SAID ABDULLAH, SOHAIL MOHAMMED & DUCHELLE NUNES

Handover research report - Sources of financial crime

Contents

[Introduction: 2](#_Toc152698085)

[What is Financial Crime? 2](#_Toc152698086)

[What Data sources are available? 3](#_Toc152698087)

[What Datasets are available? 5](#_Toc152698088)

[Why did we use the Synthetic Dataset to study financial crime? 6](#_Toc152698089)

[What was the EDA that was done? 8](#_Toc152698090)

[What are the hurdles in obtaining financial crime data? 9](#_Toc152698091)

[What are the solutions for it? 10](#_Toc152698092)

[Challenges for the coding used for this report. 11](#_Toc152698093)

[Where are the codes stored? 13](#_Toc152698094)

[Conclusion and Summary 13](#_Toc152698095)

[Reference 14](#_Toc152698096)

# Introduction:

Financial crime encompasses a wide spectrum of illicit activities aimed at gaining illegitimate financial benefits, poses a significant threat to the global economy and financial systems. To effectively combat this pervasive issue, researchers and policymakers need access to reliable and comprehensive financial crime data. This report delves into the diverse sources of financial crime data, exploring their strengths, limitations, and utilisation strategies.

**The primary sources of financial crime data include** (Nadig, 2023):

**Transaction Data:** A vast repository of financial records, encompassing bank transfers, credit card transactions, and online payments, providing insights into spending patterns, fund movements, and potential anomalies indicative of fraud.

**Customer Data:** Comprising demographic information, employment details, and transaction history, customer data aids in identifying potential risks and suspicious activities, enabling targeted fraud prevention measures.

**Third-Party Data:** External sources, such as public records, social media, and credit reports, supplement traditional customer data, providing additional insights into individuals' financial behaviors and potential risk factors.

**Regulatory Data:** Financial intelligence reports (FINTRs) and regulatory filings offer valuable insights into the nature and scope of financial crime, enabling the assessment of existing prevention measures.

**Law Enforcement Data:** Law enforcement records, including crime reports and investigative details, provide insights into criminal methodologies, modus operandi, and emerging threats.

**Whistleblower reports:** Whistleblowers are individuals who report suspected financial crimes to the authorities. Whistleblower reports can be a valuable source of information, as they often come from people who have firsthand knowledge of the wrongdoing.

**News articles and media reports:** Based on the reports in the Australian media, journalists often play an important role in exposing financial crimes. They usually do this by investigating suspicious activity, interviewing whistleblowers, and analysing public records.

# What is Financial Crime?

Financial crime is commonly characterised as offenses against property, encompassing the illicit appropriation of property owned by others for one’s personal use and gain. (Gottschalk, 2010a). Financial crime, ranging from fraud such as identity theft to large-scale operations like money laundering—estimated at around $2 trillion annually by the UNODC—disproportionately affects vulnerable populations, underscoring the urgency for implementing privacy-preserving technologies to counteract these crimes. Financial institutions combat financial crime by monitoring transactions, and adhering to AML rules, but they are constrained by privacy concerns, data breach risks, and stringent regulatory compliance, hindering the free exchange of information. Data-sharing can enhance the detection of crimes and predictive analytics to prevent financial crimes. The most effective fight against financial crime requires the collaboration of financial institutions and public authorities.

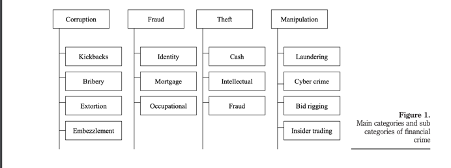
As stated by (Gottschalk, 2010) **financial crime categories** consist of:

**1.** **Corruption**

**2.** **Fraud**

**3.** **Theft**

**4.** **Manipulation**



Source: (Gottschalk, 2010a)

Financial crime in Australia encompasses a wide range of illegal activities that aim to gain financial benefit through illegitimate means. As stated above, these activities can be broadly categorised into three main types:

**Fraud:** Fraud involves deceiving or tricking someone into surrendering money or property. Examples of fraud in Australia include accounting fraud, financial planning misconduct, and cartel conduct.

