WHOIS Data Provides Context to Academic Research on ML-Powered Domain Blocking

Posted on July 25, 2022

WhoisXML API is constantly on the lookout for research collaborations with experts from various fields. Among those who heeded our call for partnership is Mohammad Ismail Daud, a student at the University of California, Davis (UC Davis), who has been working on a thesis on developing a federated learning system that can classify dangerous domains.

Patterns characteristic of malicious domains were analyzed with the help of critical Internet data provided by WhoisXML API. The data allowed the researcher to feed his project while developing his machine learning (ML) algorithm.

The Challenge: Filter Lists as Domain-Blocking Data Sources

While they play a big role in ad-blocking and other browser-filtering solutions, filter lists have their limitations. For instance, only few of them are comprehensive, such that most ad blockers need to glean data from multiple filter lists. Even with several data sources, there is no guarantee that the information is complete and up to date.

Filter lists may not also be geographically inclusive. Some countries and organizations located there may be more active in reporting malicious domains than others, and only these reported properties would typically make their way into filter lists.

The Solution: ML-Enabled Domain-Blocking System

ML can immensely improve domain classification and blocking systems. The thesis focuses on developing an algorithm to classify domains that carry malware, privacy-invading ads, and cryptocurrency miners through a federated learning system. As such, the resulting data will be more comprehensive and inclusive.

To accurately implement the system, the research involved thousands of domains found in filter lists, which then built on patterns and common characteristics found as inputs for the federated learning system.

How WHOISXML API Fits In

WhoisXML API played a critical role during the exploratory data analysis (EDA) stage. According to the researcher, “I used WHOIS API and Website Categorization API. Both tools are simple and straightforward to use. Some API providers have extra bells and whistles that they force you to use, but WhoisXML API is really easy to work with. It’s stable, and I can trust that it’s not going to flatline later.”

WHOIS API

Using WHOIS API, significant data patterns characteristic of malicious domains were found, allowing to determine better how the domains work and how the actors set them up. For instance, among the common attributes found during the EDA are that the domains are primarily under the .com top-level domain (TLD) and registered in the U.S. The most frequently linked registrar behind those domains is also a prominent one.

Website Categorization API

Website Categorization API provided other behavioral patterns that helped focus on the correct data. In particular, most of the domains on the filter list may not host enough content to enable models to decide on blocking. With this insight, the researcher concluded that threat actors behind the malicious domains in the study didn’t put effort into publishing content. These domains carry

malware, cryptominers, and privacy-invading ads, and may only be used as endpoints.

—

The thesis highlights the importance of Internet-related data, including WHOIS and website categories, in classifying malicious domains. Through projects like this, WhoisXML API can continuously work toward achieving our vision of a safer and more transparent Internet.

About Us

WhoisXML API aggregates and delivers the most comprehensive domain, subdomain, IP, and DNS data repositories. Our intelligence is accessible via different consumption models, including APIs, data feeds, monitoring tools, and Web-hosted reports. Please don't hesitate to contact us for inquiries and proposals for joint research and investigations.