Open Access Library Journal
2024, Volume 11, e11463
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Information Reconnaissance by Accumulating Public Information Data Sources

Matthew Duffy, Xueying Pan, Samuel Wilson

School of Engineering and Computer Science, Oakland University, Rochester, Michigan, USA
Email: ivypan89@gmail.com

How to cite this paper: Duffy, M., Pan, X.Y. and Wilson, S. (2024) Information Reconnaissance by Accumulating Public Information Data Sources. *Open Access Library Journal*, **11**: e11463.
https://doi.org/10.4236/oalib.1111463

Received: March 20, 2024
Accepted: April 27, 2024
Published: April 30, 2024

Abstract

In the Internet age, the most valuable pieces of data when investigating an individual are phone numbers, email addresses, and usernames. These data points, which are typically freely shared by their owners, can act as a primary key to link research on a person to other data sources dispersed throughout the web. The Internet has made obtaining a wealth of data an accessible activity requiring only creativity and persistence. An investigator with the motivation to identify information on a certain person or organization can use these data points to build links and present useful information on a target. Our paper aims to study connecting disparate data from publicly accessible sources to provide detail into someone or something is called Open Source Intelligence (OSINT). In this paper, we introduce the subject of OSINT both in a broad sense as well as through documenting hands-on experience in scraping information on one of the members of this research group as well as generic internet targets. The first focus is on presenting the overall concept including norms and techniques that may be used to investigate an individual. The second focus is on deep-diving a specific tool along with a method for adding data sources to present the broad scope of available information. Several useful tools that are common in the OSINT space are presented. Finally, we found the challenge in narrowing the focus to the intended target and in relating it to useful information.

Subject Areas

Computer and Network Security

Keywords

Open Source Intelligence, Maltego, Tor, Google Dorks, Metagoofil, theHarvester

1. Introduction

As the world grows increasingly dependent on technology, much of our lives have moved into the public domain. This is true not only in areas where we freely share personal information such as in Social Media, and intentional information sharing such as websites about our businesses, but also in ways owners of the data may not intend. For instance, every real estate transaction is a matter of public record that is available whether the parties desire it or not. This is true of many other records such as internet property ownership (domain names), metadata tied to phone numbers, historical address lists, etc. Due to this fact, investigators are using the internet to profile individuals and businesses using public information as a source. The field of study for connecting disparate data from public sources to provide detail about someone or something is called Open Source Intelligence (OSINT).

Former researchers may have faced different challenges and situations based on the scope of their investigations and the specific context in which they operate. Limitation of the sheer volume of data available and the need to sort through a ton of data tremendous locate relevant information may make them hardly obtain access to reliable and relevant information. It could be difficult for them to get the accuracy and reliability of open-source data when information may be outdated, intentionally misleading, or incomplete because they must use strict verification procedures to ensure their findings are reliable. Research efforts involving OSINT may be limited in breadth and efficacy due to a lack of finance, time, or specialized skills. To optimize the impact of their work, former researchers need to strategically allocate resources and prioritize their efforts.

1.1. Data Classifications for Intelligence

Open Source Intelligence or public information reconnaissance is the process of using publicly available information to conduct investigations. Research within this domain falls into two categories [1]:

OSINT - Review of publicly accessible information which is factual and legally accessible. This category does not include documents that have leaked from sites like wikileaks or data that has been accessed through illegal methods even if the researcher did not perform the illegal information gathering activity themselves.

NOSINT or Non-Open Source Intelligence - Review of information which also includes secret data which may be discovered via access to government intelligence or through "hacking" into data which should not be accessible.

Government intelligence agents and hackers typically engage in NOSINT investigations. Only information that is publicly available and legally obtained is included in OSINT Investigations. The United States Department of Defense (DOD) [2], Federal Bureau of Investigation (FBI) [1] and North Atlantic Treaty Organization (NATO) [3] all have official documents that define and create standards for how OSINT Investigations are to be handled by government agencies.

Examples of different sources of public information are SEC filings, publicly-facing social media data, court documents, and other government documents. Many of these documents and other records with public information are now easily accessed on the World Wide Web. Tools are created to aid in the gathering phase of these documents as the web connects massive amounts of data that can easily be accessed by accessing various sources with a properly crafted search query.

1.2. Web Layer Targets for OSINT

Three different layers of the internet are available for review. These are:

- Public Web
- Deep Web
- Dark Web

Public OSINT investigations take place at all layers of the internet. They can include a combination of data from any of the three layers of the internet. The surface web layer is made up of information that is accessible without any need for login, authentication, or a webpage that is not hidden behind a web form that needs some form of user interaction to gain access via a portal [1]. This can be thought of as any document or website that a web crawler can access by knowledge of the location alone.

Deep web OSINT describes investigations that examine information that is hidden behind a login page or web form. It is believed that 80% to 90% of the internet is made up of deep web content [4]. An example of these searches is investigations that use social media information such as Facebook or Instagram. A social media investigation is the most widely used when investigating an individual or business [4].

Dark web OSInt describes any investigations that use resources that exist on the dark web. The dark web is the smallest section of the web and it is not easily accessed since special tools and web portals are needed to access it. Much of the content on the dark web is linked to illegal activities such as human trafficking, narcotics, money laundering, and hacking. Investigators may also use the dark web to hide their tracks when conducting investigations. The dark web makes up the smallest area of the internet and it is best described as a decentralized, anonymous area with special procedures that need to be followed to access pages such as limited hours of availability per day, use of special web browsers, and non-human readable URLs [1].

1.3. Types of Data Gathering Methods Used in Investigations

OSINT can use passive, active, or semi-passive methods to acquire data: (**Figure 1**)

Active OSINT directly communicates with a target such as pulling information from servers in control of the owning organization, scraping network information off of an authoritative DNS server, or actively scanning a target for vulnerabilities. These methods of data gathering are not recommended as they

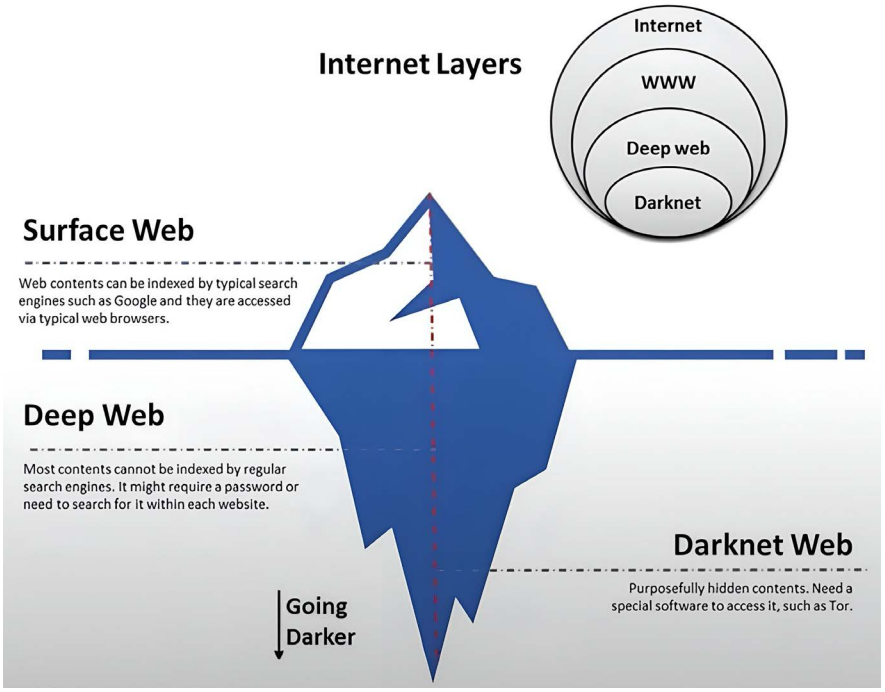

Figure 1. Layers of the internet [1].

are most likely to be detected by intrusion detection/intrusion prevention systems as they typically cause suspicious traffic alerts [1].

