Tools for Understanding Identity

Sadie Creese*, Thomas Gibson-Robinson*, Michael Goldsmith*, Duncan Hodges*, Dee Kim†, Oriana Love†,
Jason R.C. Nurse*, Bill Pike† and Jean Scholtz†
*Department of Computer Science,
University of Oxford, UK Email: firstname.lastname@cs.ox.ac.uk
†Pacific Northwest National Labs,
Richland, WA Email: firstname.lastname@pnnl.gov

***Abstract*—We present two tools for analysing identity in support of homeland security. Both are based upon the Superidentity model that brings together cyber and physical spaces into a single understanding of identity. Between them, the tools provide support for defensive, information gathering and capability planning operations. The first tool allows an analyst to explore and understand the model, and to apply it to risk-exposure assessment activities for a particular individual, e.g. an influential person in the intelligence or government community, or a commercial company board member. It can also be used to understand critical capabilities in an organization's identity-attribution process, and so used to plan resource investment. The second tool, referred to as Identity Map, is designed to support investigations requiring enrichment of identities and the making of attributions. Both are currently working prototypes.**

I. Introduction

Identity in the 21st century spans cyber and physical spaces in a complex way, with many of us maintaining multiple personas related to our personal and work lives. In truth, modern identity is a complex entanglement of expressions in both spaces, and group identity is an amplification of some common aspects and an attenuation (sometimes) of others.

Increasingly, identity is projected through on-line media – most obviously by way of social networks such as Facebook and Twitter. Societies are becoming less passive consumers of media (traditionally through television or radio) with many more people enabled to produce, share and promote their content, whether this content is for personal enjoyment, or to disseminate a political or religious 'message'. In fact we are witnessing information operations on a massive scale.

These fundamental changes in what is meant by 'identity' have changed how we should go about understanding it; no longer can we simply consider off-line identity attributes in isolation of on-line attributes (or vice-versa). We must now acknowledge a holistic identity covering **all** aspects of an individual's identity, whether projected in the natural world, cyber world or indeed in psychological domains, as this drives behavior which in turn can be considered the central precursor to all on and off-line identities.

Identity attribution, where a behavior or action is attributed to an identity, and *identity enrichment*, where a small set of an individual's identity expressions is used to derive a more in-depth picture, are common in a variety of exercises within the law-enforcement and intelligence communities, but also by a large number of commercial entities. In previous work [1–3] we have documented a model of identity and identity-enrichment capability which can be used to explore these holistic identities (or what we refer to as superidentities).

In this paper we focus on the operationalization of the model in two tools. The tools can be used to explore the application of the model in three 'modes'. Specifically, Section III focuses on the tool support for exploring identity for defensive purposes and also understanding the identity-enrichment capabilities encapsulated in the model. We then outline the use-cases that inspired the second tool's design in Section IV, followed by a description of our concept for tool support in investigations in Section V. We present our conclusions and avenues for future work in Section VI. But first we provide more background on the context of the Superidentity model.

II. Superidentity Model

The model considers *elements* of identity which are simple facets of an individual's identity. Typically, these are either Cyber (e.g. a username on an Online Social Network (OSN), email), Biographical (e.g. real name, date of birth), Biometric (e.g. fingerprints, eye colour) and Psychological (e.g. models of personality such as Big-5 [4] or dark- triad [5]). An individual's superidentity is a multi-set of these elements.

The model also includes transforms or inferences which are used to create new elements of identity from existing elements. For example, within open-source literature, inferences exist to infer an individual's height from their shoe-size [6], a user's Flickr username from their email address [7] or an individual's level of Machiavellism from their Twitter content [8]. There is a confidence associated with each of these elements and transforms and there is a framework in place for modeling both the propagation of confidence and the resolution of 'conflict' within the superidentity. These aspects are outside the scope of this paper, but we refer the reader to [1–3].

Initially, the model was developed in order to provide a mechanism for individuals to attempt to 'manage' their superidentity [1], by understanding what elements of identity they wish to keep private and being aware of the mechanisms an adversary could use to attain these. As organizations become more aware of the need to better protect their cyber-assets, increasingly attackers are targeting individuals; spear-phishing has now evolved into 'laser-guided' spear-phishing and it is becoming more important that organizations are cognizant of the attack-surface presented by its employees. In support

978-1-4799-1535-4/13/$31.00 ©2013 IEEE

of this a number of experiments were performed using the model on OPSEC-aware individuals. These experiments began to demonstrate the application of the model in an identity enrichment (intelligence gathering) mode.

