UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

MODELING CROSS-BORDER FINANCIAL FLOWS USING A NETWORK THEORETIC APPROACH

by

Chaka Patrick Sekgoka

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy (Industrial Systems)

in the

Department of Industrial and Systems Engineering
Faculty of Engineering, Built Environment and Information Technology

UNIVERSITY OF PRETORIA

February 2021

Declaration

I declare that except where due acknowledgement has been made, this thesis is my own work, which I hereby submit for the degree of Doctor of Philosophy at the University of Pretoria. It has not been submitted previously, in whole or in part, to qualify for any other academic award.

Sekgoka

Chaka Patrick Sekgoka

February 2021

Abstract

Title:	Modeling cross-border border financial flows using a network theoretic approach
Author:	Chaka Patrick Sekgoka
Supervisor:	Professor Olufemi Adetunji
Co-supervisor:	Professor V.S. Sarma Yadavalli
Department:	Department of Industrial and Systems Engineering
University:	University of Pretoria
Degree:	Doctor of Philosophy (Industrial Systems)

Criminal networks exploit vulnerabilities in the global financial system, using it as a conduit to launder criminal proceeds. Law enforcement agencies, financial institutions, and regulatory organizations often scrutinize voluminous financial records for suspicious activities and criminal conduct as part of anti-money laundering investigations. However, such studies are narrowly focused on incidents and triggered by tip-offs rather than data mining insights.

This research models cross-border financial flows using a network theoretic approach and proposes a symmetric-key encryption algorithm to preserve information privacy in multi-dimensional data sets. The newly developed tools will enable regulatory organizations, financial institutions, and law enforcement agencies to identify suspicious activity and criminal conduct in cross-border financial transactions.

Anti-money laundering, which comprises laws, regulations, and procedures to combat money laundering, requires financial institutions to verify and identify their customers in various circumstances and monitor suspicious activity transactions. Instituting anti-money laundering laws and regulations in a country carries the benefit of creating a data-rich environment, thereby facilitating non-classical analytical strategies and tools.

Graph theory offers an elegant way of representing cross-border payments/receipts between resident and non-resident parties (nodes), with links representing the parties' transactions. The network representations provide potent data mining tools, facilitating a better understanding of transactional patterns that may constitute suspicious transactions and criminal conduct.

Using network science to analyze large and complex data sets to detect anomalies in the data set is fast becoming more important and exciting than merely learning about its structure. This research leverages advanced technology to construct and visualize the cross-border financial flows' network structure, using a directed and dual-weighted bipartite graph.

Furthermore, the develops a centrality measure for the proposed cross-border financial flows network using a method based on matrix multiplication to answer the question, "Which resident/non-resident nodes are the most important in the cross-border financial flows network?" The answer to this question provides data mining insights about the network structure.

The proposed network structure, centrality measure, and characterization using degree distributions can enable financial institutions and regulatory organizations to identify dominant nodes in complex multi-dimensional data sets. Most importantly, the results showed that the research provides transaction monitoring capabilities that allow the setting of customer segmentation criteria, complementing the built-in transaction-specific triggers methods for detecting suspicious activity transactions.

In loving memory of my mother and my son

Maria Mokgadi Sekgoka

and

Kgolofelo Sekgoka

"The purpose of (scientific) computing is insight, not numbers."

Richard Wesley Hamming

Acknowledgements

I express my sincere gratitude to my mentor, Professor Sarma Yadavalli, for his continuous support, motivation, and immense knowledge. You have put an incredible effort into my personal and career development over the 25 years that I have known you. I am very grateful for the undergraduate and postgraduate courses that you have taught me at the University of the North, Polokwane. Your encouragement and belief carved a successful career path for me.

I thank my supervisor Professor Olufemi Adetunji who dedicated his time and effort to guide me throughout my research journey. I appreciate that you took the time to read my work and supported every endeavor to make matters work.

I am very grateful for the support that I received from MaryAnne DePesquo. MaryAnne, you dedicated your time to read my research and position it in a meaningful way. I have learned so much from your guidance and our interactions. You made the SAS Global Forum 2019 an experience of note for International Professional Award winners in Dallas, Texas-USA. I thoroughly enjoyed the conference proceedings and the training interventions.

Thank you to all the staff members in the Department of Industrial and Systems Engineering for the insightful discussions we had during my visits to the University of Pretoria. You have made my working environment a delightful one.

I gratefully acknowledge the support that I received from my colleagues at the South African Reserve Bank and the valuable discussions I had with them: Professor Pumela Msweli, Thoraya Pandy, Moffat Mangole, Eleanor Mashiane, Marius Stander, Herminah Chechane, Sharlay Madalane, Pinky Mokgoko, and Gerhard van der Venter. Thank you for taking the time to listen and share your thoughts with me.

I thank the Banking Sector Education and Training Authority and the University of Pretoria for funding this research, the South African Reserve Bank for providing the cross-border financial flows data set, and the SAS Institute of South Africa Incorporated for availing the SAS system. I am grateful to Ronel Dijksman, Ayanda Simelane, Precious Tshivhase, Chere Monaisa, and Sifiso Mnguni for coordinating and administering the research funding.

I thank my life coach Bronson Hlatshwayo for the fruitful coaching sessions and moral support over the years. Thank you to Andre Zitzke for all the insightful discussions and the international relationships you fostered during my research journey. I also extend my gratitude to my fellow Ph.D. students Catherine Maware and John Gbeminiyi Oyewole. I enjoyed our talks, encouragements, and laughs during our informal meetings on Campus.

Many excellent researchers at the University of Twente (The Netherlands) influenced my appreciation for mathematical sciences, namely, Professor Wim Albers, Professor Erik van Doorn, Professor Stephan van Gils, and Professor Arunabha Bagchi. I acknowledge Professor S.P Mashike (University of Limpopo), who arranged a scholarship to study Engineering Mathematics at the University of Twente.

I am much indebted to my uncle Annies Magomo Mametja for his continuous guidance and guardianship. To my Mathematics and Physical Science teachers at Molabosane High School, Laurence Dumisani Nxumalo and Mmapula Lillian Mashele, I want to thank you for stimulating and inspiring my mathematical sciences career.

A very special thank you to my wonderful wife, Irene Sekgoka, for her unconditional love and unwavering support. Irene, you have been encouraging and patient. Words cannot describe how truly grateful I am. To my beloved sons, Lehlogonolo, Tumisho, and Mahlasedi, I would like to say thank you for your understanding, playtime, and family time that you have sacrificed.

Table of Contents

List of Figures

List of Tables

Abbreviations

AML: Anti-money Laundering

API: Application Programming Interface

AUC: African Union Commission

BoP: Balance of Payments

CDD: Customer care Due Diligence

CFT: Combating of the Financing of Terrorism

DES: Data Encryption Standard

ECA: Economic Commission for Africa

FIC: Financial Intelligence Centre

FICA: Financial Intelligence Centre Act 38 of 2001

GDP: Gross Domestic Product

GFI: Global Financial Integrity

IIP: International Investment Position

IMF: International Monetary Fund

IoT: Internet of Things

ITRS: International Transactions Reporting System

NTSA: Tiny Symmetric encryption algorithm

OECD: Organisation for Economic Co-operation and Development

POCA: Prevention of Organized Crime Act, 121 of 1998

POCDATARA: Protection of Constitutional Democracy Against Terrorist and Related Activities Act number 33 of 2004

RegTech: Regulatory Technology

SARB: The South African Reserve Bank

SARS: The South African Revenue Services

SupTech: Supervisory Technology

TEA: Tiny Encryption Algorithm

TIC: Treasury International Capital

TLS: Transport Layer Security

XML: eXtensible Markup Language

Chapter 1: Introduction

1.1 Introduction

This chapter provides the research's background and scope, problem statement, and research objectives, including definitions of terminologies underpinning the problem statement. The chapter also outlines the expected contributions to knowledge and concludes with an overview of the thesis's structure.

1.2 Background and scope

1.2.1 Cross-border financial flows

The global financial system is subject to a wide range of risks and vulnerabilities exploited by criminal networks to launder the proceeds of crime and finance terrorist activities with a relatively low risk of detection. These risks and vulnerabilities include the following:

- Voluminous and volatile cross-border financial flows that obscure individual transactions and provide opportunities for criminal organizations to transfer value across country borders.
- There are limited resources in most customs agencies for detecting illegal trade transactions and limited recourse to verification procedures or programs for exchanging customs data between countries.

Cross-border flows are money transferred by a resident to a non-resident and vice versa because of financial transactions involving individuals, private and public firms, central banks, financial institutions, and legal entities such as trusts and non-profit organizations or a combination thereof, in at least two different countries. International trade transactions (imports/exports) in goods and services and remittances are examples of cross-border flows.

Cross-border financial flows are the backbone of a country's Balance of Payments (BoP) accounts. The flows comprise international business transactions between residents and non-residents involving imports and exports of goods and services, purchase and sale of financial assets (bonds and shares), real assets (factories, land, and buildings), income receipts and payments (dividends, interest), current transfers (remittances, gifts, charitable donations), as well as borrowing from and lending to the rest of the world (foreign loans/bonds). Capital flows are an essential aspect of the global monetary system and offer potential economic benefits to countries.

1.2.2 Mechanisms for recording international financial transaction

Countries use different accounts to record BoP transactions, distinguished according to the nature of the economic resources provided and received. The volatility and size of the flows also pose policy challenges for many countries.

Commercial banks and other financial intermediaries use several sources, such as electronic messaging systems, survey questionnaires and forms to report cross-border financial flows transactions to regulatory authorities for BoP reporting and regulatory purposes. For example, the U.S. Treasury International Capital (TIC) reporting system and the SARB's International Transactions Reporting System (ITRS) are examples of computerized data collection systems.

The TIC reporting system collects data on cross-border portfolio investment flows and positions between U.S. residents and foreign residents. SARB's ITRS is an electronic messaging system for collecting cross-border financial flows data between South African residents and non-residents from licensed financial institutions that trade foreign currencies.

It is a regulatory obligation for authorized dealers to report the cross-border financial transactions to the SARB, Financial Intelligence Centre (FIC), and other regulatory institutions such as the South African Revenue Services (SARS). The SARB uses the international financial transaction database along with trade statistics from SARS to compile BoP accounts.

Figure 1.1 depicts the flow of cross-border transactional data between South African residents and the rest of the world. Residents make payments to non-residents and receive payments from non-residents and vice versa through the authorized dealer network comprising commercial banks and other licensed financial institutions such as Bureau de changes (currency exchanges). The authorized dealers facilitate the payments through the corresponding bank relationships in the country of non-residents to finalize the fund's transfers.

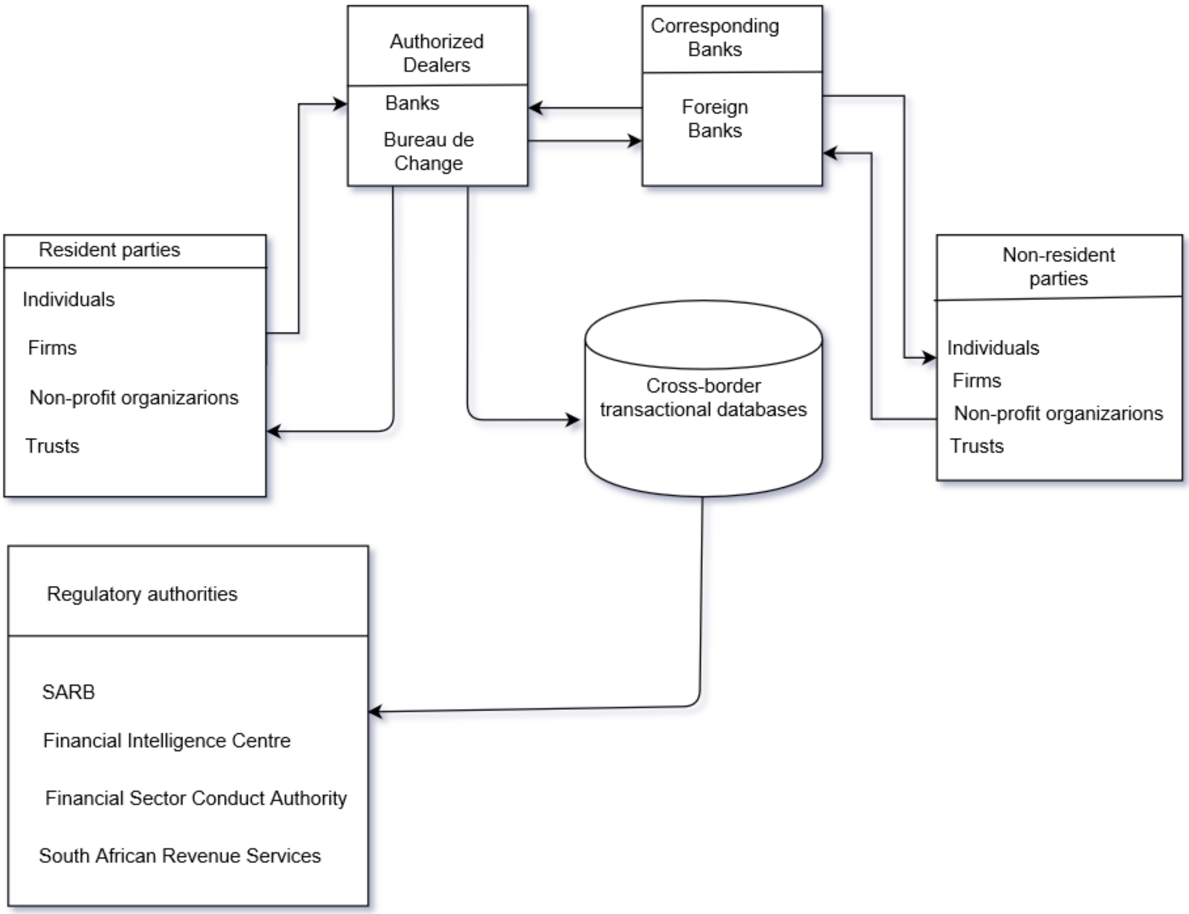


Figure 1.1: Depiction of cross-border financial transactions data flow in South Africa

The authorized dealers mostly use the eXtensible Markup Language (XML) to manage and share structured data in a human-readable text file to send cross-border financial flows data to the SARB. The design goals of XML emphasize simplicity, generality, and usability across the Internet.

Several schema systems exist to support the definition of XML-based languages, while programmers have developed many application programming interfaces (APIs) to support XML data processing. In service-oriented architectures, disparate systems communicate with each other by exchanging XML messages.

The computer code in Appendix B.1 shows an XML file example for reporting cross-border financial transactions. Each transaction comprises several components, i.e., transaction date, personal details of the resident party and the non-resident party, the authorized bank and the corresponding bank, and payment details.

1.2.3 Illegal cross-border transfer of value

Researchers have documented significant financial losses in many countries due to illegal cross-border movement of money in recent years, which is part of money laundering (GFI, 2015; United Nations, 2015). Money laundering is the process of obscuring the origins of illegally obtained money, typically by passing it through the banking system or business transactions.

The process starts with the criminal activity that gives rise to the illegal proceeds, such as bribery, drug trafficking, tax evasion, and corrupt business practices. The money launderer seeks to disguise funds earned from such unlawful activities.

In response, many countries enacted Anti-Money Laundering (AML) laws and policies, comprising a comprehensive plan of action to fight money laundering and terrorist finance. AML requires financial institutions to monitor transactions for suspicious activities and criminal conduct.

The adoption of AML software systems with built-in transaction-specific triggers allows active tracking of transactions under the risk-based approach guidelines and recommendations provided by the Financial Action Task Force (FATF) for regulatory compliance purposes.

Despite the availability of such tools coupled with the regulatory guidelines, most of the AML related investigations remain narrowly focused on incidents and triggered by tip-offs rather than data mining insights. The built-in transaction-specific triggers often produce many false positives for financial fraud and money laundering. Criminal networks often know the events that trigger suspicious transactions and circumvent them using advanced transaction layering techniques.

This research hypothesizes that the illegal transfer of funds across country borders by criminal networks exhibits relationship structures inherent in complex systems; therefore, using structural and statistical properties of network science can enhance our understanding of cross-border flows. Using the proposed network tools for analyzing cross-border transactional data can improve surveillance capabilities in regulatory authorities and regulated entities, enabling them to combat money laundering.

In recent years, researchers have obtained useful tools for describing relationships inherent in complex systems through the study of network science, which is a part of mathematical graph theory. Network science employs various techniques and ideas from many fields, including data mining and visualization, statistical inference, electrical engineering, molecular biology, statistical mechanics, and social sciences.

This research leverages advances in technology and data mining methods to develop the network structure of cross-border financial flows using a directed and weighted bipartite graph. The proposed graph-theoretic approach models transactions between residents and non-residents to identify potentially suspicious activity and criminal conduct.

Firstly, the research develops the symmetric-key encryption algorithm to preserve information privacy in multi-dimensional data sets, thereby addressing privacy concerns for governments, firms, and private individuals.

The proposed algorithm utilizes the group structure of the multi-dimensional data sets for efficient data processing using a computer program. The algorithm provides an alternative encryption method to the sophisticated non-cryptographic techniques used by researchers to query statistical databases involving multi-dimensional data sets in various fields.

The lack of research in the analysis of cross-border flows is mainly attributable to the difficulties in accessing and sharing financial transactions data, which comprise private and confidential information protected by multitudes of regulations, laws, and best practices. Financial transactions that do not involve cross-border payments/receipts are not part of this study.

Information privacy laws and policies protect the confidentiality of information in BoP data collected at the transactional level. The research's use of confidential data adheres to the requirements of section 33 of the SARB Act, No. 90 of 1989, as amended, which entails the preservation of the secrecy of the information of the SARB. As a result, this research's data set cannot be made available to the public.

The legal authority for collecting BoP data derives from a section of the law (act) applicable in a country. For example, the Breton Wood Agreements Act of 1945 is the applicable law in the United States of America.

The Exchange Control Act enforces regulatory reporting of cross-border financial transactions in countries maintaining Exchange Controls. For example, Section 9 of the Currency and Exchanges Act of 1933 is the applicable law in South Africa. Central banks and government agencies such as the country's official statistical agency are the custodians of statistical data on cross-border transactions as delegated by the Treasury Departments in many countries.

The IMF's BoP manual serves as the international standard for the conceptual framework underlying BoP statistics by providing guidelines for classifying cross-border flows data to member countries (IMF, 2013). The classification system groups international transactions that show similar behavioral patterns to facilitate their utilization and adaption for multiple purposes such as regional and global aggregations, bilateral comparisons of components of BoP statistics, policy formulations, and analytical studies.

Secondly, the research uses SAS® software to implement the encryption algorithm, visualize the directed and weighted bipartite network structure, and characterize the cross-network. Lastly, the study develops a network centrality measure to identify the highly connected nodes responsible for the most transactions in the network.

The research uses real data set comprising remittance transactions extracted from the International Transactions Reporting System (ITRS) of the South African Reserve Bank (SARB) and a hypothetical data set for network construction, visualization, and computing the proposed centrality measure.

The network theoretic approach will potentially benefit financial institutions, regulatory organizations and law enforcement agencies to combat money laundering, illicit flow flows, drug trafficking, terrorist financing and other organized crimes. The benefits will be more pronounced in mineral-rich developing countries, which researchers found to suffer large capital outflows due to illicit financial flows.

The research does not provide policy recommendations for combating money laundering using the proposed network-theoretic approach. Also, due to this research's nature, participants of this study did not agree for their data to be shared publicly, so supporting data is not available.

1.3 Problem statement and research objectives

Given the background provided thus far, the problem statement for this research is stated as follows:

"*Illegal transfer of funds across country borders reduce domestic capital resources and the tax revenue needed to fund infrastructure development and poverty alleviation programs in emerging market economies* (World Bank, 1985; Patnaik, Gupta, & Shah, 2012; GFI, 2015)."

Furthermore, illegal activities such as exploitation of mineral resources, organized crime, drugs counterfeiting, corruption, and fraud in international trade have a devastating impact on the affected communities.

Advanced computing and communication technology enables financial institutions and supervisory authorities to handle complex tasks relating to data collection, data management, and data mining. Also, the much-improved capacity and speed of data storage media have led to the heavy use of third-party services and infrastructures to host data sets. Hence, the recent adoption of cloud computing, business intelligence, and advanced analytical systems improves risk management practices within the financial services industry.

The new technologies also present several risks and challenges including user privacy, data security, data lock-in, service availability, performance and disaster recovery (Buyya, Broberg, & Goscinski, 2010; Takabi, Joshi, & Ahn, 2010). Hence, the research's consideration of protecting information privacy in the multi-dimensional data comprising cross-border financial flows.

Analysis of data sets from a wide range of sources such as financial transactions held by financial institutions, patient records maintained by healthcare systems, salary records held by employers, investigation records held by the criminal justice system, motor vehicle registration information contained by public institutions, etc., often triggers information privacy concerns. Compliance with data protection regulations by private and public firms is required whenever one processes data comprising personally identifiable information. Large data sets, such as the cross-border financial transactions data set, contain personally identifiable data protected by information privacy laws and policies.

Despite the availability of the cross-border flow data sets and advanced computing technology, illicit financial flows in developing countries keep rising. Reliance on tip-offs to support anti-money laundering fails to yield a significant impact in reducing illicit financial flows due to their hidden nature. *Based on the (GFI, 2015) report, illegal outflows of capital from the developing countries cracked the $1 trillion plateau*.

The research question is as follows:

Can the analysis of the cross-border financial flows' network structure and its statistical properties enable identifying suspicious activity transactions and criminal conduct?

The research question above leads the formulation of the following research objectives. The first objective is to develop a symmetric-key encryption algorithm to preserve information privacy in data sets comprising cross-border financial flows. The algorithm's purpose is to enable the development and implementation of a cryptosystem to circumvent information privacy concerns when analyzing cross-border financial flows data sets.

The proposed symmetric-key encryption algorithm uses temporary variables created during the computer program's compilation, making efficient use of computer memory. In addition to enabling data encryption, the temporary variables provide a mechanism to aid the computation of the underlying statistical measures of networks, such as edge weight and node degree, for network construction.

The second objective is to a develop network model for studying the transaction patterns between residents and non-residents in cross-border financial flows. The study uses a directed and weighted bipartite graph with dual weights, representing the monetary value and volume of international financial transactions, and SAS® Visual Analytics software to create the visualization of the network structure

The third objective is to develop a method of measuring node centrality based on the projection of vectors to illuminate the cross-border financial flows network structure. The centrality measure and visualization will enable financial institutions, supervisory authorities, and law enforcement agencies to identify the highly connected nodes in the network, thereby empowering them to plan their inspection programs using transaction samples from such nodes.

The fourth and last research objective is to characterize the cross-border financial flows network using the distribution of node degrees. The cross-border financial flows network is a weighted and directed network with two distinct node sets. Therefore, it comprises two degree distributions for each set of nodes, resulting in four degree distributions.

The hierarchical clustering procedure characterizes the network, reducing the number of degree distributions from four to two, one for each set of nodes. The research focusses on the degree distribution of the residents since they act as the gateway for both inward and outward financial flows of a country.

This research provides solutions to the problem statement by demonstrating the effectiveness of using network science tools to model cross-border financial flows while preserving information privacy relating to personally identifiable data.

The formulation of the research objectives is as follows:

- To leverage technology in developing a simple and efficient algorithm for preserving the privacy of personally identifiable information in multi-dimensional data sets, such as the cross-border financial flows data set.
- Use network science tools to analyze transaction patterns between residents and non-residents in cross-border transactions, thereby providing useful analytical tools for detecting and impeding the illegal transfer of funds across country borders.
- Develop and use statistical properties of networks to complement the existing classical statistical measures for cross-border financial flows to enhance our understanding of the workings of cross-border financial flows.

The research seeks to achieve the objectives stated above subject to the following limitations.

- The research focuses on analytical methods for addressing the money laundering problem. Hence, it does not attempt to provide a quantitative measure of illicit financial flows.
- The research uses a sample data set extracted from the South African database of international financial transactions to construct and visualize the cross-border financial flows network, which is not representative of the global regulatory environment.
- The study uses a hypothetical data set of cross-border financial flows to illustrate the proposed centrality measure. Hence, the test data set is not very large in contrast to the real-world data sets.
- The research does not make any policy recommendations to regulate or reduce the illegal transfer of funds across country borders.

1.4 Knowledge contributions

The output of this research will contribute to knowledge in the following ways:

- It develops a symmetric-key encryption algorithm for preserving information privacy in multi-dimensional data sets, using temporary automatic variables. Various fields, such as medical studies, official statistics and financial transactions, with multi-dimensional data sets, can use the proposed symmetric-key encryption algorithm to encrypt sensitive information. The algorithm is fast in execution and very simple to implement compared to other sophisticated and poorly performing methods.
- It develops a node centrality measure based on the matrix multiplication method to identify the highly connected nodes in the cross-border financial flows network. Network structures utilizing the directed and weighted networks in other fields can use this measure.
- It uses a hierarchical clustering procedure to reduce the number of degree distributions in a directed and weighted bipartite graph to identify important groups of nodes in the network. The resulting clusters enable customer segmentation in the cross-border data set to facilitate using the traditional analytical methods in detecting suspicious activity transactions.

The research findings show that network science can strengthen financial regulation and improve risk management practices by supporting the risk-based approach. Supervisory authorities can use the derived methodology to define and plan their inspection of regulated entities for criminal conduct and money laundering in a cost-effective manner.

1.5 Outline of the thesis

The outline of the remaining chapters is as follows. Chapter two outlines the main theories and reviews the relevant literature. The chapter discusses the money laundering process and uses trade-based money laundering to illustrate the problem. The chapter reviews the literature relating to the global regulatory response to the money laundering problem, comprising the regulatory standards, risk-based approach, and the foreign exchange controls.

Chapter two also reviews the literature relating to information privacy preservation in multi-dimensional data sets such as cross-border financial flows, financial networks, bipartite network structures and centrality measures.

Chapter 3 provides a background of mathematical graph theory, starting with a brief history of the subject and its network science applications. This chapter's primary focus is on graph theory's main results, which are useful when developing the network structure of cross-border financial flows. The results are mostly attributable to the work done by (König, 1936).

Chapter 4 discusses cryptographic and non- cryptographic methods for protecting information privacy in statistical databases, including encryption, differential privacy and data separation techniques. Most importantly, the chapter proposes the symmetric-key encryption algorithm to preserve the confidentiality of personally identifiable information in cross-border financial transactions and implements it using a SAS® computer program. The chapter concludes with an overview of the remittances data set.

Chapter 5 discusses network science as the chosen tool for answering research questions. This chapter develops the network structure of cross-border financial flows using the directed and weighted bipartite graph with dual weights and discusses the significance of the proposed network structure. Also, the chapter presents network visualization using SAS® Visual Analytics software.

Chapter 6 provides an overview of the measures for networks, including density and centrality. The chapter's primary objective is to develop the node centrality measure for the cross-border financial flows network using matrices. The proposed methodology accumulates centrality weights to resident nodes that transferred funds to the same non-resident nodes and vice versa, resulting in larger weight allocations to nodes that share connections.

The centrality measure, along with network visualization, enables the identification of the highly connected nodes. The chapter further uses a hierarchical clustering procedure to characterize the cross-border financial flows network. The clustering procedure provides a visual analysis of the cross-border financial flows network structure using the degree distributions. The clustering procedure's output enriches the cross-border financial flows data set with additional variables to enable classical analytical methods and support the risk-based supervisory framework.

Chapter 7 discusses the research findings, scientific contributions and the study's significance. The chapter summarizes the research objectives and identifies areas for further research. Also, the chapter discusses the limitations of the methods proposed in this research.

Chapter 8 provides the concluding remarks, thereby summarizing the main conclusions and areas requiring further research.

Chapter 2: Literature review

2.1 Introduction

Chapter two outlines the main theories and reviews the applicable literature. The chapter begins with an outline of the money laundering problem and discusses trade-based money laundering. Next, the chapter outlines the regulatory response to the problem, discussing the global regulatory standards, the risk-based approach, and the foreign exchange controls.

This research's primary focus is the risk-based approach, which leverages technology for the extraction, processing and analysis of risk factors in large data sets using methods that include machine learning algorithms. The study adopts a network theoretic approach to enhance surveillance of cross-border financial flows in regulated entities. Hence, this chapter reviews literature on structures and statistical properties used for identifying the important nodes in networks.

The chapter also discusses and reviews the literature regarding the preservation of information in data sets comprising private and confidential information such as cross-border financial flows. The chapter concludes with an overview of the data set used in the research.

2.2 Money laundering

Researchers and practitioners acknowledge that criminal networks continue utilizing the global financial system to launder their crime proceeds and finance terrorist activities (van Duyne, 1994; Walker, 1999; Harvey, 2004; Teichmann, 2019).

Money laundering is a mechanism that ensures the functioning of crime as an economic system, hence the global efforts to combat this illicit economy. Its economic feasibility enables criminal networks to finance terrorist activities using money deemed clean, hindering terrorism prevention policies' effectiveness. Money laundering poses significant economic challenges in emerging markets by reducing domestic resources and tax revenue needed for funding infrastructure and poverty alleviation programs (Ba & Huynh, 2018).

Criminal networks commonly use three methods to obscure the origins of illegally obtained money and integrate it back into the formal economy. The main challenge is to achieve the objective without raising the suspicion of law enforcement agencies.