Passive OSINT on the other hand does not directly communicate with resources to gather data and instead relies on obtaining information from secondary sources such as databases and search engines which already have indexes or a copy of the desired information. Passive collection is the most common method used in OSINT investigations and is also considered a best practice to remain anonymous [1]. Limited information is gathered with this method due to a lack of communication between the investigator and the target.

Semi-passive OSINT is a hybrid method that combines light communication to targeted servers that masquerade as regular internet traffic to gather data on the target while remaining anonymous in the process [1]. One drawback to this method is that despite being anonymous, the user could still alert the target party that an investigation of some sort is occurring.

1.4. Users of OSINT

A wide variety of individuals conduct OSINT investigations daily. Law enforcement officers and government agencies are most commonly associated with OSINT investigations. Many other individuals also perform investigations in their course of business such as lawyers and paralegals for case research. HR teams are now using tools to screen new hires via simple investigations into prospective candidates. Some businesses also use teams of analysts to monitor their brand and conduct research into potential issues or image challenges ahead of major public reaction [1].

Cybercriminals and stalkers also perform investigations to find potential targets for their crimes. Many different investigations can be performed by private citizens like monitoring the publication of their name, reverse image searches of pictures to see if they have been used without authorization, and researching contact information for a potential private sale (e.g. ad for a car or house). Journalists and activists also perform OSINT investigations to research trends related to their work.

2. Recommended Osint Practices

One of the most common things individuals who conduct OSINT investigations need to do is take steps to ensure they remain anonymous as they gather information. If an investigator fails to remain anonymous during an investigation it could lead to the target individual discovering they are under investigation which may trigger deletion or purging of valuable information needed by the investigator. Being discovered could also open the investigator up for attack or investigation from the other party. Individuals who conduct investigations have tools and procedures to stay anonymous during their data-gathering phase.

2.1. Privacy Protection Services

An investigator should always start by using a VPN internet connection when browsing the internet during an investigation [1]. Once on the VPN, an investigator can further add to their anonymity by using a tool like The Onion Router (Tor) or the Invisible Internet Project (I2P) to access the surface web using the dark web. It is also recommended to use a new virtual machine for each new investigation to ensure previous investigations or activities have not compromised the user or investigator before the investigation begins.

2.2. Social Media Investigation

Another best practice to follow during an OSINT investigation is to never use a social media account that is linked directly to the investigator. When using a social media platform to find data for an investigation, it is highly recommended to never use the personal account of the investigator. First, the activity could very easily be linked to the investigating party which would alert the target to the investigation. Second, the investigator may violate the terms of service of the social media platform which may lead to alerting the social media platform and causing the investigator's account to be blocked or banned from the platform. To access the wealth of knowledge on a subject provided by the social media platform an investigator is recommended to create a dummy account with an email or phone number that does not link to them for use in the investigation [1].

Many social media platforms consider both dummy accounts and using web crawlers to scan and copy information from the platform to violate terms of service. Such activities are likely to get the user banned from the platform and even

potentially litigated by the owning company [1]. It is important to note that violating the terms of service of a social media platform is not the same as breaking the law. All data found on social media that is publicly viewable and which was not accessed through unauthorized means to avoid privacy settings is considered appropriate OSINT data.

3. Foundational Tools in an Osint Investigation

The next set of tools that are suggested for serious OSINT investigations are two different operating systems and a web browser that either has features built in to aid OSINT investigations or comes preconfigured with settings and behaviors to maximize privacy. These tools are considered foundational because they are not directly used for investigating, but instead, they have been preconfigured at the system or software level to optimize a setup that makes an investigation easier.

Both Tails OS and Kali Linux come with applications and configurations that minimize the need for an investigator to spend time setting up a special operating system for OSINT investigations. This is especially helpful if an investigator chooses to follow the recommendation of using a new virtual machine for each investigation to avoid detection caused by previous behavior. The web browser Tor helps the investigator to stay anonymous when browsing online. Using I2P is another method to achieve anonymous status. In the following sections, focus is given primarily to Tor as it is the more popular method for anonymous web browsing between the two.

3.1. Tails OS

Tails OS is a privacy-hardened operating system that is ideal for investigations that rely on remaining anonymous. This Debian Linux-based operating system is purpose-built as a portable operating system to avoid censorship, surveillance, and viruses. The OS is meant to be booted from a USB flash drive. An investigator simply has to insert the USB drive and they have their operating system they can use anywhere on any device.

One key to the OS is a feature called "amnesia" which ensures a fresh instance of the OS is started on each boot and all information from previous sessions is completely erased. The OS also has Tor with Ublock for anonymous web browsing as well as a list of other preinstalled filters for maximum privacy. Tails also has some features built directly into the OS such as blocking all connections which attempt to connect without Tor [5]. Additionally, all persistent storage is encrypted on the USB boot device and the OS does not write anything to the hard drive. All user data is contained within system memory which causes a full wipe on shut down.

One con to this operating system is that it is not recommended to make configuration changes as this may risk privacy hardening built into the operating system. It is possible to save and configure files on a limited encrypted persistent portion of the drive. According to NSA whistleblower Edward Snowden, Tails

and Tor are highly recommended tools for maintaining anonymity during research operations [5].

3.2. Tor Web Browser

The Onion Router (Tor) is a tool recommended for all serious OSINT investigations as it can be installed on a host of operating systems including Windows, OS X, Linux, and Android for privacy and anonymous web browsing. Tor uses a maze of encrypted connections between each client and server to create an anonymous channel of communications between the two parties. Tor functions by using a concept called perfect forward secrecy [6]. Tor achieves this by allowing anonymous users to host relay nodes which are web servers that route traffic on the Tor network. Tor then uses these relays to anonymously route encrypted communication between entry and exit nodes [1].

First, the browser uses a Tor directory server to create a circuit made up of at least three different relay nodes between the user and their destination. The more relay nodes are used, the less likely the source traffic can be traced. However, the tradeoff is the more relay nodes are used, the slower the communication becomes between the user and the destination. After a circuit is chosen, Tor encrypts the request into a virtual envelope that only the exit node has the key to open. The virtual envelope is nested in a series of envelopes that contain the destination information of the next node which can be opened by keys held by the next node(s) which the envelope is destined for. Each node only knows the previous source node and the next destination nodes that it forwards the next virtual envelope to [6]. Both the exit and entry nodes are the most vulnerable because these nodes are aware of the address of either the user for the entry node or the destination for the exit node.

3.3. Tor Vulnerabilities and Dark Web Investigations

Part of the reason the Tails OS exists is that despite Tor being a useful tool for remaining anonymous on the web, there are still a few attacks that the browser is vulnerable to which could expose the identity of the client or host in the event the attack is successful. User traffic analysis, passive traffic analysis, circuit reconstruction, and circuit shaping are all methods or attacks used to discover the identity of either the user or owner of a request using the Tor browser [6] (**Figure 2**).

User traffic analysis involves performing an investigation into the anonymous traffic in the hopes that the user or website owner makes a mistake that somehow unintentionally reveals themselves through a non-protected connection such as opening a downloaded PDF in Acrobat which makes a call through the unprotected web browser instead of using the Tor browser to open the document.

Passive traffic analysis is a different investigation method that looks to reveal the identity of an individual who is using the Tor network by comparing some

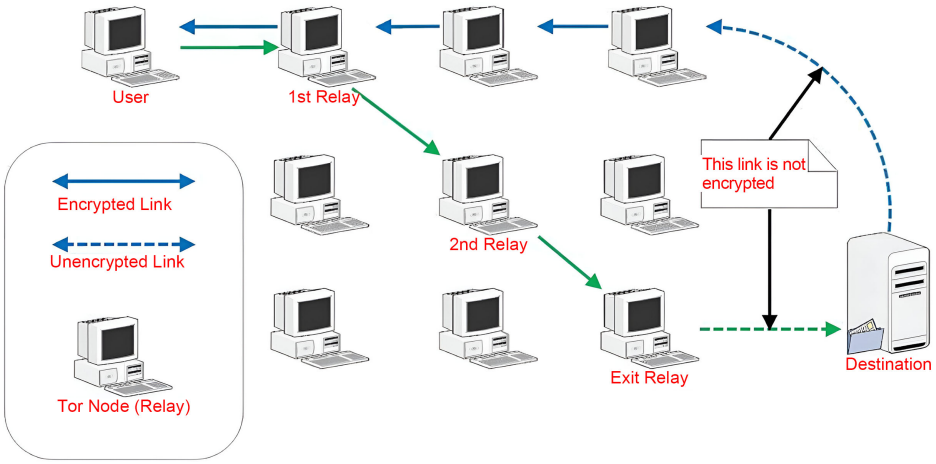

Figure 2. Logical Layout of the Tor Network [1].

form of signature of the anonymous traffic with a signature from regular internet traffic that reveals the identity of the individual through matching patterns in network, hardware, software, or operating system behavior [6]. One example of a signature could be made up of a unique combination of operating system, patch level, and hardware in both traffic examples. This type of attack can be partially mitigated by using the most up-to-date version of the Tor software.