We have also discussed previously how the model might be able to describe capability within an intelligence production environment [9]. This article particularly focused on the benefits of providing a data-level model of an organization's capability, without a cloud of business-processes or skills grouping which often obfuscate the truly important processes involved in generating actionable intelligence. The data-level modeling also provides a clear way to map collaboration across different organizations and also encapsulate the capability provided by innovative, creative counter-cultures within these sorts of organization.

These previous papers describe the three basic modes of operation, namely: defensive, information-gathering and capability planning. We use the same underlying model to support all three of these modes of operation. This is key to using the feedback from the 'investigative' and 'defensive' modes to further improve and understand our modeling of capability, within the context of the organization in which it is deployed. As we capture the behavior of users and feed this back into the model, we effectively drive capability-development from the real-life on-going use-cases.

There are a number of products in the marketplace that provide some of the capability described above, notable Maltego [10] and IBM's InfoSphere Identity Insight platform [11]. However, our approach of maintaining a simple, intuitive, consistent model for identity across these three modes of operation permitting feedback and interaction across the broad surface of identity is unique. Further, none of these existing tools are underpinned by a mathematical model with accompanying analysis potential.

III. Tool for understanding a model of identity

In order to protect an individual against a targeted attack that is attempting to deduce as much personal information about them as possible, it can be very useful to analyse the overall effect of the model's inferences (recall an inference allows new elements to be derived — e.g. an inference might specify that address can be derived from name). In particular, by considering how the inferences combine (i.e. chain to allow new derivations of elements), it becomes possible to deduce how best to defend an individual. For example, it may well be the case that there is one inference that is critical for an attacker to make, since it is the only way of deducing a certain element. Thus, if a defender is aware of this through the use of our model, they are able to assess how best to protect an individual against such attacks.

It can also be useful to consider how the inferences are able to combine in order to guide an analyst assuming the role of an attacker towards the best way to attack a certain individual, and thus to defend against the attack. However, as more inferences are added to an identity model, it often becomes difficult to reason about how the inferences combine, thus making tool support necessary.

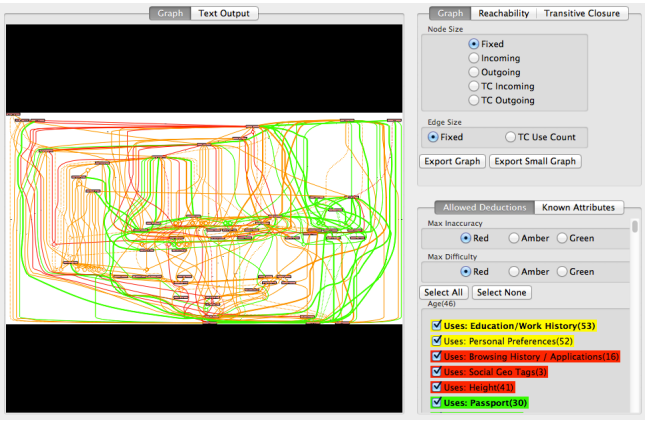

Fig. 1. The main window of the tool, showing the visualization of the inferences.

To help address this problem we have developed a tool that allows the overall effect of the model's inferences to be analyzed. In particular, the tool allows a user to easily determine, given a particular set of inferences and elements, which elements or inferences are of most value to an investigator (whether a defender or an attacker). The tool is also able to take into account the *difficulty* of the inferences (the difficulty of an inference is an estimation of what resources the inference requires), allowing the investigator (or person tasked with defending an individual) to consider different scenarios depending on the perceived resources of the attacker, or to prioritize lines of investigation. It also has a number of interesting visualizations that are designed to aid the user in understanding the model.

The tool's main visualisation, as shown in Figure 1, displays the inferences as a graph. In the simplest case, where each inference is of the form *given* A *deduce* B, each node in the graph represents an element and each edge represents an inference, as illustrated in Figure 2. The nodes are labeled with the element name, whilst the edges are labeled with the identifier of the inference. The edges are colored according to their accuracy (with green edges being the most accurate and red the least), whilst the edges are bolded according to their difficulty. More complex inferences, of the form *given* A *and* B *deduce* C, are displayed by introducing a new node that represents the inference. This node has incoming edges from each of the required elements (in the above example, A and B) and precisely one outgoing edge to the deduced element (in the above example, C).

