UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Tècnica Superior d'Enginyeria
de Telecomunicació de Barcelona

Open-Source Intelligence Profiling, Analysis and Countermeasures

Master Thesis
submitted to the Faculty of the
Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona
Universitat Politècnica de Catalunya
by

Christodoulos Tziampazis

In partial fulfillment
of the requirements for the master in
Cybersecurity

Advisor: Josep Pegueroles
Barcelona, June 2021

Contents

List of Figures

List of Tables

Revision history and approval record

Revision	Date	Purpose
0	12/03/2021	Document creation
1	15/06/2021	Document revision

DOCUMENT DISTRIBUTION LIST

Name	e-mail
Christodoulos Tziampazis	christodoulos.tziampazis@estudiantat.upc.edu
Josep Pegueroles	josep.pegueroles@upc.edu

Written by:		Reviewed and approved by:	
Date	14/06/2021	Date	14/06/2021
Name	Christodoulos Tziampazis	Name	Josep Pegueroles
Position	Project Author	Position	Project Supervisor

Abstract

In the past decades the size of the internet and digital world has grown enormously. Today businesses are operating entirely online and thus bringing with them all of their costumers' personal details. Digital profiling is a practice that immerse from different necessities such a criminal profiling for proactively identifying malicious actions but also for users profiling for marketing campaigns. In this research we tackling the different digital profiling techniques and methodologies in order to analyse and propose potential counter measures which may affect the success rate of identifying and profiling an individual. Based on that, we are analysing OSINT tools, past research papers and existing methodologies in an attempt to aggregate the techniques used and comment on possible ways that might make the profiling phase a step more difficult. Finally, we also propose potential improvements that can be applied in the future that will assist the scope of this research

1 Introduction

This master thesis focuses on the digital profiling of individuals and reports on counter measures against methodologies and techniques used to collect, aggregate and analyse the already available public data. This has been achieved by reviewing existing open-source intelligence (OSINT) tools and methodologies presented in literature and further explained in Methodology Section 4. Based on these 2 approaches we concluded the data sources available for collecting OSINT alongside with the methodologies used as well as the reasoning behind each step taken. The scope of the research is to bring together the different methods used for profiling and reflect on their level of severity and exposure.

To begin with, over the past decade the Internet became an inseparable part of our everyday lives. The use of internet does not only represent a communication medium but, nowadays, a way of living. The traffic that is created from the vast amount of people using their computers, smartphones, tablets and other internet connected devices gave life to the largest businesses and organisation that we know of today like Facebook, Instagram, Uber, Twitter and so on. The topics that people convey in such platforms ranges from leisure to work and politics but also to more sensitive matters such as finance, real estate, small business, legal and illegal exchanges of goods and many more.

The availability of such personal information raises privacy concerns since it can be collected in many cases in real-time. This information gave life to the digital profiling era. As we are about to see in this research, there exist different techniques for not only collecting but also analysing the user's behaviour. Any kind of publicly available information including pictures, text, videos, location, interests, friends, education, relationships and so on, can be now used against an individual. A well-known example is the Facebook-Cambridge Analytica scandal were Cambridge Analytica collected and used more than 80 million Facebook users' data to perform an audience segmentation and psychographic analyse in an attempt to gain an in-depth understanding of the targeted audience [1].

1.1 OSINT Investigation

Let us now introduce the term OSINT profiling. In general, OSINT includes methodologies of various kinds of techniques for gathering and analysing data that is publicly available. Specifically, publicly available data refers to sources that contain free, legally disclosed, accessible by everyone and not classified information. Such data can be found in the different social media networks (SMN) or they can be directly queried by search engines such as Google. The data collected could be then used to construct a profile for a specific target including behavioural features and characteristics.

Usually when referring to OSINT profiling, we could either target broader audience or to just one individual. Both require a large number of search cycles mainly because the terms used in a single search might not return adequate results for our target and thus some follow up sequential searches might be required to further examine the results found in the previous iterations. Furthermore, these follow up searches most of the times help us uncover even more hidden relation between our target and its connections and interests.

In general, the OSINT investigation can be divided into 3 broad categories [2]. Each

category provides a clear direction on the objectives and potential targets of the OSINT investigation.

1.1.1 Organized Crime & Military

Social media networks, among other means of communications, have been proven to be an integral part of intelligence gathering. For example, there have been numerous cases where individuals as also criminal groups have publicly indicated their intensions about their criminal acts, relations and ideologies. In other cases, targeted individuals may not publish incriminating data but still they could be of great significance by providing various types of information such as relationships, locations, other individuals and so on. However, the vast amount of available information makes the profiling and investigation process even more difficult for law enforcement since there is endless information to be gathered and analysed.

1.1.2 Cybersecurity

In the cybersecurity domain, OSINT is an essential tool since every aspect of cyber practises include a form of OSINT gathering procedures. For example, in the case of red teams and penetration testing, the team's first steps are to discover and collect intelligence about the targeted company, its employees, the infrastructure and so on. Moreover, other assessments include practices such as threat intelligence, threat hunting and social engineering and all together could be comprised by one or more of the following methodologies: Web search engines, Darknet search, reverse imaging, public documentation, personal data, emails and domains information, exploits archives and so on.

1.1.3 Business & Social Intelligence

As it is discussed in Section 2 State of the Art, various tools have been developed in order to analyse behaviours and emotions of people only by the data posted by them in SMN and forums. Researchers and criminologist can utilize these tools and methods to construct a complete profile about people and possibly predict their opinions and believes. This kind of information could be then used for various occasions such election and marketing campaigns or for the formation of the public opinion about different topics.

1.2 Work plan & Milestones

This research we conducted in a total period of 3 months. The waterfall representation of the flow of this research is depicted in Figure 1. As mentioned before, this research consists of 2 approaches namely the review of OSINT tools and the review of the digital profiling literature. Each of the approaches required approximately a period of 1 month to review, analyse, classify and conclude the collected data. Our main objective is to analyse and conclude counter measures against digital profiling, however, this field is very broad and the methodologies used are complex and diverse. Due to the massive amount of available information, our initial plan to provide detailed counter measures against digital profiling was slightly altered and instead we narrowed down our results to

the most important and noticeable findings. This diversion in our initial plan is further explained in the Conclusions Future Development Section 6.

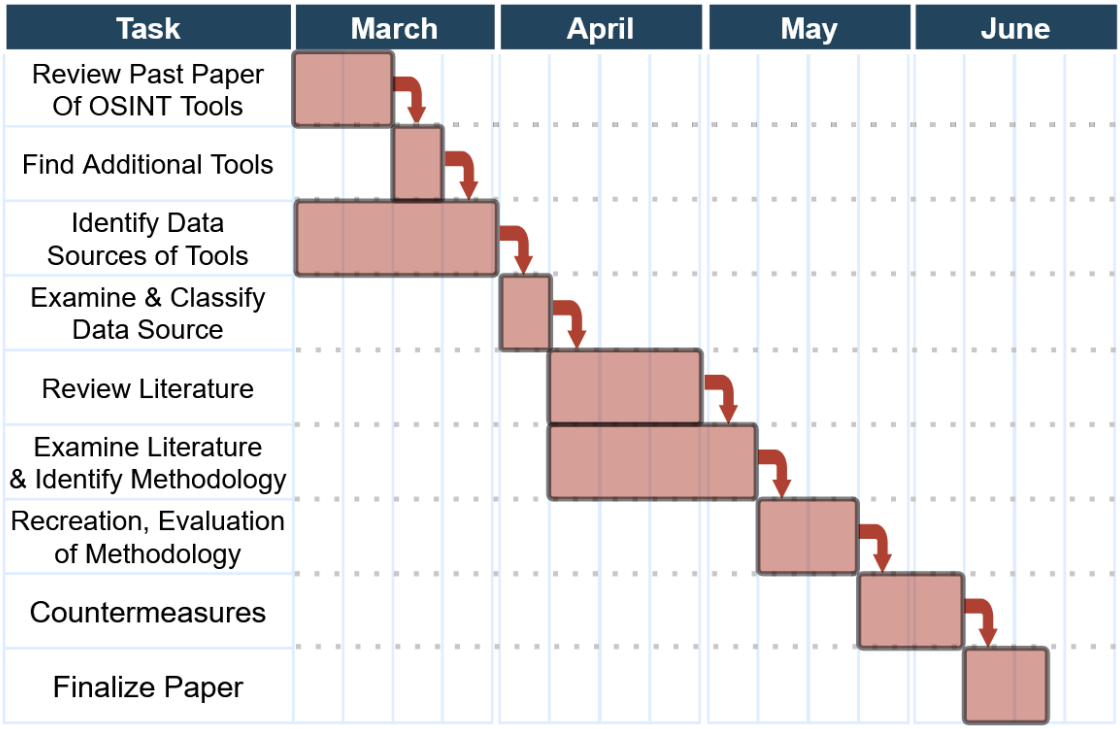


Figure 1: Thesis work plan diagram

2 State of the art

This section summarize the state-of-the-art tools that are used in open-source intelligence. The aim is to review the tools and techniques already existing and go over their scope, potentials, and ease of use.

2.1 Tools, Websites and Techniques

In this brief overview, we will go over the tools explaining their main goals, their architecture, and real-life applications.

Before reviewing the different tools, let us consider what other techniques can be combined with OSINT tools to make an attack or an investigation successful. Today, a widely known technique that is considered to be a social engineering vector is phishing. Phishing is a deceitful way of gathering sensitive information from individuals and the technique can be applied in many different scenarios such as through phone, also known as vishing, or through emails and private messages that is also known as smishing. There are different techniques out there that can be adapted and applied for all the kinds of attacks. In this particular case, using OSINT tools, one can gather enough sensitive information about a target and use the spear phishing technique to create and sent personalized and more sophisticated emails to a few specific end users.

Collecting OSINT is a very challenging process and it is necessary to know in advance the scope of our investigation so that one can look in the right direction when searching for OSINT data. There exist different kinds and types of OSINT but from a more general point of view the following attributes and methods can be used to collect OSINT as shown in Table 1.

Methods	**Attributes**
Cyber Threat Intelligence	Full Name
Reconnaissance	Emails
Search Engines	Usernames
Web scrapping	Social Media
	Metadata
	Data archives
	Geolocation
	IP Addresses & Domains

Table 1: OSINT methods & attributes

2.1.1 OSINT

The OSINT framework [3] is an online website where one can find OSINT resources. The website provides both free and paid alternatives. The resources are displayed in tree structure that is divided into categories such as username, email address, IP address and so on. Each category is divided into smaller subcategories that at end up to either links

for tools denoted by (T), to a URL that contains the search term denoted by (M), URLs and tools that need registration denoted by (R) and finally Google Dork denoted by (D).

Similar to the scope of this research, by expanding the emails addresses category, as depicted in Figure 2, one can see various reconnaissance websites and tools available for extracting email addresses.

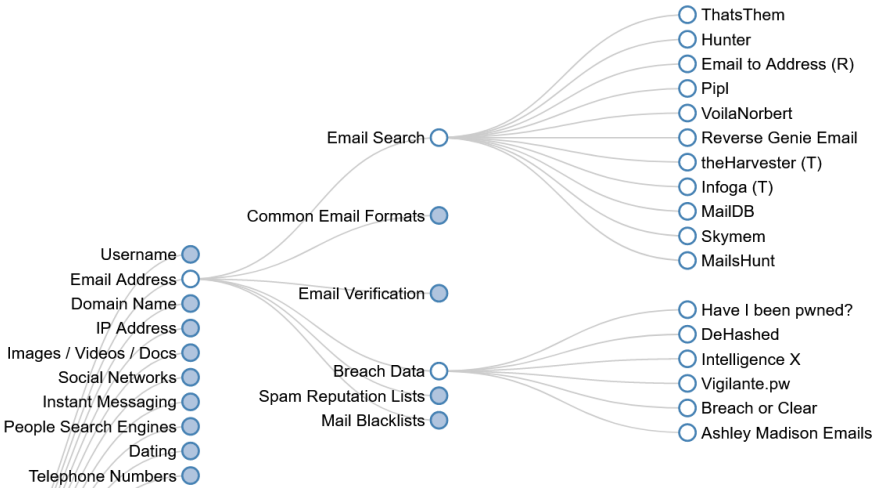


Figure 2: OSINT-framework tree

2.1.2 Internet Archive

The Internet Archive [4] is a website that holds more than 500 billion earlier version of web pages. That means that even if a web page is update or removed from the web, archives such as this might hold a copy of that page. This is extremely useful when it comes to web investigation since one can now browse older pages of forums that might include crucial information for a case.

2.1.3 Maltego

Maltego [5] it's a top of its class tool used by Cybersecurity experts and Law Enforcement for all sorts of different investigations ranging from fraud to threat hunting and digital forensics. The application provides OSINT plugins that are called Transforms which in essence they are small codes that are interacting with the data sources and help the application visualize the results. In this way, the tool can perform information gathering and data mining in real time and with the help of a node-based graph it can visualize the results by creating relationships and connections between them.

On top of that, Maltego uses 57 data sources, used by the Transforms, to extract the required information. This data sources include some of the tools and websites that are reviewed in this section such as Shodan, Internet Archive but also other like Blockchain.info and PhoneSearch.

A short example of Maltego node-based relationship graph is depicted in Figure 3 [6]. The example shows a relationship between 2 companies and their registered officers. In

this way, the graph can help investigators conclude information such as how the officers might me connected or even if some of them are the same person.

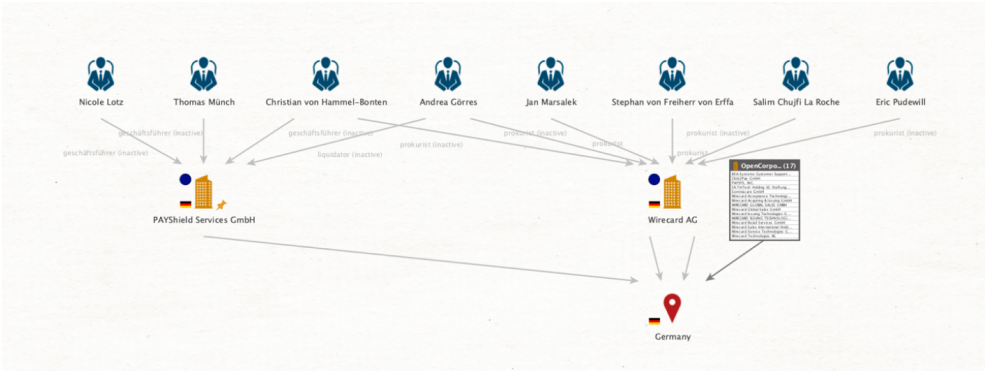


Figure 3: Maltego node-based graph

2.1.4 Shodan

Shodan [7] is search engine for all the internet-connected devices such as webcams, SCADA, routers, and many more. It has servers that are located all around the world and they are crawling the internet 24/7 for those devices. Shodan collects all this publicly available information into its database and thus anyone can query them based on their protocol, ports, software version, vendor name, firmware and so on. Moreover, Shodan evaluate the collected devices and it can conclude whether a webcam uses weak configuration or even whether a device offering a service is a honeypot.

