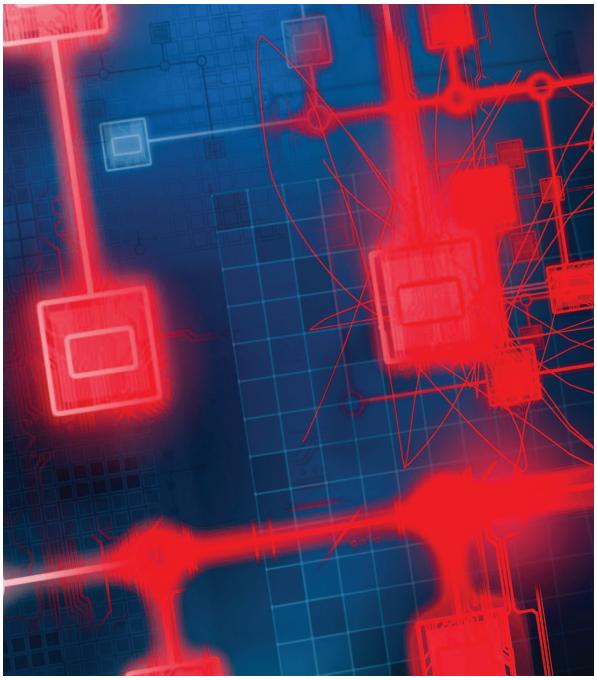

Who Am I? Analyzing Digital Personas in Cybercrime Investigations

Awais Rashid, Alistair Baron, Paul Rayson, and Corinne May-Chahal, *Lancaster University, UK*

Phil Greenwood and James Walkerdine, *Isis Forensics, Lancaster, UK*

The Isis toolkit offers the sophisticated capabilities required to analyze digital personas and provide investigators with clues to the identity of the individual or group hiding behind one or more personas.

Digital communities not only bring people closer together but also, inadvertently, provide criminals with new ways to access potential victims online. Digital personas play a key role in criminal tactics in online social media. One criminal might hide behind multiple digital personas or a group of criminals might share a single persona 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 the following:

- Child sex offenders masquerading as young people to gain their victims' trust. An offender might use multiple personas over the course of an interaction, initially posing as a young person and then introducing another persona—for example, an older relative. Alternatively, an offender group might share a single persona so that multiple people can groom a victim 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]
- Radicalization of youth in online forums through persuasive messaging.[3] Offenders sometimes use multiple digital personas as a tactic. For example, one persona is used to vigorously support a radical cause, followed by silence for a few days; 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 manual analysis difficult or even impossible. Consequently, law enforcement agencies face huge online communication data analysis backlogs during cybercrime investigations, with backlogs of six to nine months being commonplace.

Even though a range of commercial tools such as EnCase (www.guidancesoftware.com/encase-forensic.htm) and Internet Evidence Finder (www.magnetforensics.com/products/internet-evidence-finder) can assist in such investigations, they mainly focus on data extraction. Any analysis of the data is left to the investigator, who has access only to simple techniques such as keyword-based searches or phrase detection based on user-defined lists. Such techniques do not scale, and they do not include models of deceptive behavior or sharing of online personas. It is not uncommon for investigators to extract data

Published by the IEEE Computer Society

from hard disks or mobile phones using a tool such as EnCase and then manually read it to identify when an offense might have occurred and make a value judgment about 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 analyze all digital personas involved—the cognitive load is immense.

