Cybersecurity analyst's guide on how to use machine learning to show cybercriminals' true colors
Vesta Matveeva

Head of investigation department, APAC

Iaroslav Polianskii

Senior Data Scientist, APAC

Introduction
Data leaks appearing on the dark web are an actual problem of the modern world. Events of the recent past show that even world-renowned enterprise, financial and IT giants are not immune to data leaks. Equifax, British Airways, SingHealth, Marriott International, Sephora, Canva, Zynga, Microsoft, Tokopedia, T-Mobile, Linkedin, Twitch — these are just a few names that have been spotted in high-profile data breach scandals in the past several years. One would hardly attempt to question if these companies had enough resources to ensure their security, but here is the fact — they all did fall victim to cyberattacks.

Even if the company's data was leaked not as a result of its own actions, but rather the actions of its contractor or a partner with which it shared data, the result will be the same. Data leaks can both disclose sensitive information about the company's internal processes affecting business decisions or reputation and personal data of the company's customers, making them turn their back on a brand or company that used to be their favorite. Both outcomes can bring the business at the verge of collapse, which is why one of the first instincts of breached companies, burning from thirst for revenge, is to find attackers behind the network compromise and data leak to dispense justice.

These data leaks can often end in the shadow part of the Internet. Underground resources, which are closed to unauthorized visitors, are rife with discussions that are directly related to planned or previously committed crimes. These resource regulars are attackers themselves, their accomplices looking for an opportunity to make some money, but also visitors who have their own agenda, though, distinct from others. This is the case for employees of law enforcement agencies and special services, as well as corporate and private security specialists. Under the guise of bad guys, they collect cyber intelligence data and research and analyze criminal activity in order to investigate and prevent cybercrimes.

An average underground forum has huge volumes of daily text messages going through it, which makes analyzing this information manually almost impossible and also ineffective. Algorithm development and machine learning (hereinafter - ML) implementation can break this stalemate, significantly reducing the amount of manual work and enabling analysts to connect the dots in the tremendous amount of data.

This article aims to show the methods cybersecurity analysts who come to the aid of compromised companies in such cases can use to, firstly, determine if an alleged data breach did take place and a database put up for sale in the dark web was authentic, and, secondly, identify the threat actor responsible. It demonstrates how machine learning algorithms can facilitate the processes of cyber intelligence data analysis and cyber investigations, while at the same time further enriching its results. And if it happens that you somehow embarked on the path of investigating a data leak, this guide will give you the ideas from what to begin and how to further proceed.

This guide is intended for:
• cybersecurity greeners who are doing their first steps in the cybersecurity world. Even if not all the terms and techniques described in the text are familiar to you, you can get your first impression of cyber investigation process and strengthen your intention to pursue a cybersecurity career;

• cybersecurity analysts and corporate security team members. From the text you'll learn the methods that can be used to probe into a data leak, even if for the time being you're sure that your customers are reliably protected;

• machine learning algorithm developers who will get a broader perspective of the cybercrime investigation process and be able to apply this knowledge in the future to advance the cybercrime investigation industry, making the process more efficient and prompt.

Thus, we'll focus on two major points:

1. **Real breach or fake.** There are leaks that are purported to be new to the public, but in reality they turn out to be databases that were earlier released somewhere else or are new only to some extent, comprising both old and fresh data.

2. **Skilled threat actor or a newbie.** Many attackers today use multiple accounts on the underground forums in order to better hide their activities. To determine the goal of an attack and proceed with its further probe, one has to know the adversary.

**Manual analysis**

We decided to demonstrate the methods that can be used for the achievement of the aforesaid goals using a case with **RaidForums** underground platform where a set of databases was released on September 17, 2020 by a user nicknamed **ExpertData**.
First, Group-IB analysts began to manually analyze the **ExpertData** account. The information publicly available on the forum indicates that this user joined RaidForums on September 5, 2020.

The account was then renamed to repeat its user identifier 121907651 (several threads and posts of **ExpertData** were removed by admins, which is why the total number of posts differs from fig.2).
So the situation was quite typical — a new user published information about multiple data leaks from companies in different regions. It was reasonable to doubt if the databases were real leaks and if the attacker could have been trusted. Then we started to analyze threads of discussion where this account participated.

On September 17, 2020, ExpertData published a thread to sell a leaked database, allegedly containing info on over 100,000 users.
On the same day, **ExpertData** published a thread to sell 18 databases leaked from companies in Brazil, Egypt, Indonesia, India, Italy, Hong Kong, Mexico, Saudi Arabia, Singapore, Thailand, Vietnam, UK, US, and other countries (fig.1). As it can be seen from the screenshot, the original post didn't mention the names of the companies compromised.