Shodan is not only a great tool for gathering public information but also for evaluating the exposure of a company's services in the public internet. On top of that, it also supports integration with other top tools such as Maltego, Nmap, FOCA and even web browsers like Chrome and Firefox.

By simply querying "SCADA", Shodan recovered more than 2 thousand related devices from all around the world. The exhaustive crawling of Shodan can be seen in Figure 4, where each device comes alongside with its IP address and their response banner, their geolocation, other services available, their operating systems and so on.

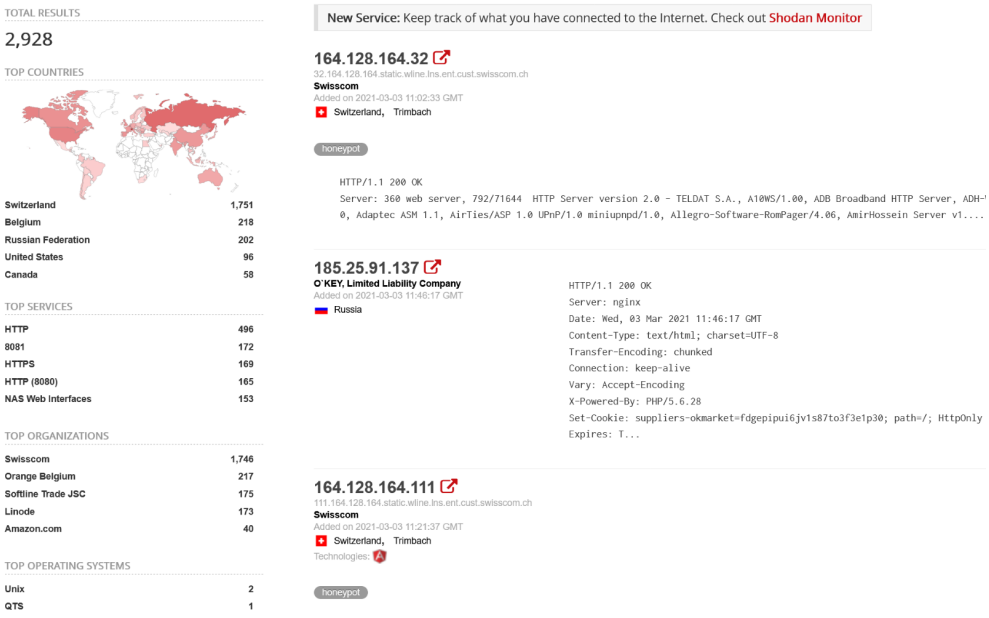


Figure 4: Shodan – Results from SCADA query

2.1.5 Recon-ng

Recon-ng [8] is a top of its kind command line interface tool used for reconnaissance. It is exclusively designed to be utilized as a web-based open-source intelligence tool and is a modular framework which allows developers to easily contribute. On top of that, it automates most of the necessary steps for gathering information. Its marketplace consists of a large number of modules, as depicted in the Figure 5, such as Recon, Discovery, Exploitation, Import and Reporting.

```
discovery/info_disclosure/cache_snoop
exploitation/injection/command_injector
import/csv_file
recon/companies-contacts/pen
recon/domains-contacts/pgp_search
reporting/csv
```

Figure 5: Recon-ng Marketplace Example - Modules Path

The recon module is the largest one available and it is formed of various categories ranging from domains, host to repositories, profiles and locations. Moreover, it also supports API for all the different application to further enhance the reconnaissance. Finally, its integrated database makes the investigation quite easy and fast and it is a great tool especially when the profiling of an individual is required.

2.1.6 The Harvester

The harvester is yet another popular tool for gathering OSINT data from public sources. It can collect all different kind of information such as IP addresses, URL, names, subdomains, email addresses and more. The harvester uses passive sources such as public search engines and services but also active methods such as reverse resolution and DNS brute forcing to obtain the needed information. The tool also supports API keys for integration with other OSINT or pen-testing tools.

As a simple example, by calling "theHarvester -d example.com -b linkedin" command the tool will look for people from example.com domain in LinkedIn. The full list of the data sources can be found at the harvesters GitHub repository [9].

2.1.7 Google Dorks

Google Hacking or google dork, as it is widely known, is a term that indicates an advanced way of using search engine such as Google. Search engines, other than the common search terms, they also accept more advanced operators such as "intext:", "inurl:", "site:", "language:" and so on. This technique allows users to exploit the web by simply using their search engine and nothing more. Dorking does not only apply to Google but also to other search engines to like Yahoo, DuckDuckgo and Bing.

As simple as it may sound, dorking is a powerful way of scraping the public web. A simple example of a dork search query could look like this intitle:"index of" "idx_config" [10]. The results of this simple advanced query contain file with passwords, usernames, authentication keys, salted and many more sensitive information.

2.1.8 ExifTool

Exiftool [11] is a command line application used to read, write and edit wide variety of different file types such as EXIF, GPS, XMP and even from digital cameras like DJI, CANON, GoPro, Kodak and so on. This tool focuses on the metadata of the files that may not be obvious from their indented use. As for example, the Figure 6 displays the EXIF information of an image taken from an iPhone. Moreover, it includes sensitive information like GPS altitude, latitude and longitude as also the time that the picture was taken.

To that end, such information can be considered quite powerful for an investigation or crucial when an attacker is able to get his hands on them.

Figure 6: EXIF Screenshot of an image [12]

2.1.9 Creepy

Creepy [13] or also known as geocreepy, is a simple social media-oriented tool that uses the available information from applications such as Flickr, Twitter and Instagram to gather OSINT data for the given users and display the results on graphical user interface (GUI) map. The application's GUI usability is straightforward and it allows users to simply configure the plugins and supply the usernames of the targets.

Aggregating and analyzing this kind of information can help analyst to visualize and understand the behavior of their target or even, in combination with other OSINT tools, provide enough information for an attacker to perform social engineering and phishing attacks. Creepy it is indeed considered a great tool; however, its support has been unfortunately discontinued.

2.1.10 Spiderfoot

Spiderfoot [14] is a great OSINT reconnaissance tool that is also free and open-source. It integrates an exceptionally large variety of modules which include almost every data source that is available out there. Specifically, it supports over 200 modules and over 100 public sources. This inclusiveness allows spiderfoot to conduct task ranging from simple port scanning and banner grabbing to meta data analysis, IP geo location, threat intelligence, web scrapping, social media account enumeration, dark web scraping, integration with other tools such as Shodan and many more. It also has an integrated database in SQLite and an enterprise version where the tool is fully automated and configured with also a GUI for the visualization of the extracted data.

This all-in-one tool would be of a significant importance for red teams at the early stages of reconnaissance as also for enterprise level applications since the tool automates all the

steps OSINT collection so that the user can focus on the analysis.

2.1.11 Foca

FOCA [15] it's an abbreviation for Fingerprinting Organizations with Collected Archives. This tool is mainly used to search, collect and then analyze documents for their metadata as also any other hidden information that they might include. It utilizes different search engines such as Google, Bing and DuckDuckGo to look for those files either through dorking, as explained in 2.1.7 Google Dorks, or through other means. On top of that, the tool can also look for potential leaks in sensitive directories like admin, manage, upload, control panel, etc. Once this is done, the metadata module will run to extract all the available information. The tool supports a wide range of documents such as Microsoft Office Documents, Adobe InDesign, Adobe PDF, Open Office files, Images, Word Perfect and SVG.

In addition to its main functionality, FOCA can also perform fingerprinting, either passive or active, for HTTP, DNS, FTP, SMTP services and also look for local and remote file inclusion when configured.

2.1.12 Tools Classification

Based on the tools that have been described in this section, the appropriate categories have been created in order to properly classify them based on the purpose they serve as depicted in Figure 7.

Category Description:

- Standalone: Tools that are known to be completely independent. For example, Shodan does not depend on other tools integrations (modules) to complete its tasks.
- Multipurpose: Tools that either support a large variety of integration with other tools or they are depended on other tools to complete theirs tasks. If a tool is marked as multipurpose, then one can find the rest of its functionality(modules) to their official webpages (see references).
- Metadata: Categories such as Metadata have been created since tools like ExifTool serve only one purpose which is to extract and manipulate metadata. However, if any other tools fall into the same category this is also mentioned in the table below.
- Social Media: This category was created based on Creepy and it serves the same purpose as the Metadata category.
- Search Engines: Following Metadata and Social Media categories, this category was created due to the notion of Google Dorking.
- Offensive: Offensive category tools that support integration(modules) with any kind of offensive tool.

Tools	Categories					
	Standalone	Multipurpose	Search Engines	Social Media	Metadata	Offensive
Internet Archive	Archiving Web					
Maltego		Yes	Module	Module	Module	Module
Shodan	Internet-Connected Devices					
theHarvester		Yes	Module	Module		API Integration
Recon-ng		Yes	Module	Module	Module	Module
Google Dorks			Advanced Search			
ExifTool					Manipulation	
Creepy				Geo Location		
Spiderfoot		Yes	Module	Module	Module	Module
FOCA		Yes	Used for Dorking		Documents Analysis	Module

Figure 7: Classification

3 Reviewed Papers

In the following section we review different papers that were focused on analyzing and identifying individuals through their social media accounts and the data publicly available on the internet.

The papers are mainly discussing and proposing methodologies to collect, aggregate and analyze the raw information gathered. The methodologies presented below are 4 main approaches that were utilized the most by the reviewed papers. Profiling and linking identities is a challenging area of OSINT and it requires a considerable amount of time to collect and preprocess the data. The collected intelligence must be of a meaningful shape and format since it will be later on fed to either a ML algorithm or be analyzed by data analyst.

3.1 Natural Language Processing

In this first article [16], research aim to predict the location of users by extracting all the publicly available information from 3 main social network platforms namely Twitter, Facebook and Instagram. To achieve their goal, they developed a toolkit that consist of different modules for collecting data. The paper covers the following methodologies for data collection: data retrieval and analysis of posts, texts and pictures that includes full-text search, hashtags, geotags (aka check-ins, location mentions) and geolocation information. In addition, the paper makes use of machine learning and NLP techniques to extract the "hidden" information in the collected data. As for example NLP can be used to process text and extract valuable information. For instance, posts might not include hashtags and geotags, but the users might expose their location through location description and mentions.

Similarly, in another paper [17], they developed an OSINT tool called TExtractor that its goal is to extract and analyse audio and video content. This tool can take not only a text input but also a video or a recording and perform a speech recognition to extract intelligence. The paper addresses the cyber threats domain and argues that the tool could be used to search for keywords that will link to malicious activities as also potential malicious actors.

In general, the use of NLP in profiling seems to be an integral part of the process. In the research world, NLP can be used in ML, statistics and linguistics in many different applications. Its goal is to process the data given, analyse and provide an understanding for the text or speech provided. NLP is of a great importance and it provides high level solutions for OSINT tools and investigations. In addition to the applications discussed above, there exist a wide variety of other applications of NLP that are both interesting and novel. Especially in cybersecurity, NLP could be used to gather data from a company's infrastructure, CVE's, cybersecurity related-articles and analyse them to identify vulnerable targets and possibly create profiles for threat actors and the methods they might uses against them.

3.2 Attributes & Nodes Similarity

The attributes selected is another crucial step in profiling process since these kinds of investigations require a multitude of different personal details. Selecting which attributes are going to be needed also depends on what methods will be used to process the data. It is crucial for a researcher to be fully aware of the attribute used since they can affect the final result directly.

In the following paper [18], the attributes selected were either extracted directly from social network profile of the target, which it falls in the category of self-disclose data, or they were derived from indirect information such as connections, interests and so on. Specifically, the aforementioned research paper uses the following attributes: Username, Name of the user, Location, Popularity, Language used, Active time zone, Connection type, through a group, Area of interest labels, other social network connection, Registration date, Common event Registration. Based on these attributes the researches utilize node similarity methods which are algorithms that take a set of nodes and compares them based on the nodes that are connected to. For example, having two nodes that share a lot of neighbouring nodes can be considered similar. For the comparison they make use of various methods such as cosine similarity method, Euclidean distance similarity method, word n-gram similarity and so on.

In yet another paper [19], they developed a profiler tool that could be used for the digital profiling of criminals. In this research they used a large variety of attributes that were further utilized to extract important information and enhance the target's profiling. The attributes used are the following: Name, Profile picture, User ID, Gender, Birthday, Religion, Political view, Education, Work Hometown, Location, Friend request, Relationship, Friend, Music, TV, Movie, Book, Activity, Group, Website, Link, Note, Event (Tagged), Photo (Tagged), Video, Message, Status update.

3.3 Behavioral & Content-based analysis

Today social media network application, especially the most famous ones, have already in place and enabled by default all the necessary privacy settings. The user is able to choose between different level of privacy that are categorized in the following way:

- Make personal details publicly available so that anyone can see their name, username, etc
- Available only to their friends
- Available to Friends of friends
- Completely private.

Despite the people's lack of knowledge and awareness about privacy, applications like LinkedIn may subliminally force the user to make his profile public so that the user has a better chance to be found. However, in the case where such user attributes are not available, one can identify and correlate user profiles by performing a behavioural analysis. The paper [20], conducts a research on Twitter with the ultimate goal of identifying

malicious or benign users. For the behavioural analysis part, they mainly make use of the SPOT (for Scoring Suspicious Profiles on Twitter) methodology and NUANCE for the content-based analysis. In a brief overview, SPOT performs the classification of the users' profiles based on set of behavioural features and the from the NUANCE framework they make use of Recentred Local Profile's algorithm in order to compare the profiles foreach account. Some of the features used as metrics are: account age, frequency tweet, average (hashtags), average @(mentiones), number of friends, number of followers, number of tweets.

In similar paper [21], they have developed a tool call SocialMatching++ which contains additional behavioural features other than the common ones like life events(eg. graduation year). They claim that these features are enhancing the behavioural analysis and their main objective is to discover if different users' profiles of the same individual are linkable in the different social media applications.

3.4 Machine Learning

Lastly, despite all the above methodologies, the most fundamental aspect of every paper reviewed is ML. It is indeed an integral part of almost all of the researches since its predictive capabilities are of great importance. Researchers use ML in order to save a lot of hours of manual work and they either use it to classify data or predict an outcome. In a profiling investigation, as explained before, this is necessary since behavioural analysis of hundreds of thousands of users is infeasible to be performed manually.

To begin with, in the paper [22], the researchers crawl a community of YouTube users and they make use of their comments, uploads, favourites and playlists in order to predict their attitude and classify them as threat against law enforcement. Based on their attitude the users are divided into categories P for negative and N for holding a non-negative attitude towards authorities. To achieve that, they trained a simple Bayesian ML classifier and they also used a dictionary-based approach that is basically a classification of comments that is performed with a list containing expressions and terms that express negative attitude.

In this next paper [23], the 2 methodologies were used. The first one was a ML algorithm that was used for behavioural analysis over the user features extracted from the identities. The second methodology consist of 2 different approaches for collecting and identifying user intelligence. The ultimate goal of this specific paper is to link the different identities of the same user in the numerous social media applications.

