

Practical Data Leakage Analysis

written by Mert SARICA | 1 June 2023

Conti, a Russian-backed cybercrime group that earned \$180 million in revenue from ransomware attacks in 2021, reached a major turning point in 2022 with Russia's invasion of Ukraine. The group publicly supported the Russian invasion, resulting in a rift among its international members. One member began leaking internal messages from 2020-2021 on a Twitter account (@ContiLeaks), including the source code for the ransomware they used in their cyberattacks. The group was considered one of the most notorious cybercrime groups in the world.

The screenshot shows the Twitter profile of 'conti leaks' (@ContiLeaks). The profile is set to public and has 28 tweets. The bio reads 'fuck ru gov' and mentions 'Şubat 2022 tarihinde katıldı'. The account has 1 follower and 6,698 followers. A tweet from March 2nd is visible, mentioning 'Ukraine will Rise!' and linking to 'anonfiles.com/zflc33Lbxf/jab...'. Another tweet from March 1st is partially visible, mentioning 'anonfiles.com/l3b7n7L6xc/con...' and 'conti source without locker src.'.

https://twitter.com/ContiLeaks

Hack 4 Career. Inf... LinkedIn Mert SARICA (mer... Inbox - mert.saric...

Anasayfa Keşfet Bildirimler Mesajlar Yer İşaretleri Listeler Profil Daha fazla

conti leaks
@ContiLeaks
fuck ru gov
Şubat 2022 tarihinde katıldı
1 Takip edilen 6.698 Takipçi
KELA, Cryptolaemus ve takip ettiğin diğer 16 kişi tarafından takip ediliyor

Tweetler Tweetler ve yanıtlar Medya Beğeni

conti leaks @ContiLeaks · 2 Mar
Ukraine will Rise! fresh jabber logs anonfiles.com/zflc33Lbxf/jab...
18 58 173

conti leaks @ContiLeaks · 1 Mar
anonfiles.com/l3b7n7L6xc/con... - conti source without locker src.
15 80 222

Mert SARICA @MertSARICA

The screenshot shows a Twitter interface with a dark theme. On the left is a sidebar with navigation options: Anasayfa, Keşfet, Bildirimler, Mesajlar, Yer İşaretleri, Listeler, Profil, and Daha fazla. Below these is a 'Tweetle' button and a user profile for 'Mert SARICA @MertSARICA'. The main content area shows the profile of 'conti leaks' with 28 tweets. Four tweets are visible, each containing a list of links to anonfiles.com. The tweets are as follows:

- conti leaks** @ContiLeaks · 1 Mar
anonfiles.com/T6U9caL6x5/Scr...
anonfiles.com/V1Uec2Ldxa/baz...
anonfiles.com/X3U4cdL6x7/baz...
anonfiles.com/ZfU0c0Lex7/Scr...
anonfiles.com/bfV9cfL5xd/Scr...
anonfiles.com/deVdcaLbx6/Scr...
anonfiles.com/f1VfcbLdxe/Scr...
anonfiles.com/lfV7c2L8xa/con...
anonfiles.com/nfVbccL9x7/baz...
16 replies, 104 retweets, 216 likes
- conti leaks** @ContiLeaks · 1 Mar
anonfiles.com/f1VfcbLdxe/Scr...
2 replies, 22 retweets, 65 likes
- conti leaks** @ContiLeaks · 1 Mar
this is the 2020 chats: anonfiles.com/H8B7b1L4x6/2_t...
3 replies, 27 retweets, 68 likes
- conti leaks** @ContiLeaks · 27 Şub
conti jabber leaks anonfiles.com/VeP6K6K5xc/1_t...
18 replies, 146 retweets, 297 likes

As a cybersecurity researcher, when data from such threat actors is leaked, one of the things that interests me the most is whether the data includes information about hacked organizations in Turkey, as well as non-Russian, English messages. If you ask me why, it's because I can have the opportunity to learn how extensively Turkey is targeted by these threat actors and which nationalities are involved in such internationally organized crime groups. To find out, I decided to conduct cybersecurity research to provide insights to cybersecurity researchers who are also interested in this topic.

First, I downloaded the files that include the Conti group's messages from the sharing area of the vx-underground website. When I extracted all the zip files, more than 11,000 files came out.

Browser window showing a directory listing for Conti/ on the website https://share.vx-underground.org/Conti/.

