FLARE

Clustering Malicious Actors: A Three Part Artificial Intelligence Story

Report

https://flare.systems

Table of Contents

Part One.

Why Clustering Malicious Actors Using Artificial Intelligence Can Help with Organizational Cyber Risk

Introduction

In today's digital world, malicious actors have access to a multitude of ways to preserve and increase their anonymity online. We recently posted an article highlighting various Open-Source Intelligence (OSINT) methods to profile cybercriminals on the darkweb, and you could consider this a follow-up and in-depth dive into the Artificial Intelligence (AI) based technologies we use to track, identify, and compare malicious actors.

As we have covered in the article mentioned above, there are a myriad of reasons a malicious actor may utilize a different username on various platforms, ranging from a desire to increase their anonymity and hide from law enforcement to attempting to escape a previously ruined reputation.

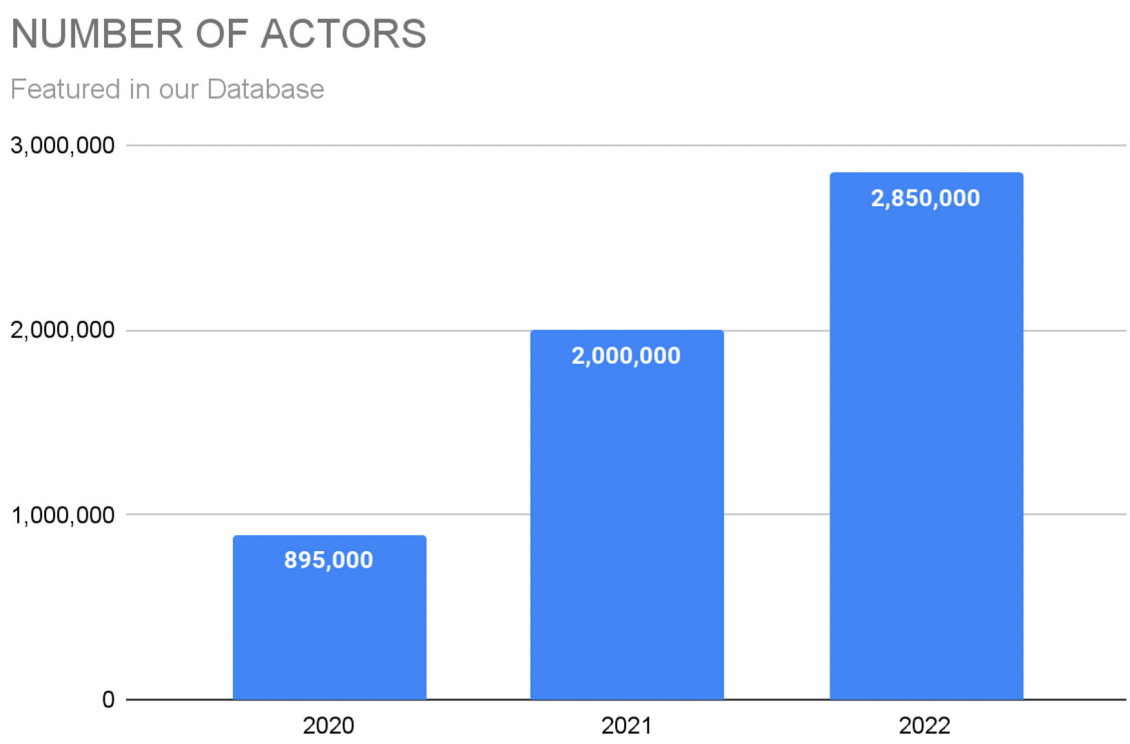
Whilst we previously covered some ways to identify the same actor across multiple platforms or various monikers, the following approach is focused on identifying similarities between actors and by doing so, revealing similar actors. Indeed, using techniques from the field of Natural Language Processing (NLP), AI allows us to easily and rapidly identify similar actors to a previously selected actor.