4 Methodology

This section includes a detail description of the methodology used to carry out this research. The methodology is comprised by a sequence of steps. Namely there are 6 steps:

- Identification and classification of the tools discussed in Section 2.
- Tool's data sources classification.
- Practical utilization of the tools including detailed examples of how they work and how they extract information.
- Reviewed literature.
- Conduct an analysis based on the methodologies from [23].
- Concluding the results by proposing countermeasures against the profiling of individuals.

The above sequence of steps can be summarized, as depicted in Figure 8, in a waterfall scheme of 3 categories.

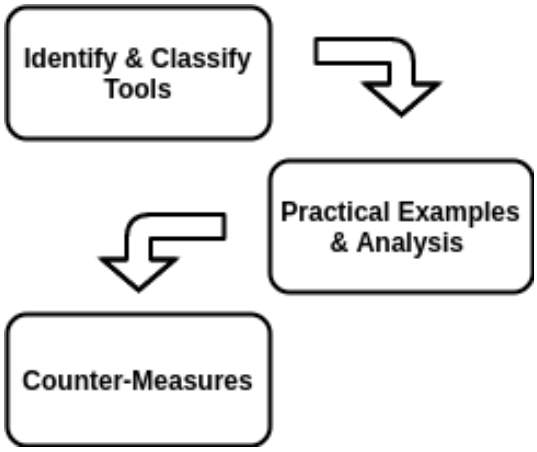


Figure 8: Methodology

4.1 Tools Identification & Classification

The identification and classification step of the tools has been discussed in detail in the State-of-the-Art Section 2 of this paper. The section plays the most important role in the development of this research paper. Briefly, it collects together and discusses the most important sources and tools for performing open-source intelligence tasks and it also provides a short description for each tool about its main goal and capabilities. Additionally, we also review the existing approaches for performing a profiling investigation and the tools used. The section's objective is to identify profiling tools and methodologies and classify them. This information will then be used to examine how these tools are collecting, analysing and presenting the collected information.

In more detail the process goes as follows:

- Review and collected OSINT tools
- Provide a brief description on the given tools
- Classify the tools
- Collect the methods and their source of information (eg. Databases, Social Media, etc)
- Review existing profiling methodologies
- Provide a brief overview of each paper

Worth mentioning, is the fact that there exist a lot of different sources and tools that can perform a wide variety of OSINT tasks, however since a lot of tools make use of the same sources and they even support integration for other tools, is thus irrelevant to include every single source and tool available out there. For example, the OSINT framework mentioned in Section 2, provides an exhaustive list of available tools, websites and sources for performing all kinds of OSINT tasks.

4.2 Tools' Data Sources & Methodology Recreation Analysis

After gathering all the necessary information about the tools and reviewing the papers we can now begin the practical examination. There are 2 practical parts, the first one is to use each tool to performing a profiling OSINT scan on a specific individual and the second one is to review the existing methodologies from the paper [23]. All the results are presented and discussed further in Section Results 5.

From the tools gathered in Identification & Classification step, we will now focus only on those specific tools used for profiling. Namely these tools are: Maltego, Recon-ng, theHarvester, Spiderfoot.

4.3 Tools & Data Sources

This step will help us understand the inner process of each tool and how it makes use of the available resources online. For example, one of the primary discoveries we are interested in is the parameters that the tools are using to target a person. Furthermore, to successfully achieve that we must first identify all of the data sources (whether are tools, websites, archives, databases, etc) that the tools are using and depend on.

The approach we conducted goes as follows. Firstly, after reviewing the OSINT tools we identify those used for profiling as explained in subsection 5.2. Then we collected and classified all the data sources that these tools are using as presented in subsection 5.1. The classification process is crucial since in the last step we have to examine and understand how and where the different data sources collect their data from. The methodologies used to collect data will then be used as a point of reference to propose potential counter measures.

4.4 Methodology Recreation

Furthermore, we reviewed the existing profiling literature in order to determine the current techniques and technologies used for collecting, aggregating and analysing OSINT information. The searched keywords can be found in Appendix 6.

Moreover, we are briefly recreating the data collection methodology presented in the paper [23]. This paper, as many others, have presented a way to collect and link user identities of their different social network applications. We are utilizing these methodologies for a simple profiling investigation of both specific and random targets in order to get an overview of the level of accessibility of such information. At this point, is important to mention that to carry out this step we created a new google account and all the queries were made through TOR browser.

The methodology consists of 5 data collection techniques presented in Table 2.

Techniques	**Description**
Advance Search Operator	This term refers to the dorking techniques used in search engines.
Social Aggregator	Business/Websites such as About.me
Cross Platform Sharing	Sharing content from one SMA to another
Self-Disclosure	Disclosing in one SMA other SMA's
Friend Finding Feature	SMA use your contact emails to suggest friends.

Table 2: Data collection techniques

Finally, both of these practical assessments will eventually provide us with an in-depth understanding with the ultimate goal of identifying potential counter-measures against the tools and the existing methodologies for profiling.

4.5 Counter-measures Proposal

The Counter Measures Section 5.4 is the last step of the is research. The counter measures are basically a conclusion of the overall findings of the methodologies used to collect, uncover and identify personal details. The results are aggregated and analysed closely since most of the methodologies used are overlapping with each other with slightly different modification based on the objectives of each research. All in all, we have examined the results from the literature review, the OSINT tools and the methodology recreation in order to propose counter measures that can be considered as privacy phase before creating and publishing content online.

5 Results

In this section we will present the results of this research. The section is divided into 3 main subsections namely the Data Sources Classification process, OSINT Tools Practical Examples & Analysis and Use case Recreation Analysis Results.

Before moving forward, is important to say that there is an immense amount of publicly available data today on the internet and that makes the data collection and the analysis a very challenging and complex process. Furthermore, the collected data is, for the most part, diverse in format and unstructured. This makes everything even more difficult since it becomes extremely difficult to distinguish between valuable and unrelated data. However, as mention in Section 3, today there exist different algorithms such as ML or NLP that can be utilized to extract and infer context from such data.

Another worth mentioning point is the fact that the most data available on the internet is in English and thus results in any other language are not considered and not processed. Valuable and important information can be also found in other languages and with the help of translator software we might be able to extract the context and increase the amount of information to be processed.

5.1 Data Sources Classification Process

In order to understand the nature of the sources used by the tools we must first classify them and analyse them in categories. OSINT tools, for the most part, make use of other OSINT tools and data sources. After going through all the data source for the aforementioned tools, we have created 5 categories. These 5 categories represent all the different types of data sources that tools and researchers can use to collect and infer data. The full list of each tool's data source can be found in Appendices. Let us now go over each category.

5.1.1 Social Media

In the first category, Social Media, we have included data sources that are based on raw SMN API calls. For instance, creating and utilizing directly the APIs of SMN such as LinkedIn, Twitter, YouTube and so on.

5.1.2 Search Engines

Next is the Search Engine category which includes all the different search engines that are utilized to crawl the web for information such as Google, Bing, DuckDuckGo, IntelligenceX, Baidu, Bing, Exalead, Yahoo and so on. The search engines are using dorking in order to perform a more advance search and thus some tools like theHarvester, as explained in Section 5.2.3, don't require API to query the different SMNs. In this way, the search results become more complicated and difficult to obtain since the use of a script is necessary in order to scrap the results.

5.1.3 Scripts Crawlers

This brings us to the next category which is Script Crawlers. This category utilizes 2 approaches, the first one is fixed URLs and the second one is automated scripts. As explained in 5.2.2, 2 straight forward examples are the Profiler and dev_driver modules which in essence they consist of a list of the SMNs URLs used for navigating to the users' application profiles. With the script included in these specific modules in Recon-ng the tool then uses the arguments passed by the researcher and it then query's all the listed URLs.

5.1.4 DB - Data Breaches Crawlers

Another type closely related to crawlers is the DB - data breach crawlers. This category consists of mainly databases that were populated either by publicly exposed breached data or by businesses that use to collect information, including personal data, about people. Some of the most known modules that fall in this category are HaveIbeenpwned and FullContact respectively.

5.1.5 Local Scripts

Finally, the last category Local Scripts is referring to any type of script that mangles the given input with the ultimate goal to uncover user details about our target. In Recon-ng the modules mangle and unmangle serve exactly this purpose since the first one makes use of the information stored in the database to create usernames and email addresses and the latter, using again the DB, to uncover individual name components.

5.2 OSINT Tools Practical Examples & Analysis

In this section we will present the results obtained from the tools. Specifically, we will make use of Kali Linux distribution and the following tools: Matelgo, theHarvester, Recon-ng, Spiderfoot.

At this point is important to mention that the tools are using a wide variety of data sources and as a result some of them require the use of a paid API or they have a paid version and thus cannot be utilized in this research. As for example in Maltego the data sources available at the community version are less than 10. As we have already explained before, from all the identified data sources used by those tools, we are only going to use the ones concerning profiling so that we can further examine them.

In more detail the total number of data sources identified from those 4 tools are 391. From these, only 66 are used in digital profiling of an individuals and from those data sources only from 37 were returning an outcome. It is also important to note here that these numbers represent the total amount of all data sources aggregated from each tool and thus the numbers are not unique.

5.2.1 Maltego

Maltego is a top of its class tool with a very straightforward user interface. The equivalent of the modules in Maltego are called Transforms, transforms are either scripts or a service that is provided by private companies. The transforms can be found in Appendix **??**.

Firstly, we have to mentioned that we used the community edition of matlego and thus the transforms available were limited. In Figure X, we list all the free transforms related with profiling. When we install Maltego, by default, only the Standard Transformations are installed and available for use but based on our free subscription we can add more by installing them manually.

Since Maltego supports more than 50 transformations, we must narrow down the used modules to only those relevant for profiling a person. Firstly, we select the categories that possibly will unveil to us the wanted information. The categories selected can be seen in the Figure 9 under the Section Data Categories.

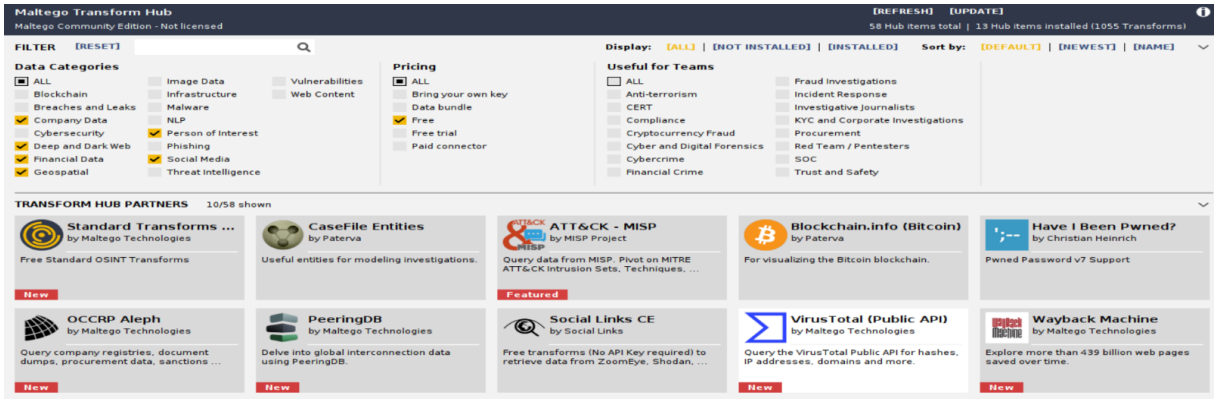


Figure 9: Matlego's Profiling Transforms

Once we have installed all the modules and acquire the necessary API's, we will now continue with the profiling. Firstly, we create a new empty graph. All the modules loaded and install in the previous phase are now available to us in the Entity Palette indicated by point 1 in Figure X. Each Entity represents an object which has its own attributes. For example, based on the information available to us, we select the Person entity from the Personal category. The entity can be dragged and dropped in our new empty graph indicated by point 2 in the Figure 10. The selected entity, that can be seen on the right indicated by point 3, requires some attributes which in this case is the person's name and surname.

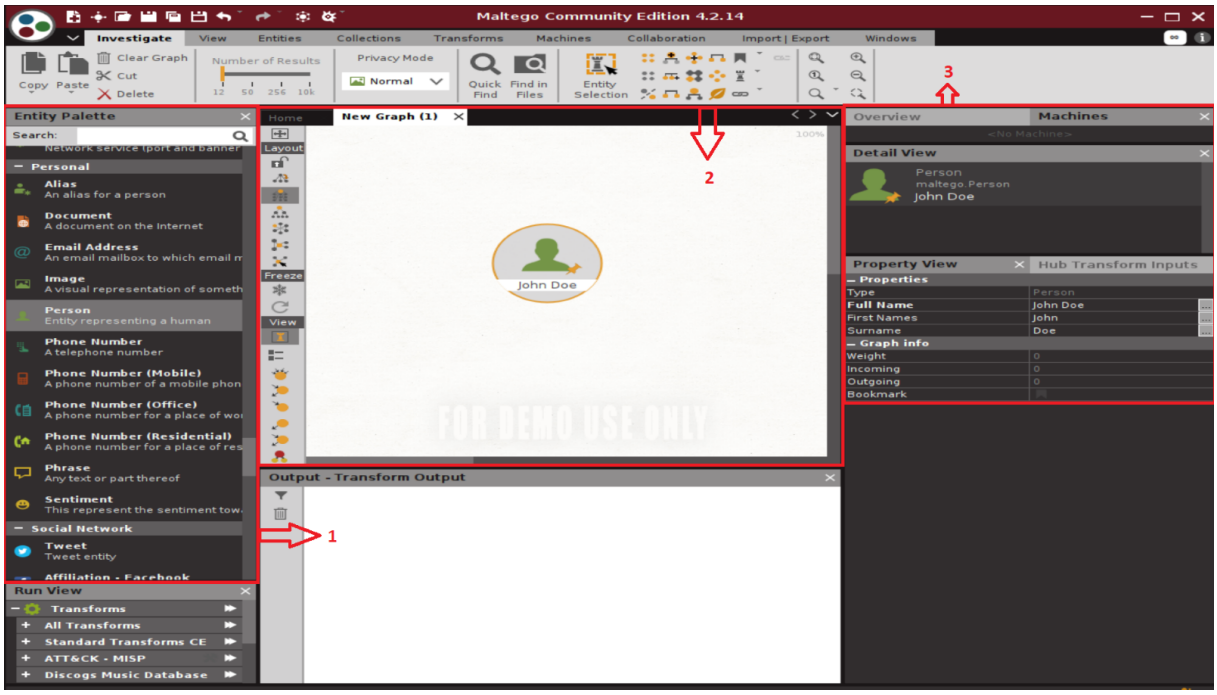


Figure 10: Maltego's main investigative page

The attributes are necessary for the process of profiling since the transformation require some kind of input depending on the case. Given the full name of our target, as in Figure 11, we can do a right click to list all available transformations categories for the selected entity. Next, we can simply choose to retrieve all the different information about our target using the category "All Transformation" and Maltego will make the connection between the results and our target for us.

