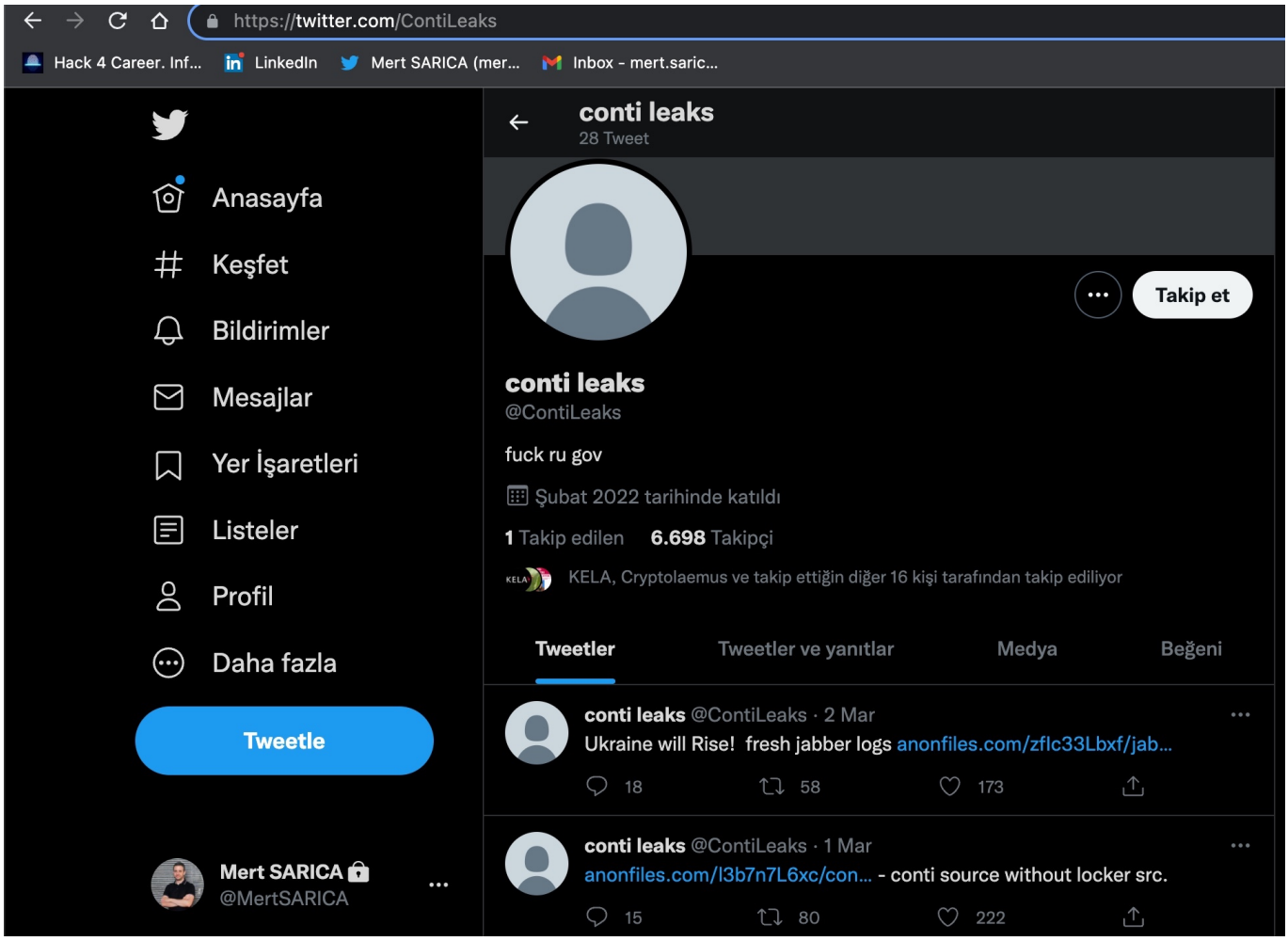
