XV Jornadas de Ingeniería Telemática.
JITEL 2021.
Universidad de A Coruña.

*Actas de las XV Jornadas
de Ingeniería Telemática
(JITEL 2021),
A Coruña (España),
27-29 de octubre de 2021.*

ISBN

Open-Source Intelligence Profiling, Analysis and Countermeasures

Christodoulos Tziampazis
ETSETB,

University of Catalonia

christodoulos.tziampazis@estudiantat.upc.edu

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 analyze and propose potential counter measures which may affect the success rate of identifying and profiling an individual. Based on that, we are analyzing open-source intelligence 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 propose potential improvements that can be applied in the future that will assist the scope of this research.

Keywords – OSINT, Digital Profiling, Counter measures, Social media networks

I. INTRODUCTION

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 organization 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. 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 such as in the Cambridge Analytica case.

In general, OSINT includes methodologies of various kinds of techniques for gathering and analyzing 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 behavioral features and characteristics. OSINT profiling investigations can be divided into the following categories: Organized Crime & Militar, Cybersecurity and Business & Social Intelligence [1].

This master thesis focuses on the digital profiling of individuals and reports on counter measures against methodologies and techniques used to collect, aggregate and analyze the already available public data. Additionally, the scope of the research is also to bring together the different methods used for profiling and reflect on their level of severity and exposure.

II. STATE OF THE ART

This section summarizes 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.

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. Additionally, other techniques can be also combined with OSINT tools to make an attack or investigation successful. 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.

In this section the 11 following OSINT tools and techniques are briefly reviewed: osintframework, Internet Archive, Matlego, Shodan, Recon-ng, theHarvester, Google dorking, ExifTool, Creepy, Spiderfoot and Foca. The tools in general have many capabilities and overall different purposes. Tools with broad scope such as Recon-ng provide modules that can perform reconnaissance, discovery and even exploitation tasks where others like ExifTool serve a single purpose. Maltego [2] it's another 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. Another exceptional tool is Shodan [3] **.** Shodan is a search engine for all the internet-connected devices such as webcams, SCADA, routers, and many more. It collects all 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. Finally, one can find an enormous list of tools based on their scope available at osintframework [4]. It provides a categorized tree structure schema that contains nearly every available OSINT tool of both free and paid alternatives.

Since this research is entirely focused on digital profiling through SMNs, we created a short capability list in order to classify the tools and emphasize only on those with social media and search engine capabilities. As explained in Methodology IV, Maltego, Recon-ng, theHarvester and Spiderfoot are the tools that were chosen to be further analyzed.

III. Literature Review

The papers reviewed 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.

A. Natural Language Processing

In this first article [5], researchers 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 [6], they developed an OSINT tool called Textractor that its goal is to extract and analyze 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.

B. 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 [7], 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, Area of interest labels, other social network connection, Registration date and Common event registration. Based on these attributes the researchers 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 neighboring 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.

C. Behavioral & Content-based analysis

Today's social media network application, especially the most famous ones, have already in place and enabled by default all the necessary privacy settings. The user indeed has the chance to choose between different levels of privacy settings.

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 behavioral analysis. The paper [8], conducts a research on Twitter with the ultimate goal of identifying malicious or benign users. For the behavioral 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 behavioral features and from the NUANCE framework they make use of Recentered Local Profile's algorithm in order to compare the profiles for each 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.

D. 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 behavioral analysis of hundreds of thousands of users is infeasible to be performed manually.

In the paper [9], the researchers crawl a community of YouTube users and they make use of their comments, uploads, favorites 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.

IV. 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:

1. Identification and classification of the tools.
2. Tool’s data sources classification.
3. Practical utilization of the tools including detailed examples of how they work and how they extract information.
4. Reviewed literature.
5. Conduct an analysis based on the methodologies from [10].
6. Concluding the results by proposing counter measures against the profiling of individuals.

The above sequence of steps can be summarized, as depicted in Fig. 1, in a waterfall scheme of 3 categories.

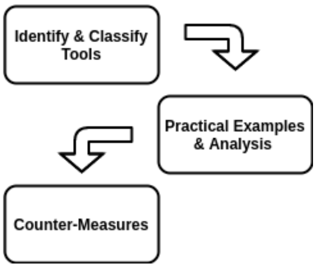


Fig. 1: Waterfall methodology summary.

A. Tools Identification & Classification

This section plays the most important role in the development of this research paper. Briefly, it collects 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, analyzing and presenting the collected information.

B. 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 [10].

As a first step, we will turn our attention only on those specific tools used for profiling. Namely these tools are: Maltego, Recon-ng, theHarvester, Spiderfoot. The approach we conducted goes as follows. Firstly, after reviewing the OSINT tools we identify those used for profiling as explained. Then we collected and classified all the data sources that these tools are using. 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.

Moreover, we are briefly recreating the data collection methodology presented in the paper [10]. 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.

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.

C. Counter-measures Proposal

The Counter Measures section 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 analyzed 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.

V. Results

A. OSINT tools & Data Sources

In order to understand the nature of the sources used by the tools we must first classify them and analyze 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.

- Social Media: Data sources that are based on raw SMN API calls.
- Search Engines: 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 etc.
- Script Crawlers: Automated scripts crawling fixed URLs such as SMN user endpoint URL.
- DB - Data breach crawlers: Databases that were populated either by publicly exposed breached data or by businesses that use to collect and aggregate information such as personal data.
- Local Scripts: Any type of script that mangles the given input with the ultimate goal to uncover user details about our target.