File Name ↓	File Size ↓	Date ↓
Parent directory/	-	-
Conti Chat Logs 2020.7z	2417273	2022-03-01 02:46:14
Conti Documentation Leak.7z	234714	2022-03-01 05:29:38
Conti Internal Software Leak.7z	3911885	2022-03-01 02:57:08
Conti Jabber Chat Logs 2021 - 2022.7z	1160294	2022-03-02 13:10:39
Conti Locker Leak.7z	6852466	2022-03-05 04:29:03
Conti Pony Leak 2016.7z	62014991	2022-03-01 02:51:14
Conti Rocket Chat Leaks.7z	3370574	2022-03-01 02:47:40
Conti Screenshots December 2021.7z	452894	2022-03-01 02:46:06
Conti Toolkit Leak.7z	94186791	2022-03-01 02:42:15
Conti Trickbot Forum Leak.7z	8542211	2022-03-01 02:50:56
Conti Trickbot Leaks.7z	955850	2022-03-01 06:52:40
Training Material Leak	0	1969-12-31 18:00:00

Terminal window titled "Leak - -zsh - 96x30" showing commands and output:

```
mertrix@Hack4Career Leak % ls -al
total 0
drwxr-xr-x  14 mertrix  staff   448 Apr 10 20:33 .
drwxr-xr-x@  48 mertrix  staff  1536 Apr 10 20:10 ..
drwx----- 150 mertrix  staff  4800 Mar  1 11:34 Conti Chat Logs 2020
drwx-----   3 mertrix  staff   96 Mar  1 14:29 Conti Documentation Leak
drwx-----  14 mertrix  staff   448 Mar  1 11:56 Conti Internal Software Leak
drwx----- 398 mertrix  staff 12736 Mar  2 22:10 Conti Jabber Chat Logs 2021 - 2022
drwx-----   3 mertrix  staff   96 Mar  1 11:48 Conti Pony Leak 2016
drwx-----  10 mertrix  staff   320 Mar  1 11:47 Conti Rocket Chat Leaks
drwx-----   7 mertrix  staff   224 Mar  1 11:35 Conti Screenshots December 2021
drwx-----   4 mertrix  staff   128 Mar  1 11:39 Conti Toolkit Leak
drwx-----  55 mertrix  staff  1760 Mar  1 11:50 Conti Trickbot Forum Leak
drwx-----   4 mertrix  staff   128 Mar  1 15:52 Conti Trickbot Leaks
drwx-----   9 mertrix  staff   288 Apr 10 20:31 conti_locker
drwx-----   4 mertrix  staff   128 Apr 10 20:31 jabber_logs
mertrix@Hack4Career Leak % find . | wc -l
11289
mertrix@Hack4Career Leak %
```

After learning that the messages are stored as readable text in JSON files (Example: 185.25.51.173-20220301.json), my first task was to use the following regex-supported GREP command to find and deduplicate all IP addresses in the files. I ended up with a total of 3819 IP addresses that match these two regex patterns, which I saved in a file named "ip.txt."

