Sekgoka et al., *Cogent Engineering* (2022), 9: 2046680
https://doi.org/10.1080/23311916.2022.2046680

cogent engineering

COMPUTER SCIENCE | RESEARCH ARTICLE

Privacy-preserving data mining of cross-border financial flows

Chaka Patrick Sekgoka[1]*, Venkata Seshachala Sarma Yadavalli[1] and Olufemi Adetunji[1]

Received: 06 October 2020
Accepted: 21 February 2022

*Corresponding author: Chaka Patrick Sekgoka, Department of Industrial and Systems Engineering, University of Pretoria, South Africa
E-mail: Chakasekgoka@gmail.com

Reviewing editor:
D T Pham, School of Mechanical Engineering, University of Birmingham, Birmingham, United Kingdom

Additional information is available at the end of the article

Abstract: Criminal networks continue to utilize the global financial system to launder their proceeds of crime, despite the broad enactment of anti-money laundering (aml) laws and regulations in many countries. Money laundering consumes capital resources and the tax revenue needed to fund infrastructure development and alleviate poverty in developing market economies. This paper, therefore, expands on the tools available for enabling privacy-preserving data mining in multi-dimensional datasets to combat cross-border money laundering. Most importantly, this paper develops a novel measure for detecting anomalies in cross-border financial networks, allowing financial institutions and regulatory organizations to identify suspicious nodes. The research used a sample dataset comprising international financial transactions and a hypothetical dataset to demonstrate the measure of node importance and the symmetric-key encryption algorithm. The results support the argument that the proposed network measure can detect node anomalies in the cross-border financial flows network, enabling regulatory authorities and law enforcement agencies to investigate financial transactions for suspicious activity and criminal conduct. The encryption algorithm can ensure adherence

Chaka Patrick Sekgoka

ABOUT THE AUTHORS

Chaka Patrick Sekgoka received his PhD in Industrial Systems from the University of Pretoria, South Africa. His research interests include data mining, predictive analytics, machine learning, complex networks, statistics and probability. He is an independent consultant at EQPlus Technologies (Pty) Ltd, where he is involved in digitization projects in the Financial Services sector.

Venkata Seshachala Sarma Yadavalli is a professor and Acting Head of the Department of Industrial & Systems Engineering at the University of Pretoria, South Africa. He is on the editorial board of various national and international journals. He has received numerous awards from several professional bodies for his research contributions.

Olufemi Adetunji is an associate professor at the Department of Industrial and Systems Engineering at the University of Pretoria, South Africa. He has published many articles in local and international journals. His research interests are in supply chain design and engineering, lean manufacturing and applied optimisation.

PUBLIC INTEREST STATEMENT

Money laundering poses a significant economic challenge in emerging markets, reducing the tax revenue needed to fund infrastructure development and poverty alleviation programs. Despite the recent advances in technology and the availability of financial transaction datasets, money laundering-related investigations are narrowly focused on incidents and generally triggered by tip-offs. This research proposes a network model to combat illegal cross-border fund transfers while preserving the privacy of personally identifiable information. The study leveraged data mining methods and regulatory policies to identify suspicious transaction patterns in financial transactions between residents and non-residents. The research used a sample dataset drawn from the South African database of international financial transactions to illustrate the proposed privacy-preserving data mining approach. Supervisory authorities can use the model to define and plan inspections of regulated entities in a cost-effective manner. Financial institutions can use the model to enhance compliance monitoring and risk management functions.

to information privacy laws and policies without compromising data reusability. Hence, the proposed methodology can improve the proactive management of money laundering risks associated with cross-border fund flows for the global financial system's benefit.

Subjects: Algebra; Statistics; Computational Logic; Computer Science (General); Data Preparation & Mining

Keywords: Information privacy; symmetric-key encryption; bipartite graph; cross-border financial flows; centrality; anti-money laundering

1. Introduction

The global financial system is subject to a wide range of risks and vulnerabilities exploited by criminal networks to launder their proceeds of crime with a relatively low risk of detection. One example of such a threat is the voluminous and volatile cross-border financial flows, obscuring individual transactions and providing opportunities for criminal networks to transfer funds across country borders.

Many countries have adopted the Financial Action Task Force (FATF)'s internationally endorsed standards, providing a comprehensive set of counter-measures against money laundering (Cox, 2014). Most financial institutions' automated AML systems embed the FATF standards, enabling the built-in transaction-specific triggers to identify suspicious transaction in real time. However, on many occasions, the flagged cases turn out to be false positives (Pourhabibi et al., 2020).

Money launderers are often aware of the events that trigger suspicious transactions and circumvent them using advanced transaction layering techniques and methods, such as the "straw man," sophisticated documentation, and consulting firms (Harvey, 2004; Teichmann, 2019; Van Duyne, 1994; Walker, 1999). The straw man fallacy disguises the beneficiary's identity, which is the focus of most compliance procedures that banks and other financial institutions implement.

Using graph-based substructures and measures for detecting money laundering activities is a large area of network theory research, with several measures already proposed (Sun et al., 2005b; Li et al., 2020; Xiong et al., 2010; Zhiguo et al., 2015). However, the proposed measures are not easily extensible to directed and dual-weighted networks such as the cross-border financial flow network (Akoglu et al., 2015). In addition, the standard metrics for weighted networks (such as degree, closeness, and betweenness centrality) have solely focused on tie weights and not on the number of ties.

Researchers have proposed the measure of node importance for social networks, combining tie weights with the number of ties (Opsahl et al., 2010). The metric takes the form:

$$C_D^{W\alpha}(i) = k_i^{(1-\alpha)} s_i^{\alpha} \tag{1}$$

where α is a positive tuning parameter. If α is set between 0 and 1, then having a high degree is favorable, whereas if it is set above 1, then a low degree is favorable. Notably, the measure proposed in Equation 1 considers one network weight. Therefore, it is not extensible to the proposed directed and dual-weighted networks.

This paper closes this research gap by proposing a network model for directed and dual-weighted networks, along with a measure of node importance that combines tie weights and the number of ties. The proposed metric addresses the question, "Which are the most important or central nodes in the cross-border financial flow network?". In Section 4, we compare the metric in

Equation 1 with the measure proposed in this paper at different levels of α. The results show that the measure proposed in this paper can detect suspicious transaction activity involving multiple high-volume flows of funds from source to destination accounts.

The lack of research about money laundering involving cross-border transactions is mainly attributable to data access and sharing restrictions. Governments and firms use multitudes of regulations, laws, and best practices to protect datasets comprising private and confidential information. Hence, researchers recommend developing techniques incorporating privacy concerns as a fruitful direction for future data mining research (Agrawal & Srikant, 2000; Qi & Zong, 2012; Xu et al., 2014).

