MINING HACKER FORUMS FOR PROACTIVE THREAT INTELLIGENCE

A Project Report

submitted by

Manogna P.V.S

(301470177)

Kajori Roy

(301549849)

Jay Garchar

(301555819)

under the supervision of

Dr.Mohammed Tayebi

in partial fulfillment of
the requirements for the degree of
Master of Professional Computer Science

December 2022

Contents

Chapter 1

Introduction

1.1 Introduction

This document is a result of continuous efforts to develop the Pipeline-based approach for identifying key actors using Cyber Threat Intelligence.

With so many organisations constantly facing sophisticated and destructive cyber threats, the threat landscape is currently growing at a rapid rate.The complexity of cybercriminals' tools and techniques makes it difficult for traditional security controls to keep up, therefore businesses are now seeking for more effective ways to improve their cyber security capabilities.Vulnerabilities are at an all-time high due to the quick development of new technology , It is essential to forecast cyberattacks, take adequate safeguards, and use cyberintelligence that enables activities in order to successfully handle this circumstance.

Many firms have historically gathered and analysed data from internal log files, leading in reactive CTI.The online hacker community can provide significant proactive CTI value by alerting organisations about threats they were previously unaware of. The clear, social, and dark web have been found as rich sources of significant cyber-security information that may be detected, crawled, and then leveraged to actionable cyber-threat intelligence.

Manually extracting meaningful threat information from these sources is a time-consuming and error-prone procedure that necessitates a large investment of resources. In order to identify important hackers participating in underground forums, it is necessary to suggest a technique or programme that will automatically evaluate these forums.

We addressed this issue by developing a end to end pipeline where a proposed architecture adopts a four-phase model which begins with information gathering from various hacker forums and pre-processing the data; Analysing the data ;Ranking the top hackers and Providing intelligence on the data.

1.2 Background

Complex information systems are used by businesses, governments, and academic institutions all over the world to run their operations.Each of these devices has the potential to be a target for a cyberattack, and their rapid advancement has produced a significant riseÂ in vulnerabilities that can be exploited.Among the most potential dangers to the country's national security are cyber threats.Malware is expected to be 50 times more prevalent today than it was ten years ago, and most of it may be exploited to launch deadly cyberattacks.[5]

As adversaries use a wide range of tools and strategies to attack their targets with motives ranging from intelligence gathering to damage or financial gain, cyberthreats have become more numerous and sophisticated over time. Due to the valuable data they manage, companies are among the potential victims that are especially high value targets for cyberattacks.

Cyber security is a complicated and diverse problem domain that is only getting more so.Every day, there are about 2,200 cyberattacks. A 600% surge in cyberattacks was spurred on by the pandemic in addition to health issues.Due to these factors, it is challenging to entirely stop all attacks, even when defences against cyberattacks are implemented. Even if 99.9% of organisations with a moderately large network successfully fight against assaults, the remaining 0.01% failures would inflict huge damage to the entire company.

Cyberthreats are still developing in an effort to defeat the security precautions taken by security experts. It is getting harder and harder to protect against sophisticated attacks or exploits in the present cybersecurity environment.Hackers are well-funded, have cutting-edge technology, and extensive experience. They are skilled in identifying the weak points in the actual enterprise network, including management and employees, in addition to being able to develop their assault strategies.

Cybercriminals are always developing more sophisticated and targeted attacks that can get beyond traditional security measures (eg: firewalls, intrusion detection and prevention systems, etc). The controls in place are also primarily reactive, meaning that they are modified "after the fact" with knowledge derived from the results of earlier successful attacks. To improve the efficacy of cyber security defence, more proactive strategies are required.To stay up with the continuously changing threat landscape, proactive security solutions like Cyber Threat Intelligence (CTI) are required.[3]

Information, which is frequently referred to as "cyber-threat intelligence,"(CTI) is typically derived from collected data and includes zero-day vulnerabilities and exploits, indicators (system artefacts or observables linked to an attack), security alerts, threat intelligence reports, as well as suggested security tool configurations .

The main goal of CTI is to enhance conventional security measures using data gathered from a variety of internal and external sources. Research on the tools, techniques, and intents of possible adversaries allows for the anticipation of future attacks.Effective CTI will deliver information that is actionable, accurate, and timely (reveals real dangers) (implies a clear course of action for threat remediation).

Open Source Intelligence (OSINT), or intelligence gathered from publicly accessible sources, can be a valuable asset to proactive CTI by warning businesses of hazards they were previously unaware of.The internet hacker community is one new source of OSINT data. The online hacker community, which consists of hacker forums, Dark-Net Markets, carding shops, and Internet-RelayChat (IRC), allows millions of hackers from many geopolitical zones such as China, the United States, and Russia to share destructive tools and expertise.[**?**]

Online hacker forums can be a useful and innovative source of data for developing proactive and thorough CTI and malware research. In an effort to identify important hackers and gather reliable cyber threat intelligence from them, many cybersecurity experts concentrate on hacker-centered research on cybercrime.

Key hackers only make up a small portion of underground forum users, despite the enormous data volume of these communities. Key hackers must be carefully analysed, which takes a lot of time and knowledge. Finding the main hackers and then mining emerging cyber threats is one technique to solve issues in the face of such sophisticated network attack and defence state

Our Project will focus on obtaining cyber threat intelligence from open source hacker forums and In this work, we present an automatic method for identifying key hackers.

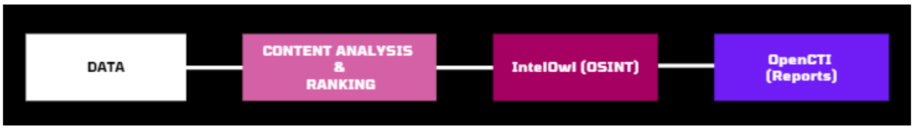
1.3 Objective of the Project

To create a pipeline that follows a four-phase paradigm, starting with information gathering from multiple hacker forums and pre-processing the data gathered; followed by Content analysis of the data, Content-based approach analyze user data based on selected evaluation metrics, such as activity and content quality; based on the content analysis , ranking of the top hackers is done ; and finally provide intelligence about the key hackers.

Chapter 2

System Architecture

2.0.1 PipeLine



Data : The information from anonymous forums and user interaction is gathered in this part. All discussions take the form of threads in underground forums (i.e., one user starts a thread and makes a post, to which other users respond, discussing various hacker-related information given by community members).

Prior to collecting any posts, we first collect all of the threads from the forum, together with each post's username, profile, content, and order.The data from the crawled raw sources are not well-formatted. Here, we undertake data preprocessing to enhance the text analysis. To maintain consistency in the data format, we start by changing all of the data to lowercase.Second, we remove punctuation and non-ASCII characters.

Content Analysis & Ranking: Content analysis is used to identify the existence of specific words, topics, or concepts in a given set of qualitative data (i.e. text). It helps to quantify and examine the occurrence, significance, and connections ofÂ specific words, themes, or concepts using content analysis.

We constructed a user evaluation metric system based on CA, and extract some features from the collected data as usersâ evaluation metric.Based on the weight of each metrics we determined the ranks, a higher rank indicates more malicious intent.

Intel Owl(OSINT): The next phase is to gather intelligence on the important actors to take the necessary steps to stop the attacks after the ranking of the key actors has been determined. IntelOwl is integrated into the pipeline in order to acquire this intelligence. Intel Owl is an Open Source Intelligence (OSINT) solution for obtaining threat intelligence data about a given file, IP address, or domain through a single API at scale.

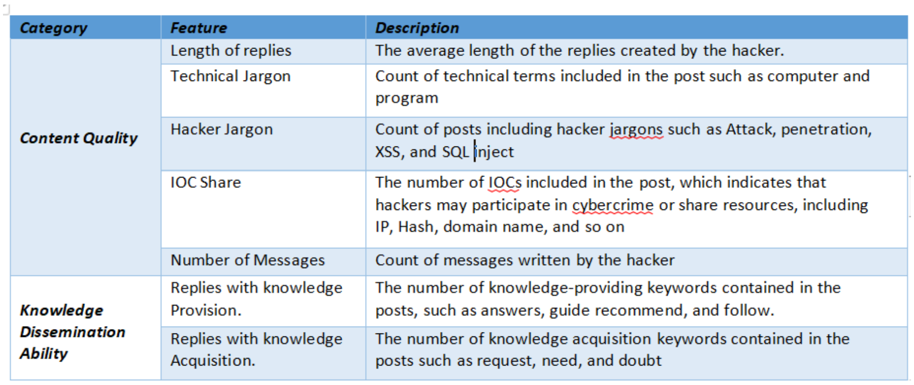
OpenCTI (Reports): The output generated by jobs executed by IntelOwl is then sent to OpenCTI for visualisation. New relations may be inferred from existing ones to facilitate the interpretation and the presenting of this information once data has been collected and processed by the analysts within OpenCTI. This allows the analysts to extract and leverage meaningful knowledge from the raw data.

2.0.2 Content Analysis

Content analysis is a research tool used to determine the presence of certain words, themes, or concepts within some given qualitative data (i.e. text). Using content analysis, researchers can quantify and analyze the presence, meanings, and relationships of such certain words, themes, or concepts.Â

In order to dig out the relevant features and behaviors of key hackers, there have been various works to explore and study the usersâ characteristics of underground forums or online forums.Â As shown in the Table, we summarize the common features. The related works mainly portray users from three aspects, including activity, content quality, and knowledge dissemination ability.

Activity is reflected by the number of posts, the more active the user, the more the number of replies and threads in the forums. Users with high-quality speeches have longer posts, and also involve a lot of hacker jargons, technical jargons, and threat intelligence. In addition, usersâ interaction is usually along with knowledge transfer (knowledge acquisition and provision), and key hackers are often the core of knowledge transfer.



Category	Feature	Description
Content Quality	Length of replies	The average length of the replies created by the hacker.
	Technical Jargon	Count of technical terms included in the post such as computer and program
	Hacker Jargon	Count of posts including hacker jargons such as Attack, penetration, XSS, and SQL Inject
	IOC Share	The number of IOCs included in the post, which indicates that hackers may participate in cybercrime or share resources, including IP, Hash, domain name, and so on
	Number of Messages	Count of messages written by the hacker
Knowledge Dissemination Ability	Replies with knowledge Provision.	The number of knowledge-providing keywords contained in the posts, such as answers, guide recommend, and follow.
	Replies with knowledge Acquisition.	The number of knowledge acquisition keywords contained in the posts such as request, need, and doubt

Indicators of Compromise (IOC) is a very important criteria to determine the key hackers in a hacker forum. The indicatiors of compromise we have considered in the scope of this project are:

- IPV4 and IPV4 CIDR addresses shared by the users in the forum
- IPV6 addresses shared by the users in the forum
- URLs shared by the users in the forum
- Domains shared by the users in the forum
- Email addresses shared by the users in the forum

2.0.3 Ranking

We utilised two factors to identify the key hackers: content quality[6] [1] and knowledge dissemination abilities. Longer messages from users who speak well contain more technical jargon, threat intelligence, and hacking jargons. Additionally, knowledge transfer (Â acquisition and provision of knowledge) generally involves user contact, with key hackers serving as the knowledge transfer's central hub. We build a method for user evaluation metricÂ based on content analysis[2], and we select specific features from the gathered information for this purpose.

Entropy is a measurement of a data-generating function's unpredictability or diversity in the context of cyber security.Data with 100% entropy is completely random, and no meaningful patterns can be detected. It is possible to forecast future generated values using low entropy data. Calculating the entropy value could assess an event's randomness and disorder, or the degree of dispersion for some metric, in accordance with the properties of entropy. The stronger the metric's influence (weight) on the overall evaluation, the more discrete the metric. As a result, we employ the entropy weight method32 to give several metrics weights in order to produce a thorough evaluation for each user.