- The first method involves transferring value through the global financial system using wire transfers and commercial transactions.
- The second method involves the physical movement of banknotes using cash couriers and bulk cash smuggling.
- The third method uses illegal means of trading, such as smuggling and false declarations of trade documents to transfer value.

This research's primary focus is the first method, involving the illegal transfer of funds across country borders through the banking system. Criminals often engage in unlawful cross-border fund transfers to conceal the proceeds of crimes and finance terrorist activities, both domestically and internationally.

Three stages comprise the money laundering process, i.e., placement, layering, and integration. The placement phase is the initial stage of the money laundering process, which involves introducing the illegitimately obtained funds from criminal activities into the financial system. The launderer moves funds from their original cash source into some other forms, such as purchasing a physical asset, say property, or any physical item of value. Subsequently, the money launderer can sell the asset to disguise the funds as legitimately obtained funds. Hence, the placement phase is not just the cash flow into the banking system since it also involves the initial transfer of assets into other forms, enabling the money launderer to undertake further layering and obscure the trail of the funds (Cox, 2014).

In most cases, money launderers will target areas with weak controls. For example, suppose the money launderer knows that a firm needs cash to avoid bankruptcy. In that case, the money launderer will seize the opportunity to place the illegally obtained funds into the business.

The second step in the money laundering process is the layering process, which involves stratifying the financial transaction, making unclear the source of the funds and the funds' current position. Adding more layers to the financial transaction will only complicate the scheme, thereby making it harder to prove the funds' illicit basis. In very complex scenarios, the money launderer will shift the funds between several accounts in many different jurisdictions to obscure the audit trail. Layers often include various financial transactions, high-value items, currency and equipment sales, real estate, and legitimate businesses' purchase.

The layering process involving the purchase or sale of a property and other physical assets pose significant risks for the money launderer since the authorities can follow the trail and discover the links with the original crime proceeds. Depending on the number of layers and the complexity of the money launderer's schemes, the layering process can become costly to the money launderer.

The final step integrates illegally obtained funds into the mainstream banking system and returns the launderer's funds as clean money. It is called the integration phase. The money launderer's main aim is to successfully integrate the funds so that it becomes difficult for law enforcement agencies and financial regulators to distinguish between legitimate and illegitimate funds, thereby enabling the money launderer to use the funds.

Detecting money activities is very difficult. It is like searching for a needle in a haystack. The transactions flagged by the financial institutions' compliance environments often turn out to be false positives on many occasions (Pourhabibi, Ong, Kam, & Boo, 2020).

Money launderers continue to find innovative ways to launder their criminal proceeds while lowering the detection risk in an era of advanced compliance regimes. Innovative mobile payments mark the beginning of a new era in money laundering and terrorist financing by allowing criminal networks to transfer funds around the world safely and conveniently with anonymity.

In a recent study, Teichman surveyed criminals and white-collar crime prevention experts to investigate how criminals laundered money and financed terrorism through the financial system (Teichmann, 2019). While the study captured the perspectives of only 70 participants, its findings were illuminating. The results suggested that it was very feasible for criminals to circumvent the existing compliance mechanisms using straw man, sophisticated documentation, and consulting firms. The straw man fallacy disguises the beneficial owner's identity, which is the focus area for most compliance procedures implemented by banks and other financial institutions.

Criminal networks take industry benchmarks into account when setting up legitimate consulting firms, providing real and fictitious services. Both terrorist financiers and money launderers use sophisticated documentation to prove the purpose of their transactions.

Not much information is available on the costs and benefits of implementing technology for detecting and impeding money laundering. Advances in AML regulations, which requires regulated entities to invest in technology appear to be a cost burden instead of enhancing the deterrence of money laundering, partly due to the difficulties of estimating the volume of money laundering (Magnusson, 2009; KANG, 2018).

2.3 Overview of trade-based money laundering

The term “illicit financial flows” emerged in the 1990s and was closely associated with capital flight, which received much attention from researchers and policymakers (World Bank, 1985). Illicit financial flows refer to the cross-border movement of money associated with illegal activity. The following conditions often determine the classification of the cross-border movement of funds as illicit flows:

- The funds are related to acts deemed unlawful, such as tax evasion and corruption; or
- The funds are themselves the proceeds of crime, for example, drug trafficking, smuggling of minerals, and human trafficking; or
- The purpose of transferring the funds is to finance an illegal activity, such as organized crime and terrorist activity.

Imports and exports misinvoicings are trade-based money laundering techniques considered the dark side of international trade and the largest component of illicit financial flows. Researchers have documented significant financial losses in many countries due to illicit financial flows (GFI, 2015; United Nations, 2015).

Researchers commonly use two main channels to measure illicit financial flows. The first channel involves deliberate manipulation of customs invoices on external trade, known as trade misinvoicing. The second is the leakages from the balance of payments, also known as the World Bank Residual Method (World Bank, 1985; GFI, 2015).

A firm interested in moving capital out of a country would misrepresent the exported goods' price in the invoice and other related documentation (stating it below the real value), thus bringing reduced foreign exchange into the country (Patnaik, Gupta, & Shah, 2012). Similarly, a firm engaging in imports may misrepresent the imported goods' price in the invoice and other related documentation (stating it above the real value). The firm transfers the difference between the higher invoice amount and the actual paid amount to other countries. Researchers used this popular technique to estimate trade-misinvoicing series for 126 countries and the extend of money laundering was concerning (Bhagwati, 1964; Gulati, 1987; Claessens & Naude, 1993; Patnaik, Gupta, & Shah, 2012).

It is also common for researchers to use partner country trade statistics for detecting and providing estimates of trade misinvoicing, using the following equation:

For country i, product k, and partner j at time t, export misinvoicing is calculated as follows:

$$EM_{ij,t}^{k} = I_{ji,t}^{k} - (\beta * X_{ij,t}^{k}) \qquad (2.1)$$

where I_{ji} represents imports by country j from country i according to country j's data, X_{ij} represents exports by country i to country j as reported by country i, and β is the associated freight, insurance and transportation factor. An estimated value of 1.10 is used for β (Bhagwati, 1964).

Figure 2.1 shows an illustration how a copper importer, ABC, can move money offshore illegally by over-invoicing its imports. Firstly, ABC sets up a shell company, XYZ copper exporter, in a tax haven country, Mauritius. ABC purchases copper for the value of US$1000 000 from XYZ, and an invoice showing the copper purchase of US$1 000 000 worth is forwarded to tax authorities by ABC. XYZ then ships copper worth US$500 000 to ABC and receives double the amount for the export to ABC. Bankers of XYZ then move US$500 000 into ABC's offshore bank account and use the remainder to pay XYZ for the copper exports.

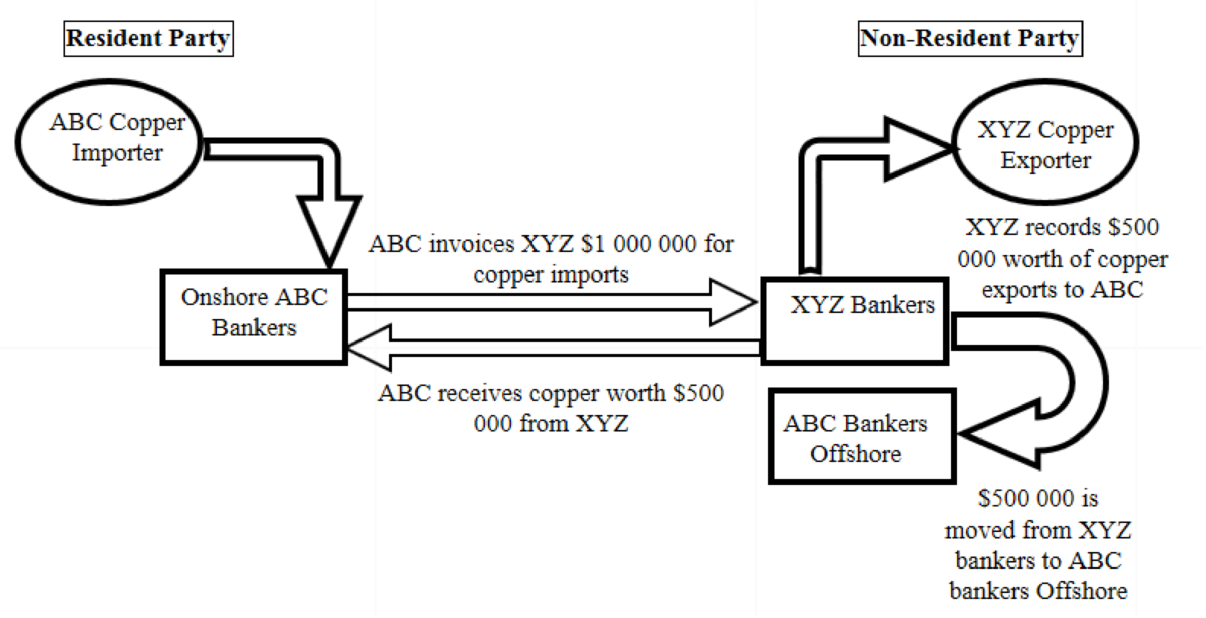

Resident Party
Non-Resident Party
ABC Copper Importer
XYZ Copper Exporter
XYZ records $500 000 worth of copper exports to ABC
ABC invoices XYZ $1 000 000 for copper imports
Onshore ABC Bankers
XYZ Bankers
ABC receives copper worth $500 000 from XYZ
ABC Bankers Offshore
$500 000 is moved from XYZ bankers to ABC bankers Offshore

Figure 2.1: Overinvoicing of imports by a copper importer

The two channels used to estimate illicit financial flows still fall short of measuring all the unrecorded flows due to the lack of bilateral trade data on services and the secretive nature of such flows. Other illicit financial flow channels include cash movements or smuggling of goods, antiques, precious gems, gold, silver, and other precious metals per the definition.

Bribery, the corruption of government officials and politicians also serve as conduits for illicit financial flows. Also, bank transfers and swap arrangements are possible channels for illegal transfers of money abroad.

Hawala is one example of the difficult to measure schemes. Under the Hawala scheme, money changes hands across national borders without any physical movement. Under the Hawala scheme, the funds move through trusted associates networks, often friends and family members. Participants use trust and code words to authorize each other to release funds without corresponding funds transfers.

The Global Financial Integrity (GFI) report on illicit financial flows indicated that of the US$1 trillion in illicit financial flows leaving developing countries annually, over 83 percent was due to trade misinvoicing, and Western economies were the beneficiaries for such funds. Illicit financial outflows exceeded combined official development aid and inward foreign direct investment in all developing countries for all but three years of the 2004-2013 time period (GFI, 2015). According to GFI, the four primary reasons for criminals to misinvoice trade transactions are money laundering, directly evading taxes and customs duties, claiming tax incentives, and dodging capital controls (GFI, 2017).

In 2011, African Ministers of Finance, Planning and Economic Development held the fourth joint conference of the African Union Commission and the United Nations Economic Commission for Africa (AUC/ECA). ECA mandated the establishment of the High-Level Panel on illicit financial flows chaired by Thabo Mbeki, the former President of the Republic of South Africa. Underlying the decision was the determination to ensure Africa’s accelerated and sustained development, emphasizing reliance on its resources (United Nations, 2015).

The ECA report estimated financial losses in Africa, resulting from illicit financial flows to over US$50 billion annually. The panel adopted a convention of breaking the illicit financial flows into three components, i.e., commercial activities, criminal activities, and corruption. The report concluded that 65% of illegal financial flow estimates for Africa were due to business activities. Multinational corporations abused transfer pricing by taking advantage of their multiple structures to shift profit across different jurisdictions and engaged in aggressive tax avoidance practices.

Studies on the financial impact of trade-based money laundering received much attention in the last two decades. However, it is not entirely clear how the practice is carried out precisely due to the problem's technical nature. The methods involve the global shadow financial systems, propped up by the tax havens and financial secrecy, corruption, and poor governance (FATF, 2006).

2.4 Regulatory response to money laundering

2.4.1 Global regulatory standards

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 (Unger & van der Linde, 2013; Cox, 2014). AML is the epitome of fraud, requiring financial institutions to undertake customer care due-diligence measures, including verifying and identifying their customers in various circumstances and monitoring transactions for suspicious activity and criminal conduct.

Global standards underpin the different AML rules and regulations within local legislations and allow countries to adopt a more flexible set of measures to target their resources effectively. The Financial Action Task Force (FATF) was established in 1989 to encourage policies to protect the global financial system against money laundering. The FATF made 40 recommendations detailing a comprehensive plan of action to fight money laundering. The FATF increased its recommendations from 40 to 49 post the September 2001 terrorist attacks in the United States of America, setting out the necessary framework to detect, prevent, and suppress the financing of terrorism (CFT) (Turner, 2011; Cox, 2014).

For example, the following Acts provide AML and CFT legislative framework in South Africa:

a) The Prevention of Organized Crime Act, 121 of 1998 (POCA).
b) The Financial Intelligence Centre Act number 38 of 2001 (FICA), as amended.
c) The Protection of Constitutional Democracy Against Terrorist and Related Activities Act number 33 of 2004 (POCDATARA).
d) Guidance Notes, Public Compliance Communications, directives, and circulars issued by the Regulators and Supervisors from time to time.

Financial regulations were reformed post the 2008 global financial crisis to enable supervisory authorities to promote the safety and soundness of financial institutions and prudential regulation (Carretta, Vincenzo, & Schwezer, 2017). Supervisory authorities adopted proactive regulatory methods based on standards, collectively known as the risk-based approach, to improve the quality of AML related risk management within regulated entities. (FATF, 2014).

The fifth recommendation of the FATF stipulates processes and procedures at the heart of anti-money laundering:

- Financial institutions should not keep anonymous accounts
- Financial institutions should undertake Customer care Due Diligence (CDD) measures, which include identifying and verifying their customers in a various circumstance.

The following CDD measures should be undertaken:

(a) Identifying and verifying the customer's identity using reliable data and information that are independent;

(b) Taking steps to identify the beneficial owner of an account;

(c) Obtaining information on the purpose and nature of the business relationship;

(d) Conducting ongoing due diligence on the business relationship and undertaking and maintaining appropriate scrutiny of transactions throughout the relationship.

This research focuses on using a combination of advanced technology and data mining methods to improve the surveillance of cross-border financial flows for financial institutions and regulatory organizations. The research proposes analytical methods for improving the surveillance of cross-border transactions, enabling financial institutions and regulatory organizations to comply with the eleventh recommendation of the FATF.

The FATF's eleventh recommendation explicitly recommends using advanced technology and data mining methods to identify suspicious transactions (Cox, 2014; FATF, (2012-2020)). No monetary thresholds apply to the reporting of suspicious and unusual transactions or activities.

Recommendation eleven states as follows:

- Draw special attention to all large, complex, unusual, and unusual transaction patterns, which have no apparent economic value or visible lawful purpose.
- The background and purpose of such transactions should be examined and made available to help competent authorities and auditors.

Recent supervisory practices leverage technology and analytical methods to enhance regulatory compliance and risk management functions in regulated entities (referred to as SupTech). Also, regulated entities adopt innovative technology to improve their compliance and risk management functions (referred to as RegTech).

In 2014, the FATF published its risk-based approach guidance paper for the banking sector, aiming to support the development of prevention and mitigation measures that are commensurate to the money laundering and terrorist financing risks identified (FATF, 2014). In the guidance paper, the FATF shared countries' supervisory experiences, seeking to illustrate the risk-based approach's implementation. The findings indicated that due to limited resources, it is practically impossible for the supervisors to inspect all banks within a calendar year. Therefore, supervisors focus their attention on AML/CFT systems through a combination of on-site examinations and off-site reviews, demonstrating their reliance on SupTech and RegTech to facilitate the risk-based approach.

2.4.2 The risk-based supervisory approach

Prudential regulation requires financial intermediaries to control risks and hold adequate capital as defined by capital requirements, liquidity requirements, by the imposition of concentration risk (or large exposures) limits, and by related reporting and public disclosure requirements and supervisory controls and processes (Morris, 2019). Hence, prudential regulation seeks to influence risk-taking in regulated entities, which is problematic since financial institutions continually make risk decisions (Carreta, Farina, & Schwizer, 2017).

The term "risk-based approach" has recently become a common phrase in the risk management and compliance functions of regulated entities. The risk-based approach is central to the effective implementation of the recommendations for preventing money laundering, terrorist financing, and financing the proliferation of weapons of mass destruction adopted by the FATF plenary in February 2012 (FATF, (2012-2020)).

SupTech and RegTech represent a move away from prudential regulation, often referred to as the traditional rules-based compliance to standards-based compliance, facilitating the risk-based approach (FATF, 2014; FATF, (2012-2020)). SupTech refers to supervisory practices leveraging technology and analytical methods to enhance regulatory compliance and risk management in regulated entities. Regulatory technology (RegTech) refers to regulated entities' adoption of innovative technology to improve their compliance and risk management functions.

The use of machine learning methods for detecting money laundering is a subject of focus by researchers in recent years. Researchers developed a supervised learning model for predicting the probability of reporting a new transaction using background information about the sender/receiver, their earlier behavior, and their transaction history as inputs. Model training used three types of historical data: regular transactions deemed legal, transactions flagged as suspicious by the bank's internal alert system, and potential money laundering cases reported to the authorities (Martin, Anders, Huseby, Geir, & Johannes, 2020).

Figure 2.2 provides an overview of the compliance and risk management environment encompassing the risk-based approach. Supervisors use data comprising activities of the regulated entities for compliance monitoring and regulatory reporting. Regulatory breaches may result from the reported activities of the regulated entities, which trigger sanctions to serve as inputs into the regulated entities' compliance function. This traditional regulatory approach focuses on the detection of non-compliance to regulations after the breach had occurred.

The variables that are associated with an increased probability of non-compliance are called risk factors. The risk factors play a central role in the prediction and prevention of regulatory breaches. In contrast to the traditional supervisory approach, the risk-based approach allows supervisors to monitor risk factors, thereby enabling them to focus on the areas of most significant risks within the regulated entities and act in advance of regulatory breaches.

In 2018, the Bank of International Settlements (BIS) surveyed the early users of SupTech to study their experiences (Broeders & Prenio, 2018). The findings showed that Central Banks use data collection methods such as APIs, machine-readable regulation, data input, data pull, and cloud computing. The data analysis methods include neural networks, supervised learning, unsupervised learning, topic modeling, random forest, and image recognition.

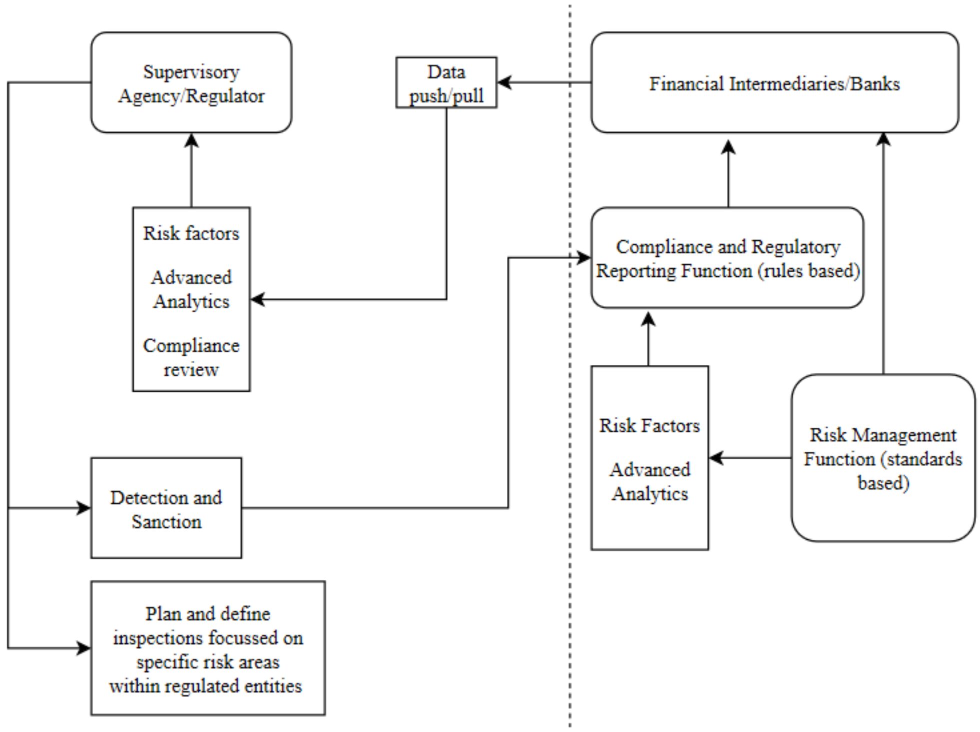

Supervisory Agency/Regulator
Data push/pull
Financial Intermediaries/Banks
Risk factors
Advanced Analytics
Compliance review
Compliance and Regulatory Reporting Function (rules based)
Risk Factors
Advanced Analytics
Risk Management Function (standards based)
Detection and Sanction
Plan and define inspections focussed on specific risk areas within regulated entities

Figure 2.2: A schematic depiction of a regulatory environment that encompasses a risk-based approach (adapted from (Lazen, 2018))

The available AML solutions and systems include ORACLE AML Express Edition, SAS AML, AML manger by Fiserv, and AIM Insight, among others. Some of these systems provide financial institutions and supervisors with advanced monitoring capabilities, which enable the setting of criteria and parameters used for customer segmentation and allocation of the clusters for customers. Also, some of the systems have built-in transaction-specific triggers to enable effective monitoring of transactions under the risk-based approach guidelines and recommendations provided by the FATF.

While AML/CFT systems may trigger suspicious activity on cross-border financial transactions above the applicable designated threshold, criminal networks circumvent these compliance mechanisms by splitting transactions and spreading them over several financial intermediaries. The consolidated data view of regulatory organizations is critical in detecting such criminal behavior.

The proposed network structure of cross-border financial flows enables the computation of measures based on behavioral patterns between residents and non-residents across industry data sources. The research proposes degree distribution and centrality measures to enable regulatory organizations and financial institutions to identify and assess risks associated with the illegal transfer of funds across country borders.

This research's scientific contribution is developing network science tools for analyzing cross-border transactions data, thereby enhancing the transaction-specific triggers based approaches to fortify the AML restraints within the risk-based approach. Foreign exchange controls and capital controls

Foreign exchange controls refer to the various forms of restrictions imposed by governments and central banks on the purchase/sale of foreign currencies by residents or on the purchase/sale of local currency by non-residents. Foreign exchange controls often accompany the capital controls to restrict speculation against currencies in developing economies.

Examples of foreign exchange controls are as follows:

1. Prohibitions for the use of foreign currency in a country.
2. Restrictions on the amount of foreign or local currency that one can import or export.
3. Quantitative limits on the quantity or value of commodities or goods that one can import or export in a given period.

Capital flight occurs when a country experiences a large-scale outflow of financial assets and capital due to political and economic instability. Policy-makers use balanced and sustainable macroeconomic policies to manage the large-scale flight of capital. In some developing countries, capital controls are necessary to manage cross-border capital flows and limit capital flight's economic impact.

In the late 1970s, free-market economists viewed capital controls as harmful to economic developments after such rules were implemented in most countries post the second World War. Such opposing views resulted in the emergence of widespread criticisms of capital controls. Later, the World Bank and the IMF persuaded countries to abandon capital controls to facilitate financial globalization. However, the Latin American debt crisis of the early 1980s, the 1997 East Asian financial crisis, the 1998 Russian financial crisis, and the 2008 global financial crisis highlighted the risks associated with capital flows' volatility.

The econometric analysis undertaken by the IMF and other economists found that, in general, countries that deployed capital controls weathered the 2008 global financial crisis better than comparable countries that did not (Gallagher, 2011; Ostry, Gosh, Chamon, & Qureshi, 2012).

In April 2011, the IMF proposed its first guidelines for using capital controls (IMF, 2011). The policy framework focused on the emerging market economies and the IMF’s low-income members. Researchers found the framework to have limited applications due to the framework's assumption that advanced economies did not need capital controls (Dierckx, 2011).

Fratzscher studied both capital controls and foreign exchange policy and concluded that capital controls appear to be less motivated by worries about financial market volatility but rather by concerns about capital inflows triggering an overheating economy (Fratzscher, 2012). The overheating of the economy can be in high credit growth, rising inflation, and output volatility. This view supports the IMF's recommendation of considering prudential measures or capital controls in response to capital inflows.

Annina Kaltenbrunner argued that given the inherent instability of international financial markets and the structural subordination assumed in developing and emerging economies, capital controls need to be permanent, comprehensive, and standardized development instruments (Kaltenbrunner, 2016).

The Organization for Economic Co-operation and Development (OECD) provides a balanced framework for countries to remove barriers to capital movement while providing the flexibility to cope with economic and financial instability situations. In 1961, the Code of liberalization of capital movements was born with the OECD when member countries were in economic recovery and development, and international movement of capital faced many barriers. The OECD recently reviewed its Code, allowing non-member countries' participation, further strengthening the instrument while providing increased flexibility to address financial stability risks. The review facilitated collective action by boosting transparency and improved decision making to assess country-specific measures and shared understandings on acceptable practices relating to managing and liberalizing capital flows (OECD, 2020).

The foreign exchange controls and capital controls restrict capital movement in and out of countries. The liberalization of capital controls and increased financial openness amplify macroeconomic management's complexity due to capital flows' volatility. Some governments often embed anti-money laundering policies and procedures within their capital flow management frameworks, further complicating their liberalization efforts and financial openness.

The anti-money money laundering policies and guidelines embedded in the capital controls provide detailed guidelines relating to the regulatory requirements and quantitative restrictions for trading foreign exchange in some countries (South African Reserve Bank, 2020). An essential aspect of the foreign exchange and capital control is the obligation imposed on the regulated entities to report the cross-border financial transactions data to the regulatory authorities.

2.5 Information privacy preservation

2.5.1 Publication of statistical data - challenges

Statistical agencies collect and publish statistical data to help researchers and policymakers make appropriate inferences and decisions to benefit the broader society (Dwork, 2011; Dwork & Roth, 2014). The primary purpose of statistics is to take findings from a sample group and generalize them to a population. Using a statistical database to learn a fact, for example, that HIV causes AIDS, enables the analyst to compute the likelihood that individuals who are not necessarily in the database will develop AIDS.

The agencies must keep individual or unit level information confidential to uphold public trust. Therefore, they often publicize the perturbed or masked version of the original data, where they remove all the explicit identifiers, such as name, address, and phone number. The practice of de-identifying data and ad-hoc generalization is not enough to render data anonymous because attributes often combine uniquely to re-identify individuals. Hence, the well-known phrase which says that ‘anonymous data’ often isn’t that anonymous.

Linkage attack occurs when adversaries collect supplemental information about an individual from multiple data sources and then combine that data to form a whole picture about their target, which is often an individual’s personally identifiable information. Figure 2.5 demonstrates linkage attack using a study by Latanya Sweeney (Sweeney, 2000) .

The leftmost circle in Figure 2.3 shows some of the ambulatory data elements collected and shared by the National Association of Health Data Organizations from hospitals, physician's offices, clinics, and so forth. The rightmost circle shows some of the data elements from the voter registration list for Cambridge Massachusetts purchased by the researcher. The data elements included the name, address, ZIP code, birth date, and gender of each voter. This information can be linked using ZIP, birth date, and gender to the medical data, thereby linking diagnosis, procedures, and medications to individuals.

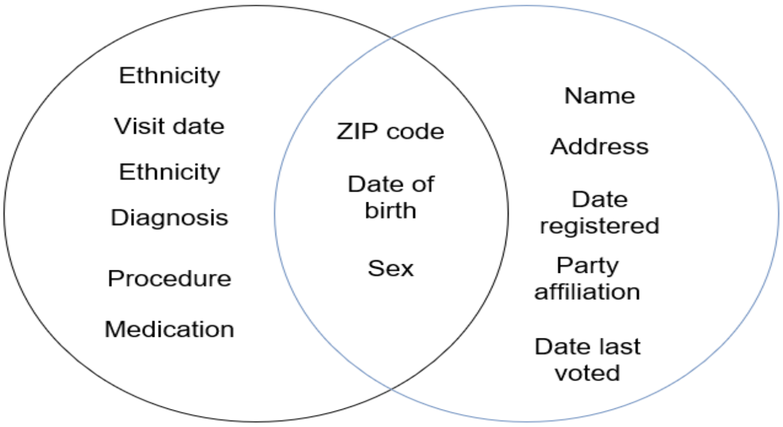


Figure 2.3: Linking to re-identify de-identified data

In 2003, researchers showed that one could reveal the entire information content of the database by only posting the results of a surprisingly small number of random queries (Dinur & Nissim, 2003). The general phenomenon is known as the Fundamental Law of Information Recovery, and its crucial insight, namely that in the most general case, one cannot protect privacy without injecting some amount of noise, led to the development of differential privacy (Dwork, McSherry, Nissim, & Smith, 2006).

In August 2006, AOL Research released data on one of its websites containing twenty million search keywords for over 650,000 users intended for research purposes. AOL did not disclose the names of the users in the report, thinking that it was enough to anonymize their names using a unique IDs. However, personally identifiable information was present in many of the queries. The New York Times newspaper was able to locate an individual from the anonymized search records by cross-referencing them with phonebook listings. Consequently, the ethical implications of using this data for research was debated (Barbaro & Zeller Jr, 2006).

In 2007, Netflix published anonymized data about movie rankings for 500,000 customers and demonstrated that an adversary with little information about an individual subscriber could identify the subscriber's record in the dataset. The researchers used the Internet Movie Database as the source of background knowledge to successfully de-identify the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information (Narayanan & Shmatikov, 2008).