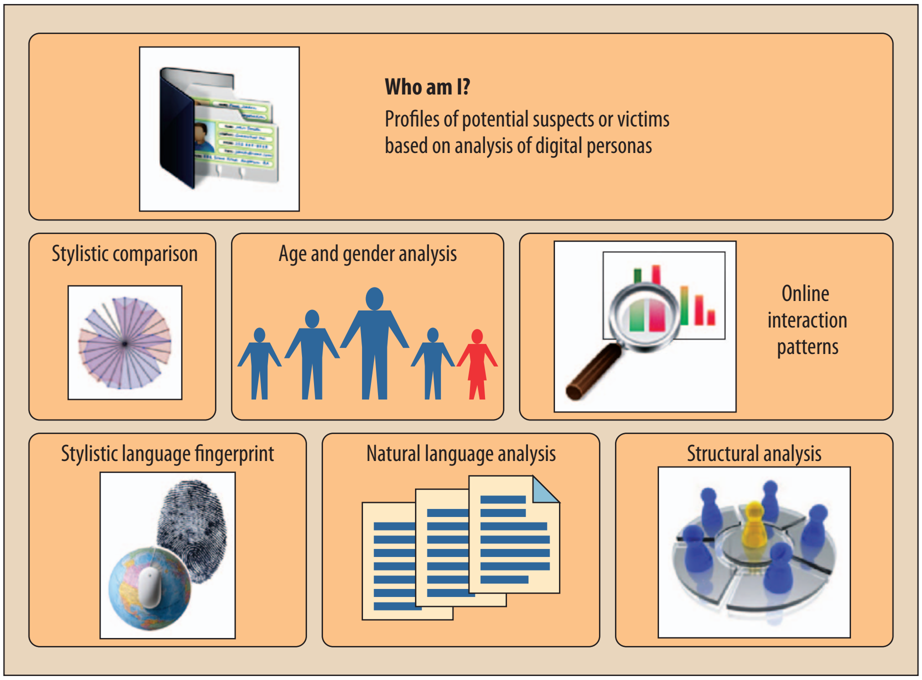
Who am I?
Profiles of potential suspects or victims based on analysis of digital personas
Stylistic comparison
Age and gender analysis
Online interaction patterns
Stylistic language fingerprint
Natural language analysis
Structural analysis

Figure 1. The Isis toolkit. The toolkit combines the use of statistical methods from corpus-based natural language processing with authorship attribution tools.

INFORMATION MINING AND ANALYSIS RESEARCH

Relevant research on the mining and analysis of information from online social media has mainly focused on extracting the key messages prevalent in such media. Hansan Davulcu and colleagues focused on detecting sentiment markers that indicate radicalization and counter messages in online forums.[4] A 2004 study demonstrated how common word use across actors can be used to derive knowledge about the structure of covert social networks and their weak points.[5] Other work revealed 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, researchers have applied techniques from the fields of corpus-based natural language processing and text mining to these problems. Corpus analysis, particularly at the semantic level, can describe 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 inherent deception tactics has not been considered to date.

The Isis toolkit (www.comp.lancs.ac.uk/isis) addresses this particular challenge by enabling efficient and sophisticated analysis of digital personas in large-scale online textual communications. Our approach complements recent research that highlights the difficulty in identifying authorship when language is intentionally obfuscated[10] as well as other work demonstrating that automatically predicting text authorship on a large scale is viable.[11] Our work shows that it is possible to predict a persona's key attributes such as age and gender with acceptable accuracy regardless of whether or not the author is obfuscating the language.

THE ISIS TOOLKIT

As Figure 1 shows, the Isis toolkit combines statistical methods from corpus-based natural language processing with authorship attribution tools. Analysis techniques from corpus linguistics and natural language processing, such as keyword profiling, offer the capability to compare word frequencies. Previous work extended this approach to extract key grammatical categories (equating to features of style) and key semantic fields (showing key concepts).[12] These techniques use large representative samples of writing or transcribed speech for training and reference comparison, have high accuracy, and are designed to be robust across various types of text. Using tools and methods from the authorship attribution field makes it possible to narrow the focus from language varieties down to the individual writer to identify the author's stylistic fingerprint.

In the past, authorship attribution techniques were mainly applied to determine the 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]

The specific challenges that we faced in implementing the Isis toolkit included integrating the statistically sophisticated but knowledge-poor techniques from authorship attribution with linguistically informed methods from corpus-based natural language analysis, combining the macro level (models of language varieties) with the micro level (models of individual's use of language). Additionally, these methods must operate on small quantities of noisy language data observed in online social networks and deal with masquerading or similarly deceptive behavior that an individual might assume in an attempt to hide his or her identity.

The Isis toolkit can compare the metric scores produced for two or more text collections to indicate how likely it is that the sources of the texts overlap or are written by people of a similar age and gender.