The three broad categories for privacy-preserving data mining are data obfuscation, summarization, and data separation (Adam & Wortmann, 1989; Cios & Moore, 2002; Clifton & Vaidya, 2004; Dwork et al., 2006; Kou et al., 2007). The development of both the cryptographic and machine learning methods and their integration with the three broad categories of privacy-preserving techniques have been a subject of research interest in recent years (Pathak et al., 2010, 2011; Wang et al., 2018).

This paper leverages advanced technology to develop a symmetric-key encryption algorithm at the intersection of classical cryptography and data obfuscation methods, encrypting the explicit variables from the financial transaction datasets, such as resident name and resident address. The motive for developing the symmetric-key encryption algorithm is two-fold: To leverage the group structure of the multi-dimensional dataset to transform the individual's personally identifiable information (PII) without losing data reusability. Second, we compute the dual weights of the cross-border financial flow network.

The organization of the paper is as follows: this section provides background on cross-border financial flows and AML. Section 2 presents the proposed encryption algorithm, followed by the proposed network structure of cross-border financial flow model along with the measure of node importance. Section 3 illustrates the workings of the symmetric-key encryption algorithm, network visualization and the proposed network measure. The last section concludes the paper.

1.1. Background

1.1.1. Cross-border financial flows

Cross-border financial flows are money transfers made by a resident to a non-resident and vice versa because of financial transactions involving individuals, private and public firms, central banks, financial institutions, as well as legal entities such as trusts and non-profit organisations or a combination thereof, in at least two different countries. The banking sector plays an important role in channelling cross-border flows in a country.

Figure 1 depicts the flow of cross-border transactional data between South African residents and the rest of the world. The authorized dealer network (comprising commercial banks and other licensed financial institutions) facilitates cross-border payments between residents and non-residents, through the corresponding bank relationships abroad. Hence, a distinctive feature of cross-border financial flows is the existence of financial transactions between residents and non-residents of a country.

Central banks and statistical agencies record cross-border flows for Balance of Payment (BoP) reporting and other regulatory purposes. The significance of the central bank's databases is that they provide transactional information concluded at different financial institutions by the same resident/non-resident. The transactional data comprise PII such as a phone number, email address, social security number, and residential address that one can use to identify a specific individual. Importantly, many countries use information privacy laws and policies to protect such sensitive data.

Most research studies focus on the impact of regulatory policies relating to cross-border flows on economic growth (Babus, 2016; Neanidis, 2019; Ostry et al., 2012; Silva et al., 2016). The Latin

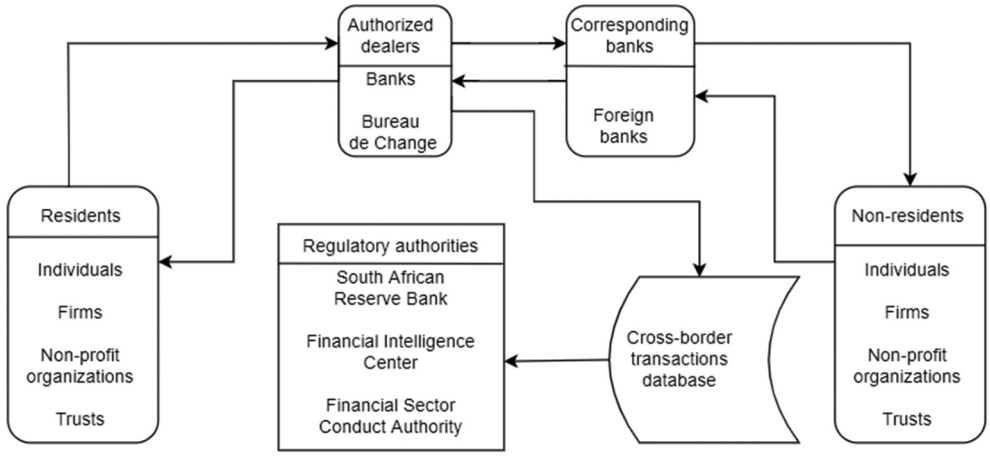

Figure 1. A depiction of cross-border financial transactions data flow in South Africa

American debt crisis of the early 1980s, the 1997 East Asian financial crisis, the 1998 Russian financial crisis, and the 2008 global financial crisis all highlighted the risks associated with cross-border flows' volatility.

Cross-border flows increased substantially in recent years due to rapid developments in financial technology (FinTech) innovations. The FinTech innovations drive productivity growth due to efficient payment systems and reduced online transaction costs (Freund & Weinhold, 2004; Meltzer, 2015; Neanidis, 2019). Migrant worker remittances also play an increasingly prominent role in the economies of many nations, enhancing financial transactions between their home and host countries. Detecting and preventing money-laundering activity in the remittances industry is a crucial area of regulatory concern and a focus of this research.

1.1.2. Anti-money laundering

AML refers to a set of laws, regulations and procedures intended to deter criminals from using the financial sector to disguise cash proceeds from illegal activities as legitimate. Global standards issued by the FATF enable countries to adopt a more flexible set of AML measures. The standards recommend using advanced technology and data mining methods to identify suspicious transactions, requiring no monetary thresholds for reporting suspicious and unusual transactions to regulatory authorities (Cox, 2014; FATF, 2014).

Limited information is available on the costs and benefits of implementing technology for detecting and impeding money laundering, partly due to the difficulties of estimating the volume of money laundering. However, many financial institutions continue to derive business value from the widely available AML systems. Researchers concluded that investment in advanced technology appears to be a cost burden instead of enhancing the deterrence of money laundering (Kang, 2018; Magnusson & Harvey, 2009).

2. Materials and methods

2.1. Symmetric-key encryption algorithm using temporary variables

Advances in wireless technology continue to create exponential growth in connected devices, leading to the internet of things (IoT) revolution. IoT comprises millions of connected devices that can sense, compute and communicate, resulting in significant information/data security concerns. Cryptographic methods are primarily used to address such data security concerns, with several proposed algorithms (Deshkar et al., 2017; Sreeja et al., 2019).

This paper proposes an encryption algorithm utilizing temporary enumeration variables generated automatically during the compilation phase of a computer program in order to derive a permutation. The derived permutation is a lookup table. Hence, the proposed technique is analogous to the Permutation Cipher (Stinson & Paterson, 2006). In addition to deriving a permutation, the algorithm uses the temporary variables to compute the weights of the directed and dual-weighted bipartite network. The proposed algorithm executes can be executed quickly due to its simplicity. Figure 2 depicts the proposed symmetric-key encryption algorithm.

Several software environments for statistical computing, such as SAS® and R programming language, provide packages for BY-group processing, a technique used to process data grouped by values of one or more common variables. This paper illustrates the proposed symmetric-key encryption algorithm using the SAS® programming language.