Another vital purpose of statistics is to offer critical guidance in producing accurate analyses and reliable predictions to assist investigators in a wide variety of fields such as marketing campaigns, law enforcement, and financial regulation, among others. The latter is the primary focus of this research – to study the statistical properties of networks to benefit regulatory compliance.

2.5.2 Cryptography applications in preserving information privacy

Advances in wireless technology continue to create exponential growth in the number of connected devices, leading to the internet of things (IoT) revolution. IoT comprises millions of connected devices that can sense, compute and communicate. The devices capture, process, and transmit large amounts of data through the IoT network, resulting in significant information/data security concerns. Cryptography is often the solution to such data security concerns (Menon, 2017; Sreeja, Varghese, Menon, & Khosravi, 2019).

At the basic level, cryptography aims to secure the communication between two parties, usually referred to as Alice and Bob, thereby enabling them to communicate over an insecure channel so that third parties, say, Oscar, cannot understand the conversation. The insecure channel could be the computer network or the telephone line, for example. The information that Alice sends to Bob can be English text, numerical data, or anything at all, which we call plaintext. Alice encrypts the plaintext using the predetermined key to result with a ciphertext, which she sends to Bob. Upon seeing the ciphertext in the communication channel, Oscar cannot determine what the plaintext was. However, Bob knows the encryption key, which he can use to decrypt the cyphertext and reconstruct the plaintext (Stallings, 1999; Stinson & Paterson, 2018).

Researchers proposed numerous encryption algorithms to ensure the security of transmitted data through the IoT network, including the single-key encryption technique called Tiny Encryption Algorithm (TEA). TEA is a block cipher known for its simplicity of description and implementation. David Wheeler and Roger Needham of the Cambridge Computer Laboratory first designed and presented TEA at the Fast Software Encryption workshop in Leuven in 1994 (Wheeler & Needham, 2004). TEA and its numerous developed versions have a few weaknesses. Notably, it suffers from equivalent keys - each key is equivalent to three others, meaning that the effective key size is only 126 bits.

Most of the symmetric algorithms use Feistel ciphers. The Feistel cipher decomposes the encrypted plaintext into two parts and is an efficient method for implementing block ciphers. A transformation function known as the round function is applied to one half using a sub-key, and the output of the round function is XOR'ed with the other half (Rebeiro, Nguyen, Mukhopadhyay, & Poschmann, 2013; Bani Baker & Al-Hamami, 2017).

Recently, (Sreeja, Varghese, Menon, & Khosravi, 2019) proposed a novel tiny Symmetric encryption algorithm (NTSA), providing enhanced security for transferring text files over the IoT network. NTSA introduces additional key confusions dynamically for each round of encryption. Experimental results showed that the NTSA algorithm was much more secure and efficient than some state-of-the-art existing encryption algorithms.

The cross-border financial data comprise personally identifiable information protected by information privacy policies and laws. Hence, this study proposes a cryptographic technique to protect information privacy for individuals, firms, and non-profit organizations in cross-border transactional records. The secrecy of the proposed symmetric-key encryption algorithm and its limitations will be the subject of discussion in chapter 4. The study will also discuss cryptanalysis, which is the process of attempting to compute the key, given the string of ciphertext.

2.6 Financial networks and graph structures

The French economist Francois Quesnay conceptualized financial network theory and provided a precise formulation of interdependent systems in economics and the multiplier theory's origin.

Researchers used Quesnay's fundamental idea and its matrix representation to provide a statistical description of the flow of money and credit in an economy (Thore, 1969; Cohen, 1987). Thore outlined some of the matrix representation limitations of the flow of funds accounts and stressed the need for a framework of study focusing on the micro-behavior of the various monetary subjects involved, including banks, insurance companies, and other financial intermediaries, leading to development of financial networks (Nagurney & Hughes, 1992)

A financial network is any collection of traders, firms, and financial intermediaries called vertices, with connections (edges) between the entities representing transactions or the ability to mediate transactions. Nagurney and Ke proposed using methodologies of network theory and variational inequality theory to construct a network depiction of the financial economy, which explicitly includes financial intermediaries along with the 'sources' and 'uses' of funds (Nagurney & Hughes, 1992).

The model proposed by Nagurney and Ke described the various economic agents' behavior, derived the optimality conditions, and defined the governing equilibrium state. Nagurney and Hughes also suggested the formulation and solution of estimating the flow of funds account as network optimization problems (Nagurney & Ke, 2001).

This research does not seek to optimize funds' flows but to detect unusual financial transactions between residents and non-residents in cross-border transactions to combat money laundering and terrorist financing. Hence, the research adopts the traditional graph theory applied in many network science applications, such as biological networks, information networks, technological and social networks. Figure 5.1 shows a visualization of various forms of networks.

(a) The network structure of the financial economy with Intermediation (Nagurney & Ke, 2001).

(b) The citation network of academic papers in which vertices are papers and the directed edges are citations of one article by another (Newman M. J., 2003).

(c) The World Wide Web, a network of text pages accessible over the Internet, in which the vertices are pages, and the directed edges are hyperlinks (Newman, 2003).

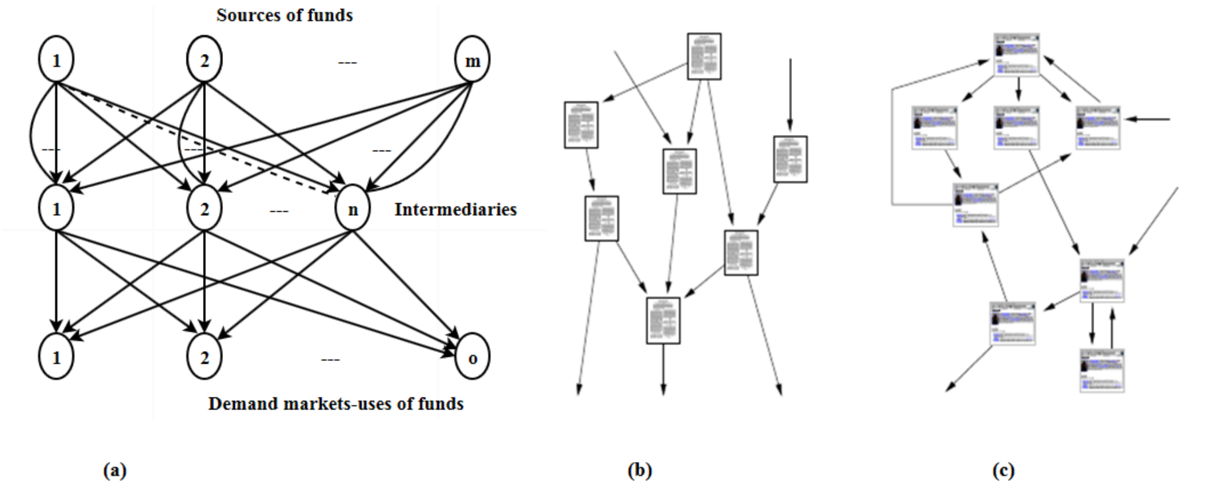


Figure 2.4: Illustrations of various forms of network structures.

Researchers use graph-based substructures for detecting potential fraudulent cases in the trading networks, consisting of a group of traders that trade with each other in specific ways to manipulate the stock market (Xiong H & Zhou, 2010). Economic policymakers paid much attention to the financial networks to better understand the root causes of the multiple financial crises that have devastated economies worldwide. (Silva, de Souza, & Tabak, 2016; Babus, 2016). (Sun, Qu, Chakrabarti, & Faloutsos, 2005) investigated methods for exploiting communities in bipartite graphs graph to identify node anomalies.

Using network science to analyze large and complex data sets to detect anomalies in the data set is fast becoming more important and exciting than merely learning about its structure. Anomaly detection is the branch of data mining concerned with discovering rare occurrences in data sets.

The challenge associated with the research problem is that there is no unique definition for anomaly detection in the cross-border financial network. The reason is that the general definition of an anomaly or an outlier is vague. Hawkins gave the first definition of an outlier, dated back to 1980: "An outlier is an observation that differs so much from other observations to arouse suspicion that a different mechanism generated it (Akoglu, Tong, & Koutra, 2015)."

(Li, et al., 2020) proposed using a multipartite network model (FlowScope) for money laundering involving high-volume flows of funds through bank accounts chains. The research considered that money launderers make fraudulent transfers from source accounts to destination accounts through one or many layers of middle accounts to conceal funds and decrease detection accuracy.

Theoretical analysis showed that FlowScope guarantees the amount of money that fraudsters can transfer without being detected, outperforming state-of-the-art baselines by accurately detecting the accounts involved in money laundering in both injected and real-world data settings (Li, et al., 2020). The research appears to improve the existing graph fraud detection approaches, which focus on dense subgraph detection (Tang & Yin, 2005; Wang & Yang, 2007).

2.7 Centrality measures for bipartite networks

Many real-world networks, including social, biochemical, World Wide Web, and other networks, divide naturally into communities. In such communities, the density of edges within a community is relatively higher than the edges' density between communities. Identifying communities is challenging as the network groups can overlap or be well hidden within the network structure.

The definition and analysis of groups or communities within networks is a large area of network theory research (Nadakuditi & Newman, 2012; Newman M. J., 2013; Peixoto, 2013; Newman & Peixoto, 2015; Hosseini-Pozveh, Zamanifar, & Naghsh-Nilchi, 2017). The cluster/community-based methods for anomaly detection in networks rely on finding densely connected node groups and identifying the nodes and edges that have connections across communities. The definition of anomaly under this setting refers to finding nodes/edges that do not directly belong to one particular community. Methods that exploit nodes' communities to identify anomalies in bipartite graphs include (Sun, Qu, Chakrabarti, & Faloutsos, 2005; Akoglu, Tong, & Koutra, 2015).

Many sophisticated algorithms for community detection have been developed, such as hierarchical clustering and spectral partitioning. Some are based on modularity, which measures the quality of a network division into groups of nodes found by a community detection algorithm.

Modularity is the fraction of the edges that fall within the given groups minus the expected fraction if edges were distributed at random. The modularity function enables evaluating the quality of a partition of a network into groups of nodes. The higher the modularity, the higher the quality of the network partition. Hence, the large values of modularity are indicative of a well-pronounced community structure in the observed network. Newman formally defines modularity as

$$Q = \frac{1}{2m}\sum_{i,j}\left[a_{ij} - p_{ij}\right]\delta(g_i, g_j) \tag{2.2}$$

where $2m$ is the sum of the degrees in the network, a_{ij} is the $(ij)^{th}$ entry of the adjacency matrix A, p_{ij} is the expected number of edges between nodes u_i and u_j if the network was random, g_i is the community with which node u_i is associated and $\delta(a,b) = 1$ if $a = b$ and 0 otherwise (Newman, 2006; Lancichinetti & Fortunato, 2011).

In unipartite networks, ties often have a strength naturally associated with them, differentiating them from each other. Operationalizing tie strength as weights is standard practice, and some proposed network measures for weighted networks include the discussed node centrality measures discussed: degree, closeness, and betweenness. However, these generalizations solely focused on tie weights and not on the number of ties, which was the central component of the original measures. (Opsahl, Agneessens, & Skvoretz, 2010) proposed generalizations that combine both these aspects and illustrated the benefits of this approach by applying one of them to Freeman's EIES dataset (Freeman L. C., 1978; Opsahl, Agneessens, & Skvoretz, 2010).

Researchers often reduce bipartite networks to unipartite networks to simplify their analysis. However, the one-mode projections often compromise the network's structural properties and can lead to imprecise network measurements (Lehmann, Schwartz, & Hansen, 2008). Moreover, there are numerous ways to obtain the one-mode projections, each with different characteristics and peculiarities. (Yang, Aronson, & Ahn, 2020) presented BiRank, an R and Python package that performs PageRank on bipartite networks directly.

In a recent study, researchers argued that there is no commonly accepted way to compare different centrality metrics' effectiveness and reliability, resulting in a newly developed theoretical framework for evaluating and comparing the metrics. Tests on a large set of networks using the new framework showed that the standard centrality metrics perform unsatisfactorily, highlighting intrinsic limitations for describing the centrality of nodes in complex networks (Sciarra, Chiarotti, Laio, & Ridolfi, 2018).

Recently-developed bipartite methods include BiRank and CoHITS (Akoglu, Tong, & Koutra, 2015). The methods proved to be significantly more robust measures of prescription drug-seeking and better predictors of subsequent opioid overdose than traditional centrality estimates, including PageRank in a one-mode network projection (Aronson, Yang, Odabas, Ahn, & Perry, 2020).

2.8 Summary

Chapter two outlined the main theories and reviewed the literature applicable to this research, covering money laundering and the global response mechanisms to the money laundering problem.

Most importantly, the chapter outlined the risk-based supervisory approach, which leverages advanced technology and analytical methods to enhance regulatory compliance and risk management in regulated entities.

This research aims to contribute meaningfully to the deterrence of money laundering by leveraging technology and data mining techniques such as visualizations and statistical properties of networks in line with the fifth and the eleventh recommendations of the FATF.

Chapter 3: Graph theory fundamentals

3.1 Introduction

This chapter provides background on the fundamental concepts and definitions from mathematical graph theory, starting with a brief history of graph theory and graphs' representations. The chapter gives an overview of some of the essential properties of graph structures, network flows, and matching results in bipartite graphs. These results are mostly attributable to the work done by König (König, 1936).

The chapter lays a solid mathematical foundation for this research, thereby enabling the introduction of the directed and weighted bipartite as a model for cross-border financial flows and the network's characterization using degree distributions in subsequent chapters.

The chapter concludes with an overview of the cross-border financial flows data set and summarizes the chapter.

3.2 A brief history of graph theory

In the earliest known graph theory paper, Euler solved the famous seven bridges of Königsberg problem in 1736 (Newman J. R., 1953). The inhabitants of Königsberg, in what was then Prussia (now called Kaliningrad), debated whether it was possible to take a walk through each part of the city and crossing each of the seven bridges of Pregel River only once. Figure 3.1 shows Euler's drawing of the bridges of Königsberg in 1736, with geographical designations (Newman J. R., 1953; Gribkovskaia, Halskau, & Laporte, 2007).

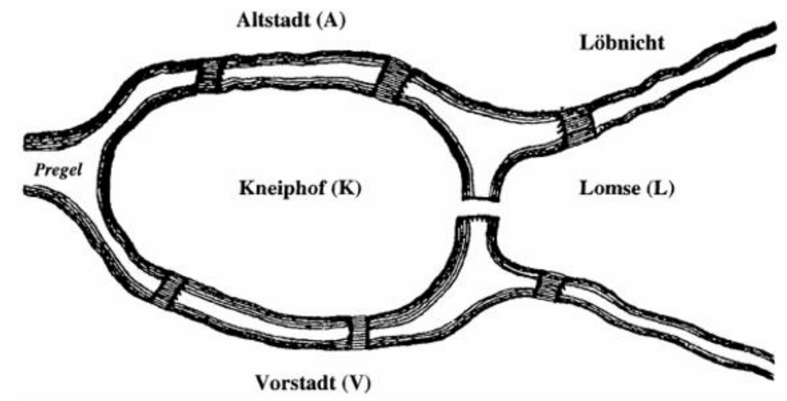


Figure 3.1: Euler's drawing of the bridges of Königsberg in 1736

Leonhard Euler was chair of mathematics at the St. Petersburg Academy of Science when he solved this significant problem. He proved that no such walk could exist, pointing out that route inside each landmass is irrelevant. The only significant feature of a route is the sequence of bridges crossed (Euler, 1736). Euler drew an undirected graph, where he assigned a node to each of the four landmasses and edges connected the node pairs if a bridge connected the two corresponding landmasses.

Euler argued that the graph must be connected, and each of the nodes must have an even degree for the solution to exist, which was not the case. Figure 3.2 shows the graph representing Euler's map of Königsberg depicted in Figure 3.1. A graph containing a closed walk using each edge exactly once is now called unicursal or Eulerian.

Euler did not provide formal proof of connectedness and evenness as sufficient conditions for unicursality. It was Hierholzer who provided the first polynomial-time "end-pairing" algorithm for detecting a unicursal walk in a connected and undirected graph (Hierholzer, 1873; Gribkovskaia, Halskau, & Laporte, 2007). Edmonds and Johnson provide a simple description of the Hierholzer algorithm, while Fleischner described several alternative procedures of the algorithm (Fleischner, 1991; Edmonds & Johnson, 1995).

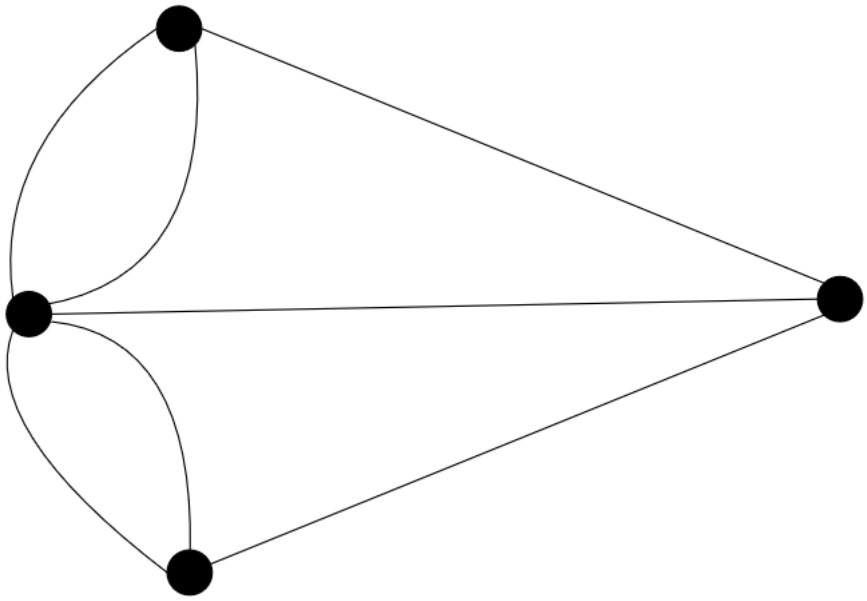


Figure 3.2: Graphical representation of Euler's map depicted in Figure 3.1

The model used by Leonhard Euler not only provided the solution to the problem but also gave birth to the mathematical discipline of graph theory as we know it today. Graph theory has emerged as a proper discipline with several books already published. Notable contributions to graph theory were due to König, who published the first book on the subject (König, 1936), and Berge published the second book on the subject in 1958 (Harary, 1979).

Graph theory was applied successfully in many network science applications, such as biological networks, information networks, technological and social networks. The focus of such studies recently shifted away from the analysis of simple structures to complex networks analysis and is mostly driven by advances in computing and communication technologies, which enable the collection and storage of large data sets that are both structured and unstructured.

In recent years, researchers use graphs for detecting substructures and hierarchies in complex network communities. Several books on network science have been published (Dorogovtsev & Mendes, 2003; Newman, 2010; Dehmer, Pickl, & Wang, 2015).

3.2.1 Fundamentals of graph theory

The combinatorial methods found in graph theory ensure that the construction techniques are considerably different from the classical computational approach. Graphs arise naturally in the study of other mathematical structures such as polyhedra, lattices, and groups. In addition to the purely structural relationships that are the defining characteristics of a graph, quantitative characteristics are imparted to the graph's vertices and edges, resulting in a network. For example, the flow of energy is the quantitative measure for electrical networks, while traffic flow is the associated quantitative measure for transportation networks.

A graph consists of a nonempty set V, a (possibly empty) set E disjoint from V, and a mapping ϕ that associates an unordered pair of distinct vertices with each edge. The elements of V and E are called vertices and edges, respectively, and ϕ is called the incidence mapping associated with a graph.

We say that an edge e joins vertices v and w if $\phi(e) = \{v, w\}$, written vw, and that e has ends v and w. An edge is incident with a vertex v if it is one of its ends, and two vertices joined by an edge are adjacent. It should be noted that adjacency is a relationship between two like elements (either vertices or edges), while incidence is a relationship between unlike elements. A graph is usually denoted by G or (V, E, ϕ), or (V, E) when the incidence mapping is implicit in the definition. To avoid ambiguities, we denote the set of vertices of a graph by $V(G)$ and the set of edges by $E(G)$. The number of elements in any set S is denoted by $|S|$. The number of vertices and edges of a finite graph $G = (V, E)$ is denoted by $|V|$ and $|E|$, respectively. A graph is simple if it has no loops or parallel edges. A simple graph in which any two vertices are adjacent is called a complete graph. An empty graph is one whose edge set is empty. A graph is finite if both its vertex set and edge set are finite.

In applications of graph theory, the need to introduce directions on the graph edges arises naturally. Directed edges may represent a relationship between vertex pairs which is not symmetric. For example, when dealing with problems of traffic flow, it is necessary to know the permitted direction of traffic flow since some roads are one-way streets. Clearly, the direction is a very important factor in such cases. Also, directed edges may be introduced in order to establish a frame of reference and thus avoid ambiguities.

A graph in which all the edges are ordered pairs (and therefore called arcs) is called a digraph. Formally, a directed graph (digraph) Dconsists of a nonempty set V of vertices, a set A of arcs or directed edges (disjoint from V), and a mapping Δ of A into $V \times V$. The mapping Δ is called the directional incidence mapping associated with the directed graph. If $a \in A$ and $\Delta(\mathrm{a}) = (\mathrm{v}, \mathrm{w})$, then arc a is said to have v as its initial vertex and v as its terminal vertex. Directed graphs are usually denoted by D or (V, A, Δ), or by (V, A) when Δ is not used explicitly. The associated undirected graph of a directed graph is obtained by disregarding the ordering of the end points of each arc.

Parallel edges, also called multi-edges, can be drawn between vertices, which refer to multiple edges between the same pair of vertices. Self-edge or loop, in which the source vertex is identical to the target can be also drawn. Multi-edges are necessary if there are different types of interactions between the same pair of vertices as is often observed in real-world systems.

In the simplest case, graphs are assumed to have no multi-edges or self-edges. Figure 3.3 represents a graph G whose vertex set V is $\{v_1, v_2, v_3, v_4, v_5\}$ and edge set E consists of edges $\{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8\}$, where vertex v_3 contains a self-edge e_5 and the vertex set $\{v_4, v_5\}$ have multi-edges $\{e_6, e_7, e_8\}$. The case when an edge joins a vertex with itself (self-edge) is not part of this research due to the nature of cross-border financial flows transactions.

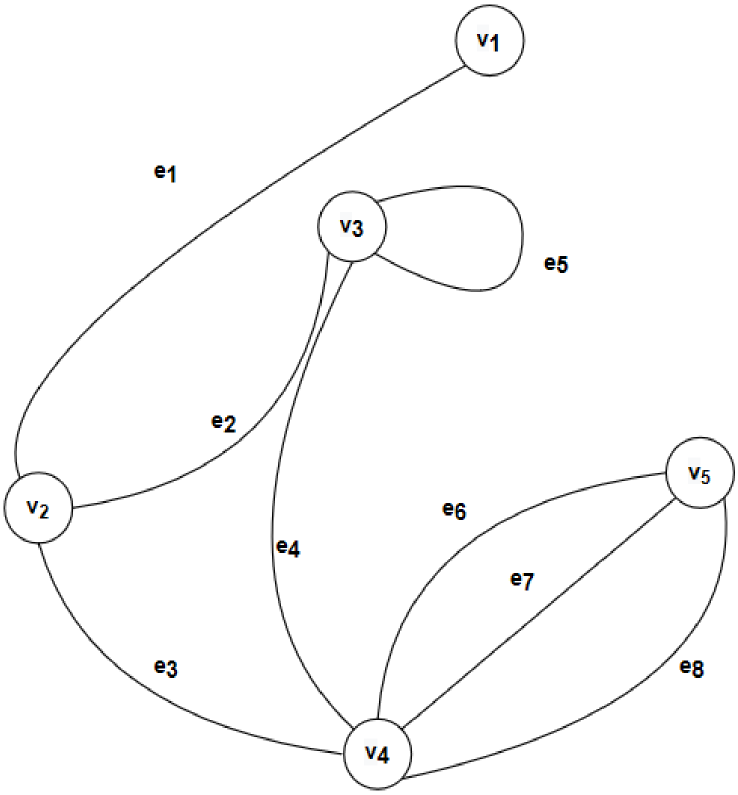


Figure 3.3: A multi-edged graph with a self-edge

An important property of graphs that is frequently used in network science is the degree of each vertex. The vertex degree is defined as the number of its incident edges, with loops counted twice. A vertex with degree zero is called an isolated vertex. A graph is called regular if every vertex has the same degree, and it is k −regular if that degree is k. In the case of a simple network consisting of N vertices, in which $A_{ij} = A_{ji} = 1$, if an edge is drawn between vertices i and j, the degree of vertex i, k_i, is expressed as

$$k_i = \sum_{j=1}^{N} A_{ij} \tag{3.1}$$

The distribution of the vertex degree provides a characterization of the network structure and was defined by Barabási and Albert, see (Barabási & Albert, 1999). This distribution is defined as follows:

$$p(k) = \frac{1}{N}\sum_{j=1}^{N} \delta(k_i - k) \tag{3.2}$$

where $\delta(x)$ is the Kronecker's delta function, which returns 1 when $x = 0$ and returns 0 otherwise (Takemoto & Oosawa, 2012). Hence, $p(k)$ represents the probability that a randomly chosen vertex on the graph will have degree k. The degree distribution can be obtained by plotting a histogram of $p(k)$ for any given network.

Given two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, their union is another graph $G_3 = (V_3, E_3)$ denoted by $G_3 = G_1 \cup G_2$, where the vertex set $V_3 = V_1 \cup V_2$ and the edge set $E_3 = E_1 \cup E_2$. Figure 3.4 shows a diagrammatic representation of the union of two graphs. The intersection of two graphs G_1 and G_2 denoted by $G_1 \cap G_2$ is a graph G_4 consisting only of those vertices and edges that are in both G_1 and G_2.

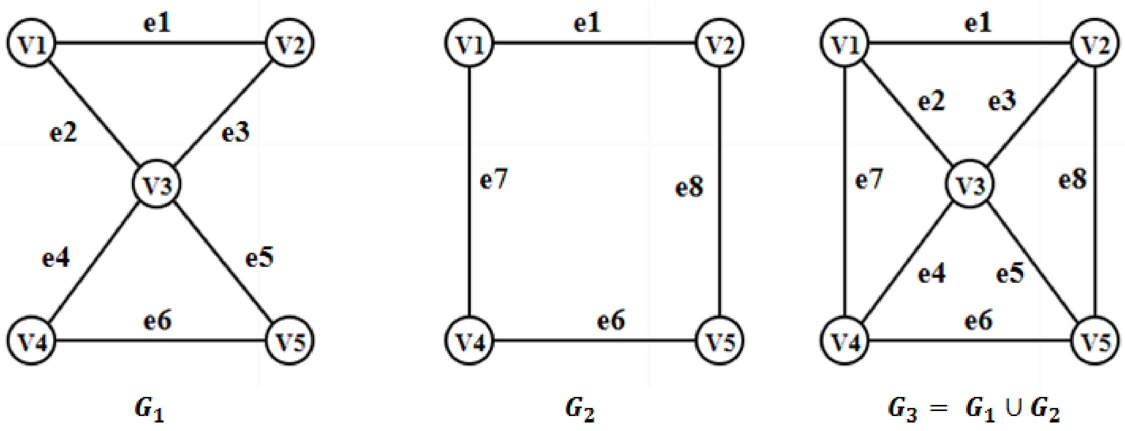


Figure 3.4: The union of two graphs

Given a graph $G = (V, E)$, the graph $F = (V_1, E_1)$ is called a subgraph of G if and only if $V_1 \subseteq V$ and $E_1 \subseteq E$. If F is a subgraph of G then the relationship is denoted by $F \subseteq G$. When $F \subseteq G$ but $F \neq G$, i.e., $V \neq V_1$ and $E \neq E_1$, then F is called a proper subgraph of G. A spanning subgraph of G is a subgraph F with $V = V_1$, meaning that it is obtained by edge deletion only. Subgraphs can be naturally derived using two operations, i.e., edge deletion and vertex deletion resulting in smaller graphs that satisfy the conditions stated above. An edge-deleted subgraph of G is denoted by $G \setminus e$.

An induced subgraph F is obtained from a subset of the vertices of G and all the edges connecting pairs of vertices in that subset. Two subgraphs H_1 and H_2 are said to be vertex-disjoint if $V(H_1) \cap V(H_{2)} = \emptyset$, and similarly H_1 and H_2 are edge-disjoint subgraphs if $E(H_1) \cap E(H_{2)} = \emptyset$. Given spanning subgraphs $F_1 = (V, E_1)$ and $F_2 = (V, E_2)$ of a graph $G = (V, E)$, we may construct a spanning subgraph of G whose edge set is the symmetric difference $E_1 \Delta E_2$ of E_1 and E_2.

Figure 3.5 shows the symmetric difference of two spanning subgraphs F_1 and F_2 of a graph with five vertices. Note the absence of the common edges $\{e1\}$ and $\{e6\}$ on $F_1 \Delta F_2$.