**Money laundering:** Money laundering is the process of disguising the proceeds of crime to make them appear legitimate. This often involves moving illicit funds through complex financial transactions to conceal their origins. Examples of money laundering in Australia include the Tabcorp, Westpac, CBA, ANZ, and NAB scandals (McKenzie, Baker and Mitchell, 2017)

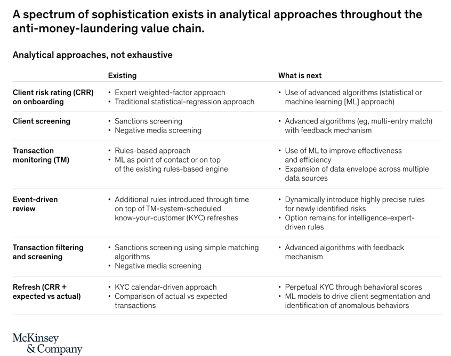
**Cartel and organized crime:** Cartels are groups of businesses that collude to manipulate markets, reduce competition, and fix prices. Organised crime groups engage in a range of criminal activities, including financial fraud, money laundering, and drug trafficking (Chen, 2023).

# What Data sources are available?

As seen in the media, there is an exponential growth in financial crimes around the world. Criminal organisations and individuals are using more sophisticated techniques to evade detection. Financial institutions and governments need to use sophisticated machine learning techniques to fight financial crime.

“Banks that innovate and adopt new technologies and techniques to address regulatory compliance demands will be industry leaders in the years to come” (Delzoppo 2022)

Hence the fight against financial crime needs voluminous and high-quality data. **More quality data enhances model performance.**



These more complex models require not only internal data sources but also external data sources. Data sources for financial crime in Australia can come from various sources, and government agencies, regulatory bodies, financial institutions, and law enforcement. These agencies are important in collecting and analysing financial crime data.

A variety of data sources can be utilised to study and develop prevention techniques for detecting financial crime. These sources provide valuable insights into criminal patterns and behaviors’, enabling the development of effective risk assessment models and fraud detection systems (CFCS, 2023).

**Transaction Data:** Bank transfers, credit card transactions, and online payments.

**Customer Data:** Demographics, employment details, and transaction history.

**Third-Party Data:** Public records, social media, and credit reports.

**Regulatory Data:** Financial crime reports and regulatory filings.

**Law Enforcement Data:** Crime reports and investigative records.

**Data sources and agencies for financial crime in Australia:**

* Australian Transaction Reports and Analysis Centre (AUSTRAC)
* Australian Securities and Investments Commission (ASIC)
* Australian Prudential Regulation Authority (APRA)
* Australian Competition and Consumer Commission (ACCC)
* Australian Federal Police (AFP)

**Data sources for the project:**

|  |  |
| --- | --- |
| **Kaggle** | [**https://www.kaggle.com/**](https://www.kaggle.com/) |
| **Australia Bureau of Statistics** | [**https://www.abs.gov.au/**](https://www.abs.gov.au/) |
| **Australian Banking Association** | [**https://www.ausbanking.org.au/data-research/data/**](https://www.ausbanking.org.au/data-research/data/) |
| **Data.gov.au** | [**https://data.gov.au/home**](https://data.gov.au/home) |
| **Research Data Australia** | [**https://researchdata.edu.au/**](https://researchdata.edu.au/) |
| **UN data** | [**https://data.un.org/**](https://data.un.org/) |
| **World Bank data** | [**https://data.worldbank.org/**](https://data.worldbank.org/) |
| **FinCEN data** | **https://www.icij.org/investigations/fincen-files/download-fincen-files-transaction-data/** |

# What Datasets are available?

There are a variety of datasets for financial crime analysis and research from a vast number of sources ranging in a variety of formats, including structured, unstructured, and semi-structured data. The choice of datasets depends on the different categories of financial crimes being investigated. However, due to privacy concerns and regulations the availability of dataset is limited.

We have collected four financial datasets from Kaggle which consist of three synthetic datasets and one real anonymised dataset.

**Synthetic Financial Dataset for Fraud Detection**: Synthetic datasets generated by the PaySim mobile money simulator.

**Synthetic Fraudulent Transaction Detection**: Financial transaction labelled as fraudulent or legitimate.

**Synthetic IBM Transactions for Anti Money Laundering (AML)**

**Credit Card Fraud Detection Dataset 2023:** Real anonymised dataset

**FinCEN Files transaction data:** financial intelligence reports that reveals the role of global banks in industrial-scale money laundering.