About a month later, on October 28, 2020, **ExpertData** published the same post, this time specifying the names of the companies breached and excluding Saudi Arabian companies in the travel industry from the list.

In the post **ExpertData** has also specified his contact details: Jabber account expertdata@jabber.ru and Telegram account @ExpertData.
The databases published by this author seemed to be unique since they had not been seen in the publications of other threat actors before. This was also noted by other forum members, who commented on one of ExpertData's posts.

Figure 6 - **ExpertData**’s account on Telegram

Figure 7 - One of the active forum users with a high reputation commenting on **ExpertData**’s post
In early November 2020, a user nicknamed **The Polaris** initiated a dispute against **ExpertData** in a separate thread. **The Polaris**, who attempted to purchase a database from **ExpertData**, then accused the latter of lying about the quality (quantity of correct password hashes) of the NitroGo (GoNitro.com) dump in a bid to get the payment worth $10,000.

The dispute thread was published on November 2, 2020. At the time of writing, the thread is deleted, but Group-IB analysts managed to retrieve it.

![Figure 8 - Screenshot of the dispute thread against ExpertData initiated by a user nicknamed The Polaris](image)

**The Polaris** joined Raidforums on August 29, 2020, six days before **ExpertData** account was created.
Surprisingly, ExpertData wasn't the first to reply to this dispute: a user with the nickname ShinyHunters stood up for him and came up with counter accusations against The Polaris.
In their post, **ShinyHunters** accused **The Polaris** of hiding under multiple accounts and included a link to a screenshot of their earlier conversation in Jabber, in which **ShinyHunters** suspected that it might be cybersecurity researcher Vinny Troia hiding behind the user nicknamed **The Polaris**, and welcomed him in a personal chat as "Hello Vinny."
ShinyHunters are a famous group of hackers who intrude their victim networks and exfiltrate data. The group with this name became famous in the APAC region after Tokopedia leak in May 2020, and is famous for selling unique databases that they exfiltrate during their hacking activity.

Vinny Troia is, in particular, known for researching hacker group ShinyHunters and having published in July 2020 a research about them. According to his findings, it is supposed that ShinyHunters is a successor to NSFW and Gnostic Players groups.

Figure 11 - Screenshot of a chat between The Polaris and ShinyHunters in Jabber

The dispute stopped on November 10, 2020. The Polaris was banned. Moreover, the account of ExpertData was labeled as a scam one due to several disputes on the forum. After the disputes were resolved, this label was removed from the ExpertData account.
which might indicate that the claims of The Polaris were not deemed relevant by forum administrators.

Typically, the ShinyHunters account doesn't join any disputes not against them, which is why we suggested that there might be a link between ShinyHunters and ExpertData. To further explore this assumption, we started comparing the threads, in which these two users participated and their posts themselves.

After the post with 18 private databases, ExpertData published two more databases (now the posts are deleted from RaidForums): GoNitro (on October 29, 2020) and AnimalJam (on November 11, 2020).

On October 21, 2020, the Nitro PDF (sold as GoNitro dump) service stated that it had been subject to a low impact security incident, however, the media then reported that user and document databases allegedly stolen as a result of the incident were offered for sale in a private auction. ExpertData published information about the GoNitro databases on October 29, 2020, which seems to be the first public post offering for sale the databases.

Later, in January 2021, GoNitro was republished again by the user nicknamed Spiral with a reference to ShinyHunters.

![Figure 13 - Screenshot of Spiral's post offering for sale the GoNitro dump](image)

Thus, we can suggest that ExpertData had access to the leaked GoNitro databases because of the connection with ShinyHunters.

On November 10, 2020, user Chandler-Bing published the AnimalJam database with a reference to ShinyHunters (the original post has been modified by the forum admins, and
now can only be found in the reply of another user).

**Figure 14** - Commentary showing the offer to sell the base by Chandler-Bing

*ExpertData* then published the same database a day later, on November 11, 2020. This could also be a sign of a connection between *ExpertData* and *ShinyHunters*.

However, it is also possible that *ExpertData* just bought the databases and published them by himself. To confirm or reject our hypothesis about affiliation between *ExpertData* and *ShinyHunters*, we decided to analyze *ExpertData* account history.

It was revealed that back in March 2020 the user with the nickname *Expert* used the same Telegram account @expertdata (later in March 2020, the account was banned).

**Figure 15** - Screenshot of *Expert*’s post offering for sale a database
We compared messages of users **Expert** and **ExpertData** and it seems that they used a similar message template to publish the leaks, except for one case only — the database published on September 17, 2020 (fig.4). The **Expert** account also focused on selling databases. However, the quality of databases posted by **Expert** and **ExpertData** (created in September 2020) accounts is different. Most of **Expert**’s databases were not authentic and represented combo lists of previously published data leaks. The changes in the quality of data offered for sale prompted us to think that the user might have started to work with some group in September 2020, presumably **ShinyHunters**.