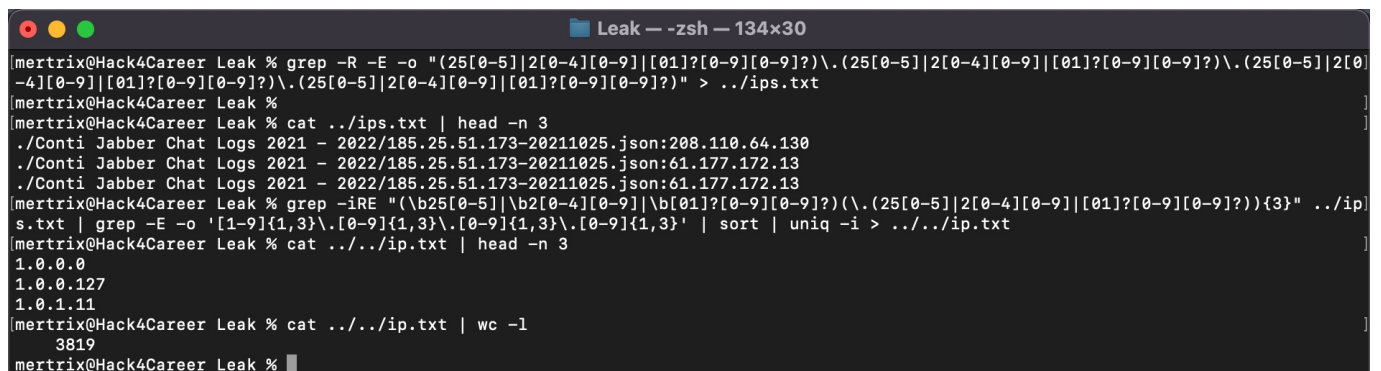
```
grep -R -E -o
```

```
"(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)" > ../ips.txt
```

```
grep -iRE
```

```
"(\b25[0-5]|\b2[0-4][0-9]|\b[01]?[0-9][0-9]?)\.(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)" ..../ips.txt | grep -E -o
```

```
'[1-9]{1,3}\.[0-9]{1,3}\.[0-9]{1,3}\.[0-9]{1,3}' | sort | uniq -i > ..../ip.txt
```



```
mertrix@Hack4Career Leak % grep -R -E -o "(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)" > ../ips.txt
mertrix@Hack4Career Leak %
mertrix@Hack4Career Leak % cat ../ips.txt | head -n 3
../Conti Jabber Chat Logs 2021 - 2022/185.25.51.173-20211025.json:208.110.64.130
../Conti Jabber Chat Logs 2021 - 2022/185.25.51.173-20211025.json:61.177.172.13
../Conti Jabber Chat Logs 2021 - 2022/185.25.51.173-20211025.json:61.177.172.13
mertrix@Hack4Career Leak % grep -iRE "(\b25[0-5]|\b2[0-4][0-9]|\b[01]?[0-9][0-9]?)\.(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)" ..../ip
s.txt | grep -E -o '[1-9]{1,3}\.[0-9]{1,3}\.[0-9]{1,3}\.[0-9]{1,3}' | sort | uniq -i > ..../ip.txt
mertrix@Hack4Career Leak % cat ..../ip.txt | head -n 3
1.0.0.0
1.0.0.127
1.0.1.11
mertrix@Hack4Career Leak % cat ..../ip.txt | wc -l
3819
mertrix@Hack4Career Leak %
```

When it came to finding out which of these IP addresses belong to Turkey, I found help in the IPinfo API and its Python library. By using this library with the IP2Geo Tool v2 that I developed, I queried all the IP addresses in my possession (ip.txt), and I learned that two of these IP addresses (31.210.111.142, 5.188.168.19) are located in Turkey.

The screenshot displays the Socradar platform interface. The top section shows the 'IP INTEL CARD' for the IP address 5.188.168.190, which is categorized as 'High Risk'. A risk score of 1000/1000 is shown, along with a map of Turkey/Istanbul. Below this, a table lists details: IP Address (5.188.168.190), Network (5.188.168.0/23), Country/City (Turkey/Istanbul), and Penalty Reasons (Threathose (100%)).

The bottom section shows search results for public code repositories. The results are filtered by date range from 07 Mar 2022 to 10 Apr 2022. Two results are shown, both from the repository 'https://share.vx-underground.org/Conti/c...'. The first result is dated 13 Mar 2022 and contains a snippet of text: '30t365BP2z0W | Turkey \ n \ n186.216.125.178 system 0kwKcCs8qJP2Z \ n177.190.69.162 admin 0l0ctyQh243063u0 \ n45.235.6.161 system 0kwKcCs8qJP2Z \ n191.241.180.55 admin 0l0ctyQh243063u0 \ n170.84.78.86 system 0kwKcCs8qJP2Z \ n170.247.15.165 system 0kwKcCs8qJP2Z \ ...'. The second result is also dated 13 Mar 2022 and contains a similar snippet: '30t365BP2z0W | Turkey \ n \ n186.216.125.178 system 0kwKcCs8qJP2Z \ n177.190.69.162 admin 0l0ctyQh243063u0 \ n45.235.6.161 system 0kwKcCs8qJP2Z \ n191.241.180.55 admin 0l0ctyQh243063u0 \ n170.84.78.86 system 0kwKcCs8qJP2Z \ n170.247.15.165 system 0kwKcCs8qJP2Z \ ...'.

When it came to my curiosity about the other topic, I decided to explore Python libraries capable of language detection from text. After a brief research, I came across several prominent libraries in this field, including fastText, langdetect and langid

While testing the libraries individually on the text from the leaked Conti data, I observed that each library made accurate language detections for some texts but produced incorrect results for others. As I pondered over which library to use, I decided to develop a tool that combines all three libraries and allows users to specify the confidence level parameter according to their

needs and preferences. This approach would provide a more reliable way to determine the language in a customizable manner.

After merging the leaked Conti data into a single file using the command `find . -type f -print -exec cat {} \; > ../logs.txt`, I used the Language Identification tool I developed to check each line in the “logs.txt” file for Turkish language detection using the three libraries (with the confidence level set to “High”).