The calculation process is as follows:

- **Data Standardization**:
- we use minimum and maximum method to standardize the data since the measurement units of various indicators are not uniform, and the data dimensions and data levels are quite different[4].

$$x_{ij} = \frac{x_{ij} - minx_j}{maxx_j - minx_j}$$

- In the above equation x_{ij}represents the jth metric of the ith user, $maxx_j$ is the maximum value of the jth metric, and $minx_j$ is the minimum value
  Calculate the information entropy of the jth metric

$$e_j = -k \sum_{i=1}^{n} p_{ij} ln\left(p_{ij}\right)$$

where k=1/ln(n) and p_{ij}=$x_{ij}/\sum_{i=1}^{n} x_{ij}$ Calculate the weight of each metric where m is the count of metrics.

$$w_j = \frac{1 - e_j}{\sum_{j=1}^{m}\left(1 - e_j\right)}$$

- Perform a weighted summation of the weights of each metric to generate a comprehensive evaluation of underground forum users as

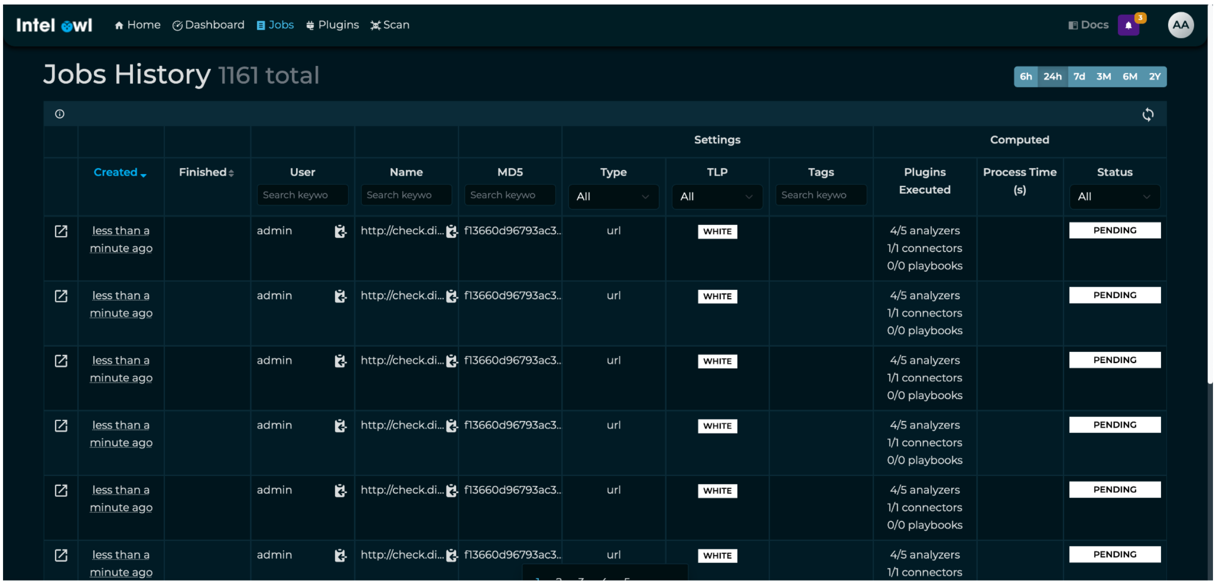
$$U_i = \sum_{j=1}^{m} x_{ij} \cdot w_j$$

.

2.0.4 OSINT Visualisation

Once the rankings are obtained, the next step is to obtain intelligence on the key actors to take required actions to prevent the attacks. In order to obtain this intelligence, IntelOwl is integrated into the pipeline. Intel Owl is an Open Source Intelligence, or OSINT solution to get threat intelligence data about a specific file, an IP or a domain from a single API at scale. It integrates a number of analyzers available online and a lot of cutting-edge malware analysis tools. It is for everyone who needs a single point to query for info about a specific file or observable.

IntelOwl integration in the pipeline is done using a python script. This script uses pyintelowl library to call IntelOwl API for obtaining OSINT on the IOCs extracted from the data of top key actors. The IOCs include URLs, IPs, Domains, Email Addresses, etc.



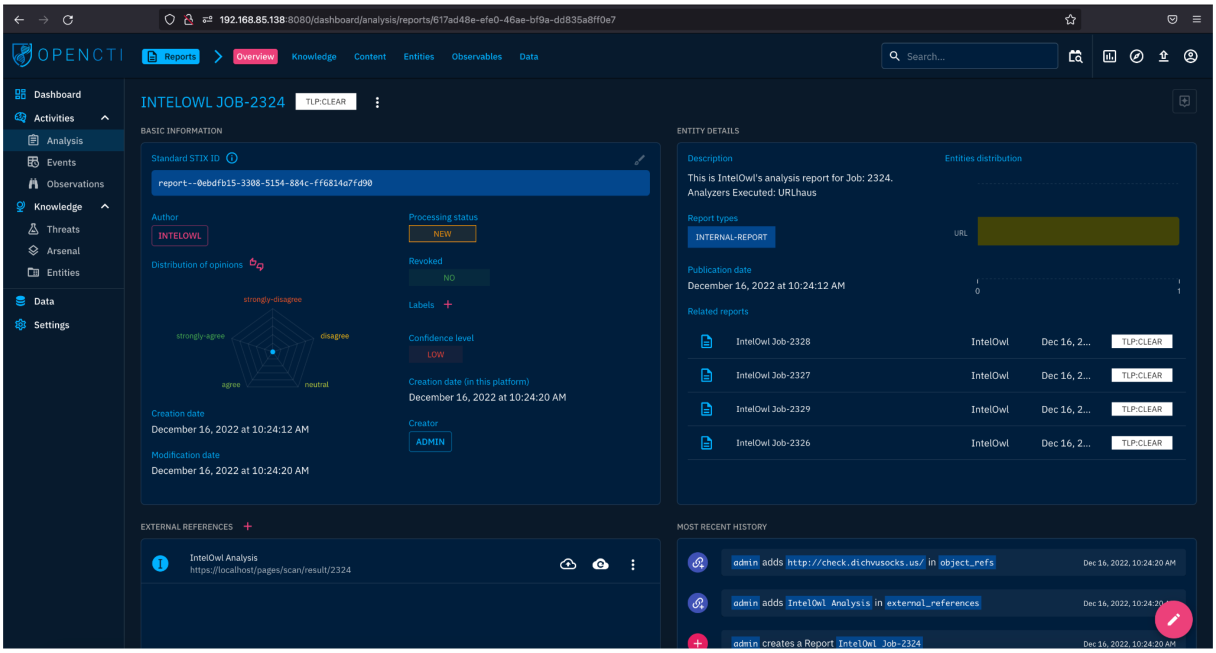
Jobs History 1161 total

							Settings			Computed	
	Created	Finished	User	Name	MD5	Type	TLP	Tags	Plugins Executed	Process Time (s)	Status
			Search keywo	Search keywo	Search keywo	All	All	Search keywo			All
	less than a minute ago		admin	http://check.di...	f13660d96793ac3..	url	WHITE		4/5 analyzers 1/1 connectors 0/0 playbooks		PENDING
	less than a minute ago		admin	http://check.di...	f13660d96793ac3..	url	WHITE		4/5 analyzers 1/1 connectors 0/0 playbooks		PENDING
	less than a minute ago		admin	http://check.di...	f13660d96793ac3..	url	WHITE		4/5 analyzers 1/1 connectors 0/0 playbooks		PENDING
	less than a minute ago		admin	http://check.di...	f13660d96793ac3..	url	WHITE		4/5 analyzers 1/1 connectors 0/0 playbooks		PENDING
	less than a minute ago		admin	http://check.di...	f13660d96793ac3..	url	WHITE		4/5 analyzers 1/1 connectors 0/0 playbooks		PENDING
	less than a minute ago		admin	http://check.di...	f13660d96793ac3..	url	WHITE		4/5 analyzers 1/1 connectors		PENDING

The output generated by jobs executed by IntelOwl is then sent to OpenCTI for visualisation. OpenCTI is an open source platform allowing organisations to manage their cyber threat intelligence knowledge and observables. It has been created in order to structure, store, organise and visualise technical and non-technical information about cyber threats. Once data has been capitalised and processed by the analysts within OpenCTI, new relations may be inferred from existing ones to facilitate the understanding and the representation of this information. This allows the analysts to extract and leverage meaningful knowledge from the raw data. In order to achieve this IntelOwl - OpenCTI connector is used in the pipeline. This connector is a part of IntelOwl which formats the data and generates a report on OpenCTI for all the jobs. This report helps in understanding the OSINT obtained for all the jobs created using python script.

The image below shows one of the reports that presents the output of IntelOwl jobs on OpenCTI. It presents information on entities, relations and knowledge and many other insights that were observed on the data.

The image below shows one of the reports that presents the output of IntelOwl jobs on OpenCTI. It presents information on entities, relations and knowledge and many other insights that were observed on the data.



192.168.85.138:8080/dashboard/analysis/reports/617ad48e-efe0-46ae-bf9a-dd835a8ff0e7

INTELOWL JOB-2324 TLP:CLEAR

BASIC INFORMATION

Standard STIX ID
report--0ebdfb15-3308-5154-884c-ff6814a7fd90

Author
INTELOWL

Distribution of opinions

Creation date
December 16, 2022 at 10:24:12 AM

Modification date
December 16, 2022 at 10:24:20 AM

Processing status
NEW

Revoked
NO

Labels +

Confidence level
LOW

Creation date (in this platform)
December 16, 2022 at 10:24:20 AM

Creator
ADMIN

ENTITY DETAILS

Description
This is IntelOwl's analysis report for Job: 2324.
Analyzers Executed: URLhaus

Entities distribution

Report types
INTERNAL-REPORT

Publication date
December 16, 2022 at 10:24:12 AM

Related reports

IntelOwl Job-2328	IntelOwl	Dec 16, 2...	TLP:CLEAR
IntelOwl Job-2327	IntelOwl	Dec 16, 2...	TLP:CLEAR
IntelOwl Job-2329	IntelOwl	Dec 16, 2...	TLP:CLEAR
IntelOwl Job-2326	IntelOwl	Dec 16, 2...	TLP:CLEAR

EXTERNAL REFERENCES +

IntelOwl Analysis
https://localhost/pages/scan/result/2324

MOST RECENT HISTORY

admin adds http://check.dichvusocks.us/ in object_refs Dec 16, 2022, 10:24:20 AM
admin adds IntelOwl Analysis in external_references Dec 16, 2022, 10:24:20
admin creates a Report IntelOwl Job-2324 Dec 16, 2022, 10:24:20 AM

Chapter 3

Tools

- **Google Collab**: To run python code for data analysis of bigger data sets with no set-up which gives free access to Google computing resources such as GPUs and TPUs.
- **Intel Owl** : is an Open Source Intelligence, or OSINT solution to get threat intelligence data about a specific file, an IP or a domain from a single API at scale.
- **Open CTI** : OpenCTI is an open source platform to Store, organize, visualize about cyber threats.
- **Jupyter Notebook** : Jupyter Notebook allows users to compile all aspects of a data project in one place making it easier to show the entire process of a project to your intended audience.Through the web-based application, users can create data visualizations and other components of a project to share with others via the platform.