The tool allows the graph display to be customized towards a particular task by manipulating the options on the right hand-side of Figure 1. In particular, it is possible to only display edges that correspond to inferences that have a certain minimum accuracy or maximum difficulty. Further, a particular desired target element can be highlighted, in which case only deductions or elements that are of use in deriving that element

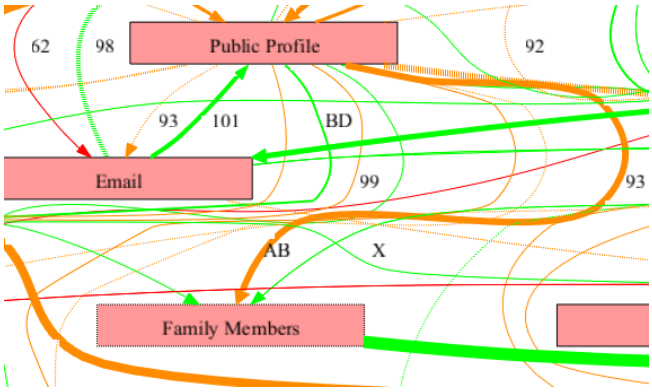

Fig. 2. A small section of the inference graph, in which the edge width is sized according to the number of times the inference is used in the transitive closure. In the above, the inference from Email to Public Profile is more important than the inference from Public Profile to Email.

are shown.

The tool also supports a number of more sophisticated analyses that are implemented using the *transitive closure* of the inferences, which is conceptually the set of all possible inferences that are formed by chaining together one or more of the original inferences. For example, suppose we denote an inference that obtains B from A as $A \rightarrow B$, and have the two inferences $A \rightarrow B$ and $B \rightarrow C$. The transitive closure of these two inferences will consist of both of the preceding inferences, and the inference $A \rightarrow C$.

The tool is able to present the information from the transitive closure in several different ways. For example, the nodes of the graph, as displayed in Figure 1, represent the elements of identity and can be sized according to the number of times they appear in the transitive closure. This provides some information as to which elements are the most crucial elements to obtain. Further, the edges, which represent the various inferences, can be sized according to the number of times they appear in the transitive closure, as illustrated in Figure 2. This provides an indication of which inferences are most useful to an attacker, and, therefore, where defensive measures should be concentrated.

We do not compute the full transitive closure since it is possible for the transitive closure to be exponential in the size of the original set of inferences. Unfortunately, our experiments confirm that this happens frequently with real-world examples. Therefore, we instead only keep the best, according to some heuristic, inference for any set of input and output elements, thus ensuring that the set of inferences never grows too large. This strategy has been highly effective in practice, and manages to prune down the transitive closure of several hundred inferences to just a few thousand, whilst retaining useful information.

Clearly, the choice of the heuristic to decide which inference is better is critical in making sure that the transitive closure is relevant. The most effective heuristic we have found compares the difficulty, accuracy and the set of required elements in order to decide which one is better. In particular, given two inferences i_1 and i_2 that deduce the same element, we say that i_1 is *better than* i_2 providing i_1's accuracy is higher, or i_1's difficulty is lower, or i_1 requires fewer elements in order to be successfully executed. If a particular analysis task required it, a different heuristic could be picked that accomplished a certain goal. For example, if the difficulty of an inference was not important (i.e. the perceived attacker has almost unlimited resources as it is a nation state), then the difficulty could be ignored when considering which inference was better.

Having presented our tool support for exploring identity for defensive aims and understanding the identity-enrichment capability encapsulated by the model, we now consider the tool support devised for investigative purposes.

IV. Use-case definition

To inform and inspire research, provide real-world context, and inform a vision of the identity attribution, use cases were gathered from individuals within the law-enforcement and intelligence communities. By conducting semi- structured interviews with nearly twenty individuals (8 US Law Enforcement, 8 US Human and Cyber Intelligence Analysts, 1 UK Missing Persons Analysts, and 1 UK Financial Services Fraud Analyst), common identity-attribution processes were derived and represented as use cases. Each use case was partitioned into: its overall identity-attribution mission, the tasks involved, the sources of data used, turnaround time needed, certainty needed in resulting identity, starting/known elements of identity, common transitional elements of identity, and ending/unknown identity elements.

Through the discussions with identity attribution practitioners, the processes used, the most common starting elements, and most desired goal elements were documented. Then, the inferences within our model that were currently being researched could be evaluated against the most common identity-attribution exercises. This evaluation exercise helped in exposing areas of strength and weakness within the inference capability under development relative to the real-world tasks of relevant user communities. Understanding the areas of weakness within the instantiated superidentity model (i.e. those that produce low confidence paths with a high number of steps) would help decisions be made surrounding research objectives between the Cyber, Biometric, Biographical and Psychological domains.

