Foraging Online Social Networks

Gijs Koot∗, Mirjam A.A. Huis in ’t Veld∗, Joost HendrickSEN†, Rianne Kaptein∗,
Arnout de Vries∗ and Egon L. van den Broek†
∗TNO, The Netherlands
Email: {gijs.koot,mirjam.huistveld,rianne.kaptein,arnout.devries}@tno.nl
†TUNIX Digital Security, The Netherlands
Email: joost@hendricksen.net
‡Utrecht University, The Netherlands
Email: vandenbroek@acm.org

Abstract—A concise and practical introduction is given on Online Social Networks (OSN) and their application in law enforcement, including a brief survey of related work. Subsequently, a tool is introduced that can be used to search OSN in order to generate user profiles. Both its architecture and processing pipeline are described. This tool is meant as a flexible framework that supports manual foraging (and not replaces it). As such, we aim to bridge science’s state-of-the-art and current security officer’s practice. This article ends with a brief discussion on privacy and ethical issues and future work.

I. INTRODUCTION

Over the last decade the Internet has become an important communication platform in the Western world. Due to the emergence of mobile broadband connections, smartphones and tablets people spend more time online. Encouraged by Online Social Networks (OSN), the ease of sharing information on the Internet has become an extension of people’s lives. Anything can be shared in communities, weblogs, social networks, and forums, 24/7. 91% of the adults use OSN regular and spent > 20% of their time online on OSN [8]. While the majority of users use communities to share experiences, people with wrong intentions use those platforms to exploit criminal or illegal activities [9].

With the growth of shared information, digital criminal investigation becomes an important part to the field of criminal investigation. OSN are obviously a valuable source of information. Law enforcement agencies are interested in utilizing this information to contribute in criminal prosecutions [9]. As part of exploring the possibilities of Open Source INTelligence (OSINT), this research focuses on investigative profiling of an individual. For years, this process has been part of classical forensic research. In this research, we explore ways to apply this technique to open information sources, specifically OSN.

At this time, a security officer (e.g., a police officer) (still) often manually searches for personal data on open information sources on the Internet. The security officer manually forage specialized public search engines and information sources to supplement a user profile. However, this makes the process of digital profiling time consuming and prone to errors. Therefore, law enforcement agencies are looking for an instrument to assist them in their job. We want to support the security officer, without taking away the human component and its analysis strengths. We will present a model that partly automates the process of online profiling and utilizes the human input. As such it targets the first of three phases of social media analytics: capturing [8]. The second phase is “understanding”, which is left to the human security officer. The third phase is presenting, for which we provide a basic functionality.

In the next section, we discuss related work. Section III discusses the characteristics of OSN. In Section IV, we will introduce a novel tool supporting the generation of people’s profiles, utilizing data from various sources. Last, in Section V, we close this article with a discussion including privacy and ethical issues and future work.

II. RELATED WORK

Several models have been proposed to gather information from open data sources for law enforcement, including OSN. Here, we will give four typical examples of such models.

Pouchard, Dobson, and Trien [13] proposed two models that use two different sources: the Internet in general and the DNI Open Source Center (i.e., an US government intelligence service that aggregates open data sources). The first model provides functionality to collect data from open sources, saving it in a local database, and focuses on the visualization of data. The second model stores and processes open source data and is able to extract metadata (e.g., topic, city, and geographical coordinates). It implements the SeRQL query language with the RDF repository to compose search queries. Search results are analyzed using a named entity recognizer.

A validated named entity recognizer for law enforcement purposes is proposed by Crawley and Wagner [6]. It is founded on rule based entity guessing, regular expressions, and machine learning. Their entity recognizer aims to recognize locations, persons, telephone, credit card numbers, simple dates, email, URLs, and IP addresses. The algorithm was trained on both English and German corpora and realized high scores on both recall and precision.

Baldini, Neri and Pettoni [2] described an extensive model to perform multi-language data mining on unstructured text. Their approach is based on Natural Language Processing (NLP) and has the ability to perform multi language lexical analysis on large sets of documents. Their model is able to extract functional relationships within a document that are indexed on a conceptual level and can be searched or browsed by term and can be visualized in a tree view. They have created a search engine based on the same functional relationships. The free text search query a security officer enters is analyzed, the system responds with the conceptual expansion of the query.
TABLE I. LIST OF ATTRIBUTES PRESENT IN OSN, ADOPTED FROM [18]

<table>
<thead>
<tr>
<th>Attribute</th>
<th>User ID</th>
<th>Networks</th>
<th>Name</th>
<th>Birthday</th>
<th>Political Views</th>
<th>Books</th>
<th>Music</th>
<th>TV Shows</th>
<th>Chats</th>
<th>Profile Picture</th>
<th>Religious Beliefs</th>
<th>Education History</th>
<th>Website</th>
<th>List of Friends</th>
<th>Movies</th>
<th>Events</th>
<th>Photos/Videos</th>
<th>Friends Requests</th>
<th>Work History</th>
<th>Notes</th>
<th>Hometown</th>
<th>Current Location</th>
<th>Tags of Photos/Videos</th>
<th>Activities</th>
<th>Messages in Inbox</th>
<th>Posts from News Feeds</th>
</tr>
</thead>
</table>

Based on the concepts extracted from the document collection, the data analyst selects relevant concepts after which a list of resulting documents is displayed to the security officer. For our case, parts of their approach can be reused to improve precision in the process of searching for personal information on regular web search engines though it is not specifically designed for extraction of personal profile attributes.

