Bruce et al. Security Informatics 2014, 3:12
http://www.security-informatics.com/content/3/1/12

Security Informatics
a SpringerOpen Journal

RESEARCH Open Access

Pathways to identity: using visualization to aid law enforcement in identification tasks

Joe Bruce[1*], Jean Scholtz[1], Duncan Hodges[2], Lia Emanuel[3], Danaë Stanton Fraser[3], Sadie Creese[2] and Oriana J Love[1]

Abstract

The nature of identity has changed dramatically in recent years and has grown in complexity. Identities are defined in multiple domains: biological and psychological elements strongly contribute, but biographical and cyber elements also are necessary to complete the picture. Law enforcement is beginning to adjust to these changes, recognizing identity's importance in criminal justice. The SuperIdentity project seeks to aid law enforcement officials in their identification tasks through research of techniques for discovering identity traits, generation of statistical models of identity and analysis of identity traits through visualization. We present use cases compiled through user interviews in multiple fields, including law enforcement, and describe the modeling and visualization tools design to aid in those use cases.

Keywords: Identity; Attribution; Enrichment; Visual analysis; Modeling; Wizard; Law enforcement; Visualization; Data transformation

Background and related work

The complexity of identities has increased dramatically in recent years, particularly with the introduction and widespread adoption of social networks and broader online activities [1]. The complexity presents challenges for law enforcement, especially given the massive volumes of data being generated [2]. Identity is a key enabling factor in almost everything in the 21st century [3]; successfully identifying individuals underpins almost every secure and private system. Important systems such as banking, international travel, and commerce all rely on identity; many interpersonal services rely on individuals identifying themselves (e.g., email, information technology services, etc.)[4].

Gathering a more complete picture of an individual (identity enrichment) and tracing an activity back to the acting party (identity attribution) are tasks complicated by the diversity of identifying information. However, the glut of data can also aid law enforcement in their investigative tasks if they can take advantage of it [5]. Furthermore, new connections and investigative paths are accessible if the research can be materialized into operational procedures [6].

Law enforcement decision makers are now recognizing the importance of using social networks in their investigations. The International Association of Chiefs of Police (IACP) conducted a survey in 2012 that showed that 92.4% of the 600 law enforcement agencies polled were using social media [7]. Law enforcement personnel can use social media for prevention of crime as well as investigation of crime. Information about plans for protests often can be found online as can discussions and photographs of activities that have occurred [8]. As identity spans both the natural and the cyber worlds, it is important that law enforcement have the tools to establish and pursue identities as they flow across the domain boundaries; no longer is it sufficient to explore identities purely in one domain [3].

The SuperIdentity project is designed to accomplish just that: provide the tools that pursue identities across domains. SuperIdentity is a collaborative effort among six UK universities and a US national laboratory conducting research in a variety of domains that can aid in attribution and enrichment investigative tasks. This paper presents a visual analytic application developed at the Pacific Northwest National Laboratory (PNNL) that provides a visual interface to a complex statistical model of identity, developed at the University of Oxford. The model encapsulates research performed at the collaborating universities in

* Correspondence: joe.bruce@pnnl.gov
[1]Pacific Northwest National Laboratory (PNNL), Richland, WA, USA
Full list of author information is available at the end of the article

the United Kingdom and generates analytic pathways that can lead an investigator from that which is known to that which is unknown but essential to the investigative task.

The pathways generated by the model, and the visualization that makes them accessible, will enable law enforcement to take full advantage of the interconnected elements of identity in our modern world. From the beginning, the SuperIdentity project has used a user-centric design approach. An early activity in the SuperIdentity project was the collection of use cases (*Developing the use cases* Section) to explore how various law enforcement roles could interact with the model. The aim of this activity was twofold. The initial aim was to understand how various types of end users undertook the task of identifying individuals, including the data used, the initial evidence that was usually available, the desired outcome of the identification, and the amount of uncertainty that could be tolerated. Secondly, the information obtained in these use cases helped to guide the development of the model for identity and the visualization to support those same end users. The use cases were collected through interviews with law enforcement personnel and illustrate difference scenarios in identifying individuals or in enriching the identification of an individual. These use cases helped researchers understand how law enforcement personnel go about doing identification tasks and problems and constraints in this work. Exemplar use cases were developed from current use cases and illustrate how new types of identity traits could facilitate identification tasks. The application has been developed using exemplar use cases that have helped identify useful visualization techniques. A user-centered evaluation is also planned for later in the year and will be discussed in the *Future Work* section.

Research in identity attribution and enrichment is beginning to explore how this process may span the physical and cyber world divide. For instance, links between biometric physical features and avatar recognition for identity authentication and enrichment are being explored [9]. There has been a focus on authorship attribution, for example, to determine if the content of multiple online social network identities belongs to a single author [10] or extracting identity features in tracing cybercrime [11].

Several active research consortiums are exploring how to mine and model identity data in more comprehensive ways. The Collaborative information, Acquisition, Processing, Exploitation and Reporting project (CAPER) is one of the larger ongoing projects [12]. This active consortium is focused on enabling law enforcement agencies across the EU to work together, with a focus on the ability to use and share open source intelligence to detect and prevent organized crime. The development process of the Caper Regulatory Model (CRM) has taken a user-centric approach in involving law enforcement users in the function, legality, and visualization of the model. The project has been guided by the goals to enable analysts to collaboratively collect, connect, and work with multiple data types (e.g., audio, text, video) to enrich identity information, while allowing analysts to fuse their own closed source data with open source on a site-by-site basis. However, much of the project's focus has been developing the former (enhancing open source web mining/analysis capabilities) to provide "early warning" trends for organized crime across cyber-space. Although there is some discussion on the inclusion of physical domain identity attributes, such as biometric data, there is no indication of cross-domain links or inference capabilities beyond big data pattern detection within CRM.

The Uncertainty of Identity multidisciplinary project has been capitalizing on the wealth of location information now available in the cyber world [13]. Specifically, this project looks at how geo-social networks can provide spatio-temporal information linking physical and virtual identities. Although this approach allows for large-scale links between physical and cyber domains, the identity attributes that can be inferred from cyber location data are currently limited to online social networks and relatively shallow identity descriptors. For instance, the project has developed a method to model and geographically visualize Twitter activity. From this, the user is able to infer names, ethnicity, gender of Twitter users, and temporal activity patterns across different physical locations (e.g., London, Paris, and New York) [14].

Much of the previous attribution work has centered on author attribution and cyber intrusion attribution, e.g., [15-17]. This work is heavily weighted toward automated, machine-learning solutions. Moreover, the attribution efforts do not cross identification domains.

For many applications, a holistic approach to understanding identity is needed. A holistic approach would span and link physical and cyber domains and be organized in such a way that individuals working with identification tools could intuitively understand, organize, manipulate, and infer unknown from known pieces of information about an individual or group. This paper introduces the SuperIdentity approach to these issues and describes the user-centric methods used in the development of the SuperIdentity model and visualization interface.

