Compliance Chief 360°

The independent Resource for Compliance Officers

Reprinted from compliancechief360.com

Looking for Compliance Risks in Oceans of Data

Leveraging machine learning and advanced anomaly detection to address the overwhelming flood of digital communications data

On September 29, 2021 By John Poulin and Ben Schein

Since email first revolutionized the way professionals communicate with each other, the proliferation of faster and more sophisticated digital communication tools for powering business has only continued to expand. It's well known that companies are facing an unprecedented spike in the volume and complexity of data they generate from digital collaboration platforms that combine elements of instant messaging, audio and video conferencing, file sharing, and social media. And it's getting much worse.

As "work-from-anywhere" models became a requirement during the COVID-19 pandemic and workers were forced to interact almost completely over digital platforms, those oceans of data rose dramatically. This tremendous growth in digital communications data is accompanied by a host of potential underlying compliance risks, even while compliance officers already report their teams are being stretched thinner than ever. Email traffic itself has nearly doubled since the emergence of the pandemic, but the majority of the explosion in company-generated data comes from platforms such as Microsoft Teams, which generates up to 18 new fields of metadata and transactional records for each of its six unique types of virtual interactions.

At the core of the issue is the fact that the flood of data has rapidly outpaced companies' ability to implement data governance strategies to effectively monitor its integrity. This limitation, however, hasn't stopped numerous regulators including the Criminal Division of the U.S. Department of Justice (DoJ), the Securities and Exchange Commission (SEC), and Financial Industry Regulatory Authority (FINRA) each from issuing their own separate guidance regarding companies' retention and utilization of internal data assets in furtherance of compliance and risk management programs. (See, *SEC Rule 17a-3 and 17a-4, SEC Rule 204-2 and 206(4)-7; and FINRA 2210, 2212-2216, 3110, 4511 and 4513.)*

The messaging from regulators since the onset of the COVID-19 pandemic made clear that the onus is on compliance program leaders not only to find ways to effectively monitor this new surge in data they're experiencing at present, but also to continually remediate, according to the DoJ, "any impediments exist that limit [compliance and control personnel's] access to relevant sources of data" for testing of policies, controls, and transactions.

It's a tall order. Many companies are finding that the legacy tools and technologies in place to conduct risk-based monitoring are inadequate or impractical for overseeing and exerting control over such large and diverse sets of data in the manner that regulators now seem to broadly expect.

Why Traditional Approaches Fail

The most traditional and widely-used approaches to digital communications monitoring center around keyword searches, which employ a series of statistical queries to identify or "count" the frequency of search term matches within a particular data set. In their most basic form, keyword search tools can take the form of using "Control+F" at the top of your web browser, while newer natural language processing-based (NLP) tools may incorporate features such as advanced search term syntax languages that enable "batching" of similar results to further isolate individual review areas. Digital communications monitoring tools based on this approach, however, are often only effective in situations where the volume of data is low and the scope of the review is limited to *known* words or phrases associated with the risk or target activity.

The unique and unpredictable nature of human language presents severe complications for compliance leaders that rely solely on keyword searches and other tools based on simple term-frequency to oversee the mass volumes emails, chat messages, customer service calls, and recorded video conference data their companies generate on a daily basis. Notably, there are three key factors that can undermine the effectiveness of traditional keyword search-based monitoring tools when deployed on digital communications datasets of today's scale and complexity:

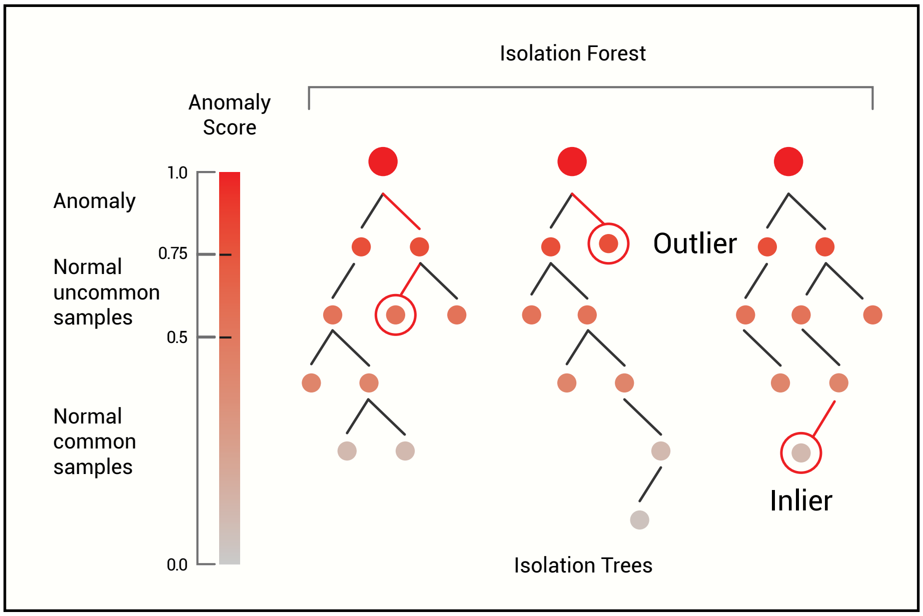
1) Adversarial Adaptation: By now, most employees are well aware that their organizations are collecting and monitoring their digital communications data. Adversarial adaptation refers to the fact that bad actors are constantly adapting or evolving their use of digital communications platforms to avoid detection by monitoring tools. By the time the compliance team has updated its search term list to include a new nickname for a prohibited activity, employees may have already moved on to using an emoji, GIF, or other chat feature signaling to take the discussion offline. Because they limit results to what the reviewer knows to include in the search terms beforehand, the detection power of tools that rely primarily on keyword searches inherently get weaker over time as user behavior evolves in an adversarial manner.