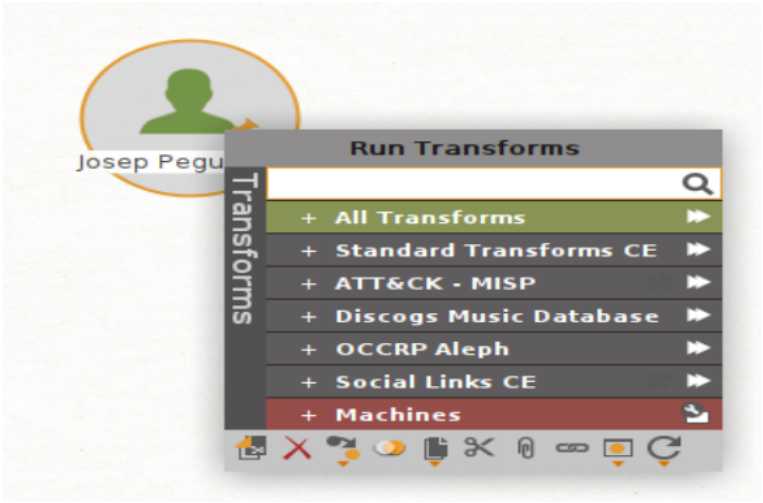


Figure 11: List of all available transform for the entity

Worth mentioning is that fact that by just providing a simple name, Maltego will poten-

tially return a large number of results. To that end, is extremely difficult to present all of our findings in this research so we are just going to focus on a few important and crucial information. In Figure 12, we present the results after executing all transforms and it is clear that they are numerous results about our person of interest including different ID's accounts, documents and more. One observation is that the name given to the entity is not used in static way but rather it has been interpreted so that it can be mangled in a more humane way. For example, the name Josep was changed into Joseph and that is why we can also see result containing the name Joseph Pegueroles also. Maltego also supports a transform called Rosette that helps with names to be interpreted and translated into different languages with different alphabets like Arabic. In a more thorough investigation, each result would have been examined in depth, however in our case we will focus our attention only on the obtained email addresses indicated by the symbol "@".

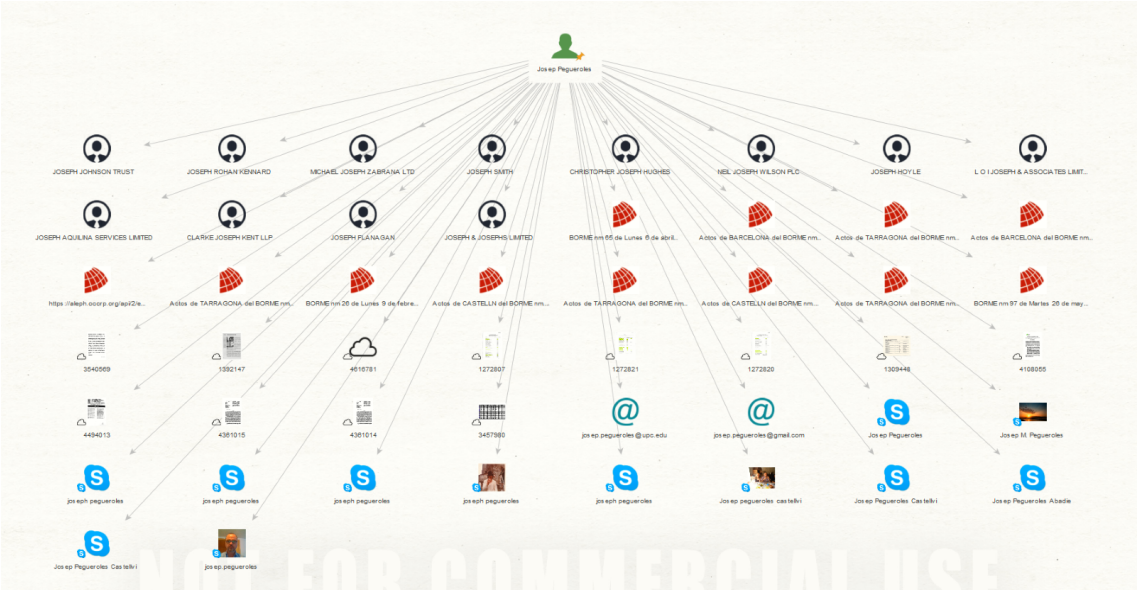


Figure 12: Findings of the person entity

For the sake of the research the results have cleaned out the results so that we can present our findings in a clear and straightforward manner. By repeating again the category "All Transforms" for the 2 obtained email addresses the results are even more interesting. We can now see how Maltego makes use of the obtained results in order to create relationships between entities.

Let us take a closer look into the findings. In the Gmail address, on the left-hand side of the Figure 13, we recover a MySpace account depicted as a tower and on the right-hand side a full name entity of our target retrieved from the UPC domain. The middle part of the graph provides the relationships between our three main entities, Gmail, UPC and our human entity. From them we recover a LinkedIn account depicted as a tower, a Skype account, and 3 other entities in the shape of a small red fire. These entities indicate that the accounts are included in leaked account lists, also the accounts are falling in the category of complainer accounts meaning that this address is frequently unsubscribing

from marketing lists and finally that the accounts are common email providers.

Figure 13: Findings of the all the email entities

Overall, even with this small community edition of Maltego we can still recover a lot of information about a target. Figure 14 depicts a small example of how chaotic the results could become for a single target. Each of the results however need a thorough examination and must be further analyzed so that an investigator can recover not only accounts and names but also context about a target as for example if the target is into politics and so on.

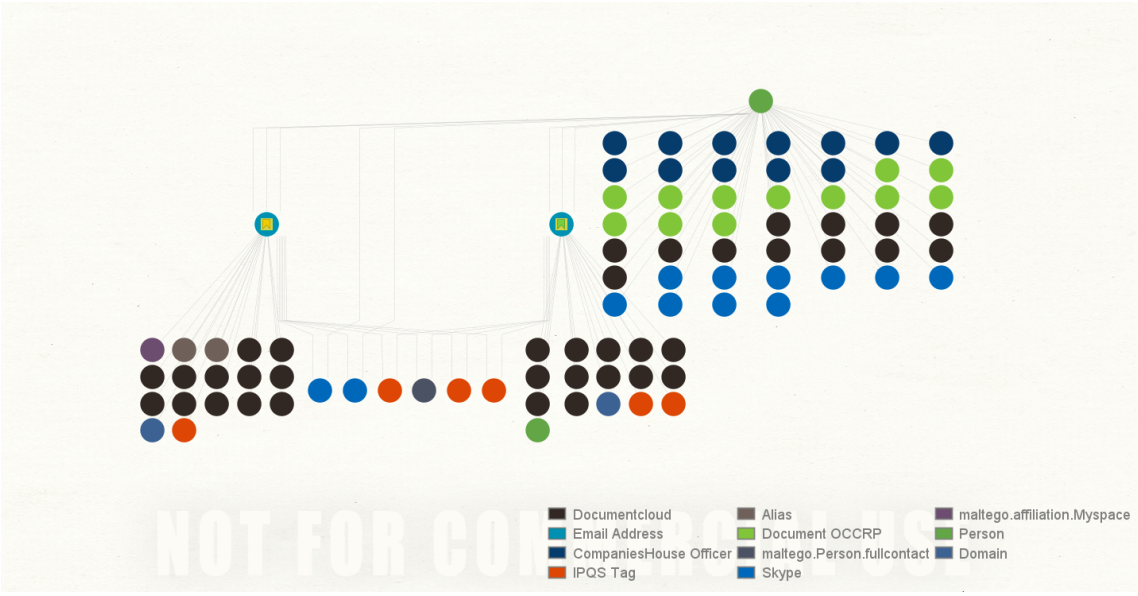


Figure 14: Small non-processed investigation

5.2.2 Recon-ng

Recon-ng is yet another powerful tool for reconnaissance. Developers can easily contribute toward the expansion of the tool since recon-ng is based on modules. Modules are divided into categories based on their purpose and each module serves a different case. The tool also supports a database scheme that allows an investigator to combine all the available data together.

The first and most important task is to identify which of the available modules are used for profiling - Appendix **??**. Once this is done, we conduct our profiling investigation with our target. The goal will not only be to recover information about our target but also analyze the script behind each module so that we can later on examine some countermeasures against them.

In summary, Recon-ng utilizes breached data, queries the API of data providers or it uses python scripts to crawl and sift the web. In Figure 15, a short section of the marketplace is depicted [24] and one can observe which modules require API keys and/or dependencies to work properly.

```
[recon-ng][default] > marketplace search
```

Path	Version	Status	Updated	D	K
discovery/info_disclosure/cache_snoop	1.1	not installed	2020-10-13		
discovery/info_disclosure/interesting_files	1.1	not installed	2020-01-13		
exploitation/injection/command_injector	1.0	not installed	2019-06-24		
exploitation/injection/xpath_bruter	1.2	not installed	2019-10-08		
import/csv_file	1.1	not installed	2019-08-09		
import/list	1.1	not installed	2019-06-24		
import/masscan	1.0	not installed	2020-04-07		
import/nmap	1.1	not installed	2020-10-06		
recon/companies-contacts/bing_linkedin_cache	1.0	not installed	2019-06-24		*
recon/companies-contacts/censys_email_address	1.0	not installed	2019-08-22		*
recon/companies-contacts/pen	1.1	not installed	2019-10-15		
recon/companies-domains/censys_subdomains	1.0	not installed	2019-08-22		*
recon/companies-domains/pen	1.1	not installed	2019-10-15		
recon/companies-domains/viewdns_reverse_whois	1.0	not installed	2019-08-08		
recon/companies-domains/whoxy_dns	1.1	not installed	2020-06-17		*
recon/companies-hosts/censys_org	1.0	not installed	2019-08-22		*
recon/companies-hosts/censys_tls_subjects	1.0	not installed	2019-08-22		*
recon/companies-multi/github_miner	1.1	not installed	2020-05-15		*
recon/companies-multi/shodan_org	1.1	not installed	2020-07-01	*	*
recon/companies-multi/whois_miner	1.1	not installed	2019-10-15		
recon/contacts-contacts/abc	1.0	not installed	2019-10-11	*	
recon/contacts-contacts/mailtester	1.0	not installed	2019-06-24		
recon/contacts-contacts/mangle	1.0	not installed	2019-06-24		
recon/contacts-contacts/unmangle	1.1	not installed	2019-10-27		
recon/contacts-credentials/hibp_breach	1.2	not installed	2019-09-10		*
recon/contacts-credentials/hibp_paste	1.1	not installed	2019-09-10		*
recon/contacts-credentials/scylla	1.3	not installed	2020-09-14		
recon/contacts-domains/migrate_contacts	1.1	not installed	2020-05-17		

Figure 15: Recon-ng Marketplace

Twitter

Another interesting module used for profiling is the Twitter. Twitter modules twitter_mentions and twitter_mentioned use the API of Twitter to query, based on a given name and surname, tweets that the user was mentioned or tweets that the user mentions someone else.

Firstly, we create and provide recon-ng with the necessary API keys as in the Figure 16. At this point is important to mention that we must provide the username of our target in the username field of the profile database. The username however must be in the exact

form of the user as depicted in Figure 15. For example, if the username of our target is a concatenation of his name and surname, then anything other than this will not return any results. Before we advance further in our investigation is important to mention that Twitter limits searchable tweet history to 7 days which means that older tweets that mention our target will not be available through the API.

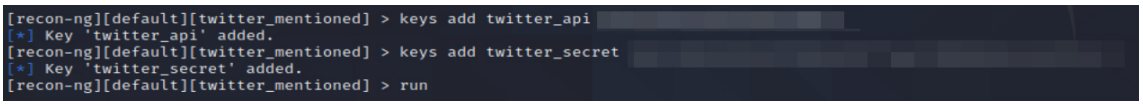


Figure 16: Recon-ng API keys

In this particular case we first verified our target's Twitter username so that we are sure the results will be accurate. After setting up everything, we can now start our investigation by typing "run" in our terminal. The module twitter_mentioned is querying the API for tweets that our target has been mentioned. As in Figure 17, this resulted in a short list of tweets that yield the user mentioned our target and their twitter account URLs. In contrast, the twitter_mentions is querying all tweets that our target has mentioned someone else. In this particular case each result is further queried to analyse the accounts that were mentioned.

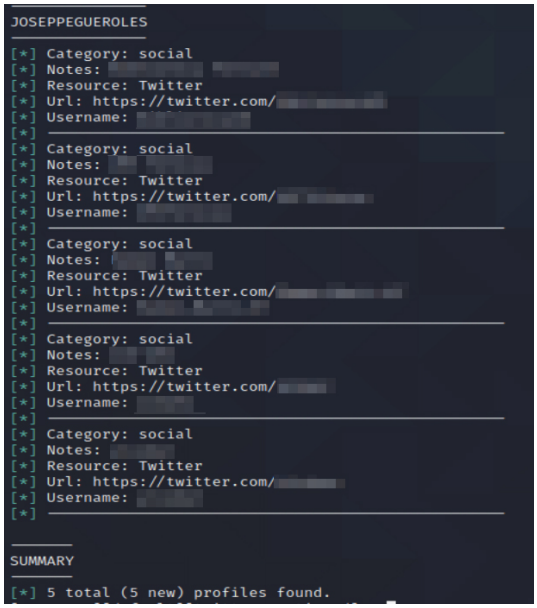


Figure 17: Recon-ng Twitter_mentioned results

After analysing further, the total number of the resulted accounts for users that were mentioned by our target is 267. Due to the substantial number of information only the summary is depicted in Figure 18. The retrieved information is all aggregated into the profile database of the tool and we can now simply type the command "show profiles" to review all the collected information from the twitter modules. The table profiles provide

information about the users that interacted with our target, either people or organisations, that would help us draw some conclusions based on the context of the results in later stages of our investigation. As for example, in Figure 19, in the notes section we can observe information like "UPC Telecos","Biblioteca Ferrate" or "ICE-UPC". In a real investigation this kind of information could lead us to conclude things about our target like occupation, visited places, friends and so on.

