being commonplace. Even though a range of commercial tools such as EnCase[1] and Internet Evidence Finder [2] exist to assist such investigations, they are mainly focused on data extraction. Any analysis of the data is left to the investigator supported only by simple techniques such as keyword-based searches or phrase detection based on user-defined lists. Such techniques do not scale. Nor do they include models of deceptive behaviour or sharing of online personas. It is not uncommon for investigators to extract data from hard disks or mobile phones using a tool such as EnCase and then manually

[1] http://www.guidancesoftware.com/encase-forensic.htm
[2] http://www.magnetforensics.com/products/internet-evidence-finder/

read it to identify when an offence may have occurred and make a value judgement on whether one or more digital personas were used as part of the offender's tactics. Given the large amounts of text and number of online participants during such investigations, it is virtually impossible for the investigator to analyse all digital personas involved – the cognitive load is immense.

Related Work. Relevant research on mining and analysis of information from online social media has mainly focused on extracting key messages prevalent in such media. Davulcu et al. [4] focus on detecting sentiment markers that indicate radicalisation and counter messages in online forums. Diesner and Carley [5] have shown how common word use across actors can be used to derive knowledge about the structure of covert social networks and their weak points. Other recent work has shown that clustering of individuals in online communities is not driven by homophily [6] and that it is possible to gain deeper insights through analysis of latent structures in online conversations [7]. In recent years, techniques from the fields of corpus-based natural language processing and text mining have been applied to these problems. Corpus analysis, particularly at the semantic level, can provide a way of describing the key features in extremist discourse [8] and authorship attribution enables automatic identification of a given writer or speaker [9]. Analysis of digital personas and the deception tactics inherent therein have not been considered to date. In this article, we present the Isis toolkit[3] which addresses this particular challenge by enabling efficient and sophisticated analysis of digital personas in large-scale online textual communications. Our approach complements recent research such as that by Afroz et al. [10], which highlights the difficulty in identifying authorship when language is intentionally obfuscated and that of Narayanan [11] which shows that it is viable to automatically predict text authorship on a large-scale. Our work shows that it is possible to predict key attributes of a persona (i.e. age and gender) with acceptable accuracy regardless of whether the author is obfuscating the language.

2 The Isis Toolkit

As shown in Figure 1, in the Isis toolkit, we combine two families of techniques which make use of statistical methods from corpus-based natural language processing and authorship attribution. Analysis techniques from corpus linguistics and natural language processing, such as keyword profiling, offer the capability to compare word frequencies. In our previous work [12] we have extended this approach to extract key grammatical categories (equating to features of style) and key semantic fields (showing key concepts). These techniques rely on large representative samples of writing or transcribed speech (corpora) for training and reference comparison, have high accuracy and are designed to be robust across various types of text. In conjunction, tools and methods from the field of authorship attribution permit us to narrow the focus from language varieties down to the individual writer in order to utilise a stylistic fingerprint for the author. In the past, authorship attribution techniques were mainly applied to ascertain authorship of historical texts. Recently, more robust evaluation techniques have been developed and authorship attribution methods have been applied to known problems with standard benchmark data [9]. Specific challenges that we faced in implementing the Isis toolkit were to integrate the statistically sophisticated but knowledge-poor techniques from authorship attribution with linguistically-informed methods from corpus-based natural language analysis, and combine the macro level (models of language varieties) with the micro level (models of individual's use of language). Additionally, these methods have to operate on small quantities of noisy language data observed in online social networks, and to deal with masquerading or similarly deceptive behaviour where an individual may be attempting to hide his or her identity.