Identifying similar actors offers a wide range of benefits, of course including potentially detecting a change of moniker in a malicious actor. Additionally, actor similarities may be of vital importance to various investigations when it comes to threat intelligence, allowing the interested party to identify potential threats before any malicious action is taken. Using this process can provide valuable information in identifying the risk a certain actor may pose to your organization. The opposite stays true as well, where a low-risk darkweb actor's activities as well as the activities of similar actors can be ignored, helping reduce noise encountered in investigations.

Another advantage that our Similar Actor Model offers, is the capacity to identify actors based exclusively on the language and wording employed, posting volume notwithstanding; this allows for easy detection of actors that like to keep a low-profile and may have otherwise stayed under the radar. All things considered, volume isn't nearly as important as specific wording and verbiage when it comes to identifying the intent of a cyber-criminal.

The objective when implementing this technology was to allow our users to easily identify malicious actors that could be related to an ongoing investigation, whether participants from various darkweb forums or vendors from darknet marketplaces. This idea stems entirely from our AI expert, Olivier Michaud. If you'd like a sneak peek of Olivier's insights on AI in 2022 watch our 2022 Cybersecurity predictions video and jump to 24:15!

To give you an idea of the challenge ahead, our new Similar Actor Model has to consolidate and analyze all the information currently available in our ever-growing database. At the present moment, this accounts to: more than 2.8 million unique forum profiles, more than 1.6 million forum discussion topics, and nearly 1.9 million darknet market listings!



Amount of unique forum profiles found in Flare's database

Meet Our AI Expert Olivier Michaud.

Olivier has been working with us since the start of his third internship, back in summer 2020. At that time, he was completing his bachelor's degree in software engineering at the École de Technologies Supérieure whilst simultaneously starting his master's in artificial intelligence at the same university. He completed his master's degree in collaboration with Flare; his research project being the automatization of data extraction from various darkweb forums using Natural Language Processing.

Olivier's experience with NLP goes a long way. In fact, his first NLP project was a personal venture of his, Star Wars. Using his recently acquired knowledge in Artificial Intelligence and Web Scraping tools, Olivier's idea was to generate a visualization of his favorite saga. Head over to Medium to read the full story!

Part Two.

How Artificial Intelligence Can Help With Clustering Malicious Actors And Decrease Organizational Cyber Risk

We will now delve deeper into Natural Language Processing (NLP), a branch of AI. More specifically, we will demonstrate how it's possible to cluster malicious actors based on the content they posted on various discussion forums. We will go over some key concepts of NLP, and provide you with some insight as to how we can use these tools and tailor them to the behavior of malicious actors. Before anything else, we need to collect some intelligence about malicious actors.

Gathering Information

For this first step, we used Flare's database, which contains multiple years of information about more than two million malicious actors from all kinds of underground forums. This database is continuously updated and the number of malicious actors is constantly growing.

Looking at each actor featured in our database, we can observe a plethora of different things, including the actor's activity on every monitored forum in which they participate. More intelligence is available, such as the actor's PGP key, their registration date, their title, and various other information that can be used in other ways to profile malicious actors. In the scope of this project, however, what interests us is the title and content of the malicious actors' forum discussion topics. Once enough data has been gathered, the next step will be to cluster the actors based on the content of their posts, for which we use NLP toolings.

Figure 1 - The posting history of a malicious actor encountered on an illicit forum

Natural Language Processing

As previously mentioned, Natural Language Processing is a field of artificial intelligence; NLP's purpose is to make a computer "understand" human language. It is used in numerous applications such as spam detection, translation, sentiment analysis, etc. NLP still is an active research area, thus many challenges are still unresolved. The truth is, humans have unique capabilities when it comes to interpreting nuances and different writing styles, capabilities that computers do not have.

Luckily, the last couple of years brought a great deal of advancements with the rise of Deep Learning: a different branch of Artificial Intelligence that attempts to emulate the human brain. Deep Learning makes it possible to feed an Artificial Intelligence system with text and have it "learn" the meaning of each word.