The novel investigative features of the Isis toolkit include the following:

- *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 multiple people are sharing a single persona.
- *Establish the age and gender of the person behind a digital persona*. Isis achieves this by synthesizing the stylistic "fingerprint" and extracting additional markers using a natural-language-analysis engine. Furthermore, the toolkit can detect masquerading tactics with a high degree of accuracy—for example, detecting when an adult is masquerading as a child.
- *Establish online interaction patterns of particular digital personas*. Isis analyzes both the conversation structure and the language used therein to determine a specific persona's key characteristics such as signature moves when signing off from a conversation or frequently used words and phrases. The toolkit also can analyze a persona's behavior—for example, identifying when a participant is typically active—not only within an average 24-hour period but also in terms of day of the week. It also can determine whether a persona becomes increasingly sexual or aggressive over a period of days or weeks.

These techniques are equally applicable either for building up a profile of potential suspects or victim identification. Investigators can use them to gain a better understanding of the digital personas involved, and their use also potentially provides clues to the identity of an individual or group in the physical world.

Stylistic language "fingerprint"

The Isis toolkit can observe and scrutinize a wide range of subtle language traits to assist in authorship analysis. Examples include the proportions of punctuation characters, the use of emoticons, and vocabulary measures. The toolkit uses these language traits to build a stylistic "fingerprint" that it can, in turn, use to represent the language of a particular user, a set of users, or a 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 the type token ratio, a vocabulary indicator. The calculation of each metric takes into account text length so that a mixture of sources can be combined where appropriate—for example, email texts are generally longer than chat room texts.

Through this process, Isis can assign a list of metric scores to a single text or collection of texts. Examples include the collated messages from a single chat room user or a sample of texts chosen to represent the language of adult female chat room users. The toolkit can then compare the metric scores produced for two or more text collections to indicate how likely it is that 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: to assist with automatic age and gender analysis, and to provide a visual impression of how close two text sources are with regard to their linguistic style.

Age and gender analysis

Isis performs this analysis in four steps. The first three steps utilize a natural-language-analysis engine, while the fourth combines the knowledge thus extracted with the metric scores from the stylistic fingerprint:

- Step 1: Tokenize an incoming text sample and tag each word with a part-of-speech (POS) label—noun, verb, adverb, adjective, and so on.
- Step 2: Assign each word or phrase within the text to one semantic field using general conceptual labels such as finance, warfare, government, sports, and so on. These first two steps rely on a set of hybrid techniques to select the most likely tag in each context.
- Step 3: Count features such as the language styles used at the word, POS, and semantic field levels.
- Step 4: Compare each level 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