Circuit reconstruction is a complex attack that involves an attacker loading compromised nodes into the network which can then eavesdrop on traffic at the node level [7]. When a victim's traffic is routed through this compromised node, this anonymity is broken. Perfect forward secrecy has redundancy built in to make this type of attack highly unlikely [6]. The more nodes in a circuit the less likely all the nodes are compromised by a single attacker.

Circuit shaping is an example of an attack that requires an attacker to place a tampered version of Tor on a user's system that sends all traffic through compromised nodes accessible to the attacker [6]. Strong security practices are the best protection against this type of attack by protecting the system from malware and verifying the signature of the Tor Software before it is installed. Investigators should verify the signature included with the user's Tor browser with the signature found on the Tor website to ensure the browser has not been tampered with [6]. When it comes to Tor, the two dark web investigation techniques and complex attacks listed above, both have a very low likelihood of being successful but the user needs to be very careful not to reveal themselves outside of Tor. This is important because user error is the easiest way for a user to reveal themselves online.

3.4. Kali Linux

Another operating system that is popular with OSINT investigators is Kali Linux. Kali comes preinstalled with tools and software that can be used for different aspects of OSInt investigations. Tools like Maltego, Nmap, theHarvester and Metagoofil come preinstalled in the environment [8]. Some of these tools

will be discussed later in the paper as they are useful for finding information on individuals. Unlike Tails, Kali Linux does not come preinstalled with Tor so the investigator must download and set up the Tor browser themselves. Kail Linux also does not come preconfigured with the same level of privacy hardening as Tails does so the investigator may need to make adjustments to the system and software settings to maximize user privacy on the operating system.

Tails OS and Tor Browser mainly focus on providing anonymity and privacy for online activities, while Kali Linux is a specialized operating system tailored for cyber security and ethical hackers' purposes. Tails OS could route all internet traffic from the Tor network by default and provide users with anonymous browsing capabilities because it includes pre-installed privacy and security tools such as Tor browser when a high level of anonymity and privacy is needed for whistleblowing, journalism, and activism. When individuals and organizations are engaged in testing cyber security and researching the security of networks, applications, and operating systems, Kali Linux would be an essential tool for them.

4. Maltego Osint Tool and Custom Development

As discussed in this paper, data availability within the OSINT space is typically not the challenge. Rather, turning that data into useful information by building meaningful relationships in any approach other than manual analysis is the primary issue. One tool that aims to ease this is the software suite Maltego, developed by the company Paterva. This software presents a dashboard for many of the open-source data providers that will be reviewed elsewhere in this paper. The differentiator for Maltego, however, is the presentation of data through a graph-database-like relational model with relationships between defined data entity types [9].

4.1. Transforms

When using the tool, several data entity types can be used as a seed item into the tool's data gathering methods called "Transforms". Transforms attempt to apply logic to the type of additional data that is requested for the currently selected seed entity. For example, within the tool, a few base data entity types are "Domain" and "Email Address". On an element of Domain entity type, transforms include network level lookups such as "To DNS Name" and "To IP Address". Whereas the Email Address entity has transformed such as "To Flickr Account" and "To Person [PGP]" to pull from public sources of registration information. The tool builds level-wise relationship data and allows a user to keep performing additional transforms on data revealed by a former transform which allows an investigator to build a graph of connected information from the starting seed element [10].

Transforms come in three varieties:

1) Built-ins provided by default in the installation

2) Installable transforms available in the Maltego Trans-form Hub within the

tool

3) Custom installable transforms that can be developed by the customer for internal use or sharing at the discretion of the creator

Within the second category of transforms, those that are installable, many of them require a purchase of the Maltego annual subscription plan from Paterva at C999 per year [11] to make them available. Many of the transforms also require an API key from the data source provider who may themselves also charge a subscription fee. Some valuable examples of this include a service by Social Links called Social Links PRO which claims to allow social media identification by individuals (including the use of face recognition), geography, and content within the Darknet [12]. While the information is public, the organizations that present this information in an organized fashion for Maltego users to consume monetize the convenience and time-saving potential of their services.

Some of the additional installables in the second category do provide valuable information while being free. One such example is a transform put together by the website HaveIBeenPwned. HaveIBeenPwned is a search engine that allows individuals to search their email addresses to see if their account has been a victim of known data breaches to date [13]. This functionality requires the user to create an Email Entity of the desired address upon which transforms can be run to return known data breaches the account has been a victim of with password exposure likely. Use of this data source within Maltego is much like the use of HaveIBeenPwned's own online service but with the advantage of a clearer relationship of hacked data and the ability to view multiple email address searches in a single view.

4.2. Identifying Additional Transforms

Maltego also offers the ability for users to introduce their own transforms through custom development which comprises the third type of transforms. This functionality is available in the free version of the software. Creating custom transforms requires an internet-connected endpoint that can run Paterva's Maltego TRX Python-based web service. Essentially, this endpoint serves as a target for the desktop client to send XML-based requests which are then processed against any data source that is Python accessible; such as scraping data from a Linux application, requesting data against any API that is not already in Maltego's Transform Hub scope, or pulling information from an organization's internal database. Once the resulting information is obtained, it can be wrapped back into Maltego's XML format within Python and passed back to the desktop client.

With the primary focus in this project on revealing information about a person (as opposed to a general network infrastructure investigation); a data source was sought that would exemplify a view of public information that may be thought of as sensitive to those not familiar with the field. With some research, a service that makes information from the Multiple Listing Service (MLS) database available was identified. The MLS provides information on real estate listings,

sales transactions, parcel data, and valuations [14]. This service is the Zillow API which is exposed via a subsidiary company known as Bridge Interactive. On contacting Bridge, a justification was required to access their API. The justification of data analysis for a graduate school research project was sufficient to provide full access to the platform. This API is available as a no-cost offering.

The Bridge Interactive site offers documentation on the use of Zillow endpoints which includes data collection in multiple areas including [15]:

1) Parcels - Base information about properties regarding assessments and transactions

2) Transactions - Current and past transactions including sales, building of properties, and title transfers

3) Zestimates - A proprietary algorithm for current property values (similar to a property appraisal)

4.3. Adding Custom Transforms

To enable useful data in the Maltego desktop client, several steps are required. First, while Maltego comes with a built-in Domain Entity and a "whois" transform for looking up domain ownership information such as listed phone number, email address, and base address information; the ability to look up full owner address is not provided. For this reason, custom development was required to look at a full address to then build further lookups to utilize the Zillow API.

For both use cases, a custom Python endpoint must be written to accept data from Maltego, unwrap it perform further lookups, and provide it back to the client. For purposes of this project, four transforms were written:

1) Whois lookup transform to pull full address from a domain name whois source

2) Zillow public records service for mortgage holder on the last transaction record for the address (if applicable)

3) Zillow public records service for balance of mortgage on the last transaction record for the address (if applicable)

4) Zillow Zestimate service for approximate current property value of the entered address

Creation of the custom listener was eased greatly by the Paterva-provided Maltego TRX Python library which abstracts much of the client knowledge requirement by automatically processing client XML on both incoming requests and outgoing response [16]. Thus, the focus of development can be on collecting data against whatever external data source is required. A logical diagram of the process for a request from the user to install these created custom transforms and utilize them along with the API is shown in **Figure 3**.

