Investigating People: A Qualitative Analysis of the Search Behaviours of Open-Source Intelligence Analysts

Sean McKeown, David Maxwell, Leif Azzopardi
School of Computing Science
University of Glasgow
Glasgow, Scotland
{s.mckeown.1, d.maxwell.1}@research.gla.ac.uk, Leif.Azzopardi@glasgow.ac.uk

William Bradley Glisson
School of Computing
University of South Alabama
Mobile, AL, USA
bglisson@southalabama.edu

ABSTRACT

The Internet and the World Wide Web have become integral parts of the lives of many modern individuals, enabling almost instantaneous communication, sharing and broadcasting of thoughts, feelings and opinions. Much of this information is publicly facing, and as such, it can be utilised in a multitude of online investigations, ranging from employee vetting and credit checking to counter-terrorism and fraud prevention/detection. However, the search needs and behaviours of these investigators are not well documented in the literature. In order to address this gap, an in-depth qualitative study was carried out in cooperation with a leading investigation company. The research contribution is an initial identification of Open-Source Intelligence investigator search behaviours, the procedures and practices that they undertake, along with an overview of the difficulties and challenges that they encounter as part of their domain. This lays the foundation for future research in to the varied domain of Open-Source Intelligence gathering.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval: Users and interactive retrieval: Task models; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval:Specialized information retrieval: Web and social media search; H.3.4 [**Information Storage and Retrieval**]: Systems and Software: Performance Evaluation

General Terms

Human Factors

Keywords

OSINT, open-source intelligence, online investigation, search behaviour, domain-specific search, investigative process

1. INTRODUCTION

The ubiquity of Internet access has transformed many facets of modern society, substantially changing how we communicate and share information. Social media platforms, blogging sites and messaging services allow individuals to broadcast their thoughts or otherwise express themselves online. A by-product of this is that we effectively publish a great deal of biographical information on the World Wide Web, potentially making it freely available to anyone wishing to look for it. Studies of Web search logs indicate that somewhere between 4% [24] and 10% [13] of Web searches contain the name of a person, suggesting that there is some degree of public interest in seeking this type of information. Indeed, there are specialised search engines built for the purpose of finding people on the Web [28]. Regardless of whether the search is conducted from a specialised or general Web search engine, most queries pertain to low profile individuals, as opposed to celebrities [24, 28].

Social media appears to play a large role in person searches, with data showing that 66% of outclicks on a person search engine lead to social media profiles [28]. These profiles offer insights in to peoples lives which may not be available via traditional means [6]. Additionally, social media users openly share sensitive information, one striking example of which is the prevalence of gang promotion by young gang members [33]. Despite this, only 1 in 4 users of the social media platform *Facebook* adopt restrictive privacy settings [7], such that this information may be publicly available – though these policies in themselves do not completely prevent information being gathered about the user [1, 16].

In the intelligence community, intelligence gathered from open data sources is referred to as *Open-Source Intelligence (OSINT)*[1]. Information discovered from these sources, which includes publicly available social media information, may be utilised for a variety of purposes. In some cases, it may be used illegitimately, such as for stalking [1], terrorism [9],

[1]'Open-Source" refers to availability of the information to the public, as opposed to a software licensing philosophy.

or identify theft [10]. However, with the appropriate legal and ethical considerations [1, 21], these sources may be exploited in legitimate OSINT investigations, examples of which pertain to: employee vetting [5, 12], fraud detection [26], counter-terrorism [9] and gang violence [33].

The types of information sought after, as well as their uses, vary depending on the context of the investigation. Ultimately, the information is collected in order to fulfil a specific task. An example of this would be vetting an applicant for a high profile legal position. In this case, it is prudent to conduct criminal background checks, as well as to inquire as to the general nature, behaviour and disposition of the individual, as to avoid future scandal. The former tasks may focus on finding news articles and legal documents, while social media and other communication platforms can be leveraged to assess the individual's character. These findings may provide good reason to reject the applicant, or provide support for their application. In the case of terrorist investigations, personal conduct is viewed in a different context, with the social element emphasising related individuals. Who are they talking to? What are they saying? How are they communicating? Where do they go and what do they do? This type of inquiry could result in the discovery that the subject associates with known terrorists, or openly posts information relating to extremist websites. In the case of a potential corporate merger, it may be prudent to investigate the track record of the company and its executives. Public business records could be utilised in this case, which may uncover a long line of failed business, suggesting that this merger would be unwise. In all cases, vast arrays of open data sources can be leveraged in order to determine a course of action or support/refute a particular assertion.

In this paper we explore a subset of OSINT investigations which focus on collecting publicly available information about individuals. In order to examine investigator search behaviours and identify generalisable search requirements in this domain, an in-depth qualitative study was conducted utilising several analysts from a well-established company. Due to the sensitive nature of this line of work, and to protect the confidentiality of the company and ongoing investigations, all examples provided in this paper are conceptually representative of actual cases but are entirely anonymised.

The remainder of this paper is structured as follows. Section 2 explores previous work on domain specific search and the behaviour of intelligence analysts. Section 3 describes the methodology used to capture the search behaviour of sampled users in this search domain, in addition to a description of the questionnaire used as part of the semi-structured interviews. Findings are presented in Section 4, which describes the process involved, as well as search behaviour, tools used and difficulties encountered in these online investigations. Section 5 compares and contrasts the findings of this research to previous work, with a summary and future work in Section 6.

2. BACKGROUND

In this section, we will provide a brief overview of the different search behaviours observed in different domains, as well as previous work on the behaviour of intelligence analysts, in order to provide context for future discussion.

Different search domains have different properties and different user/system requirements, an examination of which allows for the development of more effective domain specific search tools [14]. For this reason, substantial research has been conducted in order to characterise the ubiquitous Web search domain – a survey of which can be found in Markey [20]. This previous work has shown that the average Web user engages in short search sessions, poses short queries, makes minimal use of advanced search functionality/relevance feedback, and typically only reviews one-to-two pages of results [25, 20]. In essence, Web users desire immediate satisfaction for minimal effort [15], i.e. they appear to subscribe to Zipf's *Principle of Least Effort* [35].