We moved on with the messages of **ShinyHunters** and found another dispute on RaidForums, in which **ShinyHunters** was previously involved. On May 13, 2020, a user nicknamed **Jumbo** (account created on May 12, 2020) initiated a dispute against **ShinyHunters** and **fs0c131y/whysodank**. Jumbo claimed that he transferred 1.5 BTC to **ShinyHunters** for the databases put up for sale, but didn't get the purchase, with **ShinyHunters** denying the money receipt.

![Figure 16 - Screenshot of a dispute talk initiated by Jumbo against ShinyHunters](image)

The account **whysodank**, mentioned in a dispute, was created back in April 2020.
The history of changing usernames in the profile is provided below.

A user with the nickname *whysodankk* joined the dispute and refuted *Jumbo*'s claims saying that the screenshots allegedly confirming the money transfer in fact represented the transfer of money between two wallets belonging to *Jumbo*. *Whysodankk* also confirmed that he knew *ShinyHunters* and supported his position in the dispute.

*Jumbo* participated in only one thread on the forum — with accusations against *ShinyHunters*. It is noteworthy that it was *whysodankk* who was mainly involved in the discussion, and not *ShinyHunters*, against whom the charges were brought as well.

The dispute ended on May 14, 2020. *Jumbo* was banned from the forum over a fake scam report which shows that he didn't provide evidence in support of his accusations against *ShinyHunters*. 
From the above posts it can be concluded that **whysodankk** is either a partner of **ShinyHunters** or just a different account of the latter. To check this, we decided to find accounts on the RaidForums whose contact details overlap with **ShinyHunters’** ones.

It turns out that **ShinyHunters** and **whysodankk** specified the same contact details — shinyhunters@xmpp.jp. Examples of the posts are presented below.

![Figure 23 - Screenshot of ShinyHunters’ post with Jabber account shinyhunters@xmpp.jp](image)

![Figure 24 - Screenshot of whysodankk’s post with Jabber account shinyhunters@xmpp.jp](image)

(account **whysodankk** was used to sell the Tokopedia database back in May 2020. The post on the screenshot contains the first part of the dump, while the full dump of Tokopedia was published by this account later)

The same contact details indicate that **ShinyHunters** and **whysodankk** are related to each other. Moreover, these two accounts supported each other in the dispute.

The same ready-to-help strategy was used by **ShinyHunters** and **ExpertData**, which leads us to a conclusion that **ExpertData** might be a member or a data broker of **ShinyHunters**
hacker group, and the databases trafficked by ExpertData could have belonged to ShinyHunters.

Figure 25 - Timeline of disputes against ShinyHunters and ExpertData
Additional confirmation of this assumption can be found below.
The table below shows information about contact details of the accounts revealed during the analysis.

The above analysis revealed ties between accounts ExpertData, ShinyHunters, and whysodankk (aka whysodank, fs0c131y). The links between ShinyHunters and other groups are out of the scope of this article.

Going back to our primary task to understand the quality of databases and the level of threat actor for the post of ExpertData on October 28, 2020, we can make the following conclusions:

1. **The databases are most probably authentic.** No other posts with the same databases published earlier by other users were revealed. During the analysis links between ExpertData and ShinyHunters were uncovered. Several databases published by ExpertData were most probably exfiltrated by ShinyHunters and subsequently transferred to ExpertData. Publishing of unique data leaks is typical for ShinyHunters group members.

2. **It is supposed that ShinyHunters was behind these leaks.** ShinyHunters is a famous hacker group with a strong background and skills. It is assumed that the group consists of several people, therefore they can publish new leaks using different accounts. If our conclusion is true, then the databases were obtained as a result of the affected companies' compromise. In such situations it is highly recommended to start incident response and investigation as soon as possible.

After the link between ExpertData and ShinyHunters was uncovered, we decided to further analyze ShinyHunters' relation to other members of the forum and find other accounts that can potentially be linked to this threat actor. However, to find other links to ShinyHunters, we would have had to analyze around a hundred messages of ShinyHunters and a bunch of posts belonging to their potential members. To simplify the research, we decided to use ML-algorithms based on semantic text analysis.

**Semantic analysis**

The first part of the article represented the manual analysis of the attackers' profiles and posts on the underground forum. Group-IB, however, is trying to automate its methodologies of cyber investigations and resorts to ML-algorithms to analyze dark web and build correlations between various posts in the underground.

We, therefore, decided to check the same range of data utilizing Natural Language Processing (hereinafter - NLP) algorithms to reveal multiaccounts of the same person or various accounts of the same group of people. The analysis was carried out based on the
assumption that the semantics of messages from the same person or sometimes even members of the same group should be similar.