Chapter 4

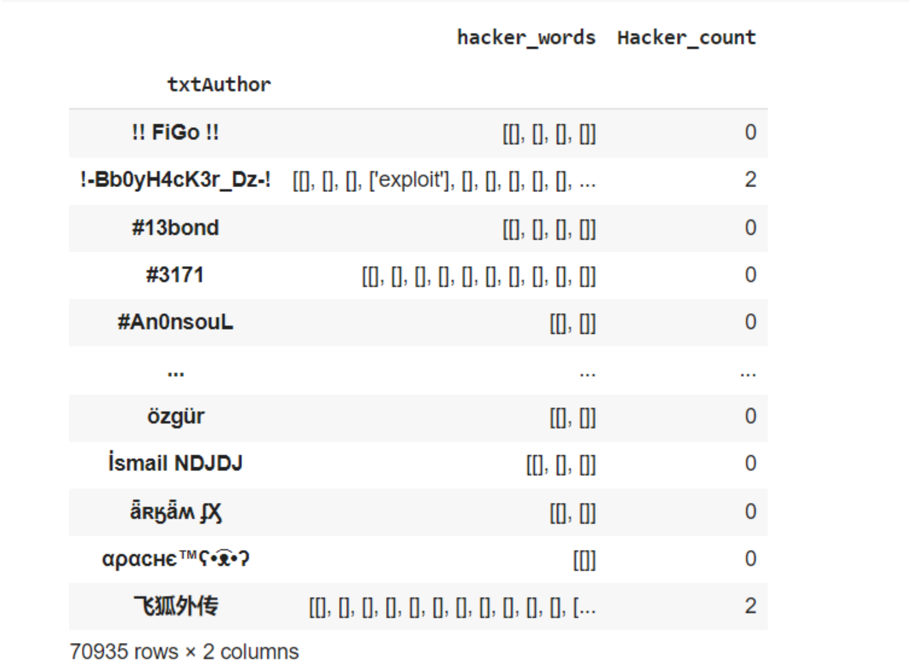
Results

We identified the key hackers using various Content Analysis metrics:

4.1 Content Analysis

Hacker Words:

Identifying and counting hacker words for each user:

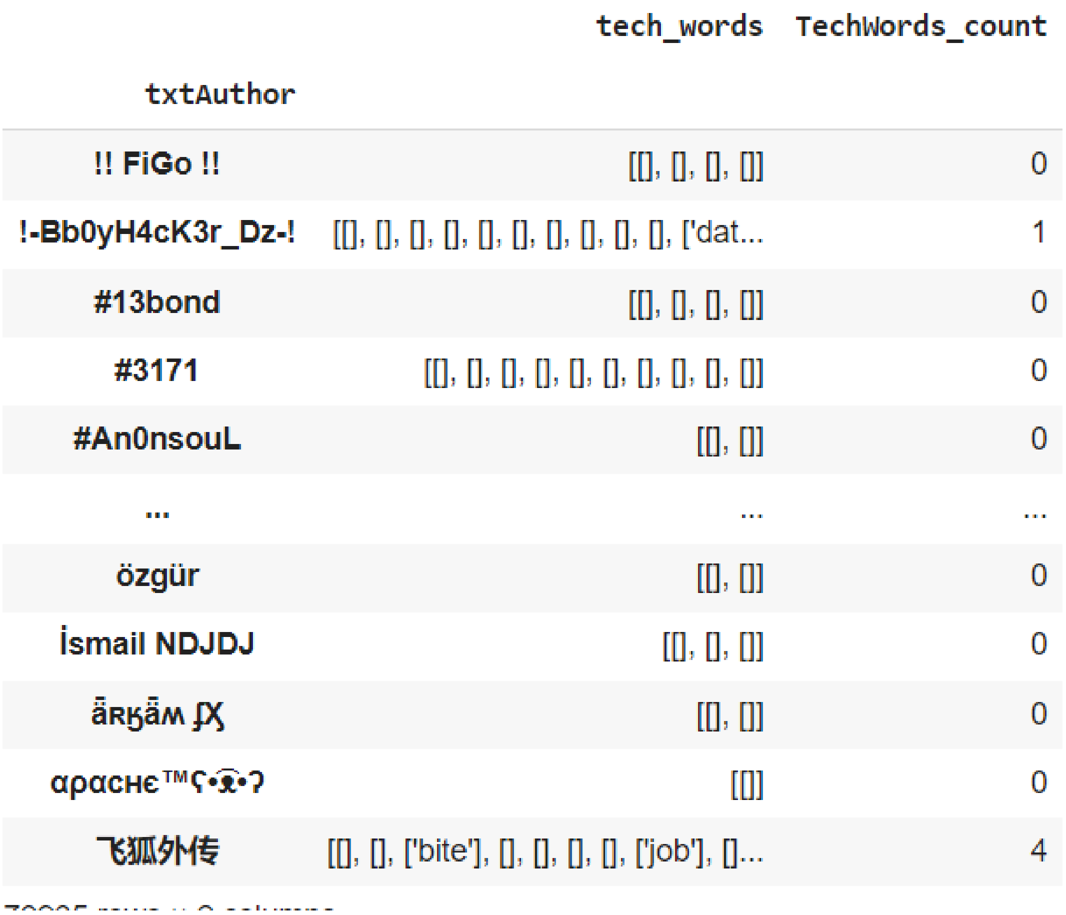


	hacker_words	Hacker_count
txtAuthor		
!! FiGo !!	[[], [], [], []]	0
!-Bb0yH4cK3r_Dz-!	[[], [], [], ['exploit'], [], [], [], [], [], ...	2
#13bond	[[], [], [], []]	0
#3171	[[], [], [], [], [], [], [], [], [], []]	0
#An0nsouL	[[], []]	0
...	...	...
özgür	[[], []]	0
İsmail NDJDJ	[[], [], []]	0
ăʀҕăм ʆχ	[[], []]	0
αρacнє™ʕ•ᴥ•ʔ	[[]]	0
飞狐外传	[[], [], [], [], [], [], [], [], [], [], [], [...	2

70935 rows × 2 columns

Technical Jargons:

Identifying and counting the Technical Jargons

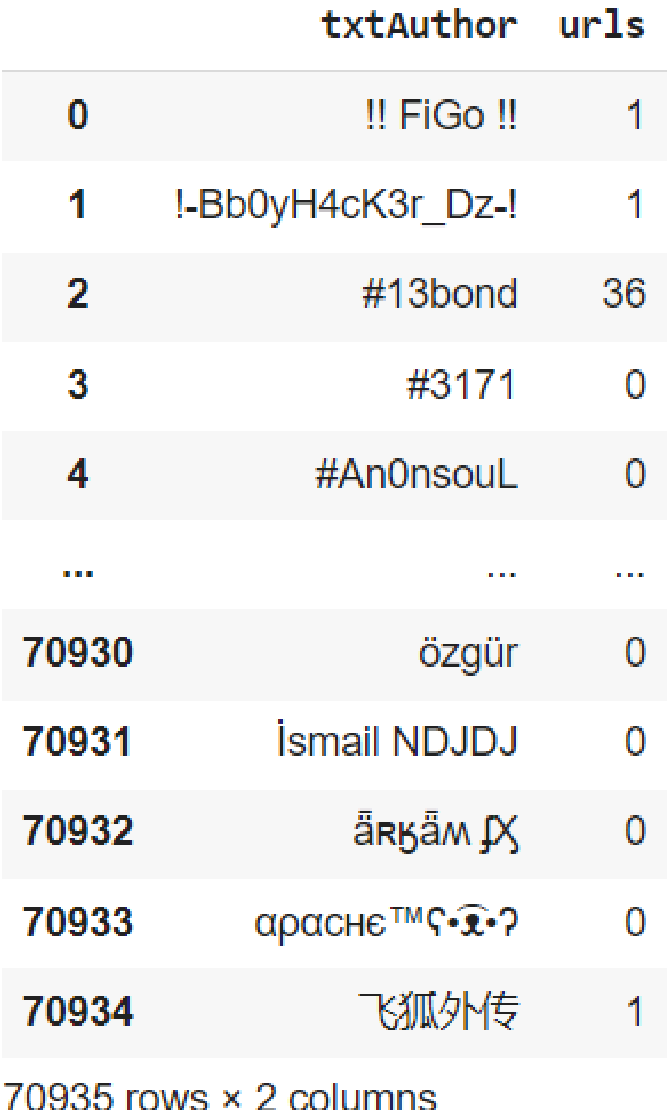


	tech_words	TechWords_count
txtAuthor		
!! FiGo !!	[[], [], [], []]	0
!-Bb0yH4cK3r_Dz-!	[[], [], [], [], [], [], [], [], [], [], ['dat...	1
#13bond	[[], [], [], []]	0
#3171	[[], [], [], [], [], [], [], [], [], []]	0
#An0nsouL	[[], []]	0
...	...	...
özgür	[[], []]	0
İsmail NDJDJ	[[], [], []]	0
ăʀӄăм ʆҲ	[[], []]	0
αρacнє™ʕ•ᴥ•ʔ	[[]]	0
飞狐外传	[[], [], ['bite'], [], [], [], [], ['job'], []...	4

70085 rows × 2 columns

URLs:

Counting and identifying the URLs which can be considered as an Indicator of Compromise

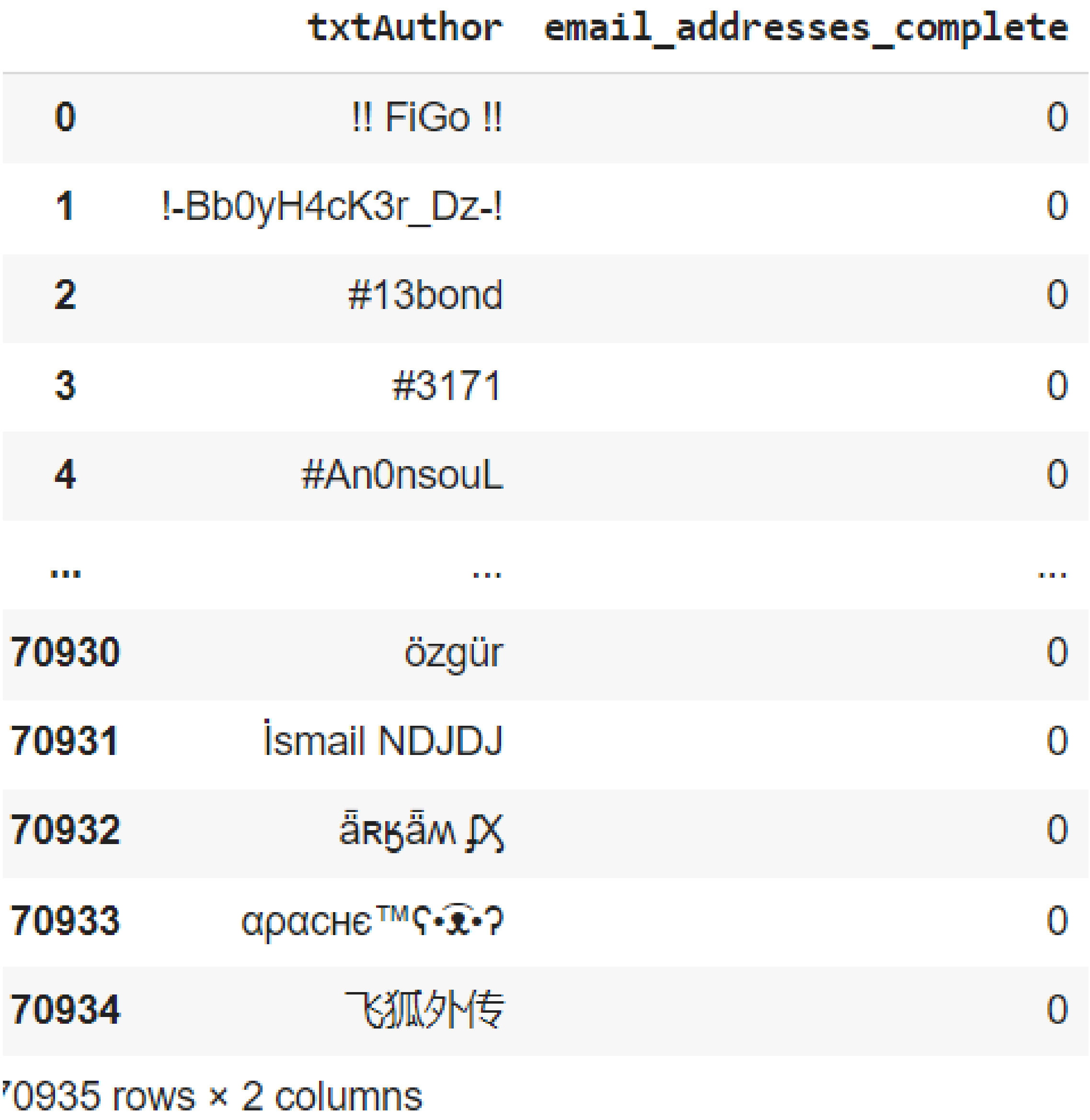


	txtAuthor	urls
0	!! FiGo !!	1
1	!-Bb0yH4cK3r_Dz-!	1
2	#13bond	36
3	#3171	0
4	#An0nsouL	0
...	...	...
70930	özgür	0
70931	İsmail NDJDJ	0
70932	ǟʀӄǟʍ ʄӼ	0
70933	αρα¢нє™ʕ•̫͡•ʔ	0
70934	飞狐外传	1