The resulting code required to enable this transform is provided in the code callout below. Python libraries must be imported to enable web requests and parse JSON within a transform class. The response from the Zillow API is provided in a JSON object which must be deserialized and passed back to the requesting client.

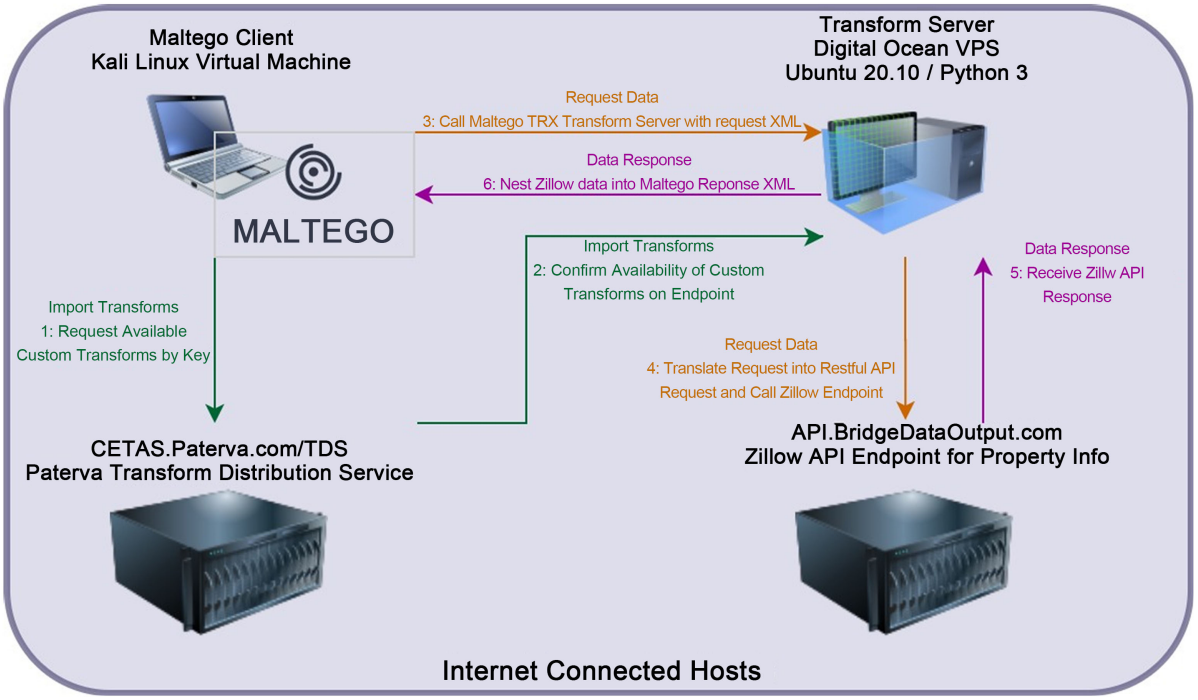

Figure 3. Logical diagram for created maltego data interchange.

The result of this custom transform approach can be seen in **Figure 4** which shows that starting from a Domain entity as the seeded data source and using these custom transforms provides easy access to data the owner may not want to expose such as a relevant bank and balance of a loan taken out.

```
from maltego_trx.entities import Phrase
from maltego_trx.transform import
      DiscoverableTransform

import json
import requests

class zillowLender(DiscoverableTransform):

@classmethod
def create_entities(cls, request, response):

      tokenFile = open("transforms/token.txt", "r")
      try:
            token = tokenFile.readline().rstrip()
      finally:
            tokenFile.close()

      baseUrl = "https://api.bridgedataoutput.com/api/
            v2/pub/transactions"
      accessToken = "?access_token=" + token
      queryType = "&address="
      inputData = request.Value
      sortAndOrder = "&limit=1"

      apiData = requests.get(baseUrl + accessToken +
            queryType + "\"" + inputData + "\"" +
            sortAndOrder)

      resultJSON = apiData.json()

      out = resultJSON['bundle'][0]['lenderName'][0]

      responsePhrase = "Mortgage Holder: " + out

      response.addEntity(Phrase, responsePhrase)
```

Python Code for Custom Transform

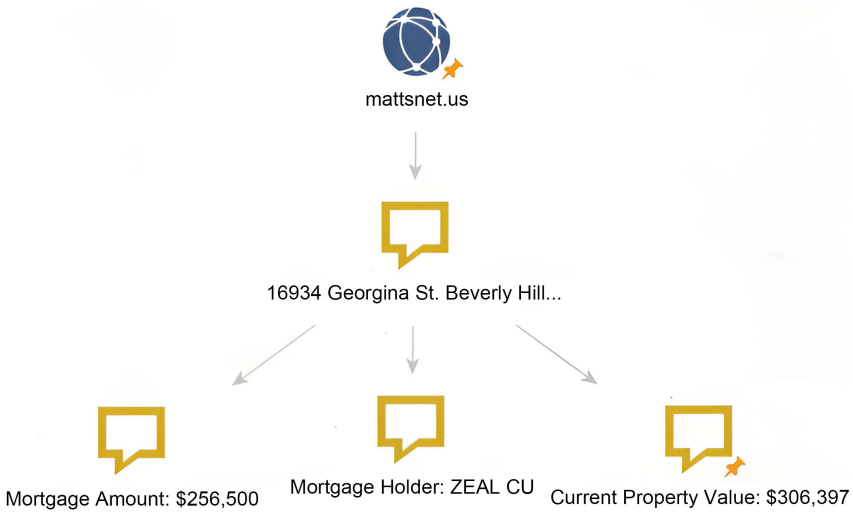

Figure 4. Resulting data from custom API transforms.

Maltego is an effective open-source intelligence (OSINT) tool for link analysis, data mining, and information collecting. A few of the tendencies are in the development of Maltego. Maintaining and enhancing data integrations is to access a variety of commercial and open-source data sources. Improving visualization tools is to give consumers more understandable and educational depictions of linkages and interconnections in the data. Integrating with tools and platforms is for cooperation to improve investigator communication and knowledge exchange. Introducing capabilities for workflow orchestration is to manage various data sources and analysis steps and automate intricate investigative processes. Putting security improvements into practice is to protect sensitive data and maintain the privacy and confidentiality of investigations. Developers use open APIs and development frameworks to meet their unique requirements for facilitating the creation and sharing of custom solutions. All in all, Increasing functionality, enhancing usefulness, and adjusting to the changing needs of OSINT community is the development trend for Maltego.

5. Other Valuable Tools in an Osint Investigation

There are many other tools that can be used for various investigations ranging from simple to complex. Investigators have the option of using prebuilt tools or they can create their own tools with scripts. Some examples include scripted programs that spawn custom web crawlers to search sites for information and then parse findings into a textual report, such as Metagoofil. Another common approach is using search engines such as Google or Facebook's graph search in creative ways. Another example is the Security and Exchange Commission's EDGAR tool used for looking up company filings from publicly traded companies at the https://www.sec.gov/ website [17].

There are also many different APIs that can be used to aid in investigations that can be used to connect to public data sources or interact with useful software applications that aid in the investigating process when coding custom

scripts or programs for investigations [4]. It is also important not to overlook the metadata that can be found embedded in web pages or hosted documents as they can be a useful source of information that can easily be overlooked.

There are a number of other prebuilt applications that help in an investigation by automating the data gathering process that runs many relevant searches on sites and data sources all at once such as PhoneInfogia or theHarvester. The following sections review tools in detail to introduce the reader to a few available OSINT activities that could help find desired information faster.

5.1. PhoneInfoga/Phone Research Databases

In OSINT one valuable piece of information in an investigation is a phone number. With a phone number, the individual can find information about the target's name, address, email or social media accounts, net worth, and their relatives or closely known associates. Many tools and resources exist to help individuals use phone numbers to find information about the person linked with the number. Websites such as whitepages.com, yellowpages.com, pipl.com, usphonebook.com, and spokeo.com exist that allow individuals to look up a phone number with name and location, email, or phone number to find out information about an unknown individual. In many cases, these websites offer a limited amount of information for free and require a paid subscription for full access to all information available with full background checks and criminal records available at additional costs. Many free sites also exist but we observed the quality of information varies between these sites and it is in the user's best interest to verify the accuracy of information returned by cross-verifying between several sources. It is also possible to combine the free data between many different phone lookup sites to create a profile of an individual without having to pay for a service.