Colombini and Colella [5] approach the process of digital profiling by mapping it to the process of traditional profiling and, consequently, bridge the gap between traditional profiling and digital profiling. They created a model to assess whether or not different mass media devices (e.g., mobile phones, laptops, and desktop computers) belong to the same person. They propose a method based on set theory, where designated features are extracted from different devices. Then, these features are compared to a sample profile (i.e., set of features) to determine whether or not they are similar. However, their approach is rather specialized to specific devices and operating systems and is not properly validated with real cases and, therefore, not applicable to our case.

For reasons of brevity, these four models compose far from an exhaustive survey. Several other interesting initiatives have been introduced. These include the Highway to Security: Interoperability for Situation Awareness and Crisis Management (HITS/ISAC) model [1], work on semantic linking and contextualization for social forensic text analysis [14], the ORCA T I and ORCA T II systems that supports OSINT via supporting tools for selecting, collecting and storing open source data [13], and tools that provide visualizations of networks (e.g., [12]). However, neither off-line data mining nor visualization is our topic of research.

Taken together, the models presented in literature propose various techniques for extracting information from a set of existing documents. They employ data mining, data extraction, and analysis on the set of documents. However, the models discussed here do not perform an ad-hoc search on online sources, which is inevitable when searching highly dynamic sources like web sites of OSN.

### III. ONLINE SOCIAL NETWORKS (OSN)

OSN are web services that allow their users to generate a profile in a system (and determine its accessibility), link with other users and share attributes, and browse through the OSN the system maintains [3]. The tool described in this article supports the process of searching OSN for specific individuals; it supports finding, gathering, aggregating, and, ultimately, analyzing and presenting personal information (cf. [8]). As such it can serve as the back-end for various tools and functions, such as profiling [4], [5], [7], [10], [13], [15], [18].

OSN provide various types of information scents and attributes. Table I provides the list of types that we will use. Given Table I, one can wonder: What is the accessibility of all these types of information over the different OSN? This question was answered by Chen, Kaafar, Friedman, and Boreli [4] for Facebook, Twitter, LinkedIn, MySpace, and YouTube. Not surprisingly, name and username were available in the vast majority of cases for all five OSN. For the other types of information there was a significant amount of variation among the OSN. Information such as status, books, music, movies, TV shows, zodiac sign, “interested in”, religion and birthday were only available for Facebook, MySpace, and/or YouTube for a minority of the users. Nevertheless, even a sample from Table I can unveil a person’s crucial information (e.g., his social network) [15].

Searches of OSN can be augmented in various ways. It can be conceived as an iterative process, in which queries are adapted based on the data already gathered, the information extracted from it, and the (partial) profile generated. As such, it can also be considered as a closed-loop system in which security officer and system interact until the security officer determines that the final state is reached. This can be either a completed profile or the observation that further iterations do not improve the process.

The iterative process has the following three phases: i) OSN search, which results in a profile overview; ii) profile selection and full profile overviews; and iii) selection of relevant attributes, which results in an aggregated profile. Subsequently, the aggregated data is analyzed and, if needed, the security officer can go back to the initial step. The security officer is essential in this process. Not only does the security officer decide when to continue with profiling and when to stop, the security officer also generates the actual profile and applies the principle of cooperative annotation [17] to structure the data, link data to each other, et cetera, where the system failed to do so automatically. Next, we will introduce a tool that supports the first of the three phases.

### IV. PROFILING TOOL

This section reports on our endeavor to develop an online profiling tool to search multiple OSN and aggregate the retrieved data into a profile overview. The current tool is a first version demonstrator that does not yet include any intelligence in the aggregation process of the data. So, the tool does not merge the data and does not check for either inconsistencies or duplicates. However, in practice, the collection of the data over multiple OSN is already a challenge, which this tool can already relieve. Moreover, thanks to its modular, standardized implementation, it can be easily extended to be able to handle additional OSN.

#### A. Architecture

The architecture was implemented in a web application, which enabled cross-platform compatibility and multi-user support. See Figures 1 (top) and (bottom) for a schematic
Fig. 1. Profiler’s general architecture (top) and web crawler specific architecture (bottom).

overview of respectively the profiler’s general architecture and the web crawler’s architecture. The foundation for the implementation was the Django web framework. This framework provided features for rapid prototype development and scalability. The data model was defined in a Django project and deployed on a SQLite database. Since Django is written in Python, we extended it with libraries for authorization on OSN (OAuth 2.0), HTML-parsing (BeautifulSoup), URL handling (urllib2), and many more. To facilitate high quality usability, we have adopted AJAX (i.e., HTML, CSS and jQuery) for the tool’s front-end.