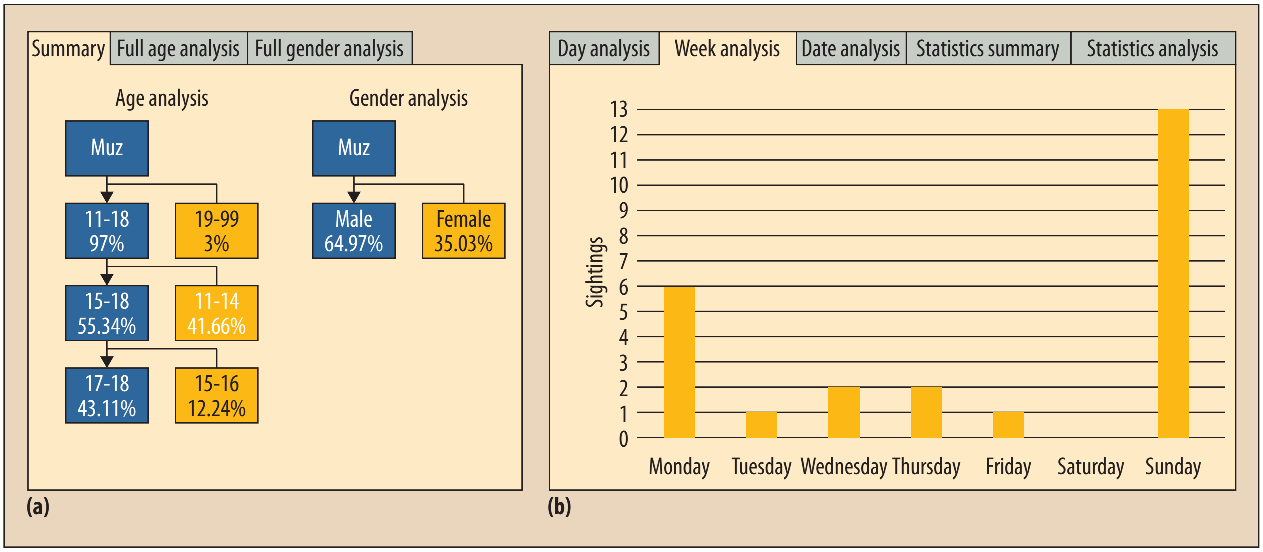

Figure 2. Specific reference datasets: (a) age and gender decision trees, and (b) online-offline time analysis.

of the three levels. Isis produces metric scores from the stylistic fingerprint for each reference dataset such as different gender groups and uses them as features for training a text classifier.

Various machine learning algorithms and methods for feature extraction are used for a range of text classification purposes. Isis uses logistic regression with the metric scores as features to classify a given text into gender groups. Probabilities are produced that 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 to derive weighted combined scores. The system then assigns a value for how likely it is that the incoming text is written by a male or female author. Similarly, Isis can prepare reference datasets by age range and compare them in the same manner.

It is possible to focus 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. As Figure 2 shows, the Isis toolkit provides this information as a decision tree that a law enforcement officer can consult and interact with.

Comparing stylistic fingerprints

While the automatic prediction of age and gender is useful in many cases, visualizing language differences and similarities also can be helpful to an investigator. The metric scores offer the ability to plot stylistic differences on a graph. While more than two lists can be compared on the same plot, here we discuss only the comparison of two lists.

Given two lists of metric scores, 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 specialized for the text comparisons being performed—for example, 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 evident, with substantial overlap indicating that the language style is similar and little overlap indicating contrasting language styles.

In addition to displaying how close a user's text is to a given age and gender dataset, the fingerprinting method also can be used to compare the text from two personas to establish whether they are actually the same individual, or to compare texts from one persona at different times to explore whether multiple individuals share the persona.

To describe this process and 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 the figure) of the fingerprints indicates that the new text's language style is similar to that of 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 chat room 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 cybercrime scenario.