**Sources of data about financial crime in Australia:**

**AUSTRAC** Financial Intelligence Reports: AUSTRAC, the Australian Transaction Reports and Analysis Centre, is a government agency responsible for combating financial crime in Australia. AUSTRAC collects financial intelligence reports (FINTRs) from financial institutions and other reporting entities, which contain information about suspicious transactions and activities. These FINTRs provide valuable insights into the nature and scope of financial crime in Australia (austrac.gov.au, 2023)

**Australian Prudential Regulation Authority** (APRA) Data: APRA, the prudential regulator for the Australian financial system, collects a wide range of data from financial institutions, including data on fraud losses, cybercrime incidents, and anti-money laundering (AML) activities. This data provides valuable insights into the risks of financial crime faced by Australian financial institutions and the effectiveness of their AML controls (Apra.gov.au, 2023)

**Australian Bureau of Statistics** (ABS) Data: The ABS, Australia's national statistical agency, collects data on a range of topics, including crime and justice. The ABS's Crime Victimisation Survey provides data on the prevalence of financial fraud among Australian households and businesses. This data can be used to track trends in financial crime over time and identify areas where prevention efforts are needed (Abs.gov.au, 2023)

**Australian Federal Police** (AFP) Data: The AFP, Australia's national police force, collects data on a range of crimes, including financial crime. The AFP's Financial Crime Data Repository contains information about financial crime investigations and prosecutions. This data can be used to identify emerging trends in financial crime and develop targeted prevention strategies (Afp.gov.au, 2023)

**Academic Research:** Researchers at universities and other research institutions conduct studies on financial crime in Australia. These studies often use data from a variety of sources, including AUSTRAC FINTRs, APRA data, ABS data, and AFP data. Academic research can provide valuable insights into the causes and consequences of financial crime and inform the development of effective prevention strategies.

# Why did we use the Synthetic Dataset to study financial crime?

According to Hendrik (2021), sourcing actual customer data is not always possible for investigation of financial crime. Our team ran into the same problems hence we used a few sources of synthetic dataset. The advantages were.

**Data Privacy:** Synthetic data eliminates the privacy concerns associated with real financial crime data, which often contains sensitive customer information. This allows researchers to safely analyses and study financial crime patterns without compromising individual privacy.

**Data Availability:** Real financial crime data is often limited and not readily available for research purposes due to privacy restrictions and data ownership issues. Synthetic data can be generated in large quantities, providing researchers with the data volume they need to conduct comprehensive studies.

**Data Control:** Synthetic data generation allows researchers to control the data distribution and characteristics, enabling them to create specific scenarios and test different hypotheses without relying on limited real-world data.

**Reproducibility:** Synthetic data can be easily reproduced and shared among researchers, facilitating collaboration and consistency in research findings.

**Machine Learning Applications:** Synthetic data is well-suited for training and evaluating machine learning models used in fraud detection and financial crime prevention. It can be used to create diverse and realistic training datasets without exposing sensitive customer information.

According to Lopez-Rojas (2023) Synthetic data is a valuable tool for studying financial crime because it provides a safe and ethical way to train AI systems to detect and prevent fraudulent activity. Traditional methods of detecting financial crime are becoming increasingly ineffective, and AI has the potential to revolutionise the fight against these crimes. Lopez-Rojas (2023) goes on to explain that training AI systems requires large, diverse, and representative datasets, which can be difficult to obtain using real-world financial data. Synthetic data can be generated to mimic real-world financial data without exposing sensitive information or introducing bias. This makes it an ideal tool for training AI systems to detect financial crimes.

**Benefits of using synthetic data in the fight against financial crime:**

* Protects sensitive information: Synthetic data does not contain real-world customer data, so it does not pose a risk of exposure to sensitive information.
* Reduces bias: Synthetic data can be generated to be representative of the population, which helps to reduce bias in AI systems.
* Simulates different scenarios: Synthetic data can be used to simulate different scenarios and edge cases, which helps to test and validate the performance of AI systems.

**Role of financial institutions in adopting synthetic data:**

According to Lopez-Rojas (2023) Financial institutions have a crucial role to play in adopting synthetic data because they have access to large amounts of financial data and the resources to train AI systems. By using synthetic data, financial institutions can:

* Improve detection of financial crimes: AI systems trained on synthetic data can detect financial crimes more effectively than traditional methods.
* Reduce costs: Using synthetic data to train AI systems can reduce the costs associated with manual review and rule-based systems.
* Stay ahead of evolving threats: As financial crimes continue to evolve; synthetic data can be used to develop AI systems that can detect and prevent new types of fraud.