2.1.1. Description of variables

A detailed description of the automatically generated temporary variables (N, FIRST.*variable*, LAST.*variable*) and other variables used for BY-group processing is as follows:

(1) **N** is a counter variable that records the record number being processed in the dataset. Its initial value is set to 1 and is incremented by one whenever a new record is processed.

(2) **BY variables** are the PII variables by which the dataset is sorted or indexed.

(3) **BY values** are the values of the BY variables.

(4) **BY groups** are distinct groups of records with the same BY values. A single BY group divides the records of a BY variable by its BY values.

(5) **FIRST.*variable*** is a Boolean mapping on the BY group variable, which has a true value if the processing is done on the first record of the BY group and false value otherwise.

(6) **LAST.*variable*** is a Boolean mapping on the BY group variable, which has a true value if the processing is done on the last record of the BY group and false value otherwise.

2.1.2. Procedure—Part one

(1) Input the original multi-dimensional dataset.

(2) Sort the dataset by the encryption variable to enable the creation of the BY groups.

(3) FIRST.*variable* and N are automatically set to true at the start of dataset processing.

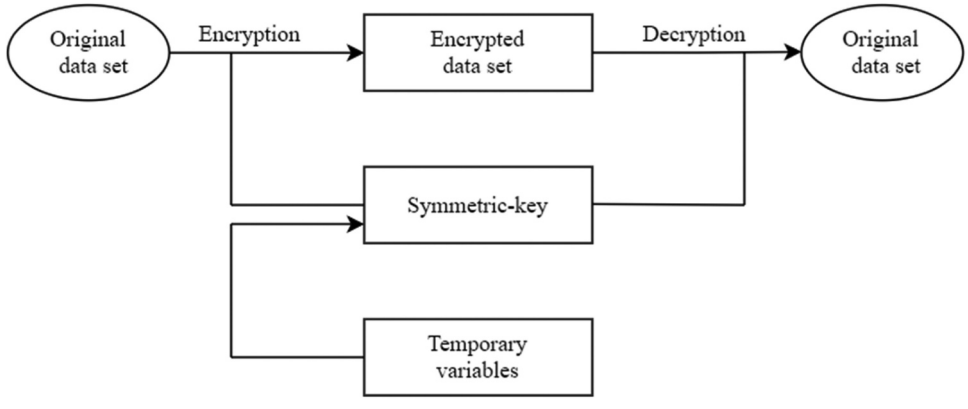

Figure 2. Symmetric-key encryption using temporary variables

(4) If the BY value of the next record equals the BY value of the current record, set LAST.*variable* to false and true otherwise.

(5) If FIRST.*variable* is true, concatenate the first letter of the encryption variable with the value of N to obtain the encrypted variable. Retain the value of the encrypted variable.

(6) N automatically increments by one.

(7) If the BY value of the current record equals the BY value of the previous record, then set FIRST.*variable* to false and true otherwise.

(8) Return to Step (4).

(9) Stop after processing the last record of the dataset.

(10) Repeat the algorithm from Step 1 through Step 9 until all the encryption variables have been encrypted. The resulting dataset is the symmetric key for both encryption and decryption of the PII variables.

(11) To obtain the encrypted dataset, drop the PII variables from the dataset in Step 10.

2.1.3. Procedure—Part two

(1) Input the dataset from Part One.

(2) Sort the dataset by the encrypted variables to enable the creation of BY groups.

(3) FIRST.*variable* and N are automatically set to true at the start of dataset processing.

(4) Initialize transaction count and transaction amount to zero.

(5) If the BY value of the next record equals the BY value of the current record, then set LAST.*variable* to false and true otherwise.

(6) Increment transaction count by one and increment transaction amount by its current value.

(7) If LAST.*variable* is true, output the current record.

(8) N automatically increments by one.

(9) Return to Step (4).

(10) Stop after processing the last record of the dataset.

2.1.4. Advantages and disadvantages of the symmetric-key encryption algorithm

The algorithm's decryption operation uses a technique similar to the Permutation Cipher; hence, it is not computationally intense. The algorithm does not provide descriptive statistics of a demographic nature, thereby reducing its susceptibility to linkage attacks.

The algorithm's safety depends on the security of the channel used to exchange the decryption key. However, it is essential to note that it is technically impossible to stop a person who is duly authorized to access confidential information from improperly disclosing that information to someone else. The proposed algorithm is not suitable for encrypting and decrypting live databases due to its requirement to store the symmetric key. Its effectiveness is limited to multi-dimensional datasets due to the group structure of such datasets.

2.2. Cross-border financial flows network

2.2.1. Network structure and representation

Most graph-based models for countering money laundering activity consider single-step transfers, ignoring the multiple financial transactions concluded through different institutions by individuals or firms (Tang & Yin, 2005; Liu et al., 2017; Lv et al., 2008; Paula et al., 2016; Prakash et al., 2010; Rajput et al., 2014). Recently, researchers proposed a multipartite network model capable of detecting money laundering involving high-volume flows of funds from source to destination accounts via layers of middle accounts (Li etal. 2020; Sun et al., 2021).

This paper proposes a directed and dual-weighted graph with weights representing the monetary value and volume of transactions, accounting for the dependencies between financial transactions while focusing on the beneficiary's identity. The cross-border financial flow network structure is similar in design to the citation networks, which enables researchers to quickly identify the important literature in a specific field within a relatively short time and with less effort. The cross-border financial flow network allows financial institutions and regulatory organizations to quickly identify the important residents/non-residents in enormous datasets comprising international financial transactions.

Formally, the cross-border financial flow network is defined as the directed and weighted bipartite graph $G = (V, A, w)$, with $V(G) = V_R \cup V_{NR}$ and $A(G) \subseteq (V_R \times V_{NR}) \cup (V_{NR} \times V_R)$, where the disjoint sets $V_R = \{r_1, \ldots, r_k\}$ and $V_{NR} = \{nr_1, \ldots, nr_p\}$ represent the resident vertex set and the non-resident vertex set with $|V_R| = k$ and $|V_{NR}| = p$, respectively. The set $A(G)$ represents the direction of financial flows, where the outward payments flow from residents to non-residents and the inward payments flow from non-residents to residents. The weight function computes the sum of transaction counts and the sum of the financial value of transactions.

Figure 3 shows a schematic depiction of the cross-border financial flow network with $k = 5$ and $p = 9$. The weight function a_{ij} denotes the total number of transactions from resident r_i to non-resident nr_j, whereas a'_{ij} denotes the total number of transactions from non-resident nr_j to resident r_i. Similarly, the network structure could be depicted with the weight functions b_{ij} and b'_{ij} representing the total financial value of transactions from residents to non-residents, respectively, vice-versa or both.