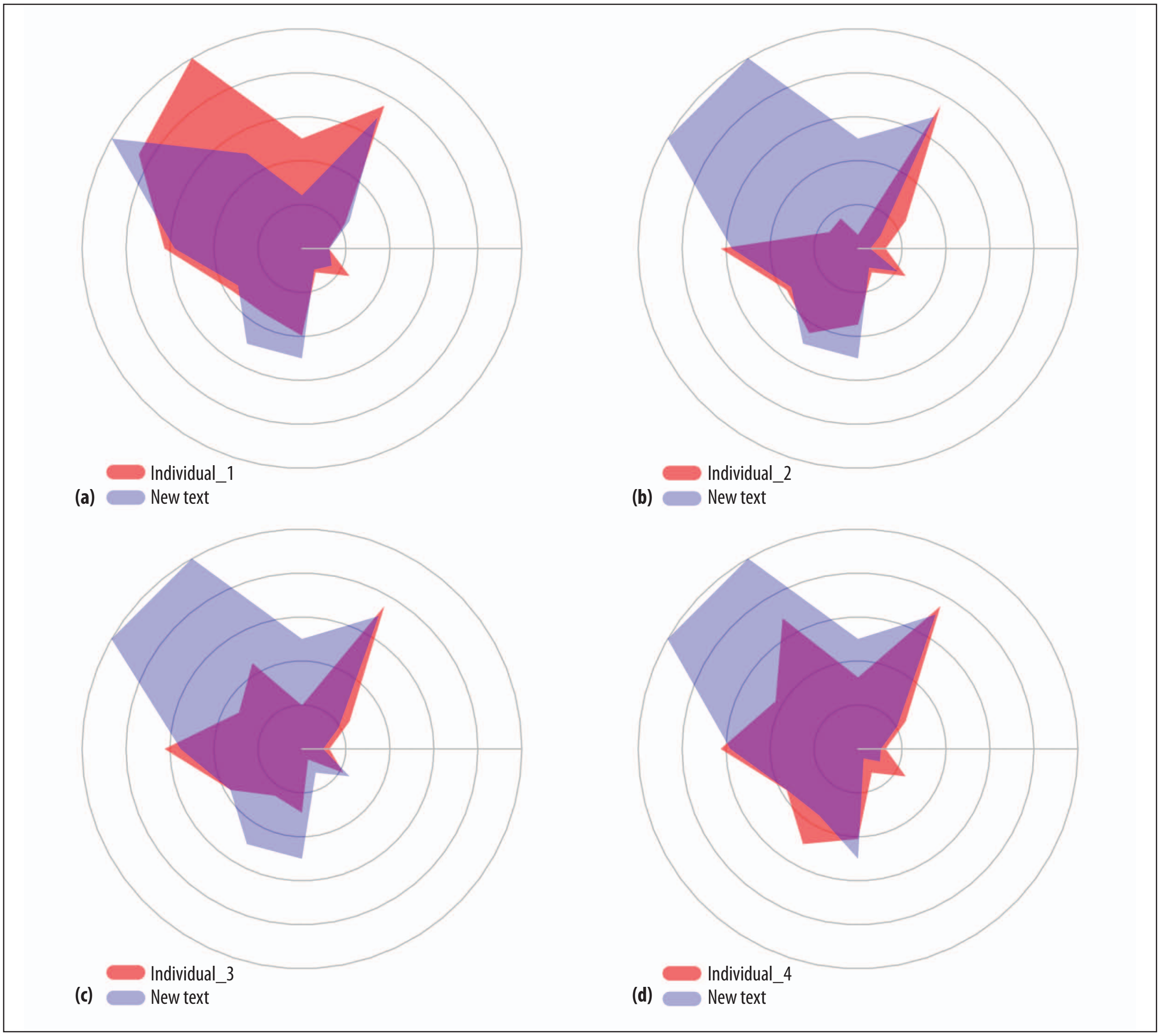

Figure 3. Comparison of language style fingerprints for a new text against four individuals' fingerprints.

Online interaction patterns

The Isis 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 usernames of those 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 entered into the toolkit are converted into a generic format. A key aspect of this is breaking down and modeling the log in terms of the participants and their respective activity, such as posting messages, sharing links, leaving or joining the conversation, and so on. Once it has built this model, the toolkit can quickly analyze and present intelligence about a particular participant. This can include an analysis of language use—for example, frequently used stylistic characteristics such as keywords, names, topics, and so on—or identifying patterns of online and offline times. Semantic categorization allows classifying parts of a conversation based on their meaning—for example, whether it is 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 might occur over the duration of the conversation, for example, to help determine whether a conversation is becoming increasingly sexualized.

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

Profiling cybercriminals and victims

The various analysis techniques combine to form a key feature of the toolkit—the ability to generate identity profiles of specific digital personas. Isis can 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.

The generated profile is built from several elements, including

- *Language usage.* This element provides a model of the persona's language use within conversations and highlights characteristics such as people or place names, dates and times, frequently used words and phrases, aggressive or sexual content, or email addresses and URLs. It also includes nondictionary words about which an investigator might or might not be aware that could indicate an attempt at disguising what is being discussed or represent unique jargon used within that domain.
- *Age and gender analysis.* Investigators can use a decision tree to provide an inferred estimation of the age and gender of the person behind the persona. By default, this provides a summary view that presents the strongest path through the tree, but they also can view the full tree, allowing them to examine the decisions the toolkit made at all points if the certainty of the decision is not clear-cut.
- *Online activity.* An analysis of overall online activity can highlight when the persona 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.

These profiles can provide investigators with additional intelligence about trends and characteristics not immediately apparent to the human eye.

DIFFERENTIATING BETWEEN GENUINE PERSONAS AND DECEPTIVE BEHAVIOR

We have used the Isis toolkit on reference datasets and in live environments to test its effectiveness in correctly detecting the attributes of an individual behind a persona. In the two test sets presented here, no deception is intended in the first, while in the second, an individual is using deceptive tactics.