Developing the use cases

Use case development was an early activity in the SuperIdentity project. It is important to note that the majority of these use cases were obtained in the United States and hence some of the constraints discussed here are based on the legal framework applicable in the United States. However, researchers in the SuperIdentity project are looking at differences in laws in different countries, and indeed the resulting model allows the encapsulation of differing legal, cultural, and ethical frameworks [4].

Use cases were collected for several reasons. First, use cases provided insight into how analysts who have the task of identifying individuals currently work, including information they usually know, information they need to know, the certainty needed in identifications, resources used, and obstacles encountered, such as time constraints and inability to access certain resources. Secondly, understanding current work provided insights about what identity attributes would be useful and how to present this information to the end-users.

Individuals were recruited from three analysis domains: intelligence analysis, cyber security investigations, and law enforcement. PNNL works with a number of different agencies and law enforcement departments, so individuals in some of these organizations were contacted to help recruit participants. In the United Kingdom, recruitment was facilitated by one of the project sponsors. In the law enforcement domain, we interviewed individuals from the county sheriff's office, a police chief, and city law enforcement officers working in a fusion center. The individuals who participated were interviewed about their work in identifying individuals.

Questionnaires were also developed to elucidate the types of information that were used and how important the various information types were to the identification work. After the semi-structured interviews, participants were asked to look at information in different domains and identify those elements they commonly used in their work. For example:

Demographics/physical attributes: age, gender, ethnicity, handedness, facial biometrics
Work and extra-curricular activities: hobbies/interests, travel plans, group affiliations
Financial information: owned assets, banking information
Court/council records: arrests, tickets/fines, current/past addresses
Cyber attributes: email addresses, social network user names, personal websites

Overall, 21 individual use cases were developed. Commonalties were identified in these individual use cases, which resulted in the generation of several generic use cases, such as going from an online user name to an actual name. In addition, a number of exemplar use cases were generated, illustrating how different types of information can be combined to augment what we know about a person, including cyber (e.g., IP address), biometric (e.g., fingerprint), biographic (e.g., home address), and psychological (e.g., personality traits) information.

In this paper, the focus is on the issues involved in law enforcement. Besides gathering information to generate use cases, additional information about policies and procedures in law enforcement was obtained, providing valuable insights into the context in which software tools need to work. The following paragraphs contain information about the context in which identification tasks often take place. It should be noted that the majority of this information is based on interviews done in the US states of Oregon and Washington.

Much of the public-facing law enforcement work is done in real time and in close proximity to the individual being identified—e.g., during traffic stops. However, for some officers there is also other investigative work that, while it does not have to be completed in real time, has a requirement for the task to be completed as efficiently as possible.

Real-time law enforcement work, such as traffic stops, can have many constraints. While a law enforcement officer may stop a car or an individual, there has to be a valid reason, such as a traffic violation. The driver of the car or the individual stopped does have to talk to the law enforcement officer, but the discussion must focus only on the reason the individual was stopped. The individual detained can refuse to answer questions, but it is illegal to lie to a law enforcement officer. The individual stopped can only be detained for a short period of time; a traffic stop that lasts for more than 20 minutes is unreasonable. If passengers are in the car, they cannot be asked for identification or questioned unless the officer has seen them break a law.

Having stopped an individual, the law enforcement officer has information about the vehicle (if driving), the driver's license of the individual driving the vehicle, and a physical description of the individual. This information can be communicated via radio or a computer and additional information obtained, such as the owner of the vehicle, the name and aliases of the driver, whether there are outstanding warrants for the arrest of the driver, and possibly information as to whether the driver has a history of being "unfriendly to law enforcement." If there is a warrant out for arrest of an individual with this name, date of birth, and matching physical characteristics, the officer is justified in taking the individual to the police station and taking fingerprints to increase the certainty of the identification. It is currently not legal to obtain a DNA sample.

In general, law enforcement officers preferred to err on the side of caution. It is better to take a person in for more questioning than to miss picking up a person who has outstanding warrants. When officers provide data on the stopped individual to a police database "near misses" are returned if there is no direct match. Sorting out a number of near misses may consume too much time, so the officer may take the individual back to the police station for full identification.

In investigative work, officers have more time to conduct their investigation but also lack the richness of information (physical and recorded) that officers in real-time situations have. There may be physical descriptions from eyewitnesses and a description of a vehicle and/or a license plate number. Items may be left behind, including notes, fingerprints, and footprints. The goals are to locate the individual(s) who were involved in the crime being investigated and to place them at the scene of the crime with enough certainty to make an arrest.

In addition to the work done by Law Enforcement Divisions, the United States has created a number of fusion centers in response to the September 11, 2001 terrorist attacks. These fusion centers, located in urban areas, comprise representatives from the major intelligence agencies and state and local law enforcement agencies. They can share information to help with law enforcement, prevent terrorist activities, and respond to emergency situations [18]. The analysts that work in these fusion centers are a very rich resource for our use cases as they deal with many different types of data, from different sources, with differing confidence levels; hence, several interviews were conducted with analysts from fusion centers.

In the 21 use cases we collected, 10 were from the law enforcement domain and the rest were from intelligence domain. Here we focus on the law enforcement use cases. Of these 10, three could be classified as strictly attribution tasks: a crime has happened and the task is to identify who did it. Four could be classified as enrichment tasks: an individual is known but it is unknown if this individual is a danger to police. Three others could be classified as both attribution and enrichment: is this person really who they say they are, and what else is known about that person? Even attribution tasks are often more complex that just going from a crime to a name. We might have several similar crimes and want to determine if the same individual is responsible for all crimes. Of the use cases involving enrichment, two were descriptions of real-time incidents; law enforcement officers have an individual in front of them and want to understand who that person is and if that person poses a danger to them.

A typical use case is the Property Crime: Someone has broken into a car and stolen some items. There is a small amount of evidence, e.g., some fingerprints, a footprint, and perhaps a description of the person from a passerby. The police want to find out who is responsible. They need to find the name of the individual and the individual's current location.

The use cases, such as the examples discussed above, provided a good understanding of the tasks, resources and constraints that law enforcement officers face. Using this information a number of exemplar use cases were developed. Similar themes in use cases were identified and merged into an exemplar use case. As the individual use cases are based on what is currently done, the exemplar case studies were augmented to utilize the work in SuperIdentity. These use cases are being used in developing the visualizations to demonstrate to end users the different possibilities for obtaining the desired information given their starting information, the various resources available to them, and the certainty required in their identification. The visualization work will be discussed in the *SuperIdentity model and visualization* Section and will use the exemplar use case below.

The exemplar use case described below is based on a real homicide case. We have added a second homicide to illustrate how the model can be used to in connecting information between two crimes. We have also added an iPhone to introduce new types of information used by the model.

Two homicides occurred in a particular town within two weeks. The police are trying to find a possible suspect or suspects and motives. They have several pieces of information. At the scene of the first homicide, a witness saw a car leave the area and is able to recall a partial license plate. An iPhone was dropped at the second scene and police believe it is reasonable to assume that the phone is not associated with the victim or the victim's friends or family.