To use the Language Identification tool, you need to provide the following parameters.

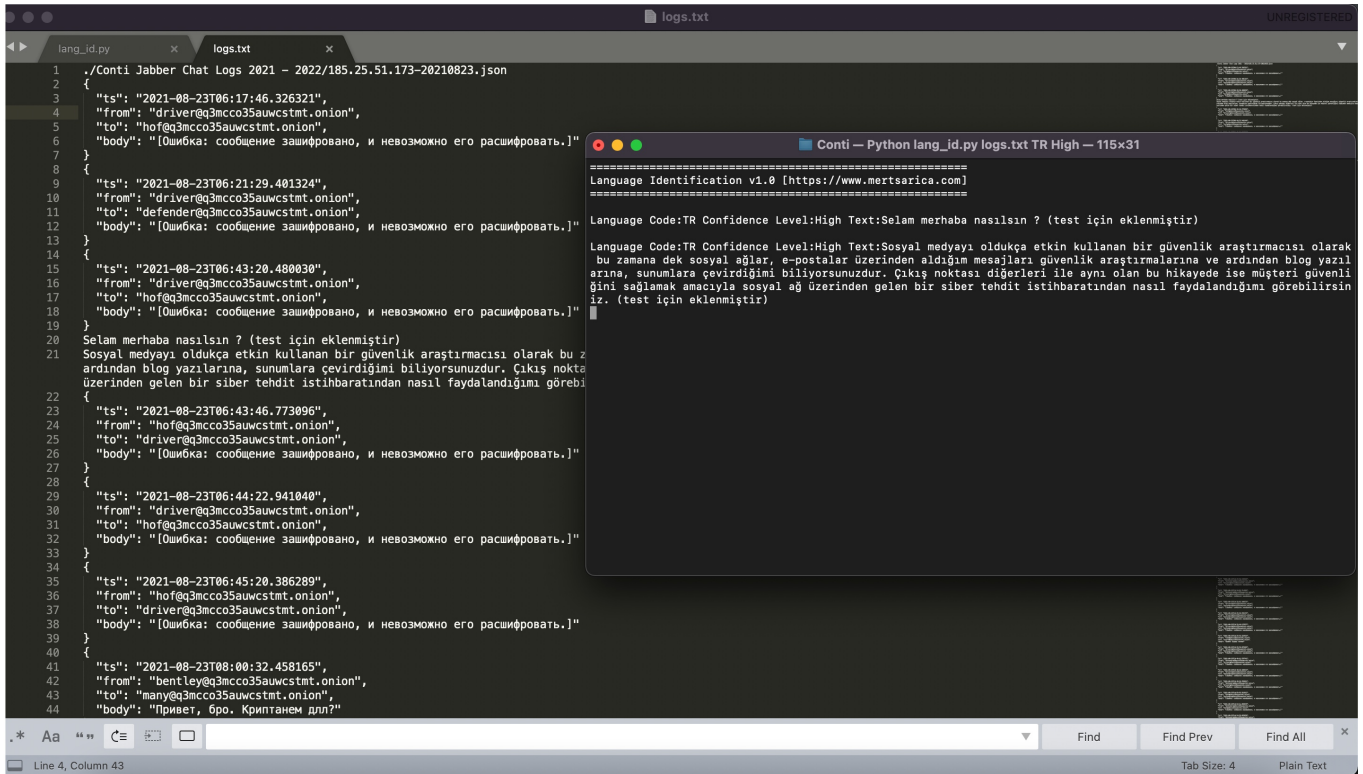
1. The first parameter is the text file you want to analyze, specifying it line by line.
2. The second parameter is the language code for the language you want to detect (e.g., “TR” for Turkish, “EN” for English).
3. The optional third parameter determines the confidence level. If you set it to “High,” when all three libraries detect the language code you specified, it will indicate it on the screen.

Here’s an example command using the tool:

```
python3 lang_id.py logs.txt TR High
```

This command will analyze each line in the “logs.txt” file for Turkish language detection with a high confidence level.

Since there were no Turkish words or sentences used in the text files, there was no language detection indicating the usage of Turkish language by any of the three libraries. However, to test the tool’s functionality, I added three fake Turkish texts to the “logs.txt” file. As a result, I successfully observed that the program detected them correctly. Through this analysis, I learned from the leaked Conti data that there was no Turkish conversation among the group members, thereby clarifying my final curiosity.