To obtain the adjacency matrix representation of the cross-border financial flow network, we denote the set of $k \times p$ matrices with non-negative real entries by $R^{k\times p}$ and arrange the node set $V_R \cup V_{NR}$ in the order $r_1, \ldots, r_k, nr_1, \ldots, nr_p$. The adjacency matrix comprises elements of $R^{k\times p}$ with entries $A = \{a_{ij}\}$ and $B = \{b_{ij}\}$, such that:

$$A = \begin{cases} a_{ij} & \text{if resident } r_i \text{ transferred funds to non} - \text{resident } nr_j \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$\text{where } a'_{ij} = \text{total number of transactions associated with the edge}(nr_j \rightarrow r_i) \tag{3}$$

$$B = \begin{cases} b_{ij} & \text{if } a_{ij} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$\textit{where } b_{ij} = \text{total financial value of transactions associated with the edge}(r_i \rightarrow nr_j) \tag{5}$$

Similarly, entries of matrices $A' \in R^{p\times k}$ and

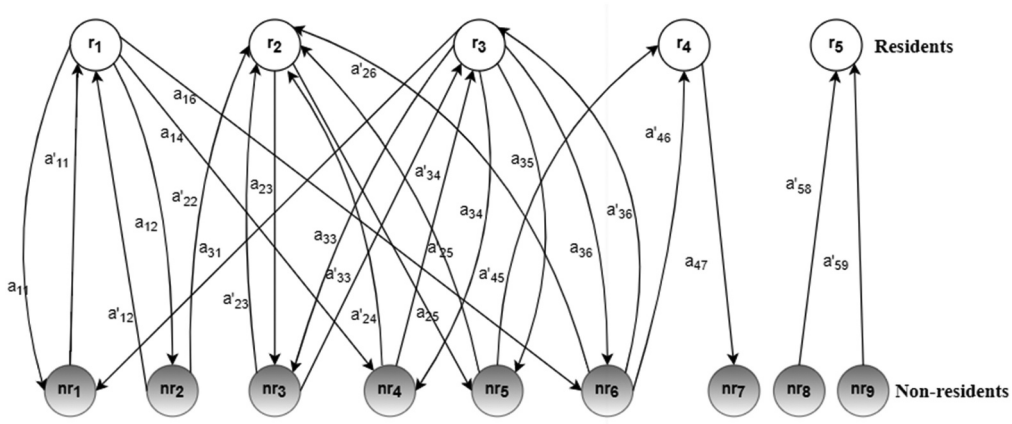

Figure 3. Schematic depiction of the cross-border financial flows network

$B \in R^{p\times k}$ are such that:

$$A' = \begin{Bmatrix} a'_{ij} & \text{if non} - \text{resident } nr_j \text{ transferred funds to resident} r_i \\ 0 & \text{otherwise} \end{Bmatrix} \quad (6)$$

where

$$a'_{ij} = \text{total number of transactions associated with the edge}(nr_j \rightarrow r_i) \quad (7)$$

and (7)

$$B' = \begin{Bmatrix} b'_{ij} & \text{if } a'_{ij} > 0 \\ 0 & \text{otherwise} \end{Bmatrix} \quad (8)$$

$$\text{where} b'_{ij} = \text{total financial value of transactions associated with the edge}(nr_j \rightarrow r_i) \quad (9)$$

Using the transaction counts as weights, the adjacency matrix F of the cross-border financial flows network is of the form:

$$F = \begin{bmatrix} 0_{k,k} & A \\ A & 0_{p,p} \end{bmatrix} \quad (10)$$

where $0_{k,k}$ and $0_{p,p}$ represent the $k \times k$ and $p \times p$ zero matrices.

The adjacency matrix representation of the cross-border financial flow network is inefficient due to the large number of zero entries. Hence, this paper uses a list of financial transaction records, discarding the zero entries.

2.2.2. Centrality measure based on two weights

The proposed centrality measure uses the matrix multiplication method to identify the nodes responsible for unusual transaction patterns. Such nodes include the multiple residents who transfer funds to the same non-resident and vice versa, as well as resident nodes transacting large financial values using high transaction volumes. Criminal networks often use these strategies to avoid thresholds and triggering alerts.

Formally, consider the dual weights of the cross-border financial flow network, A and B. Let

$$W_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj} = \sum_{k=1}^{n} A_{ik} B_{jk}^{T} \tag{11}$$

where $n = k \times p$. Hence, $W = AB^T$ is the matrix with entries W_{ij} that are the product of the rows of A and the columns of B^T.

Define the centrality measure for node i as

$$C_i = \frac{1}{W} \sum_{j} W_{ij} \tag{12}$$

where W is the sum of all the entries of the matrix W and $\sum_{i} C_i = 1$.

The diagonal entries C_{ii} are the sum of the product of transaction volume and financial value for Resident i. Large diagonal entries indicate node dominance, which could be due to significant fund transfers or multiple transactions (or a combination thereof). Hence, the diagonal elements provide a mechanism for identifying nodes with i) large volume and sizeable financial value, ii) large volume and low financial value and iii) low volume and sizeable financial value. The onus is on the financial institutions and regulatory organizations to verify the financial flows in the event of extreme importance.

If $C_{ij} > 0$ for $i \neq j$, then Resident i and Resident j transferred funds to the same non-resident during the period. In that case, C_{ij} equals the product of the number of transactions for Resident i and the financial value of transactions for Resident j. Hence, the centrality measure for Resident i increases with Resident i′s increasing neighbours in the network. It is not the absolute value of the measure that matters but the high or low centrality measure of each node.

The centrality measure based on matrices A and B can shed some light on the importance of each of the resident nodes in the cross-border financial flow network. The centrality measure for non-resident nodes is similarly defined, where $C = A'B'^{T}$, measuring the importance of non-resident nodes in the cross-border financial flow network.

3. Results

This section illustrates the symmetric-key encryption algorithm as well as the proposed centrality measure. It makes use of a hypothetical dataset comprising cross-border financial flows, structurally similar to the dataset extracted from the South African Reserve Bank (SARB)'s international financial transaction database. We also present a visualization of the cross-border financial flow network using SARB's dataset. The section concludes with a comparative analysis of the proposed centrality measure with the metric defined in Equation 1, but for directed networks (Opsahl et al., 2010).

$$C_{D-Out}^{W\alpha}(i) = k_i^{Out} \times \left(\frac{s_i^{Out}}{k_i^{Out}} \right)^{\alpha} \tag{13}$$

3.1. Encrypting the cross-border financial flows dataset

A computer program generated using SAS® Enterprise Guide 7.1, Copyright© 2014, SAS Institute Inc., Cary, NC, USA, was used to encrypt both the hypothetical dataset and the dataset drawn from the SARB. The SARB data extract contained 28,649,763 financial transaction records for the 2014–2015 calendar years, comprising six data fields, namely: resident name, transaction date, flow date, non-resident name, BoP category and transaction amount.