Lopez-Rojas (2023) also advocates that just as a traditional vaccine trains the immune system to recognize and fight a specific virus, synthetic data can be used to train AI systems to recognise and prevent financial crimes. By providing a safe and effective way to train AI systems, synthetic data will help to protect our financial well-being from the ever-evolving threat of financial crime.

# What was the EDA that was done?

After the code to create some synthetic data was written, our team worked on ways to applying the below privacy-preserving techniques, as part of EDA:

**Differential Privacy**

We have used the synthetic datasets and written code to apply privacy preserving techniques. The code demonstrates the application of differential privacy (DP) to protect the privacy of sensitive data, specifically deposit amounts in this case. It involves adding controlled noise to the data to make it difficult for an attacker to re-identify individual transactions while still preserving the overall statistical properties of the data.

**Federated Learning**

Similarly, The code was applied to the synthetic dataset to illustrates the concept of federated learning, a distributed machine learning approach where multiple parties collaboratively train a model without sharing their individual data.

**Homomorphic Encryption**

Again, the code was applied to the Synthetic dataset to demonstrates the basic operations of homomorphic encryption, a technique that allows performing computations on encrypted data without decrypting it.

**Secure Multiparty Computation (SMC)**

The code was again applied to the synthetic dataset which implements Shamir's Secret Sharing, a technique for secure multiparty computation.

**Bank Transaction Data**

Code to start analysis on this or similar dataset. This code reads an excel file named 'bank.xlsx' and filters the data to keep only rows where the deposit amount is greater than 8000 and less than 10,000. It then saves the filtered data to a new CSV file named '001.csv'.

This code can be used to study financial crime detection by identifying potential suspicious activity based on transaction amounts. By filtering the data to include only transactions where the deposit amount is between 8,000 and 10,000, the code can identify transactions that are outside of the normal range for most account holders. This type of activity could be indicative of money laundering or other financial crimes.

**Here are some specific ways in which this code could be used to study financial crime detection:**

**Identifying potential money laundering:** Money launderers often use large cash deposits to try to hide the source of their illicit funds. By identifying transactions where the deposit amount is between 8,000 and 10,000, the code can flag these transactions for further investigation.

**Identifying potential fraud:** Fraudsters may also use large deposits to try to scam victims. By identifying transactions where the deposit amount is between 8,000 and 10,000, the code can flag these transactions for further investigation.

**Understanding patterns of suspicious activity:** By analysing the data from the filtered transactions, researchers can identify patterns of suspicious activity. This information can then be used to develop new and more effective methods of detecting financial crimes.

# What are the hurdles in obtaining financial crime data?

Many financial crimes go underreported, businesses may hesitate to report financial crime too concerned to lose reputation or public trust. Public authorities and private entities find compiling comprehensive meaningful datasets challenging. According to Pattara (2023) The main reason that a comprehensive dataset on financial crime is difficult to acquire is privacy and regulatory constraints. The second reason is data fragmentation. This means data is fragmented across many institutions.

Furthermore, obtaining financial crime data for research and analysis can be challenging due to several hurdles, including:

**Data Availability and Accessibility:** Financial institutions and law enforcement agencies often have strict data privacy and confidentiality policies that restrict access to financial crime data. This can make it difficult for researchers to obtain the data they need to conduct meaningful studies.

**Data Quality and Consistency:** The quality and consistency of financial crime data can vary depending on the source and the methods used to collect and store the data. This can make it difficult for researchers to generalize their findings and draw accurate conclusions.

**Data Integration and Harmonization:** Financial crime data can be siloed across different institutions and agencies, making it difficult to integrate and harmonize the data for comprehensive analysis. This can hinder efforts to identify patterns and trends across different types of financial crime.

**Data Sensitivity and Confidentiality:** Financial crime data often contains sensitive and confidential information, such as customer financial records and investigative details. This raises concerns about data privacy and the potential for misuse of the data.

# What are the solutions for it?

Creating synthetic datasets for financial crime when real datasets are unavailable due to privacy concerns and regulations can be a viable solution. However, it’s crucial to note that the effectiveness of synthetic datasets depends on the quality of the generation process. The synthetic data should accurately capture statistical properties and patterns present in real financial data to ensure that models trained on such data are relevant and reliable in real-world applications.