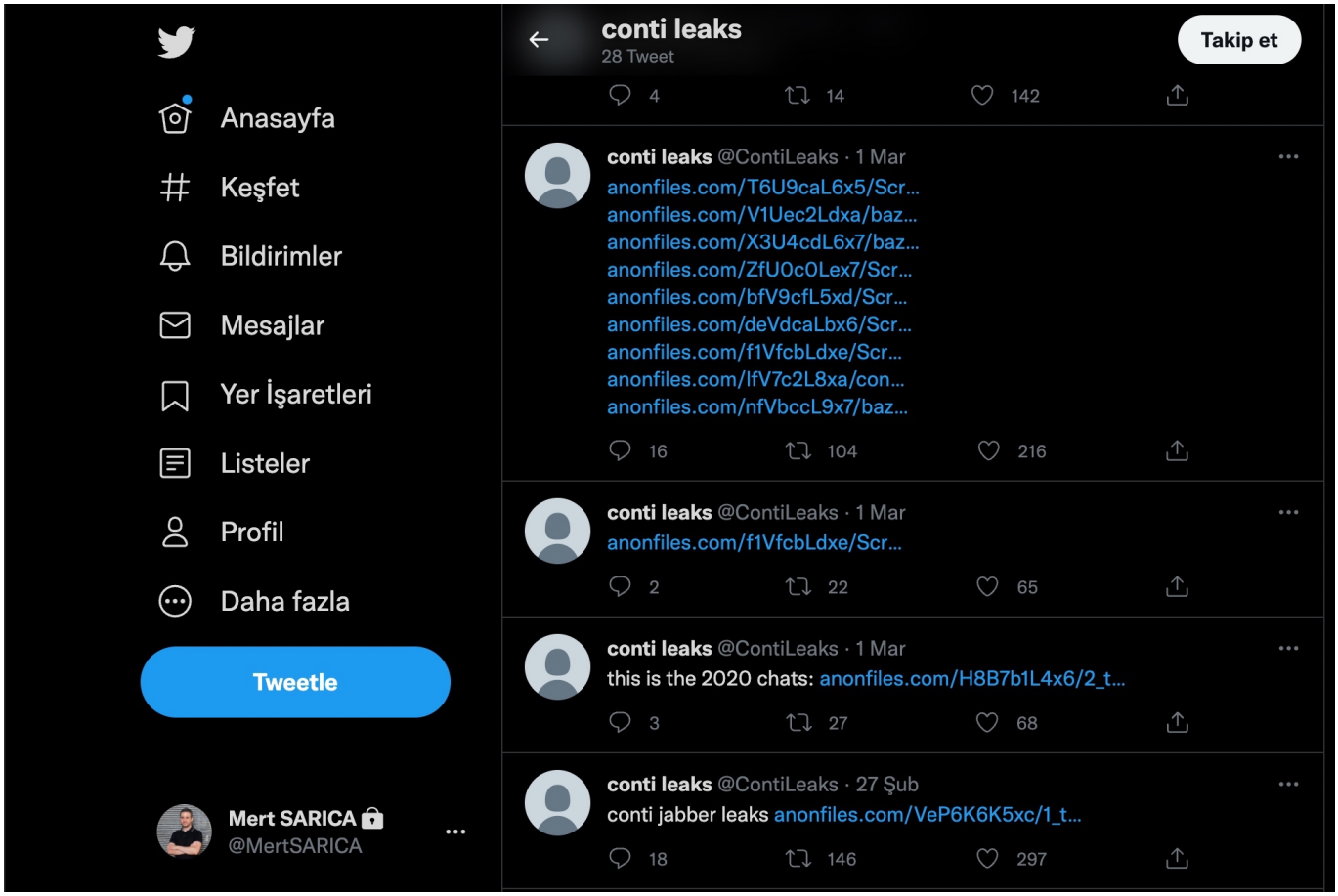
Pratik Veri Sızıntısı Analizi