Using the description of the car and the partial license plate number, the police have located a set of a dozen suspects. Of these, only four were anywhere close to the area of the crime scenes. None of these four have any police records. However, the phone number of one of these persons was listed as a recent call on the cell phone left at the scene.

The license plate information and car description are typically what would be expected for information as are the contents of the cell phone, such as recent online activity and fingerprints on the phone. In the next section, the model will suggest other information that may be obtained. If a suspect can be identified and brought in for questioning, it may be possible to determine if the suspect was responsible for both homicides.

SuperIdentity model and visualization

The exemplar use case described in the previous section demonstrates the requirement to consider identity as complex phenomena crossing many different domains. SuperIdentity considers identity across four complementary domains: an individual's biological makeup, biographical information, online presence (cyber domain), and mental makeup (psychological domain). These cross-domain characteristics are used to evaluate the holistic identity associated with individuals.

There are projections of identity in the natural world; the physical projections of identity include things like biometrics, an individual's description (height, sex, weight, hair color, etc.) are important as these are used by humans to help recognize individuals in the natural world. In addition to these physical projections of identity, biographical information provides factual information about the individual, such as home address, occupation, social security numbers, work address, etc. This information is often used to locate or identify a single individual.

Over the last 20 years, online projections of identity have emerged; these projections in cyber-space are now pervasive throughout society. Society uses cyber-space for everything from very personal activities (such as engaging with friends, documenting our lives, expressing our creativity), to learning and developing opinions, to very practical concepts such as travel planning, commerce, and banking [19]. All online activities leave a trail of identity pieces scattered throughout cyber-space, whether through conscious disclosures (e.g., on social networks), subconscious disclosures (such as exploited in textual-content analysis), or through technology-level leakages (e.g., through cookies) [20]. Meta-data associated with online activities also exists, such as IP addresses, account names, etc., which can all be related to an individual's identity.

The final identification domain that is important to consider when discussing identity is the psychological domain. This involves the psychological profile of an individual. Measures like the Big-5 [21] and the Dark Triad [22] give insight into concepts such as extraversion, conscientiousness, Machiavellianism, etc. These personality elements often drive behaviors in different spaces and as such are important to understand when investigating an individual.

All four of these identification domains are important to consider. This is clear from the exemplar use case in the *Developing the use cases* Section. Starting with a mobile phone, much of the data available is likely to be related to the cyber domain, and much of the investigation may involve gathering online data. But the investigator likely wants to explore more about the motive of the individual and indicators of intent. Both psychological traits and content posted on the Internet will be contributing factors. The license plate will provide contradictory information to that of the phone: the owners are not the same person. This investigation will draw in social contacts, physical residence, and temporal location. Leveraging all this data to yield a more complete picture of the individuals involved requires the collective use of data from all four domains.

To this end, we use an intuitive modeling approach that, in an investigative mode, allows the capture of these rich identities in addition to documenting both the processes by which the pieces of identity were derived and the confidence associated with each piece. To illustrate, consider the partial license plate gathered as evidence from the exemplar use case. Any identification performed will have a moderate level of uncertainty attached to it because the correctness of the license plate is in question. Inferences suffer the same potential uncertainty, depending on the nature of the inference. An investigator might infer a relationship between a phone's owner and a person listed in the contact list, but uncertainty exists about the nature and strength of that relationship. Knowing how the inferences were made (their derivation) and their associated confidence is necessary to yield a complete picture of what is known, and what is unknown, about a particular identity.

The model is built around two simple concepts: elements of identity and inferences. An element of identity is a single piece of information that is representative or indicative of an identity; it does not need to be unique (i.e., an individual's height), it may be unique in a given context (i.e., a Twitter user name is unique on Twitter, but the same user name may be used by another individual on Instagram) or it may be globally unique (e.g., a MAC address). Each element has an associated confidence that is related to the uncertainty associated with an element.

As part of the SuperIdentity project, a simple taxonomy of these elements of identity has been created with the top layer of the hierarchy representing the four identification domains discussed previously (biological, biographical, cyber, and psychological). The folksonomy was created as a joint-design exercise amongst the SuperIdentity consortium. The cross-disciplinary expertise within the consortium was exploited in order to identify both the elements of identity across all projections of identity and then place the elements into a hierarchy. The hierarchy that resulted from this exercise contained four distinct levels the top-level representing the four domains (as discussed previously), the level below this representing a sub-division within this domain (for example within Cyber there are subdivisions for devices (e.g. smartphone, laptop, etc.), generic technologies (e.g. e–mail) and personas (an individual's use of a site)). The penultimate level of the hierarchy represents the space where the element of identity is exposed (for example elements are exposed in a Facebook account, other elements are exposed in a hand). The final element of the hierarchy is the actual data point – this represents the element that is explicitly measured or captured (for example within Facebook, the username, the avatar, the friendslist etc. all represent elements that can be captured, for a hand the fingerprints can be captured, the fingerlength can be measured, etc.) Figure 1 illustrates Facebook and hand identity elements as examples within the context of the SuperIdentity hierarchy. Summarizing statistics related to the model as a whole can be found in Table 1.

This hierarchy allows control of the fidelity around the elements of identity, e.g., an investigator may be interested in anything about the individual's work-life rather