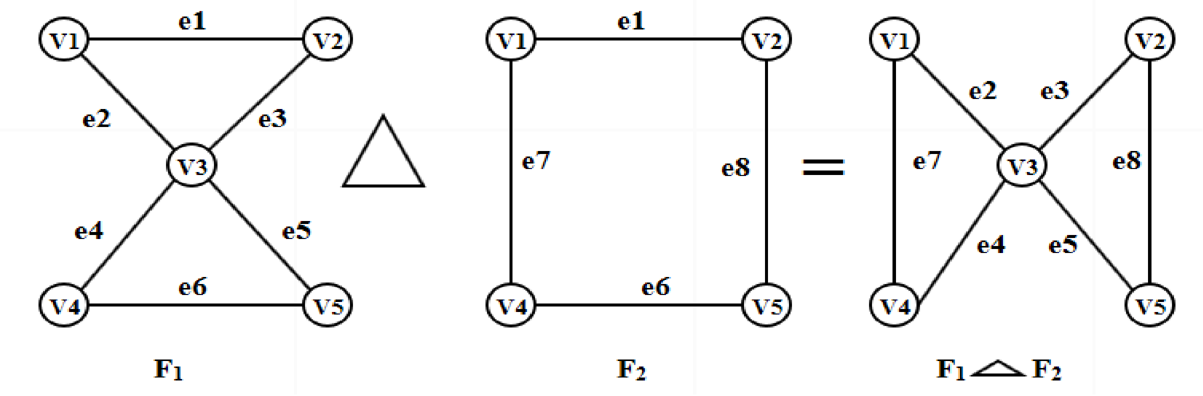


Figure 3.5: The symmetric difference of two graphs

Given a graph G, a walk in G is a finite sequence of vertices and edges of the form $v_0e_1v_1 \ldots v_{n-1}e_nv_n$ where $\phi(e_i) = v_{i-1}v_i$ for $1 \leq i \leq n$. When it is clear which edges are involved, a walk $W = v_0e_1v_1 \ldots v_{n-1}e_nv_n$ will be denoted by $W = v_0v_1 \ldots v_{n-1}v_n$. A walk from vertex u to vertex v is called a (u, v)-walk. The number of edges in a walk is called its length. Given a (u, v)-walk W, the (v, u)-walk obtained by traversing W in the opposite direction is denoted by W^{-1}.

A walk in which all the edges are distinct is a trail. If, in addition, the vertices $v_0, v_1, \ldots, v_n$ are distinct (except, possibly, $v_0 = v_n$), then the trail is a path. A trail that traverses every edge of a graph is called a Eulerian trail, named after Leonhard Euler. A walk in a graph is closed if its initial and terminal vertices are identical. A closed trail in which the origin and internal vertices are distinct is called a cycle.

A graph is connected if and only if there is a path between each pair of vertices. In a connected graph G, the length of the shortest path between two vertices u and v is called the distance between u and v, denoted by $d_G(u, v)$. The set of vertices at distance i from v is denoted by $N_{i,G}(v)$. $N_{1,G}(v)$ is called the neighbourhood of v. Analogously for a subset of vertices $S \subset V$, we denote the set of all the vertices of G which are adjacent to at least one vertex in S by $N_G(S)$.

The distance between two vertices in a connected graph is a very important property used in network science to characterize the structure of networks. In large networks, the length of the shortest path between a given pair of nodes (vertices) is known to be surprisingly small. This property is referred to as the "*small-world property*". This property was demonstrated through experiments conducted by Stanley Milgram, who was an American social psychologist and a professor at Yale University in the 1960s (Newman, 2013).

A directed graph is weakly connected if the underlying undirected graph obtained by replacing all directed edges of the graph with undirected edges is a connected graph. A graph is strongly connected if it contains a directed path from u to v and a directed path from v to u for every pair of vertices (u, v). A tour of a connected graph G is a closed walk that traverses each edge of G at least once, and a Eulerian tour one that traverses each edge of G exactly once.

One can obtain a disconnected graph by deleting edges from a connected graph as discussed earlier. Such subsets of edges are called edge separators or edge cuts. Many important properties of graphs can be successfully investigated by using derived subgraphs since the derived subgraphs have fewer vertices than the original graph. In this research, identification of such subsets is critical to the implementation of an effective cross-border financial surveillance model.

The edge connectivity $\lambda(G)$ of a graph G is defined to be the minimum number of edges whose removal disconnects G. Furthermore, the graph G is k-edge connected if $\lambda(G) \geq k$. Thus, a non-trivial graph is 1-edge connected if and only if it is connected. Analogously, a separating set in a connected graph G is a set of vertices whose deletion disconnects G. When a vertex is deleted, its incident edges are also removed. If a separating set contains only one vertex v, we call v a cut-vertex.

Given a graph $G,$ a matching is a subgraph of G where every vertex has degree 1. In particular, the matching represents a collection of edges that do not have a common vertex. A matching of a graph G is perfect if it is a spanning subgraph of G. A perfect matching has $\frac{|V|}{2}$ edges. A matching with the largest possible number of edges is called a maximum matching. The matching number $\nu(G)$ of a graph G is the size of the maximum matching. A maximal matching is a matching M of a graph G with the property that if any edge not in M is added to M, it is no longer a matching. Examples of maximal matching and maximum matchings are indicated in Figures 3.6a and 3.6b, respectively.

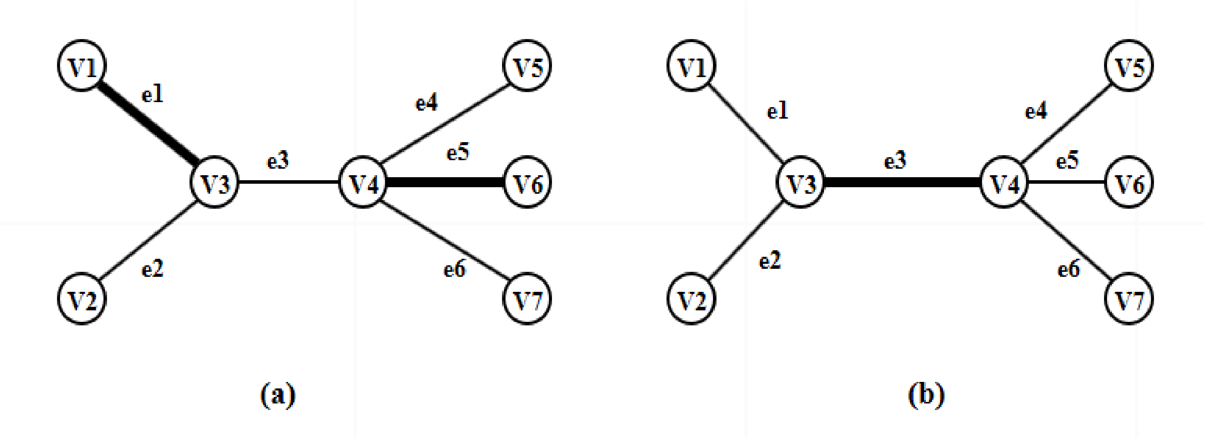


Figure 3.6: (a) A maximum matching of a graph **G**. (b) A maximal matching of graph **G**

A vertex is said to be covered or saturated by a matching M if it is incident with an edge of M. An M-alternating path of cycle in G is a path or cycle whose edges are alternating in M and $E \setminus M$. An M-alternating path might or might not start or end with edges of M. If neither its origin nor its terminus is covered by M then the path is called an M-augmenting path. The augmenting path is relevant in the study of maximum matchings.

3.2.2 Representations of graphs

Graphs (networks) are mathematically represented in several different ways. Much research effort has gone into the development of efficient methods for network related computations (algorithms) and storage (data structures) in computer science. Such developments facilitate the transition from graphs as purely mathematical objects to graphs as practical tools for use in the analysis of networks.

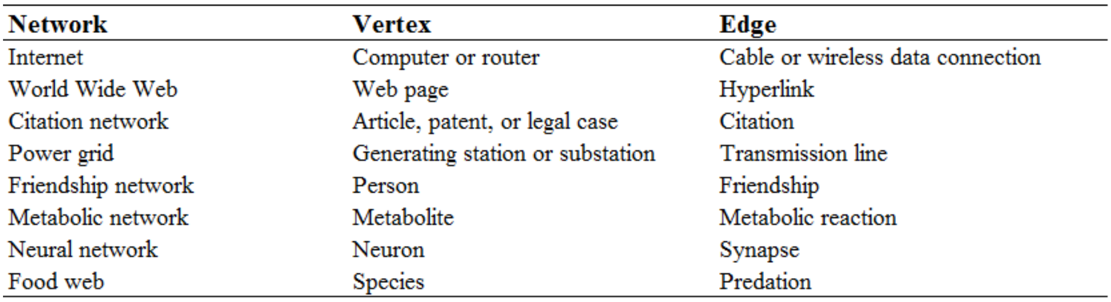
Edge lists and adjacency matrices are the two most common ways of representing networks. An edge list is a simple two-column list of all adjacent node pairs. The adjacency matrix is a square matrix whose elements indicate whether pairs of nodes are adjacent or not in the network. The adjacency matrix is the most preferred method for network representation, given that matrices are fundamental data objects in most programming and software environments.

In the simple network case, the adjacency matrix $\boldsymbol{A}$ is the matrix with elements A_{ij} such that

$$A_{ij} = \begin{cases} 1 \; if \; there \; is \; an \; edge \; between \; node \; i \; and \; node \; j, \\ 0 \; otherwise \end{cases} \qquad (3.3)$$

In such networks the adjacency matrix is symmetric, and the diagonal elements are all zero. Multi-edges are represented by setting the corresponding matrix A_{ij} equal to the multiplicity of the edge. A single self-edge from node i to itself is represented by setting the corresponding diagonal element A_{ii} of the adjacency matrix equal to 2.

In a directed network each edge has a direction. The two directions are counted as being distinct directed edges and parentheses are used to denote the ordered pairs. The arrowheads on the edges represent the direction. The adjacency matrix for such networks is a $n \times n$ matrix $\boldsymbol{A}$ with $A_{ij} = 1$ if and only if $(i, j) \epsilon E$ or zero otherwise. In addition to direction, the weight associated with a connection is also important in real-world systems. For example, in the World Wide Web, the weight of hyperlinks for famous sites may be different from those of personal sites that are visited by only a few people. A weighted network may be represented using the weights as the entries of the adjacency matrix. Figure 3.7 shows a few examples of networks representing real-world systems.



Network	Vertex	Edge
Internet	Computer or router	Cable or wireless data connection
World Wide Web	Web page	Hyperlink
Citation network	Article, patent, or legal case	Citation
Power grid	Generating station or substation	Transmission line
Friendship network	Person	Friendship
Metabolic network	Metabolite	Metabolic reaction
Neural network	Neuron	Synapse
Food web	Species	Predation

Figure 3.7: Vertices and edges in network models based on Newman (Newman, 2013).

3.3 Properties of bipartite graphs

Some real-world problems can be modelled as a graph where the edges represent compatibility and the goal is to create the maximum number of compatible pairs. In such situations, bipartite graphs arise naturally.

The first systematic investigation of the properties of bipartite graphs was begun by König between 1914 and 1916 and documented in his famous book (König, 1936). This section provides the theoretical background on several characterizations of bipartite graphs, including the most widely used result from König. It also discusses matchings in bipartite graphs and the results from Hall (Hall, 1935).

The first book on bipartite graphs alone was published based on graduate courses taught by A.S. Asratian at Yerevan State University and lectures given by R. Häggkvist at Umea University (Asratian, Denley, & Häggkvist, 1998).

Formally, a bipartite graph $G(V_1, V_2)$ is a graph in which the vertex set V can be partitioned into two sets of vertices, V_1 and V_2 with the following properties:

1. If $v \epsilon V_1$, then it may be only be adjacent to vertices in V_2.
2. If $v \epsilon V_2$, then it may be only be adjacent to vertices in V_1.
3. $V_1 \cap V_2 = \emptyset$.
4. $V_1 \cup V_2 = V$.

The sets V_1 and V_2 may be thought of as a colouring of the graph with two colours: if one colours all vertices in V_1 black, and all the vertices in V_2 white, each edge has endpoint of differing colours. Hence, the sets V_1 and V_2 are called the colour classes of G and $(V_1 \cup V_2)$ is a bipartition of G. Therefore, for a graph to be bipartite it must be possible to colour the vertices with at most two colours, so that no two adjacent vertices have the same colour.

A bipartite graph with bipartition (V_1, V_2) is denoted by $G(V_1, V_2)$. If $G(V_1, V_2)$ is a simple graph and every vertex in V_1 is joined by to every vertex in V_2, then G is called a complete bipartite graph. A graph is called m by n bipartite if $|V_1| = m$ and $|V_2| = n$, and a balanced bipartite graph if $|V_1| = |V_2|$. Figure 3.8 shows a 3-regular graph with coloured vertices, which is also represented as a balanced bipartite graph with bipartitions $V_1 = (X, Y, Z)$ and $V_2 = (K, L, M)$.

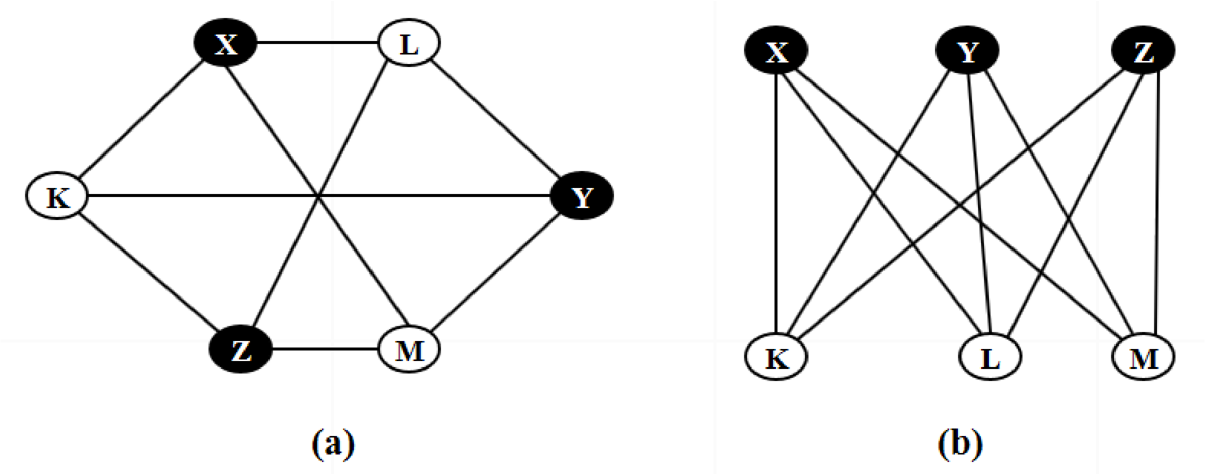


Figure 3.8: (a) A 3-regular graph with coloured vertices, (b) a unique bipartition of (a).

Cycles of bipartite graphs are of even length. Some simple observations about the structure of bipartite graphs are stated below and followed by some widely used results that characterises bipartite graphs. There are many characterisations of bipartite graphs, and therefore many ways to recognise them algorithmically.

Property 3.3.1. *A connected bipartite graph has a unique bipartition*

Property 3.3.2. *A bipartite graph, without isolated vertices, which has t connected components, has* 2^{t-1} *bipartitions.*

For example, the complete bipartite graph on Figure 3.2 has one bipartition because it consists of one connected component.

Lemma 3.3.3. *If* G *is a bipartite graph, and the bipartition of* G *is* (V_1, V_2)*, then*

$$\sum_{v \in V_1} deg(v) = \sum_{v \in V_2} deg(v)$$

Proof. By induction on the number of edges of the graph G. Suppose $|V_1| = m$ and $|V_2| = n$ for some $m, n > 0$. The case when both m and n are equal to one is trivial since only one edge can be drawn between the vertices.

Now take the spanning subgraph of G without the edges and proceed with induction as follows: add one edge from any vertex in V_1 to any vertex in V_2. Then

$$\sum_{v \in V_1} deg(v) = \sum_{v \in V_2} deg(v) = 1$$

Now suppose this is true for $n - 1$ edges and add one more edge. Since this edge adds exactly one to both $\sum_{v \in V_1} deg(v)$ and $\sum_{v \in V_2} deg(v)$, then it is true for all $n \in N$.

Theorem 3.3.4. *If* G *is a k-regular bipartite graph with* $k > 0$ *and the bipartition of* G *is* (V_1, V_2)*, then the number of elements in* V_1 *is equal to the number of elements in* V_2.

Proof. Since the graph is k –regular, $\sum_{v \in V_1} deg(v) = k|V_1|$ and $\sum_{v \in V_2} deg(v) = k|V_2|$. From the previous lemma, we have

$\sum_{v \in V_1} deg(v) = \sum_{v \in V_2} deg(v) \Longrightarrow k|V_1| = k|V_2| \Longrightarrow |V_1| = |V_2|$.

Lemma 3.3.5. *A graph G is a bipartite graph if and only if G contains no closed walk of odd length.*

Proof. Since an odd cycle is also an odd walk the condition is certainly sufficient. Let G be a bipartite graph and $W = v_0v_1v_2 \dots v_nv_0$ be a closed walk in G. Consider the level representation of G with respect to v_0. Define the sequence $\alpha_1\alpha_2, \dots \alpha_{k+1}$ by

$$\alpha_i = \begin{cases} 1\ if\ 1 \leq i \leq k\ and\ the\ level\ of\ v_{i-1}\ is\ less\ than\ the\ level\ of\ v_i \\ 0\ otherwise \end{cases}$$

Then, since W is closed, the sequence must contain equal numbers of 1's and 0's, hence must be of even length. Therefore W is also of even length.

The theorem below was derived by König and is one of the most widely results used to characterise bipartite graphs. As stated earlier, there are many other ways to characterise bipartite graphs.

Theorem 3.3.6. *A graph G is a bipartite graph if and only if G has no cycle of odd length.*

Proof. Suppose that G is a bipartite graph with bipartition (V_1, V_2), and $C = v_0v_1v_2 \dots v_nv_0$ is a cycle of G. Without loss of generality, we may assume that $v_0 \in V_1$. Given that G is a bipartite graph, $v_1 \in V_2$ since the vertices of C must alternately be in V_1 and V_2. Hence n must be odd and, and C is an even cycle.

It suffices to prove the converse when G is connected. Assume that G contain only even cycles and let v be an arbitrary vertex of G. Partition all other vertices of G based on the parity distance (even or odd) from vertex v. That is, let

$$V_1 = \{u \in V : d_G(u, v)\ is\ even\}$$

$$V_2 = \{u \in V : d_G(u, v)\ is\ odd\}$$

Clearly, $V_1 \cap V_2 = \emptyset$ and $V_1 \cup V_2 = V$. It remains to show that (V_1, V_2) is indeed a bipartition of G. Suppose that x and y are two vertices of V_1, and that $xy \in E$. It follows that

$$x \in V_1 \Longrightarrow \exists (\mathrm{v}, \mathrm{x}) - \text{path } P_1\ of\ even\ length$$

$$y \in V_1 \Longrightarrow \exists (\mathrm{v}, \mathrm{y}) - \text{path } P_2\ of\ even\ length$$

Concatenate (v, x)-path P_1, the edge xy, and the (y, v)-path P_2^{-1} to obtain a closed odd walk. By the previous lemma, the graph must contain an odd cycle. Hence, it must be the case that (V_1, V_2) is a valid bipartition, so G is bipartite.

3.4 Matching in bipartite graphs

In many applications of graph theory, many questions of practical interest amount to finding a matching in a bipartite graph G with bipartition (V_1, V_2), which covers every vertex in V_1. For example, suppose that a certain number of jobs are vacant. Given a group of applicants for these jobs, fill as many jobs as possible, assigning applicants only to jobs for which they are qualified. A bipartite graph G with bipartition (V_1, V_2) can be used to represent this situation where V_1 represents the set of applicants, V_2 the set of jobs, and an edge vw with $v \in V_1$ and $w \in V_2$ signifies that applicant v is qualified to do a job w. An assignment of applicants to jobs, one person per job, corresponds to a matching problem in G, and the problem of filling as many vacant jobs as possible amount to finding a maximum matching in G.

The M-augmenting path defined earlier plays an important role in the structure of matchings. Consider the following two properties.

Property 3.4.1. *Let M be a matching and Pan augmenting path relative to M then $M\Delta P$ is also a matching of G and $|M\Delta P| = |M| + 1$.*

Property 3.4.2. *Let M and N be matchings in G. Then each connected component of $G[M\Delta P]$ is one of the following:*

(1) An even cycle with edges alternately in $M \setminus N$ and $N \setminus M$, or

(2) A path whose edges are alternately in $M \setminus N$ and $N \setminus M$.

The following theorem, which builds on these observations, due to Berge (1957), points out the relevance of augmenting paths to the study of maximum matchings.

Theorem 3.4.3. *A matching M in a graph G is a maximum matching if and only if G contains no M*-augmenting path.

Proof. Let M be a matching in G. Suppose that G contains an M-augmenting path P. Then, the set $M^* = \big(M \backslash E(P)\big) \cup (E(P) \backslash M) = M \Delta E(P)$ is a matching with larger cardinality than M, i.e., $|M^*| = (|M| + 1) > |M|$. Thus, M is not a maximum matching.

Conversely, suppose that M is not a maximum matching. Therefore, there exists a matching M' with $|M'| > |M|$. Consider the subgraph $H \subseteq G$ with $V(H) = V(G)$ and $E(H) = M \cup M'$. Each vertex of H has degree one or two in H, for it can be incident with most one edge of M and one edge of M'. Consequently, each component of H is either an even cycle with edges alternately in M and M', or else a path with edges alternately in M and M'. Since $|M'| > |M|$, there is a component of H which is a path with more edges in M' than in H. Then P is an M-augmenting path.

Hall (1935) derived the necessary and sufficient conditions for the existence of a matching in a bipartite graph G with bipartition (V_1, V_2), which covers every vertex in V_1.

Theorem 3.4.4. *A bipartite graph $G(V_1, V_2)$ has a matching which covers every vertex in V_1 if and only if $|N(S)| \geq |S|$ for $S \subseteq V_1$.*

Proof. Let $G(V_1, V_2)$ be a bipartite graph which has a matching M covering every vertex in V_1. Consider the subset S of V_1. The vertices in S are matched under M with distinct vertices in $N(S)$. Therefore, $|N(S)| \geq |S|$ and the matching condition holds.

Conversely, suppose that $G(V_1, V_2)$ is a bipartite graph which as a matching covering every vertex in V_1. Let M' be a maximum matching in G and v a vertex in V_1 not covered by M'. Denote by Z the set of all vertices reachable from v by M' $-$alternating paths. Since M' is a maximum matching, it follows from the previous theorem (Berge's) that v is the only vertex in Z not covered by M'. Set $R = V_1 \cap Z$ and $B = V_2 \cap Z$. Clearly, the vertices of $R \backslash \{v\}$ are matched under M' with the vertices of B. Therefore, $|B| = |R| - 1$ and $N(R) \supseteq B$. In fact $N(R) = B$, because every vertex in $N(R)$ is connected to v by an M' $-$alternating path. These two equations imply that $N(R) = |B| = |R| - 1$. Hence, Hall's theorem holds.

3.5 Network flows

Network flow theory is essentially the study of digraphs. The models and algorithms introduced by L.R Ford, Jr., and D.R. Fulkerson in their classic book titled 'Flows in Networks' have set the foundation for the study of network flow problems (Ford & Fulkerson, 1962). Network flow problems arise in many applications of graph theory such as internet traffic, transportation systems, communication systems, road traffic flow, and power supply networks. This section presents an overview of network flow theory. In particular, the maximum flow problem and its relationship to the minimum cut are discussed together with Ford and Fulkerson approach to finding the maximum flow in a network. Most results from this section were quoted from a well written book by Jonathan Gross and Jay Yellen (Gross & Yellen, 2006).

A digraph $N = (V, A)$ is called a network, if V is a set of vertices with two distinguished vertices called the source and the sink, A a set of arcs or directed edges (disjoint from V), and to each $a \in A$ a non-negative real number $cap(a)$ is assigned which is called the capacity of arc a. The out-set of vertex v denoted $Out(v)$, is the set of all arcs that are directed from vertex v and the in-set of vertex v denoted $In(v)$, is the set of all arcs that are directed to vertex v. If $s \in V$ is the source then s is not incident to any element of the in-set, and $t \in V$ is the sink if t is not incident to any element of the out-set. A single source-single sink network with source s and sink t is referred to as a $s - t$ network. For any two vertex subsets X and Y of a digraph N, we denote $\langle X, Y \rangle$ the set of all arcs in N that are directed from a vertex in X to a vertex in Y.

A **(feasible) flow** f in a network N is an assignment to each arc $a \in A$ a non-negative real number $f(a)$ such that:

1. (Capacity constraint) $f(a) \leq cap(a)$;

2. (Conservation constraint) For every vertex $v \in V$, except s and t, the following flow conservation law holds:

$$\sum_{a \in In(v)} f(a) = \sum_{a \in Out(v)} f(a)$$

An arc for which $f(a) = cap(a)$ is called saturated; if $f(a) < cap(a)$, then arc a is called unsaturated. The value of the network flow in a capacitated network, denoted by $val(f)$ is the sum of all the flows leaving the source s; the flow conservation law implies that it equals the sum of the flows arriving at t:

$$val(f) = \sum_{a \in Out(s)} f(a) = \sum_{a \in In(t)} f(a)$$

Figure 3.9 shows a flow for a 6-vertex capacitated $s - t$ network. The adopted convention on the drawing to distinguish capacity from flow when both numbers appear is to place capacity in bold and to the left of the flow. Note that the total amount of flow leaving the sources vertex s equals 15, which equals the net flow entering the sink vertex t.

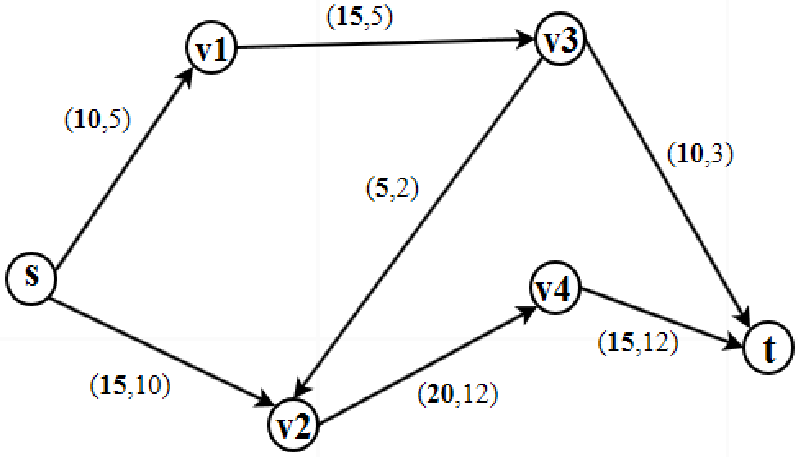


Figure 3.9: Example of a flow for a 6-vertex capacitated $\mathbf{s} - \mathbf{t}$ network.

A flow in a network N is a maximum flow if there is no flow in N of greater value. The problems of finding the maximum flow in a capacitated network N is closely related to the problem of finding the minimum cut in N. Maximum flows are very important in many applications of network flow theory such as traffic flow networks and transportation networks.

Consider the subset of arcs $S \subseteq A$ that partitions the vertices of the network N into two disjoint sets V_s and V_t such that the source $s \subseteq V_s$ and the sink $t \subseteq V_t$. Then the set of all arcs that are directed from a vertex in set V_s to a vertex in set V_t is called a $s - t$ cut of the network N and is denoted by $\langle V_s, V_t \rangle$.

Figure 3.10 shows an example of a $s - t$ cut where $V_s = \{s, v1, v2\}$ and $V_t = \{v3, v4, t\}$. The capacity of a cut is the sum of the capacities of the arcs in the cut. The minimum cut of a network N is defined as the cut with the minimum capacity. Different cuts have different capacities and no flow can exceed the smallest capacity over all cuts in the $s - t$ network.

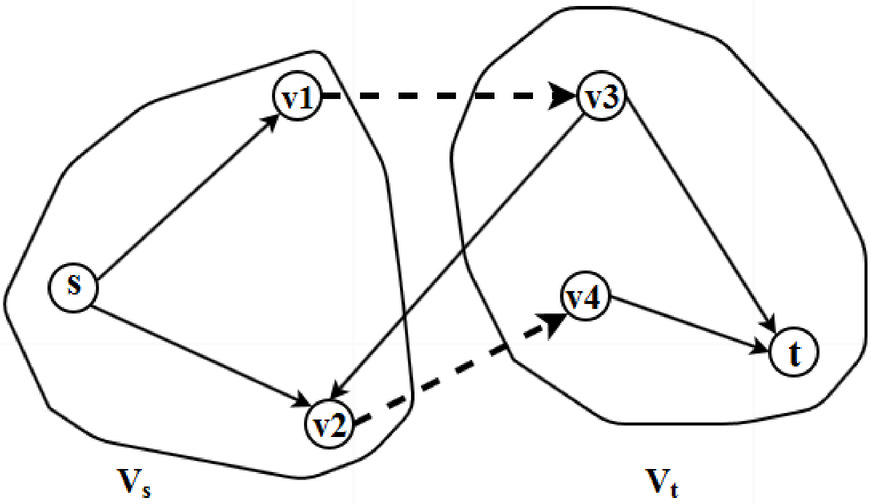


Figure 3.10: Example of a **s** − **t** cut

Proposition 3.5.1. *Let* $\langle V_s, V_t\rangle$ *be a* $s - t$ cut of a network N. Then every directed $s - t$ path in N contains at least one arc in $\langle V_s, V_t\rangle$.