In contrast to general Web searching, some domains are characterised by the need to conduct exhaustive, time consuming, searches [15]. Some instances of exhaustive search are so large scale that they occupy multiple searchers for months at a time [2]. In exhaustive search, the goal is to find all documents which are relevant to the given information need, such that high recall is emphasised. In the patent search domain [17], this is especially important as failing to find a relevant document may have significant legal repercussions. Perhaps as a result of this, patent searchers place greater emphasis on search control, utilising more advanced search functionality [17]. This high recall requirement also appears to be a feature of the E-Discovery domain [2], where extensive document review is cited as being the primary investigative overhead. Conversely, searches conducted by software engineers [11] have relatively few relevant results, with queries containing many technical terms and acronyms. Users in this domain also place a strong emphasis on the authority of the source, which is derived from document metadata, such as site reputation, author and publish date.

Several works have focused on identifying differences between the behaviour of experts and non-experts. White and Morris [31] found that those with search expertise submit fewer queries, spend more time searching, and make more use of advanced search functionality. Additionally, these users also exhibit a higher degree of discrimination when viewing the results page, viewing fewer documents, spending less time reviewing each document, and clicking on lower ranked results.

When the emphasis is shifted from search expertise to domain expertise/knowledge, findings are similar with regards to query and session length, with the addendum that a higher number of unique sites are visited by domain experts [29, 30]. This suggests that experts are typically more invested in the outcome of the search process, even when the domain does not necessarily require exhaustive methodologies. In addition, experts have been shown to prefer more technical sources [29], while non-experts frequent commercial and consumer orientated sites [29, 30]. Liu et al. [18] also suggest that users with high domain knowledge attempt to leverage their expertise to extract information from indirect sources for easy tasks, rather than attempting to locate documents which are more concise and to the point. Similarly, Bhavnani [4] has shown that domain experts identify and utilise a variety of authoritative sources, while non-experts rely on general Web search engines. Further, it was demonstrated that those with domain expertise appear to have some high level understanding of the search process for the task at hand, while non-experts do not. Regardless of whether the difference in expertise is search or domain specific, experts typically perform more effective searches than non-experts [4, 30, 31].

Domain specific or technical vocabulary appears to be an important element of domain specific search, with experts using more of these terms per query than non-experts [29, 30], while also selecting query concepts more effectively [32]. The use of this vocabulary can be acquired by non-experts over time [27, 30]. However, Vakkari et al. [27] suggest that there is a minimum threshold of search competency which must be met, for a boolean search system, before an extended vocabulary can be effectively utilised.

Intelligence Analysis.

Intelligence analysts are tasked with fusing multiple, diverse, streams of data in order to rapidly arrive at a conclusion, often without a well defined starting point for the analysis [34]. This section first describes the recommended approach to Open-Source Intelligence (OSINT) searching and analysis, before discussing previous work relating to the behaviour of closed intelligence analysts.

OSINT investigations begin with the communication of an information need, with the first step being to identify potential sources of information which fulfil this need. The sources utilised for criminal inquiries will differ from those used for planning humanitarian aid, or assessing the damage done by a natural disaster. At this point, data collection is strategic, in that the search sources are chosen based on potential return on time invested, with the disclosure of findings focusing on addressing the given task [22]. The steps involved in the tactical collection of data are critical, as the volume of information available online is vast, and the analyst could easily be overwhelmed. Time management and prioritisation is therefore paramount if the OSINT analyst is to be effective. Additionally, subtasks may be automated in order to save large quantities of time [1, 22]. Appel [1] likens the process to playing slot machines at a casino, where, after a point, there is more to be lost than to be gained. Other concerns for OSINT searching may also include the relative freshness of the information, particularly with tasks relating to breaking news or events [3], or concerns over digital footprints which could potentially tip off a subject [22].

Chin et al. [8] analysed the behaviour of national security intelligence analysts by exposing five of them, each with previous intelligence or military experience, to two mock cases which were created by a sixth analyst. Several analysis techniques were discovered, not all of which are compatible. A popular strategy is to map all pieces of evidence to all possible hypotheses in order to determine which is the most likely. However, one analyst disagreed with this approach, as it may cause data to be interpreted in such a way as to fit a particular hypothesis. Another scheme involved interpreting the evidence in terms of high level attributions, such as intent and motivation. Irrespective of the strategy, a key feature of the analysis is that it is iterative; new questions are derived from viewed documents. Analysts described this iteration as never being finished, such that arbitrary deadlines are required to produce results in a timely fashion. As a result of this, analysts developed varying document prioritisation and triage approaches, primarily involving physical copies of documents. Analysts were not found to use any specialised software tools when identifying patterns in the data, instead opting to use *Microsoft Office* products to map out relations. Additionally, varying levels of credibility are assigned to the findings, with facts not necessarily being considered to be "concrete truths" [8].

Pirolli and Card [23] conducted a cognitive task analysis and think aloud study with intelligence analysts in order to identify leverage points for future improvement. The authors characterise analyst behaviour as a kind of expert behaviour, where existing schemas are built from experience and applied to new scenarios. The process was characterised in terms of two primarily loops, an information foraging loop and a sensemaking loop. Foraging involves searching for information, filtering documents and reading the documents with the goal of extracting information. Sensemaking involves the iterative development of a mental model, with the conceptualisation of a schema which fits the evidence. At each stage there is potential for feedback between the various sub-tasks, both inter- and intra-loop. These loops can be used to support both bottom-up (data to theory) or top-down (theory to data) models, with the authors suggesting that either model can be invoked depending on the task at hand. Key leverage points were found in both loops. The foraging loop contains typical precision-recall trade-offs as analysts move from a larger set of documents to a more narrow one. Individual cognitive loads in this loop are associated with scanning, reading and extracting information from documents, as well as those associated with the iterative process in follow-up queries. Leverage points in the sensemaking process include the attention span available for the evidence and hypotheses, as well as the generation of alternative hypotheses while avoiding confirmation bias.