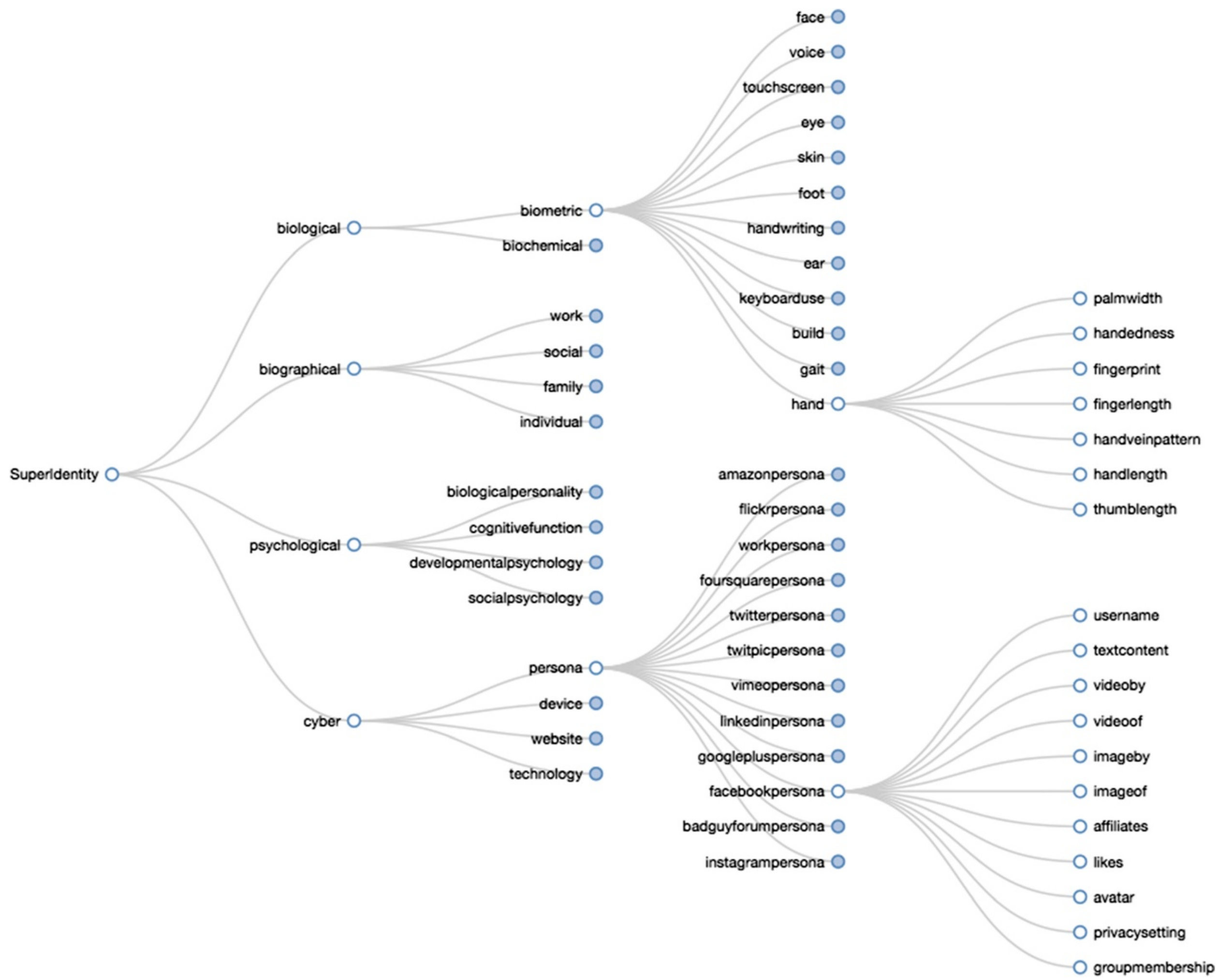

Figure 1 Hierarchical view of the SuperIdentity model. The penultimate level of *facebookpersona* and *hand* are revealed. White nodes represent fully expanded portions of the hierarchy and blue nodes represent collapsed portions of the hierarchy.

than his or her specific occupation [23]. The model is completely agnostic of any particular taxonomy—meaning any taxonomy of elements can be used, and it is expected that individual organizations will tailor the taxonomy of elements to their context.

Simply having a bag of elements of identity is not enough to satisfy the exemplar use case previously outlined. To this end, we introduce the second core concept of the model: inferences. These allow the creation of a new, previously unknown element of identity from a known element of identity; these can be simple, automated transforms (e.g., using a Twitter username to infer the corresponding Twitter avatar), more complex inferences that use statistical correlations (e.g., the estimation of height from foot-length), or other inferences that involve using databases of information the investigator may have available (e.g., the ability to infer a name from a home address using a local authority database). Each inference has an associated description, which explains how to perform the inference; in essence, what process does an investigator need to perform the inference?

Confidence can be propagated along an inference. In other words, given an input element, it is possible to calculate the probable confidence of the result of the inference. The discussion of this is beyond the scope of this paper, but the reader is referred to [24,25].

The model results in a directed graph with the vertices representing elements of identity and the edges representing inferences. The inferences are annotated; these annotations are dimensions that are used to describe a number of the inferences' characteristics, for example, whether a transform can be automatable, how long it takes to perform the inference, whether the inference uses classified technology, etc.

The model provides a guide as to how to perform the identity attribution or enrichment tasks; while the model

Table 1 General statistics from the super identity model

Number of elements	**297**
Elements in the Biographical Domain	56
Elements in the Biological Domain	50
Elements in the Cyber Domain	157
Elements in the Psychological Domain	34
Number of transforms	**1853**
Source Element in the Biographical Domain	275
Source Element in the Biological Domain	74
Source Element in the Cyber Domain	1413
Source Element in the Psychological Domain	91
Average size of SuperIdentity*	
At 70% Confidence	5.81
At 50% Confidence	14.45
At 20% Confidence	24.27

*The number of elements that can be populated beginning from each element and following transforms until the confidence is lower than the given confidence.

can provide a description as to how to perform the task, the model will not, at present, perform an inference. The model can be thought of as a recipe book that allows users to explore identity in a number of different ways.

Casting an identification activity into the model results in a set of known elements (representing an investigator's starting knowledge, e.g., the evidence from an investigator) and a set of desired elements (representing the final knowledge required). Given this bound, the model queries the graph to work out all routes through the graph from the known elements to the unknown elements. Individual paths can then be chosen based on the paths that provide the desired elements with the greatest confidence or by using the vertex labels (e.g., the route that provides the fastest answer, routes that don't require the use of the internet, etc.).

Once a set of possible routes has been chosen, it is then possible to use the routes and the descriptions associated with each inference to lead an investigator through the steps required to perform the tasks. This output from the model then provides the guide that the investigator can use.

The model has a number of other uses in both privacy and capability management [23,25], including the ability to allow investigators' gut instinct to jump around the graph, effectively creating their own ad-hoc inferences [24,26].

The SuperIdentity visualization complements the SuperIdentity model by helping the user explore and traverse the model, encapsulating the computations performed over the model, and organizing gathered data for review and dissemination. The visualization is a workflow management application and is designed to lead the user through a step-by-step process with the model supporting the underlying computation. In this way, using the application is much like route planning. In fact, the visualizations used draw inspiration from transit maps (e.g., the London Tube Map) [27]. The user has a starting point, a desired destination, and potentially many intermediate points. By using such a metaphor, we hope to help the user anticipate the flow of the application.

The application comprises a series of screens that represent stages of an investigative process: establishing a context for the inquiry, recording known quantities and desired quantities, exploring routes from that which is known to that which is desired, navigating those routes to arrive at a result, and reviewing the results of the inquiry. Each of these is depicted in order from top to bottom in Figure 2. The application leads users through these screens with animated transitions (panning from left to right) as the investigation progresses from one stage to the next. More detailed images of each stage are provided below as the stages are discussed.

The context for an inquiry is termed a *project* in the application. A project has no associated data to begin with but accumulates data as the user begins to record his or her findings. Any data discovered, either as a desired result or an intermediate result, is recorded in the context of the project and recalled whenever the project is opened. Computations performed by the model are stored and recalled as well, but they have no effect on any other project.

Primarily, data is collected as discrete entities called *nodes*. Nodes that are constructed from values known a priori are called *seeds*. They represent the knowledge with which the user begins. Nodes constructed without a known value but that represent what the user would like to know are called *targets*. The user's interaction with the application will be a progression from seeds to targets. The first screen, then, represents the dichotomy of seeds and targets (Figure 3). This dichotomy is reinforced visually by the vertical divide in the central circle of the second screen; the circle represents the whole of the identity, with seeds on the left and targets on the right. An investigative process is a movement, using information transformations from left (the seeds) to right (the targets).