The sequence of data collection and processing is shown below:

- The collection of all posts from the RaidForums (topic, message, nickname, date and time);
- The filtering of messages with a minimum length of less than 100 characters to exclude semantically insignificant texts;
- The vector representation of texts obtained using the BERT neural network model (12-layer, 768-hidden, 12-heads, 110M parameters).

For further analysis, all message vectors (N) belonging to user ShinyHunters were selected. For each such vector, the top 10 nearest vectors of other users’ messages were found (by the Euclidean distance). As a result, we obtained a matrix of dimension N*10 (example of the matrix is in the table below), with the nicknames of published messages as well as its elements. Then we filtered out repeated messages and messages whose Euclidean distance exceeds the specified threshold. The threshold was calculated empirically by comparing the proximity of different messages.

As a result, we received a list of messages semantically close to the messages of ShinyHunters. All the messages were grouped by the users who published them (matches). Users who have only one semantically close message (1 match) were excluded from the subsequent analysis to decrease false correlations. The last step was averaging the Euclidean distance across all messages for each of the users. The resulting values will be the scores we were looking for.

It is important to note that this metric is not symmetric: if, for example, a User A account analysis determined their link to User B, such a link might not be found as a result of the analysis of User B. This is due to the fact that we have limited the search to the top 10 nearest vectors for each user message. Surely, such sets of the nearest vectors, even for semantically close messages, may differ. For this reason, the similar analysis for the found users may not display the link with ShinyHunters.
Here is an example of calculation for User_3, based on the results of the first 3 messages in the table above:

\[ S = \frac{(3.3 + 1.1 + 2.2)}{3} = 2.2 \]

Matches (semantically close messages): 3

Average distance (score): 2.2

The results of the analysis of the ShinyHunters messages are presented in fig. 27 (the total number of user posts at the time of writing is indicated in square brackets).

---

<table>
<thead>
<tr>
<th>ShinyHunters</th>
<th>Distances to the semantically closest messages (the User who published the message)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Message 1</td>
<td>0.1 (User_1)</td>
</tr>
<tr>
<td>Message 2</td>
<td>1.1 (User_3)</td>
</tr>
<tr>
<td>Message 3</td>
<td>0.4 (User_2)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Message N</td>
<td>0.25 (User_1)</td>
</tr>
</tbody>
</table>

== ShinyHunters [118] ==

- ShinyHunters: score: 0.0 matches: 118
- fs0c131y: score: 2.452 matches: 5
- MyBiggyBruteBolt: score: 2.564 matches: 3
- J4ckd0x: score: 2.794 matches: 2
- Megadimarus: score: 2.948 matches: 2
- johnlockejrr: score: 2.972 matches: 5
- Omnipotent: score: 3.055 matches: 2
- Troy Hunt: score: 3.194 matches: 2
- Databases: score: 3.228 matches: 4
- FluffyBunnyFufu: score: 3.26 matches: 2

*Figure 27 - The result of the analysis of ShinyHunters posts*
The analysis showed that accounts fs0c131y, MyBiggyBruteBolt, J4ckd0x, Megadimarus, johnlockejrr, Omnipotent, Troy Hunt, Databases and FluffyBunnyFufu published several messages that are semantically close to ShinyHunters' posts.

In fig.27, we see that the number of semantically close messages (matches) between ShinyHunters and other users is small, so it is incorrect to unequivocally state that there is a connection between them. However, the values of this metric give us a reason to take a closer look at possible relationships.

It is noteworthy that the semantic analysis doesn't reveal the strong link between ExpertData and ShinyHunters accounts. Below is the metric for the ExpertData account.

```
============= ExpertData [15] ==============
ExpertData   score: 0.0   matches: 15
DrDastan     score: 3.825  matches: 2
The404       score: 3.914  matches: 2
ambins       score: 4.09   matches: 2
killerstr    score: 4.116  matches: 3
Proto        score: 4.399  matches: 2
Momondo      score: 4.468  matches: 2
DonJuji      score: 4.924  matches: 3
ShinyHunters score: 6.238  matches: 2
```

*Figure 28 - The result of the analysis of ExpertData posts*

The distance to ShinyHunters messages is too high, which can be explained by three assumptions:

1. **ExpertData** is a data broker who worked for this group. Big groups usually change data brokers quite often and don't interact with them closely. That is why the format of the messages of a data broker could be different from ShinyHunters posts.

2. **ExpertData** has few posts;

3. Semantic analysis may not work for dialogues in which each message carries its own semantic meaning. Let's say that threat actors might discuss the same topic, but each of the messages in this thread will have their unique idea, and semantic analysis will not be able to reveal links between these separate messages.
As you can see from fig.27, there is a link between ShinyHunters and fs0c131y (aka whysodankk) accounts. The link was earlier shown in the part of manual analysis and confirmed by semantic analysis.