The screenshot shows a terminal window with two tabs: 'lang_id.py' and 'logs.txt'. The 'logs.txt' tab is active, displaying a JSON array of chat messages. The messages are from a user with the email 'driver@q3mcco35auwcstmt.onion' to a user with the email 'hofeq3mcco35auwcstmt.onion'. The messages are in Russian and contain encrypted data. The 'lang_id.py' tab is also visible, showing the same JSON array. A window titled 'Conti - Python lang_id.py logs.txt TR High - 115x31' is open, displaying the output of the language identification tool. The output shows the language code 'TR' and confidence level 'High' for the text 'Selam merhaba nasilsin ? (test için eklenmiştir)'. The tool also provides a detailed analysis of the text, including a summary of the content and a list of keywords.

```
1 ./Conti Jabber Chat Logs 2021 - 2022/185.25.51.173-20210823.json
2 {
3   "ts": "2021-08-23T06:17:46.326321",
4   "from": "driver@q3mcco35auwcstmt.onion",
5   "to": "hofeq3mcco35auwcstmt.onion",
6   "body": "[Ошибка: сообщение зашифровано, и невозможно его расшифровать.]"
7 }
8 {
9   "ts": "2021-08-23T06:21:29.401324",
10  "from": "driver@q3mcco35auwcstmt.onion",
11  "to": "defender@q3mcco35auwcstmt.onion",
12  "body": "[Ошибка: сообщение зашифровано, и невозможно его расшифровать.]"
13 }
14 {
15  "ts": "2021-08-23T06:43:20.480030",
16  "from": "driver@q3mcco35auwcstmt.onion",
17  "to": "hofeq3mcco35auwcstmt.onion",
18  "body": "[Ошибка: сообщение зашифровано, и невозможно его расшифровать.]"
19 }
20 Selam merhaba nasilsin ? (test için eklenmiştir)
21 Sosyal medyayı oldukça etkin kullanan bir güvenlik araştırmacısı olarak bu z
22 arından blog yazılarına, sunumlara çevirdiğimi biliyorsunuzdur. Çıkış noktası
23 diğerleri ile aynı olan bu hikayede ise müşteri güvenli
24 ğini sağlamak amacıyla sosyal ağ üzerinden gelen bir siber tehdit istihbaratından nasıl faydalandığımı görebil
25 ir. (test için eklenmiştir)
26 {
27   "ts": "2021-08-23T06:43:46.773096",
28   "from": "hofeq3mcco35auwcstmt.onion",
29   "to": "driver@q3mcco35auwcstmt.onion",
30   "body": "[Ошибка: сообщение зашифровано, и невозможно его расшифровать.]"
31 }
32 {
33   "ts": "2021-08-23T06:44:22.941040",
34   "from": "driver@q3mcco35auwcstmt.onion",
35   "to": "hofeq3mcco35auwcstmt.onion",
36   "body": "[Ошибка: сообщение зашифровано, и невозможно его расшифровать.]"
37 }
38 {
39   "ts": "2021-08-23T06:45:20.386289",
40   "from": "hofeq3mcco35auwcstmt.onion",
41   "to": "driver@q3mcco35auwcstmt.onion",
42   "body": "[Ошибка: сообщение зашифровано, и невозможно его расшифровать.]"
43 }
44 {
45   "ts": "2021-08-23T08:00:32.458165",
46   "from": "bentley@q3mcco35auwcstmt.onion",
47   "to": "many@q3mcco35auwcstmt.onion",
48   "body": "Привет, бро. Криптанем длл?"
49 }
```

```
Conti - Python lang_id.py logs.txt TR High - 115x31
=====
Language Identification v1.0 [https://www.mertsarica.com]
=====
Language Code:TR Confidence Level:High Text:Selam merhaba nasilsin ? (test için eklenmiştir)
Language Code:TR Confidence Level:High Text:Sosyal medyayı oldukça etkin kullanan bir güvenlik araştırmacısı olarak
bu zamana dek sosyal ağlar, e-postalar üzerinden aldığım mesajları güvenlik araştırmalarına ve ardından blog yazıl
arına, sunumlara çevirdiğimi biliyorsunuzdur. Çıkış noktası diğerleri ile aynı olan bu hikayede ise müşteri güvenli
ğini sağlamak amacıyla sosyal ağ üzerinden gelen bir siber tehdit istihbaratından nasıl faydalandığımı görebilirsin
ir. (test için eklenmiştir)
```

I hope this method I have followed and the two tools I have developed will be beneficial for security researchers and experts in data leakage analysis. Hope to see you in the following articles.