To succinctly represent the key processes and commonalities among these attribution tasks, we summarized the use cases gathered within two exemplar scenarios: *From Username to Person* and *Individual within a Crowd*. The exemplar scenarios were motivated and crafted from the use cases in order to highlight the domain diversity of the identity-attribution tasks of law enforcement and intelligence.

A. Exemplar Scenarios

From Username to Person. A suspicious article was posted online that catches the attention of the intelligence community. The IP address was tracked to an internet café in a large city.

At this café, several incomplete data points (elements) were collected: low-quality surveillance video from the past two weeks, hundreds of fingerprints, and credit card information of several café patrons. In addition, the username of the article author, the text written, the blog site and several user comments were collected. The host of the blog was not able to share any additional information. The investigator wishes to understand who this person is (and quickly). In particular, the analyst would like to know the identity of the user, if the account is shared or individually owned, the associates of this person, the individual's IT skill level, age, gender, and ideology.

Individual Within a Crowd. A public protest has just begun unexpectedly at a well-known area of downtown. Law enforcement is working to identify the individuals of interest within a crowd in an effort to mitigate any issues, but only know about this group's views and leadership based on their vocal and unsettling on-line presence within public on-line discussion forums. Low quality video surveillance is being leveraged to help with monitoring and is doing a good job capturing the features of most participants within the crowd. Law enforcement's challenge is to understand how the participants within the crowd map to the actors within the group's on-line discussion forum.

The first exemplar user scenario, *From Username to Person*, explores tracking suspicious cyber-activity to the individual's actual name. As several of our user scenarios started within the Cyber domain – with a username and an activity attributed to this username, we deemed it necessary to have an exemplar scenario which addressed the transition from the Cyber to Biographical domain.

In contrast, *Individual within a Crowd* explores the transition from the Biometric domain to the Cyber domain. Given a person (or several persons) of interest within a crowd, can it be determined who this person is within cyberspace? In this scenario, we explore how low-quality surveillance video – which detects several individuals' faces, walking gait and other Biological markers of individual – may be leveraged in the Cyber domain to understand if a person is a member of a radical protesting group, in order to mitigate violent actions.

Exemplar Scenario Name	Domains
From Username to Person	• Cyber (username, writing samples – look at this over a period of time, account sharing); • Biographical (location, associates, real name, expertise, age, gender, credit card information); • Psychological (ideology – look at this over a period of time); • Biometric (fingerprint, gait, face).
Individual within a Crowd	• Biometric (gait, height, facial features, observable features); • Cyber (discussion groups, social friend/follower network analysis); • Psychological (ideology); • Biographical (arrest record, real name)

TABLE I
SUPER IDENTITY ELEMENT DOMAINS REFERENCED BY THE SCENARIOS

The exemplar user scenarios allow the research of the identity model to be directed towards the most relevant tasks, domains, and identity elements for our stakeholder analyst communities. Table I highlights the domains and the correlating identity elements of each of the two scenarios. The most apparent utility of the exemplars has been realized within the user interface for the model's investigative uses, Identity Map.

V. TOOL FOR USING A MODEL OF IDENTITY

To help law enforcement and intelligence analysts exploit the model to perform identity attribution, the Identity Map tool was designed. Drawing inspiration from subway map designs, the user interface, as shown in Figure 3, adopts a suitable, directional path paradigm. By presenting the user with different paths from known to unknown elements of an individual's identity, users can select the workflow of highest interest based on any set of route attributes: confidence, number of steps, or route domain diversity/homogeneity. In its first iteration, Identity Map will not connect to databases or APIs to populate element values. Instead, Identity Map is intended to allow users to explore paths from the known to the unknown elements while adjusting confidence along the way.

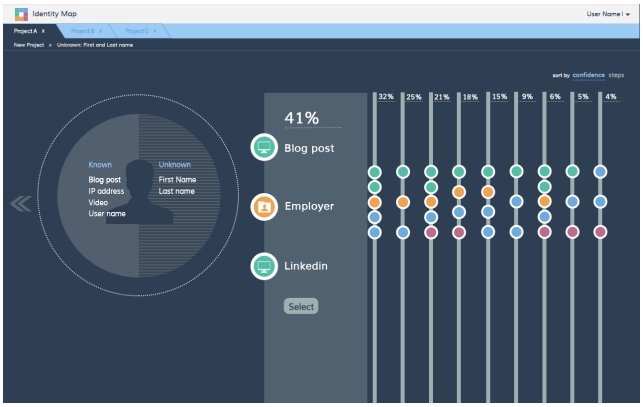

Fig. 3. User Interface Mock-up of the Identity Map currently in development.