Classifying age and gender of genuine personas

For this test we used the British National Corpus (BNC), a reference dataset with 100 million words of written and spoken language that represents a wide cross-section of British English. We utilized the portion of BNC (1,684 people, which constitutes 10 percent of the entire collection) where metadata about an individual, including age and gender, was available. We used "leave-one-out cross-validation" to train our system using the texts from all 1,684 individuals except the text from the individual being used as a test subject—the person whose age and gender was being classified. We repeated this classification for all 1,684 individuals as test subjects.

For each classification, the Isis toolkit provides probabilities that the individual belongs to a specific age band; for example, an individual might be predicted to be age 11 to 18 with a probability of 74 percent, and over age 18 with a probability of 26 percent. The prediction then moves down a level, that is, to between ages 11 to 14 with a probability of 49 percent, and between ages 15 to 18 with a probability of 25 percent, and so on, with gender probabilities also calculated.

As the age and gender classification decision trees in Figure 2a show, the age group or gender with the highest probability is taken as the prediction at each decision point. A probability threshold is also used to decide whether a prediction is used. If the highest probability is below this threshold, no prediction is made, and the age or gender is marked as "unknown."

Our analysis performs better when deception tactics are being used, thus demonstrating the effectiveness of digital persona analysis as a valuable tool in the investigator's workbench.

Recall and precision are used to measure the algorithms' ability to correctly classify the age and gender of each individual in the test set. Recall is the proportion of individuals tested for which the correct prediction is made; precision is the proportion of predictions made—that is, not unknown—that are correct according to the metadata.

Figures 4a and 4b give the recall and precision values obtained for age classification at different specificity levels, which are outlined in Figure 4c. Recall is 72.15 percent, and precision is 72.24 percent at Level 1, that is, distinguishing between children (ages 11–18) and adults (over age 18). This is based on a probability threshold of 50 percent. By increasing the threshold, greater precision can be achieved at the cost of fewer classification decisions being made—that is, more unknowns are returned. With a higher threshold of 80 percent, the precision of adult and