Unfortunately, most of the available tools rely on a pre-defined vocabulary, often based on well-written content such as news articles or the famous Wikipedia encyclopedia. This means these tools are not exactly adapted to the vocabulary used in underground forums, of which a great portion can be considered jargon and fraud-related lingo. Malicious actors not only have their own way of communicating, but are also located all around the globe and don't limit themselves to the use of the English language alone. This brings us to our next challenge, building a vocabulary specific to the one used on underground illicit forums.

Building a Vocabulary

With underground communities continuously growing all around the world, as previously mentioned, these underground communities do not conform to a single language. The most spoken languages on these platforms are without a doubt English and Russian, but we also have encountered forums exchanging in a multitude of other languages, such as Polish, French, Vietnamese, to name a few.

Taking this into account, our first challenge was to compose a language tailored to these communities in order to train our artificial intelligence. This challenge can be further broken down into two aspects; we don't want our vocabulary to be too large, as this will hinder the model's learning process, and we don't want it to be too small since there wouldn't be enough information to learn from.

The first step in order to build our vocabulary is to generate a text corpus representing the underground wording. In order to do this, we use the textual data we have from each actor in Flare's database. Consequently, this text corpus contains years of malicious actors' publications on various underground forums.

The second step is to determine which words will be part of our vocabulary, and in order to do so, we need to tokenize our corpus. Tokenization refers to the concept of separating text into a series of words and subwords, known as tokens, and associating a unique identifier to each of these. This action is essential for every NLP application, as even if the model's input is simply text, we need to transform it in order to create a representation that the computer will understand since as previously mentioned, computers don't understand language as we do.

Whilst tokenization is not in the scope of this article, it's important to note that there are several ways to tokenize textual content. With this in mind, Flare's AI team developed their tokenization method to interpret text originating from underground vocabularies.

This method used the Byte-Pair Encoding (BPE) algorithm in order to identify words and subwords in the malicious actors' text corpus. In this context, uncommon words are decomposed into subwords in order to reduce the vocabulary size, whilst frequent words are kept intact. Under these conditions, words like hacking and cracking are kept intact, while words like webcam are separated as two tokens, web and cam.

As another example, with Flare's tokenization process, a sentence containing "Scam-Official-Request" will result in three main tokens [Scam, Official, Request] and will be classified as similar to another sentence containing these three words, but without the hyphens. Simpler techniques would not be able to separate this hyphenated sentence efficiently, as they might assume they are part of the same word.

The vocabulary used by malicious communities is constantly evolving, for example, NFTs are the talk of the hour whereas they never were mentioned a year ago. Hence why the algorithm is computed daily on a fresh text corpus in order to capture new trends in malicious actors' discussions.

The following figure shows a subset of this vocabulary, represented in a two-dimensional (2D) space with a t-Distributed Stochastic Neighbor Embedding (TSNE) algorithm, which helps in visualizing high density data, something that would otherwise not be possible to represent in two dimensions. Each point in the representation is an approximation based on the understanding of the Artificial Intelligence for every word in the vocabulary. Therefore, we can see that some words with similar meanings, such as "hack" and "crack" are situated close to each other, the same goes for "Fullz" and "Carding".

Figure 2 - a subset of the vocabulary used on illicit forums, clustered in a 2D space

With our vocabulary defined, and our Artificial Intelligence learning a representation from it, now officially comes the time to cluster similar actors based on the content of their posts.

Clustering Actors

The final component of this experiment was to position each actor into a spatial dimension. In other words, our AI model will "read" every actor's publications, and will distribute them according to the content they are writing about as well as their writing style.

As new actors appear each day, every actor's position is updated each day based on the new information we have, as mentioned above. In the following figure, we can represent a small subset of malicious actors using a TSNE algorithm, as we did for the vocabulary.