Our web framework uses a Model View Controller (MVC) architecture pattern. This pattern separates distinct aspects of an application’s implementation. The Model consists of: i) the data model, ii) operations regarding the model, and iii) validation rules. The View describes an output representation of the data, such as HTML and JSON. The Controller translates security officer’s input to the Model or View. The initial data set is founded on the features presented in Table I. However, in practice not all attributes appeared usable. Hence, depending on the OSN crawled, the set of attributes was dynamically adapted. The search interface, and the search results interface are shown in Figure 2 (top) and (bottom) respectively.

B. Processing pipeline

In this architecture the search, extraction, loading and transformation for each OSN is realised in the target specific crawler. The target specific crawler will perform a set of OSN specific search strategies to find relevant data on the designated target. Its processing pipeline contains the following modules: search, pre-filter, profile crawler and data extractor, and relevance calculator. When the target specific crawlers are initiated they will authenticate with the OSN API, since this authentication (OAuth 2.0) will eventually time out the system will ask each security officer to log in and grant the profiler application permission to access the OSN account.

The search module uses the initial data to perform a search query on the OSN API, depending on the target OSN it applies different search strategies. Strategies to improve recall include user name parsing and specific web search engine searches; those strategies are implemented in the search module.Parsed user names are appended to the result list. For web search engines, the search results in the HTML source are placed in class identified DIV elements. After performing a search query on a web search engine, the result page is parsed to extract the URLs of the search result. To extract user IDs from each search result the page is parsed and all hyperlinks are extracted, further examination of the URL classify whether or not a URL is linking to a user profile. If so, the user ID is extracted and appended to the result list, which is sent to the pre-filter module. The list of user IDs from the search process contains duplicate user names and user IDs. The pre-filter will create a distinct list of unique IDs that is passed on to the profile crawler.

Depending on the target the parser will either use the API or parse the HTML content of an URL to extract user profile data from a page or a profile. The HTML parsing scheme is hardcoded in the application. Each found attribute type on the
target would be translated to our general types by an array of dictionaries. Finally, the user profiles are sent to the relevance calculator. Because different strategies are used to find user profiles the results might not all be relevant. To calculate the relevance of a profile we used the following approach: the presence of the terms from the search query in the resulting profiles are calculated and normalized by dividing it by the total number of terms in the profile. The relevance ratio (i.e., 0 ... 1) is used to order all results.

V. DISCUSSION

OSN have been discussed in the context of public security. Existing models apply various techniques for extracting information from a set of existing documents. They employ data mining, data extraction, and analysis on the set of documents [2], [5], [6], [1], [12], [13], [14]. The profiling tool introduced here deviates from this practice in that it can perform ad-hoc searches on online sources, which is inevitable when searching highly dynamic sources like web sites of OSN. Moreover, its modular, standardized implementation allows an straightforward extension to be able to handle additional OSN.

Last year, the penetration of the PRISM electronic surveillance program of the U.S. National Security Agency (NSA) was unveiled in its full extent [7]. On the one hand, this reveals the importance of OSN, and more generally online “open” sources, which calls for OSINT. On the other hand, this illustrates a new threat, a threat to our privacy [3], [7]. Further, it should be noted that it is rather naive to restrict data mining efforts to open sources, where closed sources are at least as important. The tool presented here solely uses true open access resources and, as such, remains within the legal boundaries. Such considerations are crucial to maintain consumer’s trust in OSN [3], [10].

Although foraging OSN is often discussed, its importance is generally acknowledged. Here, we presented a tool that can aid this (traditionally manual) process. The Needle Custom Search engine1 [11] is envisioned to be integrated with the current tool, to enable analyses exploiting semantic annotations (e.g., temporal annotations, named entities, and domain context) [11], [14]. Other key techniques that should be integrated in the current tool, include opinion mining, sentiment analysis, topic modeling, trend analysis, and visual analytics [8]. These techniques could aid the second phase of social media analytics: understanding [8]; so far, left to the security officer.

In sum, we conclude with acknowledging that this article does not reveal a huge scientific progress. And this is not what this work was meant to be. It provides a bridge between science, computer engineering, and security officer’s current (manual) practice. A tool is presented to aid exactly this, in an intuitive, easy accessible manner. As such, this tool has been valued by several security officers. Therefore, we will continue and integrate it with other existing tools (e.g., [1], [11], [12], [13]), open source libraries (e.g., Stanford’s Natural Language Processing (NLP) software)2, and the latest academic advancements (e.g., complexity and content analysis [14], [16]).

REFERENCES


1Online Needle Custom Search (NCS) demonstrator: http://www.mediaminer.nl/topic_3_context/ [Last accessed on July 22, 2014]