One tool that automates the search of a phone number across many websites is known as PhoneInfoga. This tool searches several sources including Google, social media platforms, and other APIs to find information about the number. PhoneInfoga was created by Sundowndev and can be downloaded from the PhoneInfoga GitHub repository. The tool can be run from the command line or with a user interface in a web browser. To run from the command line enter the scan –n command with the desired number. Searching functionality is represented pictorially in **Figure 5**. The first step of the process is utilizing the Numbiverify API to find the country, area, city, carrier, and time zone of the number [18]. In the next phase, the tool footprints the number on Google using automated Google Dork searches that explore if the number appears on any disposable phone number sites, phonebook sites, reputation sites, social media sites, documents, and other webpages [18]. Finally, the tool uses the OVH API to search if the number is a voice-over IP number and return the VOIP type and provider. PhoneInfoga does this in a matter of seconds which is a huge advantage to individually performing each search manually which would take substantially longer. Users may even run the tool from a web browser for a better user experience (**Figure 6** and **Figure 7**).

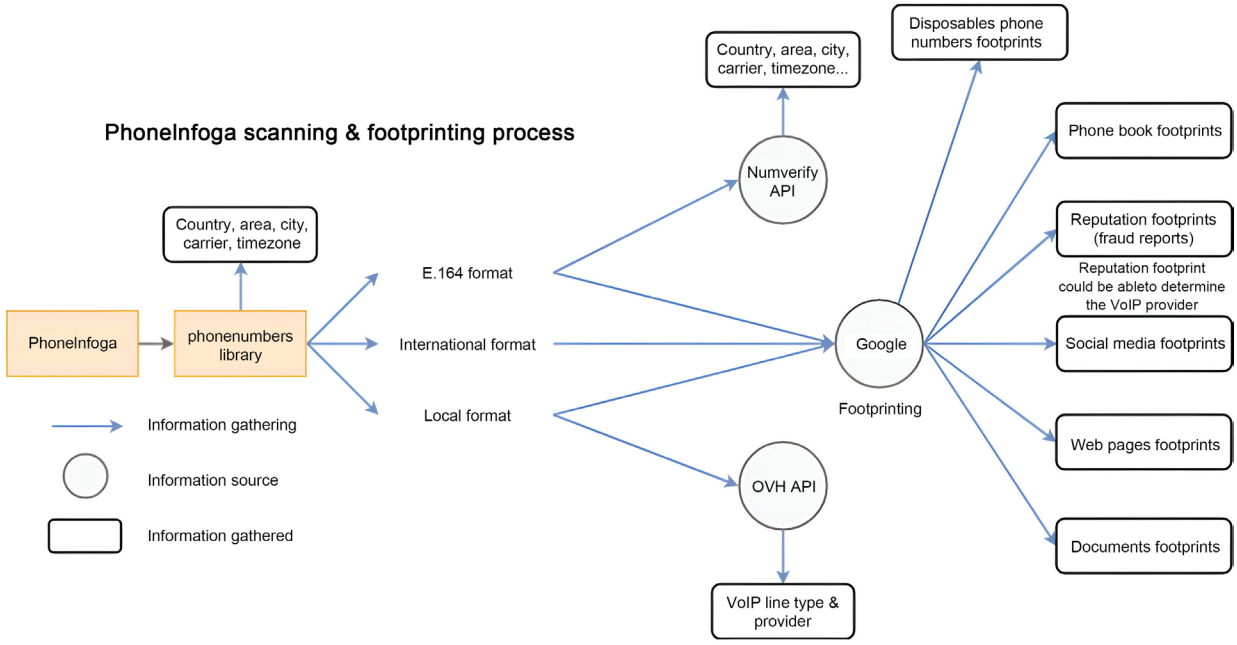

Figure 5. PhoneInfoga data map [18].

As you can see the data is not only organized in a way that makes it easier to read and access the link but there is a button that can be clicked to open all links in each section. All the user has to do is enter the number and then sort through the tabs the tool opens in a web browser. This helps the individual rapidly explore the returned results without manually searching across all the different services. The quality and quantity of information vary dramatically between different phone numbers as well so it is not uncommon for a search not to provide valuable results for a particular query depending on what type of number it is.

One example of investigation an individual can use this tool for is to look up an unknown number that has been called or look into a number that appears in a classified ad for an item they are looking to purchase. It is very important to verify all information that is returned by these services because it does a broad public record search and in many cases the data may be outdated or untrustworthy. Accuracy results fluctuate between different tools and websites.

5.2. Google Dorks

Google "Dorking" is a method for querying search engines for vulnerabilities exposed by public information sharing [19]. This creative search method combines words and phrases that would be expected to be in sensitive files along with filenames, directory paths for expected files, and keywords related to the attack target to use in a search engine for discovery purposes. As discussed in other sections of this paper, it is wise to utilize anonymity tools such as a VPN or Tor to perform such investigations depending on the laws within the country where this activity is performed. Such methods will ensure that if searches are recorded they will not be traceable back to the person performing reconnaissance (**Figure 8**).

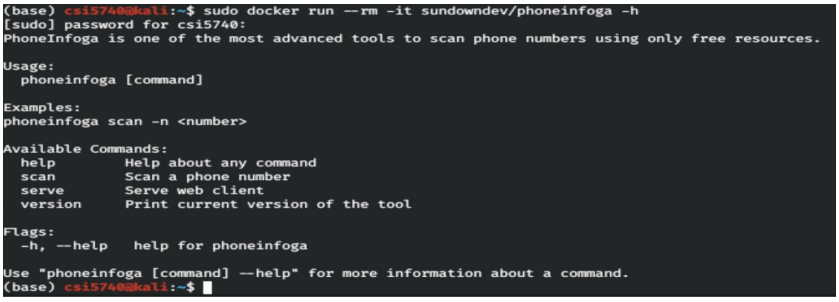

Figure 6. Help menu for PhoneInfoga.

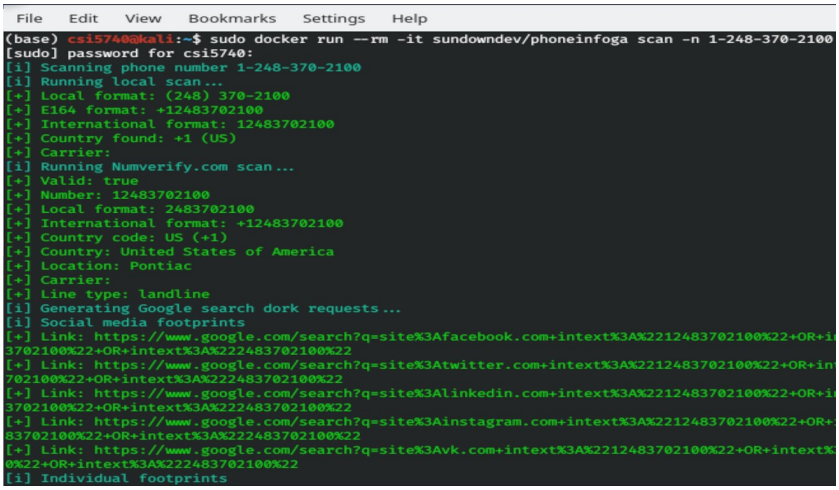

Figure 7. Example of a PhoneInfoga query.

