Symphony AYASDI

Catching the Unknown Unknowns in Cyber Attacks Using Unsupervised Machine Learning

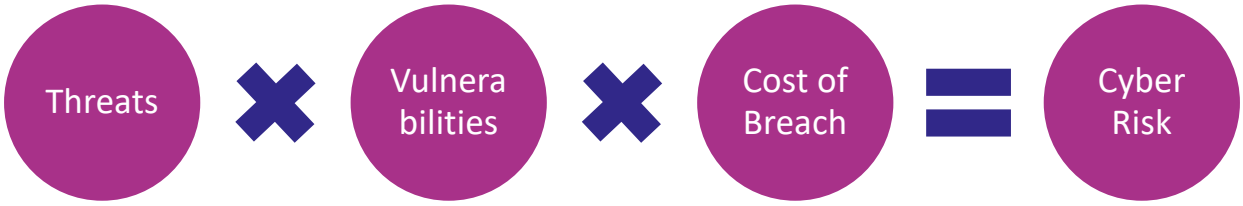
WHITEPAPER

Introduction

It's a truism that criminals try to stay one step ahead of the law. In financial crimes, that is definitely the case. And what's more, yesterday's techniques to combat that crime are woefully inadequate today. Rules-based–or signature-based–cyber detection techniques have been a source of concern among crime prevention practitioners for the past several years due to the inability of those techniques to keep pace with emerging new attack vectors. Most in the financial industry have embraced various AI/ML techniques to essentially allow the model to learn the rules required to detect attacks. The hypothesis has been that with enough sample data, you can build supervised models to detect never-before-seen-attacks. But if you're trying to train models to detect unknown attacks, you need data representative of known attacks.

So how do you discover and detect unknown attacks whose signatures are also unknown? This is the conundrum security experts face today.

Unknown unknowns are attacks no one has detected and have not yet developed approaches to detect and stop with preventative measures. Finding better ways to identify this sort of activity (and the data trail it creates) with low errors is a very valuable endeavor because it may protect against an organization's biggest threat.



Threats are entities with the means and intent to attack your organization. What's at risk ultimately is the value of whatever your enemies can damage or steal, keeping in mind they focus on your perceived or detected vulnerabilities. Juniper Research forecasts that the annual cost of data breaches will increase from $3 trillion in 2019 to $5 trillion in 2024. For financial institutions, the costs of risk mitigation are well known; 6-14% of annual IT budgets, or around 0.2% to 0.9% of company revenue. (Source: Deloitte 2019) But even with all of this expenditure, unknown unknowns are a source of concern because although you may think you are doing everything you can, it only takes a single breach to cause potentially significant damage.

Many new approaches to anomaly detection use supervised learning techniques. Clustering and dimension reduction techniques are among the most common methods for improving detection. However, these approaches have limits since they perform global optimization when the necessary task is understanding rare events and emerging trends. They also lack explainability, and this leads practitioners to spend an enormous amount of time tweaking models. Is there a better way?

Using Topology to Detect Unknown Unknowns in Complex Data

A newer and more effective approach to detecting unknown unknowns use a subset of mathematics called topology, otherwise known as the study of shape. About a decade ago, Stanford University mathematicians discovered that many of the concepts within the realm of topology can be used to make a much-improved sense of highly complex data, defined by datasets with a large number of columns. The practice of applying these techniques to complex data is called topological data analysis, or TDA.

Real World TDA Successes

In one case of cybersecurity implementation, a Japanese bank enlisted Symphony AyasdiAI to find bad actors in log-in data. There was no ground truth for what a bad actor looked like, so the technology had to prove why the anomalies Symphony AyasdiAI found were malicious. Once the technology identified a bad actor, bank employees investigated to validate the work done by the machine. Symphony AyasdiAI looked across hundreds of different features and used unsupervised machine learning to find one anomalous group with over 13K logs, which have 23 high-risk requests from amongst 40 million total logs. All 23 requests in this group originated from two IP addresses, making this a unique group, and a likely malicious actor. It is this level of precision and explainability that is required for cybersecurity to make an impact and be the defense wall enterprises need.