written by Mert SARICA | 27 May 2023
If you are looking for an English version of this article, please visit here.

Gerçekleştirdikleri fidye yazılım (ransomware) saldırıları ile 2021 yılında 180 milyon dolar gelir elde etmiş ve dünyanın sayılı siber suç çeteleri arasında yer almayı başarmış Rusya destekli Conti grubu, 2022 yılında Rusya'nın Ukrayna'yı işgal etmesi ile önemli bir dönemece gelmişti. Conti grubunun Rus işgalini desteklediğini açıklaması ile uluslararası grup üyeleri arasında oluşan fikir ayrılıkları neticesinde üyelerden biri @ContiLeaks isimli Twitter hesabından grubun kendi aralarında 2020-2021 yıllarında gerçekleştirdiği yazışmaları sızdırmaya başlamıştı ve sızıntılar, siber saldırılarda kullandıkları fidye yazılımlarının kaynak kodunun da sızdırılması ile devam etmişti.

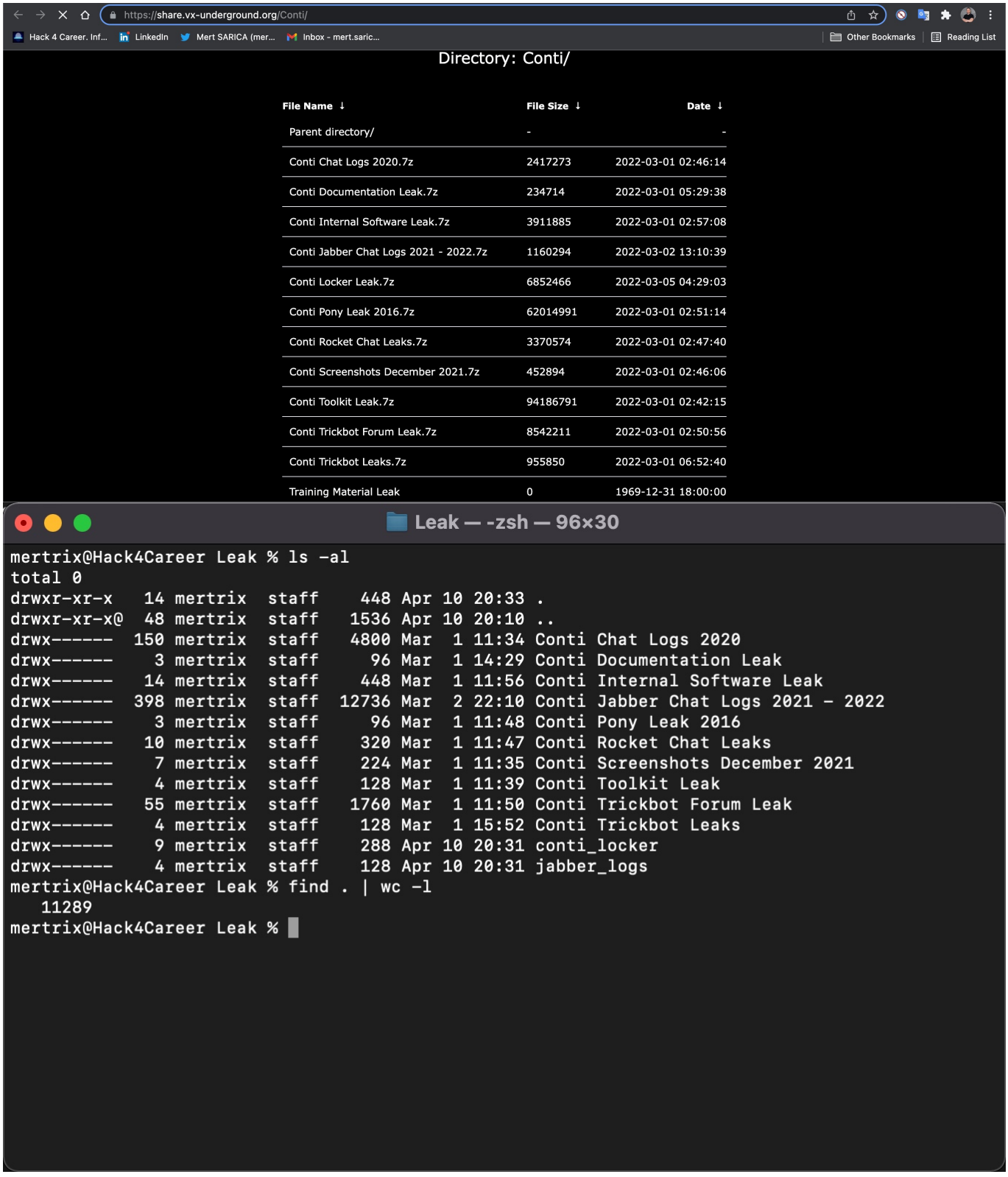




Bir siber güvenlik araştırmacısı olarak bu gibi tehdit aktörlerine ait veriler sızdığında en çok merakımı cezbeden konuların başında bu veriler arasında Türkiye'de hacklenmiş bir sisteme ait bilgilerin ve de Rusça olmayan, özellikle Türkçe yazışmaların, metinlerin yer alıp, almadığı oluyor. Neden diye soracak olursanız, bu tür sayılı tehdit aktörleri tarafından Türkiye'nin ne kadar geniş çapta hedef alındığını ve de bu tür uluslararası, organize, gruplarda hangi milletlerden üyelerin olduğunu öğrenme şansım olabiliyor.

Bu zamana dek bu merakımı gidermek için diğer güvenlik araştırmacılarının analiz çalışmaları ile yetinmeyi tercih etmiş biri olarak buna bir dur demeye, çok daha hızlı bir şekilde bu bilgiye nasıl ulaşabileceğimi öğrenmeye ve benim gibi bu konuya ilgili duyan siber güvenlik araştırmacılarına fikir vermek için kolları sıvamaya karar verdim.

İlk iş olarak Conti grubuna ait yazışmaların da yer aldığı dosyaları vx-underground isimli web sitesinin paylaşım alanından indirdim. Tüm dosyaları teker teker açtığımda içinden 11000'den fazla dosya çıktı.