The novel investigative features of the Isis toolkit are as follows:

- **Establish a stylistic language "fingerprint"** of potential suspects or victims. These fingerprints can then be overlaid on each other and compared to study whether one person might be hiding behind a single persona or if a single persona is being shared by multiple people.
- **Establish the age and gender of the person behind a digital persona** – this is achieved by synthesising the stylistic "fingerprint" with additional markers extracted using our natural language analysis engine. Furthermore, the toolkit can detect masquerading tactics with a high accuracy, for instance, detecting when an adult is masquerading as a child.
- **Establish online interaction patterns of particular digital personas** – analysis of the conversation structure as well as the language used therein is used to determine key

[3] `http://www.comp.lancs.ac.uk/isis/`

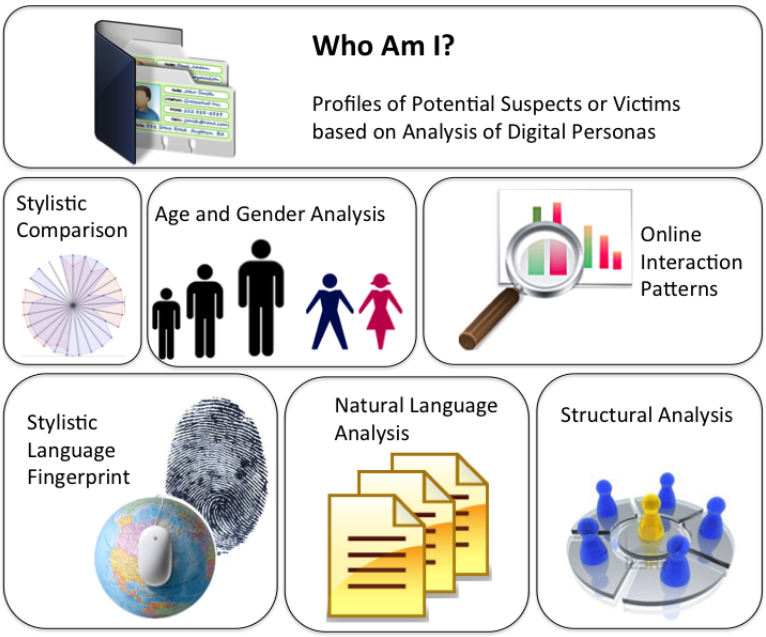

Figure 1: The Isis Toolkit

characteristics of a specific persona, e.g., signature moves when signing off from a conversation or specific words and phrases used frequently. The toolkit also enables analysis of a persona's behaviour, e.g., identifying when a participant is typically active – not only within an average 24 hour period but also in terms of day of the week and if a persona becomes increasingly sexual or aggressive over a period of days or weeks.

Using the above features, the toolkit supports building up a profile of potential suspects or victims (the techniques are equally applicable for victim identification), enabling investigators to gain a better understanding of the digital personas involved and also potentially providing clues to an individual's or group's identity in the physical world.

2.1 Stylistic Language "Fingerprint"

Within the Isis toolkit a wide range of subtle language traits are observed and scrutinised to assist in authorship analysis; examples include the proportions of punctuation characters, the use of emoticons and vocabulary measures. These language traits are used to build a stylistic "fingerprint" which can, in turn, be used to represent the language of a particular user, set of users or collection of texts.

Metrics used to construct the stylistic "fingerprint" range from simple counts, such as the number of exclamation marks present, to more complicated measures, such as Type Token Ratio (a vocabulary indicator). The calculation of each metric takes into account text length, allowing for a mixture of sources to be combined where appropriate; for example, email texts will generally be longer than chat room texts. Through this process a list of metric scores can be assigned to a single text or collection of texts; examples include the collated messages from a single chatroom user or a sample of texts chosen to represent the language of adult female chatroom users. The metric scores produced for two or more text collections can then be compared to indicate how likely the sources of the texts overlap, or are written by people of a similar age and gender. The Isis toolkit uses the metric scores in two ways; firstly to assist with the automatic age and gender analysis, and secondly to provide a visual impression of how close two text sources are with regards to their linguistic style. We discuss these next.

2.2 Age and Gender Analysis

This analysis is performed in four steps. The first three steps utilise our natural language analysis engine while the fourth combines the knowledge thus extracted with the metric scores from the stylistic "fingerprint".