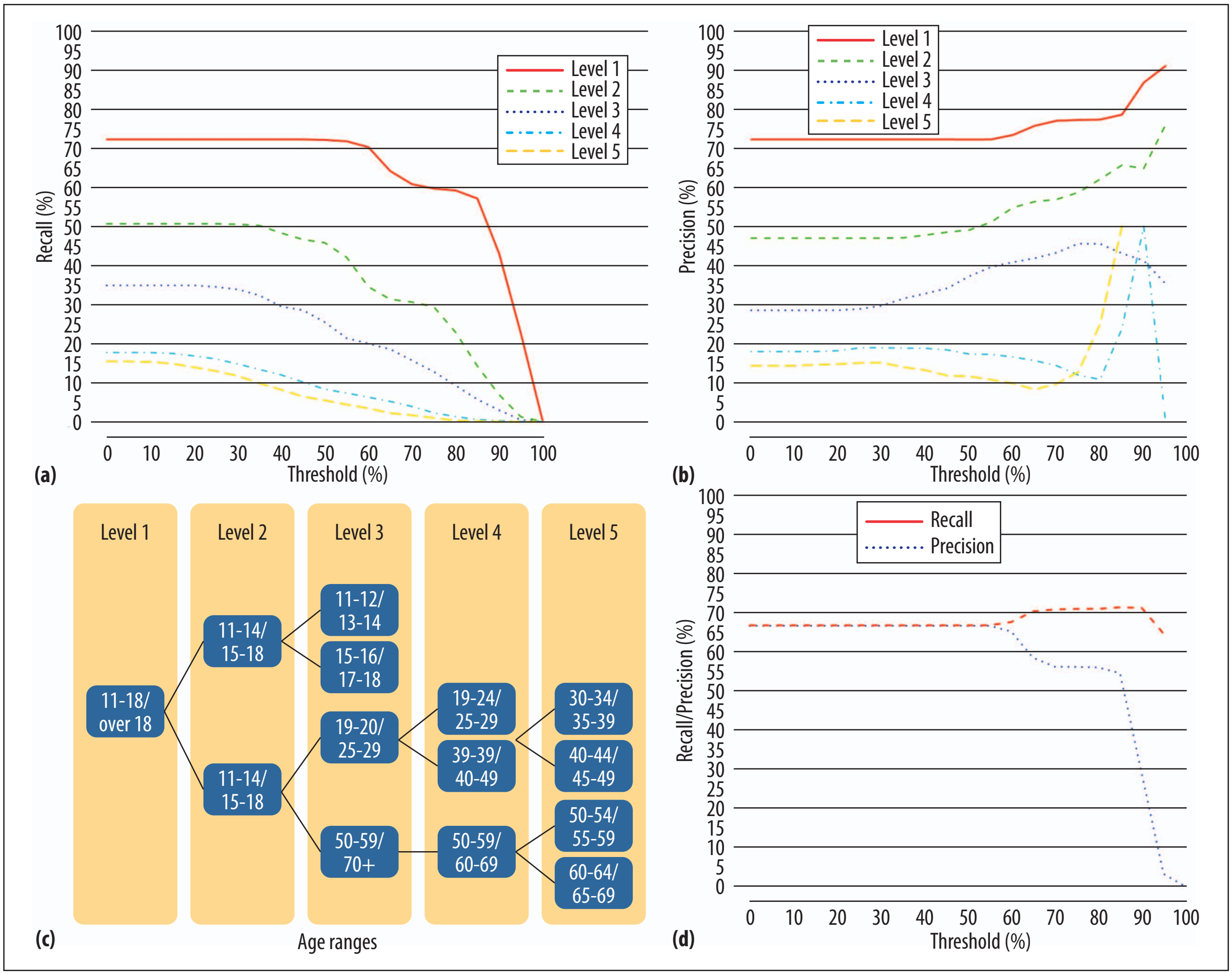

Figure 4. Age and gender classification: (a) recall and (b) precision for age classification at (c) different specificity levels, and (d) gender classification.

child classification increases slightly to 77.35 percent, but recall drops to 59.20 percent. Naturally, the precision and recall drop at higher levels of the age decision tree where the age ranges are more specific.

As Figure 4d shows, for gender classification with a threshold of 50 percent, recall is 66.74 percent, and precision is 66.86 percent. Again, a higher threshold can be set; increasing the threshold to 80 percent improves precision to 71.07 percent, but recall drops to 56.08 percent.

Classifying deceptive personas

We tested our toolkit on the BNC data to determine the accuracy of our algorithms when individuals are not being deceptive. However, given our focus on detecting misuse of digital personas, we tested the toolkit on detecting masquerading behavior—when an individual hides behind a false persona, for example, pretending to be a child.

We set up a "live" environment, similar to a Turing test, in two schools. Subjects ages 11 to 18 years chatted online with 10 individuals behind the scenes in sessions divided by age group. In each session, one-half of the individuals behind the scenes were children or 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 or young people or masquerading as being in those age groups.

For deciding whether an individual is a child or an adult masquerading as a child, the age classification algorithm achieves precision and recall of 84.29 percent, with a probability threshold of 50 percent. The precision can be increased with a higher threshold; at an 80 percent threshold, precision increases to 93.18 percent, but recall drops to 58.57 percent, with fewer predictions being made.

These results obtained using our toolkit are in stark contrast to the accuracy of the children's responses, with only 18 percent of children across the year groups able to correctly identify whether they were chatting with an adult or a child. For gender, precision is 80.6 percent with a 50 percent threshold, while recall is 77.14 percent. Increasing the threshold to 80 percent improves precision to 84.09 percent, but recall drops to 52.86 percent. These results are in contrast to the children correctly identifying the gender of the person with whom they were chatting in 58.8 percent of the cases.

Online social media affords "connectedness" that enables individuals and groups from various geographical, cultural, and socioeconomic backgrounds to interact and share experiences. However, the very nature of identity in online social media—a fluid and dynamic concept 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 that cybercriminals use. This new digital world and these sophisticated criminal tactics call for new tools to aid 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, thus demonstrating the effectiveness of digital persona analysis as a valuable tool in the investigator's workbench.