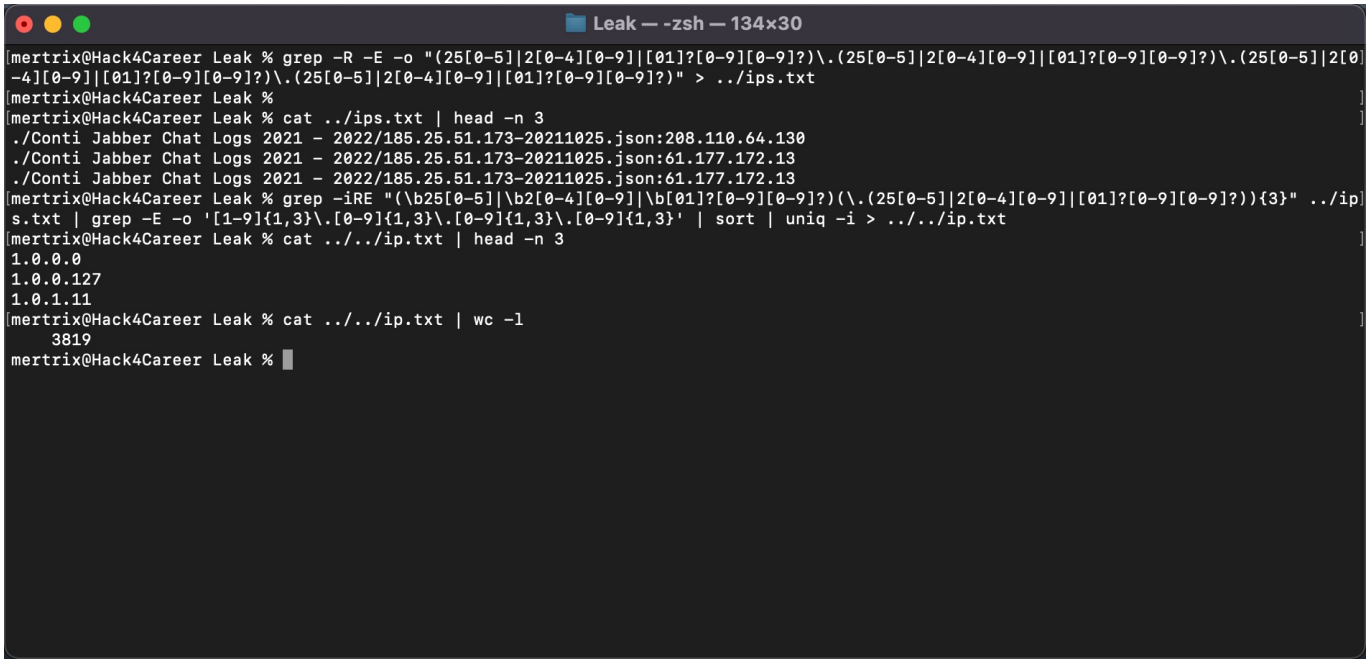


Yazışmaların okunabilir metin olarak JSON uzantılı dosyalarda saklandığını (Örnek: 185.25.51.173-20220301.json) öğrendikten sonra ilk iş olarak tüm dosyalardaki IP adreslerini aşağıdaki regex destekli GREP komutu ile