Proof. Let $Y = \langle s = v_0, v_1, v_2, \dots, v_k = t\rangle$ be the vertex sequence of a directed $s - t$ path in network N. Since $s \in V_s$ and $t \in V_t$, there exists a first vertex v_j on this path that is in the set V_t. Then the arc from vertex v_{j-1} to v_j is in $\langle V_s, V_t\rangle$.

The following proposition demonstrates the relationship between flows and cuts.

Proposition 3.5.2. *Let* f *be a flow in an* $s - t$ network N, and let $\langle V_s, V_t\rangle$ be any $s - t$ cut of N. Then

$$val(f) = \sum_{a \in \langle V_s, V_t\rangle} f(a) - \sum_{a \in \langle V_t, V_s\rangle} f(a)$$

The upper bound for the maximum flow problem is provided by the following proposition.

Proposition 3.5.3. *Let* f *be a flow in an* $s - t$ network N, and let $\langle V_s, V_t\rangle$ be any $s - t$ cut of N. Then

$$val(f) \leq cap\langle V_s, V_t\rangle$$

The techniques presented by Ford and Fulkerson (1962) spurred the development of computational tools for analysing network flow problems. The basic idea presented by Ford and Fulkerson is to increase the flow in a network iteratively based on suitably chosen alternating sequence of vertices and arcs until it cannot be increased any further, resulting in the maximum flow in a network. The alternating sequence of vertices and arcs is called the augmenting flow path.

An $s-t$ **quasi-path** in a network N is an alternating sequence $P = \langle s = v_0 a_1 v_1, \ldots, v_{k-1} a_k v_k = t \rangle$ of vertices and arcs that forms and $s-t$ path in the underlying graph of N where arc a_i is called a forward arc if it is directed from vertex v_{i-1} to vertex v_i, and arc a_i is called a backward arc if it is directed from vertex v_i to vertex v_{i-1}. If the flow on each forward arc can be increased and the flow on each backward arc can be decreased the $s-t$ quasi-path is called the f**-augmenting path** P. Thus, if $\triangle_p$ denotes the flow increase, then

$$\triangle_p = \begin{Bmatrix} cap(a) - f(a), if\ a\ is\ a\ forward\ arc \\ f(a), if\ a\ is\ a\ backward\ arc \end{Bmatrix}$$

The quantity $\triangle_p$ is called a **slack on arc** a. Its value on a forward arc is the largest possible increase in the flow, and on a backward arc, the largest possible decrease in the flow, disregarding conservation of flow. The following proposition summarises how the f-augmenting path is used to increase the flow f in a network and Theorem 2.4.5 associates the maximum flow in a network N to the f-augmenting path. The relationship between the maximum flow and minimum cut is summarised in Theorem 2.4.6.

Proposition 3.5.4. *Let* f *be a flow in an* $s-t$ network N, and let P be an f-augmenting path with minimum slack $\triangle_p$ on its arcs. Then the augmented flow $\hat{f}$ given by

$$\hat{f} = \begin{Bmatrix} f(a) + \triangle_p, if\ a\ is\ a\ foward\ arc\ of\ P \\ f(a) - \triangle_p, if\ a\ is\ a\ backward\ arc\ of\ P \\ f(a),\ \ otherwise \end{Bmatrix}$$

is a feasible flow in network N, and $val(\hat{f}) = val(f) + \triangle_p$.

Theorem 3.5.5 [Characterisation of maximum flow]. *Let f be a flow in a network N. Then f is a maximum flow in N if and only if there does not exist an f-augmenting path in N.*

Theorem 3.5.6 [Max-Flow Min-Cut]. *For a given network, the value of a maximum flow is equal to the capacity of a minimum cut.*

The Ford-Fulkerson approach to finding the augmenting path in the augmented flow $\hat{f}$ (residual graph) is not specified. An improvement to the Ford-Fulkerson algorithm was achieved by Edmonds-Karp (Edmonds & Karp, 1972), by specifying the search order when finding the augmenting path. The Edmonds-Karp algorithm uses breadth-first search techniques to traverse the network to find the shortest path that has available capacity.

3.6 Summary

Chapter 3 introduced the fundamentals of mathematical graph theory, beginning with the subject history, the concepts, definitions, representation of graphs, and proofs of significant results relating to bipartite graphs.

The research uses the directed and weighted bipartite graph as a model for cross-border financial flows. The next chapter introduces the symmetric-key encryption algorithm that utilizes the multi-dimensional data set's group structure to compute the edge weights of the proposed directed and weighted bipartite graph model for cross-border financial flows.

Chapter 4: Preservation of information privacy using encryption

4.1 Introduction

Researchers use three broad classes to categorize techniques for preserving the privacy of personally identifiable information in statistical databases, i.e., data obfuscation, summarization, and data separation (Adam & Wortmann, 1989; Cios & Moore, 2002; Clifton & Vaidya, 2004; Kou, Peng, Shi, & Chen, 2007).

Encryption is part of cryptography, viewed as one of the most complex data obfuscation techniques by many. The encryption techniques are widely used by governments to protect their sensitive political, economic, law enforcement, and military information from foreign governments with hostile interests. The U.S. government endorsed the first publicly available cryptographic algorithm called the Data Encryption Standard in the early 1970s (Smid & Branstad, 1988).

The underlying complexities and robustness of encryption techniques are evident in distributed ledger technologies such as Blockchain, which rely on sophisticated cryptographic algorithms to secure the sender of transactions' identity and ensure no tampering of records. Current encryption algorithms in use include the Triple-DES, IDEA, AES, RSA, RC6, Serpent, and Elliptic curve (Stinson & Paterson, 2018).

This chapter proposes a cryptographic approach to privacy-preservation by developing a symmetric-key encryption algorithm to preserve personally identifiable information in multi-dimensional data sets. The proposed symmetric-key encryption algorithm leverages the multi-dimensional data sets' group structure to create temporary variables during a computer program's compilation phase. The algorithm uses the temporary variables to perform the algorithmic operations relating to data processing and computing the cross-border financial flows network's weights. The objective is to avail data comprising personally identifiable information for analysis without revealing the identities of individuals.

The chapter begins with a background on privacy preservation methods, emphasizing the cryptographic approach and other techniques for preserving privacy in statistical databases such as differential privacy and data separation techniques.

The algorithm comprises two distinct parts. The first part performs elementary algebraic computations to derive a non-random permutation, a significant component of the encryption and decryption key. The second part reorganizes the data set and computes the descriptive statistics for deriving the network structure of cross-border financial flows. The chapter concludes with the implementation of the algorithm using the SAS® software program.

4.2 Overview of cryptographic techniques

Symmetric-key encryption algorithms are generally categorized in two categories, i.e., block ciphers and stream ciphers. In block ciphers, the successive plaintext elements are encrypted using the same key, K. That is, the ciphertext string $\boldsymbol{y}$ is obtained as follows:

$$\boldsymbol{y} = y_1 y_2 \ldots = e_K(x_1) e_K(x_2) \ldots \tag{4.1}$$

The basic idea of stream ciphers is to generate a keystream $\boldsymbol{z} = z_1 z_2 \ldots$, and use it to encrypt plaintext string $\boldsymbol{x} = x_1 x_2 \ldots$ according to the rule

$$\boldsymbol{y} = y_1 y_2 \ldots = e_{z_1}(x_1) e_{z_2}(x_2) \ldots \tag{4.2}$$

Block ciphers often operate as critical components in the design of many cryptographic protocols and widely used for encryption of bulk data. In contrast, stream ciphers are simple to implement in hardware and fast in execution, especially in applications where plaintext comes in quantities of unknowable length like in a secure wireless connection.

More formally, a cryptosystem is a five-tuple (P, C, K, E, D), which satisfies the following conditions:

1. P is the finite set of possible plaintexts;
2. C is a finite set of possible ciphertexts;
3. K, the keyspace, is a finite set of possible keys;
4. For each $k \in K,$ there is an encryption rule $e_k \in E$ and a corresponding decryption rule $d_k \in D.$ Each $e_k: P \to C$ and $d_k: C \to P$ are functions such that $d_k\big(e_k(x)\big) = x$ for every plaintext element $x \in P.$

Condition 4 says that if a plaintext is encrypted using an encryption key, and the resulting ciphertext is subsequently decrypted using the decryption key, then the original plaintext results. This condition is the main property of encryption.

Continuing with the analogy described in chapter two, Alice and Bob will choose an encryption key when they are in the same place and are not being observed by Oscar, or alternatively, when they do have access to a secure channel, in which case they can be in different places.

The use of the chosen key will enable Alice and Bob to communicate over an insecure channel as illustrated in Figure 4.1

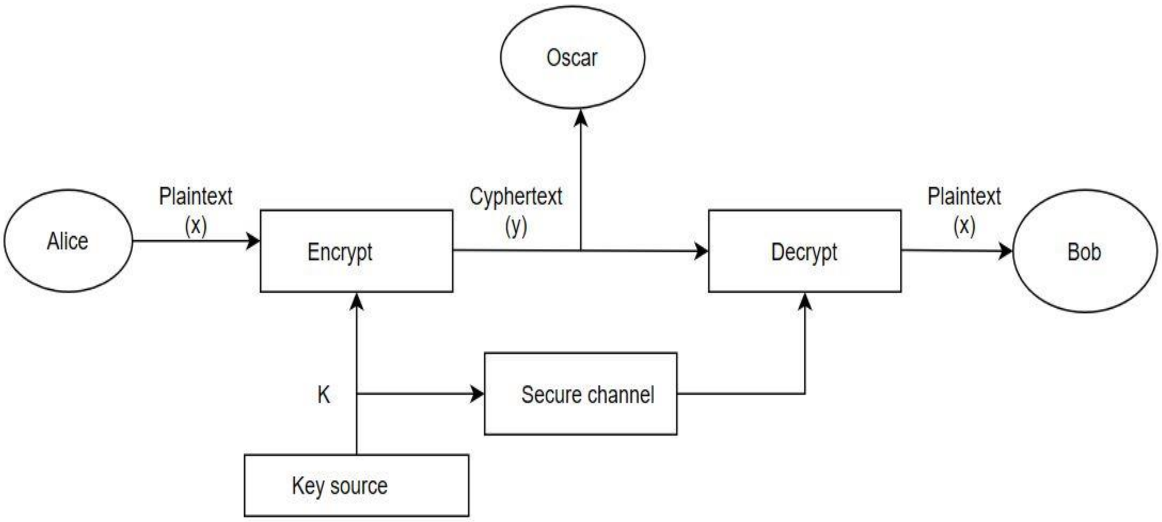


Figure 4.1: The communication channel

The encryption algorithm which utilizes a single key is called a symmetric-key encryption algorithm. In such cryptosystems, the decryption rule d_k and the encryption rule e_k are either the same or e_k can be derived from d_k. For example, DES decryption is identical to encryption, but the key schedule is reversed. Such algorithms execute fast and are simple to implement. However, if the encryption process requires too many steps to execute, then the algorithm may not be suitable for encrypting large amounts of data (Ayushi, 2010).

Some of the well-known single-key encryption algorithms include the Shift Cipher and the Substitution Cipher. To encrypt the plaintext in a Caesar Cipher, one moves each letter a specific number of positions to the left or right. For example, with a shift of 5, E would be replaced by F, G would become B, and so on. The Caesar Cipher is an example of the Shift Cipher, named after Julius Caesar, who used it with a shift of three to protect messages of military significance.

The Substitution Cipher uses permutations of alphabetic characters for encryption and decryption. Hence, the Substitution Cipher includes all the $26!$ permutations, while the Shift Cipher uses the 26 elements only. Its formal definition is as follows:

Let $P = C = \mathbb{Z}_{26}$, where $\mathbb{Z}_{26}$ is the set $\{0, \dots, m-1\}$ equipped with two operations, $+$ and $-$, called the arithmetic modulo n. K consists of all possible permutations of the 26 symbols $0,1, \dots ,25$. For each permutation $\varphi \in K$, define

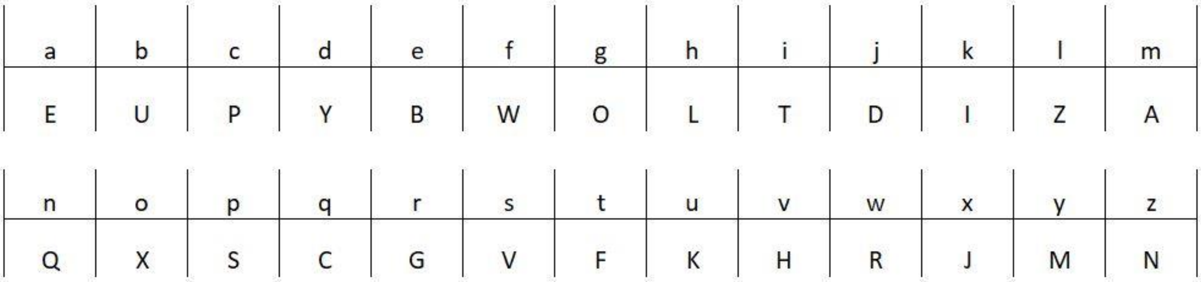
$$e_\varphi(x) = \varphi(x),$$

and define

$$d_\varphi(x) = \varphi^{-1}(x),$$

where φ^{-1} is the inverse permutation to φ.

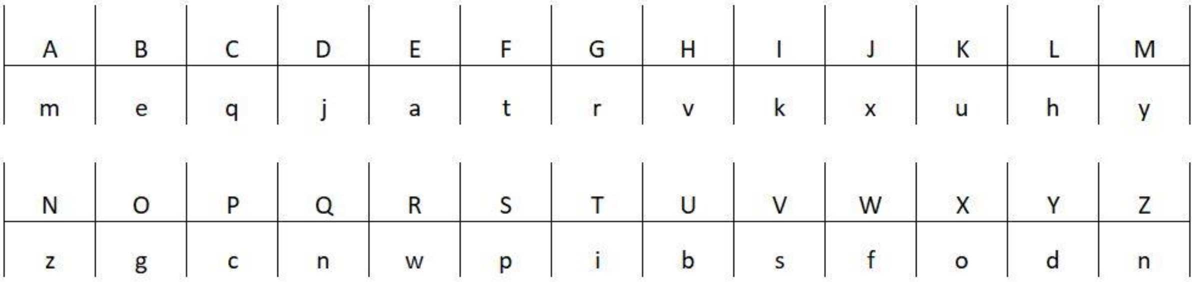
It is more convenient to use alphabetic characters as opposed to residues modulo 26, since the encryption and decryption operations do not require computations. Figure 4.2 shows an example of a random permutation, $\varphi,$ which could comprise an encryption function. Lowercase characters represent plaintext and uppercase letters represent cyphertext.



a	b	c	d	e	f	g	h	i	j	k	l	m
E	U	P	Y	B	W	O	L	T	D	I	Z	A
n	o	p	q	r	s	t	u	v	w	x	y	z
Q	X	S	C	G	V	F	K	H	R	J	M	N

Figure 4.2: A random permutation comprising an encryption function

Thus, $e_\varphi(a) = E$, $e_\varphi(b) = U$, etc. The decryption function is the inverse permutation formed by writing the second line first, and then sorting in alphabetic order to obtain Figure 4.3. The decryption operations are $d_\varphi(A) = m$, $d_\varphi(B) = e$, etc.



A	B	C	D	E	F	G	H	I	J	K	L	M
m	e	q	j	a	t	r	v	k	x	u	h	y
N	O	P	Q	R	S	T	U	V	W	X	Y	Z
z	g	c	n	w	p	i	b	s	f	o	d	n

Figure 4.3: An inverse permutation comprising the decryption function

Another special case of the Substitution Cipher is the Affine Cipher, which uses encryption functions of the form $e(x) = (ax + b)\ mod\ 26,$ where $a, b \in \mathbb{Z}_{26}$. Some simple cryptosystems include the Vigenere Cipher, Hill Cipher, Permutation Cipher, and Autokey Cipher (Stinson & Paterson, 2018).

Asymmetric encryption, which is also called public-key encryption, utilizes two keys. The message encrypted using the public key requires the same algorithm and a matching private key to decrypt. While the public key can be freely shared, the paired private key must remain a secret well kept.

Diffie and Hellman put forward the idea of public-key encryption in 1976, leading to the invention of the well-known RSA algorithm in 1977 (Diffie & Hellman, 1976; Rivest, Shamir, & Adleman, 1978). Public key algorithms are fundamental security ingredients in modern cryptosystems, applications, and protocols assuring the confidentiality, authenticity, and non-reputability of electronic communications and data storage.

The public-key algorithms underpin various Internet standards, such as Transport Layer Security (TLS). Some algorithms provide key distribution and secrecy (e.g., Diffie-Hellman key exchange). In contrast, some provide digital signatures (e.g., Digital Signature Algorithm), and others offer both (e.g., RSA).

Asymmetric encryption tends to be slower in execution than symmetric encryption, mainly due to complex algorithms with a high computational burden. Figure 4.4 is a graphical depiction of asymmetric encryption.

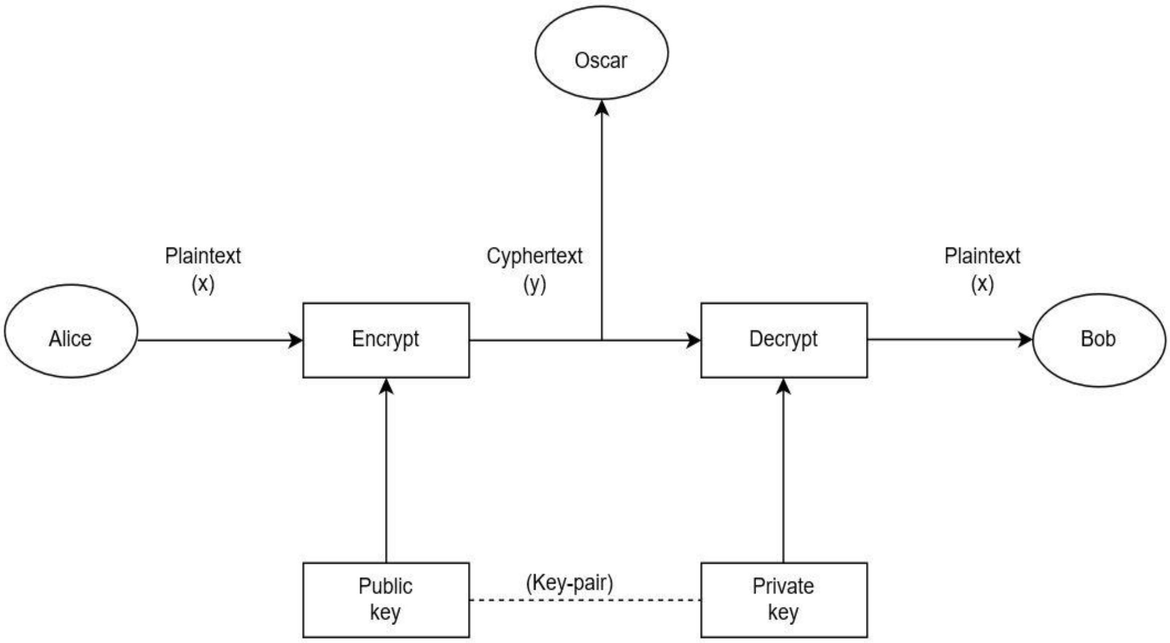


Figure 4.4: Asymmetric key encryption

Classical cryptography was synonymous with encryption. Modern-day cryptography concerns various aspects of information security, such as data confidentiality, data integrity, and user authentication.

Cryptography exists at the intersection of the disciplines of mathematics, computer science, electrical engineering, communication science, and physics. Applications of cryptography include electronic commerce, chip-based payment cards, digital currencies, computer passwords, and military communications. Distributed ledger technologies such as Blockchain rely on sophisticated cryptographic algorithms to secure the identity of the sender of transactions and to ensure the safekeeping of financial records. Triple-DES, IDEA, AES, RSA, RC6, Serpent, and Elliptic curve are some examples of encryption algorithms that are currently in use (Stinson & Paterson, 2018).

4.3 Non-cryptographic approaches to privacy-preservation

Recent advances in privacy-preservation methods include DataSifter and Personalised Differential Privacy (PDP). These methods effectively provided useful information about sensitive data without revealing much about any individual (Ebadi, Sands, & Schneider, 2015; Marino, et al., 2019).

Data separation-based methods such as vertical partition and horizontal partition are also robust against side-knowledge, allowing them to perform better than the widely used summarization techniques (Vaidya & Clifton, 2004; Kou, Peng, Shi, & Chen, 2007).

Researchers have also integrated cryptography and machine learning into the three broad categories of privacy-preserving methods (Pathak, Rane, & Raj, 2010; Pathak M. , Rane, Sun, & Raj, 2011; Wang, Wang, Bi, & Xu, 2018). Multiple parties can use deep learning based on artificial neural networks to model, classify, and recognize complex data such as images, speech, and text without sharing their input datasets (Shokri & Shmatikov, 2015). The next section provides an overview of differential privacy.

4.3.1 Differential privacy

Privacy issues are usually associated with failure to control access (authentication) to information, to control the flow of information, or the purposes for using the information.

Differential privacy arose in a context in which ensuring privacy is a challenge even if all these control problems are solved; thus, differential privacy is a definition of privacy tailored to the problem of privacy-preserving data analysis (Dwork, 2011).

In a study to examine the trade-off between the privacy and usability of statistical databases, researchers found that a small amount of disturbance suffices to result in a substantial violation of privacy (Dinur & Nissim, 2003). This insight led to the development of differential privacy in 2006 (Dwork, McSherry, Nissim, & Smith, 2006).

Differential privacy uses mathematical tools such as Laplace and Gaussian mechanisms, among others, to randomize the individual's responses, which effectively hides the presence or absence of the individual's data throughout the lifetime of the database. Hence, it separates the utility of the database from the increased risk of harm for an individual due to joining the database (Dwork, McSherry, Nissim, & Smith, 2006).

Differential privacy ensures that the outcome of any analysis is equally likely, independent of whether an individual takes part or refrains.

Formally, a randomized function $\mathbb{Q}$ gives $\varepsilon - differential\ privacy$ if for all data sets D and D' differing on at most one row, and all $R \subseteq Range(\mathbb{Q})$,

$$\frac{Pr[\ \mathbb{Q}(D) \in R]}{Pr[\mathbb{Q}(D') \in R]} \leq e^{\varepsilon} \tag{4.3}$$

where the probability space in each case is over the coin flips of $\mathbb{Q}$, and ε is small, say 0.01, 0.1, or in some cases, ln 2 or ln 3.

Research and applications of differentially private algorithms continues to grow, with large corporations such as Apple Inc. and Google LLC using differential privacy (Erlingsson, Pihur, & Korolova, 2014; Tang, Korolova, Bai, Wang, & Wang, 2017). Founded in 2014, Privitar is a company specializing in the development and adoption of privacy engineering technology with a global client-base across North America, Europe, and Asia.

4.4 Single-key encryption algorithm using temporary variables

The proposed symmetric-key encryption algorithm uses elementary arithmetic operations to derive a permutation, which is independent of the plaintext. The purpose of the permutation is to enable encryption and decryption. Hence, a critical component of the symmetric-key.

The algebraic operations begin by sorting the database records using the target encryption variable and utilizing the multi-dimensional data sets' group structure to determine the permutation. The procedure uses temporary variables to determine the start and end of record groups while computing the required descriptive statistics for constructing the directed and weighted bipartite network structure. Several software environments for statistical computing such as SAS® and R programming language, among others, provide packages for By-Group processing. This research uses the contents of the PDV variables in SAS® software to create new variables, which comprise descriptive statistics to construct the bipartite network structure.

The idea of the proposed symmetric-key encryption algorithm is first to encrypt the plaintext using the non-random permutation, which creates the cyphertext. Secondly, to store the unique relationship between the plaintext and the cyphertext in a secure environment for decryption purposes. At the same time, compute the descriptive statistics to enable network representation.

A permutation of a finite set S is a bijective function $\varphi: X \rightarrow X$, meaning that the function φ is both one-to-one (injective) and onto (surjective). It follows that for every $x \in X$, there is a unique element $x' \in X$ such that $\varphi(x') = x$. The inverse permutation is a map $\varphi^{-\mathbf{1}}: X \rightarrow X$ with the rule $\varphi^{-\mathbf{1}}(x) = x'$ if and only if $\varphi(x') = x$, thus, $\varphi^{-\mathbf{1}}$ is also a permutation of X.

The derived permutation enables the decryption technique that is analogous to the Permutation Cipher, which is a lookup table. Hence, it is not computationally intense. The approach allows the algorithm to execute fast due to its simplicity, thereby making it possible to encrypt large data sets.

Figure 4.5 shows the graphical illustration of the process underpinning the single-key encryption algorithm. During the compilation time of a computer program, the automatic variables are created, converting high-level language into machine language. The algorithm does not store the temporary variables in computer memory, but only make the variables available during execution.

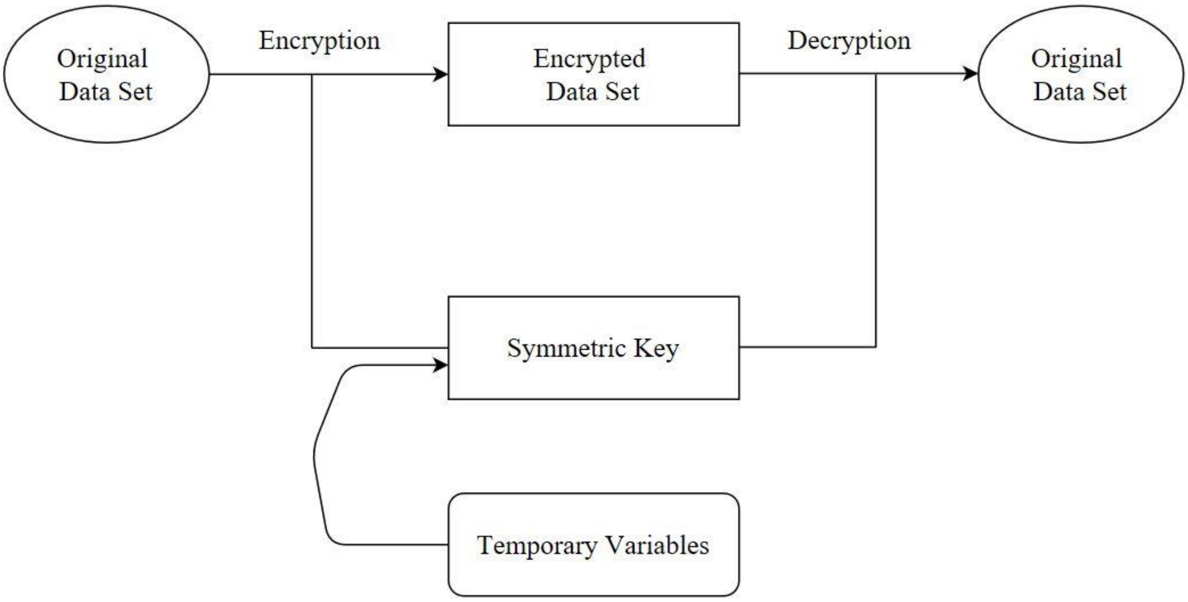


Figure 4.5: Symmetric key encryption using the temporary variables

4.4.1 Description of variables

(1) N; a counter variable which records the record number being processed in the dataset. Its initial value is set to 1 and is incremented by one each time a new record is processed.

(2) Encryption variables; are the names of variables comprising personally identifiable information.

(3) BY-variables; are the names of the encryption variables by which the dataset is sorted or indexed.

(4) BY-values; are the values of the BY-variables.

(5) 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.

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

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

4.4.2 Procedure – part one

(8) Input the original dataset with multiple records per subject.

(9) Sort the dataset by the encryption variables to enable the creation of BY-groups.

(10) FIRST.BY-variable and N are automatically set to true at the start of dataset processing.

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

(12) Concatenate a chosen prefix for the encryption variable with N to obtain the encrypted variable,

(13) Retain the encrypted variable and initialize all other dataset variables.

(14) N automatically increments by one.

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

(16) Return to step (4).

(17) Stop after processing the last record of the dataset to complete the encryption of the first variable.

(18) Repeat the algorithm from step 1 through step 10 until all the encryption variables have been encrypted.

(19) Drop the encryption variables from the data set to remain with the encrypted dataset.

4.4.3 Procedure – part two

(1) Input the dataset from part one.

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

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

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

(5) If FIRST.BY-variable is true, initialize all the variables comprising summary statistics to zero, otherwise increment the variables comprising the summary statistics.

(6) If LAST.BY-variable is true, output the current record.

(7) N automatically increments by one.

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

(9) Return to step (4).

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

4.5 Implementation of the algorithm a using SAS® programming language

This section provides an overview of the components of the SAS® programming language, followed by the encryption of the illustration of the hypothetical data to illustrate the workings of the proposed symmetric-key encryption algorithm. The section concludes with a presentation of the research results after applying the proposed symmetric-key encryption algorithm to encrypt the cross-border financial flows data set.

Table A.4 (see Appendix A) shows the hypothetical data set comprising cross-border financial transactions, for illustrating the workings of the proposed symmetric-key encryption. The resident name and the non-resident name are the variables consisting of private and confidential information. Each financial transaction represents a record of cash received/paid by a country's resident individual/firm from/to a non-resident individual/firm residing in another country. The goal is to replace personally identifiable information (resident name and non-resident name) with labels before analyzing the data set.