3. METHODOLOGY

While previous work addresses a wide range of search domains, there is a lack of research into the search behaviour of OSINT investigators. However, the potential use cases for open-source information are varied, such that it is difficult to address the entire domain at once. The processes involved in capturing information relating to natural disasters or breaking news is likely quite different to those used for employee background investigations. For this reason, this research focuses on a subset of OSINT use cases, subject centric investigations, where the primary focus is on gathering information about a single individual.

In order to gain insight in to these types of investigations, an in depth exploratory study was conducted with a leading investigation company. This involved input from a total of three participants over two stages: ***i)*** an initial interview with a company executive and ***ii)*** semi-structured interviews [19] with two analysts who carry out the Web searching for the investigations. Intelligence analysis and personal investigations are confidential by nature, as they deal with sensitive information, with a relatively small number of individuals carrying out this work. This means that there is a small pool of subjects to draw upon, in contrast to domains such as general Web search, which is evident in previous work [23]. This exploratory case study sheds light on the relatively unexplored domain of Open-Source Intelligence investigations.

Several research questions were explored in order to facilitate this research:

RQ1 How do analysts who conduct subject centric OSINT investigations perform Web searches? What is their process?

RQ2 Which types of information are they interested in? What are they looking for?

RQ3 Which types of sources do they use to get this information? Do they use any specialised tools to acquire or manage this information?

RQ4 How do the needs of these OSINT investigators compare to those of searchers in other domains?

3.1 Capturing Analyst Behaviour

The executive interview provided an understanding of the overall business practices of the company, which types of cases they investigate, and why. This also provided the context in which to place the analyst's search behaviour. Interview discussion topics included the types of tasks analysts undertake, how they generally approach the process and the nature of the end-product deliverable.

Using this sketch of the problem space and processes, a questionnaire was produced, which was used to guide semi-structured interviews with the analysts, with each analyst being interviewed in isolation. The questionnaire was designed to establish how the analysts currently perform their investigation, as well as which search techniques and tools they utilise. This design attempts to address the axes of domain specific search as discussed in Hanbury [14]. Responses were recorded by the researchers, allowing the analyst to focus on their reply.

3.2 Questionnaire Contents

The questionnaire utilised in the semi-structured analyst interviews was comprised of the following sections.

Personal Background: This section contained questions relating to: demographics; education and employment history; and a self description of their current role.

Example Cases: The analysts were asked to describe three investigations which they had conducted in the past. In order to aid the analyst in spontaneously recalling a particular case, it was suggested that it could have been one of the following: memorable, routine, challenging, unsuccessful, or collaborative. Particular questions related to the high level description of the case, what they were provided with at the beginning, what their approach was, which sources were used, how the information was corroborated, and what challenges were encountered.

Case Generalisations: The analysts were asked to attempt to describe cases in a more abstract way, generalising from particulars to features which are common to most, or all, investigations. They were also asked if there are any particular kinds of cases which do not fit the norm. This section also attempted to associate value ranges to the number of results examined, number of sources used, time spent formulating queries, etc.

Types of Search: Analysts were asked about the various types of search they perform and which sources they use to get this information. Examples of the types of search are: social media; geolocation/map; car/vehicle; news; and multimedia.

Core Functionality: Core functionality includes search operators and other features of a service, or search engine, such as filtering, bookmarking, aggregation, or query expansion/suggestions. The analysts were asked which kinds of functionality they make use of, or which they would find useful, but do not necessarily make use of.

Other Tools and Services: This section pertained to any software or resources the analysts used which were external to the Web, such as software tools for case management, analysis, visualisation, or report generation.

Investigative Process: The interviewees were then asked general questions on the investigative process, such as which aspects of a case take the longest, or are repeated most often. This section was also used to reflect/elaborate on, as well as confirm our understanding of, the processes and techniques which had been described in previous sections.

Features for Improvement: The final section involved questions relating to any functionality which the analysts thought would be useful as part of a hypothetical search system.

4. FINDINGS

This section presents the findings of the exploratory study. First, investigations are placed in context before describing the process which is applied to all subject-centric investigations which are carried out by the analysts. Following this, particular search behaviours are highlighted, with discussion on the generic task which underlies the search process. Specific difficulties associated with these types of investigation are then presented. Finally, the search domain is summarised in terms of the five axes of search in Hanbury and Lupu [14].

Each case begins with a client requesting that a particular investigation be carried out, due to particular internal flags being raised. The client provides a document, or, *briefing sheet*, which describes the purpose of the investigation, as well as known background details regarding the subject of the investigation and any relevant context. This information may vary depending on the particular context, as well as what information is available to the client. However, the briefing sheet typically includes items such as subject name, date of birth, known address, telephone number, etc. Where possible, the entire investigation is carried out using Web resources, however, when this proves insufficient, costly physical surveillance may be employed in order to gather further information. Physical surveillance operatives are provided with updated information and pertinent finding from the analyst's online investigation, allowing them to operate more effectively. While physical surveillance is an important part of the overall process, its impact on the Web search process carried out by the analyst is limited. Therefore, this research only focuses on the work carried out by the analyst, who gathers information using Web resources. An overview of the parties involved in the investigation, as well as their relationships, can be found in Figure 1.

Each investigation is conducted by a single analyst, who spends anywhere between two and eight hours gathering and reporting information, a duration which is consistent with intelligence analysts in Chin et al. (four to eight hours) [8] and NATO's OSINT guidelines (four hours) [22]. This work need not be carried out on the same day[2] and the overall duration may vary in line with the profile of the case at hand. Similarly, cases are typically not revisited once closed, however this is also subject to considerations of proportionality.