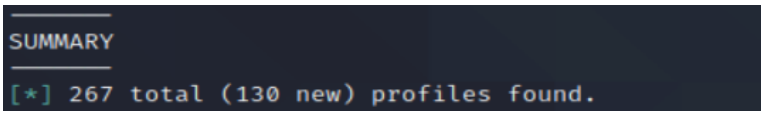


Figure 18: Recon-ng Twitter_mentions total

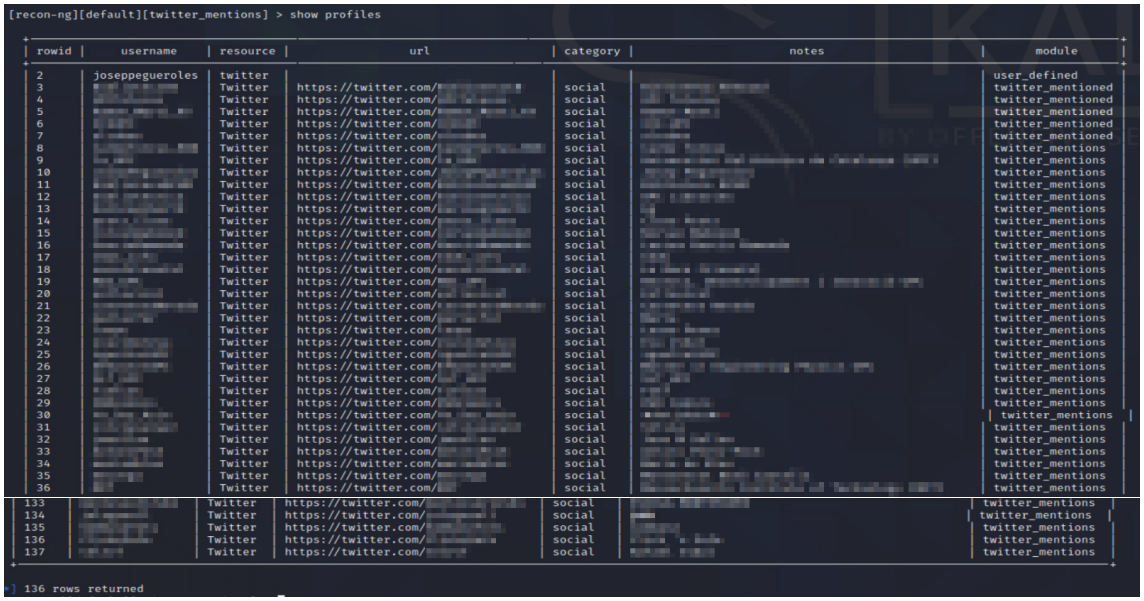


Figure 19: Recon-ng profile table

Github_users & Dev_Diver

These 2 modules, github_users and dev_diver, are both focused on the repository web application information and since they fall in the same category of profile-contacts in recon-ng we are going to provide some short examples of how they work here in the same section.

The first module, github_users, requires an API to work. It queries the GitHub API to collect all the publicly available information about the target. Again, we only have to provide a username for our target and we must make sure beforehand that the username exists.

First, we test an account that we know it exist. From the Figure 20, we see that if the account exists, we can retrieve all the available information.

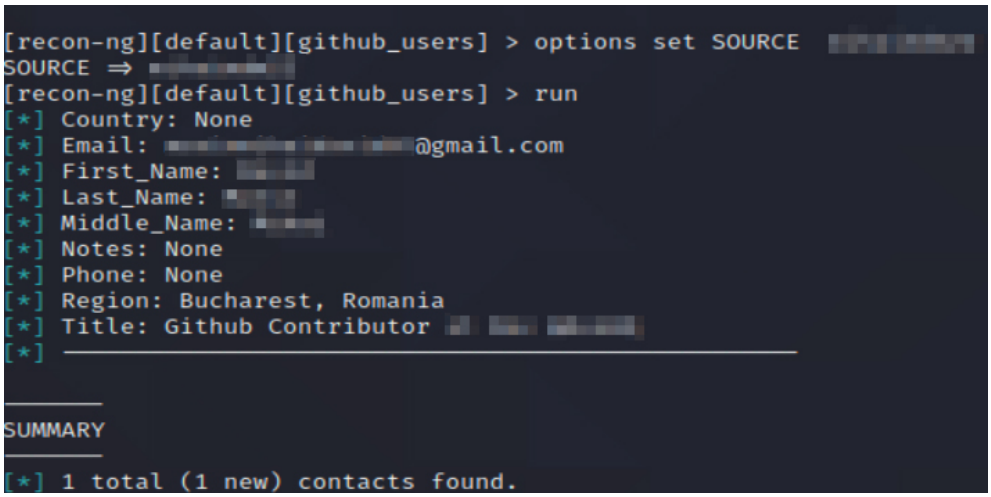


Figure 20: Recon-ng results GitHub users

Finally, by performing a few tests with some usernames we obtain the following table including information like name, surname, email, country, phone and so on as depicted in Figure 21.

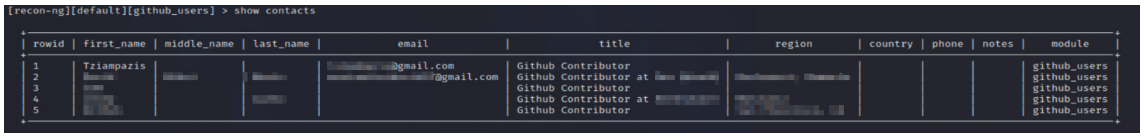


Figure 21: Recon-ng profile table

The next module called dev_diver, is a short and simple script that only requires a username to work. This module, as also many other modules in recon-ng, operates with URLs that are specifically crafted to query the existence of a usernames in code repositories. If available the script returns all possible information about the user. Figure 22 depicts the 4 different repositories that are quired in this module.

```
url = f"https://api.github.com/users/{username}"
url = f"https://bitbucket.org/api/2.0/users/{username}"
url = f"http://sourceforge.net/u/{username}/profile/"
url = f"http://www.codeplex.com/site/users/view/{username}"
```

Figure 22: Recon-ng dev_diver module URLs

Profiler

The profiler is a very simple script which makes use of certain URL username formats from the different websites listed in [25]. The script requires again a specific username to work. The json containing the websites' URL, used by the script, consist of a couple hundred websites from across different categories. An example of both is depicted in Figure 23.

```
"categories" : ["art","blog","business","coding","dating","finance","gaming","health",
                "hobby","images","misc","music","news","political","search","shopping","social",
                "tech","video","XXXPORNXXX"],
"sites" : [
   {
      "name" : "7cup",
      "check_uri" : "https://www.7cups.com/@{account}",
      "account_existence_code" : "200",
      "account_existence_string" : "Profile - 7 Cups",
      "account_missing_string" : "Oops! The content you're attempting to access could not be found.",
      "account_missing_code" : "404",
      "known_accounts" : ["john", "jbob"],
      "category" : "social",
      "valid" : true
   },
   {
      "name" : "asciinema",
      "check_uri" : "https://asciinema.org/~{account}",
      "account_existence_code" : "200",
      "account_existence_string" : "s profile - asciinema",
      "account_missing_string" : "This page doesn't exist. Sorry!",
      "account_missing_code" : "404",
      "known_accounts" : ["john", "red"],
      "category" : "coding",
      "valid" : true
   },
```

Figure 23: Recon-ng profiler module URLs

After we provide the username, we execute the module and the results can be found in profiles database as depicted in Figure 24.

rowid	username	resource	url	category	notes	module
1		LinkedIn	linkedin.com/in/joseppegueroles			user_defined
2		LinkedIn	https://es.linkedin.com/in/joseppegueroles			user_defined
3		LinkedIn	joseppegueroles			user_defined
5	joseppegueroles	issuu	https://issuu.com/joseppegueroles	shopping		profiler
6	joseppegueroles	Blogspot	http://joseppegueroles.blogspot.com	blog		profiler
7	joseppegueroles	GitHub	https://github.com/joseppegueroles	coding		profiler
8	joseppegueroles	Pinterest	https://www.pinterest.com/joseppegueroles/	social		profiler
9	joseppegueroles	slideshare	https://www.slideshare.net/joseppegueroles	social		profiler
10	joseppegueroles	Twitter	https://shadowban.eu/.api/joseppegueroles	social		profiler

Figure 24: Recon-ng profiler module resutls

Flickr - Location

Using an API, this module is retrieving geolocation information from Flickr. The module requires a latitude and longitude coordinates and also the radius in kilometres. As in Figure 25, it returns all the posts within the given radius providing all the necessary information like timestamp, the URL of the post, the profile name of the user and so on. Similarly, other recon-ng modules like YouTube, use the same technique to retrieve information about the geolocation of their posts.

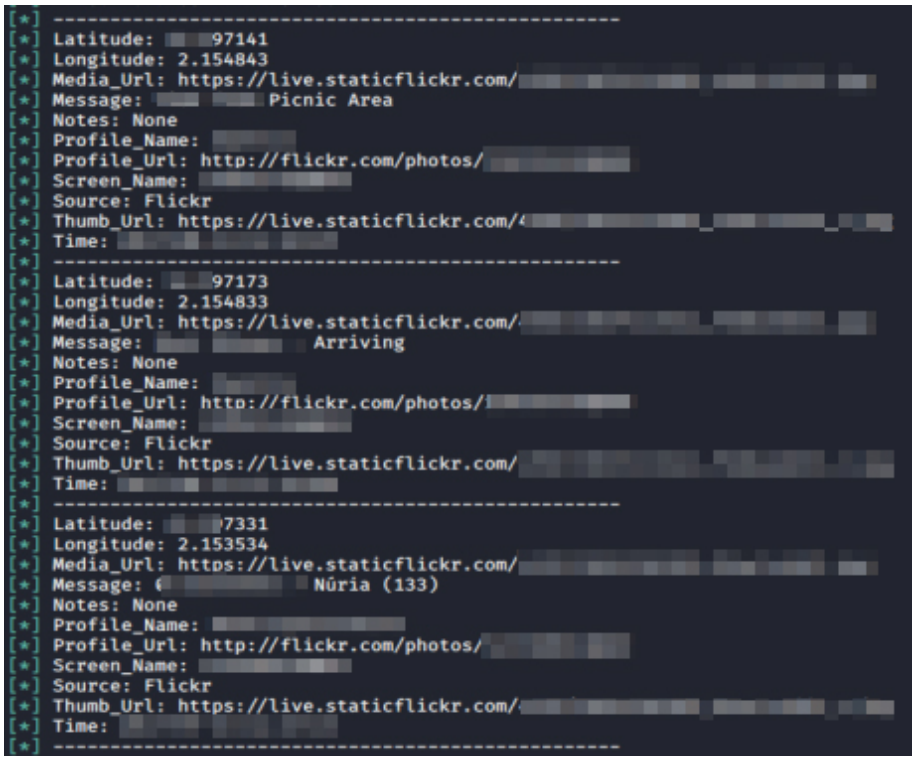


Figure 25: Recon-ng Flickr geolocation

5.2.3 theHarvester

This tool is yet another powerful one but with less capabilities compared to the other tools we are reviewing. Concerning profiling, all the modules listed in the Appendix ?? seems to be utilizing, at the most part, search engines to retrieve their results.

The Parser

Since the modules used by theHarvester are based on search engines and social media, the tool has its own parser to parse the results. By taking a closer look at the source code of the tool [26], we can see different scrappers for different scenarios. An example is depicted in Figure 26, which is responsible for parsing emails. Similarly, theHarvester uses regular expressions to collect also profile names, emails, URLs, Social Media accounts and so on. As we are going to see below, the results obtained using the LinkedIn module, is basically a search engine query that depends on google dorking and thus the results should be parsed in order to collect only the relevant information.

```
async def emails(self):
    await self.genericClean()
    # Local part is required, charset is flexible.
    # https://tools.ietf.org/html/rfc6531 (removed * and () as they provide FP mostly)
    reg_emails = re.compile(r'[a-zA-Z0-9.\-_+#~!$&\',;=:]+' + '@' + '[a-zA-Z0-9.-]*' + self.word.replace('www.', ''))
    self.temp = reg_emails.findall(self.results)
    emails = await self.unique()
    true_emails = {str(email)[1:].lower().strip() if len(str(email)) > 1 and str(email)[0] == '.'
                   else len(str(email)) > 1 and str(email).lower().strip() for email in emails}
    # if email starts with dot shift email string and make sure all emails are lowercase
    return true_emails
```

Figure 26: theHarvester example of its parser

LinkedIn

Let us demonstrate how the harvester utilizes public resources and how it achieves to get its results. Let us assume that we make use of the LinkedIn social networking application.

The command line of theHarvester is quite straightforward. One has to provide a domain name and the source to query that domain name. Executing the command in Figure 27 using our target's email address, it will look for any possible results in LinkedIn for josep.pegueroles@upc.edu:

```
┌──(kali㉿kali)-[~]
└─$ theHarvester -d josep.pegueroles@upc.edu -b Linkedin -l 500
```

Figure 27: theHarvester LinkedIn

Now let us further examine how the results are obtained by theHarvester in more detail. By looking closely at code provided in [27], the query of the python script discovery/linkedin-search.py is crafting the following:

http://'+self.server+'/search?num=100start='+str(self.counter)+'hl=enmeta=q=site

To begin with, the first variable "self.server" indicates the search engine to be used which in this particular case, predefined in the constructor, is google.com. Secondly the "self.counter" represents the number of results to be shown and it is incrementing based on the total number of results. Finally, the variable "self.word" is the provided domain which in this case is our target's email address. Executing the command given above, the final URL will eventually have the following form:

https://www.google.com/search?num=100start=0hl=enmeta=q=site

To verify that indeed the theHarvester executes this specifically crafted URL, we execute the URL in browser and the command in the terminal using theHarverster. Upon the URL execution, we can observe that this crafted URL triggers the google dork query in Figure 28.

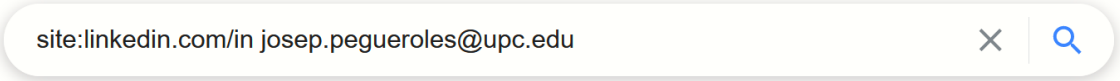


Figure 28: theHarvester LinkedIn google dork

The results in Figure 29 are an example of the google dork of the URL and they are including the following keywords: josep + pegueroles + upc + edu. These results are then filtered out, sorted and aggregated by the parser of the tool and they are presented as the output of theHarvester.

Figure 29: theHarvester google dorking results

From theHarvester's perspective the results from LinkedIn search look like the following Figure 30. Now it is obvious that theHarvester has some kind of parser since the results from the tool are much less than the results from the browser. This parser, as explained above, has as main goal to discard the results that have no correlation with the parameters passed.

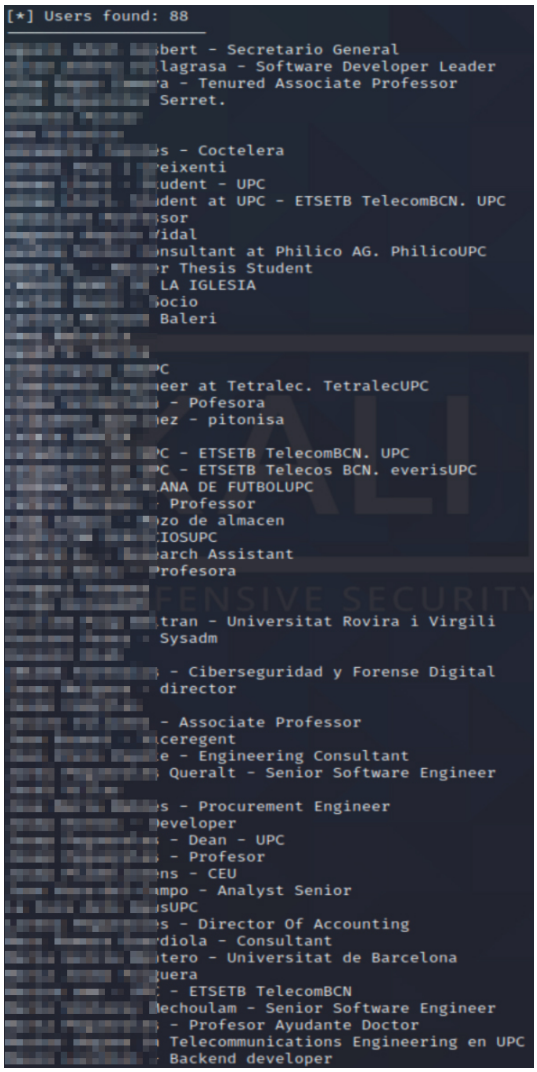


Figure 30: theHarvester LinkedIn results

Twitter

In a similar manner, querying for Twitter users, the Twitter module require no API to work. It utilizes again the power full results of google dorking as in the previous LinkedIn case. In Figure 31, a similar google dorking is used to query the domain email in Twitter using google search engine.