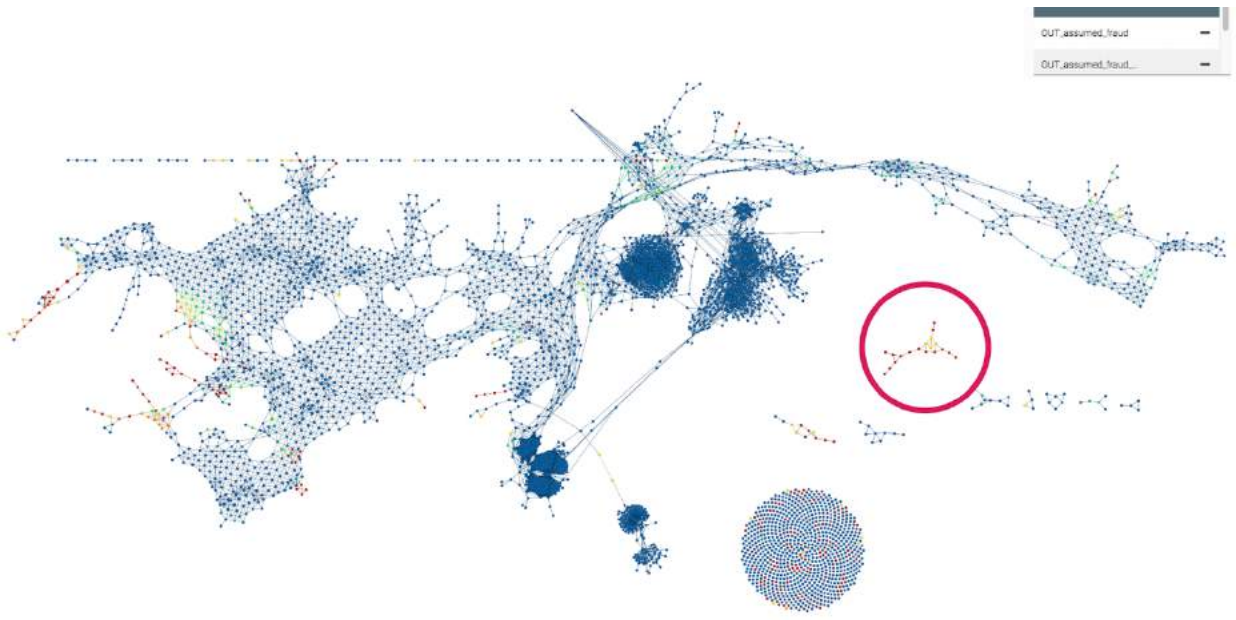


Figure 1: The anomalous group in the log in data TDA frame shows it does not have any connections to the rest of the data.

One of the largest, state-of-the-art telecom companies in Asia was concerned about unknown unknowns and enlisted Symphony AyasdiAI to solve it. Similar to AyasdiAI's previous work, there was no ground truth of what to look for. The goal was to find unknown unknows in a vast dataset, and to identify why particular behavior is suspicious. To the company's surprise, Symphony AyasdiAI's platform was able to find 15 criminals out of 11.5 million accounts in just the first

month, while detecting additional new groups with very similar patterns to those 15 criminals identified earlier.

Recently, the US government researchers studied TDA for its potential impact in the cyber domain of log analysis (Login/out data of employees). Insider threat is a particularly tricky problem to solve because you have no idea what you are looking for. With no prior learning data to train your model, emerging anomaly trends are the only way to understand new threats and use them for future detection.