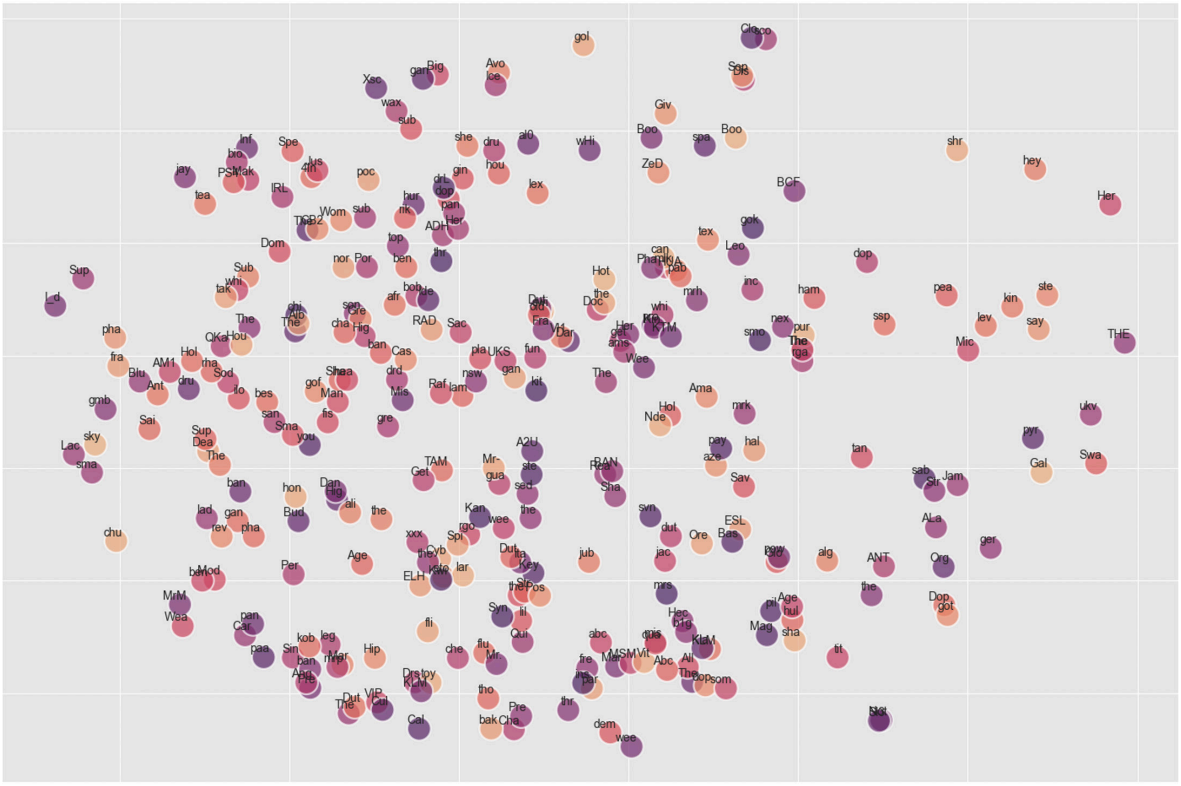


Figure 3 - a subset of actors encountered on illicit forums, clustered in a 2D space
(note: only the first three characteres of the usernames are visible)

This concludes Part 2, giving you hopefully a better view and understanding of the various Artificial Intelligence technologies used in the development of Flare features.

Part Three.

The Result of Clustering Malicious Actors Through Artificial Intelligence

Welcome to the final installment of our AI report, in this section we will go over the results presented in Flare of our new Similar Actor Model. As a reminder, in part one, we discussed why we believe that using artificial intelligence for clustering malicious actors could help organizations in monitoring the dark web more effectively, and part two covered the methods employed to reach our objective; the how if you will.

Integration to Flare

In order to integrate the Similar Actors Model to our Digital Risk Protection (DRP) platform, Flare, we have added a new section to the current "actor profile" interface. This is where all currently known information about said actor is displayed, including, when applicable, their profile on all sources we collect. From there, if any actors are identified as similar to the one presently selected, they will be displayed under the Similar Actors section and can be interacted with to browse their profiles.