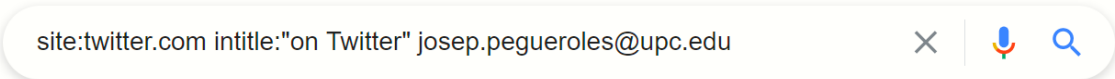


Figure 31: theHarvester Twitter dorking

To conclude, the results obtained in this section from the theHarvester are not requiring

the use of APIs but rather the tool tries to be as inclusive as possible by scrapping the web results from the search engines using its parser. The tool in general is quite powerful but when it comes to profiling theHarvester might not be the best option. The tool requires a domain to be able to provide you with IPs, emails and hosts and thus using only a name or username will not provide you in such detail results as the tools that were reviewed before.

Spiderfoot

Spiderfoot is yet another opensource and very powerful tool that supports a large variety of modules. The tool can be installed setup by simply following the steps at [28]. Spiderfoot can be either executed through the terminal by providing a simple one-line command or by using its web-based version which runs at localhost:5000 that helps visualize all step of the process your actions.

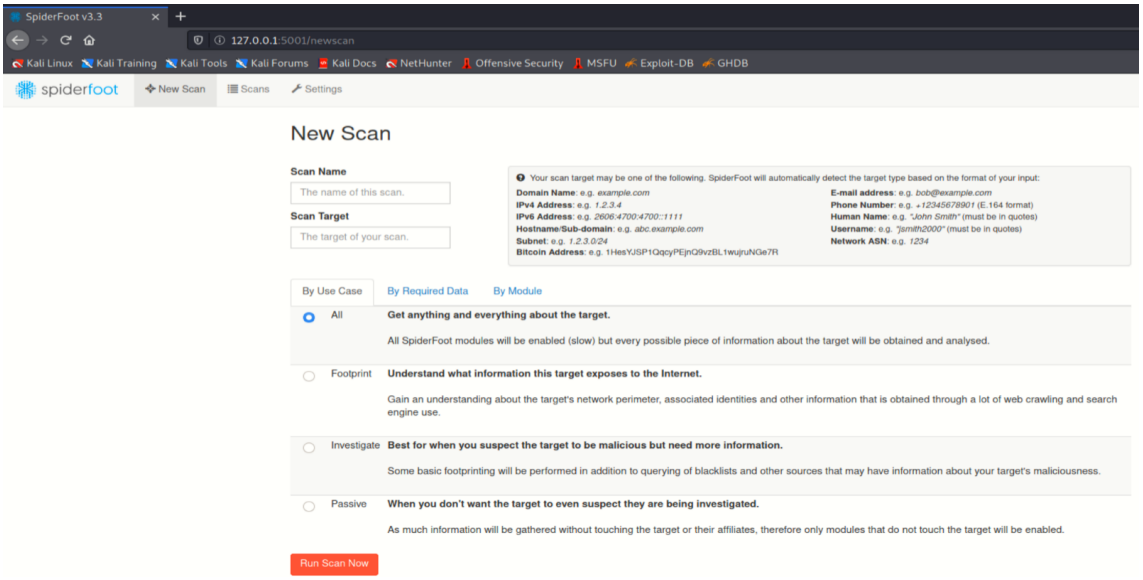


Figure 32: Spiderfoot main page

Spiderfoot is a very rich tool, in term of options, that offers not only a large variety of modules but also all the necessary categories based on the intensions of the research as depicted in main page of the tool in Figure 32. To begin with there are many different scenarios one can follow with Spiderfoot. For example, there 4 different Use Cases to choose from based on your intentions, either “Passive Scan”, “Investigative”, “Footprinting” or including all modules and all scenarios by choosing the “All option”. Additionally, we can see all the valid formats that are available as inputs for our target in Figure 33.

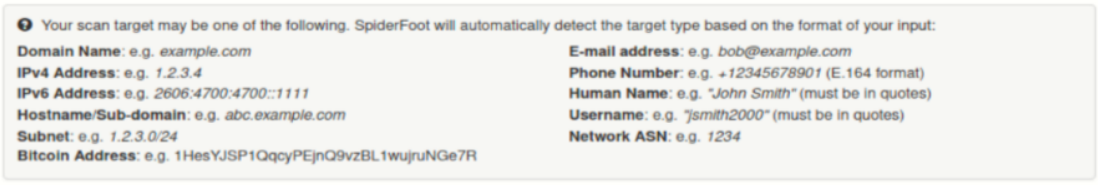


Figure 33: Spiderfoot valid input formats

Furthermore, one can also select specifically the modules to be used in the scan by navigating to the "By Module" section and deselecting the unnecessary modules as in the Figure 34. Moreover, some of the modules that have the small black lock on their right, as for example AbuseIPDB, indicate that they require an API key to work.

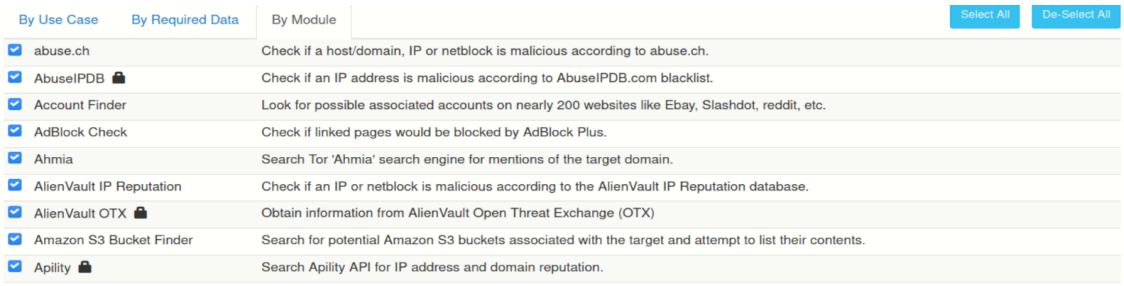


Figure 34: Spiderfoot "By Module" Section

Let us now run a simple test with our given modules from the list in Appendix **??**. In this test we are going to make use of our target's full name and email address. It is important to include only our predefined modules for this test since, after the test is completed, we will examine the source code of the modules used against the results obtained in order to determine how the tool is working internally.

Firstly, we provide the full name of our target which is Josep Pegueroles. Upon the completion of the test, we can find the summary of it in Figure 35.

Figure 35: Spiderfoot test summary

Looking into our results, by clicking on "Profiling", we can observe in Figure 36 all the different types of categories that Spiderfoot managed to retrieve results. For example, the type "Account on External Site" refers to web-app that have accounts with the username of our target.

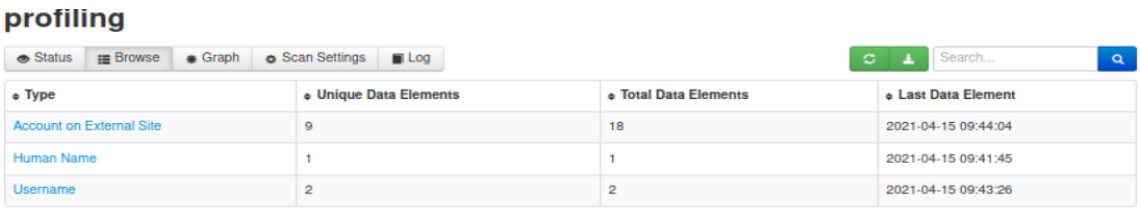


Figure 36: Spiderfoot profiling overview

In order for the tool to be able to mangle the given full name and query all these different modules it must somehow recognize the given input. The Spiderfoot recognizes that our input is a name, as explained above, and it tries to reform the give input into different formats. One of the python scripts used, named sfp_accounts.py [29], is called to reform the provided input in different ways so that it can be queried from the different sources.

As for example, by navigating to the "Username" section in Figure 34, we can observe how our input was manipulated to get its different variation. In Figure 37, under the "Data Element" section we can observe that our input was separate with a dot in the first line and concatenated in the second one.

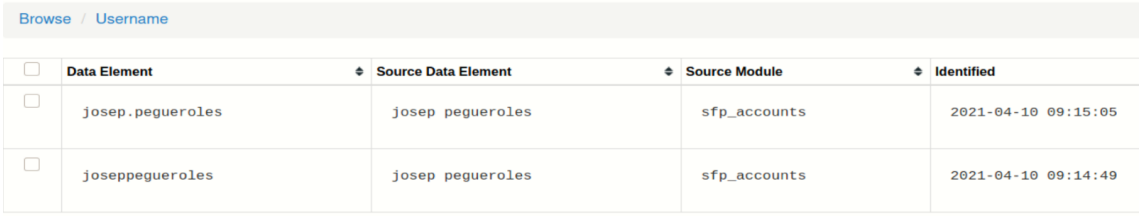


Figure 37: Spiderfoot profiling "Username"

The sfp_account.py script, under the "Source Modules" in Figure 38, is querying a json file containing a large number of social media URLs for the targeted user [25]. Now by further looking into the script we can verify indeed the manipulation of the name as depicted in the Figure 36.

```
246         if eventName == "HUMAN_NAME":
247             names = [eventData.lower().replace(" ", ""), eventData.lower().replace(" ", ".")]
248             for name in names:
249                 users.append(name)
```

Figure 38: Spiderfoot profiling input manipulation

In the results "Account on External Site", in Figure 39, we can observe all the web applications that our target's name was found as an existing account. Similarly, by looking into the "Source Module" section of all the results obtained we will see which sfp scripts have been loaded and used for each module and what kind of information each script is

responsible for. Once we have that information, we can examine further the scripts to determine how they work based and how we can benefit from them.

Data Element	Source Data Element	Source Module	Identified
Blogspot (Category: blog) http://joseppegueroles.blogspot.com	josep pegueroles	sfp_accounts	2021-04-15 09:42:16
Blogspot (Category: blog) http://joseppegueroles.blogspot.com	joseppegueroles	sfp_accounts	2021-04-15 09:42:48
Eyeem (Category: art) https://www.eyeem.com/u/josep.pegueroles	josep pegueroles	sfp_accounts	2021-04-15 09:43:26
Eyeem (Category: art) https://www.eyeem.com/u/josep.pegueroles	josep.pegueroles	sfp_accounts	2021-04-15 09:44:04
Facebook (Category: social) https://www.facebook.com/joseppegueroles	josep pegueroles	sfp_accounts	2021-04-15 09:42:16
Facebook (Category: social) https://www.facebook.com/joseppegueroles	joseppegueroles	sfp_accounts	2021-04-15 09:42:48
GitHub (Category: coding) https://github.com/joseppegueroles	josep pegueroles	sfp_accounts	2021-04-15 09:42:16
GitHub (Category: coding) https://github.com/joseppegueroles	joseppegueroles	sfp_accounts	2021-04-15 09:42:48
PinkBike (Category: hobby) https://www.pinkbike.com/u/josep.pegueroles/	josep pegueroles	sfp_accounts	2021-04-15 09:43:26
PinkBike (Category: hobby) https://www.pinkbike.com/u/josep.pegueroles/	josep.pegueroles	sfp_accounts	2021-04-15 09:44:04
Pinterest (Category: social) https://www.pinterest.com/joseppegueroles/	josep pegueroles	sfp_accounts	2021-04-15 09:42:16

Figure 39: Spiderfoot profiling results

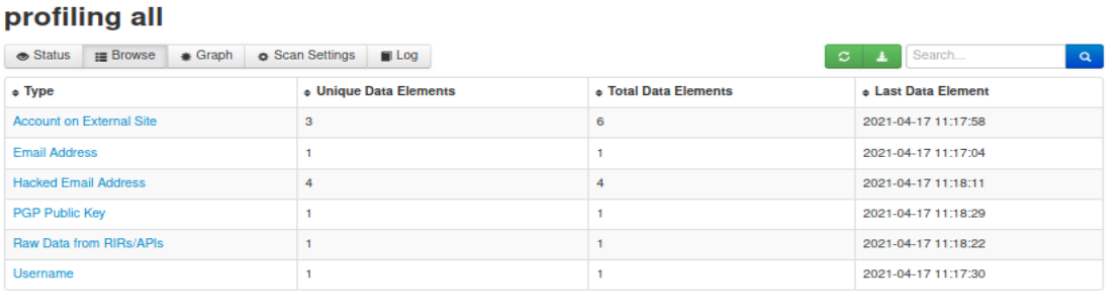
Finally, the same test has been performed for also the email addresses of our target. Equivalently, the process described above is also applied to all the given inputs like emails, usernames, etc and they can be determined by just looking into their sfp script. The total number of tests performed are depicted in Figure 40. Worth mentioning is the fact that using email addresses the results are less but more accurate and important in comparison with full name results. As in Figure 41, we can observe that we have more types of results than the ones from the full name test and more important information like “Hacked Email Address”, “PGP Public Key” and so on.

Scans

Filter: None

Name	Target	Started	Finished	Status	Elements	Action
profiling all	josep.pegueroles@gmail.com	2021-04-17 11:16:54	2021-04-17 11:18:32	FINISHED	14	
profiling	josep pegueroles	2021-04-15 09:41:43	2021-04-15 09:44:06	FINISHED	21	
profiling	josep.pegueroles@upc.edu	2021-04-15 09:39:22	2021-04-15 09:40:39	FINISHED	11	

Figure 40: Spiderfoot summary all tests



Type	Unique Data Elements	Total Data Elements	Last Data Element
Account on External Site	3	6	2021-04-17 11:17:58
Email Address	1	1	2021-04-17 11:17:04
Hacked Email Address	4	4	2021-04-17 11:18:11
PGP Public Key	1	1	2021-04-17 11:18:29
Raw Data from RIRs/APIs	1	1	2021-04-17 11:18:22
Username	1	1	2021-04-17 11:17:30

Figure 41: Spiderfoot profiling all overview

In conclusion, Spiderfoot is a very powerful tool when it's used with its full capabilities and it can be of a great importance for these kinds of investigations. The tool has very high potentials and of course it can be further expanded with our own modules and be modified based on our intentions.

5.2.4 Today's OSINT tools Limitation

As we have already saw in the previous sections the tools reviewed are unable to gather and realize context. The literature methodologies however, using ML techniques, are enabling the passivity of uncovering context from the gathered data. In today's internet social media images are the center of the attention and the most posted content. Tools like the ones above cannot infer information from the pictures uploaded. These tasks are tackled with ML techniques like the ones described in Section 3.