The application aids the user in recording seeds and targets by listing the representative elements from the model. Once a suitable element is chosen, if the node is a seed, the user is given the opportunity to record the known value and an approximate confidence in the accuracy of the data. Users' confidence may vary with their confidence in the source. No further data is recorded for targets because none is known; the application is designed to aid in its discovery.

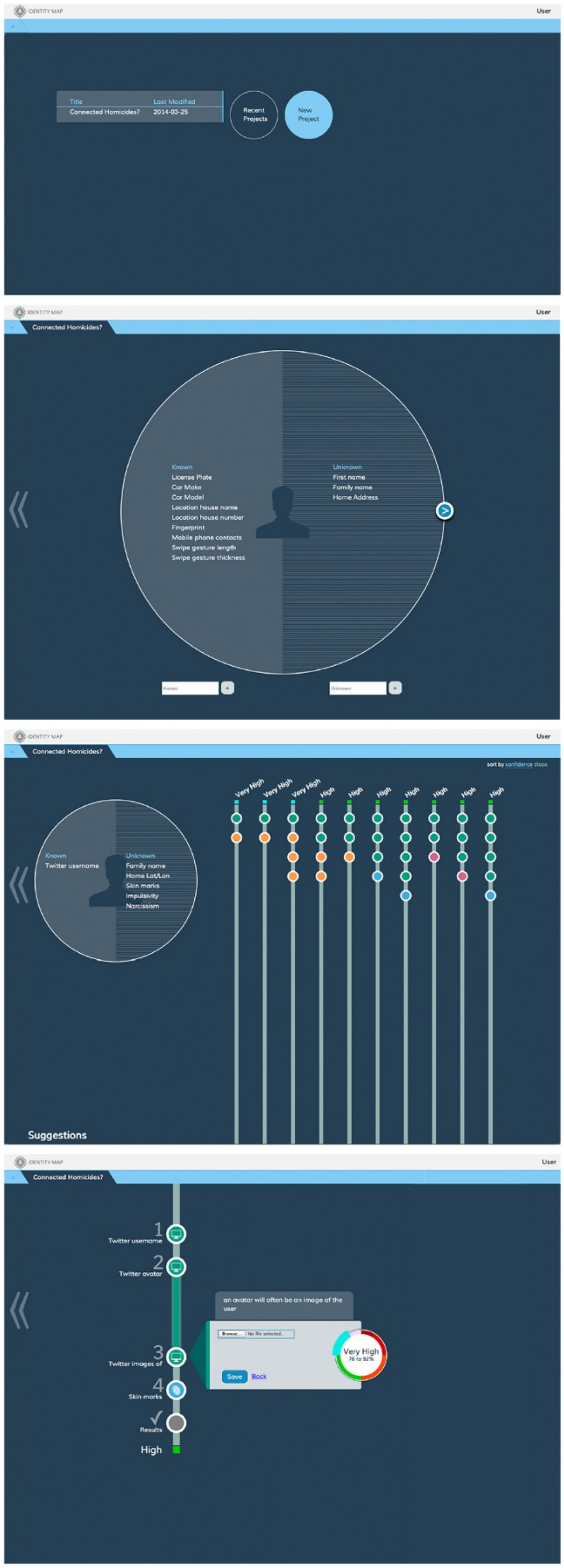

Figure 2 All four screens of the SuperIdentity application. From top to bottom: Project Create/Open, Seeds & Targets, Path Overview, and Path Walkthrough. The user progresses from left to right within the application, returning to the Overview after completing a walkthrough.

Confidence is stored in the model as a percentage value. The visualization maps these percentages into ranged categories of confidence: Very Low, Low, Medium, High, Very High, and Certain. The ranges and the terminology are configurable. Kesselman [28] shows that users make poor use of numeric ratings, so we have adjusted our visualization accordingly. To record their confidence, users select a category from a dial. The selected confidence is highlighted. The same terminology and representation of confidence are repeated throughout the interface.

Once the endpoints are established, the application can suggest paths through the model (traversals of the graph) that will lead the user from a seed to a target. Every path is composed of two or more nodes; a seed and a target are required. Nodes may be of any identification domain, as long as a transform exists between the source and the immediate target. If a direct transform exists (from the seed directly to the final target), then the path will be two nodes long. Otherwise, the model will employ intermediate nodes to complete the path.

Every transform from one node to another incurs some measure of error and thus some loss in confidence in the result. By their nature, longer paths tend to yield a lower confidence result. Paths are generated for all seeds and targets provided as inputs and a default final confidence is computed based on the encoded loss in confidence incurred by each transform. Paths are then sorted by their final confidence and presented in descending order. The paths that are likely to result in the highest confidence result are presented first.

There are times when no paths will be available from the provided seeds to the provided targets or when the paths provided are low confidence, undesirable, or nonexistent. In such cases, the model is able to suggest other seeds that may yield a higher confidence result. The user can engage those seeds by adding them on the previous screen and seeking out the necessary data, if

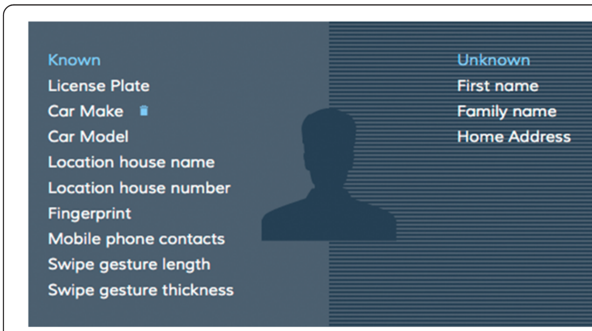

Figure 3 Seeds and targets entry screen (Also called Knowns and Unknowns). Examples of seeds and targets enumerated to populate the visualization. Seeds can have associated data and confidences; Targets have none and will be discovered through interaction.

accessible. The model will then recalculate the paths, including the added seeds, returning the higher confidence paths. Alternate seed suggestions will prefer seeds that share the closest possible relationship to existing seeds (e.g., a Facebook username seed, *John Doe*, might generate a suggestion for a Twitter username seed, *@johndoe*). If close relationships are unfruitful, the suggestions will reach out as far as a related sub-domain but no farther. For example, the Facebook username might generate a suggestion of website address but not physical address, because physical address is considered to be of a different identification domain.

The color of the nodes encodes the domain of origin (i.e., cyber, biometric, biographic, and psychological), which is reinforced with a symbolic representation when the user selects a path. A selected path presents the name of each node, its identification domain, and the order in the path. After inspecting the details, the user can progress to a detailed walkthrough of the path or select an alternate path. A selected path, with other paths juxtaposed, is depicted in Figure 4. The path is read from the top down, starting with the expected final confidence, followed by a seed, intermediate elements, and concluding at a target. "Select" takes the user to the Path Walkthrough screen for that path.