In more detail the total number of data sources identified(not unique) from those 4 tools are 391. From these, only 66 are used in digital profiling of an individual and from those data sources only 37 were returning an outcome.

All 4 tools are been examined alongside with their data sources in an attempt to examine their scripts and techniques of collecting data. In general, 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 to the fullest in this research. Maltego provides less than 10 data sources for the community version license. Even with this small community edition of Maltego we can still recover a lot of information about our target as depicted in Fig. 2.

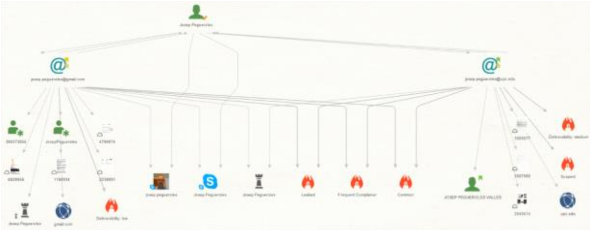


Fig. 2: Meltego single target output.

Recon-ng, Spiderfoot and theHarvester are open-source tools and thus a more thorough analysis has been performed. These 3 tools share a similar procedure for collecting OSINT and thus each of them has for every module available a parser that is responsible to scrap the return data and present only relevant results.

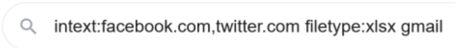
Finally, and most importantly, the tools reviewed are unable to infer context and that is a crucial draw back in digital profiling. In today's internet social media images are the most posted content and unfortunately the tools described cannot infer information from the pictures. However, the aforementioned literature machine learning methodologies, are enabling techniques for uncovering context from the gathered data.

B. Methodology Recreation

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 will also help us conclude on more concrete counter measures. The following methods can be used in many different ways and with combination of other not mention methodologies.

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. Our approach is divided into 2 parts. The first part includes queries concerning a specific target and the second task queries for random targets. The first part provided us with information such as full name, username, mentioned tweets and events and many more. This shows how simple dorking can be used to create an initial map for our target. In a larger scale, our scope is to collect as much as possible sensitive details that are publicly available on the internet with the desired data to be email addresses. 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 Fig. 3.



intext:facebook.com,twitter.com filetype:xlsx gmail

Fig. 3: Dorking query

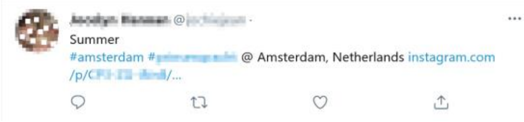
Even though it falls outside the scope of this research, alongside with emails we also recovered some very crucial information about random people and businesses such as personal and business emails, passwords, office addresses and so on.

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 at its essence another SMN that people can create a profile listing all of their means of contact. 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 and Knowem.

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 Fig. 4, 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.



[redacted]

Summer
#amsterdam #[redacted] @ Amsterdam, Netherlands instagram.com /p/[redacted]/...

Fig. 4: Twitter & Instagram linkage

4. Self-Disclosure

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.

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 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. We then signed up in the 3 SMNs mentioned before and as a result only LinkedIn's friend finder feature disclosed a user account.

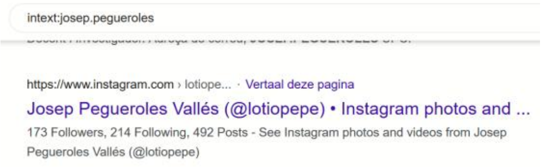
We used a very small number of accounts since our goal was not to linked accounts or uncover personal details about our targets but rather to explore the methodology and the process of uncovering identities through the friend finding feature.

C. Counter Measures

The measures presented in this section can be considered as 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 the publicly available information and the rapid change of techniques and technologies used for collection and analysis.

To begin with, from the methodologies of all the different reviewed 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.

Let us now look further into the 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 are plenty of available scripts, 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. 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 such as names, would become extremely difficult. Another important aspect is the reusability of usernames in different SMNs. For example, in Fig. 5 one can observe 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.



intext:josep.pegueroles

https://www.instagram.com › lotiope... · Vertaal deze pagina

Josep Pegueroles Vallés (@lotiopepe) • Instagram photos and ...

173 Followers, 214 Following, 492 Posts - See Instagram photos and videos from Josep Pegueroles Vallés (@lotiopepe)

Fig. 5: Simple Google dorking

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 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 behavioral analysis since 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 behavioral 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 analyzed figure of speech and predict if the 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, 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 finding everything instantly; does include unfortunately documents exposing personal details as discussed above.

VI. Conclusion & future development

This research has as main goal to explore and discover digital profiling techniques and methodologies and analyze 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 discussed above, 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 human behavior analysis they manage to give context and create relationships in an automated way. This makes our task even more complicated since analyzing 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.

This research has broad scope and thus collecting, analyzing 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. 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 the nature of 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 & authenticator key (TOTP) and even custom fields. In our case, it would be a great contribution if a tool can be developed that would have an anti-profiling capability. 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.

ACKNOWLEDGEMENTS

This thesis and paper would not have been possible without the help and support of my supervisor Josep Pegueroles which with his expertise and knowledge, he guided me throughout the research and helped me in every single obstacle I countered.

REFERENCIAS

[1] Bc Ondˇrej Zoder. Automated collection of open-source intelligence.

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

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

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

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

[6] Antonio Magal̃haes and Joao Paulo Magalhaes. Textractor: An osint tool to extract and analyse audio/video content. In Jose Machado, Filomena Soares, and Germano Veiga, editors, Innovation, Engineering and Entrepreneurship, pages 3–9, Cham, 2019. Springer International Publishing.

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

[8] 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.

[9] 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.

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