4.5.1 Overview of the SAS® program

The SAS® programming language builds a data set one observation at a time using the Program Data Vector (PDV), which is a logical area in computer memory. SAS® reads the record into the PDV and then write it to a target data set during the program execution phase. The PDV contains both permanent and temporary variables.

The SAS® program structure encompasses the DATA step processing, along with BY-group processing. The BY statement used along with the SET statement in a DATA step instructs SAS® to create the automatic variables FIRST.BY-variable and LAST.BY-variable. Each completed iteration of the DATA step increments the automatic variable _N_ by one (SAS Institute Inc., 2010).

Figure 4.6 shows an overview of the DATA step processing in SAS®, starting with the compilation phase, which converts high-level language into machine language (SAS Institute Inc., 2001).

The execution phase in Figure 4.6 shows the state of the PDV variables and the automatic variables as SAS® processes observations iteratively during the execution of the program. The program carries out the algebraic operations of the proposed symmetric-key encryption algorithm during the program execution phase.

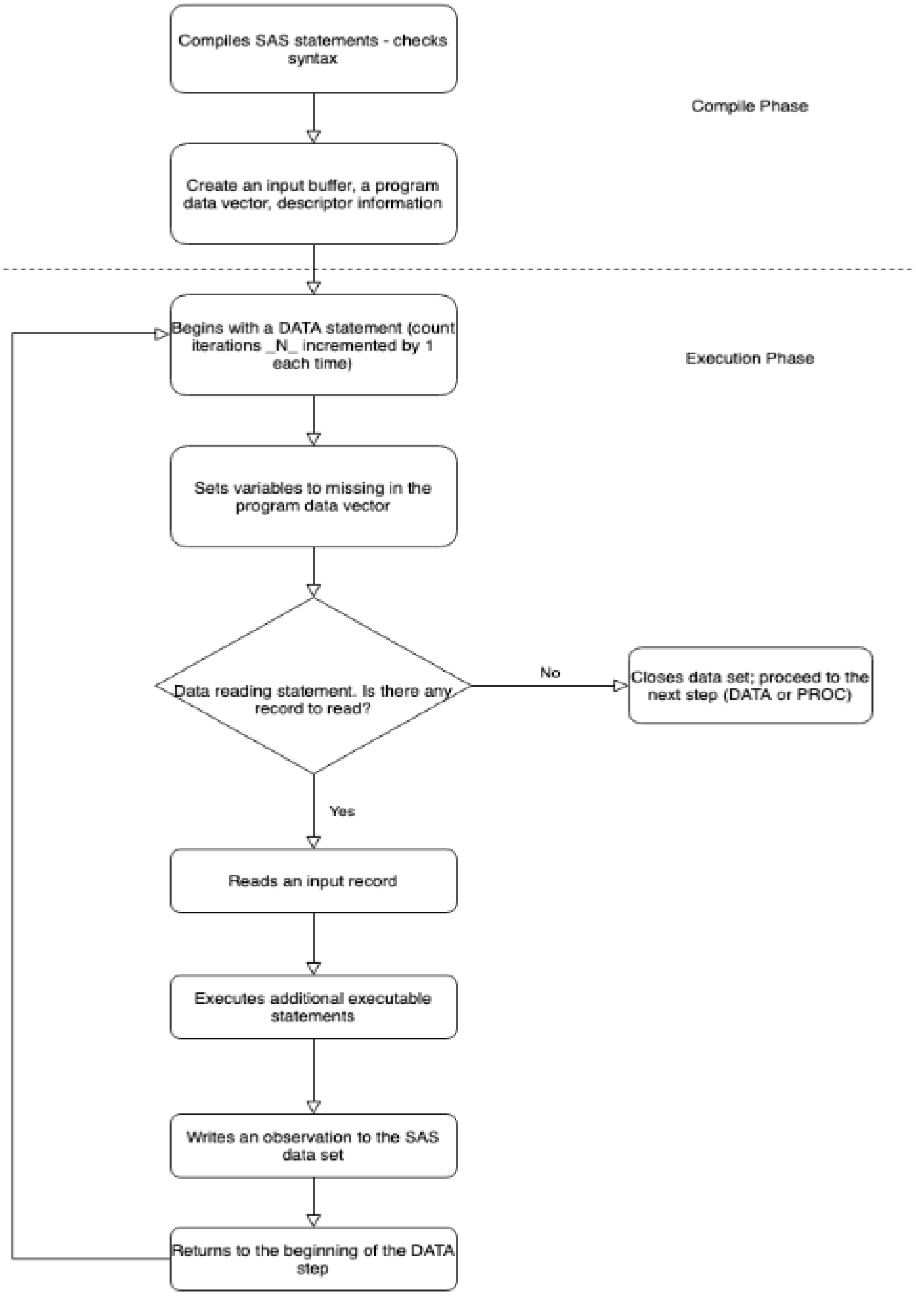

Compiles SAS statements - checks syntax
Create an input buffer, a program data vector, descriptor information
Compile Phase
Begins with a DATA statement (count iterations _N_ incremented by 1 each time)
Execution Phase
Sets variables to missing in the program data vector
Data reading statement. Is there any record to read?
No
Closes data set; proceed to the next step (DATA or PROC)
Yes
Reads an input record
Executes additional executable statements
Writes an observation to the SAS data set
Returns to the beginning of the DATA step

Figure 4.6: DATA step processing from the compile phase to execution phase.

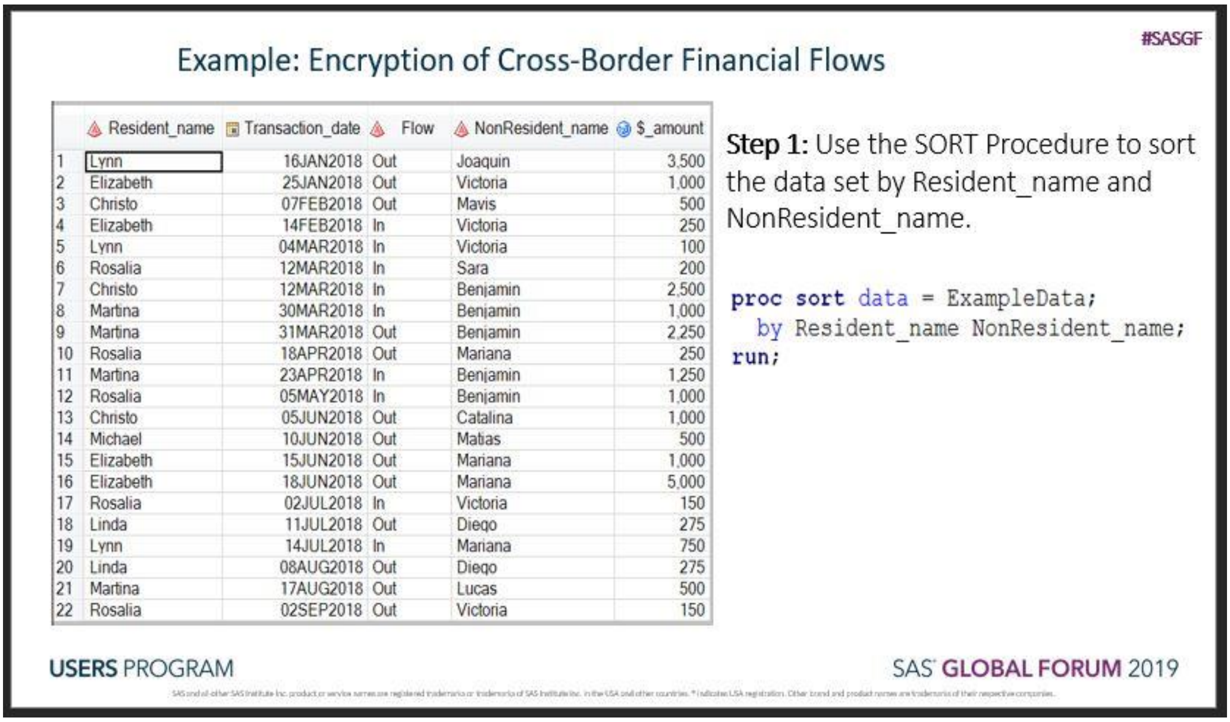
4.5.2 Implementation using SAS® software – hypothetical data set

The SAS® program in Appendix B1 uses the DATA step processing, along with the BY-group processing, to obfuscate two variables comprising personally identifiable information, i.e., “resident name” and “non-resident name,” in the hypothetical data set shown in Table A.4 (Appendix A). The program was generated using SAS® Enterprise Guide 7.1. Copyright © 2014, SAS Institute Inc., Cary, NC, USA.

The program uses the PDV variables and the RETAIN statement, which is a compile-time-only statement to retain the values of the encrypted variables in the PDV across iterations of the DATA step. Hence, the RETAIN statement avoids the reinitialization of the PDV variables.

The program executes basic algebraic operations consisting of two parts, for each of the variables comprising personally identifiable information to generate the encryption key. The execution of both parts of the program completes the first part of the proposed symmetric-key encryption algorithm. The two parts comprising the algebraic operations are as follows:

1. Sort the observations in ascending order by the variable comprising personally identifiable information. Refer to Figure 4.7 for an illustration of the first part of the SAS® program.
2. Identify the first and the last occurrence of the variable in the data set using the PDV variables and create a new corresponding variable, which becomes a member of the non-random permutation set.



#SASGF

Example: Encryption of Cross-Border Financial Flows

	Resident_name	Transaction_date	Flow	NonResident_name	$_amount
1	Lynn	16JAN2018	Out	Joaquin	3,500
2	Elizabeth	25JAN2018	Out	Victoria	1,000
3	Christo	07FEB2018	Out	Mavis	500
4	Elizabeth	14FEB2018	In	Victoria	250
5	Lynn	04MAR2018	In	Victoria	100
6	Rosalia	12MAR2018	In	Sara	200
7	Christo	12MAR2018	In	Benjamin	2,500
8	Martina	30MAR2018	In	Benjamin	1,000
9	Martina	31MAR2018	Out	Benjamin	2,250
10	Rosalia	18APR2018	Out	Mariana	250
11	Martina	23APR2018	In	Benjamin	1,250
12	Rosalia	05MAY2018	In	Benjamin	1,000
13	Christo	05JUN2018	Out	Catalina	1,000
14	Michael	10JUN2018	Out	Matias	500
15	Elizabeth	15JUN2018	Out	Mariana	1,000
16	Elizabeth	18JUN2018	Out	Mariana	5,000
17	Rosalia	02JUL2018	In	Victoria	150
18	Linda	11JUL2018	Out	Diego	275
19	Lynn	14JUL2018	In	Mariana	750
20	Linda	08AUG2018	Out	Diego	275
21	Martina	17AUG2018	Out	Lucas	500
22	Rosalia	02SEP2018	Out	Victoria	150

Step 1: Use the SORT Procedure to sort the data set by Resident_name and NonResident_name.

```
proc sort data = ExampleData;
  by Resident_name NonResident_name;
run;
```

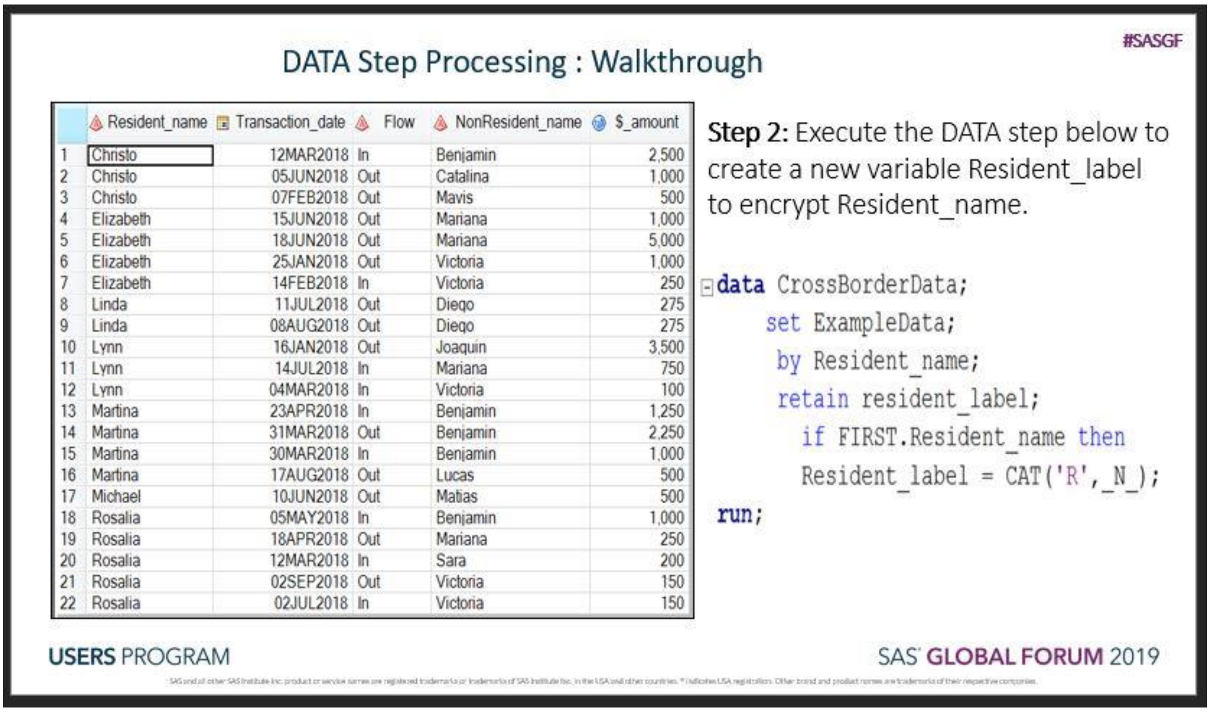
USERS PROGRAM

SAS GLOBAL FORUM 2019

Figure 4.7: Illustration of part one of the encryption program.

The second part of the program uses the PDV variables to compute the aggregated number of transactions and their aggregated financial value. The two aggregates are the weights of the directed and weighted bipartite graph.

Depictions from Figure 4.8 to Figure 4.14 illustrate the workings of the second part of the program. The figures show the state of the PDV variables during the next iterations of the DATA step. The computer code in Appendix B.2 was used to generate the encryption key. The resulting encryption key is shown in Table A.5 (see Appendix A) and should be safely stored. A DROP Statement within the SAS® code was used to drop all the variables comprising personally identifiable information from the encryption key to remain with the encrypted variables.



#SASGF

DATA Step Processing : Walkthrough

	Resident_name	Transaction_date	Flow	NonResident_name	$_amount
1	Christo	12MAR2018	In	Benjamin	2,500
2	Christo	05JUN2018	Out	Catalina	1,000
3	Christo	07FEB2018	Out	Mavis	500
4	Elizabeth	15JUN2018	Out	Mariana	1,000
5	Elizabeth	18JUN2018	Out	Mariana	5,000
6	Elizabeth	25JAN2018	Out	Victoria	1,000
7	Elizabeth	14FEB2018	In	Victoria	250
8	Linda	11JUL2018	Out	Diego	275
9	Linda	08AUG2018	Out	Diego	275
10	Lynn	16JAN2018	Out	Joaquin	3,500
11	Lynn	14JUL2018	In	Mariana	750
12	Lynn	04MAR2018	In	Victoria	100
13	Martina	23APR2018	In	Benjamin	1,250
14	Martina	31MAR2018	Out	Benjamin	2,250
15	Martina	30MAR2018	In	Benjamin	1,000
16	Martina	17AUG2018	Out	Lucas	500
17	Michael	10JUN2018	Out	Matias	500
18	Rosalia	05MAY2018	In	Benjamin	1,000
19	Rosalia	18APR2018	Out	Mariana	250
20	Rosalia	12MAR2018	In	Sara	200
21	Rosalia	02SEP2018	Out	Victoria	150
22	Rosalia	02JUL2018	In	Victoria	150

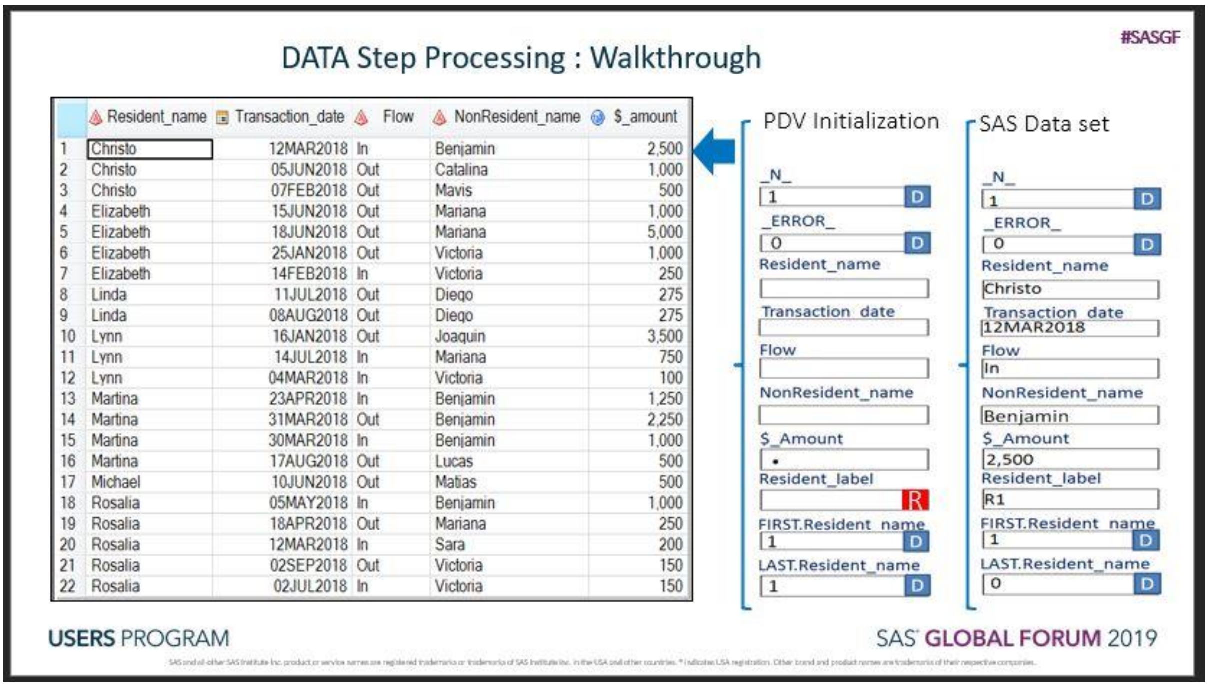
Step 2: Execute the DATA step below to create a new variable Resident_label to encrypt Resident_name.

```
data CrossBorderData;
    set ExampleData;
     by Resident_name;
     retain resident_label;
       if FIRST.Resident_name then
       Resident_label = CAT('R',_N_);
run;
```

USERS PROGRAM

SAS GLOBAL FORUM 2019

Figure 4.8: Illustration of the second part one of the encryption program.



#SASGF

DATA Step Processing : Walkthrough

	Resident_name	Transaction_date	Flow	NonResident_name	$_amount
1	Christo	12MAR2018	In	Benjamin	2,500
2	Christo	05JUN2018	Out	Catalina	1,000
3	Christo	07FEB2018	Out	Mavis	500
4	Elizabeth	15JUN2018	Out	Mariana	1,000
5	Elizabeth	18JUN2018	Out	Mariana	5,000
6	Elizabeth	25JAN2018	Out	Victoria	1,000
7	Elizabeth	14FEB2018	In	Victoria	250
8	Linda	11JUL2018	Out	Diego	275
9	Linda	08AUG2018	Out	Diego	275
10	Lynn	16JAN2018	Out	Joaquin	3,500
11	Lynn	14JUL2018	In	Mariana	750
12	Lynn	04MAR2018	In	Victoria	100
13	Martina	23APR2018	In	Benjamin	1,250
14	Martina	31MAR2018	Out	Benjamin	2,250
15	Martina	30MAR2018	In	Benjamin	1,000
16	Martina	17AUG2018	Out	Lucas	500
17	Michael	10JUN2018	Out	Matias	500
18	Rosalia	05MAY2018	In	Benjamin	1,000
19	Rosalia	18APR2018	Out	Mariana	250
20	Rosalia	12MAR2018	In	Sara	200
21	Rosalia	02SEP2018	Out	Victoria	150
22	Rosalia	02JUL2018	In	Victoria	150

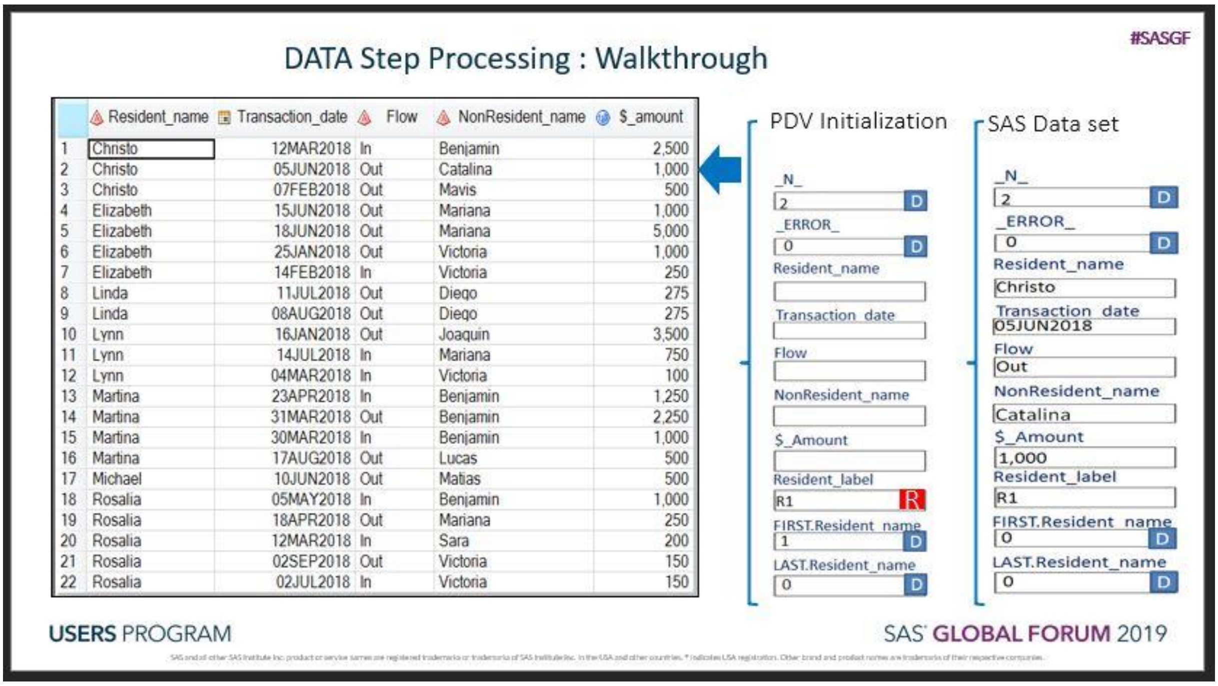
	PDV Initialization	SAS Data set
_N_	1 D	1 D
ERROR	0 D	0 D
Resident_name		Christo
Transaction_date		12MAR2018
Flow		In
NonResident_name		Benjamin
$_Amount	•	2,500
Resident_label	R	R1
FIRST.Resident_name	1 D	1 D
LAST.Resident_name	1 D	0 D

USERS PROGRAM

SAS GLOBAL FORUM 2019

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Figure 4.9: State of the PDV variables during program execution of the first observation.



#SASGF

DATA Step Processing : Walkthrough

	Resident_name	Transaction_date	Flow	NonResident_name	$_amount
1	Christo	12MAR2018	In	Benjamin	2,500
2	Christo	05JUN2018	Out	Catalina	1,000
3	Christo	07FEB2018	Out	Mavis	500
4	Elizabeth	15JUN2018	Out	Mariana	1,000
5	Elizabeth	18JUN2018	Out	Mariana	5,000
6	Elizabeth	25JAN2018	Out	Victoria	1,000
7	Elizabeth	14FEB2018	In	Victoria	250
8	Linda	11JUL2018	Out	Diego	275
9	Linda	08AUG2018	Out	Diego	275
10	Lynn	16JAN2018	Out	Joaquin	3,500
11	Lynn	14JUL2018	In	Mariana	750
12	Lynn	04MAR2018	In	Victoria	100
13	Martina	23APR2018	In	Benjamin	1,250
14	Martina	31MAR2018	Out	Benjamin	2,250
15	Martina	30MAR2018	In	Benjamin	1,000
16	Martina	17AUG2018	Out	Lucas	500
17	Michael	10JUN2018	Out	Matias	500
18	Rosalia	05MAY2018	In	Benjamin	1,000
19	Rosalia	18APR2018	Out	Mariana	250
20	Rosalia	12MAR2018	In	Sara	200
21	Rosalia	02SEP2018	Out	Victoria	150
22	Rosalia	02JUL2018	In	Victoria	150

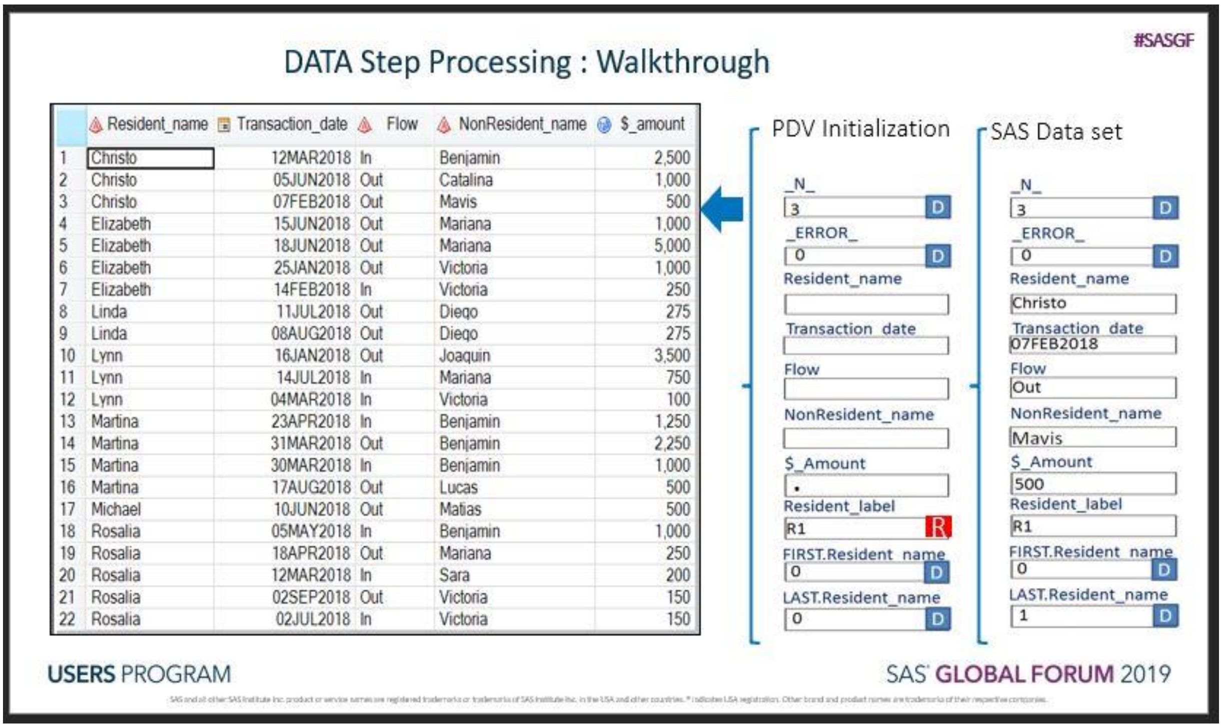
	PDV Initialization	SAS Data set
_N_	2 D	2 D
ERROR	0 D	0 D
Resident_name		Christo
Transaction_date		05JUN2018
Flow		Out
NonResident_name		Catalina
$_Amount		1,000
Resident_label	R1 R	R1
FIRST.Resident_name	1 D	0 D
LAST.Resident_name	0 D	0 D

USERS PROGRAM

SAS GLOBAL FORUM 2019

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Figure 4.10: State of the PDV variables during program execution of the second observation.