Step 1 An incoming text sample is tokenised and each word is tagged with a part-of-speech (POS) label (noun, verb, adverb, adjective etc).

Step 2 Each word or phrase within the text is assigned to one semantic field (general conceptual labels such as finance, warfare, government, sports etc). Both of these steps rely on a set of hybrid techniques to select the most likely tag in each context.

Step 3 The features (i.e., the language styles used) at the word-level, POS-level and semantic field-level are counted.

Step 4 Each level is compared to standard reference datasets that have previously been processed through the same pipeline. In the case of gender, we prepare two reference datasets, one for males and one for females. A distance metric then calculates the similarity between the incoming text sample and each of the two reference corpora for each of the three levels. Metric scores from the stylistic "fingerprint" are produced for each reference dataset i.e. for different gender groups and used as features for training a text classifier. There are various machine learning algorithms and methods for feature extraction which are used for a range of text classification purposes. We use logistic regression with the metric scores as features to classify a given text into gender groups. Probabilities are produced which indicate the likelihood that the given text should be classified as each gender group. These are then combined with the word, POS and semantic field analyses – a weighted combined score is derived and the system then assigns a value for how likely the incoming text is to be written by a male or female author. Similarly, reference datasets by age range can be prepared and compared in the same manner. It is possible to focus in on smaller age ranges by preparing specific reference datasets. This allows the toolkit to present an overview of the likelihood that a text is written by an adult or a child, and then drill down to results for more precise age ranges. The Isis toolkit provides this information as a decision tree that a law enforcement officer can consult and interact with as shown in Figure 2(a).

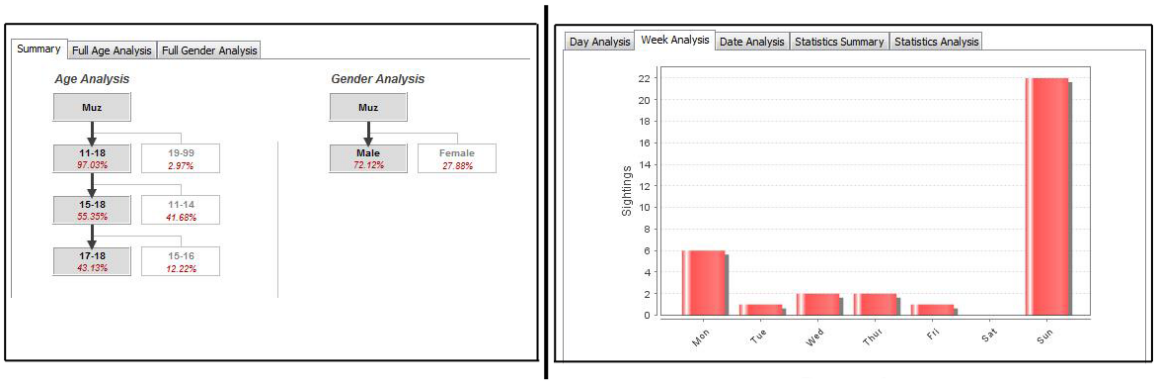

Figure 2: (a) Age and gender decision trees (b) Online/offline time analysis

2.3 Comparing Stylistic "Fingerprints"

Whilst an automatic prediction of age and gender is useful in many cases, it may also be useful for an investigator to be able to visualise language differences and similarities. The metric scores offer the ability to plot stylistic differences on a graph. Given two lists of metric scores[4], each score is divided by the maximum of the two scores for that metric. Hence one adjusted score for each metric is now 1 and the other is a fraction of that (between 0 and 1). The adjusted scores are then multiplied by metric weights derived through machine learning, which can be specialised for the text comparisons being performed, e.g., comparing a user's text against age group datasets. Radar plots of the adjusted and weighted metric scores can then be used to visually represent language style "fingerprints". When the two plots are overlaid, the similarity or difference between the two text sources represented is clear to see, with substantial overlap indicating that the language style is similar and little overlap indicating contrasting language styles.