70935 rows × 2 columns

Email Addresses

The number of email addresses identified for each user

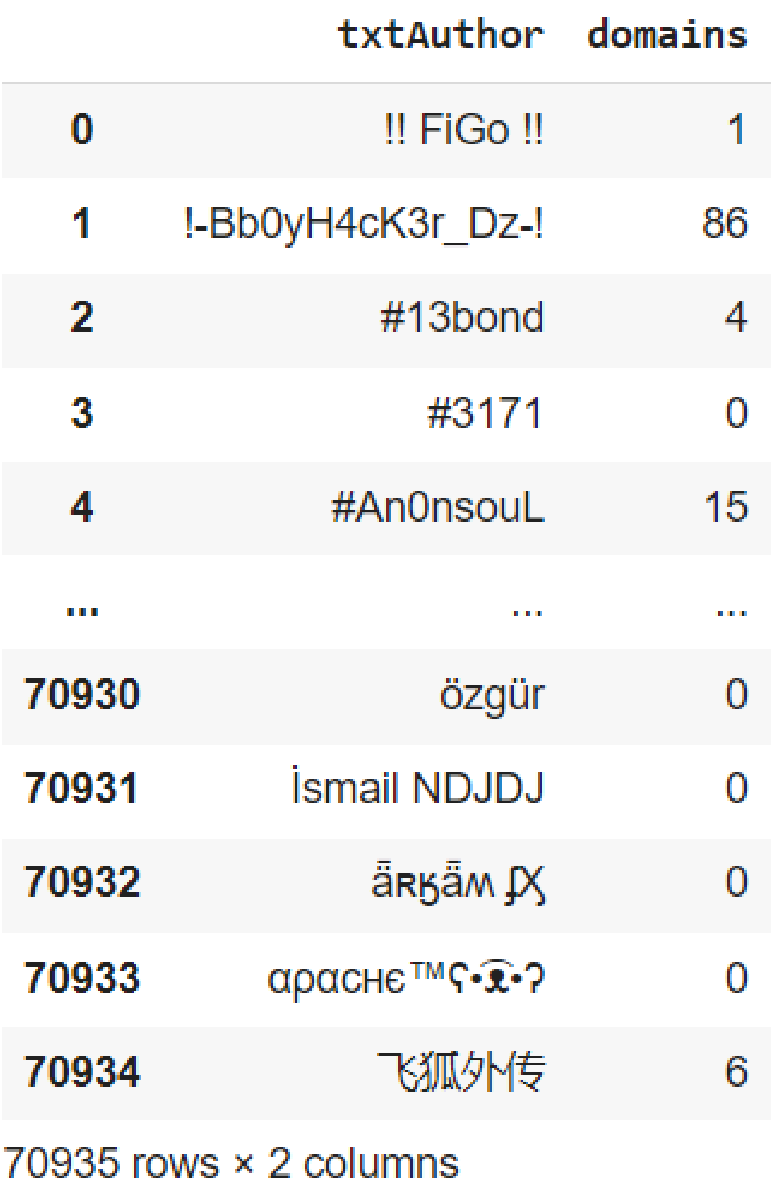


	txtAuthor	email_addresses_complete
0	!! FiGo !!	0
1	!-Bb0yH4cK3r_Dz-!	0
2	#13bond	0
3	#3171	0
4	#An0nsouL	0
...	...	...
70930	özgür	0
70931	İsmail NDJDJ	0
70932	ǟʀӄǟʍ ʄҲ	0
70933	αραcнє™ʕ•ᴥ•ʔ	0
70934	飞狐外传	0

70935 rows × 2 columns

Domains:

The number of domains identified for each user:

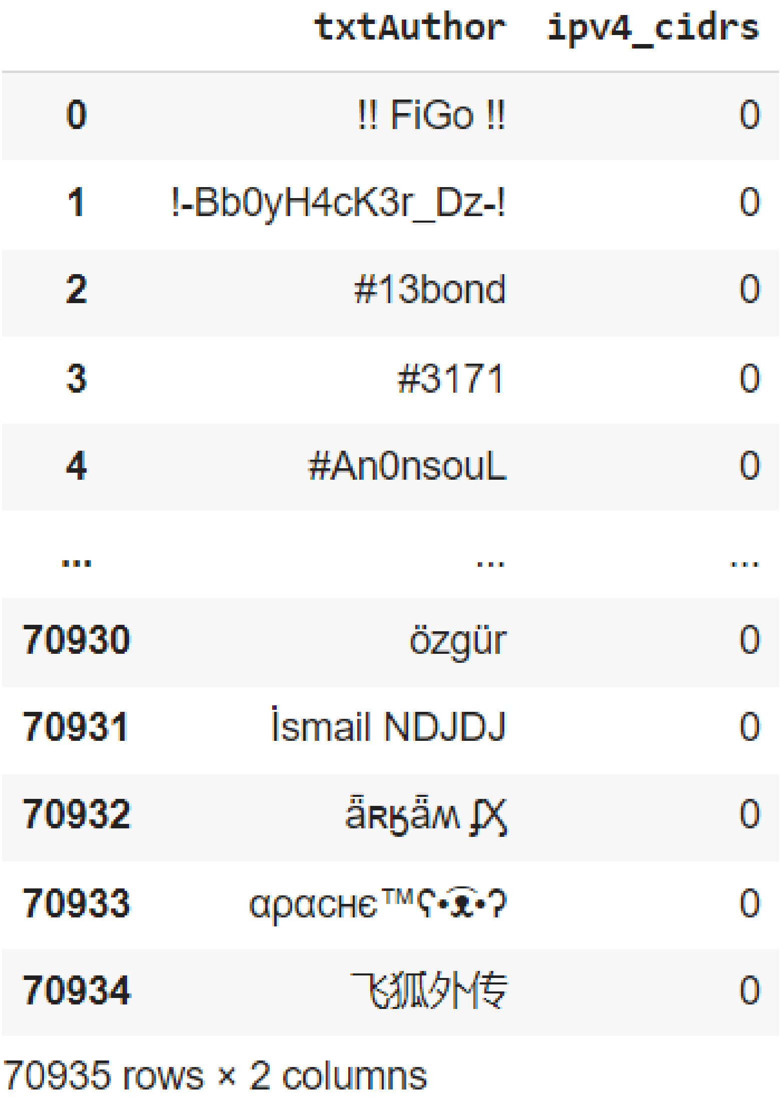


	txtAuthor	domains
0	!! FiGo !!	1
1	!-Bb0yH4cK3r_Dz-!	86
2	#13bond	4
3	#3171	0
4	#An0nsouL	15
...	...	...
70930	özgür	0
70931	İsmail NDJDJ	0
70932	ǟʀӄǟʍ ʆX	0
70933	αρacнє™ʕ•͡ᴥ•ʔ	0
70934	飞狐外传	6

70935 rows × 2 columns

IPV4 CIDR:

The number of IPV4 CIDRs obtained for each user

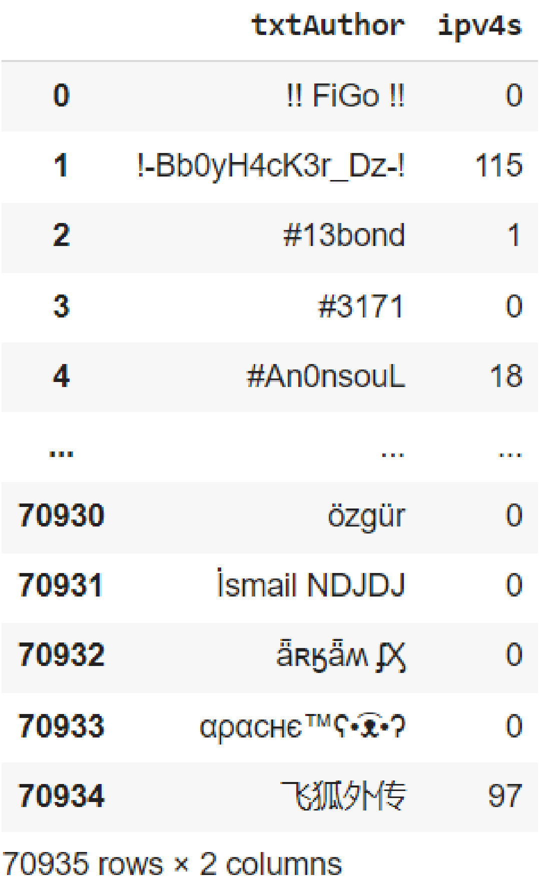


	txtAuthor	ipv4_cidrs
0	!! FiGo !!	0
1	!-Bb0yH4cK3r_Dz-!	0
2	#13bond	0
3	#3171	0
4	#An0nsouL	0
...	...	...
70930	özgür	0
70931	İsmail NDJDJ	0
70932	ǟʀӄǟʍ ʄӼ	0
70933	αρacнє™ʕ•ᴥ•ʔ	0
70934	飞狐外传	0

70935 rows × 2 columns

IPV4:

The number of IPV4 Addresses obtained for each user

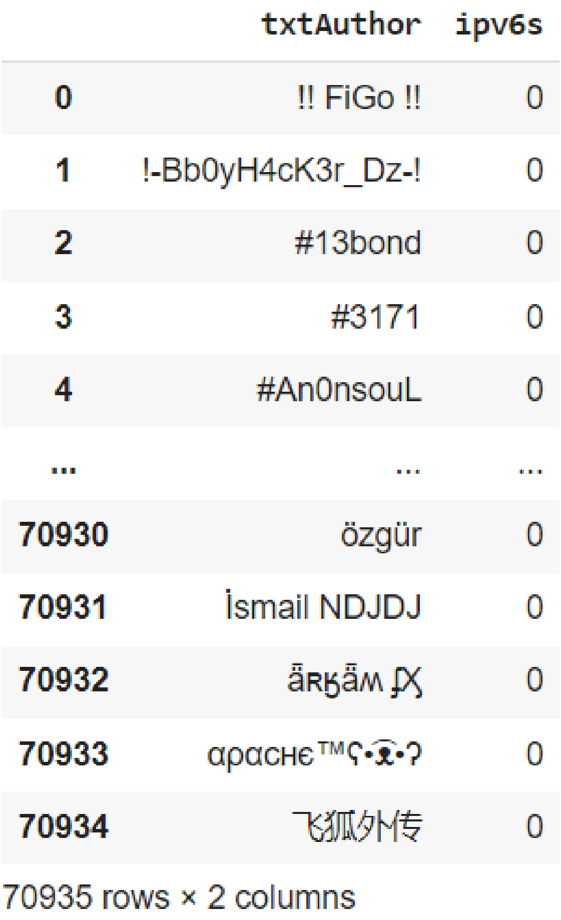


	txtAuthor	ipv4s
0	!! FiGo !!	0
1	!-Bb0yH4cK3r_Dz-!	115
2	#13bond	1
3	#3171	0
4	#An0nsouL	18
...	...	...
70930	özgür	0
70931	İsmail NDJDJ	0
70932	ā̈ʀҕā̈м ʆҲ	0
70933	αρacнє™ʕ•ᴥ•ʔ	0
70934	飞狐外传	97

70935 rows × 2 columns

IPV6:

The number of IPV6 Addresses obtained for each user

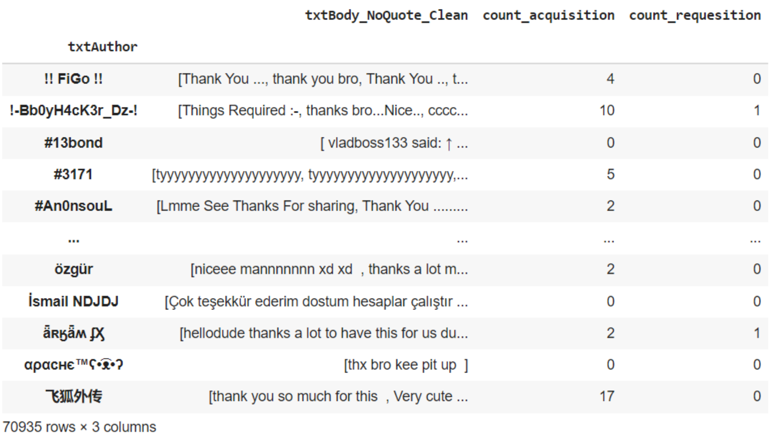


	txtAuthor	ipv6s
0	!! FiGo !!	0
1	!-Bb0yH4cK3r_Dz-!	0
2	#13bond	0
3	#3171	0
4	#An0nsouL	0
...	...	...
70930	özgür	0
70931	İsmail NDJDJ	0
70932	ăʀȣăм ʆϪ	0
70933	αρacнє™ʕ•ᴥ•ʔ	0
70934	飞狐外传	0

70935 rows × 2 columns

Knowledge Decimation:

The number of knowledge Acquisition and Requisition keywords used by each user

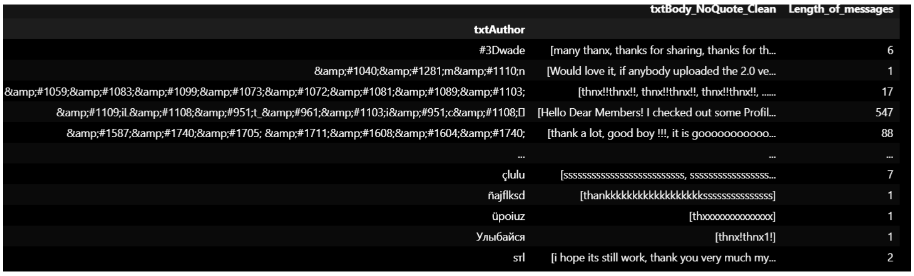


	txtBody_NoQuote_Clean	count_acquisition	count_requesition
txtAuthor			
!! FiGo !!	[Thank You ..., thank you bro, Thank You .., t...	4	0
!-Bb0yH4cK3r_Dz-!	[Things Required :-, thanks bro...Nice.., cccc...	10	1
#13bond	[vladboss133 said: ↑ ...	0	0
#3171	[tyyyyyyyyyyyyyyyyyy, tyyyyyyyyyyyyyyyyyy,...	5	0
#An0nsouL	[Lmme See Thanks For sharing, Thank You	2	0
...	...	...	...
özgür	[niceee mannnnnnn xd xd , thanks a lot m...	2	0
İsmail NDJDJ	[Çok teşekkür ederim dostum hesaplar çalıştır ...	0	0
ăʀȣăм ʆϪ	[hellodude thanks a lot to have this for us du...	2	1
αρacнє™ʕ•ᴥ•ʔ	[thx bro kee pit up]	0	0
飞狐外传	[thank you so much for this , Very cute ...	17	0

70935 rows × 3 columns

Number of Messages for each user

The Number of messages for each user is identified:

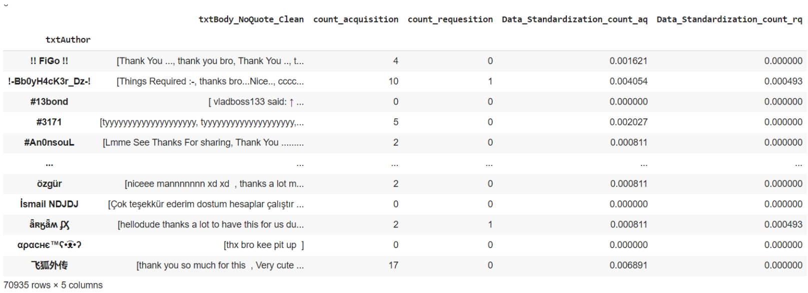


	txtBody_NoQuote_Clean	Length_of_messages
txtAuthor		
#3Dwade	[many thanx, thanks for sharing, thanks for th...	6
&#1040;&#1281;m&#1110;n	[Would love it, if anybody uploaded the 2.0 ve...	1
&#1059;&#1083;&#1099;&#1073;&#1072;&#1081;&#1089;&#1103;	[thnx!!thnx!!, thnx!!thnx!!, thnx!!thnx!!, ...	17
&#1109;iL&#1108;&#951;t_&#961;&#1103;i&#951;c&#1108;􏿾	[Hello Dear Members! I checked out some Profil...	547
&#1587;&#1740;&#1705; &#1711;&#1608;&#1604;&#1740;	[thank a lot, good boy !!!, it is goooooooooo...	88
...	...	...
çlulu	[sssssssssssssssssssssssss, ssssssssssssssss...	7
ñajflksd	[thankkkkkkkkkkkkkkkkkksssssssssssssss]	1
üpoiuz	[thxxxxxxxxxxxxxx]	1
Улыбайся	[thnx!thnx1!]	1
sтl	[i hope its still work, thank you very much my...	2

4.2 Ranking

Data Standardization:

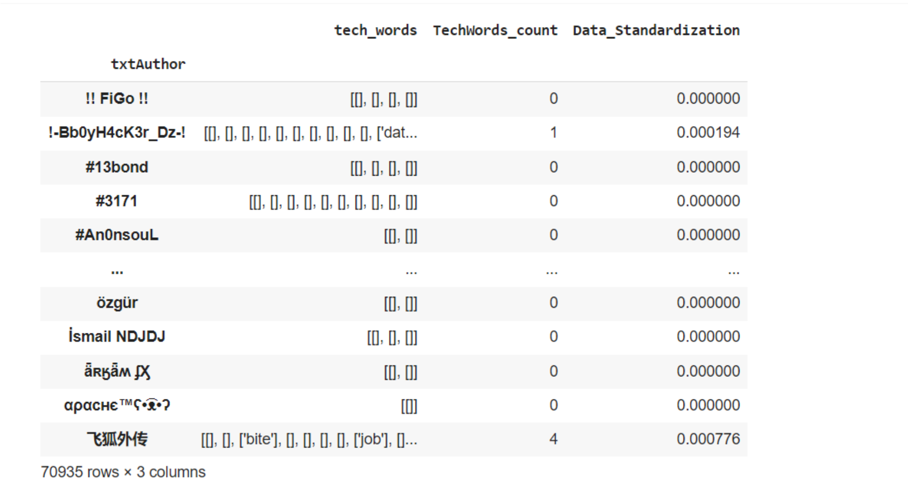
Data Standardization is done on all the metrics



	txtBody_NoQuote_Clean	count_acquisition	count_requesition	Data_Standardization_count_aq	Data_Standardization_count_rq
txtAuthor					
!! FiGo !!	[Thank You ..., thank you bro, Thank You .., t...	4	0	0.001621	0.000000
!-Bb0yH4cK3r_Dz-!	[Things Required :-, thanks bro...Nice.., cccc...	10	1	0.004054	0.000493
#13bond	[vladboss133 said: ↑ ...	0	0	0.000000	0.000000
#3171	[tyyyyyyyyyyyyyyyyyyyy, tyyyyyyyyyyyyyyyyyyyy,...	5	0	0.002027	0.000000
#An0nsouL	[Lmme See Thanks For sharing, Thank You	2	0	0.000811	0.000000
...	...	...	...	...	...
özgür	[niceee mannnnnnn xd xd , thanks a lot m...	2	0	0.000811	0.000000
İsmail NDJDJ	[Çok teşekkür ederim dostum hesaplar çalıştır ...	0	0	0.000000	0.000000
ăʀɮăм ʆҲ	[hellodude thanks a lot to have this for us du...	2	1	0.000811	0.000493
αρacнє™ʕ•ᴥ•ʔ	[thx bro kee pit up]	0	0	0.000000	0.000000
飞狐外传	[thank you so much for this , Very cute ...	17	0	0.006891	0.000000

70935 rows × 5 columns

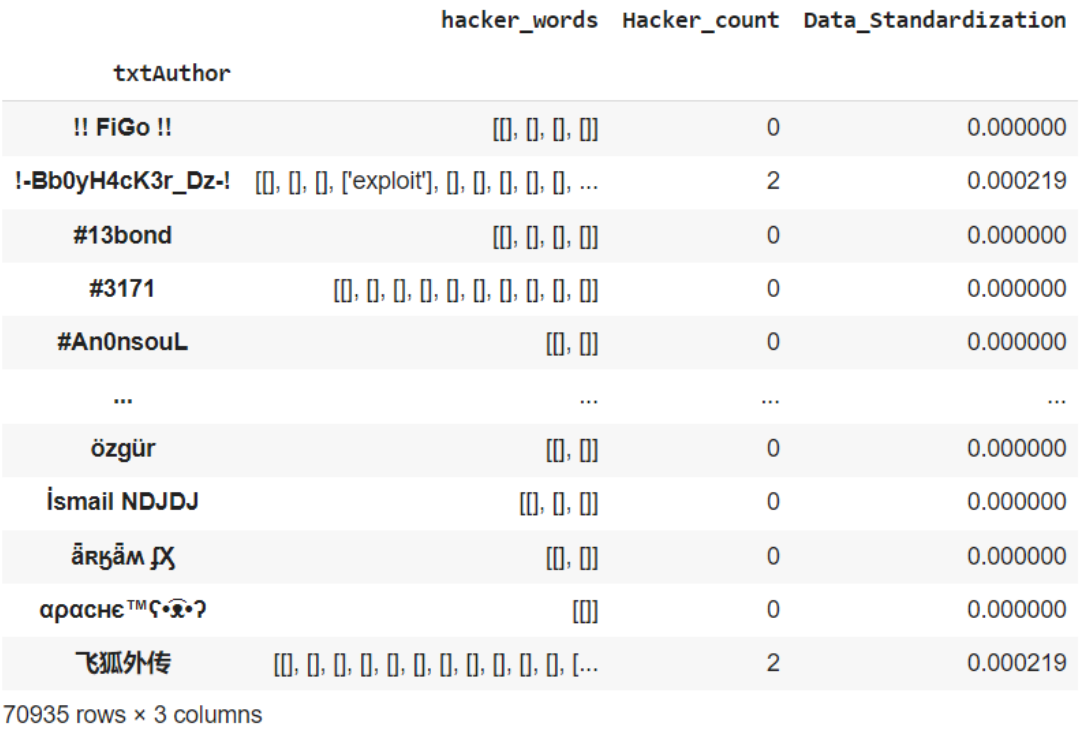
Knowledge Acquisition and Requisition Keywords



	tech_words	TechWords_count	Data_Standardization
txtAuthor			
!! FiGo !!	[[], [], [], []]	0	0.000000
!-Bb0yH4cK3r_Dz-!	[[], [], [], [], [], [], [], [], [], [], ['dat...	1	0.000194
#13bond	[[], [], [], []]	0	0.000000
#3171	[[], [], [], [], [], [], [], [], [], []]	0	0.000000
#An0nsouL	[[], []]	0	0.000000
...	...	...	...
özgür	[[], []]	0	0.000000
İsmail NDJDJ	[[], [], []]	0	0.000000
ăʀɮăм ʆҲ	[[], []]	0	0.000000
αρacнє™ʕ•ᴥ•ʔ	[[]]	0	0.000000
飞狐外传	[[], [], ['bite'], [], [], [], [], ['job'], []...	4	0.000776