[2]In some cases this was out of necessity, as one resource relating to upcoming court cases only makes information available for a week in advance of the court date. For this reason, this resource was checked at regular intervals.

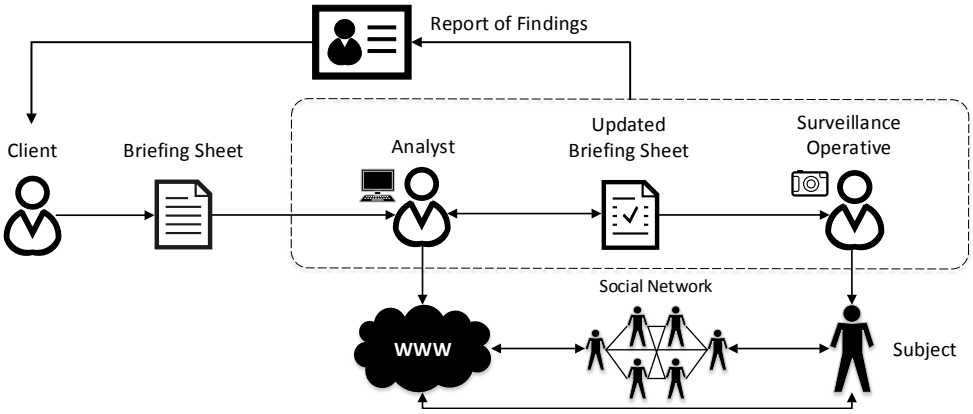

Figure 1: The parties involved in the investigation and their relationships.

Both analysts interviewed were in the 50–60 age bracket, one male, one female, both of whom have 30 years of experience in the police force. During their time with the police, one analyst specialised in criminal intelligence, while the other occupied several roles which leveraged their ability to locate persons of interest. Both analysts had been working in their current role for 3–4 years at the time this research was conducted. While neither analyst claimed to have any extensive IT training, one analyst had previous experience querying national law enforcement databases.

4.1 Process Overview

The approach to the investigation can be split into two main search phases: Background Verification and Open-Ended Web Search. The process involved in these search stages is discussed in turn, before elaborating on the case management and reporting processes.

Background Verification: In this stage, information provided on the briefing sheet is verified, updated and expanded upon. The records provided by the client may be outdated or erroneous; the subject may have moved residence or their personal circumstances, such as employment or marital status, may have changed. It is therefore critical for the investigation, as it moves forward, that this initial information be scrutinised.

The defining characteristic of this stage is that the analysts make use of a small, relatively fixed, set of services in order to verify this information. These sources are authoritative, though not necessarily guaranteed to be accurate, and, as such, are corroborated with each other.

Which sources are used will be dependant on the particular investigation. The most common services relate to residence and property verification as well as car registration checking. Cases with substantial corporate, financial or legal elements will necessitate the consultation of services relating to corporate structures and company information; credit checking; and information relating to previous convictions, respectively.

Open-Ended Web Search: The verified briefing sheet from the Background Verification stage is then used as seed information for the open-ended portion of the investigation. This stage still makes use of a small set of pre-defined resources, but the goal is to learn new things about the subject beyond simple information which could be held in a registry database. This includes utilising resources relating to news, e-commerce and multimedia, but places a particularly strong emphasis on social media and general Web search. Social media platforms, such as `www.facebook.com`, `www.myspace.com`, `www.bebo.com` and `www.twitter.com`, play a large role in the majority of these subject-centric investigations as they allow for individuals to express themselves openly. In particular, there is the potential to gather information directly pertaining to the investigation, either from the subject themselves, or their relations. The leading Web search engine is utilised as a portal to the wider Web, allowing for the gathering of new information from previously unknown sources. Search engines are typically utilised after the social media gathering phase. One analyst noted that, occasionally, competing Web search engines are used, with one competitor being cited as producing more up to date results for Internet news articles.

The open-ended search phase constitutes the bulk of the investigation, with the majority of this time being spent using search engines to find new pieces of information. It should also be noted that the stages are not entirely independent, as some information discovered in the latter stage may inspire new information to be sought, or verified, using services from the Background Verification portion of the investigation.

Case Management and Reporting: The analysts were not found to use any specialised software throughout the investigation, utilising a Web browser as the primary tool, with a popular rich text editor serving as a means of managing and reporting findings. A PDF file is generated from this text editor as part of the final report, which is used in conjunction with a Customer Relationship Management (CRM) system.

Rather than drawing conclusions in the report, the information found throughout the investigation is presented as is, with the client being left to interpret the findings and determine a course of action. That is not to say that the findings are presented in isolation, with one interviewee noting that information is qualified in order to acknowledge that it is fallible. For instance, there may be a chance that the information found belongs to an individual other than the subject of the investigation, as they happen to share names and cities. These qualifications, then, allow for various weights to be assigned to pieces of evidence, based on the degree of cor-

roborating information. The second interviewee noted that the word 'fact' is not appropriate when referring to findings, instead, opting to use the term 'indicator'.

One reason that no specialised case management or reporting facilities are utilised is that, typically, only a small number of relevant documents are found for a given investigation. Hence, there is not a large volume of information to manage at the report level. The report typically consists of highly relevant nuggets of information, with associated source URLs, screenshots and analyst notes (qualifications, context). However, one analyst refrains from including screenshots of Web pages in reports out of concern that copyright will be infringed upon were it to be included. Neither query logs nor visited Web pages are logged to include in the report, with viewed documents/sources only being included in the report when relevant information is found. That is, there is no strict auditing requirement in place, with only information which is relevant to the case being reported, together with its source.