Table 1 shows a sample of 10 network observations from the encrypted list of international financial transactions, together with a display of the transient variables generated during the compilation phase of the SAS® program. To interpret the network data in Table 1, consider observation number 2000 in the network data (first observation in Table 1). In this case, a resident labelled r10075604 paid a non-resident labelled nr15213089 an amount of USD 46.40 in a single transaction. The second observation in Table 1 shows that the same resident paid non-resident nr15557964 a total amount of USD1430.84 in 19 financial transactions during the period. The table shows the state of the transient variables for illustration purposes.

Table 2 shows the encrypted hypothetical dataset. The table contains the original PII variables; hence, it is the symmetric key for decryption. The final step of the encryption phase drops the PII variables from Table 2.

2.2. Adjacency matrix representation and network visualization

Table 3 shows the adjacency matrix representation of the cross-border financial flow network constructed from the hypothetical dataset. The table entries are the co-ordinate pairs (x, y), representing the total transaction volume and the total financial value, respectively.

To interpret the adjacency matrix, consider the entry in the ninth column of the third row (2,550). This entry indicates that the resident labelled R8 paid the non-resident labelled NR7 the total amount of USD 550 in two transactions, corresponding to the total number and financial value of transactions made by Linda in Table 2.

The cross-border financial flows network's visualization depicts the directed and weighted bipartite graph comprising two disjoint nodes. Figure 4 shows the visualization based on the dataset drawn from the SARB's database, created using SAS® Visual Analytics software, 7.4. Copyright 2014–2017, SAS Institute Inc., Cary, NC, USA. The links connect the nodes of different colors.

3.3. Network measure for the cross-border financial flows network

We compute the product matrix W from the two matrices A and B shown on the top right block of Table 3, representing the network weights of the hypothetical dataset. The entries of the normalised product matrix are as follows:

$$W = \begin{bmatrix} 0.0493 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.4276 & 0 & 0 & 0 & 0 & 0.0214 \\ 0 & 0 & 0.0362 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.12 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.0905 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.0067 & 0 \\ 0 & 0.2303 & 0 & 0 & 0 & 0 & 0.0132 \end{bmatrix} \tag{14}$$

Adding the entries of each row of the matrix W yields the proposed centrality measure as follows:

Table 1. The encrypted hypothetical dataset of cross-border financial flows

Obs	Resident node	Flow status	BoP reporting category	Non-resident node	FIRST. Resident	FIRST. Non-resident	LAST. Resident	LAST. Non-resident	Number of transactions	Total amount
2000	r10075604	OUT	416-00	nr15213089	1	1	0	1	1	46.40
2001	r10075604	OUT	416-00	nr15557964	0	0	0	1	19	1430.84
2002	r10075604	OUT	416-00	nr15753545	0	0	0	1	3	242.47
2003	r10075604	OUT	416-00	nr16060245	0	1	0	1	1	144.94
2004	r10075604	OUT	416-00	nr19536646	0	0	0	1	4	520.38
2005	r10075604	OUT	416-00	nr20319362	0	1	0	1	1	20.62
2006	r10075604	OUT	416-00	nr23995122	0	0	0	1	34	2355.24
2007	r10075604	OUT	416-00	nr26584871	0	0	1	1	2	62.03
2008	r10075669	OUT	416-00	nr13538886	0	0	1	1	35	5718.05
2009	r10075708	OUT	416-00	nr15905119	1	1	0	1	1	302.03

cogent engineering

Table 2. The symmetric key for decrypting the hypothetical dataset

Obs	Resident name	Transaction date	Flow type	Non-resident name	Amount	Resident label	Non-resident label
1	Christo	2018/03/12	In	Benjamin	2500	R1	NR1
2	Martina	2018/03/30	In	Benjamin	1000	R13	NR1
3	Martina	2018/03/31	Out	Benjamin	2250	R13	NR1
4	Martina	2018/04/23	In	Benjamin	1250	R13	NR1
5	Rosalia	2018/05/05	In	Benjamin	1000	R18	NR1
6	Christo	2018/06/05	Out	Catalina	1000	R1	NR6
7	Linda	2018/07/11	Out	Diego	275	R8	NR7
8	Linda	2018/08/08	Out	Diego	275	R8	NR7
9	Lynn	2018/01/16	Out	Joaquin	3500	R10	NR9
10	Martina	2018/08/17	Out	Lucas	500	R13	NR10
11	Elizabeth	2018/06/15	Out	Mariana	1000	R4	NR11
12	Elizabeth	2018/06/18	Out	Mariana	5000	R4	NR11
13	Lynn	2018/07/14	In	Mariana	750	R10	NR11
14	Rosalia	2018/04/18	Out	Mariana	250	R18	NR11
15	Michael	2018/06/10	Out	Matias	500	R17	NR15
16	Christo	2018/02/07	Out	Mavis	500	R1	NR16
17	Rosalia	2018/03/12	In	Sara	200	R18	NR17
18	Elizabeth	2018/01/25	Out	Victoria	1000	R4	NR18
19	Elizabeth	2018/02/14	In	Victoria	250	R4	NR18
20	Lynn	2018/03/04	In	Victoria	100	R10	NR18
21	Rosalia	2018/07/02	In	Victoria	150	R18	NR18
22	Rosalia	2018/09/02	Out	Victoria	150	R18	NR18

Table 3. Adjacency matrix representation of the cross-border financial flow network with dual weights

	R1	R4	R10	R13	R17	R18	NR1	NR6	NR7	NR9	NR10	NR11	NR15	NR16	NR17	NR18
R1	0	0	0	0	0	0	0	(1,1000)	0	0	0	0	0	(1,500)	0	0
R4	0	0	0	0	0	0	0	0	0	0	0	(2,6000)	0	0	0	(1,1000)
R8	0	0	0	0	0	0	0	0	(2,550)	0	0	0	0	0	0	0
R10	0	0	0	0	0	0	0	0	0	(1,3500)	0	0	0	0	0	0
R13	0	0	0	0	0	0	(1,2250)	0	0	0	(1,500)	0	0	0	0	0
R17	0	0	0	0	0	0	0	0	0	0	0	0	(1,500)	0	0	0
R18	0	0	0	0	0	0	0	0	0	0	0	(1,250)	0	0	0	(1,150)
NR1	(1,2500)	0	0	(2,2250)	0	(1,1000)	0	0	0	0	0	0	0	0	0	0
NR6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NR7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NR9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NR10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NR11	0		(1,750)	0	0	0	0	0	0	0	0	0	0	0	0	0
NR15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NR16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NR17	0	0	0	0	0	(1,200)	0	0	0	0	0	0	0	0	0	0
NR18	0	(1,250)	(1,100)	0	0	(1,150)	0	0	0	0	0	0	0	0	0	0