70935 rows × 3 columns

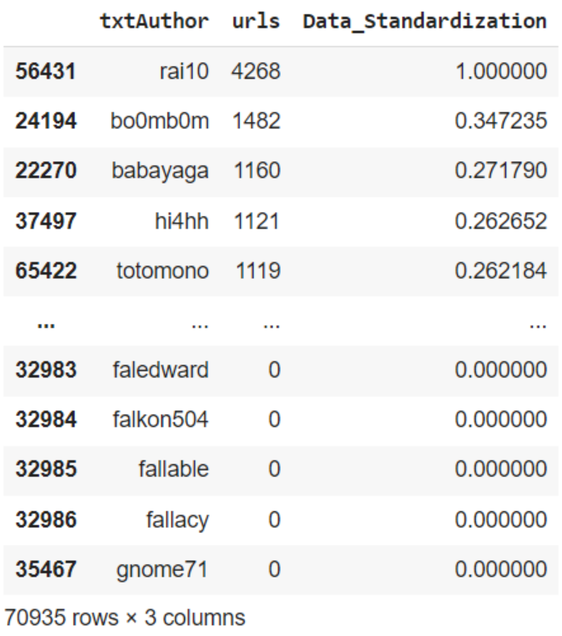
Technical Jargon Used by the hackers



	hacker_words	Hacker_count	Data_Standardization
txtAuthor			
!! FiGo !!	[[], [], [], []]	0	0.000000
!-Bb0yH4cK3r_Dz-!	[[], [], [], ['exploit'], [], [], [], [], [], ...	2	0.000219
#13bond	[[], [], [], []]	0	0.000000
#3171	[[], [], [], [], [], [], [], [], [], []]	0	0.000000
#An0nsouL	[[], []]	0	0.000000
...	...	...	...
özgür	[[], []]	0	0.000000
İsmail NDJDJ	[[], [], []]	0	0.000000
ăʀƙăʍ ʆҲ	[[], []]	0	0.000000
αρα¢нє™ʕ•ᴥ•ʔ	[[]]	0	0.000000
飞狐外传	[[], [], [], [], [], [], [], [], [], [], [], [...	2	0.000219

70935 rows × 3 columns

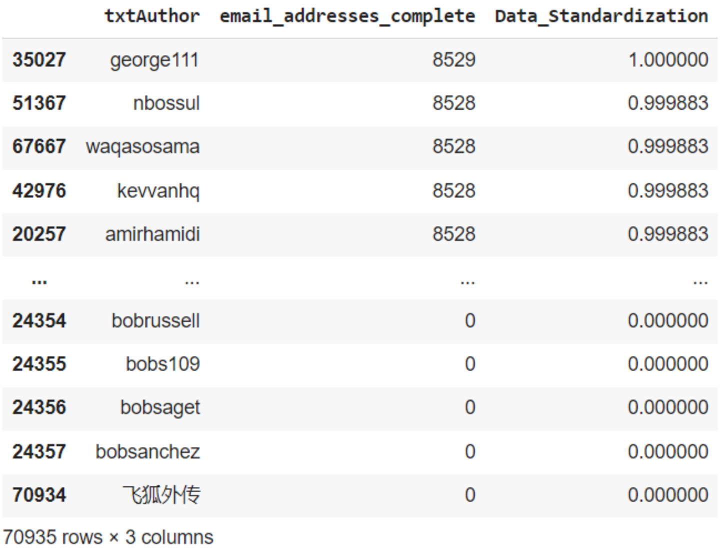
Hacker Jargon used by the hackers



	txtAuthor	urls	Data_Standardization
56431	rai10	4268	1.000000
24194	bo0mb0m	1482	0.347235
22270	babayaga	1160	0.271790
37497	hi4hh	1121	0.262652
65422	totomono	1119	0.262184
...	...	...	...
32983	faledward	0	0.000000
32984	falkon504	0	0.000000
32985	fallable	0	0.000000
32986	fallacy	0	0.000000
35467	gnome71	0	0.000000

70935 rows × 3 columns

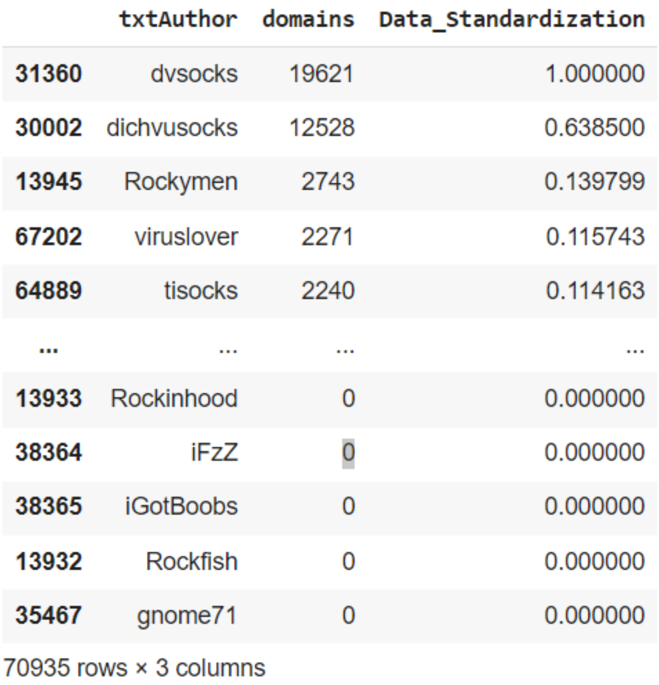
URLs shared by the hackers



	txtAuthor	email_addresses_complete	Data_Standardization
35027	george111	8529	1.000000
51367	nbossul	8528	0.999883
67667	waqasosama	8528	0.999883
42976	kevvanhq	8528	0.999883
20257	amirhamidi	8528	0.999883
...	...	...	...
24354	bobrussell	0	0.000000
24355	bobs109	0	0.000000
24356	bobsaget	0	0.000000
24357	bobsanchez	0	0.000000
70934	飞狐外传	0	0.000000

70935 rows × 3 columns

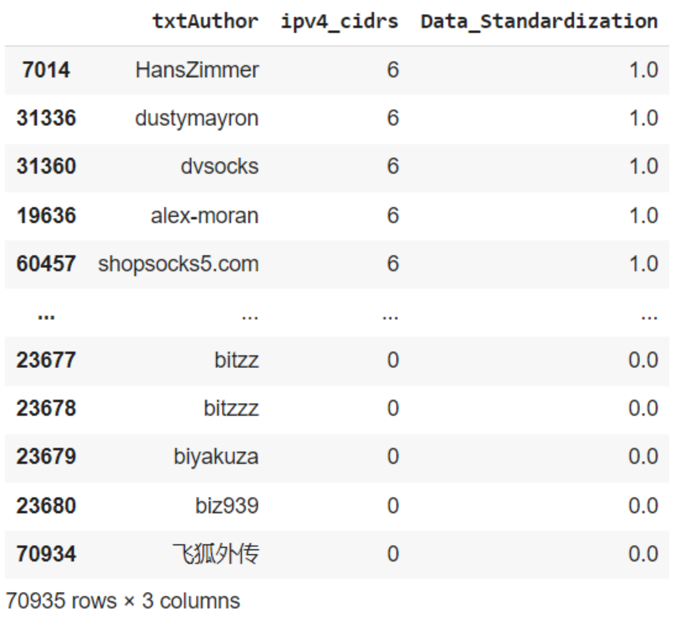
Email Addresses shared by the Hackers



	txtAuthor	domains	Data_Standardization
31360	dvsocks	19621	1.000000
30002	dichvusocks	12528	0.638500
13945	Rockymen	2743	0.139799
67202	viruslover	2271	0.115743
64889	tisocks	2240	0.114163
...	...	...	...
13933	Rockinhood	0	0.000000
38364	iFzZ	0	0.000000
38365	iGotBoobs	0	0.000000
13932	Rockfish	0	0.000000
35467	gnome71	0	0.000000

70935 rows × 3 columns

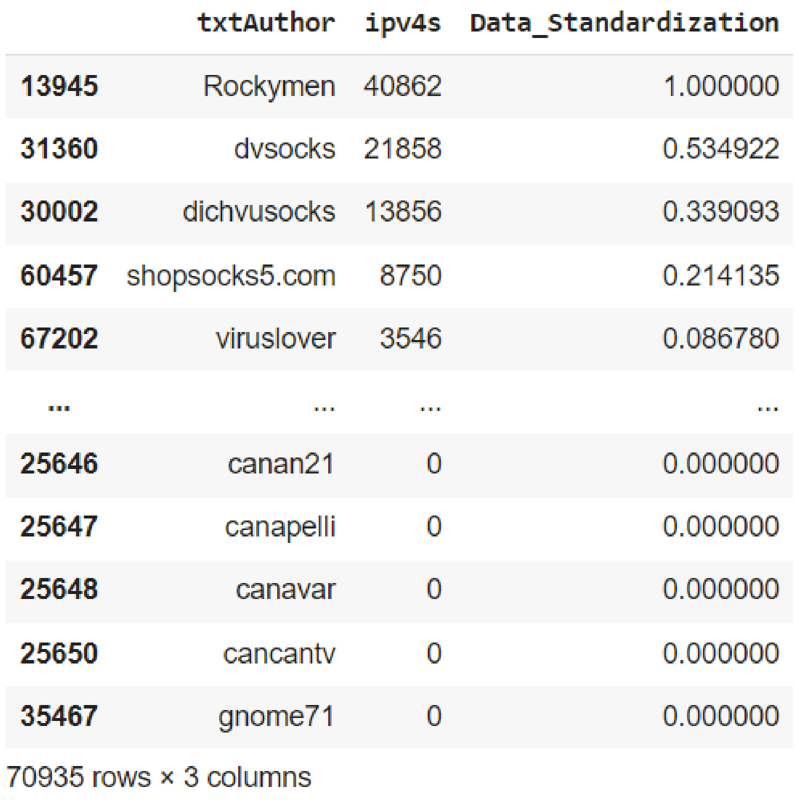
Domains shared by the hackers



	txtAuthor	ipv4_cidrs	Data_Standardization
7014	HansZimmer	6	1.0
31336	dustymayron	6	1.0
31360	dvsocks	6	1.0
19636	alex-moran	6	1.0
60457	shopsocks5.com	6	1.0
...	...	...	...
23677	bitzz	0	0.0
23678	bitzzz	0	0.0
23679	biyakuza	0	0.0
23680	biz939	0	0.0
70934	飞狐外传	0	0.0

70935 rows × 3 columns

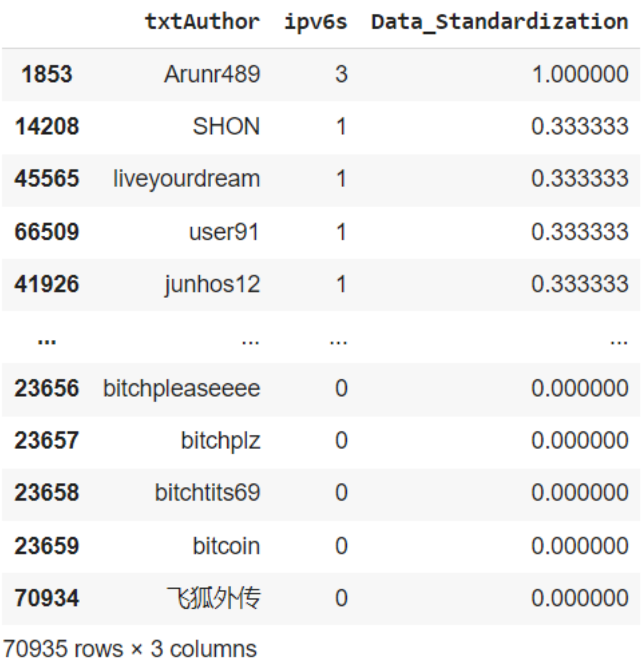
IPv4 CIDR Addresses shared by the hackers



	txtAuthor	ipv4s	Data_Standardization
13945	Rockymen	40862	1.000000
31360	dvsocks	21858	0.534922
30002	dichvusocks	13856	0.339093
60457	shopsocks5.com	8750	0.214135
67202	viruslover	3546	0.086780
...	...	...	...
25646	canan21	0	0.000000
25647	canapelli	0	0.000000
25648	canavar	0	0.000000
25650	cancantv	0	0.000000
35467	gnome71	0	0.000000