Once a path is chosen, the user walks through the path, following one transform after another, acquiring data for each node until arriving at the target node. The user interface aids in this process by presenting each node in a linear progression mimicking the Path Overview screen described earlier but providing more detail and isolating the selected path. In this view, a dialog is presented at each step that allows the user to record findings and rate confidence in the result (Figure 5).

Users are not required to enter data or record a confidence to progress. The data field is left blank; a blank data field only prevents automatable transforms from performing their task and later information retrieval (e.g., for reporting). In every other respect, the model can still function. The default confidence set is that provided by the model based on the confidence of the previous node in the path and the confidence loss due to the transform. If the user elects not to specify a confidence, this default confidence is used.

To aid the user in performing non-automated transforms, the model supplies a descriptor of means for performing the transform. As an example, to transform from the length of a person's hand to their gender, the model supplies "*Handlength and Gender are correlated—men tend to have longer hands,*" which provides a basis for inference. Some transforms are more complicated than others, but all transforms are intended to be a single step. This is where the linear nature of the walkthrough is most clearly beneficial: it dramatically simplifies the reasoning process

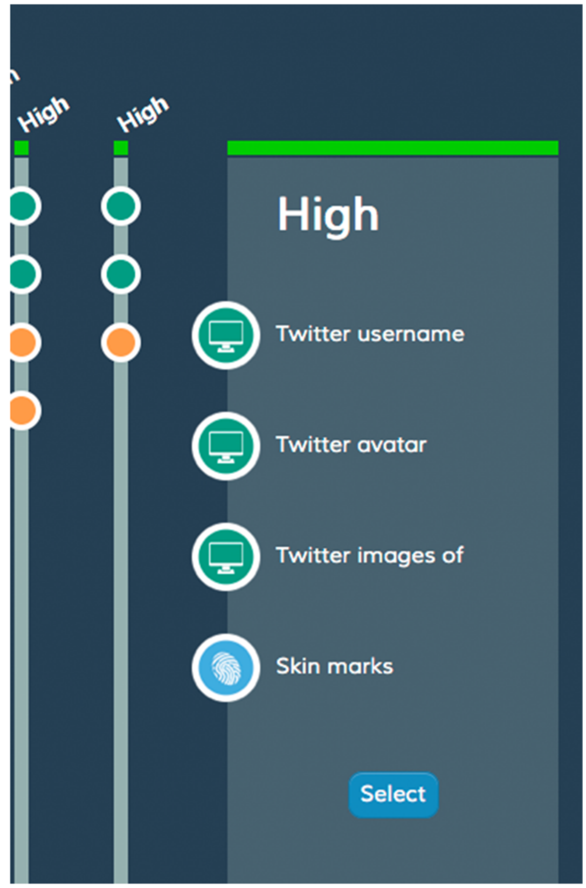

Figure 4 Paths as presented in the overview. A user can see the expanded view (right side) by clicking on a collapsed path (left side). "Select" progresses the user to the Walkthrough.

for the user, allowing them to focus on the transform tasks (which a machine cannot perform).

It may be prudent here to discuss transforms that require multiple input elements but yield a single output element. For example, biometrics on multiple attributes of a person's voice can yield inferences about certain personality traits [29]. In such cases, the path is a composite path and begins at bifurcated sub-paths that join at a later node. The linear progression is preserved by leading the user down one sub-path, then the other, and finally completing the transform at the merge node before progressing to the target node.

At the end of the walkthrough, the user is presented with a summary of the result: the user-populated target node and its confidence, other target nodes that have been populated in previous walkthroughs, and the opportunity to return to the Path Overview to pursue a new target. (The Path Overview is updated to give preference to unpopulated target nodes.) A path can be visited more than once, and a target node can be populated multiple times with potentially conflicting results.

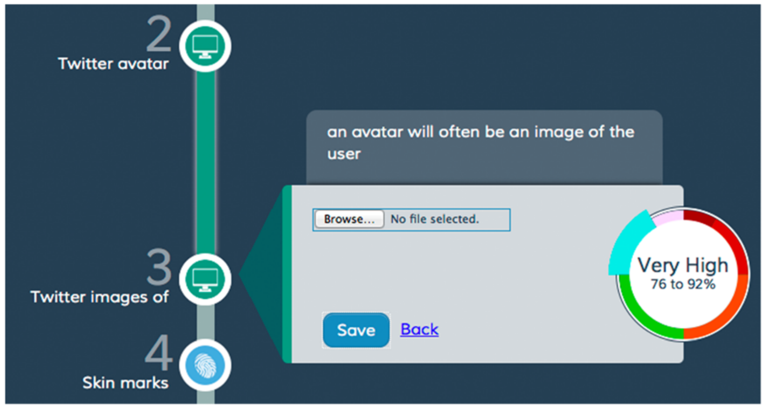

Figure 5 One transform in the Walkthrough, from "(#2) Twitter avatar" to "(#3) Twitter images of." The value and confidence are recorded in the dialog. The means of performing the transform are described above the dialog. After the final transform in a path, the dialog displays a summary of the acquired targets and their confidences.

This capability should encourage the user to explore further paths to high confidence data for the target node.

Application to use cases and future work

The construction of the model and the design of the interface are intended to support the identity attribution use cases identified through interaction with the targeted communities. To illustrate, we now consider the role of the visualization in support of the "Connected Homicides?" use case.

The connected homicides use case does not have a strong temporal element as other law enforcement cases typically have. There is sufficient information that can be considered *known items* including the partial license plate, description of the car, and geolocation, as well as data that can be retrieved from the phone recovered: swipe gesture, fingerprints, contacts, user names, and other content. The unknown items of identity revolve around identifying information to discover any correlations between the perpetrators. In particular, the user wants to know if the perpetrator in each case is really the same person.

These known items form the start points in the model; the model is queried to provide the possible paths from the start points to required information. The output of the model is the set of paths that can be followed in order to solve the problem, as described in the *SuperIdentity model and visualization* Section. The current instantiation of the model emphasizes cross-domain transforms—that is, making connections from one identity domain to another. For example, the model suggests connections from the contents of the phone to a probable home address. To accomplish this, the model presents a path leading from the phone to photos taken by the user that are geo-tagged to a location or area that has a strong grouping. Other examples of paths provided by the model, with the given seeds and targets, are in Table 2. This set of paths provides not only simple paths that an investigator is used to using but also unconventional paths that an investigator may not be used to using. This is particularly useful should an investigation stall or an investigation not proceed as expected.

The phone and the car establish an interesting pair. From the license plate, the user should be able to acquire the vehicle registration, and then the name of the person to whom the car is registered. The phone should also yield a name, stemming from the account with the phone company. Given the scenario, these names will not match, which will lead the investigator down an uncertain path: Were both people involved? How? Who should be brought in for questioning? What other evidence can be gathered in support of one or another suspect?