Table 4. Degree centrality measure based on the two individual network weights and the combined weights

					C_D^{Wa} when $\alpha =$			
Node	Out-degree	W_1	W_2	C_i	0	0.5	1	1.5
R1	2	2	1 500	0.049	2	54.8	1 500	41,079
R4	2	3	7 000	0.449	2	118.3	7 000	414,126
R8	1	2	550	0.036	1	23.5	550	12,899
R10	1	1	3 500	0.115	1	59.2	3 500	207,063
R13	2	2	2 750	0.090	2	74.2	2 750	101,973
R17	1	1	500	0.016	1	22.4	500	11,180
R18	2	2	400	0.243	2	28.3	400	5 657

$$C_i \begin{bmatrix} R_1 \\ R_4 \\ R_8 \\ R_{10} \\ R_{13} \\ R_{17} \\ R_{18} \end{bmatrix} = \begin{bmatrix} 0.049 \\ 0.449 \\ 0.036 \\ 0.115 \\ 0.090 \\ 0.016 \\ 0.243 \end{bmatrix} \tag{15}$$

Table 4 shows the results of four centrality measures used for computing node importance in the cross-border financial flow network, based on node degree, transaction volume, transaction value, as well as a combination of transaction volume and monetary value. We obtained the latter results from Equation 15. In addition, Table 4 includes the results obtained using Equation 13 for comparison purposes.

All the measures indicate that R4 (Elizabeth) is more important than others in the hypothetical network due to the large financial value of transactions as expected. The exciting node is R18 (Rosalia), considered to have minor centrality points due to its low monetary value, using the measure $C_D^{W\alpha}$. However, C_i considers node R18 as the second-most important node due to its connection with R4. Furthermore, note that $C_D^{W\alpha}$ allocates more centrality points to R10 than R13 when $\alpha = 0.5$, but quickly reverses their level of importance as the tuning parameter (α) approaches one (1). The desirable property of the proposed centrality measure is its ability to allocate centrality points to nodes connected to other vital nodes in the network.

The proposed measure's ability to identify highly connected nodes can enable analysts to exploit the cross-border financial flow network structure to understand node behaviors. For example, it can be possible to locate sub-networks in Figure 4. Figure 5 shows one of the sub-networks of the cross-border financial flow network constructed using the sample dataset extracted from the SARB's database of international financial transactions.

4. Discussion and conclusion

The need to analyze cross-border financial transactions to extract meaningful insights from the multi-dimensional dataset while preserving PII motivated this study. The paper proposed a symmetric-key encryption algorithm at the intersection of classical cryptography and data obfuscation methods, leveraging advanced technology and the dataset's group structure to gain computational efficiencies.

Performance studies comparing the proposed algorithm with other privacy-preserving techniques are a subject for further research. The algorithm's lack of suitability to function in live databases is its primary deficiency. In addition, its secrecy is a topic for further investigation.

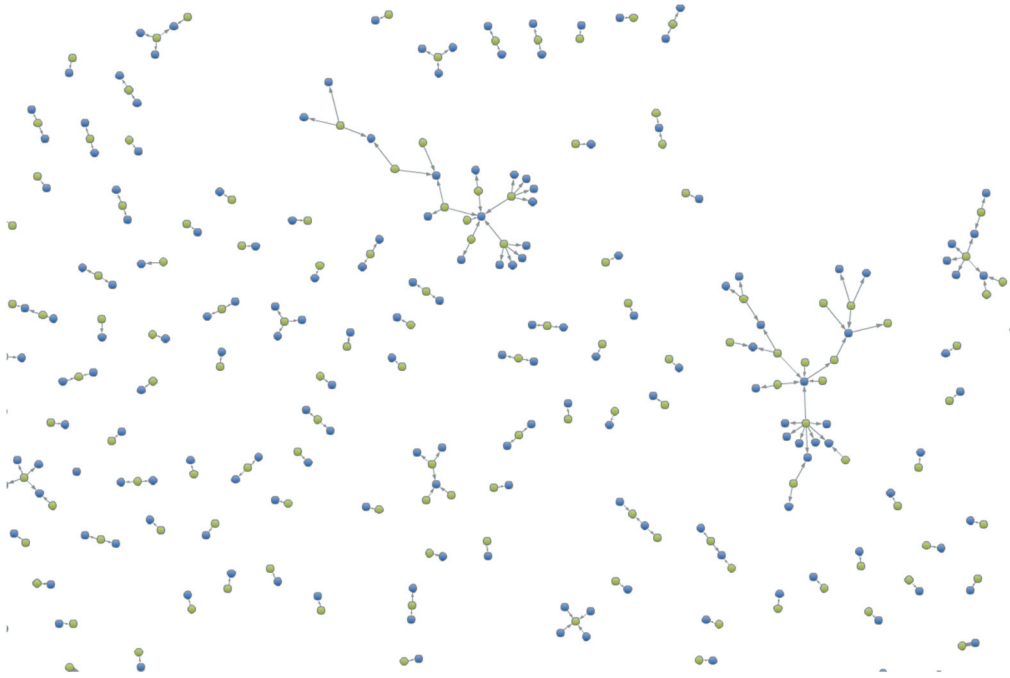

Figure 4. A snapshot of the visualization of the cross-border financial flows network based on the SARB dataset