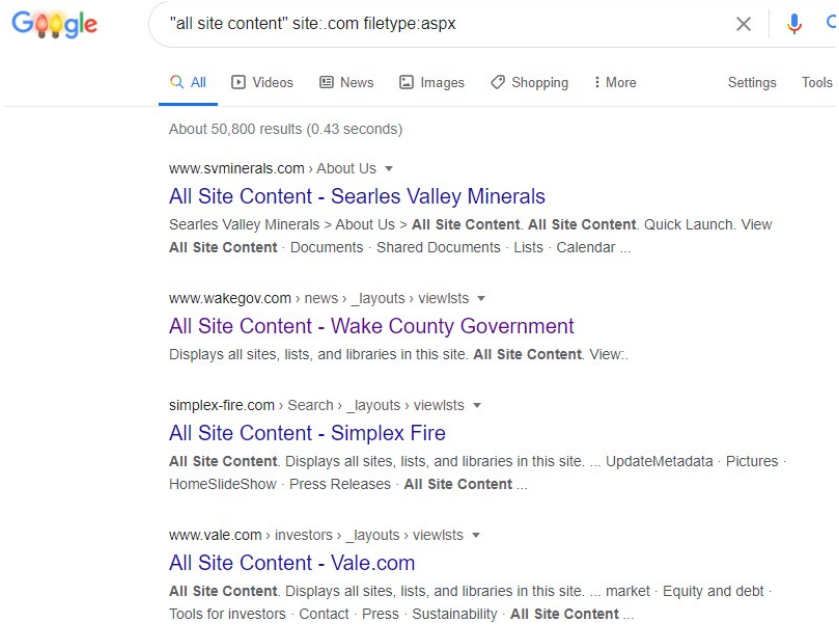

Figure 8. Example results for dorks Example 3.

Dorking makes use of special search parameters within the Google query language such as “site”, “filetype”, “meta”, “inurl”, and so on. Several examples of useful Google Dork searches follow:

1) filetype:xlsx invite zip code - Search for personal data via Excel files that are related to event planning

2) filetype:xls email - Identify Excel files that have the keyword email in them for possible data mining activities

3) "all site content" site:.com filetype:aspx - Identify open SharePoint sites that allow view of the All Site Content page which may reveal sensitive information

4) "'System' + 'Toner' + 'Input Tray' + 'Output Tray' inurl:cgi" - Identify HP printers exposed on the public internet for vulnerability scanning

5) inurl:top.htm inurl:currenttime - Identify webcams available online without login by using a known in URL structure

5.3. Metagoofil

Metadata is information about a file that is embedded within the file by the creating application [20]. Much of this information is not of particular value such as file creation/modification timestamps or the software title which created the file. However, there are valuable elements within certain file types such as PDFs, and Microsoft Office Documents. These file types typically provide information about the author as well such as the username logged in when the file was created, creator's email address, and the editing time for a document.

Metagoofil is a tool that takes advantage of metadata within files made available on the internet which the author may or may not realize they are sharing [21]. This tool is a Python script that acts as an extension of the Google Dorks concept presented previously. Essentially this tool uses information from a search engine to find public files matching the entered domain name and file type entered to extract embedded metadata information. The tool does so through the following steps:

1) Search Google by domain name and entered filetype (using Google Dorks "filetype:_filetype_site:_domain_" per the source code)

2) Download the number of matching files specified by the user to a local folder (defaults to 200)

3) Parse downloaded files one-by-one to extract available metadata using dependent tools such as hachoir and pdfminer

4) Create an output report (html or XML format based on user input) and output the resulting data on the command line

This tool exists as a Python script which was last updated in August of 2015. Unfortunately, the tool is no longer compatible with modern Python libraries and does not function with the new Python3 which renders the version that ships with Kali non-functional [22]. Instead, a new version has been forked and maintained by the GitHub user Hackndo.

[23] which is functional. To experiment with this tool, the below command is issued to trigger a download of 100 PDF and DOC files which can be found via Google on a corporate target of "Apple.com".

```
metagoofil -w outFiles -t pdf,doc -d apple.com  -f
    100
```

The resulting output as seen in Figure 9 (redacted to protect privacy) shows a variety of email addresses, many of which clearly belong to the organization. In addition, the tool also provides names of content originators and the software used to create the files. All of this information may be valuable in a social engineering approach.

5.4. theHarvester

Another useful front-end for discoveries via search engines is an application called "theHarvester". The goal of this tool is to identify public-facing information related to either a domain name or a company including email addresses, subdomains within the provided domain, IPs, and additional URLs for review [24]. This information can then be used to discover information about the company such as Test/QA URLs that may reveal product information. It can also be utilized in the social engineering perspective by identifying resources within the organization to approach to learn more.

The tool is provided as a Python3 command line application which provides both passive and active information gathering methods. On the passive side of functionality, the application provides a method to scrape many popular search engines such as Bing, Google, Baidu, DuckDuckGo, GitHub, LinkedIn, and Shodan for results. On the active side of the application, functionality for performing a brute force DNS dictionary lookup to identify subdomains as well as a capability for grabbing screenshots of each (which requires a direct connection to the target) is also provided.

Of note, similar to the caveats listed in the Maltego section of this paper, many of the services that theHarvester can scrape also require an API key. Keys are freely available for each service but require an extra step by the user to establish a normal browser session to the target search page and extract strings from the created session cookies to allow the command line application to appear to be using the same session. To experiment with the tool once installed and loaded with the relevant API keys, a scrape of a domain can be triggered to identify 100 results from all search candidates against "Apple.com" by issuing the command below. The resulting output can be seen in Figure 10.

```
theHarvester -d apple.com -l 100 -b duckduckgo
```

Of note, when using the "-b all" flag to search every available search tool available within theHarvester, the tool identified 16,518 matching IP addresses, 3 matching email addresses, and 32,308 matching domain names with references as obscure as platformengineering.dashboard.iso.apple.com. This data source along with a shell script to verify the presence of listening web servers at the discovered domain names could certainly be an attack vector to locate data elements that may be unintentionally shared. Such an example may be information from the company ahead of intended release on QA sites as has been a source of product information in the past.

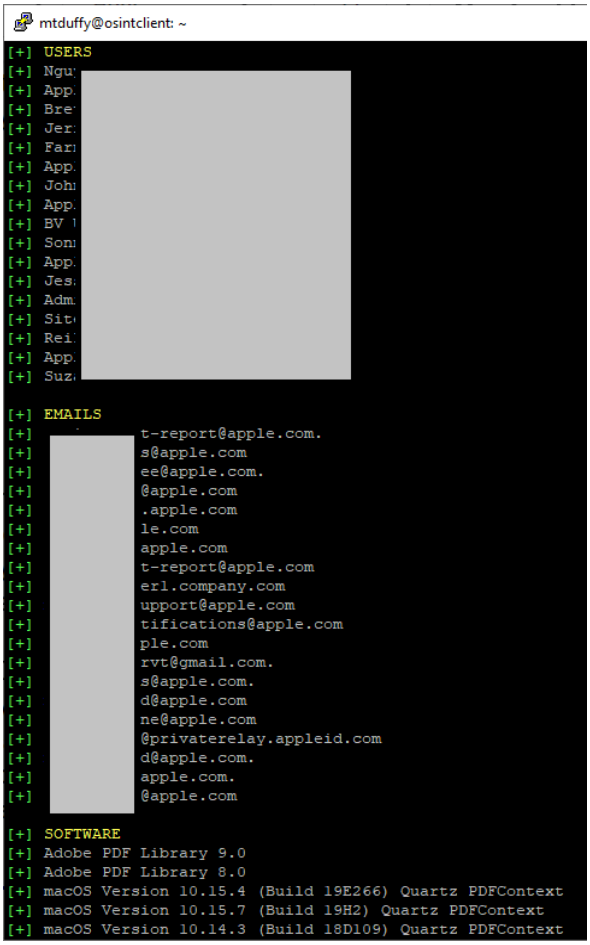

Figure 9. Example Findings using Metagoofil for Apple.com.

Figure 10. Example Findings using theHarvester for Apple.com.

Finally, as with several other tools encountered in this project, it can be noted that theHarvester does trigger some of the rate-limiting features of the Google search platform. When using any combination of theHarvester, Metagoofil, and other Google Dorks-related tools, when creating enough search traffic Google will temporarily block the IP address in use from automated scrapers and require the completion of a captcha to proceed as shown in Figure 11. This is another reason why the use of anonymization services is recommended for such activities.