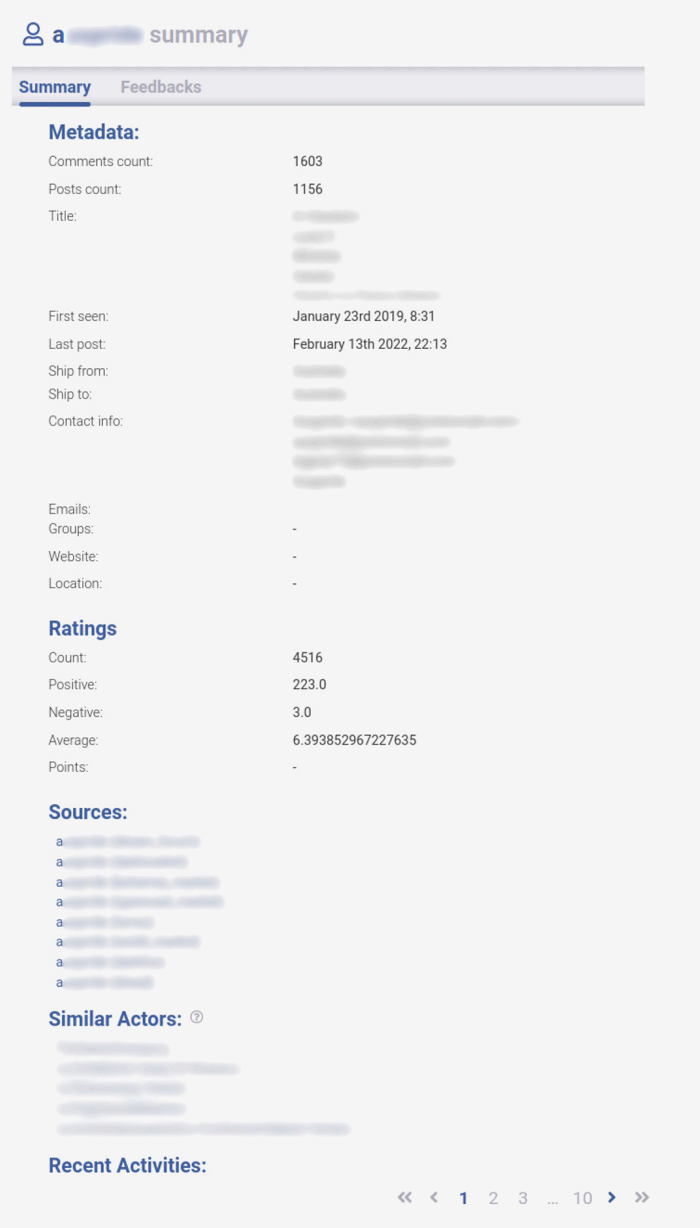
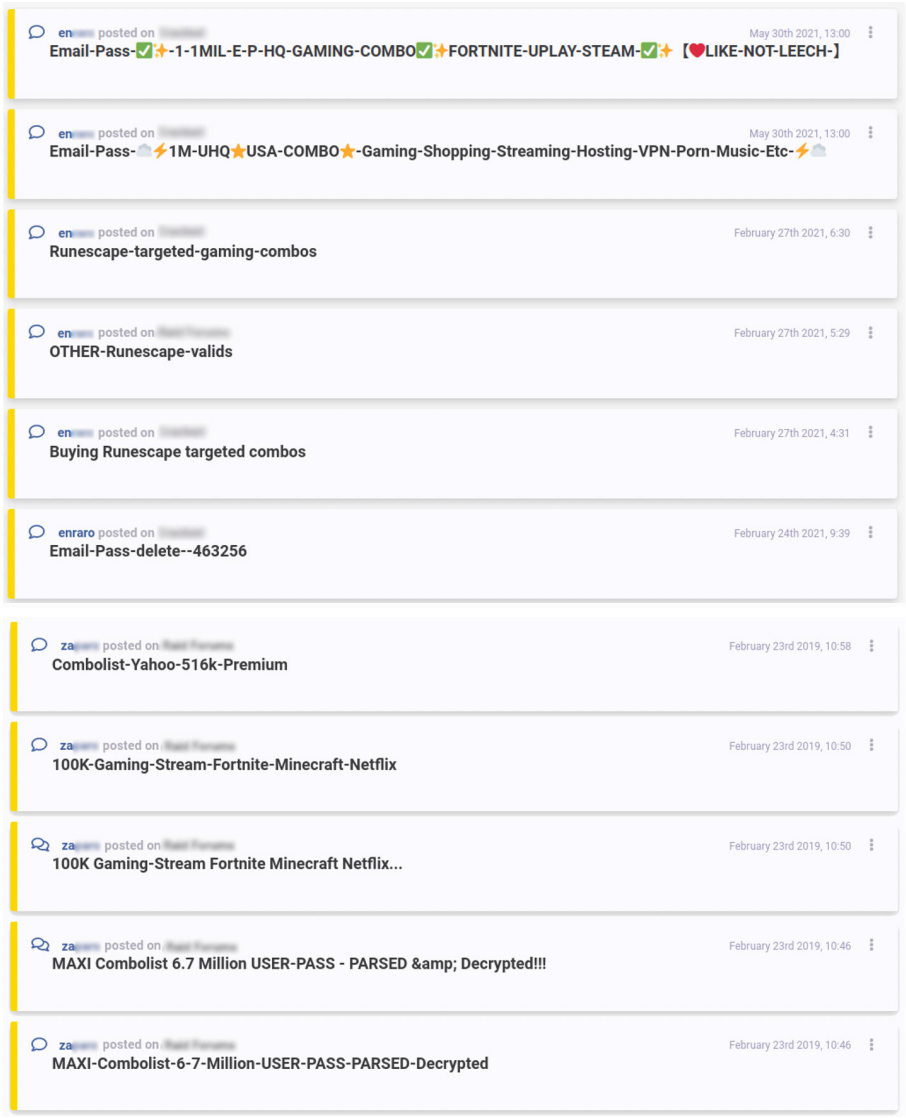