#SASGF

DATA Step Processing : Walkthrough

	Resident_name	Transaction_date	Flow	NonResident_name	$_amount
1	Christo	12MAR2018	In	Benjamin	2,500
2	Christo	05JUN2018	Out	Catalina	1,000
3	Christo	07FEB2018	Out	Mavis	500
4	Elizabeth	15JUN2018	Out	Mariana	1,000
5	Elizabeth	18JUN2018	Out	Mariana	5,000
6	Elizabeth	25JAN2018	Out	Victoria	1,000
7	Elizabeth	14FEB2018	In	Victoria	250
8	Linda	11JUL2018	Out	Diego	275
9	Linda	08AUG2018	Out	Diego	275
10	Lynn	16JAN2018	Out	Joaquin	3,500
11	Lynn	14JUL2018	In	Mariana	750
12	Lynn	04MAR2018	In	Victoria	100
13	Martina	23APR2018	In	Benjamin	1,250
14	Martina	31MAR2018	Out	Benjamin	2,250
15	Martina	30MAR2018	In	Benjamin	1,000
16	Martina	17AUG2018	Out	Lucas	500
17	Michael	10JUN2018	Out	Matias	500
18	Rosalia	05MAY2018	In	Benjamin	1,000
19	Rosalia	18APR2018	Out	Mariana	250
20	Rosalia	12MAR2018	In	Sara	200
21	Rosalia	02SEP2018	Out	Victoria	150
22	Rosalia	02JUL2018	In	Victoria	150

PDV Initialization

_N_ 3 D
ERROR 0 D
Resident_name
Transaction_date
Flow
NonResident_name
$_Amount .
Resident_label R1 R
FIRST.Resident_name 0 D
LAST.Resident_name 0 D

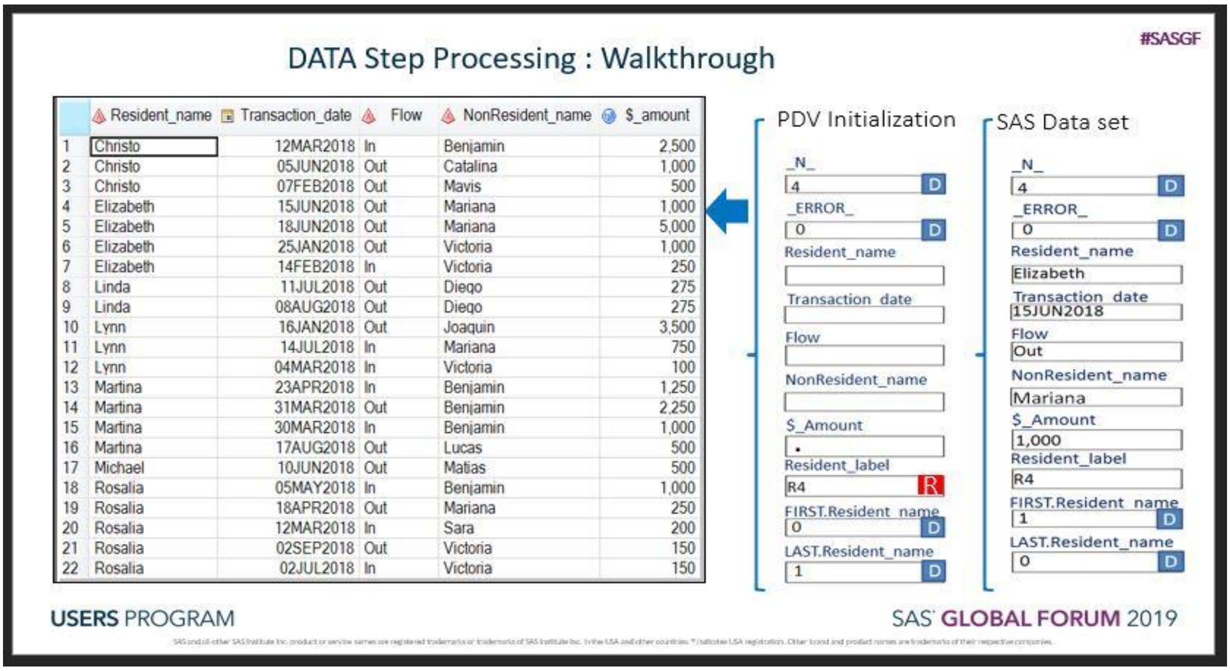
SAS Data set

_N_ 3 D
ERROR 0 D
Resident_name Christo
Transaction_date 07FEB2018
Flow Out
NonResident_name Mavis
$_Amount 500
Resident_label R1
FIRST.Resident_name 0 D
LAST.Resident_name 1 D

USERS PROGRAM

SAS GLOBAL FORUM 2019

Figure 4.11: State of the PDV variables during program execution of the third observation.



#SASGF

DATA Step Processing : Walkthrough

	Resident_name	Transaction_date	Flow	NonResident_name	$_amount
1	Christo	12MAR2018	In	Benjamin	2,500
2	Christo	05JUN2018	Out	Catalina	1,000
3	Christo	07FEB2018	Out	Mavis	500
4	Elizabeth	15JUN2018	Out	Mariana	1,000
5	Elizabeth	18JUN2018	Out	Mariana	5,000
6	Elizabeth	25JAN2018	Out	Victoria	1,000
7	Elizabeth	14FEB2018	In	Victoria	250
8	Linda	11JUL2018	Out	Diego	275
9	Linda	08AUG2018	Out	Diego	275
10	Lynn	16JAN2018	Out	Joaquin	3,500
11	Lynn	14JUL2018	In	Mariana	750
12	Lynn	04MAR2018	In	Victoria	100
13	Martina	23APR2018	In	Benjamin	1,250
14	Martina	31MAR2018	Out	Benjamin	2,250
15	Martina	30MAR2018	In	Benjamin	1,000
16	Martina	17AUG2018	Out	Lucas	500
17	Michael	10JUN2018	Out	Matias	500
18	Rosalia	05MAY2018	In	Benjamin	1,000
19	Rosalia	18APR2018	Out	Mariana	250
20	Rosalia	12MAR2018	In	Sara	200
21	Rosalia	02SEP2018	Out	Victoria	150
22	Rosalia	02JUL2018	In	Victoria	150

PDV Initialization

_N_ 4 D
ERROR 0 D
Resident_name
Transaction_date
Flow
NonResident_name
$_Amount .
Resident_label R4 R
FIRST.Resident_name 0 D
LAST.Resident_name 1 D

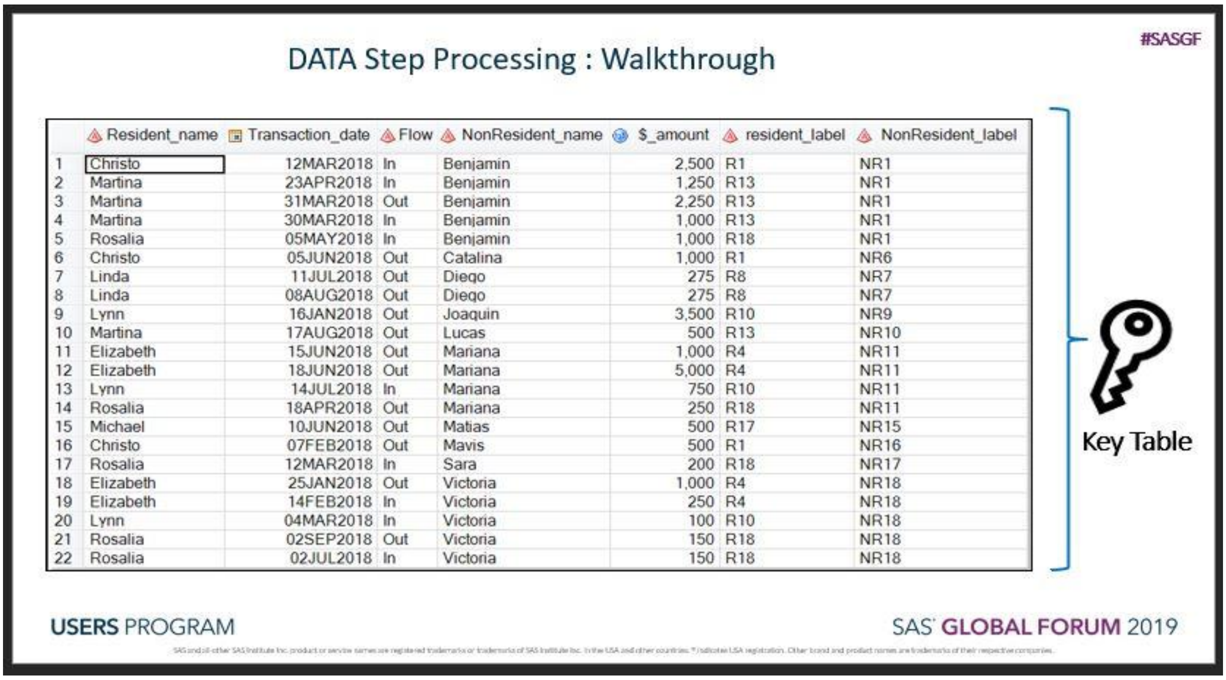
SAS Data set

_N_ 4 D
ERROR 0 D
Resident_name Elizabeth
Transaction_date 15JUN2018
Flow Out
NonResident_name Mariana
$_Amount 1,000
Resident_label R4
FIRST.Resident_name 1 D
LAST.Resident_name 0 D

USERS PROGRAM

SAS GLOBAL FORUM 2019

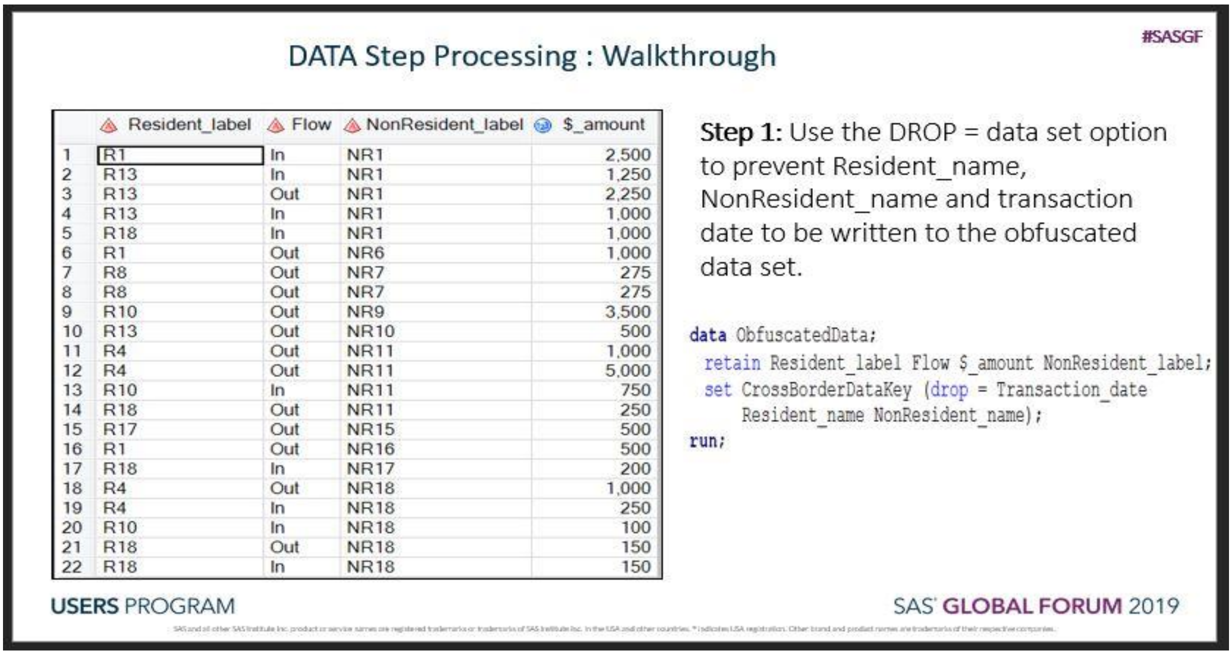
Figure 4.12: State of the PDV variables during program execution of the fourth observation.



DATA Step Processing : Walkthrough

	Resident_name	Transaction_date	Flow	NonResident_name	$_amount	resident_label	NonResident_label
1	Christo	12MAR2018	In	Benjamin	2,500	R1	NR1
2	Martina	23APR2018	In	Benjamin	1,250	R13	NR1
3	Martina	31MAR2018	Out	Benjamin	2,250	R13	NR1
4	Martina	30MAR2018	In	Benjamin	1,000	R13	NR1
5	Rosalia	05MAY2018	In	Benjamin	1,000	R18	NR1
6	Christo	05JUN2018	Out	Catalina	1,000	R1	NR6
7	Linda	11JUL2018	Out	Diego	275	R8	NR7
8	Linda	08AUG2018	Out	Diego	275	R8	NR7
9	Lynn	16JAN2018	Out	Joaquin	3,500	R10	NR9
10	Martina	17AUG2018	Out	Lucas	500	R13	NR10
11	Elizabeth	15JUN2018	Out	Mariana	1,000	R4	NR11
12	Elizabeth	18JUN2018	Out	Mariana	5,000	R4	NR11
13	Lynn	14JUL2018	In	Mariana	750	R10	NR11
14	Rosalia	18APR2018	Out	Mariana	250	R18	NR11
15	Michael	10JUN2018	Out	Matias	500	R17	NR15
16	Christo	07FEB2018	Out	Mavis	500	R1	NR16
17	Rosalia	12MAR2018	In	Sara	200	R18	NR17
18	Elizabeth	25JAN2018	Out	Victoria	1,000	R4	NR18
19	Elizabeth	14FEB2018	In	Victoria	250	R4	NR18
20	Lynn	04MAR2018	In	Victoria	100	R10	NR18
21	Rosalia	02SEP2018	Out	Victoria	150	R18	NR18
22	Rosalia	02JUL2018	In	Victoria	150	R18	NR18

Figure 4.13: Encryption key depicting the permutation derived from executing the algorithm.



DATA Step Processing : Walkthrough

	Resident_label	Flow	NonResident_label	$_amount
1	R1	In	NR1	2,500
2	R13	In	NR1	1,250
3	R13	Out	NR1	2,250
4	R13	In	NR1	1,000
5	R18	In	NR1	1,000
6	R1	Out	NR6	1,000
7	R8	Out	NR7	275
8	R8	Out	NR7	275
9	R10	Out	NR9	3,500
10	R13	Out	NR10	500
11	R4	Out	NR11	1,000
12	R4	Out	NR11	5,000
13	R10	In	NR11	750
14	R18	Out	NR11	250
15	R17	Out	NR15	500
16	R1	Out	NR16	500
17	R18	In	NR17	200
18	R4	Out	NR18	1,000
19	R4	In	NR18	250
20	R10	In	NR18	100
21	R18	Out	NR18	150
22	R18	In	NR18	150

Step 1: Use the DROP = data set option to prevent Resident_name, NonResident_name and transaction date to be written to the obfuscated data set.

```
data ObfuscatedData;
 retain Resident_label Flow $_amount NonResident_label;
 set CrossBorderDataKey (drop = Transaction_date
     Resident_name NonResident_name);
run;
```

Figure 4.14: The encrypted hypothetical data set comprising cross-border financial flows.

4.5.3 Encryption of the cross-border financial flows data set

Table A.2 (Appendix A) shows the encrypted sample of the remittances data set using the proposed symmetric-key encryption algorithm. To interpret the data set, consider the first observation in the data set, i.e., r10. The encryption algorithm allocated the label r10 to a resident whose name is private and confidential information. The resident labeled r10 entered into a single financial transaction with a marked non-resident nr639259, whereby r10 received an amount of USD 3,942 from nr639259 in a single business transaction.

The lookup-table comprising both the resident name and the resident label enables decryption, while the encryption algorithm along with the data set, enables encryption. The algorithm dropped the resident name, and the non-resident name to arrive at the encrypted data set.

Most importantly, encryption occurs at a granular level to enable the summarization of the data set. Hence, the algorithm avails all the data for analysis while preserving the privacy and confidentiality of information. For example, consider observation number 6 in Table A.2 (Appendix A), i.e., r100129. The sample data set shows that r100129 received a total number of 13 payments from two non-residents.

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

4.6.1 Advantages

The advantages of the proposed symmetric-key algorithm are as follows:

(1) The algorithm executes fast due to its simplicity, thereby making it possible to encrypt large data sets.

(2) The decryption operation uses a similar technique to the Permutation Cipher, which is a lookup table. Hence, it is not computationally intense.

(3) Massive data sets with many encryption variables complicate the symmetric-key derivation, thereby making the algorithm safer.

(4) The algorithm does not provide statistical summaries of demographic nature, reducing the chances of making it susceptible to linkage attacks. A linkage attack occurs when adversaries collect supplemental information about an individual from multiple data sources and then combine that data to form a whole picture about their target, which is often an individual's personally identifiable information

4.6.2 Disadvantages

(1) The algorithm's safety depends on the security of the channel used to exchange the decryption key. However, it is technically impossible to stop a person who is duly authorized to access confidential information from improperly disclosing that information to someone else.

(2) The algorithm is not suitable for encrypting and decrypting a live database due to the symmetric-key storage requirement.

(3) The effectiveness of the algorithm is limited to multi-dimensional data sets only due to their desired group structure.

However, no single tool can be enough to eliminate information vulnerabilities. For example, it is technically impossible to stop a person who is duly authorized to access confidential information from improperly disclosing that information to someone else.

4.7 Description of the real data set

4.7.1 Overview of remittances data set

A remittance is a transfer of money by a foreign worker to an individual in their home country, including current transfers in cash or in-kind between resident and non-resident individuals. The area of regulatory concern in the remittances industry is the detection and prevention of money laundering activity.

According to the World Bank, remittance flows to developing country regions were estimated at US$440 billion in 2015 (World Bank, 2017). Migrant worker remittances compete with international aid as one of the most massive inflows in developing countries' economies.

Table 1 shows the frequency table of remittances outflows for 2015, which forms part of the cross-border financial flows data set collected by the SARB for BoP reporting and regulatory purposes. Most of the transactions averaged under 100 US$ since migrants tend to send smaller amounts more frequently. Remittances play an increasingly large role in the economies of many countries surveyed by the World Bank Group (World Bank, 2017).For example, remittances contributed as much as 2.7% of India's Gross Domestic Product (GDP) in 2018 (Rebecca, 2019)

Table A.1 (see Appendix A) shows the alphabetic list of the few variables used in this illustration, created using the SAS® CONTENTS Procedure. Table A.2 (see Appendix A) shows the sample data set. The extracted data set comprised 3,823,732 financial transaction records, categorized for BoP reporting purpose as "migrant worker remittances" and "foreign national contract worker remittances" in 2015. The estimated number of unique parties to the recorded transactions were 427,322 residents and 231,173 non-residents.

Payments from residents to non-residents comprised 3,495,114 financial transactions. A total of 328,351 residents made these payments. Payments from non-residents to residents comprised 8.6% of the total transactions count, and 108,564 non-residents concluded those payments. Therefore, a total of 98,971 residents received at least one payment from a non-resident without making a single payment to a non-resident. Similarly, 122,909 non-residents received at least one payment from a resident without making any cross-border payment.

Table 1: Frequency distribution table of remittances outflows

$_Amount	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0 - 100	2,370,563	67.83	2,370,563	67.83
100 - 200	642,438	18.38	3,013,001	86.21
200 - 400	339,531	9.71	3,352,532	95.92
400 - 600	60,436	1.73	3,412,968	97.65
600 - 1,000	28,922	0.83	3,441,890	98.48
1,000 - 2,000	26,701	0.76	3,468,591	99.24
2,000 - 3,000	9,554	0.27	3,478,145	99.51
Over 3,000	16,969	0.49	3,495,114	100

The right-skewed distribution of the remittance outflows data set is presented using a bar chart instead of a histogram due to the unequal bin sizes in Figure 2.6

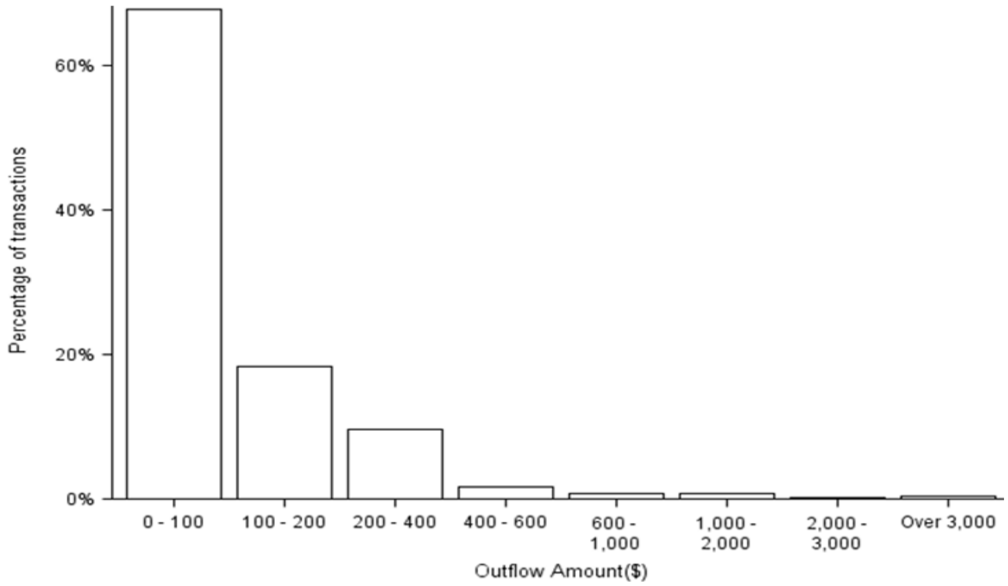


Figure 4.15: Bar graph of the distribution of remittance outflows

4.7.2 Data set validation

In the absence of unique identifiers for both resident and non-resident parties on the extracted data set, validating the number of distinct residents and non-residents is a complicated task.

For example, using a name for identifying a resident or non-resident may not result in a unique individual partly due to inconsistent spelling of names in the data set. For example, a resident named "Ndlovu" could be misspelled as "Ndhlovu" on the data set, resulting in more than one name for the same resident on the network. However, the international transaction reporting system uses validation rules for authenticating the uniqueness of residents/non-residents on the database using the identity/passport number and residential address or a combination thereof.

4.8 Summary

Chapter 4 provided an overview of cryptographic and non-cryptographic techniques used to preserve the privacy of personally identifiable information in a statistical database. Most importantly, this chapter introduced the proposed symmetric-key encryption algorithm to protect the privacy of personally identifiable data.

The algorithm was presented in two parts. The first part performed algebraic operations on the multi-dimensional data set for encryption purposes, while the second part computed the edge weights to facilitate the construction of the cross-border financial flows network from records of international transactions.

Chapter 4 implemented the proposed privacy-preserving algorithm on real data set comprising remittances flows using SAS® software and outlines its advantages and disadvantages. The chapter concluded with an overview of the data set used to illustrate the research's central theories.

Chapter 5: Network structure of cross-border financial flows

5.1 Introduction

This chapter provides a mathematical definition of the cross-border financial flows network, comprising the resident and non-resident vertex sets. A first step in analyzing the network structure is to create a visualization of it. However, network visualizations are mostly useful when the number of nodes and edges are not quite large. Hence, the two visualizations presented use a sample of real data set extracted from the database of international financial transactions of the SARB and the hypothetical data set, respectively.

The research uses SAS® Visual Analytics software to present the network visualizations. Eyeballing the visualizations reveals network features suggesting suspicious transaction patterns between residents and non-residents. For example, it is uncommon for remittance transactions to involve multiple residents and non-residents transactions, which may indicate the existence of some illicit financial schemes.

5.2 Network structure of cross-border financial flows

FATF's recommendation number five requires financial institutions to identify the beneficial owner and take reasonable measures to verify the beneficial owner's identity to its satisfaction. Each cross-border financial transaction comprises two distinct nodes, resident and non-resident. The edge links between the two nodes are the existing financial transactions between residents and non-residents.

A directed and weighted bipartite network serves as a convenient model to depict cross-border financial flows. The presence of directed and weighted arcs differentiates the proposed model from other bipartite graphs used in several studies, including the structure used for describing the world trade web (Ermann & Shepelyansky, 2013). The undirected and weighted bipartite graph model was proposed as a model to measure influence diffusion in online social networks (Zhiguo, Jingqin, & Liping, 2015).

Formally, the cross-border financial flows network is defined as the directed and weighted bipartite graph $G = (V, A, w)$, with $V(G) = V_R \cup V_{NR}$ and $A(G) \subseteq (V_R \times V_{NR}) \cup (V_{NR} \times V_R)$. The disjoint vertex sets $V_R = \{r_1, \ldots, r_k\}$ and $V_{NR} = \{nr_1, \ldots, nr_l\}$ represent the resident vertex set and the non-resident vertex set with $|\, V_R| = k$ and $|V_{NR}| = l$. The directed arcs set $A(G)$

represent the direction of the financial flows, where outflows are from residents to non-residents and inflows are from non-residents to residents.

The weight function reduces the parallel multi-edges between pairs of vertices into a single edge by computing the total number of transactions and the total value of transactions. The weight function is defined by a pair of matrices used to measure the intensity of the flows between vertex pairs.

Figure 5.1 shows a schematic depiction of the network structure of cross-border financial flows with single edge weights indicating the total financial value of transactions between node pairs.

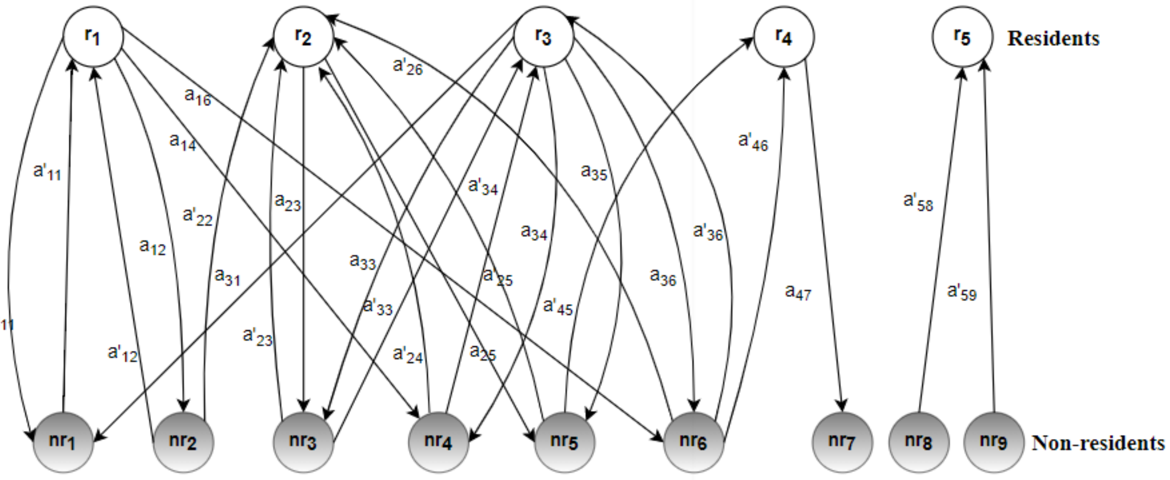


Figure 5.1: Schematic depiction of the cross-border financial flows network

5.3 Adjacency matrix representation

Denote by $\mathbb{R}^{k \times l}$ the set of $k \times l$ matrices with non-negative real entries. We arrange the vertex set $V_R \cup V_{NR}$ of the cross-border financial flows network in the order $r_1, \ldots, r_k, nr_1, \ldots, nr_l$. The matrices comprising weights are $\boldsymbol{A}$ and $\boldsymbol{B}$, which are elements of $\mathbb{R}^{k \times l}$ with entries $\boldsymbol{A} = \{a_{ij}\}$ and $\boldsymbol{B} = \{b_{ij}\}$, respectively, such that

$$A_{ij} = \begin{Bmatrix} a_{ij} \text{ if resident } r_i \text{ transferred funds to non} - \text{resident } nr_j \\ 0 \text{ otherwise} \end{Bmatrix} \tag{5.1}$$

where

$$a_{ij} = \sum\nolimits_{transaction\ count}(r_i \rightarrow nr_j) \tag{5.2}$$

and

$$B_{ij} = \begin{Bmatrix} b_{ij} \text{ if } a_{ij} > 0 \\ 0 \text{ otherwise} \end{Bmatrix} \tag{5.3}$$

where

$$b_{ij} = \sum\nolimits_{transaction\ amount} \left(r_i \rightarrow nr_j\right) \tag{5.4}$$

Similarly, entries of matrices $\boldsymbol{A'} \in \mathbb{R}^{l\times k}$and $\boldsymbol{B'} \in \mathbb{R}^{l\times k}$ are such that

$$A'_{ij} = \begin{Bmatrix} a'_{ij} \text{ if non} - \text{resident } nr_j \text{ transferred funds to resident } r_i \\ 0 \text{ otherwise} \end{Bmatrix} \tag{5.5}$$

where $$a'_{ij} = \sum\nolimits_{transaction\ count} \left(nr_j \rightarrow r_i\right) \tag{5.6}$$

$$B'_{ij} = \begin{Bmatrix} b'_{ij} \text{ if } a'_{ij} > 0 \\ 0 \text{ otherwise} \end{Bmatrix} \tag{5.7}$$

where

$$b'_{ij} = \sum\nolimits_{transaction\ amount} \left(nr_j \rightarrow r_i\right) \tag{5.8}$$

The adjacency matrix $\boldsymbol{F}$ of the cross-border financial flows network is of the form

$$\boldsymbol{F} = \begin{bmatrix} 0_{k,k} & A \\ A' & 0_{l,l} \end{bmatrix} \tag{5.9}$$

where $0_{k,k}$ and $0_{l,l}$ represent the $k \times k$ and $l \times l$ zero matrices. The matrices A and A' in equation 5.9 are replaced with B and B' when using the total financial value of transactions as the network weights.

The adjacency matrix is a simple and convenient method of representing the cross-border financial flows network on a computer. However, it uses computer memory inefficiently due to its large number of zero entries (Newman M. E., 2010). Network representation methods such as the adjacency lists, and edge lists do not suffer from the same weaknesses as the adjacency matrix.

Table A.2 (see Appendix A) shows a representation of the cross-border financial flows network on a computer using the summary list of transactions. To interpret the network data in Table A.2, consider observation number 2009. A resident labeled r10075708 entered into a single financial transaction with a non-resident labeled nr15905119, whereby r10075708 paid non-resident nr15905119 USD 302.03. Observation number 2010 indicates that the same resident entered into 26 financial transactions with a non-resident labeled nr15908732, whereby r10075708 paid non-resident nr15908732 the sum of USD 4 982.70. The table shows the temporary variables generated by the encryption algorithm for illustration only. Note that 'First. Resident Indicator' equals zero instead of one in the latter scenario due to the multiple transactions. The outward payments by residents represent inward receipts by non-residents.