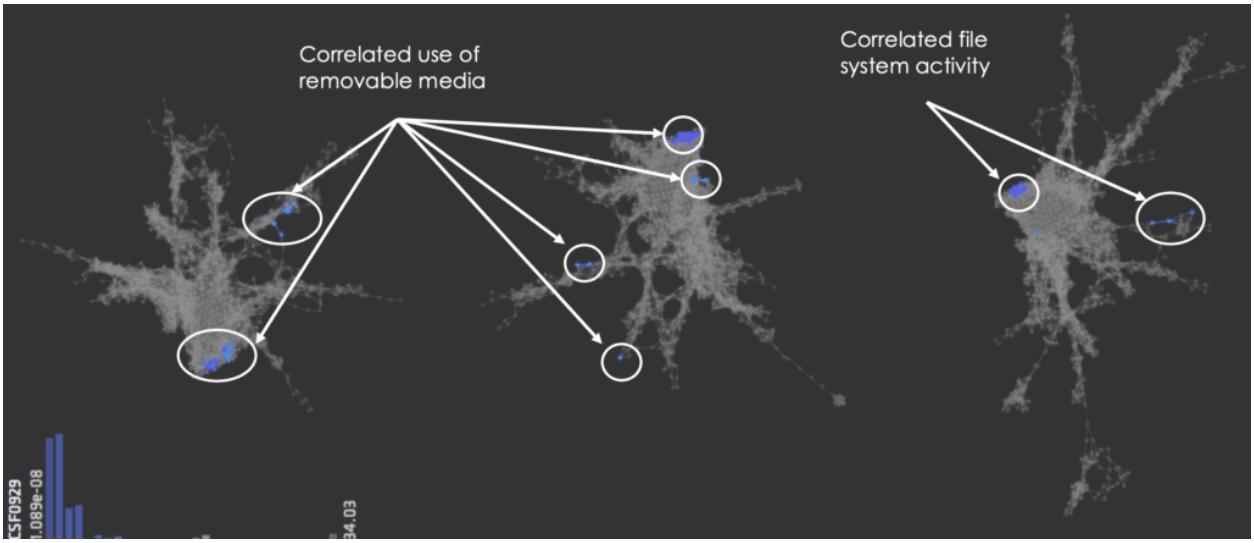


Figure 2: Using TDA to visually show the correlation of suspicious employee log-on activity with removable media usage

The research concluded that with Symphony AyasdiAI TDA, they could spot unknown unknowns, "which would have been missed by traditional means" and that the overall approach "is reproducible and generalizable to other datasets." (Source: http://hdl.handle.net/10125/63977). As a best-of-breed technique to identify unknown unknowns, TDA is at the same time an excellent technique to study any data pattern—known or unknown—and uncover and understand anomalies.

Summary

TDA has a decade of maturity with attention-grabbing successes in detecting previously unseen patterns in medical research, financial forensics, payment integrity, fraud detection, clinical healthcare, and now cybersecurity. These range from uncovering $170 million in liquidity, 45% increase in fraud detection, 25% decrease in false positives, and saving one small hospital more than $2 million per year with clinical variation management. Additionally, academia has been publishing TDA insights at a growing clip, partially cataloged here: https://www.ayasdi.com/resources/publications/

The utilization of AI is simple when dealing with known problems and training a machine to perform tasks sequentially. What's hard is asking a machine to understand or interpret behavior it has never seen before. Unsupervised learning and TDA enable AI to form relationships within the weakest of signals and isolate outlier anomalies that do not correlate to the rest of the data. This is where cybersecurity stands to benefit the most: because you are not only using a system that is trained to catch known attacks but can also alert you to unknown/new attacks as they happen.

About Symphony AyasdiAI

Symphony AyasdiAI, part of the SymphonyAI Group, is the world's most advanced artificial intelligence software company. Symphony AyasdiAI helps organizations discover new and valuable insights in enterprise data. With unprecedented accuracy, transparency, and speed. Built upon over a decade of research and experience, Symphony AyasdiAI delivers insights to Fortune 500 companies and public sector organizations to capture growth, avoid risks and manage inefficiencies.
www.ayasdi.com

A Symphony Group Company

The SymphonyAI Group is the fastest growing and most successful group of B2B AI companies, backed by a $1 billion commitment to build advanced AI and machine learning applications that transform the enterprise. Symphony AI is a unique operating group of over 1,600 skilled technologist and data scientists, successful and proven entrepreneurs, and accomplished professionals, under the leadership of one of Silicon Valley's most successful serial entrepreneurs, Dr. Romesh Wadhwani.

Symphony AYASDI
555 Twin Dolphin Dr, Suite 370
Redwood City, CA 94065 USA
+1 650.704.3395
sales@ayasdi.com
ayasdi.com | @ayasdi