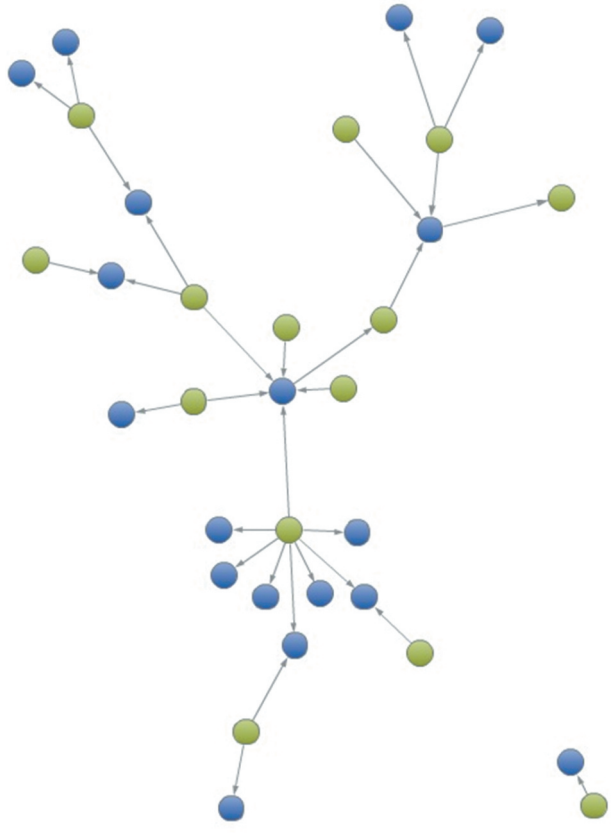

Figure 5. Sub-network of the cross-border financial flows network

The literature's lack of metrics incorporating the weights of the directed and dual-weighted network motivated the development of a new measure of node importance. The proposed metric uses the network's dual weights to allocate centrality scores. This study compared the proposed centrality measure with the existing one in literature, $C_D^{W\alpha}$, considering the number of edges and the edge weights, and a tuning parameter for controlling the relative importance of these two aspects (Opsahl et al., 2010). Experimental results indicated that while $C_D^{W\alpha}$ can identify the important nodes of the cross-border financial flow network, it fails to recognize the nodes connected to other more important nodes.

Financial institutions and regulatory organizations can use the proposed privacy-preserving data mining approach to improve the analysis and monitoring of cross-border financial transactions and curb money laundering. In particular, the tools suggested in this study can enable the implementation of the FATF recommendation stipulating that financial institutions must draw specific attention to all complex, large, and unusual transaction patterns that have no apparent economic value or visible lawful purpose.

The existing literature on exploiting nodes' communities to identify anomalies in bipartite graphs does not apply directly to the dual-weighted networks such as the cross-border financial flow network (Sun et al., 2005a; Akoglu et al., 2015). Therefore, further research is necessary to understand the network structure and dynamics of the directed and dual-weighted bipartite networks.

Acknowledgements
The authors would like to thank the SAS Institute (Pty) Ltd for providing the computer software and the SARB for availing the cross-border financial flow dataset. C. P Sekgoka is particularly grateful to Professor V.S.S. Yadavalli for his mentorship and creating a conducive working environment to conduct this research.

Author details
Chaka Patrick Sekgoka[1]
E-mail: Chakasekgoka@gmail.com
ORCID ID: http://orcid.org/0000-0002-0549-667X
Venkata Seshachala Sarma Yadavalli[1]
ORCID ID: http://orcid.org/0000-0002-3035-8906
Olufemi Adetunji[1]
ORCID ID: http://orcid.org/0000-0002-3305-5310
[1] Department of Industrial and Systems Engineering, University of Pretoria, South Africa.

Disclosure statement
No potential conflict of interest was reported by the author(s).

Funding
This research was supported by The Banking Sector Education and Training Authority Bankseta [475.4710.640000] and the University of Pretoria Postgraduate Bursary for Doctoral Students.

Correction
This article has been republished with minor changes. These changes do not impact the academic content of the article.

Citation information
Cite this article as: Privacy-preserving data mining of cross-border financial flows, Chaka Patrick Sekgoka, Venkata Seshachala Sarma Yadavalli & Olufemi Adetunji, *Cogent Engineering* (2022), 9: 2046680.

References

Adam, N. R., & Wortmann, J. C. (1989, December). Security-Control methods for statistical databases: A comparative study. *ACM Computing Surveys*, *21*(4), 515–556. https://doi.org/10.1145/76894.76895

Agrawal, R., & Srikant, R. (2000). Privacy-preserving data mining. *ACM SIGMOD Record* 29 2 , 439–450. doi: https://doi.org/10.1145/335191.335438

Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description. *Data Mining and Knowledge Discovery*, *29*(3), 626–688. https://doi.org/10.1007/s10618-014-0365-y

Babus, A. (2016). The formation of financial networks. *The RAND Journal of Economics*, *47*(2), 239–272. https://doi.org/10.1111/1756-2171.12126

Cios, K. J., & Moore, W. G. (2002, September). Uniqueness of medical data mining. *Artificial Intelligence in Medicine*, *26*(1–2 1–24 doi:10.1016/S0933-3657(02)00049-0).

Clifton, C., & Vaidya, J. (2004). Privacy-Preserving data mining: why, how, and when. *IEEE Security & Privacy Magazine* 2 (IEEE) , (pp. 19–27).

Cox, D. (2014). *Handbook of Anti-Money laundering*. John Wiley & Sons Ltd.

Deshkar, S, Thanseeh, R. A, Menon, V. G. (2017). A review on IoT based m-Health systems for diabetes. *International Journal of Computer Science and Telecommunications*, *8* 1 , 13–18.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. *Theory of cryptography conference*, (pp. 265–284). Springer.

FATF. (2014). *Guidance for a risk-based approach: The Banking Sector*. https://protect-za.mimecast.com/s/crgICElv6pC5GRwiNofra?domain=fatf-gafi.org

Freund, C. L., & Weinhold, D. (2004). The effect of the Internet on international trade. *Journal of International Economics*, *62*(1), 171–189. doi:https://doi.org/10.1016/S0022-1996(03)00059-X

Harvey, J. (2004). Compliance and reporting issues arising for financial institutions from money laundering regulation: A preliminary cost benefit study. *Journal of Money Laundering Control*, *7*(4), 333–346. https://doi.org/10.1108/13685200410810047

Kang, S. (2018). Rethinking the global anti-money laundering regulations to deter corruption. *INTERNATIONAL*

AND COMPARATIVE LAW QUARTERLY, *67*(3), 695–720. https://doi.org/10.1017/S0020589318000106

Kou, G., Peng, Y., Shi, Y., & Chen, Z. (2007, July). Privacy-Preserving Data Mining of Medical Data Using Data Separation-Based Techniques. *Data Science Journal*, *6*, S429–S434. https://doi.org/10.2481/dsj.6.S429

Li, X., Liu, S., Li, Z., Han, X., Shi, C., Hooi, B., … Cheng, X. (2020). Flowscope: Spotting money laundering based on graphs. *Proceedings of the AAAI conference on artificial intelligence*, (pp. 4731–4738). AAAI.

Liu, S., Hooi, B., & Faloutsos, C. (2017). HoloScope: Topology-and-Spike aware Fraud Detection. *Proceedings of the 2017 ACM on conference on information and knowledge management* 6-10 Nov (ACM) Singapore, (pp. 1539–1548 https://doi.org/10.1145/3132847.3133018 doi:).