bulup, tekilleştirip ip.txt isimli bir dosyaya kaydettiğimde elimde toplamda bu 2 regex paternine uyan 3819 tane IP adresi oldu.

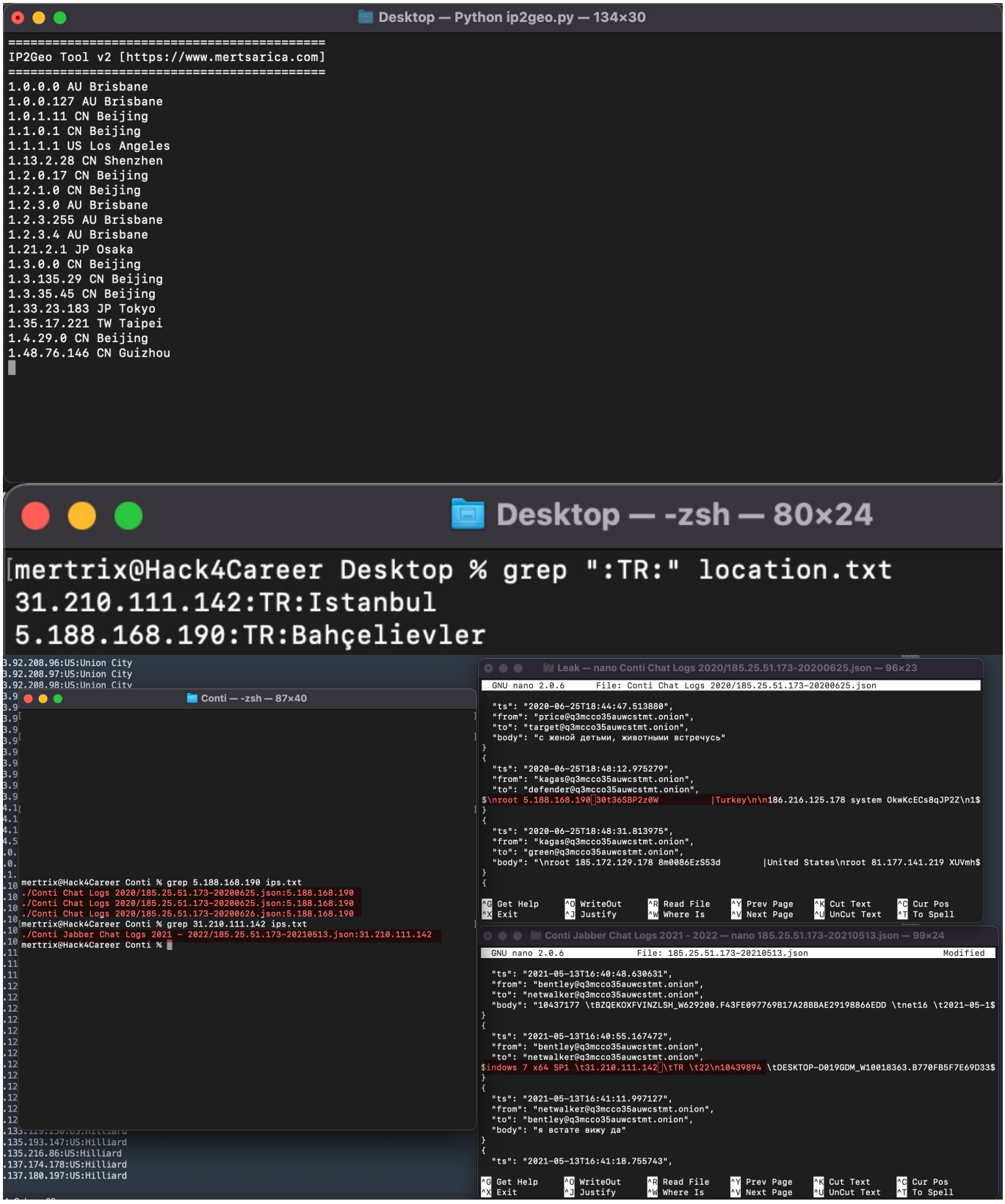
```
grep -R -E -o
```

```
"(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
```