6. Issues in the Age of Information

6.1. Ethics

Ethical issues arise from using public data sources to collect and distribute information. For instance, the intellectual capital of human beings is impaired if they have lost personal information without being paid for it or if they are not allowed to access some information that may be valuable to them. To better live in the world, people may reveal information which is to keep intimate relationships with others. Even if people knew the information that was relying on the error for their own life, they may ignore it to keep current good situations. Sometimes, policymakers may covet information even if obtaining it by invading the privacy of other people due to increasingly valuable information [25].

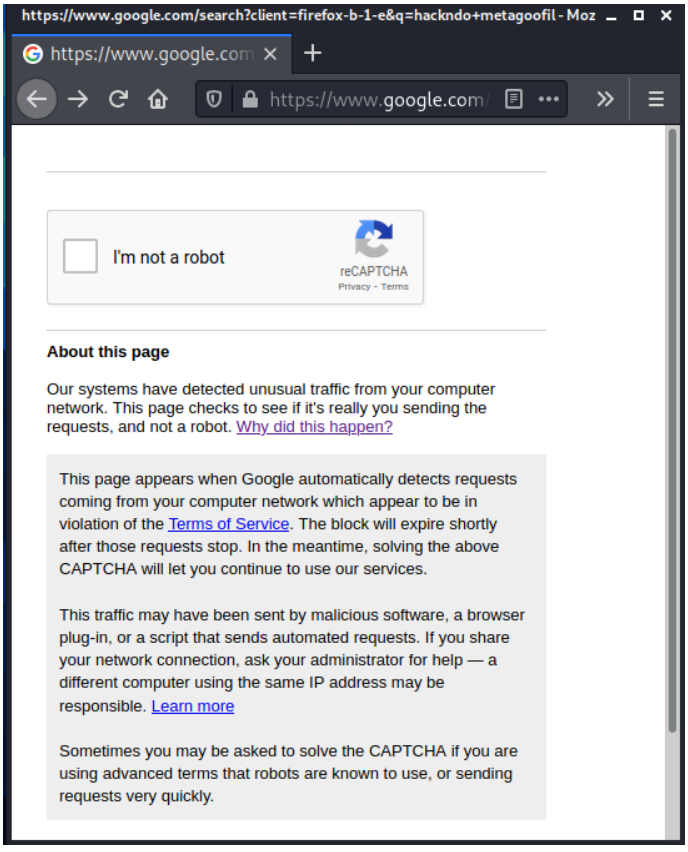

Figure 11. Google's response to suspected abuse.

We could take some countermeasures to address ethics issues such as getting informed consent before collecting any personally identifiable information. Removing or encrypting personally identifiable data from the database is to prevent the identification of individuals. Preventing unauthorized access or data breaches is using encryption, access control, and securing storage practices. Taking respecting privacy rights into account, we should not access or share sensitive personal information without legal justification or proper authorization.

6.2. Cyber Law

Particular case on Clearview AI uses robots/spyders to scan social media then uses AI to analyze and create records of people then sells access to its database to law enforcement and select third parties. This activity has not complied with the Computer Fraud and Abuse Act (CFAA) which is a United States cyber security bill to states opposition to it on the grounds, that it would make kind of routine activities via the internet illegal [26].

The development of cyber law for the use of AI tools in data collection and profiling is establishing principles that prioritize transparency, fairness, privacy, and accountability in AI-driven processes. We would provide explanations for promoting Algorithmic transparency and improving decision-making processes to let people understand how their information is being used to generate profiles. Profiling processes should be fair, transparent, and accountable to avoid bias and discrimination in data collection and analysis. People should stay informed about existing and emerging regulations governing data privacy and AI usage.

6.3. Computer Fraud

Not everything on the internet is true. To obtain unlawful use of a computer system, a hacker may distribute hoax emails to targeted users by using public data sources to get their email addresses and names. Phishing as a typical social engineering is getting sensitive data from victim users including credit card info, username, password, and Social security number when taking the users to a fake webpage that looks like a legitimate site. Email spoofing and instant messaging are two frequent methods of fraudulent attempts in the public network communication environment [27].

It is crucial to implement robust countermeasures to prevent computer fraud against unauthorized access or misuse of information. Restricting access to sensitive data is only to authorized personnel. Using strong authentication mechanisms like multi-factor authentication to verify the identification of the user who accesses data. Making or redacting sensitive data from public data sources is to prevent exposure of personally identifiable information or other sensitive information. Only allowing the minimum amount of data is accessing for legitimately necessary purposes. Deploying data loss prevention (DLP) solutions is to monitor and prevent unauthorized sharing or leakage of sensitive information.

People could use DLP policies to enforce data protection regulations and

block illegal attempts to obtain or transmit sensitive data.

6.4. Future Works

With the fast development of technology, we could be able to find more valuable transforms running on the various operating systems not only working on Windows and Linux operating systems as this project researched. Due to limitations of time in this project, we are unable to research more coalition operations by combining intelligence, and surveillance with reconnaissance assets such things like human intelligence, data fusion, sensing platforms, sensors, and network elements [28].

During the age of information operations, we should pay more attention to social network analysis of information sharing and situational awareness for improving effectiveness in the organization in case data overload is the result of a mix of diverse huge open sources, multiple information formats, and large info volumes. There is an existing network-enabled operations framework as a positive tool to make us have greater situational awareness and develop the effectiveness of missions in military operations [29].

Another area of future research is to increase actional timeline and mission effectiveness since current data overload and constraint of high-capacity communication bandwidth issues exist between intelligence, surveillance, and reconnaissance (ISR) assets which could not timely distribute information. If this issue could be resolved, it would have played critical roles in supporting current and future military operations [30].

7. Conclusions

Open Source Intelligence investigations are not a new concept, however, the tools and methods continue to evolve. Privacy issues in this space have existed long before the Internet age. With the massive increase in information accessibility provided by the internet, information is obtainable with ease if an investigator understands both the proper tools and data sources that may be used. The challenge as demonstrated by research results is not in revealing data, but in narrowing the focus to the intended target and in relating it to useful information.

Within this paper, the OSINT space has been generally introduced with several nuances to consider before starting such investigations. In addition, numerous tools have been reviewed that collecting data from several different data sources includes open APIs, closed but easily accessible APIs, and open internet search sources. Readers of this paper are encouraged to keep the ease of information accessibility in mind when volunteering information online and to be conscious of their digital footprint.

8. Tools Used

During our presentation, Dr. Fu requested a view of the tools that were analyzed and used in this project. Please see **Figure 12** for a summarized view of these tools.

Tool	Interface	Description	Primary data available	Free or Paid	Download / Access URL
Maltego	Thick client Java application	Thick client Java application with graph based data aggregator	Addresses Phone numbers Email addresses Custom additions based on custom transforms	Free for basic functionality €999/year for full functionality	https://www.maltego.com/downloads/
Google Dorks	Website	Use of existing web search tools with advanced query elements to reveal sentitive data (Typically uses Google, but opportunities available on Bing and others)	Many - no limitation	Free	Google.com
Search Digity	GUI Application	The primary attack tool of the Google Hacking Diggity Project. It is Stach & Liu's MS Windows GUI application that serves as a front-end to the most recent versions of GoogleDiggity, BingDiggity, Bing LinkFromDomainDiggity, Google CodeSearchDiggity, DLPDiggity, FlashDiggity, and MalwareDiggity	Many - no limitation	Free	https://resources.bishopfox.com/resources/tools/google-hacking-diggity/attack-tools/
PhoneInfoga	Scripted Application	Tool used to find information on phone numbers often called reverse phone number look up, aggregates many data sources into a single view	Phone number background information including location, service provider, ownership information, available use of the number	Free	https://github.com/sundowndev/PhoneInfoga
Metagoofil	Scripted Application	Scripted approach for locating files on the internet by type (MS Office; XLS, PPT, DOC & PDF) via a Google Dorks approach which then downloads content and scrapes it with metadata review tools (hachoir, pdfminer). Information scraped is provided in the shell.	Names, e-mail addresses, and creating applications from embedded data within documents found	Free	Original Python2 (defunct): https://github.com/laramies/metagoofil Maintained Python3: https://github.com/Hacknbo/metagoofil
theHarvester	Scripted Application	Gather emails, subdomains, hosts from public sources such as Google and Shodan to create a footprint of endpoints exposed for a given company or domain name	Emails, subdomains, hosts, and active web sites at identified subdomains	Free	https://github.com/laramies/theHarvester
Kali Linux	Operating System	Kali Linux is a purpose build penetration testing operating system that comes preinstalled with many useful OSint applications Like Maltego and theHarvester	N/A Operating system	Free	https://www.kali.org/downloads/
Tails OS	Operating System	as a portable operating system to protect against surveillance, censorship, advertising, and viruses. It is not only protecting privacy online for customers, but also helping them avoid censorship. When shutting down, it leaves no trace on the computer.	N/A Operating system	Free	https://tails.boum.org/install/
Tor	Anonymizing Browswer	Tor browser is to protect searching results if Google Dorks in use for any search engine because it would mask users' internet traffic and divorce their computers' identifying information from the websites that they are accessing. There has detailed guidelines about how to use the Tor Browser on different operating systems such as Windows and Linux OS. This tool always makes their searching more difficult.	N/A Dark Web Browser Application	Free	https://www.torproject.org/download/
I2P	Anonymizing Browswer	As the invisible internet project tool which is a fully encrypted private network layer to offer protection for user activity, location, and identity. Under I2p network, the user could connect with other people without the worry of being tracked or their data being collected.	N/A Dark Web Browser Application	Free	https://geti2p.net/en/download