Lv, L.-T., Ji, N., & Zhang, J.-L. (2008). A RBF neural network model for anti-money laundering. *2008 International Conference on Wavelet Analysis and Pattern Recognition*, *1* (IEEE), (pp. 209–215). https://doi.org/10.1109/ICWAPR.2008.4635778

Magnusson, D., & Harvey, J. (2009). The costs of implementing the anti-money laundering regulations in Sweden. *Journal of Money Laundering Control*, *12*(2), 101–112. https://doi.org/10.1108/13685200910951884

Meltzer, J. P. (2015). The internet, cross-border data flows and international trade. *Asia & the Pacific Policy Studies*, *2*(1), 90–102. https://doi.org/10.1002/app5.60

Neanidis, K. C. (2019). Volatile capital flows and economic growth: The role of banking supervision. *Journal of Financial Stability*, *40* February 2019 , 77–93. doi: https://doi.org/10.1016/j.jfs.2018.05.002

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, *32*(3), 245–251. doi:https://doi.org/10.1016/j.socnet.2010.03.006

Ostry, J. D., Gosh, A. R., Chamon, M., & Qureshi, M. S. (2012). "Tools for managing financial stability risks from capital inflows. *Journal of International Economics*, *88*(2), 407–421. https://doi.org/10.1016/j.jinteco.2012.02.002

Pathak, M., Rane, S., & Raj, B. (2010). Multiparty differential privacy via aggregation of locally trained classifiers. *International conference on Neural Information Processing Systems (NIPS)*, *2*, (pp. 1876–1884). Association for Computing Machinery (ACM).

Pathak, M., Rane, S., Sun, W., & Raj, B. (2011). Privacy preserving probabilistic inference with hidden Markov models. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 5868–5871). IEEE.

Paula, E. L., Ladeira, M., Carvalho, R. N., & Marzagao, T. (2016). Deep learning anomaly detection as support fraud investigation in Brazilian exports and anti-money laundering. *15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, (pp. 954–960). IEEE. https://doi.org/10.1109/ICMLA.2016.0172

Pourhabibi, T., Ong, K.-L., Kam, B., & Boo, Y. (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, *133* June 2020 , 113303. doi:https://doi.org/10.1016/j.dss.2020.113303

Prakash, B. A., Sridharan, A., Seshadri, M., Machiraju, S., & Faloutsos, C. (2010). EigenSpokes: Surprising patterns and scalable community chipping in large graphs. *Pacific-Asia conference on knowledge discovery and data mining*, (pp. 435–448). Springer. https://doi.org/10.1007/978-3-642-13672-6_42

Qi, X., & Zong, M. (2012). An overview of privacy preserving data mining. *2011 International Conference on Environmental Science and Engineering*, *12*, (pp. 1341–1347). Elsevier B.V. https://doi.org/10.1016/j.proenv.2012.01.432

Rajput, Q., Khan, N. S., Larik, A., & Haider, S. (2014). Ontology-based expert-system for suspicious transactions detection. *Computer and Information Science*, *7*(1 103–114). https://doi.org/10.5539/cis.v7n1p103

Silva, T. C., de Souza, S. R., & Tabak, B. M. (2016). Structure and dynamics of the global financial network. *Chaos, Solitons and Fractals*, *88* C , 218–234. https://doi.org/10.1016/j.chaos.2016.01.023

Sreeja, R., Varghese, P., Menon, V. G., & Khosravi, M. R. (2019). A secure and efficient lightweight symmetric encryption scheme for transfer of text files between embedded IoT devices. *Symmetry*, *11*(2), 293. https://doi.org/10.3390/sym11020293

Stinson, D. R 2006 Cryptography theory and practice Discrete mathematics and its applications (Chapman & Hall/CRC)1-58488-508-4

Sun, Q., Qu, Chakrabarti, Chakrabarti, H., Faloutsos, D., & Faloutsos, C. (2005, December). Relevance search and anomaly detection in bipartite graphs. *ACM SIGKDD Explorations Newsletter*, *7*(2), 48–55. doi: https://doi.org/10.1145/1117454.1117461

Sun, J., Qu, H., Chakrabarti, D., & Faloutsos, C. (2005). Neighborhood formation and anomaly detection in bipartite graphs. *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM)*, (pp. 418–425). IEEE Computer Society.

Sun, X, Zhang, J, Liu, S, Chen, J, Zhuang, R, Shen, H, Cheng, X 2021 CubeFlow: Money Laundering Detection with Coupled Tensors. PAKDD 2021: Advances in Knowledge Discovery and Data Mining. 12712 (Springer)78–90 May 2021 https://doi.org/10.1007/978-3-030-75762-5_7 Part of the Lecture Notes in Computer Science book series (LNCS, volume 12712)

Tang, J, Yin, J 2005 Developing an intelligent data discriminating system of anti-money laundering based on SVM. 2005 International Conference on Machine Learning and Cybernetics. Guangzhou 18-21 Aug (IEEE)doi:10.1109/ICMLC.2005.1527539

Teichmann, F. M. (2019). Recent trends in money laundering and terrorism financing. *Journal of Financial Regulation and Compliance*, *27*(1), 2–12. https://doi.org/10.1108/JFRC-03-2018-0042

van Duyne, P. (1994). Money-Laundering: Estimates in fog. *Journal of Financial Crime*, *2*(1), 58–74. https://doi.org/10.1108/eb025638

Walker, J. (1999). How big is global money laundering? *Journal of Money Laundering Control*, *3*(1), 25–37. https://doi.org/10.1108/eb027208

Wang, A., Wang, C., Bi, M., & Xu, J. (2018). A review of privacy-preserving machine learning classification. *International Conference on Cloud Computing and Security (ICCCS)*, (pp. 671–682). Springer. Cham.

Xiong, H., Z, L., & Zhou, L. Y. (2010). Detecting blackhole and volcano patterns in directed networks. *Proceedings of the 10th IEEE International Conference On Data Mining (ICDM)*, (pp. 294–303). IEEE.

Xu, L., Jiang, C., Wang, J., Yuan, J., & Ren, Y. (2014). Information security in big data: Privacy and data mining. *IEEE Access*, *2*, 1149–1176. https://doi.org/10.1109/ACCESS.2014.2362522

Zhiguo, Z., Jingqin, S., & Liping, K. (2015). Measuring influence in online social network based on the user-content bipartite graph. *Computers in Human Behavior*, *52*, 184–189. https://doi.org/10.1016/j.chb.2015.04.072

Cogent Engineering **(ISSN: 2331-1916) is published by Cogent OA, part of Taylor & Francis Group.**

Publishing with Cogent OA ensures:

- Immediate, universal access to your article on publication
- High visibility and discoverability via the Cogent OA website as well as Taylor & Francis Online
- Download and citation statistics for your article
- Rapid online publication
- Input from, and dialog with, expert editors and editorial boards
- Retention of full copyright of your article
- Guaranteed legacy preservation of your article
- Discounts and waivers for authors in developing regions

Submit your manuscript to a Cogent OA journal at www.CogentOA.com