Analysts are given the opportunity to adjust the confidence they have in the accuracy of the known elements. Additionally, they can increase or decrease the accuracy of the transforms between elements. An example of this is shown in Figure 4 with the percentage accuracy of the blog post easily entered.

The tool's workflow is as follows. To begin, analysts enter the known elements of identity. Using the exemplar scenario *From Username to Person* as our inspiration, Figure 3 shows that an analyst entered her known elements as Blog Post, IP Address, Video and Username. As each element is added, analysts have the option to add in the value of the elements: analysts can copy and paste an entire blog post, add a username, the value of an IP Address, or upload a video. In lieu of connecting directly to the data sources, we are allowing analysts to add this evidence for convenience and direct reference within Identity Map. Additionally, analysts can document

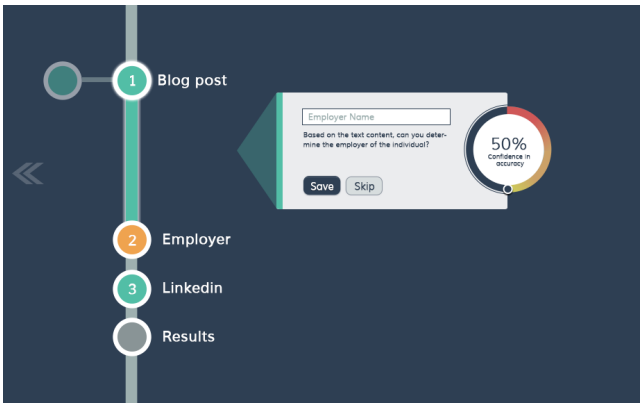

Fig. 4. Confidence configuration options offered throughout Identity Map.

the perceived accuracy of the data given. In *From Username to Person*, an analyst may have high confidence in the username and blog post elements as those are the two pieces of evidence prompting the investigation. Conversely, lower confidence may be associated with an IP Address (which can be spoofed) or in a video as it may be incomplete or obscured.

Once all the known elements are entered, the desired, but unknown, elements are added. The use cases strongly suggest that there is a primary unknown element (e.g. real name) and periphery unknown elements that will help further describe the person but are easier to obtain after getting the primary element.

Based on the priority of the primary unknown element, the ten highest confidence paths through the graph are returned (see Figure 3). The nodes within the paths are colored to represent the different domains (green – Cyber, yellow – Biographical, blue – Biometric and pink – Psychological). Each line between the nodes represent an investigation step (i.e. model transform) needed to transition close to the goal unknown element. The first node represents the known element leveraged while the last node of the path represents the resulting discovery of the unknown identity element.

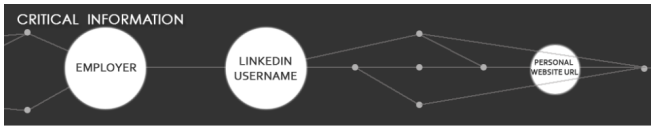

Fig. 5. Critical information path.

The 'Critical Path' section of the user interface provides an overview of the common transitional identity elements needed across different paths. For example, in Figure 5, it can be deducted that one must discover the values for the elements Employer and LinkedIn Username to progress through any of the top ten confident paths. Many of the paths also share other elements, like Personal Website URL, but are not required across all paths. The critical path will help users understand if the investigation through this approach to superidentities is worth pursing as potential bottlenecks are highlighted prior to investing time working a path.

Once an analyst selects a path, the screen animates to reveal more details surrounding that path. Another key benefit of this approach is that the instantiated identity model aims to reveal insightful paths that analysts may not always consider, but still reveal a means to the end element. Instructions regarding transitioning from the known to the first derived element are shown to help analysts get started. Each researched transition in the applied identity model impacts the overall confidence of the accuracy of the end unknown element. The analyst can adjust this default confidence to test the impact of accuracy, reflect the analyst's intuition, or knowledge external to Identity Map. If the analyst is able to find the value for the next node, they can add the value and adjust the confidence of the given transition, and proceed to the next screen. If at any point, a user is aware of the value (or wants to hypothesize a value) of a node later in the route, they can click Skip or click directly on the node to progress. As the user transitions through the screens, they will eventually arrive at a hypothesized value for the unknown element of interest, decorated with confidence of its accuracy. At this point, they may choose to corroborate this value by working alternative paths through the instantiated identity model or may transition to discover a different unknown element of interest.