As well as displaying how close a user's text is to a given age and gender dataset, the fingerprinting method described can also be used to compare the text from two personas, to establish whether they are the same individual, or to compare texts from one persona at different times to explore if

[4]More than two lists can be compared on the same plot, but we only discuss comparison of two lists here.

the persona is shared by multiple individuals. In order to describe this process and to demonstrate the fingerprint comparison technique, Figure 3 shows the language style fingerprint comparison of a previously unseen text (the messages of a single user in one chat session) against the collated texts of four individuals (for each individual the messages are taken from six chat sessions). A larger overlap (shaded purple in Figure 3) of the fingerprints indicates that the language style of the new text is similar to that of the previous texts for an individual, hence the new text is more likely to be from that user. In Figure 3, the overlap is most marked for Individual 1 (top left), so a judgment could be made that the new collection of chatroom messages is likely to be from that user – in this case that judgment would be correct (the fingerprints are from real chat sessions conducted in a simulated real-life cybercrime scenario).

Figure 3: Comparing language style fingerprints for a new text against four individuals' fingerprints.

2.4 Online Interaction Patterns

The toolkit also supports identification of patterns typical to a persona's online presence and its interaction with other participants. This is achieved through structural analysis of the text, which extracts details such as the user names of those who are participating in the chat, or date and time information that can be used to model the conversation flow to identify patterns and trends over time. All conversation logs that are entered into the toolkit are converted into a generic format. A key aspect of this is breaking down and modelling the log in terms of the participants and their respective activity (posting messages, sharing links, leaving/joining, etc). Once this model

has been built it is possible for the toolkit to quickly analyse and present intelligence about a particular participant. This can include analysis of his or her language use, for example, stylistic characteristics such as keywords, names, topics, etc. that are frequently used or identifying patterns of online/offline times. Use of semantic categorisation allows parts of a conversation to be classified based on their meaning (for example, whether sexual or aggressive in nature). By applying these techniques to the model of a participant's conversation it is possible to view any trends that may occur over the duration of the conversation, for example, to help determine if a conversation is becoming increasingly sexualised.

Analysis of online-offline windows (see Figure 2(b)) becomes particularly relevant in online social media where many participants are active. By cross-referencing different participant models the toolkit is able to show when participants are online together and the content of their conversation at those times. This can be used to make inferences about who tends to communicate with whom (and about what). It can also help determine if a suspect is switching between multiple user accounts (personas) - a trend that is frequently seen when online personas are exploited for criminal purposes.

2.5 Profiling Cyber Criminals and Victims

The various analysis techniques discussed above combine to form a key feature of the toolkit – the ability to generate identity profiles of specific digital personas. The toolkit is able to automatically create profiles for a specified digital persona, drawing upon the conversations in which it has participated to produce an overall analysis of its online activity, language and identity characteristics. These profiles can provide investigators with additional intelligence about trends and characteristics not immediately apparent to the human eye.

The generated profile is built from a number of elements, including:

- *Language usage.* A model of the persona's language use within conversations highlighting characteristics such as people/place names, dates/times, frequently used words/phrases, aggressive/sexual content, email addresses/URLs, as well as non-dictionary words which may indicate an attempt at disguising what is being discussed or represent unique jargon used within that domain (of which an investigator may or may not be aware).
- *Age/Gender analysis.* Utilising the decision tree to provide an inferred estimation of the age and gender of the person behind the persona. By default investigators are provided with a summary view which presents the strongest path through the tree, but they are also able to view the full tree allowing them to examine the decisions the toolkit made at all points should the certainty of the decision not be clear cut.
- *Online activity.* An analysis of the persona's overall online activity, highlighting when it has appeared online within relevant conversations. This analysis can take many forms including indicating when a persona is most likely to be online over a 24 hour window and on which days during the week.