---

**fs0c131y [17]**

<table>
<thead>
<tr>
<th>Account</th>
<th>Score</th>
<th>Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>fs0c131y</td>
<td>0.0</td>
<td>17</td>
</tr>
<tr>
<td>whysodankk</td>
<td>1.164</td>
<td>2</td>
</tr>
<tr>
<td>ShinyHunters</td>
<td>2.206</td>
<td>5</td>
</tr>
<tr>
<td>johnlockejrr</td>
<td>2.928</td>
<td>2</td>
</tr>
<tr>
<td>MyBiggyBruteBolt</td>
<td>2.984</td>
<td>3</td>
</tr>
</tbody>
</table>

---

**whysodankk [5]**

<table>
<thead>
<tr>
<th>Account</th>
<th>Score</th>
<th>Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>whysodankk</td>
<td>0.0</td>
<td>5</td>
</tr>
<tr>
<td>fs0c131y</td>
<td>1.164</td>
<td>2</td>
</tr>
</tbody>
</table>

---

Figure 29 - The result of the semantic analysis of fs0c131y and whysodankk posts

As you can see different nicknames of the same account are identified by the algorithm.

The link of ShinyHunters with users MyBiggyBruteBolt and johnlockejrr was found, because format of the messages is similar, however most of the posts contain the same note: "This forum account currently has an ongoing scam report, please beware trading. Details: Please respond to this scam report."

![Figure 30 - Semantically close posts of ShinyHunters/MyBiggyBruteBolt users](image)

Figure 30 - Semantically close posts of ShinyHunters/MyBiggyBruteBolt users

Metrics for MyBiggyBruteBolt and johnlockejrr are presented below.
Figure 32 - The result of the analysis of MyBiggyBruteBolt posts

<table>
<thead>
<tr>
<th>Account</th>
<th>Score</th>
<th>Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>MyBiggyBruteBolt</td>
<td>0.0</td>
<td>20</td>
</tr>
<tr>
<td>Leftyy</td>
<td>2.266</td>
<td>2</td>
</tr>
<tr>
<td>Diamondeye</td>
<td>2.322</td>
<td>4</td>
</tr>
<tr>
<td>Asset</td>
<td>2.343</td>
<td>2</td>
</tr>
<tr>
<td>Lvl3-Noob</td>
<td>2.405</td>
<td>2</td>
</tr>
<tr>
<td>DaveCrouse</td>
<td>2.414</td>
<td>3</td>
</tr>
<tr>
<td>BlockDev</td>
<td>2.462</td>
<td>3</td>
</tr>
<tr>
<td>senseye</td>
<td>2.533</td>
<td>2</td>
</tr>
<tr>
<td>Alpaca</td>
<td>2.543</td>
<td>5</td>
</tr>
<tr>
<td>TRGBeast</td>
<td>2.568</td>
<td>4</td>
</tr>
<tr>
<td>abels</td>
<td>2.62</td>
<td>2</td>
</tr>
</tbody>
</table>

In our opinion, these accounts are not relevant to ShinyHunters and could be considered as false positives.

According to fig.27, there is a link between J4ckd0x and ShinyHunters accounts. The account J4ckd0x was created back in 2016.

---

Figure 33 - The result of the analysis of johnlockejrr posts

<table>
<thead>
<tr>
<th>Account</th>
<th>Score</th>
<th>Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>johnlockejrr</td>
<td>0.0</td>
<td>129</td>
</tr>
<tr>
<td>n1js</td>
<td>1.705</td>
<td>4</td>
</tr>
<tr>
<td>imlesor</td>
<td>2.185</td>
<td>2</td>
</tr>
<tr>
<td>GreenHat77</td>
<td>2.306</td>
<td>3</td>
</tr>
<tr>
<td>noname888</td>
<td>2.367</td>
<td>2</td>
</tr>
<tr>
<td>ONYYXX</td>
<td>2.395</td>
<td>3</td>
</tr>
<tr>
<td>xD1ous77</td>
<td>2.712</td>
<td>2</td>
</tr>
<tr>
<td>brokoli777</td>
<td>2.773</td>
<td>4</td>
</tr>
<tr>
<td>Raid-Kingdom</td>
<td>2.814</td>
<td>5</td>
</tr>
<tr>
<td>megalos66</td>
<td>3.009</td>
<td>2</td>
</tr>
<tr>
<td>2020</td>
<td>3.013</td>
<td>3</td>
</tr>
</tbody>
</table>
According to the content of J4ckd0x's messages, his interests lie mostly in selling databases. The metric for this account is presented below.