4.2 Analyst Search Behaviour

Verification Tasks: In the Background Verification stage much of the searching is simply rote querying of services with existing information in order to verify its veracity. The same query is repeated to several different services in order to compare, contrast, and corroborate the results. However, in some cases, this core information is what is actually sought after by the client. An example of this would be if an individual accumulated a large amount of debt before absconding, with the task of the analyst being to locate this individual. The context then shifts, such that the focus of the investigation is on discovering this fundamental information, with much greater feedback between the two stages of the investigation.

Both analysts demonstrated a preference for free services over paid services when verifying background information. One interviewee described a micro-transactional process in which small pieces of information, given at no cost by various services, are posed as queries to other services in order to gather different pieces of free information. By using this iterative querying process, these free pieces of information can be expanded upon, and combined, forming a more complete picture while simultaneously avoiding the need for payment. This, of course, inflates the number of queries and places more load on the analyst, with much repetition of query terms across related services.

Open-Ended Searching: Differing strategies were adopted by each analyst in the second phase of the investigation. The first analyst treated social media searching as an intermediate between the Background Verification and Open-Ended Web Search stages, as it potentially enables the corroboration of information such as employment status, marital status, and location. The general Web search process adopted is an experimental one, a 'random, calculated, Web search' in which the subject's name is leveraged as the primary source of query content. Queries are submitted and reformulated at will, with no significant formulation period. Results are then assessed for relevance using background knowledge, previous findings and contextual clues. A relatively small number of documents were viewed by this analyst, numbered in the tens, rather than hundreds per query, indicating a high degree of discrimination of the results. No particular use was made of advanced query formulation techniques, which the interviewee attributed to a personal lack of skill in this area.

The second analyst also placed a similar emphasis on the importance of social media, however it was not described in terms of an intermediate step in the investigation. Queries posed to general search engines were described as simply containing 'the basics', that is, the types of information found on the analyst's briefing sheet. In this case, a larger emphasis is placed on an exhaustive search methodology of the results lists. Initially, approximately 15 pages of results are assessed for each query. If this initial assessment does not yield relevant results, then an iterative process is employed for each query, delving deeper into the results lists in the search for small pieces of relevant information. In this case, hundreds of results, across tens of results pages, may be scrutinised. That is, while the first analyst is more query orientated, the latter analyst is more document orientated. Additionally, while the first analyst made no use of advanced query functionality, the second interviewee expressed knowledge of boolean operators and phrase queries, as well as indicating an interest in learning more about additional operators. This analyst also placed some emphasis on the use of image and reverse image searching. One example given involved the utilisation of a reverse image search engine, which was used in order to scrutinise a social media profile image. From the results, it was discovered that the profile image was a stock photograph, which suggested that the account in question was a fabrication.

In neither case was there a well defined stopping point identified for the gathering of information. In cases where evidence is found, the analysts appear to use their own judgement regarding how much evidence is enough. This judgement is based on what has been looked at already, what has been found and how well the primary task has been addressed. In some cases, evidence may be hard to locate, as it may be found deep down in the search results, hosted on obscure sources. This problem may be exaggerated if little initial information is provided, or if findings are poor during the Background Verification stage of the investigation. However, once an investigative thread is discovered in a new piece of relevant information, new lines of investigation open up, facilitating the discovery of yet more relevant nuggets. That is, once a 'foothold' is obtained, the remainder of the investigation tends to progress more smoothly. For cases where this investigative stagnation is not overcome, the search is carried on until the rough cut off point of eight hours, out of concern for due diligence. One analyst noted that such cases may be revisited at a later date, during periods of lower work loads.

Both analysts voiced concerns regarding the reliability of information, though their general approaches differed. One interviewee accepted most information at face value, unless there was a reason to think that the information was not to be trusted. These reliability determinations were made using personal experience and background knowledge, as well as other pieces of information/context from the case. Particular types of information were emphasised as being particularly unreliable, such as social media relationship statuses and potential terms of endearment. The second interviewee was generally more sceptical of all information until it was corroborated in some way. Again, an example of a potentially unreliable source of information was given in news

Table 1: Comparison of analyst behaviours.

	Search Strategy	Adv. Features	Reliability	Digital Footprint	Reporting
Analyst 1	– Query orientated. – Experimental queries. – Discriminatory when examining documents.	– None used.	– Generally taken at face value unless suspicious. – Wary of particular types of information.	– Uses services requiring user accounts if known to be safe.	– Screenshots, qualifiers, URLs.
Analyst 2	– Document orientated. – Examines many documents per query.	– Some use of boolean operators and phrase queries. – Reverse image searching.	– Generally sceptical. – Distinction between 'facts' and 'reported facts'.	– Categorically avoids services which require user accounts.	– Descriptive notes, qualifiers and URLs. – 'Indicators' as opposed to 'facts'.

media, with a distinction being made between 'facts' and 'reported facts', citing media 'sensationalism' as a potential problem.

A recurring concern was that of the digital footprint left behind by an online investigation, which plays a particularly large role when social media resources are utilised. One analyst categorically avoids using services which require user logins/accounts, while the other only does so when it is known that the subject will be unable to detect their use of the account. This can be a problem for social media platforms such as `www.linkedin.com` and `www.keek.com`, as well as dating sites such as `www.match.com`, as they actively notify the account holder when their profiles has been viewed. The primary reason for avoiding this sort of interaction is that it may compromise the investigation; tipping off the subject allows them to change their behaviour, which may be detrimental to the case. The integrity of the investigation must be preserved, not just operationally, but also ethically. It was stressed, by both the executive and analysts, that only publicly available information was to be used throughout the investigation. That is, it is completely out of the question, to, say, send the subject a friend request on a social media platform for the purposes of bypassing privacy restrictions.

Neither analyst indicated that they use any tools which automate search tasks, nor do they make use of meta-search engines which automatically pose a single query to multiple services. A single aggregation service was cited, pertaining to announcements relating to deaths, marriages, birthdays and other personal or family events. Additionally, some reference was made to services which facilitate the identification of a particular individual's social media profiles, as they may not necessarily directly contain the individual's given name. Table 1 summarises and contrasts the behaviour of both analysts.