3 Differentiating Genuine Personas from Deceptive Behaviour: The Isis Toolkit in Practice

We have used the toolkit on reference datasets and in live environments to test its effectiveness in correctly detecting attributes of an individual behind a persona. Here we discuss insights from two such tests - one where no deception is intended and the other where an individual is using deceptive tactics.

3.1 Classifying age and gender of genuine personas

For this test we used the British National Corpus (BNC), which is a reference dataset with a 100 million words of written and spoken language and represents a wide cross-section of British English. We utilised the portion of BNC (1,684 people, which constitutes 10% of the entire collection) where meta-data about an individual, including age and gender, was available. We used "leave one out cross validation" whereby we trained our system using the texts from all 1,684 individuals except the text from the individual being used as a test subject, i.e., the one whose age and gender was being classified. We repeated this classification for all 1,684 individuals as test subjects. For each classification, our toolkit provides probabilities that the individual belongs to a specific age band; for example an individual may be predicted to be 11–18 with a probability of 74% and over 18 with a probability of 26%. The prediction then moves down a level, e.g., predicted to be between 11–14

with a probability of 49% and 15–18 with a probability of 25%, and so on with gender probabilities also calculated. At each decision point (as shown in the age and gender classification decision trees in Figure 2(a)) the age group or gender with the highest probability is taken as the prediction. A probability threshold is also used to decide whether a prediction is used. If the highest probability is below this threshold then no prediction is made, i.e. the age or gender is marked as 'unknown'. Precision and recall are used to measure the ability of our algorithms to correctly classify the age and gender of each individual in the test set; precision being the proportion of predictions made (i.e. not 'unknown') that are correct according to the metadata, recall being the proportion of individuals tested for which the correct prediction is made.

Figure 4(a&b) gives the recall and precision values obtained for age classification at different specificity levels, which are outlined in Figure 4(c). Recall is 72.15% and precision is 72.24% at Level 1, that is distinguishing between children (11–18) and adults (over 18). This is based on a probability threshold of 50%. By increasing the threshold, greater precision can be achieved at the cost of fewer classification decisions being made (more 'unknowns' are returned). With a higher threshold of 80%, precision of adult/child classification increases slightly to 77.35%, but recall drops to 59.20%. Naturally, the precision and recall drops at higher levels of the age decision tree (with more specific age ranges).

For gender classification (Figure 4(d)) with a threshold of 50%, recall is 66.74% and precision is 66.86%. Again, a higher threshold can be set; increasing it to 80% improves precision to 71.07%, but recall drops to 56.08%.

3.2 Classifying Deceptive Personas

Testing our toolkit on the BNC data allows us to determine the accuracy of our algorithms when individuals are not being deceptive. However, given our focus on detecting misuse of digital personas, we have tested the toolkit on detecting masquerading behaviour – when an individual hides behind a false persona, for example, pretending to be a child. For our purposes, we set up a "live" environment, similar in nature to a Turing Test, in two schools. Children and young people (11–18 years old) chatted online with 10 individuals behind the scenes in sessions divided by age group. In each session half of the individuals behind the scenes were children/young people of the same age as the chat participant while the other half were masquerading behind personas purporting to be of that age. We then employed a similar evaluation process as for the BNC dataset to test the effectiveness of our toolkit in classifying whether the people behind the scenes were children/young people or masqeurading as children/young people.

For deciding if an individual is a child or an adult masquerading as a child, the age classification algorithm achieves precision and recall at 84.29% with a probability threshold of 50%. The precision can be increased with a higher threshold; at 80% threshold, precision rises to 93.18% but recall drops to 58.57% with fewer predictions being made. These results obtained through our toolkit are in stark contrast to the accuracy of the children's responses, with only 18% of children across the year groups able to correctly identify if they were chatting to an adult or a child. For gender, with a 50% threshold, precision is 80.6% while recall is 77.14%. Increasing the threshold to 80% improves precision to 84.09% but recall drops to 52.86%. These results are in contrast to the children correctly identifying the gender of the person with whom they were chatting in 58.8% of the cases.