Figure 12. Brief table of tools used in this project as requested during the presentation.

9. Team Contributions by Member

In general, the three members of this group worked very well together throughout the semester with a weekly meeting cadence of 2 hours on Tuesday evenings. All members participated with a focus on identifying tasks and accomplishments to enable each team member to work independently for the next week to ease scheduling challenges. Any time group members were assigned with specific scope, there was always progress toward the task or a collaborative sharing of challenges to be worked through. All agree that they would work together with this group again in the future.

9.1. Common Items

Identification of the Project: During the first week of collaboration, the group shared a spreadsheet for ideas to pursue this semester. This was reviewed again with feedback on the second week of our regular cadence and the OSINT topic was chosen in a consensus among the group.

Progress Reports: Similarly, all progress reports were written collectively due to individual ownership of sub-tasks. Generally, they were created together while on our weekly call and turned in once all approved.

Deliverable Creation: Both the presentation and the paper have been a collective activity with content added from each group member.

9.2. Individual Scope Areas

The areas which were uniquely owned are detailed below by group members.

Matthew Duffy

1) Research and presentation on OSINT tools Maltego, Metagoofil, theHarvester

2) Writer of final paper sections IV, V.C-D, VII, IX

3) Creation of the custom Maltego transforms for whois lookups and Zillow API interaction using Python

4) Conversion of our paper into the LaTex format

Xueying Pan

1) Research and presentation on OSINT tools Search Diggity and Google Dorks

2) Writer of final paper sections I, III, IV, V.B, VI

3) Creation of the project website on wix.com

4) Coordination/deliverable hand-in and keeping all on track with upcoming due dates

Samuel Wilson

1) Research and presentation on OSINT tools PhoneInfoga, and web tools for phone number search

2) Writer of final paper sections I, II, III, V.A,

3) Research into the background of our topic and norms in the space through two primary textbooks

4) Creator of initial paper outline for the group

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Hassan, G.N.A. and Hijazi, R. (2018) Open Source Intelligence Methods and Tools: A Practical Guide to Online Intelligence. Apress, Berkeley. https://doi.org/10.1007/978-1-4842-3213-2

[2] (2020) Open Source Intelligence (Osint). https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/311512p.pdf?ver=2019-03-06-093811-687

[3] (2021) Nato osint handbook v1.2. https://www.studocu.com/en-us/document/collin-college/operating-system-security/nato-osint-handbook-v12-jan-2002/87709361

[4] Gibson, H. (2016) Acquisition and Preparation of Data for OSINT Investigations. In: Akhgar, B., Bayerl, P. and Sampson, F., Eds., *Open Source Intelligence Investigation*, Springer, Cham, 69-93. https://doi.org/10.1007/978-3-319-47671-1_6

[5] Tails. https://tails.net/

[6] Chapple, M. (2019) Tor and the Dark Web: Exploring the Basics. https://www.linkedin.com/learning/tor-and-the-dark-web-exploring-the-basics?trk=lynda_redirect_learning

[7] Kalpakis, G., Tsikrika, T., Cunningham, N., Iliou, C., Vrochidis, S., Mid-dleton, J. and Kompatsiaris, I. (2016) OSINT and the Dark Web. In: Akhgar, B., Bayerl, P. and Sampson, F., Eds., *Open Source Intelligence Investigation*, Springer, Cham, 111-132. https://doi.org/10.1007/978-3-319-47671-1_8

[8] Kali Linux Tools Listing. https://tools.kali.org/tools-listing

[9] (2020) Introduction to Maltego Standard Entities. https://docs.maltego.com/support/solutions/articles/15000035722-introduction-to-maltego-standard-entities

[10] (2020) What Is Open Source Intelligence (OSINT) and How to Conduct OSINT Investigations in Maltego.

https://www.maltego.com/blog/what-is-open-source-intelligence-and-how-to-conduct-osint-investigations/

[11] Make It Your Own. https://www.maltego.com/pricing-plans/

[12] Social Links PRO. https://docs.maltego.com/support/solutions/articles/15000054072-social-links-pro#overview-0-0

[13] Who, What & Why. https://haveibeenpwned.com/About

[14] Multiple Listing Service (MLS): What Is It. https://www.nar.realtor/nar-doj-settlement/multiple-listing-service-mls-what-is-it

[15] Bridge API Documentation. https://bridgedataoutput.com/docs/platform/

[16] (2020) TRX Transform Library Guide. https://docs.maltego.com/support/solutions/articles/15000024277-trx-transform-library-guide

[17] (2020) Edgar—Search and Access. https://www.sec.gov/edgar/search-and-access

[18] PhoneInfoga. https://github.com/ExpertAnonymous/PhoneInfoga

[19] Tech, T. (2020) Smart Searching with Google Dorking. https://exposingtheinvisible.org/guides/google-dorking/

[20] Riley, J. (2017) Understanding Metadata. https://digital.library.unt.edu/ark:/67531/metadc990983/

[21] Martorella, C. (2015) Metagoofil. https://github.com/laramies/metagoofil

[22] Metagoofil Does Not Find Any Results #13. https://github.com/laramies/metagoofil/issues/13

[23] (2020) Metagoofil—Python 3. https://github.com/Hackndo/metagoofil

[24] (2020) The Harvester. https://github.com/laramies/theHarvester

[25] Mason, R.O. (1986) Four Ethical Issues of the Information Age. https://www.gdrc.org/info-design/4-ethics.html

[26] 18 U.S. Code § 1030—Fraud and Related Activity in Connection with Computers. https://www.law.cornell.edu/uscode/text/18/1030

[27] Wong, C. (2024) What Is Phishing? Examples, Types, and Techniques. https://www.csoonline.com/article/514515/what-is-phishing-examples-types-and-techniques.html

[28] Pham, T., Cirincione, G.H., Verma, D. and Pearson, G. (2008) Intelligence, Surveillance, and Reconnaissance Fusion for Coalition Operations. https://apps.dtic.mil/sti/pdfs/ADA520498.pdf

[29] Buchler, N., Fitzhugh, S.M., Marusich, L.R., Ungvarsky, D.M., Lebiere, C. and Gonzalez, C. (2016) Mission Command in the Age of Network-Enabled Operations: Social Network Analysis of Information Sharing and Situation Awareness. *Frontiers in Psychology*, **7**, Article 937. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4916213/ https://doi.org/10.3389/fpsyg.2016.00937

[30] Hershey, P., Wang, M.C., Graham, C., Davidson, S., Sica, M. and Dudash, J. (2012) A Policy-Based Approach to Automated Data Reduction for Intelligence, Surveillance, and Reconnaissance Systems. *MILCOM* 2012: 2012 *IEEE Military Communications Conference*, Orlando, 29 October-1 November 2012, 1-6. https://doi.org/10.1109/MILCOM.2012.6415574