\[
\begin{array}{lll}
\text{J4ckd0x} & \text{score: 0.0} & \text{matches: 33} \\
\text{theb3ard} & \text{score: 2.447} & \text{matches: 2} \\
\text{ShinyHunters} & \text{score: 2.794} & \text{matches: 2} \\
\text{yametal03} & \text{score: 2.894} & \text{matches: 2} \\
\text{Malkian} & \text{score: 2.999} & \text{matches: 2} \\
\text{Omnipotent} & \text{score: 3.164} & \text{matches: 2} \\
\text{Peas} & \text{score: 3.194} & \text{matches: 2} \\
\text{irakliss} & \text{score: 3.318} & \text{matches: 2}
\end{array}
\]

Figure 35 - The result of the analysis of J4ckd0x's posts

On the screenshot below, you can see that the format of J4ckd0x's posts offering databases for sale is similar to those of ShinyHunters.
There are two matches in messages of **J4ckd0x** and **ShinyHunters** that relate to two databases. **J4ckd0x** reposted (September 2020) one database that was first published by **ShinyHunters** (April 2020). Dunzo database was published by **ShinyHunters** (July 2020) before the post of **J4ckd0x** (September 2020). The format of the message posted by **ShinyHunters** about Sonicbids database is similar to the format of **J4ckd0x**'s post about another leakage. The format of the message posted by **ShinyHunters** about GGumim.com.kr database is similar to the format of the **J4ckd0x**'s post about the Dunzo leakage.

Based on the foregoing, we concluded that accounts **ShinyHunters** and **J4ckd0x** could belong to the same group or at least be partners.

According to fig.27, there is a link between **ShinyHunters** and **Megadimarus** accounts. **Megadimarusr** account was created on May 4, 2020 on the RaidForums and is banned at the time of writing.
The metric for Megadimarius is presented below and is symmetric with ShinyHunters.

```
--------------------- Megadimarius [157] ---------------------
Megadimarius         score: 0.0    matches: 157
Databases           score: 2.892   matches: 2
ShinyHunters         score: 2.948   matches: 2
```

Figure 39 - The result of the analysis of Megadimarus posts

Most of his messages were removed from the visible content of the forums however thanks to the capabilities of Group-IB’s proprietary products the removed contents are preserved and analyzed as well. Most of the Megadimarus posts are related to selling databases. Moreover, some of them are associated with the activities of ShinyHunters predecessors. In the report of cybersecurity researcher Vinny Troia the relationship between ShinyHunters and Megadimarus was also shown.
The metric shows a similarity of posts' format for **ShinyHunters** and **Megadimarus**. Moreover, in this particular case every match is related to the same database (**Appen.com** and **HomeChef** leakages).

**ShinyHunters**

- Description: HomeChef is a unique lifestyle network that connects viewers to the power and joy of food.

**Megadimarus**

- Description: HomeChef is a meal delivery website.

Based on the analysis of **Megadimarius's** posts and interests we concluded that there could be a link between this account and **ShinyHunters**. It could be members of the same group or they could be partners.

Another account for the analysis from fig.27 is **Omnipotent**. It is one of the administrator accounts created back in 2015.

**Omnipotent** profile on RaidForums

One of the matches between **ShinyHunters** and **Omnipotent** is due to the following post.
Figure 42 - Post of Omnipotent about an uploaded database

It was reposted by Omnipotent from GnosticPlayers1 account which is now banned. The format is similar to a format of ShinyHunters post about selling databases. GnosticPlayers1 interest lies in selling databases and this nickname is similar to the name of the group Gnostic Players.

Two messages of Omnipotent have similar format with the format of ShinyHunter’s posts. Since Omnipotent account has thousands of messages, and only one (excluding the reposted one) having the format similar to ShinyHunters, as well as knowing that Omnipotent is the forum administrator, we consider this match as a false positive.

The metric for Omnipotent is presented below.

<table>
<thead>
<tr>
<th>Username</th>
<th>Score</th>
<th>Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Omnipotent</td>
<td>0.0</td>
<td>710</td>
</tr>
<tr>
<td>Mimi</td>
<td>0.0</td>
<td>4</td>
</tr>
<tr>
<td>BoringApe</td>
<td>0.0</td>
<td>3</td>
</tr>
<tr>
<td>RawCat</td>
<td>0.128</td>
<td>6</td>
</tr>
<tr>
<td>Istrator</td>
<td>0.154</td>
<td>5</td>
</tr>
<tr>
<td>raidflacs2k</td>
<td>0.193</td>
<td>4</td>
</tr>
<tr>
<td>Mr-Clair</td>
<td>0.385</td>
<td>2</td>
</tr>
<tr>
<td>September</td>
<td>0.5</td>
<td>7</td>
</tr>
<tr>
<td>2a-45</td>
<td>0.516</td>
<td>5</td>
</tr>
<tr>
<td>September-11-2001</td>
<td>0.546</td>
<td>5</td>
</tr>
<tr>
<td>2a45</td>
<td>0.558</td>
<td>6</td>
</tr>
</tbody>
</table>

Figure 43 - The result of the analysis of Omnipotent posts
Another account from fig.27 is Troy Hunt. This account was created in June 2019 and impersonated the name and photo of the famous cybersecurity researcher Troy Hunt.