4 Conclusion

The "connectedness" afforded by online social media enables individuals and groups from various geographical, cultural and socio-economic backgrounds to interact and share experiences. However, the very nature of identity in online social media – a fluid and dynamic notion that can be created, adapted and discarded with ease – makes such identities prone to misuse. Exploitation of digital personas has become an integral part of the tactics used by cyber criminals. This new digital world and these sophisticated criminal tactics call for new tools to aid the investigators of online crime. Our experience with the Isis toolkit demonstrates that it is possible to detect key characteristics of individuals or groups behind digital personas with a high degree of accuracy by combining techniques from corpus-based natural language analysis with those from authorship attribution. In fact, our analysis performs better when deception tactics are being used - hence demonstrating the effectiveness of digital persona analysis as a tool in the investigator's workbench. Naturally such linguistic analysis cannot reach a 100% accuracy owing to the intricacies of human language and its use. Besides, our experience in on-going trials of the toolkit in UK law enforcement agencies

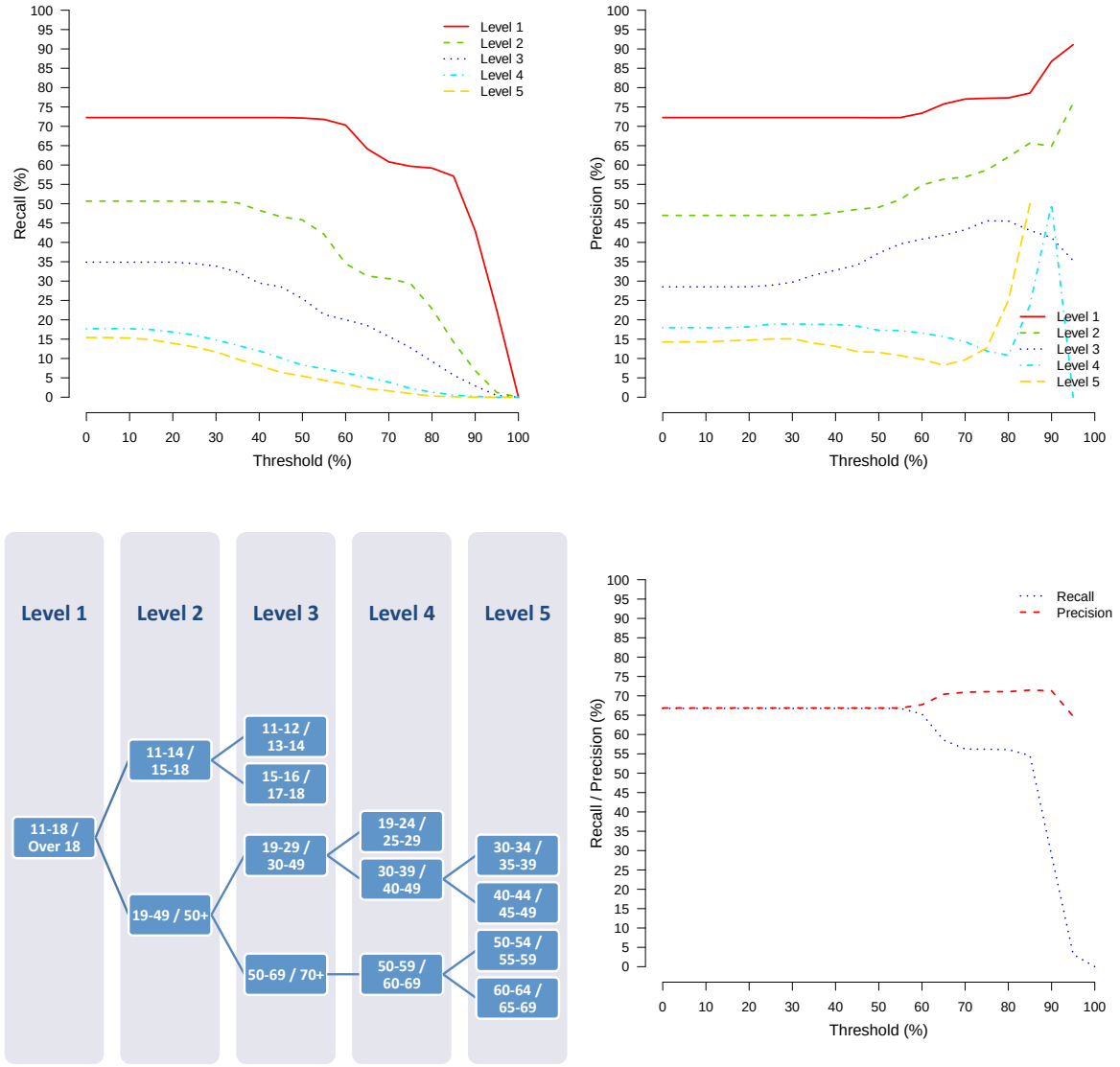

Figure 4: Recall (a - top left) and precision (b - top right) for age classification at different specificity levels (outlined in c - bottom left) and gender classification (d - bottom right).

shows that expert investigator knowledge is indispensible to the investigative process. The toolkit is, therefore, intended as a means to support the work of investigators rather than full automation. Only by combining such sophisticated tools with the expert knowledge of investigators can we hope to understand and nullify the tactics deployed by criminals online.

References

[1] Awais Rashid et al. *Technological Solutions to Offending*, pages 228–243. Willan, 2012.

[2] M. T. Whitty and T. Buchanan. The online dating romance scam: A serious crime. *CyberPsychology, Behavior, and Social Networking*, 15(3):181–183, 2012.

[3] Gabriel Weimann and Katharina Von Knop. Applying the notion of noise to countering online terrorism. *Studies in Conflict and Terrorism*, 31(10), 2008.