The model yields further paths that can aid in this process, and alternative paths can be suggested that use alternate start points. This capability can aid stalled investigations by suggesting other evidence that can be gathered to unlock new paths. Alternatively, the model

Table 2 Conceptual paths that demonstrate prototypical model output

Path	Confidence
Phone contents → Geo-tagged photos → Home address	Medium
Phone number → Cellular account → Person name	Certain
License plate → Car registration → Person name	Certain
Phone contents → Social media account names → Content of recent posts* → Indicators of mental health	Low
Phone contents → Social media account names → Content of recent posts* → Indicators of motive or intent	Low
Phone → Swipe gesture arc → Handedness	High
Phone contents → Contact List → Association with victim	Medium

*This assumes the information is either publicly available or legally accessible.

can suggest other paths that can be used to clarify who was involved in the homicides by taking multiple independent paths and providing consensus over a particular element of identity.

We have explored the use of this application to address a use case provided by law enforcement. In the process, both its strengths and weaknesses were revealed. In revealing cross-domain opportunities to complete either attributive or enriching identity tasks, the results were good. These are investigative pathways not often used by law enforcement. The application also benefits from its simple, linear progression, reducing the complexity of the statistical model to a form friendly to all users.

It is good to recall, however, that the application is designed with more than just law enforcement in mind. The collection of use cases came from analysts in intelligence, cyber security, and law enforcement. While there is some overlap in the use cases and similarity in the identification needs of each community, there are also significant differences. This application seeks to benefit all attribution and enrichment tasks in a general way. This generic approach can have negative impacts on particular use cases. As an example, in the "Connected Homicides?" exemplar use case considered, the car and the phone produced two individuals of interest, but the application provided no support for dividing the investigation or otherwise associating some pieces of identity with one individual, and others with another.

The model and the visual representation are sufficiently modular and general to be repurposed for applications targeted at particular communities or even particular investigative methodologies. Considering the use case example presented here, one could envision customized applications that contain legal aspects pertinent to a particular law enforcement region or task. This approach would facilitate the choice of paths based on what information is acceptable as evidence. We did not investigate those possibilities in this body of work and consider them an exercise for future work. The following sections discuss future work with the aim to consider further iterations of the model in the modular or general sense, rather than a strictly law enforcement focus.

Critical path

An overview of the collection of paths leading from all seeds to all targets is essential to grasp the broader picture of an investigation; it is termed the *critical path*. It highlights the nodes that are essential for a user to complete his or her task and reveals transforms that suffer greater losses in confidence. Some nodes have a greater number of paths that pass through them. Some edges have a higher possible confidence when transforming to the next node in the path. Critical information for the user might be intermediate nodes that are required for all paths to a target or collections of transforms that are necessarily weak, with no alternatives. For instance, Table 1 showing conceptual paths for the "Connected Homicides?" use case suggests that social media account names may be an essential node for inferring information about phone content, albeit with a low confidence. Such information reveals both the strengths and weaknesses of the investigators' position and can help them adjust to accommodate.

Automated transforms

Many transforms can be automated. Some transforms are simple, like approximating height from hand length or acquiring a social media avatar given a user name; they're simple functions that are easily performed by a computer. More importantly, they are rudimentary tasks for a user, and they should be relieved of such duties. Some transforms are not simple but can still be automated—for example, performing facial recognition or social network traversals. From our use case, automating the transformation from the swipe gestures on a phone to indicate handedness is automatable but would require specialized software. Transforms of this category require more sophisticated automation engines and likely more sophisticated interfaces for the user to interact with. Some transforms cannot be reasonably automated.

Every automated transform incurs some loss in confidence, just as when a user performs the same transform, they may be more or less confident in the result. A username to an account avatar is a high-confidence transform, while facial recognition software is beneficial but far from perfect, yielding perhaps a moderate to low confidence depending on the quality of the inputs. When users engage an automated transform, they are also notified of the resulting confidence, allowing them to compensate or reconsider their investigative trajectory.

The present work has planned for the presence of automated transforms, but they have not yet been introduced to the user interface and exposed as an executable option. As more research from the UK institutions in the consortium matures, we hope to introduce more opportunities for users to engage automated processes to complete portions of their work.

Review screen(s)

At present, the SuperIdentity application provides for a shallow review of gathered data for the user: the seed-target summary at the end of a path walkthrough. While this is a beneficial reminder to the user of overall progress, it is insufficient for important tasks like reconciling element discrepancies and report generation. When the user encounters multiple results for a single element, there will need to be some means of exploring the evidence supporting or refuting each. Little work has been done yet to design an approach, and it is unclear what level of detail is required for effective use. Requirements may be gathered in further interactions with the various user bases.

Reporting is supported in the sense that the model retains all the nodes and their relationships to each other. It also preserves node provenance, including metadata (e.g., whether the node was populated by the user or an automated transform). So the means are there to recall any and all data users require to generate reports of their findings, but no user interface requirements have been compiled or considered as yet. This is another topic for follow-up interactions with users.

Evaluation

Both the described exemplar use cases as well as others developed will be used in evaluating the utility and usability of this research. For each of the three analysis domains (intelligence, law enforcement, and cyber security) five to seven experts will be recruited for each evaluation study. In the studies, the experts will be asked how they would currently do the task, the resources they would use, the certainty they would have in the result assuming the necessary information needed could be obtained. Using the SuperIdentity model and visualizations, a set of experts will explore and evaluate the model's capabilities through solving the exemplar use cases. During the exercise, the experts will be asked to talk out loud, providing a more immediate evaluation of the model as they move through the exemplar use cases. They will be asked which paths they will take and the rationale for their decision, including their views on the confidence levels associated with pathways in the model. After the walkthrough, participants will be asked to rate both the utility of the model and the usability of the various visualization components. This information will be used by the visualization and model teams to make appropriate changes. It is also important to get reactions as to which information the experts feel comfortable using now, what information they deem to be problematic, and how they may be able to use their valuable tacit knowledge while using the model.

Conclusion

In conclusion, we have presented a collaborative effort from PNNL and Oxford, as well as other UK institutions, that seeks to provide a capability to investigators for enriching identities and attributing actions to individuals. Oxford has developed a robust model for traversal of paths between interconnected identity elements, statistical computation of approximate levels of confidence in data, and maintenance of investigative provenance. PNNL has developed a visual analytic interface to make the model accessible, so that investigators can leverage its capabilities without fully comprehending the complex nature of the model.

These pieces were designed and built from requirements gathered through interviews with analysts from multiple domains. We explored those for law enforcement, highlighting one particular use case that could demonstrate the capabilities of the model. It not only revealed particularly how the model can cross identity domains but also revealed opportunities for specialization of the tool in future iterations. We also discussed plans for improvement and evaluation going forward.