4.3 The Underlying Task

Not only is the high level process fixed across the various investigation types, but the underlying task is as well. That is, the analyst seeks to discover small pieces of information relating to the subject, building a network of inter-related informational items. These pieces of information potentially pertain to: relationships, associations/memberships, locations, subject attributes, interests, hobbies and other contextual information. In essence, these small pieces of information are the operating unit of the investigation, allowing the investigation to progress as well as serving as evidential items in the report.

These informational units are used in two ways. The first is to expand the investigation, using pieces of information to generate new query terms, or to give clues as to possible areas of exploration, taking the investigation in a new direction. One example of this may be the discovery that the subject plays, or has previously played, on a darts team. The analyst can then use `darts` or `darts team` as new query terms in conjunction with the subject's name or location. This could lead to the discovery of a locally managed darts league, or event, which lists the subject as a participant. The impact of this information varies depending on the particular investigation, however it may be especially relevant if the investigation pertains to an injury which would prevent the suspect from participating in such physical activities.

The second way in which these pieces of information can be used is to narrow, or focus, the investigation. The same piece of information may be used to provide filtering opportunities in order to better weed out irrelevant results, or to disambiguate person namesakes in the results. A trivial example of this is the discovery that the subject lives in London, which, at least generally, allows for the filtering of persons living in Amsterdam. These pieces of information can also provide context for existing documents, enabling better extraction of information or more efficient relevance determinations at the document level. Additionally, cross-referencing information allows the analyst to better understand, and better utilise, previous findings. One example given was the use of a social media post about dog walking, which contained an image of the subject's partner. By scrutinising this image, and cross-referencing it with other resources, the analyst was able to determine an approximate postcode of where the image was taken. This postcode eventually led to the discovery of the location of the subject, successfully ending the investigation.

4.4 Difficulties

Person disambiguation was a concern for both analysts, as it is not always straightforward to identify a namesake given the available information, or context. In most cases, a claim that a particular piece of information, or social profile, is related to the subject, requires qualification due to the uncertainty involved. Various levels of confidence can be placed in this assertion, depending on the degree to which other ev-

idence corroborates it. For instance, a social media profile which uses the subject's name may contain a profile image of a person in front of a place of residence. This profile image can then be corroborated with geographical images found on a popular publicly available satellite image provider, in order to provide support that the individual in the image is the subject, or a personal relation. This knowledge can then be used to corroborate or contextualise further pieces of information. Similarly, if the same images or other contextual clues are present across various social media platforms, it is likely that the accounts belong to the same individual. The use of these contextual clues, in concert with human judgement and experience, appears to play a key role in the process of person disambiguation, and, indeed, in determining the reliability of any piece of information.

A related problem is that of name variants, be they colloquial derivatives, erroneous spellings, nicknames, pet names or even simply the favouring of a middle name over a first name. Both analysts noted that this is not entirely uncommon, with one example being given of a court document which had incorrectly transcribed the subject's name. In cases where name variants are used, information may simply be missed. The alternative is an increased load on the analyst as many more queries are generated in order to compensate for poor findings. This problem may be compounded when Web services require payment per query, though, as previously noted, these services are not favoured by the interviewees.

In addition to concerns relating to digital footprints, social media platforms also present challenges for information acquisition, as users are typically able to control to which degree their information is accessible. That is, when the service allows it, they may opt to make their information visible only to friends, or friends of friends, rather than the general public. In these cases, rather than obtaining information from the subject's profile directly, it is obtained laterally, via social media relations, such as friends or family, who may not operate with the same level of discretion. One analyst suggested that this lateral acquisition occurs more often than not, emphasising the importance of mapping out the social relations of the subject.

When asked how a hypothetical system could improve the overall process, suggestions included aid for the above tasks of person disambiguation, name variant querying and lateral social media information gathering. However, the overall onus for improvement was on improving the general Web search process, as conducted via the leading Web search engine. The most time consuming portions of the investigation involve reviewing results, identifying relevant documents, reading documents and extracting information from them, i.e. the information foraging loop in Pirolli and Card [23]. As such, streamlining this Open-Ended Web Search phase has the most potential to reduce the overall cost of the investigation. The precision of the results seems to be the primary factor, with the extraction of key pieces of information being the priority. Increasing precision would not only reduce the time spent finding and extracting relevant information but would also reduce the time spent on investigative stagnation, where new leads are not presenting themselves. One analyst succinctly described the ideal system as one which facilitates 'minimum time in for maximum information out', i.e. one which maximises utility.

4.5 Domain Specific Search Summary

The following is a summary of the findings of this research in terms of the five axes of domain specific search found in Hanbury and Lupu [14].

Users: Both analysts interviewed were ex-police officers with 30 years of service each – with specialisations in criminal intelligence and finding persons of interest. Neither analyst possesses advanced information technology skills, instead leveraging their extensive experience and ability to utilise contextual clues and links between information.

Tasks: In subject-centric online investigations, the task is to associate different pieces of information with the subject, with the process being divided into two main stages: Background Verification and Open-Ended Web Search. Verification sub-tasks involve assessing the accuracy of fundamental information, such as: name, address, date of birth, telephone number, marital status, employment status, etc. In the Open-Ended Web Search phase, analysts seek new pieces information, such as interpersonal relations, hobbies, interests, associations, and activities. This information can then be used to address the primary goal of the case. The focus of the investigation may shift, depending on the reason for which it is being conducted. For example, corporate embezzlement investigations will have a different focus than those pertaining to employee background vetting or fraud.

Subject Area: A wide range of open data sources are utilised, such as: Web search engines; social media platforms; database resellers, such as the electoral roll or vehicle registration information; property information; e-commerce and credit checking services; public record corporate information; news and legal archives; social announcements; and map/geolocation services. However, in practice, any OSINT resource may be utilised, with the scope being limited only by the task at hand.