Figure 1 - Example of an Actor's Profile page

Examples of Similar Actors

Now for the exciting parts, the actual results! We have identified a few actors that our Similar Actor Model flagged as similar, in various categories. As previously mentioned in part 2 of this blog series, we know our model identified words such as “hack” and “crack” or “Fullz” and “Carding” to be considered similar, which we know to be true. Let’s see how that translates into actual posts from actors on dark web discussion forums.



In this first example, we can see two different actors sharing what are known as “combolists”, which are essentially documents with thousands of combinations of various email addresses and passwords, typically used for credential stuffing. These actors not only frequent different dark web forums, but also were active at different points in time, yet our Similar Actor Model identified them as being similar from the content of their posts.

The following example shows two actors sharing various fraud methods on dark web forums, mostly related to carding - a type of fraud that exploits stolen credit cards.

Lastly, below are examples of fraudsters sharing money making methods. Once again, worth noting that these two actors were active at different points in time, but were identified as similar nonetheless. One could make an argument that time periods should be taken into consideration when comparing actors, however, some investigations span over long periods of time, and focusing on the content of the messages published by malicious actors online was the top priority when designing this new feature.

The previous snippets of Flare constitute good examples of what our Similar Actor Model can do; we believe it will serve as a powerful investigative tool for Flare users.

To try it out yourself, why not head over to our website and **book a demo!**

Where Do We Go From Here

At the moment, the new Similar Actor Model is still in a sort of beta-phase and is publicly available for Flare users, however, it is limited to actors with a high number of forum discussion topics contributed. We are still very pleased with the results we are seeing so far and plan on expanding the scope of the project to cover vendors / dark web market sellers, as well as regular users that interact with forum topics through forum posts, and not necessarily creating new discussion topics.

In conclusion, the completion of this project marks yet another great Artificial Intelligence integration into Flare, providing users with more intelligence when it comes to threat actors, and additional tools for investigation purposes. As always, we have plenty of other exciting features coming soon to Flare; subscribe to our newsletter to stay in the know about everything Digital Risk Protection related!

Free Trial **Book a Demo**

www.flare.systems
hello@flare.systems

1751 Rue Richardson, Unit 3.107
Montreal, Quebec, H3K 1G6