![Figure 44 - Screenshot of Troy Hunt (aka Jnx3cx) account](image)

The history of the account's nickname change is presented below.

![Figure 45 - Username history for Troy Hunt/Jnx3cx account](image)

The main interest of the account is selling databases some of which are associated with the activities of ShinyHunters predecessors. The format of the messages is also correlated with the one of ShinyHunters.

ShinyHunters

"Hello RaidForums Community,

Today I have uploaded the Swvl Database for you to download for free, thanks for reading and enjoy!

In June 2020, the Egyptian bus operator Swvl suffered a data breach which impacted over 4 million members of the service. The exposed data included names, email addresses, phone numbers, profile photos, partial credit card data (type and last 4 digits) and passwords stored by bcrypt hashes, all of which was subsequently shared extensively throughout online hacking communities.

Compromised data: Email addresses, Names, Partial credit card data, Passwords, Phone numbers, Profile photos

This dump consists of 1 .json file, this dump's passwords are hashed using bcrypt. The .7z file's MD5 Hash is EB68C073A5E47942E87D840E8F95389. Total record count of 4195918.

Unlock for 8 credits

Database Index <- How To Get Credits"

***

Troy Hunt

"Hello RaidForums Community,

Today I have uploaded the Zynga Database for you to download for free, thanks for reading and enjoy!

In September 2019, one of the world's most successful social game developer Zynga suffered a data breach that exposed 217M accounts. The compromised data included email addresses, names, usernames, phone numbers, and passwords stored as salted SHA1 hashes.

Uncompromised data: Email addresses, Names, Usernames, Phone numbers and Passwords

This dump uses SHA1/DASH hashing Furthermore the data is stored in full SQL format. The .7z file's MD5 Hash is D9773F2C6B8B4BE0A87C08AAA1C2867A2. Total record count of 216595850.

Unlock for 8 credits

Database Index <- How To Get Credits"
Figure 46 - Posts published by Troy Hunt and ShinyHunters and written using the same format

The algorithm shows the similarity of messages of ShinyHunters with the posts of Troy Hunt about selling Canva and Zynga (now deleted from the visible contents of RaidForums). Moreover, Troy Hunt posted a message on behalf of Gnostic Players (is supposed to be a predecessor of ShinyHunters according to the research of Vinny Troia).

Figure 47 - Replies to the deleted post of Troy Hunt with a reference to GnosticPlayers

Figure 48 - Post of Troy Hunt with a reference to GnosticPlayers and Zynga database

The metric for the Troy Hunt account is presented below.
After analyzing the TroyHunt account, his posts and interests, we concluded that its link with ShinyHunters could be relevant meaning the accounts could belong to the same group or at least be partners.

Next account from fig.27 is Databases created back in 2015.
The history of name changes is presented below.

The account started to post databases only in 2019. In August 2021, the format of their messages about selling databases started to resemble ShinyHunter's posts. That is why we suppose that Databases probably became a partner of ShinyHunters in 2021. ShinyHunters themselves have used this format of messages since 2020.

The metric for Databases is presented below.
Figure 53 - The result of the analysis of Databases' posts

As you can see from the metric Troy Hunt (aka Jnx3cx) and Databases are close to each other as well.

The last account for the analysis is FluffyBunnyFufu, created back in 2018.

Figure 54 - FluffyBunnyFufu (aka Ceech, FunnyBunnyHere) profile on RaidForums
Almost all mutual threads of the users relate to databases. One of FluffyBunnyFufu's posts with a leak publication is similar in format to one of ShinyHunters' posts. Additionally, FluffyBunnyFufu joined a dispute initiated against ShinyHunters with fs0c131y/whysodankk. FluffyBunnyFufu joined the talk to defend the position of ShinyHunters. This tactic is common for the members of ShinyHunters group as we noted during the Manual analysis part of this research.