Media: No particular limitations are placed on the types of media used in the investigation. Examples were given which involved text, image, video, and audio files, suggesting that analysts make use of a wide variety of media types.

Tools: The primary tool used by the analysts was a Web browser, with a text editor and a CRM system being utilised for the purpose of reporting findings and case management. No particular browser add-ons or extensions appear to be used, with the use of advanced query functionality being minimal. The Web browser is used as mechanism for accessing a variety of sources, as discussed in the Subject Area section above.

Several potential leverage points were identified which may be addressed at the software level: person disambiguation; name variants; repeat queries; lateral information gathering of social media platforms; results precision; and information extraction.

5. DISCUSSION

While these subject centric investigations possess properties of exhaustive searches, the huge volume of documents available on the Web make finding all information relating to a subject impractical. This means that, in contrast to patent searching [17] and E-Discovery [2], high recall is not as critical as high precision, since the goal is to find a small number of highly relevant documents in a large number of

results. Therefore it is appropriate to characterise this domain in terms of exploratory searching, rather than exhaustive searching, using the definitions in Hogan et al [15]. That is, the analyst is attempting to find information about the subject but does not necessarily know in advance what may be useful, nor is there a requirement that all pieces of information be found. While there are no legal repercussions for failing to find documents, as there may be in the patent searching domain [17], there may be other costs. For example, the client may suffer a loss of reputation in a scandal relating to an employment decision, or be forced to pay large insurance settlements. Some degree of personal investment in the search is also apparent, as old cases with little information are occasionally revisited, despite the lack of necessity to do so.

As with E-Discovery searching [2] and intelligence analysis in Chin et al. [8], the Open-Ended Web Search phase appears to be non-linear in nature. Findings discovered may prompt new lines of investigation, with new queries, which divert the analysts' attention from the original search.

It is clear that the interviewees demonstrate domain expertise, rather than general search expertise. The analysts do not appear to use advanced search features more often than the average Web user, though they do view more pages of results [20, 25]. However, as with domain experts in Bhavnani [4], the interviewees demonstrated the existence of a high level structure in their overall investigative process, while also making use of pre-defined, authoritative, domain specific resources. Similarly, the general preference for avoiding commercial sites is consistent with domain expert behaviour in White et al [30].

In contrast to the intelligence analysts in Pirolli and Card [23], the report produced by the interviewees does not contain hypotheses or conclusions. However, it is possible that hypotheses are represented internally to some degree in order to facilitate the cross-referencing of pieces of information and to provide a contextual framework for the investigation. The lack of explicit hypotheses suggests that a bottom-up approach is taken, although in some circumstances, the case provided by the client may warrant a top-down approach. In this latter case, the client may wish to identify a particular type of behaviour, such as embezzlement or fraud, such that an explicit hypothesis is given from the outset.

Parallels can be drawn to intelligence analyst behaviour in Chin et al. [8]. Firstly, the interviewees did not make any use of particularly specialised tools, instead opting to use only a Web browser and text editor in order to search, manage and create reports for the case. Secondly, the interviewees' approach to the reliability of information is consistent with two of the analysts in Chin et al [8]: the first being generally accepting unless there is a reason to think otherwise; and the second being generally more sceptical of information until it is corroborated. In all cases, there is a similar reluctance to attribute the concept of a "concrete truth" [8] to the findings, instead opting to qualify fallible findings.

While all interviewees possessed a strong desire to minimise their digital footprint during the investigation, the primary means of effecting this was to avoid the use of services which require accounts, or those which could unintentionally communicate with the subject. Neither IP address obfuscation or cookie management presented themselves as issues to the interviewees. The NATO OSINT [22] handbook gives suggestions for minimising the digital footprint of the investigation by making use of proxy services and engaging in cookie management. The reasoning is that a high volume of searches relating to a particular individual/group, such as a terrorist or dictator, may be noticed and have an impact on the investigation. However, it may be unreasonable to expect that the subjects of investigations carried out by the interviewees would have the resources necessary to acquire this information via traffic analysis, meaning that they may be justified in their current behaviour.

The freshness of the information was not a primary concern for the interviewees, with information which is weeks, or even months, old being useful. This is in contrast to a large number of use cases for Open-Source Intelligence gathering which rely on the immediacy of the information, such as real time event monitoring, news summarisation or crisis management. That is not to say that the interviewees were completely unconcerned with the recency of documents, as a competing search engine was cited as being potentially more up to date when searching for Internet news articles.

6. SUMMARY

This work presented the findings from a study of a particular sub-domain of OSINT investigations, relating to subject centric investigations. Several research questions were considered in order to ground this research. In relation to **RQ1**, it was found that the search process is comprised of two main stages: Background Verification and Open-Ended Web Search. This process overview, in conjunction with a deeper analysis of how they conduct each stage addressed the search questions in **RQ2**. Analysts search for small pieces of information which relate to the subject, ranging from social relations and hobbies to employment history and residence, in order to conduct a wide variety of investigation types. While no specialised software tools were found to be used to carry out searching and reporting, or facilitate case management, a variety of Web service archetypes have been identified as being used frequently, addressing **RQ3**. Finally, **RQ4** is explored by contrasting this domain to previous work. This showed that the analysts exhibit behaviour consistent with experts in other domains, and approach evidential reliability in a similar manner to national security intelligence analysts.

Several difficulties were identified with this type of investigation, involving name variants, person disambiguation, digital footprints, social media information access, and large volumes of search engine results. The findings from this research are significant in so far as they provide insight in to the overall behaviour and search needs of these analysts. Further research could investigate practical solutions for tackling these difficulties, in order to better facilitate these types of investigations. Additional work could expand on the OSINT search domain in order to provide insight into other use cases, such as those pertaining to natural disaster response and humanitarian aid, or investigations pertaining to groups of people, rather than individuals. This may lead to the discovery that certain elements of the investigation are common across the entire OSINT domain, or that techniques utilised in domains with similar characteristics could be adapted for OSINT purposes.