Deep profiling and Behavioural Analysis off course require more than OSINT tools can offer. To study an individual, we need to investigate much more information than can be collected by the tools. The tools provide an initial mapping of an individual that can reveal relationships, occupation, location, friend, relatives, colleagues and so on. In a forensic analysis for online SMN, this information might need to be further studied and analysed.

5.3 Methodology Recreation Analysis

In this section we are going to present and discuss the results of the digital profiling methodology presented in paper [21]. The main goal of recreating this methodology is to perform another digital profiling investigation without any OSINT tools but rather with other manual means. At the end it will enrich our understanding about the different approaches' that other researches have been using, the linking and correlation data points of SMNs and also help us conclude on more concrete counter measures.

Before we begin it is important to mention that these methods can be used in many different ways and with combination of other not mention methodologies. For example, if one has already a specific target list then the steps to be performed are different than in the case of having a random target. In general, searching for random targets would result in a vast amount of data and thus processing, analysing and classifying that data for future use will be extremely challenging and time consuming.

In this particular research we are going to apply the methods in both targeted individuals but also at random targets. The methodology we are recreating consist of 5 different approaches and for each one we analyse and present the results in the following sections.

5.3.1 Advanced Search Operator

Search engines are holding vast amount of information and they are updating their databases daily. Bad configurations and wrong access control mechanism can leak very sensitive data to the public internet. In this section we are taking advantage of advanced search operators which is also known as dorking. For privacy reason all the search queries were performed through the TOR browser using Google and DuckDuckGo.

Our approach is divided into 2 parts. The first part includes queries concerning a specific target and the second task queries for random targets. In table 3, the operators used for our dorking queries are presented. The parameters of each operator have been selected based on desired information of our targets.

Operators
Intext
Site
Filetype
Inurl
Inbody
Intitle

Table 3: Advanced Search Operator parameters

In the first part we assume that we already know the name of our target. Using a simple query with the full name of our target we can recover all the related information that the search engine returns. However, one might use different names and nicknames as usernames in the SMN. For that reason, we can also mangle our input in different ways, as depicted in Figure 42, and execute again the query. A simple query such as this one provided us with another important detail which is our target's username.

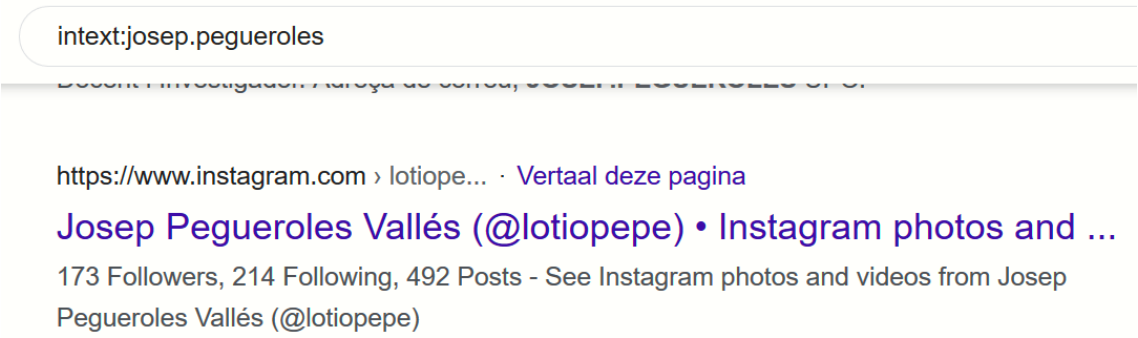


Figure 42: Dorking example intext

Now we can take this piece of information and reapply the search query used before. In this way we recycle our results to look deeper in our target. Performing again the same dork with the newly found information we can observe in Figure 43 that we uncovered new SMN for our target. These different types of personal information can be again mangled and used against the target.

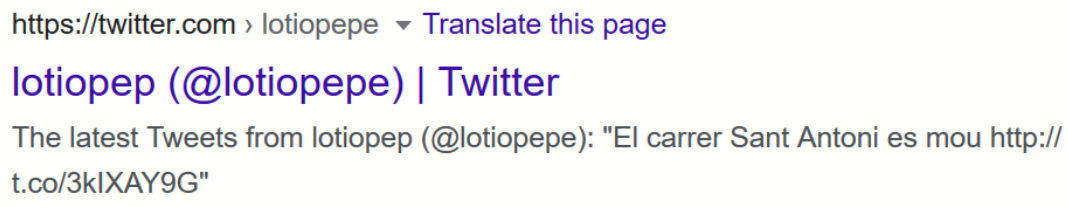


Figure 43: Dorking username results

Additionally, the above dorking does not only provide us with SMNs since '@' character is a tag widely used in SMNs for mentioning other users and thus the search engines will also return public results that this given nickname has been mentioned. From this we can conclude that mentions are a great point of references and they can also be used for profiling. In the Figure 44, we uncovered a source where our target has been mentioned. Here we can see that our target has been participated in some kind of book competition. The hashtags on the bottom provide us with important information such as the geolocation, the year this competition took place and it reveals another SMN account of our target. Such information is not only crucial for exploring our target's identity but it also helps us understand the hobbies and interest of our target which might be used later on as input in behavioural analysis.

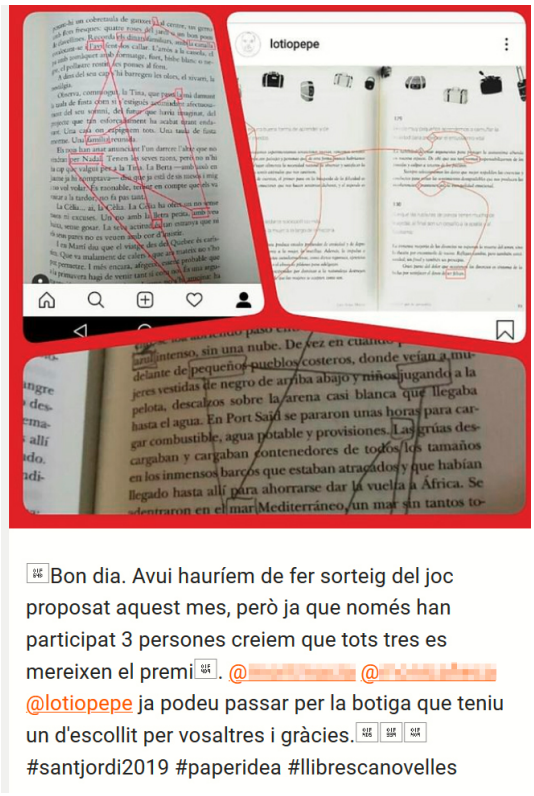


Figure 44: New SMN www.picuki.com/profile/xxxx

Moving on to the second part, we are now going to present the results collected from the random targets search queries. The scope of this part is to collect as much as possible sensitive details that are publicly available on the internet with the desired data to be email addresses. From the search we collected more than 10 file of csv type. Due to the time restriction of this research and the large amount of data to be analysed we limited our search results to only Gmail accounts. An example of a query that was used is depicted in Figure 45. This query is simply looking for any file with type xlsx that has in its text the keywords facebook.com and twitter.com and the parameter Gmail.

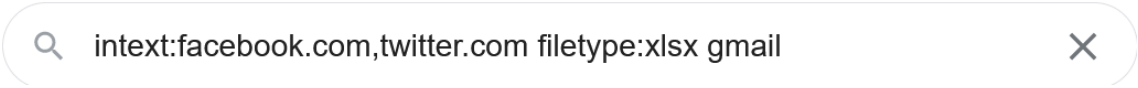


Figure 45: Example of an advnace query

Performing the above query with some alternations in file types and SMNs we managed to recover not only private accounts but also corporate and business related. With just a few csv files we manage to recover a few hundred emails. The full list of aggregated SMN accounts recovered can be found at Appendix 6. Even though it falls outside the scope of this research, alongside with emails we also recovered some very crucial information about

random people and businesses. The data available in some of the files were including all kinds of personal details as listed below:

- First and last names
- Phone numbers
- Personal and business emails
- Secondary emails
- Usernames
- Political parties
- Office addresses
- Hidden login web pages with passwords
- Social media accounts with passwords
- Multiple social media for 1 individual
- MySQL passwords
- Admin passwords for WordPress

From the aforementioned available data, we can observer how easy it is to uncover crucial details about random people. In cases where there is no professional consultation on the privacy matters of a company such leaked information can be a disastrous mistake. An example of such csv file is presented in Figure 46. Here we see not only business email addresses but also all the corresponding personal emails, phone numbers and social media profiles.

M	N	O	P
SRM Mail ID	Personal Mail ID	Contact Number	Social Profile
r	@gmail.com	669	https://www.facebook.com/
j			
s	s@gmail.com	0	
j	bnd@gmail.com	834	
s	riya@outlook.com	886	
l	dsl@gmail.com	400	https://www.facebook.com
r	gmail.com	225	https://www.facebook.com
			1.Facebook
j	s@gmail.com	652	https://www.facebook.com/
l	ee.bkh@gmail.com		
s	lec@gmail.com		
s	1@gmail.com	000	https://www.facebook.con
a			
r	w@gmail.com	466	https://www.facebook.com/
r	109@gmail.com		
a	@gmail.com	421	https://www.facebook.com
r	@gmail.com	354	
r	96@gmail.com	803	
			https://plus.google.com/u/0/
l	tud@gmail.com	265	https://www.facebook.con
g	@gmail.com	748	
l	gmail.com	371	
l	@gmail.com	623	
			Instagram:
			Facebook:
			Whatsapp:
l	@gmail.com	17	Gmail/google+: @gmail.com

Figure 46: Example of recovered CSV.

5.3.2 Social Aggregator

The next approach is rather simple and easy to perform. Social aggregators are websites that can provide details about an individual and its accounts in the various SMNs. An example of such website is About.me which is its essence another SMN that people can create a profile listing all of their means of contact. As depicted in Figure 47, using one of our random targets full name we recovered from its personal page; 5 SMNs that he is making use of.

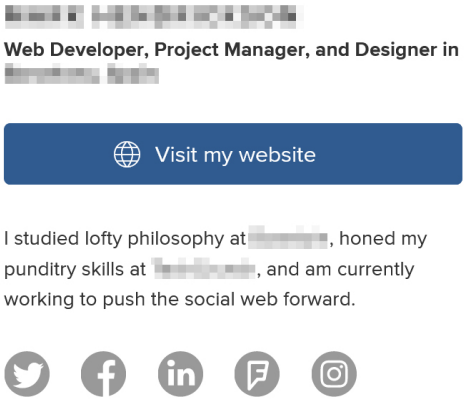


Figure 47: About.me target results

Similarly, we can take the results from the advance search operator and feed them into such websites and create a detail list of all the people and the SMN their using. Other similar websites are Webmii.com, Nameck, Checkusernames, Knowem and others that can be found listed in the OSINT framework in Section 2. An example of such website is depicted in Figure 48. In this case, we used the nickname found above for our individual target. Using multiple of these websites one can aggregate the results and manually of by using a script go over the real SMNs that our target owns.

Figure 48: Results of Knowem

5.3.3 Cross-Platform Sharing

The cross-platform sharing is another approach of linking identities between the different SMNs. Again, using the results from the advanced search operator we managed to retrieve some linkage between SMNs of a single user's account. Cross-platform sharing is basically posts shared from one SMN to another. In Figure 49, using a full name from our already collected data, we present a valid linkage between Twitter and Instagram. We were able to verify the linked identity since both of the SMNs were already provided in the previously recovered CVSs. From the Figure 49 we can observer 2 important details. The first one is the URL used to share the Instagram content to Twitter, which has the form of instagram.com/p/, and second the hashtags used that again provide us with the location of the post.

Figure 49: Twitter & Instagram link

5.3.4 Self-Disclosure

It is very common and at the same time important in the entertainment industry to share all of the SMNs that a user is active in. In this way, self-disclosure will most probably benefit the end user by providing its followers with different ways of communication. In a digital profiling investigation this step could be perform first in order to collect all the different usernames and SMNs for a specific target or last so that the identities and relationships found earlier can be verified.

In the following Figure 50, we provide an example of a Facebook page that discloses its website, Messenger, Instagram, YouTube, Twitter, Twitch and Snapchat. In a similar way, the aforementioned random targets can be fed as input to an automated script in order to uncover possible linked identities.

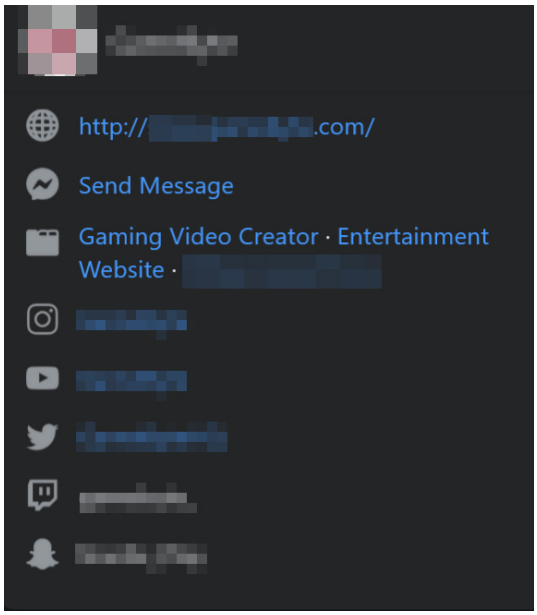


Figure 50: Self-Disclosure in Facebook

5.3.5 Friend Finder Feature

In the last method, we explore the friend finder feature of Facebook, Twitter and LinkedIn. This feature provides newly signed up users with suggest friends based on their email

contact list.

In this step there are a lot of different scenarios to consider. One can use all the collected information, such as full names, to form email addresses. These emails will be then added to the contact list of our email provider account. Finally, a new SMN account must be created so that the friend finder feature will be triggered. If that occurs then the SMN will possibly disclose any username linked with any of the accounts in our contact list. Moreover, this can be done in a more automated and faster way. As discussed before, by having a simple name we can try to reconstruct different email formats such as name.surname, name_surname, namesurname and so on. By using a simple API one can verify the existence of the accounts before moving on to the next step.

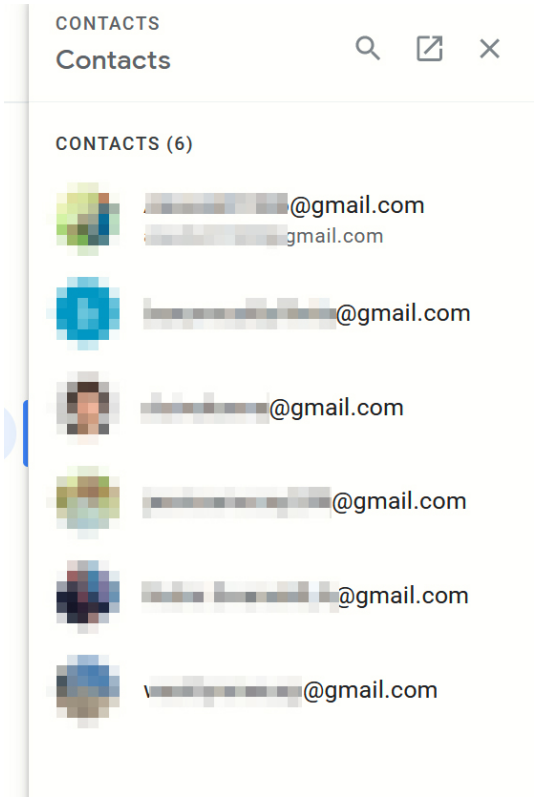


Figure 51: Gmail contact list

In our case, we utilized the already recovered email addresses from the advanced search operator step and we selected at random 10 email addresses. From these 10 accounts only 6 were discoverable in Gmail's contact list as its depicted in Figure 51. We then signed up in the 3 SMNs mentioned before and as a result only LinkedIn friend finder feature disclosed a user account as depicted in Figure 52.