ShinyHunters
"Hello RaidForums Community,\nToday I have uploaded the UImon Database for you to download for free, thanks for reading and enjoy!\nIn January 2020, the travel app creator UImon suffered a data breach. The service had almost 1.3M records with 777k unique email addresses, names, passwords stored as bcrypt hashes and in some cases, social media profile IDs, telephone numbers and bios.\n\nCompromised data: Bios, Email addresses, Names, Passwords, Phone numbers, Social media profiles\n\n\nContents Spoiler\n\nThis download consists of 1 .SQL file, this dump's passwords are hashed using bcrypt. The .7z File's MD5 Hash is 97178B410611FB2D31CA62922D4C009C. Total record count of 777 769.\n\n\n\nDownloads: Hidden Content\n\nUnlock for 8 credits\n\nDatabase Index <> How To Get Credits"

FluffyBunnyFufu
"Hello RaidForums Community,\nToday I have uploaded the CryptoStar.Asia Database for you to download for free, thanks for reading and enjoy!\nIn 2020, the cloud miming website Cryptostar.Asia was hacked and 22.3k accounts were exposed. The data included email,IP addresses and hashed passwords(md5).\n\n\nCompromised data: Email addresses, Passwords, IP Addresses\n\n\nContents Spoiler\n\nThis download consists of 1 TXT containing colon separated columns, hashing method used seems to be MD5. The .7z File's MD5 Hash is A2E37748C097AC7B9A17CFC4A414DBE. Total record count of 222327.\n\n\n\nDownloads: Hidden Content\n\nUnlock for 8 credits\n\nDatabase Index <> How To Get Credits"

According to the available data, it cannot be stated for sure if there is a clear connection between these two accounts, and it may just be a coincidence. However, it is clear that FluffyBunnyFufu keeps an eye on ShinyHunters' activities.

We can conclude that accounts J4ckd0x, Megadimarus, Troy Hunt, and Databases either occasionally used the same message format to sell databases, or they are related to each other and ShinyHunters. However, the majority of ShinyHunters' posts about selling databases have been using the same format since 2020, so we tend to think that this format is typical for the group.
Another argument is that the databases which were posted by these accounts are associated with ShinyHunters or their predecessors. According to Vinny Troia’s research, ShinyHunters is a successor to NSFW and Gnostic Players groups.

The semantic analysis along with our manual research has also shown that the higher the distance between the messages of two different accounts is, the weaker their connection is. This sounds reasonable from a theoretical point of view and our analysis has confirmed it.

Conclusions

In this article, we wanted to show several methods of analysis using which researchers can establish possible links between various accounts in the underground. To showcase them, we used the example of infamous group ShinyHunters known for their unique and resonant leaks.

The starting point of our research was a message from ExpertData, a relatively new account with several unique leaks that appeared in September 2020. The investigation could have stopped there if we treated ExpertData as just another new actor of RaidForums. However, our manual analysis suggested a link between this account and ShinyHunters based on their conversational tactics and databases posted for sale.

The suggested NLP algorithm seems to provide a relevant metric to find connected accounts although with some false positives. Combined manual and automatic analysis showed a potential link of ShinyHunters with the following accounts: fs0c131y (aka whysodankk, whysodank), J4ckd0x, Megadimirus, Databases (aka Eutropius, 012, 0 12), Troy Hunt (aka Jnx3cx). Moreover, accounts belonging to the same user can be automatically...
detected, like in cases of fs0c131y, Databases, and Troy Hunt. It is also noteworthy that accounts fs0c131y, J4ckd0x and Jnx3cx have similar format of the name.

**Figure 58 - The links between ShinyHunters and other accounts in the underground**

This article is not about the ShinyHunters activity, the group’s active behavior in the dark web is just a good example for demonstrating automatic techniques that can be used for analyzing discussions. In this case, we deliberately employ a well-known approach for searching semantically close messages to demonstrate the advantages of even common ML methods in cybersecurity investigations.

As you can see, the results of the manual method and the method based on the semantics of the messages both overlap and complement each other. The examination of behavior patterns of accounts in the dark web can show the correlation between accounts used by the same group, with automatic analysis based on the proposed algorithms allowing to increase the depth of research, while reducing the list of suspects. In the article we've examined the following connections:

- between ShinyHunters and fs0c131y (aka whysodankk, whysodank), whose link with ShinyHunters is quite obvious due to the common contact details;

- between ShinyHunters and J4ckd0x, Megadimarus, Databases, Troy Hunt (aka Jnx3cx) where the link was uncovered using semantic analysis;
• between ShinyHunters and ExpertData, whose link with ShinyHunters was not obvious, and could't be revealed by the semantic analysis alone and required some manual work.

False positives were related to standard phrases added by the forum administrators to the messages of other members. Moreover, admins reposted the messages of other members which are semantically close to what is searched. Depending on the goal of the analysis these results can be improved by applying more specific filters. The techniques proposed can be used to speed up and widen the results of the cybercrime investigations related to dark web activities. Once automated, it can even help in the deanonymization process.

Try Group-IB Threat Intelligence & Attribution right now

Detect public leaks, identify affected accounts in breached databases and try advanced threat actor profiling with best-in-class threat intelligence

Group-IB Threat Intelligence & Attribution