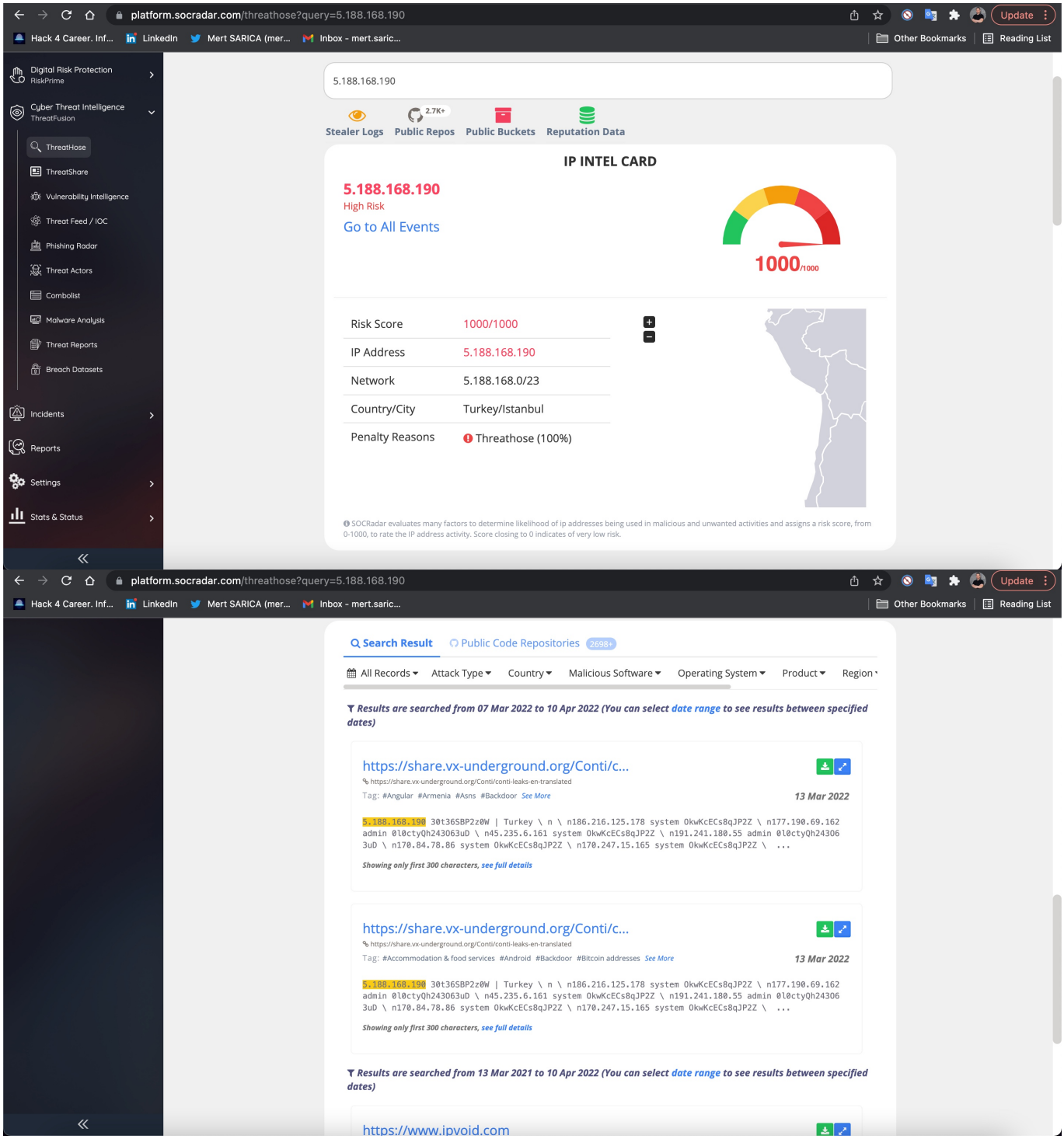
```
grep -iRE "(\b25[0-5]|\b2[0-4][0-9]|\b[01]?[0-9][0-9]?)(\.(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)){3}" ../ips.txt | grep -E -o '[1-9]{1,3}\.[0-9]{1,3}\.[0-9]{1,3}\.[0-9]{1,3}' | sort | uniq -i > ../../ip.txt
```