**Potential solutions to address these challenges** (Jacob, 2023)**:**

**Establish Clear Data Sharing Guidelines and Protocols:** Develop clear and consistent guidelines and protocols for data sharing between financial institutions, law enforcement agencies, and research institutions. These guidelines should outline data access procedures, data use restrictions, data privacy and confidentiality measures, and mechanisms for dispute resolution.

**Promote Data De-identification and Anonymization:** Encourage the use of data de-identification and anonymization techniques to protect the privacy of individuals while preserving the utility of the data for research and analysis. De-identification involves removing or modifying personally identifiable information, while anonymization involves transforming data to render it statistically unidentifiable.

**Invest in Data Infrastructure and Expertise:** Allocate resources to develop and maintain robust data infrastructure that can securely store, manage, and process large volumes of financial crime data. Invest in training and development programs to enhance the data management and analysis expertise of researchers and data analysts working with financial crime data.

# Challenges for the coding used for this report.

**Differential Privacy:**

The code that was written used Differential privacy which is a technique for releasing statistical information about a dataset while protecting the privacy of individuals in the dataset. It works by adding noise to the data in a way that makes it difficult for an attacker to re-identify any individual in the dataset, even if they have access to the noisy data. The amount of noise added to the data is a trade-off between privacy and accuracy. More noise means greater privacy protection, but it can also make it harder to get accurate results from the data (Nguyen, 2019).

**Federated Learning:**

The code that was written used Federated learning which is a technique for training machine learning models on private data without sharing the data itself. It works by having each party train a local machine learning model on their own data and then sending the model parameters to a central server. The central server aggregates the model parameters and updates the global machine learning model. This process is repeated multiple times until the global machine learning model has learned from all the local data without any of the data ever leaving the local devices or servers. Thus, as federated learning continues to evolve, new challenges will likely emerge. These may include dealing with non-standard data formats, ensuring fairness and accountability in model decisions, and adapting to new privacy regulations and technological advancements. (Ibm.com, 2022)

According to Boesch (2023), the key challenges of Federated Learning are:

**Communication Efficiency:**

* High communication overhead: Federated networks often involve a massive number of devices with limited bandwidth and communication speeds. Sending large model updates across the network can be slow and inefficient.
* Heterogeneous systems: Devices in the network may have varying computational capabilities and network connectivity, leading to stragglers and uneven participation.

**Systems Heterogeneity:**

* Hardware variability: The storage, computational power, and communication capabilities of devices can differ significantly, leading to uneven participation and performance.
* Network connectivity: Devices may be connected through different networks (3G, 4G, 5G, Wi-Fi) with varying speeds and reliability.

**Statistical Heterogeneity:**

* Non-IID data: Devices often generate data that is not identically distributed (IID). This can lead to biased models and inaccurate predictions.
* Data distribution variability: The number of data points across devices can vary significantly, further complicating modelling and analysis.

**Homomorphic Encryption:**

The code uses Homomorphic encryption which is a technique for performing computations on encrypted data without decrypting it first. This means that you can perform computations on sensitive data without ever exposing the data itself to the untrusted environment where the computations are being performed. Homomorphic encryption is based on the idea that you can encrypt a computation and then decrypt the result. The result will be the same as if you had performed the computation on the unencrypted data (Yackel, 2021)

According to Sean (2023), the key challenges of Homomorphic Encryption are:

**Complexity and Performance:**

* Computational Overhead: Encrypting, decrypting, and performing operations on ciphertexts require significantly more resources than plaintexts, leading to slow performance and high costs. This hinders real-time applications and high-throughput scenarios.
* Limited Operations: Partially homomorphic schemes limit the types of operations supported, while fully homomorphic schemes require noise management to prevent corruption, further impacting performance.

**Security and Correctness:**

* Security Risks: Leakage of information from ciphertexts, even if limited, can pose serious security risks. Ensuring absolute security is challenging and depends on the specific scheme and its implementation.
* Correctness Concerns: Noise management techniques can introduce errors or distortions in the results, impacting the reliability and accuracy of computations on ciphertexts.

**Usability and Interoperability:**

* Lack of Standardized Frameworks: The absence of common frameworks, libraries, protocols, and benchmarks makes homomorphic encryption adoption difficult for users and developers.
* Limited Interoperability: Different schemes often lack compatibility, hindering collaboration and data sharing across platforms.