70935 rows × 3 columns

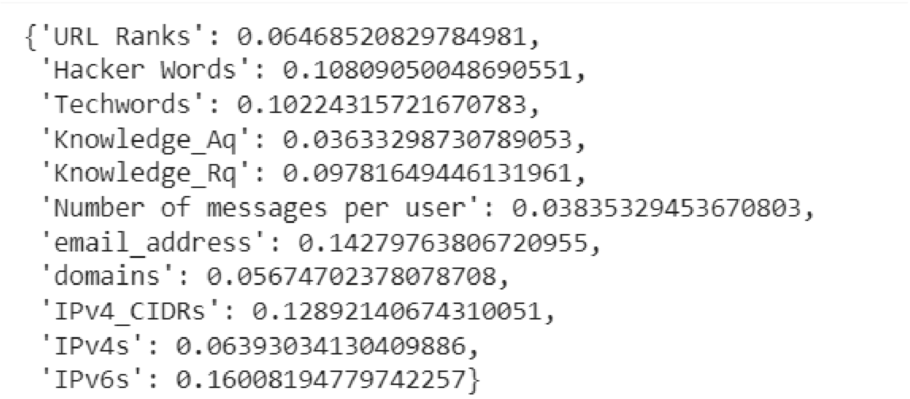
IPv6 Addresses shared by the hackers



	txtAuthor	ipv6s	Data_Standardization
1853	Arunr489	3	1.000000
14208	SHON	1	0.333333
45565	liveyourdream	1	0.333333
66509	user91	1	0.333333
41926	junhos12	1	0.333333
...	...	...	...
23656	bitchpleaseeee	0	0.000000
23657	bitchplz	0	0.000000
23658	bitchtits69	0	0.000000
23659	bitcoin	0	0.000000
70934	飞狐外传	0	0.000000

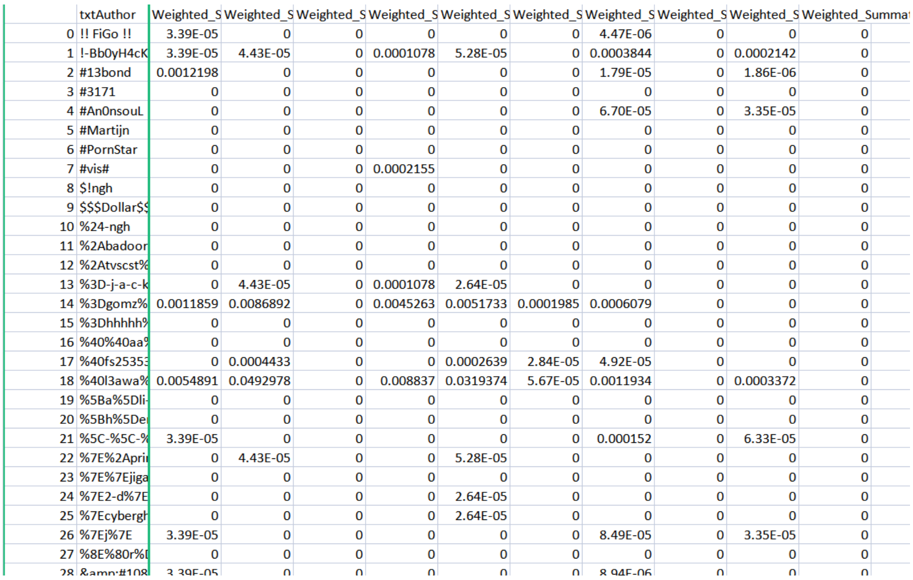
70935 rows × 3 columns

Entropy of each Metric



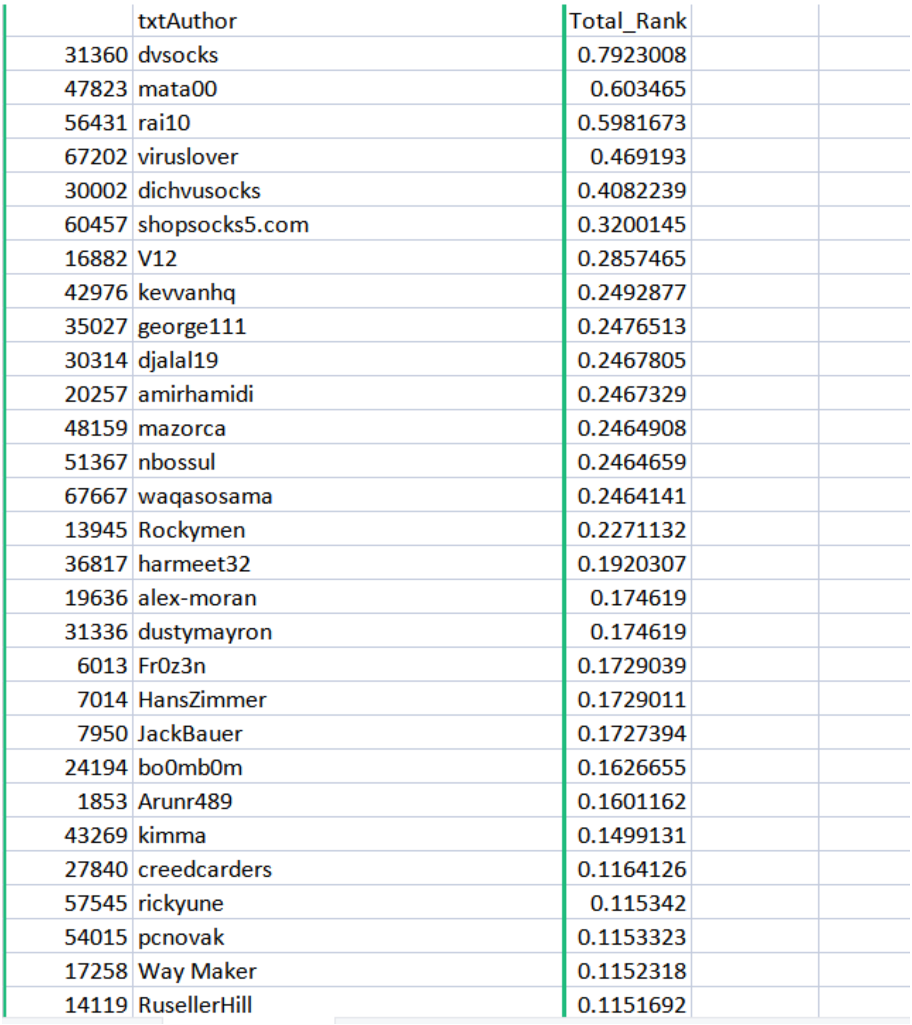
```
{'URL Ranks': 0.06468520829784981,
 'Hacker Words': 0.10809050048690551,
 'Techwords': 0.10224315721670783,
 'Knowledge_Aq': 0.03633298730789053,
 'Knowledge_Rq': 0.09781649446131961,
 'Number of messages per user': 0.03835329453670803,
 'email_address': 0.14279763806720955,
 'domains': 0.05674702378078708,
 'IPv4_CIDRs': 0.12892140674310051,
 'IPv4s': 0.06393034130409886,
 'IPv6s': 0.16008194779742257}
```

Weight of each Metric



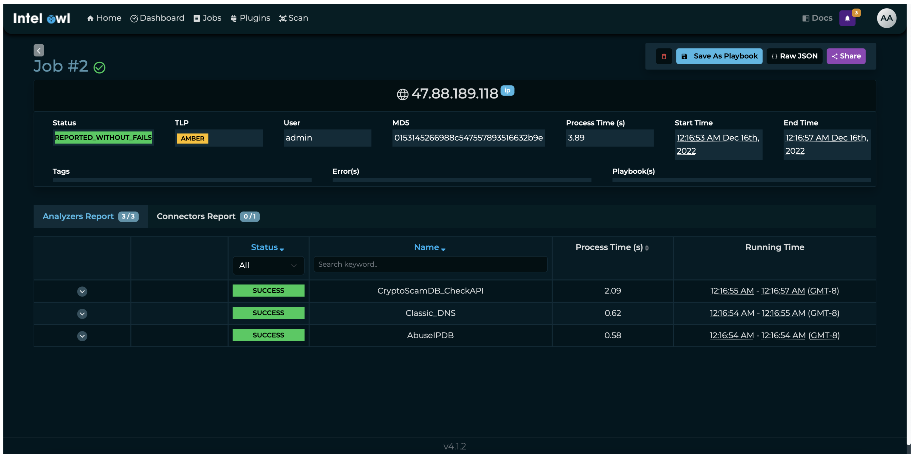
	txtAuthor	Weighted_S	Weighted_S	Weighted_S	Weighted_S	Weighted_S	Weighted_S	Weighted_S	Weighted_S	Weighted_S	Weighted_Summat
0	!! FiGo !!	3.39E-05	0	0	0	0	0	4.47E-06	0	0	0
1	!-Bb0yH4cK	3.39E-05	4.43E-05	0	0.0001078	5.28E-05	0	0.0003844	0	0.0002142	0
2	#13bond	0.0012198	0	0	0	0	0	1.79E-05	0	1.86E-06	0
3	#3171	0	0	0	0	0	0	0	0	0	0
4	#An0nsouL	0	0	0	0	0	0	6.70E-05	0	3.35E-05	0
5	#Martijn	0	0	0	0	0	0	0	0	0	0
6	#PornStar	0	0	0	0	0	0	0	0	0	0
7	#vis#	0	0	0	0.0002155	0	0	0	0	0	0
8	$!ngh	0	0	0	0	0	0	0	0	0	0
9	$$$Dollar$$	0	0	0	0	0	0	0	0	0	0
10	%24-ngh	0	0	0	0	0	0	0	0	0	0
11	%2Abadoor	0	0	0	0	0	0	0	0	0	0
12	%2Atvscst%	0	0	0	0	0	0	0	0	0	0
13	%3D-j-a-c-k	0	4.43E-05	0	0.0001078	2.64E-05	0	0	0	0	0
14	%3Dgomz%	0.0011859	0.0086892	0	0.0045263	0.0051733	0.0001985	0.0006079	0	0	0
15	%3Dhhhhh%	0	0	0	0	0	0	0	0	0	0
16	%40%40aa%	0	0	0	0	0	0	0	0	0	0
17	%40fs25353	0	0.0004433	0	0	0.0002639	2.84E-05	4.92E-05	0	0	0
18	%40l3awa%	0.0054891	0.0492978	0	0.008837	0.0319374	5.67E-05	0.0011934	0	0.0003372	0
19	%5Ba%5Dli-	0	0	0	0	0	0	0	0	0	0
20	%5Bh%5Der	0	0	0	0	0	0	0	0	0	0
21	%5C-%5C-%	3.39E-05	0	0	0	0	0	0.000152	0	6.33E-05	0
22	%7E%2Aprii	0	4.43E-05	0	0	5.28E-05	0	0	0	0	0
23	%7E%7Ejiga	0	0	0	0	0	0	0	0	0	0
24	%7E2-d%7E	0	0	0	0	2.64E-05	0	0	0	0	0
25	%7Ecybergh	0	0	0	0	2.64E-05	0	0	0	0	0
26	%7Ej%7E	3.39E-05	0	0	0	0	0	8.49E-05	0	3.35E-05	0
27	%8E%80r%I	0	0	0	0	0	0	0	0	0	0
28	&amp;#108	3.39E-05	0	0	0	0	0	8.94E-06	0	0	0