Naturally, such linguistic analysis cannot provide 100 percent accuracy because of the intricacies of human language and its use. In addition, our experience in ongoing trials of the toolkit in UK law enforcement agencies shows that expert investigator knowledge is indispensable to the investigative process. The toolkit is, therefore, intended as a means to support the work of investigators rather than offering full automation. Only by combining such sophisticated tools with the expert knowledge of investigators can we hope to understand and nullify the online tactics that criminals deploy.

References

1. A. Rashid et al., *Technological Solutions to Offending*, Willan, 2012, pp. 228-243.
2. M.T. Whitty and T. Buchanan, "The Online Dating Romance Scam: A Serious Crime," *CyberPsychology, Behavior, and Social Networking*, vol. 15, no. 3, pp. 181-183.
3. G. Weimann and K. von Knop, "Applying the Notion of Noise to Countering Online Terrorism," *Studies in Conflict and Terrorism*, vol. 31, no. 10, 2008, pp. 883-902.
4. H. Davulcu et al., "Analyzing Sentiment Markers Describing Radical and Counter-Radical Elements in Online News," *Proc. 2nd Int'l Conf. Privacy, Security, Risk and Trust* (PASSAT 10), IEEE, 2010, pp. 335-340.
5. J. Diesner and K. Carley, "Using Network Text Analysis to Detect the Organizational Structure of Covert Networks," *Proc. Conf. Computational Analysis of Social and Organizational Systems* (CASOS 04), Nat'l Assoc. Computational Social and Organizational Science, 2004; www.andrew.cmu.edu/user/jdiesner/publications/NAACSOS_2004_Diesner_Carley_Detect_Covert_Networks.pdf.
6. H. Bisgin et al., "A Study of Homophily on Social Media," *World Wide Web*, vol. 15, no. 2, 2012, pp. 213-232.
7. P. Greenwood et al., "Udesignit: Towards Social Media for Community-Driven Design," *Proc. Int'l Conf. Software Engineering* (ICSE 12), IEEE, 2012, pp. 1321-1324.
8. S. Prentice et al., "The Language of Islamic Extremism: Towards an Automated Identification of Beliefs, Motivations and Justifications," *Int'l J. Corpus Linguistics*, vol. 17, no. 2, 2012, pp. 259-286.
9. E. Stamatatos, "A Survey of Modern Authorship Attribution Methods," *J. Am. Soc. Information Science and Technology*, vol. 60, no. 3, 2009, pp. 538-556.
10. S. Afroz et al., "Detecting Hoaxes, Frauds, and Deception in Writing Style Online," *Proc. IEEE Symp. Security and Privacy* (S&P 12), IEEE, 2012, pp. 461-475.
11. A. Narayanan et al., "On the Feasibility of Internet-Scale Author Identification," *Proc. IEEE Symp. Security and Privacy* (S&P 12), IEEE, 2012, pp. 300-314.
12. P. Rayson, "From Key Words to Key Semantic Domains," *Int'l J. Corpus Linguistics*, vol. 13, no. 4, 2008, pp. 519-549.

Awais Rashid *is the director of Security Lancaster at Lancaster University, UK. His research interests include intelligent analysis of online data, digital identity, and cybersecurity behaviors. Contact him at marash@comp.lancs.ac.uk.*

Alistair Baron *is a research fellow at Security Lancaster at Lancaster University, UK. His research deals with cybersecurity challenges posed by the noisy, irregular, and multilingual nature of online data. Contact him at a.baron@lancaster.ac.uk.*

Paul Rayson *is director of the University Centre for Computer Corpus Research on Language at Lancaster University, based in both the Department of Linguistics and English Language and the School of Computing and Communications. His applied research is in online child protection, learner dictionaries, and text mining. Contact him at p.rayson@lancaster.ac.uk.*

Corinne May-Chahal *is a professor of applied social science and cochair of the UK College of Social Work at Lancaster University, UK. She leads the child protection work within Security Lancaster. Contact her at c.may-chahal@lancaster.ac.uk.*

Phil Greenwood *is the chief technology officer for Isis Forensics, Lancaster, UK. His expertise is in developing solutions to assist child protection investigations. Contact him at p.greenwood@isis-forensics.com.*

James Walkerdine *is the CEO and cofounder of Isis Forensics, Lancaster, UK. His expertise is in online forensics and child protection. Contact him at j.walkerdine@isis-forensics.com.*