Abbreviations

EPSRC: Engineering and physical sciences research council; PNNL: Pacific northwest national laboratory; CAPER: The collaborative information, acquisition, processing, exploitation and reporting project; CRM: Caper regulatory model; IACP: International association of chiefs of police.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JB: Development of user interface, writing of introduction, visualization, application, and future work sections, and conclusion; lead author. JS: Conducted and compiled user interviews, writing of use case section, and general editing. DH: Development of the model, writing of the model section. LE: Conducted and compiled user interviews, writing of use case and related work sections. DSF: User interviews, minor writing contributions. SC: Research oversight. OL: User interviews, user interface design, research oversight. All authors read and approved the final manuscript.

Acknowledgments

Work performed at PNNL (user interviews and visualization) was supported by the US Department of Homeland Security. PNNL is managed for the US Department of Energy by Battelle under Contract DE-AC05-76RL01830. The work performed at Oxford (user interviews and model development) and Bath (user interviews) was performed under the Engineering and Physical Sciences Research Council (EPSRC) grant EP/J004995/1. The SuperIdentity project is investigating the interactions between offline and online identities; the cross-disciplinary consortium ranges from innovative new biometric measures through to management of online identities. The authors would like to thank our colleagues in the Cyber Security Centre at the University of Oxford, particularly Michael Goldsmith, Jason Nurse, Thomas Gibson-Robinson, and Elizabeth Phillips, whose early work developing a transitivity model for relating identity elements has been instrumental in developing the SuperIdentity model. We would also like to thank Chris Bevan, who helped with the early design of the protocol for the interviews. Our thanks also to colleagues at PNNL: Bill Pike for his innovative leadership, Dee Kim for her contributions to user interface design, and Art McBain for his development efforts.

Author details

[1]Pacific Northwest National Laboratory (PNNL), Richland, WA, USA. [2]Oxford University, Oxford, UK. [3]University of Bath, Bath, UK.

Received: 2 April 2014 Accepted: 16 August 2014
Published online: 18 September 2014

References

1. AE Marwick, Online Identity, in *A Companion to New Media Dynamics*, ed. by J Hartley, J Burgess, A Bruns (Wiley-Blackwell, Oxford, UK, 2013), pp. 355–364
2. J James, How Much Data is Created Every Minute? in *Domosphere*, 2012. http://www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/
3. The Government Office for Science, *Foresight Future Identities (2013) Executive Summary*, 2013. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/273968/13-524-future-identities-changing-identities-summary.pdf
4. S Saxby, The 2013 CLSR-LSPI seminar on electronic identity: The global challenge – Presented at the 8th International Conference on Legal, Security and Privacy issues in IT Law (LSPI) November 11–15, 2013, Tilleke & Gibbins International Ltd., Bangkok, Thailand. Comput. Law Secur. Rev. **30**, 112–125 (2014)

5. Caper, in http://www.fp7-caper.eu/
6. Secure Identity Across Borders Linked (STORK), in https://www.eid-stork.eu/)
7. 2012 IACP Social Media Survey, in http://www.iacpsocialmedia.org/Portals/1/documents/2012SurveyResults.pdf
8. Social Media and Tactical Considerations For Law Enforcement, in https://info.publicintelligence.net/COPS-SocialMedia.pdf
9. ML Gavrilova, RV Yampolskiy, Applying Biometric Principles to Avatar Recognition, in *2010 International Conference on Cyberworlds (CW)*, 2010, pp. 179–186
10. K Gani, H Hacid, R Skraba, Towards Multiple Identity Detection in Social Networks, in *Proceedings of the 21st International Conference Companion on World Wide Web* (ACM, New York, NY, USA, 2012), pp. 503–504. WWW '12 Companion
11. R Zheng, Y Qin, Z Huang, H Chen, Authorship Analysis in Cybercrime Investigation, in *Proceedings of the 1st NSF/NIJ Conference on Intelligence and Security Informatics* (Springer-Verlag, Berlin, Heidelberg, 2003), pp. 59–73. ISI'03
12. C Aliprandi, A Marchetti, Introducing CAPER, a Collaborative Platform for Open and Closed Information Acquisition, Processing and Linking, in *HCI International 2011 – Posters' Extended Abstracts*, ed. by C Stephanidis (Communications in Computer and Information Science, vol. 173, Springer Berlin Heidelberg, 2011), pp. 481–485
13. The Uncertainty of Identity Multidisciplinary Project, in http://www.imprintsfutures.org/about/
14. M Adnan, G Lansley, PA Longley, A geodemographic analysis of the ethnicity and identity of Twitter users in Greater London, in *Proceedings of the 21st Conference on GIS Research UK (GISRUK)*, 2013
15. P Juola, Authorship Attribution. Found. Trends. Inf. Retrieval. **1**, 233–334 (2007)
16. DA Wheeler, GN Larsen, *Techniques for Cyber Attack Attribution* (Institute for Defense Analyses, Alexadria, VA, 2003)
17. J Hunker, B Hutchinson, J Margulies, Role and challenges for sufficient cyber-attack attribution, in *Institute for Information Infrastructure Protection*, 2008
18. State and Major Urban Area Fusion Centers, in http://www.dhs.gov/state-and-major-urban-area-fusion-centers
19. UK Cabinet Office, *Cyber Security Strategy*, 2011
20. S Creese, M Goldsmith, JRC Nurse, E Phillips, A Data-Reachability Model for Elucidating Privacy and Security Risks Related to the Use of Online Social Networks, in *2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 2012, pp. 1124–1131
21. RR McCrae, PT Costa, Validation of the five-factor model of personality across instruments and observers. J. Pers. Soc. Psychol. **52**, 81–90 (1987)
22. DL Paulhus, KM Williams, The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. J. Res. Pers. **36**, 556–563 (2002)
23. D Hodges, S Creese, Building a better intelligence machine: A new approach to capability review and development, in *2013 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 2013, pp. 113–115
24. D Hodges, S Creese, M Goldsmith, A Model for Identity in the Cyber and Natural Universes, in *Intelligence and Security Informatics Conference (EISIC), 2012 European*, 2012, pp. 115–122
25. D Hodges, S Creese, Breaking the Arc: Risk control for Big Data, in *2013 IEEE International Conference on Big Data*, 2013, pp. 613–621
26. D Hodges, J Nurse, M Goldsmith, S Creese, *Identity attribution across CyberSpace and the Natural Space*, 2012
27. J Vertesi, Mind the Gap: The London Underground Map and Users' Representations of Urban Space. Soc. Stud. Sci. **38**, 7–33 (2008)
28. RF Kesselman, Verbal Probability Expressions in National Intelligence Estimates: A Comprehensive Analysis of Trends from the Fifties through Post 9/11, in *MCIIS Theses in Intelligence Studies. Mercyhurst College Institute for Intelligence Studies (MCIIS)*, 2008
29. G Mohammadi, A Vinciarelli, M Mortillaro, The voice of personality: Mapping nonverbal vocal behavior into trait attributions, in *Proceedings of the 2nd international workshop on Social signal processing. ACM*, 2010, pp. 17–20

doi:10.1186/s13388-014-0012-6
Cite this article as: Bruce *et al.*: Pathways to identity: using visualization to aid law enforcement in identification tasks. *Security Informatics* 2014 **3**:12.