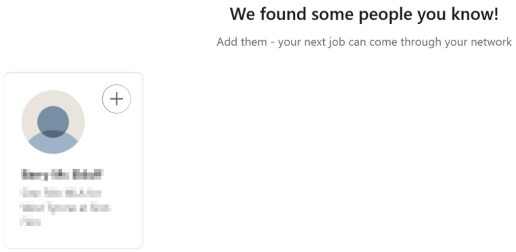


Figure 52: LinkedIn Friend Finder Feature

We used a very small number of accounts since our goal was not to linked accounts or uncover personal details about our targets but rather to explore the methodology and the process of uncovering identities through the friend finding feature. There are of course endless combinations of email addresses based on someone's personal details and thus such process will be time consuming and resource depended.

5.4 Counter measures

In this section we are going to present the counter measures against digital profiling based on the reviewed literature, OSINT tools and the methodology recreation. As explained before the measures can be consider as a privacy recommendations. These proposed ways do not prevent digital profiling but they rather play an informative role in the world of SMN. Moreover, the counter measures discussed below cannot be considered as preventing mechanism due to the erratic nature of publicly available information and the rapid change of techniques and technologies used for collection and analysis.

To begin with, is important to mention that from the reviewed methodologies used in all the different scenarios, in a very large extend, the personal attributes collected are overlapping. In addition, the methodologies of ML, NLP and Behavorial Analysis are for the most part unique in their nature since research objectives differ from one another and the technologies are advancing day by day.

As explained above since a lot of information are overlapping, we are going to present the aggregated version of the results from all the 3 methods reviewed namely OSNIT tools, literature review and methodology recreation.

5.4.1 OSINT tools, Literature & Recreation results

What has been observed a lot is that fact the researchers are trying as a first step to guess with different ways some of the basic attributes about their targets. There plenty of available scripts, also presented in the OSINT tools Section 2, that can be used to mangle and create different formats of emails addresses based on full names and interests. The most common option is to use the combination of the full name in some of the most default formats like a concatenation of johndoe@example.com or other formats such as john.doe@example.com, john_doe@example.com, jdoe@example.com and so on. Some examples of such scripts can be found in the section 5.2 of Recon-ng and Spiderfoot.

We assume that this would not be possible if users try to keep their real identity details to themselves. However, one can argue that in platforms such LinkedIn its inevitable to provide your full name. In this case, there are other mechanism in place to handle that. A good practice will be to utilize the privacy by design settings provided by the SMN and the second one is to avoid any correlation of your email address with any of your personal details. In that way scrapping the web for public information with personal details would become extremely difficult. Another important aspect is the reusability of usernames in different SMNs. For example, in Figure 41 and onwards we can observer that the user has provided the full name as also his username/nickname in a single platform which then helped us look for the presence of this username in other platforms as well.

Let us now look further into the results. Automation is probably the most important tool when it comes to Big Data. Technologies such a ML and NLP are key since data can be inputted and sensitive information can be automatically extracted. An advantage of the today's internet is that the data available is unstructured and it changes dynamically without any predicted patterns. This makes the job of such technologies even more difficult since they might become obsolete in a very short period of time. However, the results from the reviewed literature look promising and in some cases with high probability of success. The accepted input for such tools ranges from the simplest to the most complicated bit information. For example, images provide not only geolocation information but also relationship, interests and so on. Images however are hard to interpret automatically since nowadays the largest SMN companies have a policy of stripping EXIF data before the images are uploaded. Researchers have found alternative ways of doing this by utilizing mentions and mentioned features, as seen in Section 5.2 Twitter-Recon-ng, but also hashtags, geolocation tags, check-ins, regular post times and activity and so on. From such small pieces of data one can infer the time zone and even the exact location of the posts as also their prefer time of using SMNs. Furthermore, other identification elements such as cross platform sharing, event participation, graduations, schools and so on can provide important context and assist in creating relationships.

The above however is nothing compared to what can be done with the help of NLP and behavioural analysis. As we have seen in Section 3, there exist tools that can listen and extract speech from the uploaded videos or sound clips. In combination with text that one might find in comments and in the different forums a behavioural analysis can be performed as also different NLP algorithms in order to provide context to the collected data. Even more fascinating is the fact that is also possible to create relationships between 2 entities based on the analysed figure of speech and predict if these 2 entities might be the same person.

Finally, there are some situations that are out of the user's control. For example, the friend finding feature discussed in section 5.3.5, falls entirely under the company's security provisions. Inevitably the companies today give away some security for exchange of convenience. Similarly in search engines, even though is not the providers fault, the convenience of find everything instantly does include, unfortunately, the documents exposing personal details found in the Section 5.2.

6 Conclusions and future development

This research has as main goal to explore and discover digital profiling techniques and methodologies and analyse them with the intentions of providing an overview of the potential counter practices and measures. This might assist in making the process of aggregation and context derivation from public available data even more difficult.

There exist different methodologies of doing so and rather complicated ones. OSINT tools is one way of collecting data but due to some limitations, as described in Section 5.2.4, they can only be used in the early stages of profiling. These limitations have been addressed by many different research papers and with the help ML, NLP and behaviour analysis they manage to give context and create relationships in an automated way. This makes our task even more complicated since analysing such methods is time consuming and due to the fact that the researchers don't make their tools publicly available there is no way to study in–depth their inner process and inputs.

All in all, at the end of this paper we aggregate and present our recommended practices and measures against potential profiling. These methods are the results of analysing the data from OSINT tools, the reviewed literature and the methodology recreation. We are only analyzing what appears to be the attributes utilized the most and the reasoning behind their use.

This research has broad scope and thus collecting, analysing and commenting on all the counter measures against digital profiling is not feasible. With that been said, there are a lot of potential future developments that can be applied in this research. Firstly, a more thorough analysis must be applied on the collected CSVs from the advance search operator Section 5.3.1. A potential path would be to scrap all the email addresses, names, phone numbers, social media accounts and so on from the CVSs and reapply the methodology in a more automated way so that a larger number of results could be examined. In this way one would be able to further identify new ways of collecting and aggregating digital profiling attributes. Then based on the results, conduct a validity and accuracy test. Secondly, one can also measure the effectiveness of the different components of this research such as OSINT tools profiling and counter measure effectiveness. Furthermore, since OSINT tools and methodologies for digital profiling are relatively static - such as the scripts used - and the internet of today is vast and unstructured, one can also measure the effectiveness of the discussed methodologies on dynamic content on the web.

Another possible path is to develop a tool similar to a modern password manager. For example BitWarden is a password manager that can store for each entry-account, the URL for login, the name & username, the password and even custom fields. In our case, it would be a great contribution if a tool can be developed that would have a anti-profiling capabilities. This tool could assist the user to keep track of his different usernames,emails and passwords for each SMN and also, with an automated profiling script, assess and alert the user if his different SMN accounts can be correlated based on the content he posts.

Finally, digital profiling is a complicated matter that requires a lot of attention. There are always 2 sides of the story and the nature of this research is to fight against the malicious one.

References

[1] Wikipedia. Cambridge analytica. https://en.wikipedia.org/wiki/Cambridge_Analytica. Accessed: 2021-03.

[2] Bc Ondřej Zoder. Automated collection of open source intelligence.

[3] Osint framework. https://osintframework.com/. Accessed: 2021-03.

[4] Wayback machine. "https://web.archive.org/".

[5] Maltego. "https://www.maltego.com/". Accessed: 2021-03.

[6] Maltego blog. "https://www.maltego.com/blog/investigating-companies-with-opencorporates-and-maltego/". Accessed: 2021-03.

[7] What is shodan. "https://help.shodan.io/the-basics/what-is-shodan". Accessed: 2021-03.

[8] Lanmaster53. Recon-ng. "https://github.com/lanmaster53/recon-ng". Accessed: 2021-03.

[9] Laramies. theharvester. "https://github.com/laramies/theHarvester". Accessed: 2021-03.

[10] Alexandros Pappas. Google dorking. "https://www.exploit-db.com/ghdb/6758". Accessed: 2021-03.

[11] Exif tool. "https://exiftool.org/". Accessed: 2021-03.

[12] Wikipedia exif tool. "https://en.wikipedia.org/wiki/Exif#/media/File:DigiKam_EXIF_information_screenshot.png". Accessed: 2021-03.

[13] Geocreppy. "https://www.geocreepy.com/". Accessed: 2021-03.

[14] Spiderfoot. "https://www.spiderfoot.net/documentation". Accessed: 2021-03.

[15] Foca. "https://www.elevenpaths.com/innovation-labs/tools/foca". Accessed: 2021-03.

[16] Hector Pellet, Stavros Shiaeles, and Stavros Stavrou. Localising social network users and profiling their movement. *Computers & Security*, 81:49–57, 2019.

[17] António Magalhães and João Paulo Magalhães. Textractor: An osint tool to extract and analyse audio/video content. In José Machado, Filomena Soares, and Germano Veiga, editors, *Innovation, Engineering and Entrepreneurship*, pages 3–9, Cham, 2019. Springer International Publishing.

[18] Ahmet Anıl Müngen, Esra Gündoğan, and Mehmet Kaya. Identifying multiple social network accounts belonging to the same users. *Social Network Analysis and Mining*, 11(1):1–19, 2021.

[19] Joost Hendricksen. Profiler:deriving a digital profile from open source information.

[20] Perez Charles, Birregah Babiga, Layton Robert, Lemercier Marc, and Watters Paul. Replot: Retrieving profile links on twitter for malicious campaign discovery. *AI Communications*, 29(1):107–122, 2016.

[21] Hussein Hazimeh, Elena Mugellini, Omar Abou Khaled, and Philippe Cudré-Mauroux. Socialmatching++: A novel approach for interlinking user profiles on social networks. In *PROFILES@ ISWC*, 2017.

[22] Miltiadis Kandias, Vasilis Stavrou, Nick Bozovic, and Dimitris Gritzalis. Proactive insider threat detection through social media: The youtube case. In *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society*, pages 261–266, 2013.

[23] Rishabh Kaushal, Vasundhara Ghose, and Ponnurangam Kumaraguru. Methods for user profiling across social networks. In *2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/-SocialCom/SustainCom)*, pages 1572–1579. IEEE, 2019.

[24] Recon-ng marketplace. `"https://github.com/lanmaster53/recon-ng-marketplace"`. Accessed: 2021-03.

[25] Web accounts list. `"https://raw.githubusercontent.com/WebBreacher/WhatsMyName/master/web_accounts_list.json"`. Accessed: 2021-03.

[26] theharvester parser. `"https://github.com/laramies/theHarvester/blob/master/theHarvester/parsers/myparser.py"`. Accessed: 2021-03.

[27] theharvester linkedin search script. `"https://github.com/laramies/theHarvester/blob/master/theHarvester/discovery/linkedinsearch.py"`, May 2021.

[28] Spiderfoot github repo. `"https://github.com/smicallef/spiderfoot"`. Accessed: 2021-03.

[29] Spiderfoot github sfp script. `"https://github.com/smicallef/spiderfoot/blob/master/modules/sfp_accounts.py"`. Accessed: 2021-03.

Appendices

Maltego - Data Source	Category
ATT&CK – MISP	DB
FullContact	DB
Clearbit	DB
Orbis - Bureau Van Dijk	DB
Pipl	DB
TinEye	DB
Wayback Machine	DB
Flashpoint	DB
Have I Been Pwned?	DB
PeopleMon	DB
PhoneSearch	DB
Social Links CE	DB
SocialNet	DB
ZeroFOX	DB
PGP	DB

Recon-ng - Data Source	Category
recon/contacts-contacts/abc	DB
recon/contacts-contacts/mailtester	DB
recon/contacts-contacts/mangle	Local Script
recon/contacts-contacts/unmangle	Local Script
recon/contacts-profiles/fullcontact	DB
recon/locations-pushpins/flickr	Social Media
recon/profiles-profiles/twitter$_m$*entioned*	Social Media
recon/profiles-profiles/twitter$_m$*entions*	Social Media
recon/profiles-contacts/bing$_l$*inkedin*$_c$*ontacts*	Social Media
recon/profiles-contacts/dev$_d$*iver*	Scripts Crawlers
recon/profiles-profiles/namechk	Scripts Crawlers
recon/profiles-profiles/profiler	Scripts Crawlers
recon/domains-credentials/pwnedlist/account$_c$*reds*	DB

theHarvester - Data Source	Category
Baidu	Search Engine
Bing	Search Engine
Exalead	Search Engine
Google	Search Engine
Hunter	Script Crawler
Intelx	Search Engine
Linkedin	Social Media
linkedin$_{l}inks$	Social Media
Qwant	Search Engine
Trello (Uses Google search.)	Social Media
Twitter	Social Media
yahoo	Search Engine
github-code	Social Media

Spiderfoot - Data Source	Category
Account Finder	DB
Apple iTunes	Social Media
Bing	Search Engine
clearbit	DB
DuckDuckGo	Search Engine
EmailCrawlr	Scripts Crawlers
EmailRep	DB
Flickr	Social Media
FullContact	DB
Github	Social Media
Google	Search Engine
HaveIBeenPwned	DB
Hunter.io	DB
Instagram	Social Media
IntelligenceX	Search Engine
Keybase	DB
MySpace	Social Media
PGP Key Servers	DB
Psbdmp	Script
Scylla	DB
Skymem	DB
SlideShare	Social Media
Social Media Profile Finder	Scripts Crawlers
Social Network Identifier	Scripts Crawlers
Twitter	Social Media
Venmo	

Research Keywords

- OSINT algorithms for profiling
- OSINT digital profiling
- OSINT digital footprint
- OSINT profiling online social networks
- Forensics profiling.
- OSINT counter measures
- OSINT digital Identities

Aggregated SMN

- The aggregated social media list from all the sources xlsx:
- Facebook
- Instagram
- Twitter
- Tumblr
- Flickr
- YouTube
- Google+ (shut down 2019 for social media-personal use-only workspace company corporations)
- Pinterest
- LinkedIn
- Other user owned websites

Glossary

- A list of all acronyms and what they stand for.
- SMN = Social Media Networks
- API = Application Programmable Interface
- Data Source = Any kind of resource that a tool is using to collect information
- DB = Database
- TOR = The Onion Router
- ML= Machine Learning
- NLP = Natural Language Processing