Sıra bu IP adreslerinden hangilerinin Türkiye'ye ait olduğunu bulmaya geldiğinde imdadıma IPinfo API'si ve Python kütüphanesi yetişti. Elimdeki tüm IP adreslerini (ip.txt) bu kütüphaneden faydalanarak geliştirdiğim IP2Geo Tool v2 isimli araç ile sorguladığımda, bu IP adreslerinden 2 tanesinin (31.210.111.142, 5.188.168.19) Türkiye'de bulunduğunu öğrenmiş oldum.



Elde etmiş olduğum sonucu doğrulamak için ise veri sızıntılarını analiz ederek kurumları anlık olarak haberdar eden Siber Tehdit İstihbarat Platformu SOCRadar'a göz attığımda, sonuçların örtüştüğünü görmüş ve merak ettiğim ilk konuyu açıklığa kavuşturmuş oldum. :)



Sıra merak ettiğim diğer konuya geldiğinde metinden dil tespiti yapabilen Python kütüphanelerine göz atmaya karar verdim. Kısa bir araştırma yaptıktan sonra karşıma bu alanda öne çıkan fastText, langdetect, langid kütüphaneleri çıktı.

Kütüphaneleri teker teker Conti'nin sızan verilerindeki metinler üzerinde test ederken kütüphanelerin kimi metinler için doğru kimi metinler için hatalı dil tespiti yaptığını gördüm. Hangisini kullanacağım konusunda kara kara düşünürken üçünden de faydalanan bir araç geliştirmeye ve kullanan kişinin ihtiyacına, tercihine göre güvenirlik seviyesi (confidence level)