VI. Conclusion and Reflection

We have presented an overview of two tools currently under development which are designed to help analysts understand the nature of identities that span cyber and physical spaces. This is done through exploring the vulnerability that individuals and related organizations may face through exposure of identity, and through providing support to investigations seeking to enrich identities and make better and more confident attributions. We believe tools such as these are crucial in analyzing identity in the support of homeland security.

The initial focus of the tool for understanding the model of identity was the development of algorithms for efficiently calculating the transitive closure of the set of inferences. The prototype interface has been constructed to help understand the potential of the approach by visualizing the inferences within the model. The tool has been used for initial explorations of example superidentities, however the design is still in a prototype stage and in order to interpret the analysis results a significant understanding of the model is required.

Identity Map has been created using a user-centric design process, developing example use-cases and personas to provide a clear vision for the final design. The tool is a working prototype designed to help harness the SuperIdentity model in support of live investigations. This will require some final validation through engaging with the interviewees who provided the use-cases outlined above.

As with all tool support, usability is key. In the situations where these tools will be used it is even more important to minimize the cognitive load and maximize situational awareness through the efficient and effective use of tool support

– this allows analysts to focus their attention and use their intuition where they add the most value. Future work for both tools will incorporate usability assessments.

In addition, as we evolve the underlying mathematical model to incorporate notions of context (such as freshness of data, source etc.) so too will the tools evolve. We believe that there is growing evidence of the fragility of identity and a necessity to understand the risks faced by people and organizations so they can better defend themselves. We also believe that delivery of security in cyberspace necessarily depends upon an ability to attribute malicious actions to personas and that our tools offer a real possibility for enhancing such capability.

ACKNOWLEDGMENT

We would like to thank our partners in the SuperIdentity project for their continued help and support and notably their input into creating the taxonomy of the elements of identity. The SuperIdentity project is funded by the EPSRC under grant number EP/J004995/1 under the umbrella of the Global Uncertainties initiative, and by the Department of Homeland Security in the US.

REFERENCES

[1] S. Creese, M. Goldsmith, J. R. C. Nurse, and E. Phillips, "A data-reachability model for elucidating privacy and security risks related to the use of online social networks," in *2012 International Workshop on Trust, Security and Privacy in e-Government, e-Systems and Social Networking (eGSSN-12) at 11th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (IEEE TrustCom-12)*. IEEE, 2012, pp. 1124–1131.

[2] D. D. Hodges, J. R. C. Nurse, M. Goldsmith, and S. Creese, "Identity attribution across cyberspace and naturalspace," in *International Crime and Intelligence Analysis Conference*, December 2012.

[3] D. D. Hodges, S. Creese, and M. Goldsmith, "A model for identity in the cyber and natural universese," in *European Intelligence and Security Informatics Conference, 2012. EISIC 2012*, August 2012.

[4] A. Poropat, "A meta-analysis of the five-factor model of personality and academic performance." *Psychological Bulletin*, vol. 135, no. 2, pp. 322–338, March 2009. [Online]. Available: http://psycnet.apa.org/doi/10.1037/a0014996

[5] D. L. Paulhus and K. M. Williams, "The dark triad of personality: Narcissism, Machiavellianism, and psychopathy," *Journal of research in personality*, vol. 36, no. 6, pp. 556–563, 2002.

[6] G. E. Vallandigham, "Height estimation from foot and shoeprint length," *Journal of Forensic Sciences*, vol. 36, no. 4, pp. 1134–1151, 1991.

[7] Flickr. (2012) Filckr API documentation. [Online]. Available: http://www.flickr.com/services/api/flickr.people.findByEmail.html

[8] C. Sumner, A. Byers, R. Boochever, and G. J. Park, "Predicting dark triad personality traits from Twitter usage and a linguistic analysis of tweets," in *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, vol. 2. IEEE, 2012, pp. 386–393.

[9] D. D. Hodges and S. Creese, "Building a better intelligence machine: A new approach to capability review and development," in *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on*, June 2013.

[10] Paterva. (2013) Maltego. [Online]. Available: http://www.paterva.com/web6/products/maltego.php

[11] IBM. (2013) Identity insight. [Online]. Available: http://www-03.ibm.com/software/products/us/en/infosphere-identity-insight