Weighted Summation of each Metric



	txtAuthor	Total_Rank
31360	dvsocks	0.7923008
47823	mata00	0.603465
56431	rai10	0.5981673
67202	viruslover	0.469193
30002	dichvusocks	0.4082239
60457	shopsocks5.com	0.3200145
16882	V12	0.2857465
42976	kevvanhq	0.2492877
35027	george111	0.2476513
30314	djalal19	0.2467805
20257	amirhamidi	0.2467329
48159	mazorca	0.2464908
51367	nbossul	0.2464659
67667	waqasosama	0.2464141
13945	Rockymen	0.2271132
36817	harmeet32	0.1920307
19636	alex-moran	0.174619
31336	dustymayron	0.174619
6013	Fr0z3n	0.1729039
7014	HansZimmer	0.1729011
7950	JackBauer	0.1727394
24194	bo0mb0m	0.1626655
1853	Arunr489	0.1601162
43269	kimma	0.1499131
27840	creedcarders	0.1164126
57545	rickyune	0.115342
54015	pcnovak	0.1153323
17258	Way Maker	0.1152318
14119	RusellerHill	0.1151692

Final Ranks



Job #2

47.88.189.118

Status	TLP	User	MD5	Process Time (s)	Start Time	End Time
REPORTED_WITHOUT_FAILS	AMBER	admin	0153145266988c547557893516632b9e	3.89	12:16:53 AM Dec 16th, 2022	12:16:57 AM Dec 16th, 2022

Tags

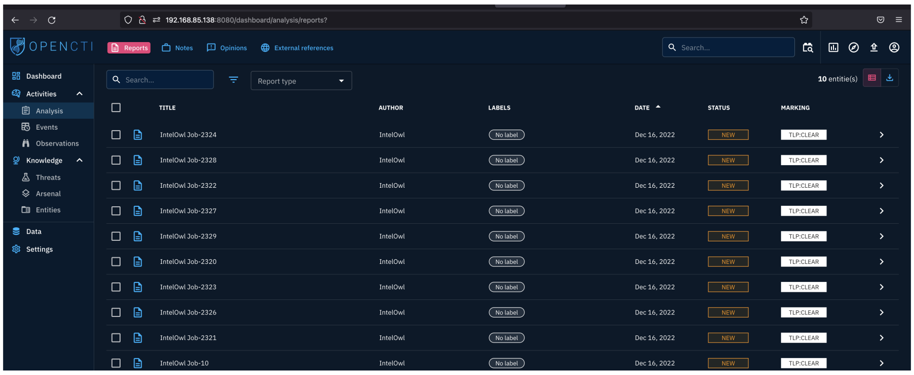
Error(s)

Playbook(s)

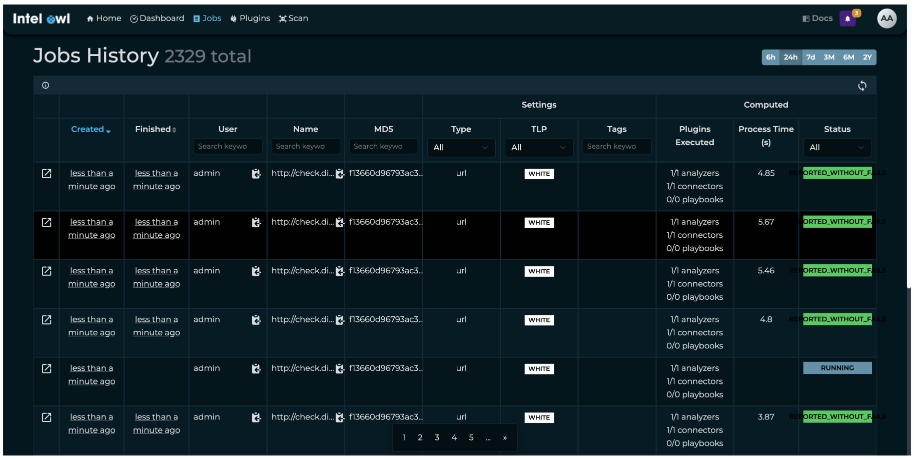
Analyzers Report 3/3 Connectors Report 0/1

		Status All	Name Search keyword..	Process Time (s)	Running Time
		SUCCESS	CryptoScamDB_CheckAPI	2.09	12:16:55 AM - 12:16:57 AM (GMT-8)
		SUCCESS	Classic_DNS	0.62	12:16:54 AM - 12:16:55 AM (GMT-8)
		SUCCESS	AbuseIPDB	0.58	12:16:54 AM - 12:16:54 AM (GMT-8)

4.3 OSINT and Visualization

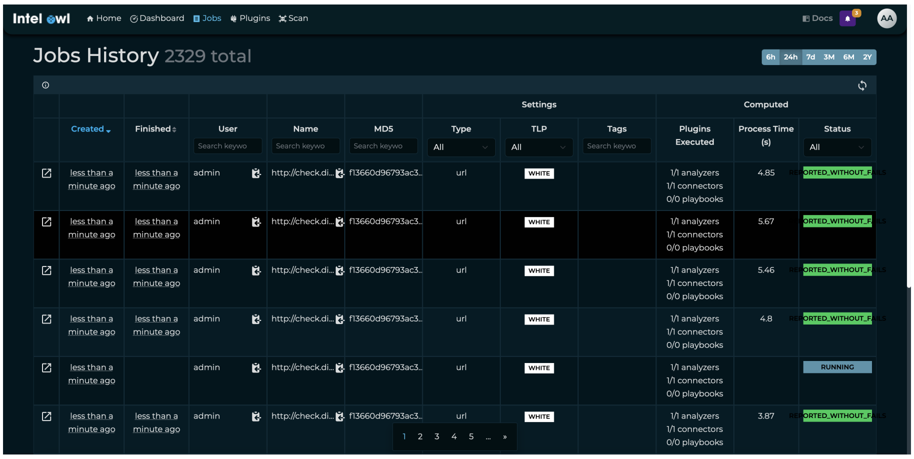


□		TITLE	AUTHOR	LABELS	DATE	STATUS	MARKING
□		IntelOwl Job-2324	IntelOwl	No label	Dec 16, 2022	NEW	TLP:CLEAR
□		IntelOwl Job-2328	IntelOwl	No label	Dec 16, 2022	NEW	TLP:CLEAR
□		IntelOwl Job-2322	IntelOwl	No label	Dec 16, 2022	NEW	TLP:CLEAR
□		IntelOwl Job-2327	IntelOwl	No label	Dec 16, 2022	NEW	TLP:CLEAR
□		IntelOwl Job-2329	IntelOwl	No label	Dec 16, 2022	NEW	TLP:CLEAR
□		IntelOwl Job-2320	IntelOwl	No label	Dec 16, 2022	NEW	TLP:CLEAR
□		IntelOwl Job-2323	IntelOwl	No label	Dec 16, 2022	NEW	TLP:CLEAR
□		IntelOwl Job-2326	IntelOwl	No label	Dec 16, 2022	NEW	TLP:CLEAR
□		IntelOwl Job-2321	IntelOwl	No label	Dec 16, 2022	NEW	TLP:CLEAR
□		IntelOwl Job-10	IntelOwl	No label	Dec 16, 2022	NEW	TLP:CLEAR



Jobs History 2329 total

						Settings			Computed		
	Created	Finished	User Search keywo	Name Search keywo	MD5 Search keywo	Type All	TLP All	Tags Search keywo	Plugins Executed	Process Time (s)	Status All
	less than a minute ago	less than a minute ago	admin	http://check.di...	f13660d96793ac3...	url	WHITE		1/1 analyzers 1/1 connectors 0/0 playbooks	4.85	REPORTED_WITHOUT_FAILS
	less than a minute ago	less than a minute ago	admin	http://check.di...	f13660d96793ac3...	url	WHITE		1/1 analyzers 1/1 connectors 0/0 playbooks	5.67	REPORTED_WITHOUT_FAILS
	less than a minute ago	less than a minute ago	admin	http://check.di...	f13660d96793ac3...	url	WHITE		1/1 analyzers 1/1 connectors 0/0 playbooks	5.46	REPORTED_WITHOUT_FAILS
	less than a minute ago	less than a minute ago	admin	http://check.di...	f13660d96793ac3...	url	WHITE		1/1 analyzers 1/1 connectors 0/0 playbooks	4.8	REPORTED_WITHOUT_FAILS
	less than a minute ago		admin	http://check.di...	f13660d96793ac3...	url	WHITE		1/1 analyzers 1/1 connectors 0/0 playbooks		RUNNING
	less than a minute ago	less than a minute ago	admin	http://check.di...	f13660d96793ac3...	url	WHITE		1/1 analyzers 1/1 connectors 0/0 playbooks	3.87	REPORTED_WITHOUT_FAILS



Jobs History 2329 total

						Settings			Computed		
	Created	Finished	User	Name	MD5	Type	TLP	Tags	Plugins Executed	Process Time (s)	Status
			Search keywo	Search keywo	Search keywo	All	All	Search keywo			All
	less than a minute ago	less than a minute ago	admin	http://check.di...	f13660d96793ac3...	url	WHITE		1/1 analyzers 1/1 connectors 0/0 playbooks	4.85	REPORTED_WITHOUT_FAILS
	less than a minute ago	less than a minute ago	admin	http://check.di...	f13660d96793ac3...	url	WHITE		1/1 analyzers 1/1 connectors 0/0 playbooks	5.67	REPORTED_WITHOUT_FAILS
	less than a minute ago	less than a minute ago	admin	http://check.di...	f13660d96793ac3...	url	WHITE		1/1 analyzers 1/1 connectors 0/0 playbooks	5.46	REPORTED_WITHOUT_FAILS
	less than a minute ago	less than a minute ago	admin	http://check.di...	f13660d96793ac3...	url	WHITE		1/1 analyzers 1/1 connectors 0/0 playbooks	4.8	REPORTED_WITHOUT_FAILS
	less than a minute ago		admin	http://check.di...	f13660d96793ac3...	url	WHITE		1/1 analyzers 1/1 connectors 0/0 playbooks		RUNNING
	less than a minute ago	less than a minute ago	admin	http://check.di...	f13660d96793ac3...	url	WHITE		1/1 analyzers 1/1 connectors 0/0 playbooks	3.87	REPORTED_WITHOUT_FAILS

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this Project, we propose a key hacker identification framework for underground forums . This framework uses Content Analysis. First, we mine the user characteristics of underground forums and construct a comprehensive evaluation. Through user influence ranking, we can identify key hackers in underground forums and use OSINT tool Intel Owl to get intelligence on Indicators of Compromise. We visualize the data and report it using OpenCTI.

5.2 Future Work

This project can be further developed to integrate Machine Learning and Deep Learning techniques to obtain insights on the relations between the members of the hacker forums by analysing their interactions. Another extension that would provide better insights could be to integrate ML/DL techniques on classifying critical and high vulnerabilities and attacks that are being discussed.

Bibliography

[1] A. Abbasi, W. Li, V. Benjamin, S. Hu, and H. Chen, *Descriptive analytics: Examining expert hackers in web forums*, in 2014 IEEE Joint Intelligence and Security Informatics Conference, IEEE, 2014, pp. 56–63.

[2] V. Benjamin and H. Chen, *Securing cyberspace: Identifying key actors in hacker communities*, in 2012 IEEE international conference on intelligence and security informatics, IEEE, 2012, pp. 24–29.

[3] I. Deliu, C. Leichter, and K. Franke, *Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks*, in 2017 IEEE International Conference on Big Data (Big Data), IEEE, 2017, pp. 3648–3656.

[4] C. Huang, Y. Guo, W. Guo, and Y. Li, *Hackerrank: identifying key hackers in underground forums*, International Journal of Distributed Sensor Networks, 17 (2021), p. 15501477211015145.

[5] S. Samtani, K. Chinn, C. Larson, and H. Chen, *Azsecure hacker assets portal: Cyber threat intelligence and malware analysis*, in 2016 IEEE conference on intelligence and security informatics (ISI), Ieee, 2016, pp. 19–24.

[6] X. Zhang, A. Tsang, W. T. Yue, and M. Chau, *The classification of hackers by knowledge exchange behaviors*, Information Systems Frontiers, 17 (2015), pp. 1239–1251.