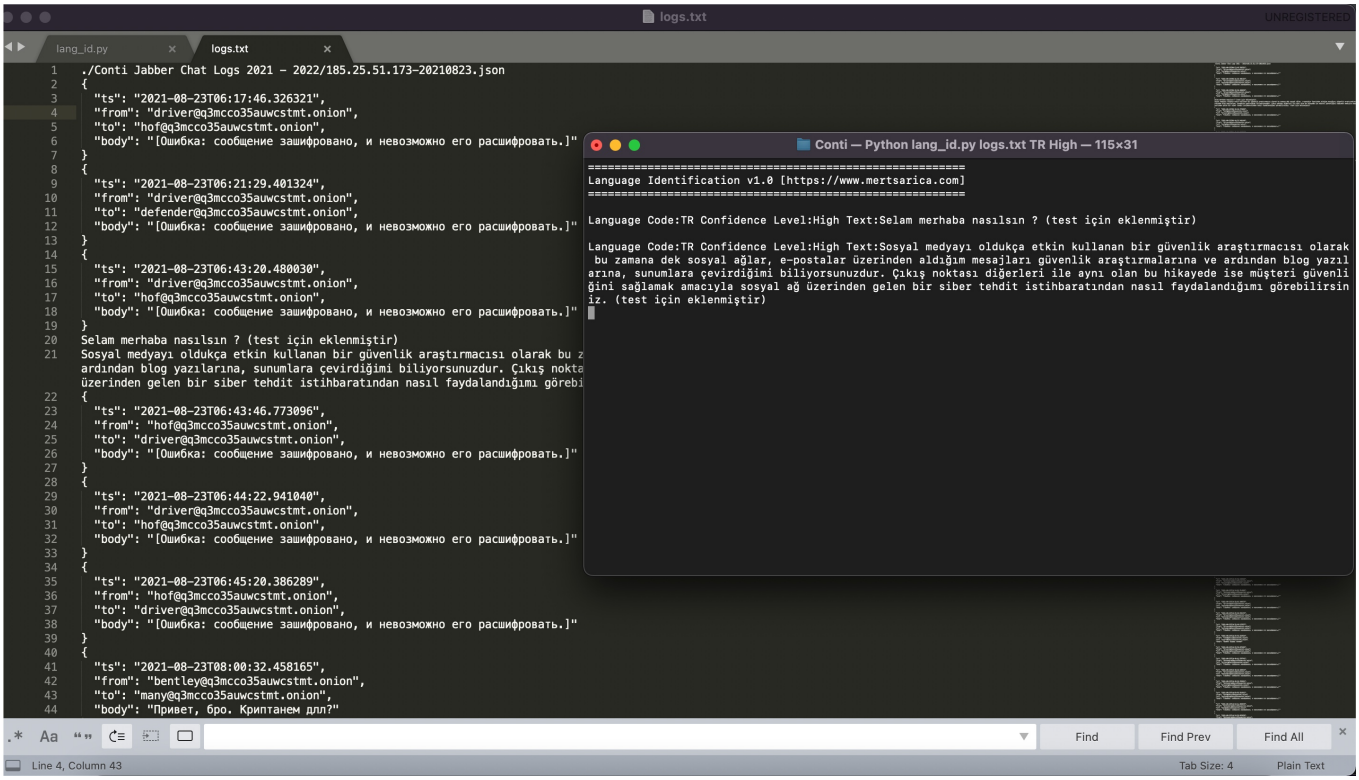
parametresi ile bu aracı kullanmasının daha doğru bir yol olacağında karar kıldım.

find . -type f -print -exec cat {} \; > ../logs.txt komutu ile Conti'nin sızan verilerini tek bir dosyada birleştirdikten sonra geliştirdiğim Language Identification aracı ile logs.txt dosyasındaki her bir satırın Türkçe dil tespitine yönelik olarak üç kütüphane tarafından (güven seviyesi olarak High parametresini belirttim) kontrol ettim.

Language Identification aracını kullanmak için ilk parametre olarak öncelikle satır satır analiz edilmesini istediğiniz metin dosyasını belirtmeniz gerekmektedir. İkinci parametre olarak ise hangi dile dair tespit gerçekleştirmesini istiyorsanız o dilin kodunu (Türkçe ise TR, İngilizce ise EN gibi) belirtmeniz gerekmektedir. Opiyonel parametre olan 3. parametre ise güven seviyesini belirlemektedir. Bu seviyeyi High olarak belirlerseniz üç kütüphanenin de belirttiğiniz dil kodunu tespit etmesi durumunda bunu ekranda belirtecektir.

Örnek kullanım: python3 lang_id.py logs.txt TR High

Metin dosyalarında herhangi bir Türkçe kelime, cümle kullanılmadığı için üç kütüphane tarafından Türkçe dil kullanımına dair herhangi bir tespit olmadı. Aracın düzgün çalıştığını test etmek için ise logs.txt dosyasına Türkçe sahte 3 metin eklediğimde ise programın başarıyla bunları tespit ettiğini görmüş oldum. Bu analiz sayesinde Conti'nin sızan verilerine göre grup üyeleri arasında Türkçe konuşulmadığını öğrenmiş ve merak ettiğim son konuyu da netleştirmiş oldum.



İzlediğim bu yöntemin ve geliştirmiş olduğum iki aracın bu gibi veri sızıntılarında güvenlik araştırmacıları, uzmanları için fayda sağlayacağını ümit ederek bir sonraki yazıda görüşmek dileğiyle herkese güvenli günler dilerim.