# Where are the codes stored?

All codes are saved and stored here:

<https://github.com/DataBytes-Organisation/Privacy-Technologies-for-Financial-Intelligence>

# Conclusion and Summary

The report concludes that synthetic datasets and privacy-preserving techniques offer promising solutions for addressing the challenges of obtaining and analysing financial crime data. By using these techniques, researchers and policymakers can gain valuable insights into financial crime patterns and develop more effective detection and prevention methods.

# Reference

Austrac.gov.au (2023), Accessed 4th December 2023, Link: [www.austrac.gov.au](http://www.austrac.gov.au)

Abs.gov.au (2023), Accessed 4th December 2023, Link: [www.abs.gov.au](http://www.abs.gov.au)

Afp.gov.au (2023), Accessed 4th December 2023, Link: [www.afp.gov.au](http://www.afp.gov.au)

Apra.gov.au (2023), Accessed 4th December 2023, Link: [www.apra.gov.au](http://www.apra.gov.au)

Boesch, G. (2023), ‘An Introduction to Federated Learning: Challenges and Applications’, accessed 4th December 2023, Link: <https://viso.ai/deep-learning/federated-learning/>

Chen, J. (2023), ‘What Is a Cartel? Definition, Examples, and Legality’, accessed 4th December 2023, Link: <https://www.investopedia.com/terms/c/cartel.asp>

Gottschalk, P. (2010a), “Categories of financial crime”, Journal of Financial Crime, Vol. 17 No.4, pp. 441-458.

Delzoppo, B (2022), ‘How data analytics is leading the fight against financial crime’, accessed 4th December 2023, Link: <https://www.ey.com/en_au/consulting/how-data-analytics-is-leading-the-fight-against-financial-crime>

Hendrik (2021), ‘FinML — The Promise of Synthetic Data in the Fight against Financial Crime’, medium.com, accessed 4th December 2023, link: <https://medium.com/tide-engineering-team/finml-the-promise-of-synthetic-data-in-the-fight-against-financial-crime-5e7a51240864>

Ibm.com (2023), ‘What is federated learning?’, Accessed 4th December 2023 Link: <https://research.ibm.com/blog/what-is-federated-learning>

Jacob, A. (2023), ‘Financial Analysis: A Guide for Financial Controllers & Business Leaders’, accessed 4th December 2023, Link: <https://www.linkedin.com/pulse/tackling-challenges-financial-analysis-guide-business-anil-jacob/>

Rusi.org (2023), The Centre for Financial Crime and Security Studies, accessed 4th December 2023, link: <https://www.rusi.org/explore-our-research/research-groups/centre-for-financial-crime-and-security-studies>

Lopez-Rojas, E. (2023), ‘Fighting Financial Crime with Synthetic Data​: The Financial Crime Vaccine’, accessed 4th December 2023, link: <https://www.linkedin.com/pulse/fighting-financial-crime-synthetic-data-vaccine-lopez-rojas-phd/>

McKenzie, N. Baker, R. and Mitchell, G. (2017) ‘It's not just CBA: all the banks are exposed to millions in money laundering’, Accessed 4th December 2023 Link: <https://www.smh.com.au/business/banking-and-finance/its-not-just-cba-all-the-banks-are-exposed-to-millions-in-money-laundering-20170914-gyhhpi.html>

Nadig M, (2023), ‘Best Practices in Financial Crime Data Management’, flagright.com accessed 4th December 2023 Link: <https://www.flagright.com/post/best-practices-in-financial-crime-data-management>

Nguyen, A. (2019), ‘Understanding Differential Privacy’, towardsdatascience.com, accessed 4th December 2023 Link: <https://towardsdatascience.com/understanding-differential-privacy-85ce191e198a>

Pattara, M. (2023), ‘4 challenges facing FCC teams, and how to tackle them’, accessed 4th December 2023 Link: <https://www.napier.ai/post/challenges-financial-crime-compliance>

Sean, G. (2023),’What are the main challenges and limitations of homomorphic encryption in practice?’, accessed 4th December 2023 Link: <https://www.linkedin.com/advice/0/what-main-challenges-limitations-homomorphic>

United Nations: Office on Drugs and Crime.<https://www.unodc.org/unodc/en/money-laundering/overview.html#:~:text=The%20estimated%20amount%20of%20money,Terrorist%20Financing>

Yackel, R. (2021), ‘What is homomorphic encryption, and why isn’t it mainstream?’, keyfactor.com, accessed 4th December 2023 Link: <https://www.keyfactor.com/blog/what-is-homomorphic-encryption/>