[4] Hasan Davulcu et al. Analyzing sentiment markers describing radical and counter-radical elements in online news. In *SocialCom/PASSAT*, pages 335–340, 2010.

[5] J. Diesner and K. Carley. Using network text analysis to detect the organizational structure of covert networks. In *Proceedings of the NAACSOS 2004 Conference, Pittsburgh, PA.*, 2004.

[6] Halil Bisgin et al. A study of homophily on social media. *World Wide Web*, 15(2):213–232, 2012.

[7] Phil Greenwood et al. Udesignit: Towards social media for community-driven design. In *International Conference on Software Engineering*, pages 1321–1324, 2012.

[8] Sheryl Prentice et al. The language of islamic extremism: towards an automated identification of beliefs, motivations and justifications. *International Journal of Corpus Linguistics*, 17(2):259–286, 2012.

[9] Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.

[10] S. Afroz et al. Detecting hoaxes, frauds, and deception in writing style online. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, pages 461–475, 2012.

[11] A. Narayanan et al. On the feasibility of internet-scale author identification. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, pages 300–314, 2012.

[12] Paul Rayson. From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4):519–549, 2008.

Professor Awais Rashid is Director of Security Lancaster, one of the UK's Centres of Excellence in Cyber Security. His research interests are in intelligent analysis of online data, digital identity and cyber security behaviours.

Dr Alistair Baron is a Research Fellow in Security Lancaster. His research deals with cyber security challenges posed by the noisy, irregular and multi-lingual nature of online data.

Dr Paul Rayson is director of the UCREL centre on corpus linguistics and natural language processing. His applied research is in online child protection, learner dictionaries and text mining.

Professor Corinne May-Chahal is Professor of Applied Social Science and co-chair of the UK College of Social Work. She leads on child protection work within Security Lancaster.

Dr. Phil Greenwood is Chief Technology Officer for Isis Forensics Limited. His expertise is in developing solutions to assist child protection investigations.

Dr James Walkerdine is CEO and co-founder of Isis Forensics Limited. His expertise is in online forensics and child protection.