7. REFERENCES

[1] E. J. Appel. *Internet Searches for Vetting, Investigations, and Open-source Intelligence*. 2011.

[2] S. Attfield and A. Blandford. Improving the cost structure of sensemaking tasks: Analysing user concepts to inform information system design. In *Proc. of INTERACT '09*, page 532–545, 2009.

[3] C. Best. Open source intelligence. *Mining Massive Data Sets for Security: Advances in Data Mining, Search, Social Networks and Text Mining, and Their Applications to Security*, 19:331–344, 2008.

[4] S. K. Bhavnani. Domain-specific search strategies for the effective retrieval of healthcare and shopping information. In *Proc. CHI'02 Extended Abstracts*, page 610–611, 2002.

[5] A. Broughton, B. Foley, S. Ledermaier, and A. Cox. The use of social media in the recruitment process. *acas*, 2013.

[6] V. Brown and E. Vaughn. The writing on the (facebook) wall: The use of social networking sites in hiring decisions. *Journal of Business and Psychology*, 26(2):219–225, 2011.

[7] S. Catanese, P. Meo, E. Ferrara, G. Fiumara, and A. Provetti. Extraction and analysis of facebook friendship relations. In A. Abraham, editor, *Computational Social Networks*, pages 291–324. 2012.

[8] G. Chin,Jr., O. A. Kuchar, and K. E. Wolf. Exploring the analytical processes of intelligence analysts. In *Proc. of CHI '09*, page 11–20, 2009.

[9] Department of Homeland Security. DHS terrorist use of social networking facebook case study | public intelligence, 2010.

[10] E. Finch. What a tangled web we weave: Identity theft and the internet. *Dot. cons: Crime, Deviance, and Identity on the Internet. Cullompton, England: Willan*, 2003.

[11] L. Freund and E. G. Toms. Enterprise search behaviour of software engineers. In *Proc. of SIGIR '06*, page 645–646, 2006.

[12] J. Grasz. Thirty-seven percent of companies use social networks to research potential job candidates, according to new CareerBuilder survey - CareerBuilder, 2012.

[13] R. Guha and A. Garg. Disambiguating people in search. In *Proc. of the 13th World Wide Web Conference,*, 2004.

[14] A. Hanbury and M. Lupu. Toward a model of domain-specific search. In *Proc. of OAIR '13*, page 33–36, 2013.

[15] C. Hogan, D. Brassil, and M. Marcus. Human aided computer assessment for exhaustive search. In *SMC '09*, pages 108–112, Oct. 2009.

[16] B. R. Holland. *Enabling Open Source Intelligence (OSINT) in private social networks*. PhD thesis, Iowa State University, 2012.

[17] H. Joho, L. A. Azzopardi, and W. Vanderbauwhede. A survey of patent users: An analysis of tasks, behavior, search functionality and system requirements. In *Proc. of IIix '10*, page 13–24, 2010.

[18] C. Liu, J. Liu, M. Cole, N. J. Belkin, and X. Zhang. Task difficulty and domain knowledge effects on information search behaviors. *Proc. of the American Society for Information Science and Technology*, 49(1):1–10, 2012.

[19] K. Louise Barriball and A. While. Collecting data using a semi-structured interview: a discussion paper. *Journal of advanced nursing*, 19(2):328–335, 1994.

[20] K. Markey. Twenty-five years of end-user searching, part 1: Research findings. *Journal of the American Society for Information Science and Technology*, 58(8):1071–1081, 2007.

[21] A. Markham and E. Buchanan. Ethical decision-making and internet research: Recommendations from the aoir ethics working committee (version 2.0). Technical report, 2012.

[22] NATO. NATO open source intelligence handbook, 2001.

[23] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proc. of International Conference on Intelligence Analysis*, volume 5, page 2–4, 2005.

[24] A. Spink, B. J. Jansen, and J. Pedersen. Searching for people on web search engines. *Journal of Documentation*, 60(3):266–278, 2004.

[25] A. Spink, D. Wolfram, M. B. Jansen, and T. Saracevic. Searching the web: The public and their queries. *Journal of the American society for information science and technology*, 52(3):226–234, 2001.

[26] L. Šubelj, Š. Furlan, and M. Bajec. An expert system for detecting automobile insurance fraud using social network analysis. *Expert Syst. Appl.*, 38(1):1039–1052, Jan. 2011.

[27] P. Vakkari, M. Pennanen, and S. Serola. Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information Processing & Management*, 39(3):445 – 463, 2003.

[28] W. Weerkamp, R. Berendsen, B. Kovachev, E. Meij, K. Balog, and M. de Rijke. People searching for people: Analysis of a people search engine log. In *Proc. of SIGIR '11*, page 45–54, 2011.

[29] R. W. White, S. Dumais, and J. Teevan. How medical expertise influences web search interaction. In *Proc. of SIGIR '08*, page 791–792, 2008.

[30] R. W. White, S. T. Dumais, and J. Teevan. Characterizing the influence of domain expertise on web search behavior. In *Proc. of WSDM '09*, page 132–141, 2009.

[31] R. W. White and D. Morris. Investigating the querying and browsing behavior of advanced search engine users. In *Proc. of SIGIR '07*, page 255–262, 2007.

[32] B. M. Wildemuth. The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology*, 55(3):246–258, 2004.

[33] R. Wolff, J. McDevitt, and J. Stark. Using social media to prevent gang violence and engage youth, 2011.

[34] H. Wu, M. Mampaey, N. Tatti, J. Vreeken, M. S. Hossain, and N. Ramakrishnan. Where do i start?: Algorithmic strategies to guide intelligence analysts. In *Proc. of ISI-KDD '12*, page 3:1–3:8, 2012.

[35] G. Zipf. Human behavior and the principle of least effort. *Addison Wealey*, 1949.