2) Unbalanced Activity: Of the terabytes of digital communications records that companies generate each year of operations, only a minute proportion is likely to exhibit characteristics that put the company at risk. While crucial to executing operations across the business, the overwhelming majority of a company's emails, chats, and meeting logs tend to be – from a compliance and internal control perspective – mundane and of no significance to achieving objectives. Unfortunately, this aspect of digital communications data is in direct conflict with traditional monitoring tools based on the principle that a document's relevance is based unilaterally on the number of times that search terms appear within its content. As the volume of messages being examined increases, using this approach to separate the red flags from rest of the data set becomes less reliable and more prone to blind spots where isolated, one-off issues are more likely to remain undetected.

3) Linear Growth of False Positives: More often than not, compliance audits and internal investigations are not limited to one single focus area or review criteria. As these reviews evolve in scope and complexity, teams using keyword search-based tools have no choice but to tack on additional search terms, casting an ever-wider net for each new risk element added to the list of targets. Because traditional term-frequency algorithms return all matching iterations of search terms regardless of their relevance, the quantity of false positives that review teams will have to sift through increases each time an additional search term gets added to the list. As a result, compliance and internal audit teams are often forced to go through a lengthy trial-and-error process with search term lists, modifying or removing terms until the number of potential "hits" is considered manageable for the budget and resources available to actually perform the review.

Machine Learning to the Rescue

Long before the emergence of COVID-19 accelerated the adoption of digital workspace platforms, experts in data science and computer science had already been developing new mathematical and analytical approaches to outlier detection that address the shortcomings of keyword searches. In 2018, a team of Paris-based researchers published the results of their experimental comparison and analysis of fourteen different algorithms and techniques for detecting outliers within datasets, otherwise known as unsupervised anomaly detection. The results of the study, which compared of each algorithm's ability to flag outliers across 15 unique real-world datasets with sample sizes ranging from 723 to 20,000, showed one technique that consistently outperformed all other algorithms on multiple datasets, known as *isolation forests.*



First developed in 2008 by a group of academics including Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou, the isolation forest technique is based on the statistical reality that outliers in datasets are "few and different," meaning that they occur less often and have attribute values that vary greatly compared to inliers. The algorithm uses a series of binary splits, called isolation trees, which can be trained with machine learning techniques and model datasets to rapidly calculate an "anomaly score" between 0 and 1 for each unique value in a data set. The closer a data point's anomaly score is to 1, the more closely that data point resembles the traits of an outlier.

In addition to its superior ability to detect outliers, other features of isolation forests such as low memory usage and fast computation time led another team of data scientists, including Rémi Domingues, to conclude in 2018 that isolation forest "is an excellent method to efficiently identify outliers while showing an excellent scalability on large datasets along with an acceptable memory usage for datasets up to one million samples." The demonstrated reliability and scalability of this approach to detecting outliers has since given rise to renowned social media companies, including LinkedIn, to adopt isolation forests as their primary tool for detecting and preventing various types of online abuse including fake accounts, member profile scraping, automated spam, and account takeovers.

Using Advanced Anomaly Detection

When dealing with mass volumes of digital communications data, utilizing an advanced anomaly detection algorithm such as the isolation forest technique has a number of advantages that can translate to increased efficiency and more reliable results for compliance team members conducting reviews. Unlike keyword searches, anomaly detection algorithms account for the fact that language is repetitive, and the frequency of a word within a set of communications data does not always imply relevance. Thus, monitoring tools that leverage anomaly detection algorithms trained via machine learning can help greatly reduce the volume of false positives flagged for review, allowing compliance teams to complete reviews in shorter times using less resources.

In addition to saving time and improving efficiency, however, the fact that anomaly detection tools do not rely on delimited search term lists can neutralize the effects of search term bias and adversarial adaptation. When compliance teams are forced to limit their monitoring of digital communications data to *known* risks contained in a delimited search term list, the likelihood of *unknown* risks remaining undetected is much higher. Prioritizing review efforts based on a statistical measure of uniqueness such as anomaly score, as opposed to the number of search term hits, can provide a much more robust and uniform method of executing compliance audits when the number of documents to be reviewed greatly exceeds the number of reviewers.

As the volume of digital communications data continues to grow, meeting the heightened monitoring expectations of regulators will likely prove to be a demanding task, especially while many compliance departments are still working through the backlog of action items emerging from the shift to remote work. To keep an eye on the risks that may be buried in the overwhelming volume of digital communications, organizations must reevaluate their monitoring strategies and equip themselves with the proper tools capable of digesting this data into actionable insights for detecting and responding to activities that put the company in harm's way.

John Poulin is Chief Technology Officer at Helio Argos, the developer of HelioPDR, a proprietary automated compliance data analytics tool.

Ben Schein, CFE, is an Associate at Helio Health Group LLC, a data science-focused compliance consulting organization for the life sciences sector.

ARGOS
AN AFFILIATE OF HELIO HEALTH

www.helioargos.com

Amy Pawloski
Vice President and General Manager, Argos
Helio Argos, LLC
apawloski@helioargos.com
215.518.0880