5.4 Cross-border financial flows network visualization

The visualization of the cross-border financial flows network depicts the directed and weighted bipartite graph, comprising two disjoint sets of nodes that are differentiated by their colours, with the condition that no two connected nodes are of the same colour. Figure 5.2 shows the visualization based on a sample of real data drawn from the ITRS of the SARB. SAS ® Visual Analytics software version 7.4 was used to create the visualization.

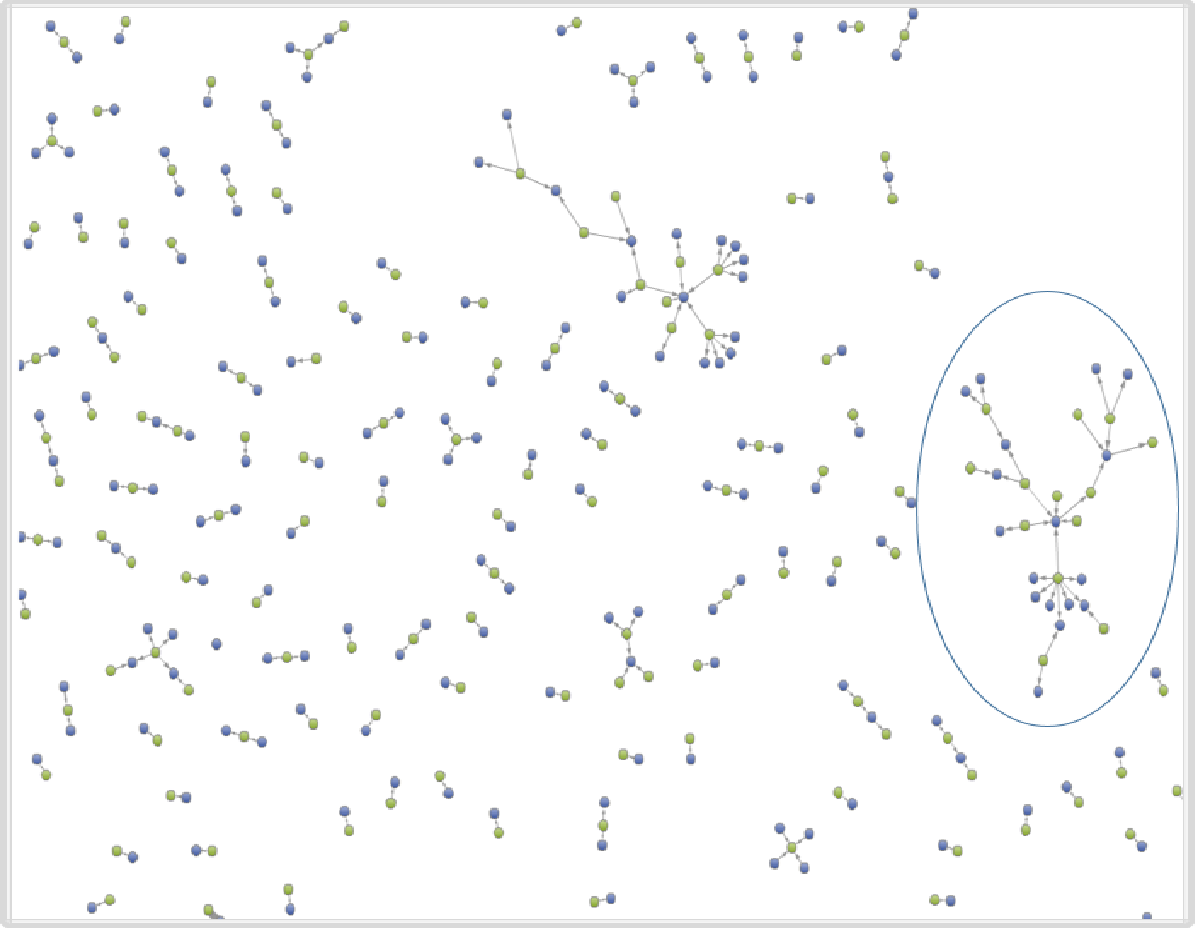


Figure 5.2: Cross-border financial flows network visualization

The visualization of Figure 5.3a shows the highly connected nodes highlighted in Figure 5.2 while Figure 5.3b shows the visualization of the network based on the example data set. The visualization was created using SAS® Visual Analytics software. 7.4. Copyright 2014-2017, SAS Institute Inc., Cary, NC, USA. The "ungrouped" node-link pairs, which require the network data to be structured to fit the basic data roles being "source" and "target" was used to generate the visualization.

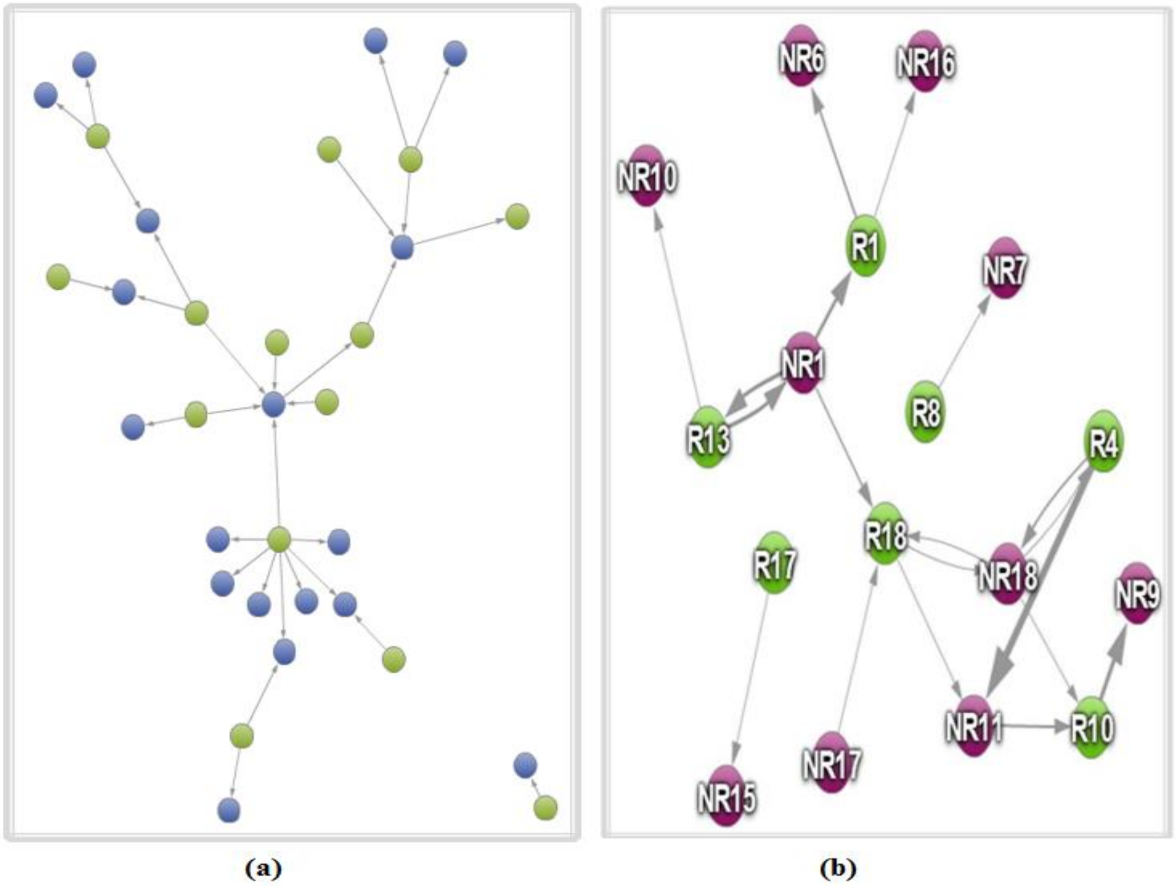


Figure 5.3: (a) The highly connected nodes of Figure 5.2, (b) Network visualization based on the example data set

The outward payment flows had residents as the source nodes and non-residents as the target nodes while the inward payment flows had non-residents as the source nodes and residents as the target nodes. An additional binary variable was created to differentiate between the resident nodes and the non-resident nodes, which facilitates the depiction of the bipartite structure of the cross-border financial flows network.

The create the visualization depicted Figure 5.4b in the SAS® Visual Analytics Explorer window, load the transformed data set into the SAS® LASR Server, and use the following network roles (SAS Institute Inc, 2017; Sekgoka & Adetunji, 2019):

- Network type = Ungrouped
- Source = Source Node (determines the direction of payment flows)
- Target = Target Node
- Node size set to “empty.”
- Node color = Resident indicator (binary 1 = yes, 0 = no)
- Link width = Amount (thick links for large payment flows)

- Link color and data tips set to “empty.”

5.5 Significance of the cross-border financial flows network

The research proposed a network representation of cross-border financial transactions, with a formal mathematical definition. The network model used the directed and weighted bipartite graph. Upon gathering network data on the network structure, what insights can be drawn from the data? What lessons can be learned about the function of the international financial transactions in an economy? What are the statistical properties of the network related to the issues of concern for this research?

The next chapter develops a centrality measure, an important and useful class of network measures to answer the questions posed in this research. Furthermore, the research characterizes the network using degree distributions to provide answers to the research question.

5.6 Summary

Chapter 5 presented the directed and weighted bipartite graph as a model for cross-border financial transactions between residents and non-residents of a country. The adjacency matrix representation of the network proved cumbersome due to a large number of zero entries. Hence, the research proposed using a list of transactions to represent the number on a computer.

Also, the chapter provided the visualization of the network using SAS® Visual Analytics software. The visualization simplifies identifying the network's highly connected nodes and motivates using the network's statistical properties to understand the network structure better.

Chapter 6: Cross-border financial flows network measures

6.1 Introduction

The chapter develops the centrality measure for the cross-border financial flows network using a technique based on matrix multiplication to answer the question, "Which resident/non-resident nodes are the most important in the cross-border financial flows network?" The answer to this question provides data mining insights about the cross-border financial flows network structure. In particular, the proposed measure identifies the dense subgraphs of the cross-border financial flows network.

The chapter begins with some network measures, including node centrality, the random network model, and the configuration model. The configuration model serves as the benchmark in calculating network modularity, which measures the extent to which nodes connect to their types in a network.

Supervisory authorities can use the proposed centrality measures to identify the residents/non-residents nodes responsible for transferring large volumes of financial flows and large transaction volumes. The measure can provide some leads in identifying unusual patterns of interaction between nodes, such as transactions with substantial volumes but low financial value, and most importantly, the identification of nodes that transfer funds to common nodes.

According to FATF's anti-money laundering recommendation, the identified nodes trigger investigations of the apparent economic or visible lawful purpose of the transactions, which may necessitate examinations of their background and purpose.

The second part of this chapter uses a hierarchical clustering technique to derive the resident node set's approximate degree distribution. Using cluster analysis aims to group the network residents by their degrees so that residents within groups are similar to one another and dissimilar between the groups. The clustering procedure reduces the degree dimensions of the resident node-set from two to one. The resulting clusters characterize the cross-border financial flows network, making it easier for the regulatory organizations to plan their inspections as part of the risk-based supervision of regulated entities.

6.2 Network density and centrality measures

After network data gathering, visualization is the first step in analyzing the network structure to provide more information about the system. However, direct visualization with a human eye is only useful for small to medium-sized networks. The case of vast networks with millions of edges and thousands of vertices requires network measures to obtain helpful information about the network structure and its properties.

A useful and essential class of network measures is centrality, which measures nodes' importance in a network. Centrality addresses the question, "Which are the most important or central nodes in a network?" Several centrality measures were developed based on different concepts and definitions of what it means to be important in a network.

To exemplify the idea of node centrality, (Freeman L. C., 1978) used a network consisting of five nodes and four edges, see Figure 6.1. The size of the nodes corresponds to the node's degree. Node A can reach all the others more quickly; it controls the flow between the others; and has more ties, giving it three advantages over the other nodes. Based on these three features, he formalized three different node centrality measures: degree, closeness, and betweenness.

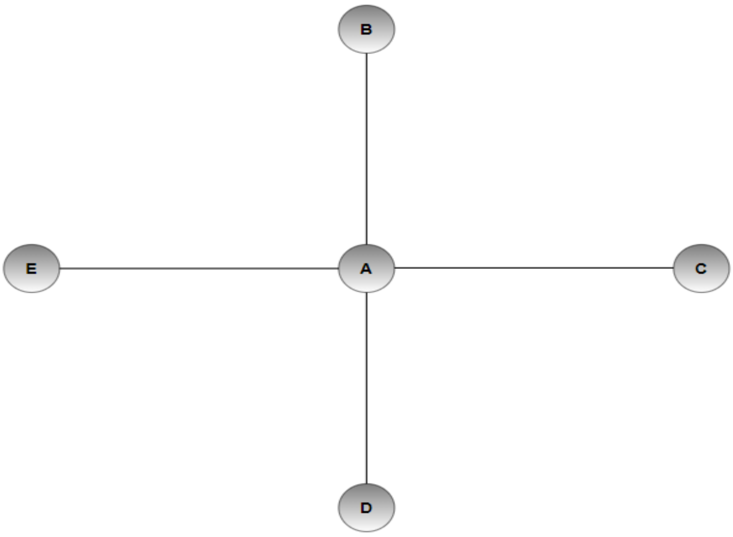


Figure 6.1: A star network with 5 nodes and 4 edges, adapted from (Freeman L. C., 1978).

This section discusses network density, the widely used centrality measures, followed by the proposed centrality measure for the cross-border financial flows network.

6.2.1 Network density

The edge density ρ of a network is the fraction of the number of edges $|E|$ with respect to the maximum possible edges. For simple directed network ((i.e., one with no multi-edges or self-edges), the edge density is

$$\rho = \frac{|E|}{|V|(|V-1|)} \tag{6.1}$$

where E is the number of edges and V is the number of nodes in the network. The density lies strictly in the range $0 \leq \rho \leq 1$. A network or subgraph with a density of one is called a clique. The opposite of a dense network is a sparse network, one with only a few edges.

The density of a bipartite network as computed by the equation above can never reach one. Therefore, the alternative definition for the density of a bipartite network is:

$$\rho = \frac{|E|}{(|U||V|)} \tag{6.2}$$

where E is the number of edges and U and V are the number of nodes in the bipartite network.

6.2.2 Degree centrality

Perhaps the simplest of these measures is the degree centrality of a node, which is the number of edges connected to the node. Although degree centrality is a simple measure, it can illuminate the network structure and capture its features. For instance, it seems reasonable to suppose that individuals with connections to many others in social networks might have more influence, more access to information, or more prestige than those with fewer connections.

A non-social network example is the use of citation counts in the evaluation of scientific papers. The number of citations a paper receives from other papers, which is its in-degree in the citation network, gives a crude measure of whether the paper has been influential or not and is widely used to measure the impact of scientific research (Newman, 2010).

Degree centrality applies to the directed bipartite networks such as the cross-border financial flows network without modifications. The cross-border financial flows network comprises two sets of nodes; each node has two different degrees, resulting in four different degrees. The nodes with unusually high degrees are called hubs.

6.2.3 Eigenvector centrality

A natural extension of degree centrality is eigenvector centrality (also called prestige score). This measure assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. A high eigenvector score means that a node is connected to many nodes who themselves have high scores.

For a given graph vertices and adjacency matrix $G = (V, E)$ with $|V|$ vertices and adjacency matrix $A = (a_v, t)$, the relative centrality, x, score of vertex v can be defined as:

$$x_v = \frac{1}{\lambda} \sum_{t \epsilon M(v)} x_t = \frac{1}{\lambda} \sum_{t \epsilon M(v)} a_v, t\, x_t \tag{6.3}$$

where $M(v)$ is a set of the neighbors of v and λ is a constant. In theory, this measure can be calculated for either undirected or directed networks. However, it works best for undirected networks. It requires some modifications before applying it to directed networks. A variation of the eigenvector centrality called Katz centrality addresses the eigen vector centrality problems (Katz, 1953; Newman, 2010).

6.2.4 Closeness centrality

Closeness centrality is based on the measure of distance in a network. It measures the average distance from a node to other nodes in a network. The closeness centrality of node u_i is given by

$$c_i = (|U| - 1) \Big/ \sum_{j \neq i} d(u_i, u_j) \tag{6.4}$$

Hence, the closeness centrality of a node is the inverse of its average distance to all other nodes in the network. The minimum distance between two nodes of a bipartite network is two if the nodes are for the same type, otherwise the minimum distance equals one.

6.2.5 Betweenness centrality

A complex measure of centrality is betweenness centrality. It measures the extent to which a node lies on paths between other vertices. This measure is attributed to Freeman (Freeman, 1977). However, Freeman pointed out that this measure was independently proposed some years earlier by Anthonise in an unpublished report. The betweenness centrality of node u_i is given by

$$b_i = \frac{1}{\lambda} \sum_{j,k} \frac{g^i{}_{jk}}{g_{jk}} \tag{6.5}$$

where g_{jk} is the number of shortest paths between nodes u_j and u_k and $g^i{}_{jk}$ is the number of shortest paths between nodes u_j and u_k that contain u_i. If $g_{jk} = g^i{}_{jk} = 0$ then $\frac{g^i{}_{jk}}{g_{jk}} = 0$ by definition.

6.3 Random networks

The random network model is one of the classical network models which provide a basis for describing networks using probability distributions. Paul Erdős and Alfred Rényi developed the popular random network model, also known as the ER model (Erdős & Rényi, 1959). Edgar Gilbert also introduced this model independently and contemporaneously (Edgar, 1959).

The random network model is considered in several fields, including sociology and mathematical biology, due to its simplicity. The significance of statistical properties observed in real-world systems can be evaluated based on the discrepancy between the random and real-world networks. Hence, the random networks used for comparison must have the same order and size, and the same degree distribution or sequence as the observed network. This research provides an overview of two popular techniques to generate random networks that can be useful for comparison purposes, i.e., the configuration model and the Curveball algorithm.

6.3.1 The configuration model

The most widely studied random network model is the configuration model, with applications in both one-mode and two-mode networks. The configuration model is a random network model with a given degree sequence, rather than degree distribution. Hence, it allows the user to incorporate arbitrary degree distributions (Newman, 2010).

The configuration model begins with a network of a given order and size zero. Each node has a specified number of half-edges or stubs corresponding to its desired degree. Hence, fixing the number of nodes and the number of edges in the random network. Next, the model chooses pairs of stubs at random and connects them to form an edge. The result is a network in which every vertex has exactly the desired degree. Figure 6.2 depicts the configuration model.

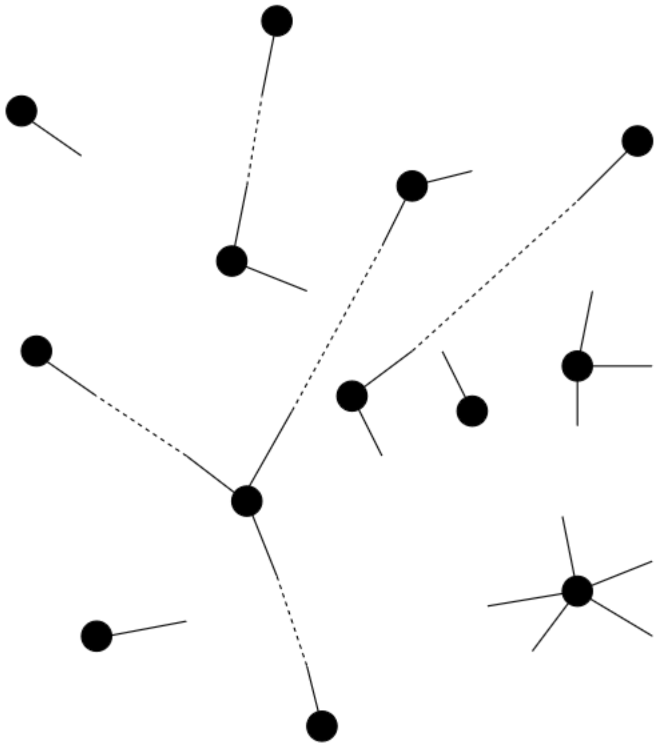


Figure 6.2: The configuration model

The uniform distribution over matchings in the configuration model has the necessary consequence that any sub is equally likely to connect to any other stub. This crucial property makes the configuration model solvable for many of its characteristics.

In the directed configuration model (DCM), each node is given as a number of half-edges called tails and heads. Researchers have used the DCM to model complex real-world networks such as neural networks, finance, and social networks (Amini & Minca, 2013; Li H. , 2018).

6.3.2 The Curveball algorithm

The Curveball algorithm randomizes the adjacency matrix while preserving its row and column totals of a network (Strona, Nappo, Boccacci, Fattorini, & San-Miguel-Ayanz, 2014). Preserving the row and column totals of the adjacency matrix is equivalent to fixing the degree sequence. The Curveball algorithm is not computationally intensive and produces uniformly distributed matrices with fixed row and column totals (Casterns, 2015).

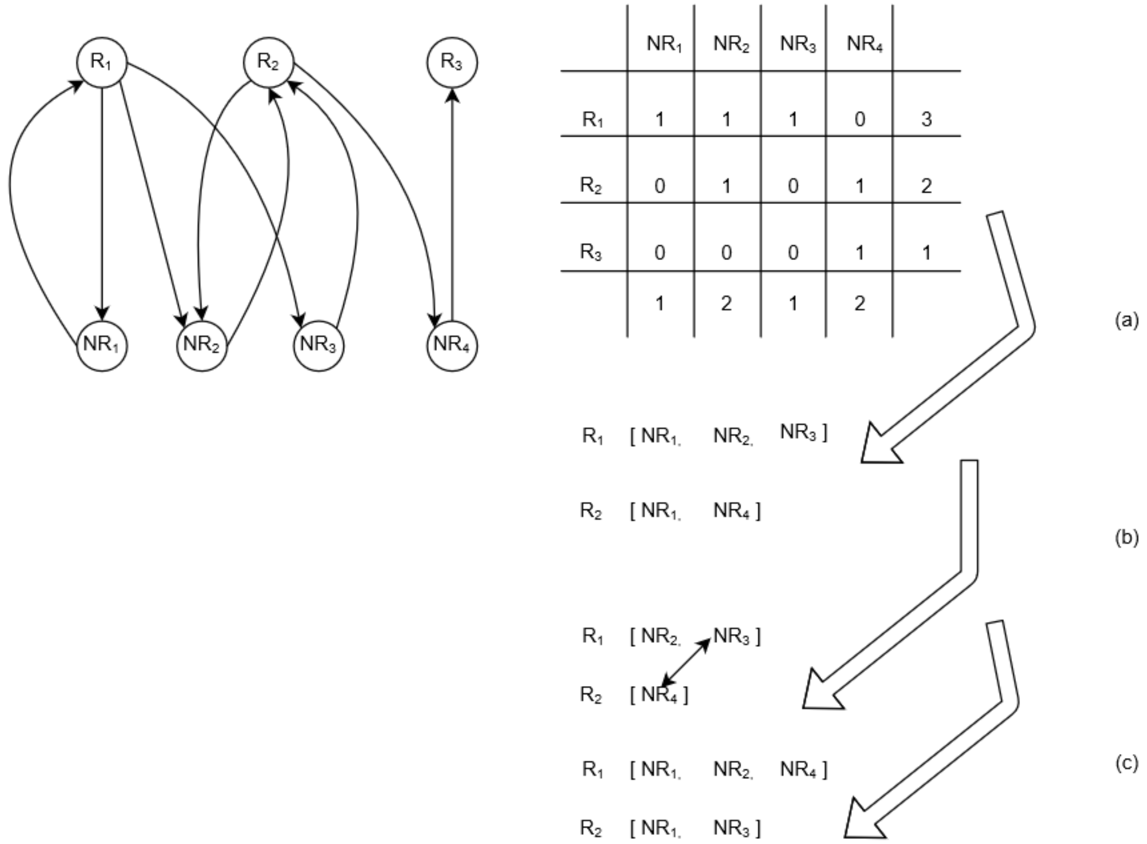
To randomize the adjacency matrix A, the algorithm randomly extracts two rows of A, say A_i and A_j, to create two lists. Next, the algorithm compares the two lists to identify the elements present in A_i but not in A_j, and the elements present in A_j but not in A_i. The procedure then selects the identified elements at random from A_i and trades them with the identified elements in A_j. In general, for any pair of lists, a certain number of elements exclusive of a list are traded with an equal number of elements exclusive of the other list. The number of trades for each pair of lists will vary from 0 to n, where n is the size of the smaller of the two sets of exclusive elements.

Figure 6.3 illustrates the functioning of the Curveball algorithm. It starts with the directed bipartite network diagram and its adjacency matrix, which comprises the outflows only, i.e., (directed edges from Resident nodes (R) nodes to Non-resident nodes (NR)).

(a) Randomly choose two resident nodes, say R_1 and R_2.

(b) The two lists are compared to identify the set of non-residents present in R_1 but not in R_2 and the set of non-residents present in R_2 but not in R_1.

(c) Randomly choose one member of R_1, say NR_3 and trade with the only member of R_2 resulting with a newly formed list.



	NR₁	NR₂	NR₃	NR₄	
R₁	1	1	1	0	3
R₂	0	1	0	1	2
R₃	0	0	0	1	1
	1	2	1	2	

R₁ [NR₁, NR₂, NR₃]
R₂ [NR₁, NR₄]

R₁ [NR₂, NR₃]
R₂ [NR₄]

R₁ [NR₁, NR₂, NR₄]
R₂ [NR₁, NR₃]

Figure 6.3: Illustration of a ‘trade’ in the Curveball algorithm.

6.4 Cross-border flows network centrality

This research proposes a centrality measure based on the principle of matrix multiplication, commonly used to obtain the dot product of vectors. Formally, consider the weights of the cross-border financial flows network, i.e., $\boldsymbol{A}$ and $\boldsymbol{B}$.

Let $\boldsymbol{C} = \boldsymbol{AB}^T$ be the asymmetric matrix with entries $\boldsymbol{C}_{ij}$ that are the sum product of the rows of $\boldsymbol{A}$ and columns of $\boldsymbol{B}^{\boldsymbol{T}}$. Define the centrality measure for resident i as follows:

$$\boldsymbol{N}_i = \frac{\sum_j \boldsymbol{C}_{ij}}{\|\boldsymbol{C}\|} \tag{6.6}$$

where $\|\boldsymbol{C}\|$ is the sum of all the entries of the matrix $\boldsymbol{C}$ and

$$\sum_i N_i = 1 \tag{6.7}$$

The diagonal entries C_{ii} are the sum product of transaction volumes and financial values for resident i. Large diagonal entries indicate large transfers of funds across country borders by residents. The onus is on the financial institutions to verify the financial flows in the event of extreme values indicated by the centrality measure.

If $C_{ij} > 0$ for $i \neq j$ it means that resident i and resident j transferred funds to the same non-resident. If $C_{ij} \neq 0$ then $C_{ji} \neq 0$ and $C_{ij} \neq C_{ji}$. C_{ij} is equal to the product of the number of transactions for resident i and the financial value of transactions for resident j while C_{ji} is equal to the product of the number of transactions for resident j and the financial value of transactions for resident i. Hence, the centrality measure for resident i and resident j increase when the two residents transfer funds to the same non-resident, but the measure does not increase by the same amount.

It should be noted that it is not the absolute value of the centrality measure that matters but the high or low measure value. The centrality measure based on matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ provides an indication of the importance of each of the resident nodes in the cross-border financial flows network. It makes it easier to visually analyze the cross-border financial flows network.

The centrality measure for non-residents is similarly defined, where $\boldsymbol{C} = \boldsymbol{A}'\boldsymbol{B}'^{T}$. It measures the importance of non-residents as the source of funds in the network.

6.4.1 Illustration of the adjacency matrix representation

The entries of the asymmetric adjacency matrix of the cross-border financial flows derived from the example data set are represented by matrices $\boldsymbol{A}$ and $\boldsymbol{B}^T$ below. To interpret the two matrices together, consider the first row of matrix $\boldsymbol{A}$ and the first column of matrix $\boldsymbol{B}^T$. The entry in the second row and first column of $\boldsymbol{B}^T$ indicates that the total financial value of 1500 was transferred in two cross-border transactions shown in the first row and second column entry of the matrix $\boldsymbol{A}$ by resident "$R1$" to non-resident "$NR6$".

A simpler interpretation of the adjacency matrix is provided in Table A.6, where the entries of the matrix are the co-ordinate pairs $(\boldsymbol{x}, \boldsymbol{y})$, representing the transaction volumes and the associated financial values based on the example data set, respectively.

$$\boldsymbol{A} = \begin{bmatrix} 0 & 2 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 3 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 2 \end{bmatrix}, and \tag{6.8}$$

$$\boldsymbol{B}^T = \begin{bmatrix} 0 & 0 & 0 & 0 & 2250 & 0 & 0 \\ 1500 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 550 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3500 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2750 & 0 & 0 \\ 0 & 6000 & 0 & 0 & 0 & 0 & 250 \\ 0 & 0 & 0 & 0 & 0 & 500 & 0 \\ 500 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 7000 & 0 & 0 & 0 & 0 & 400 \end{bmatrix} \tag{6.9}$$

6.4.2 Illustration of network centrality measure

The entries of the matrix **Q** represent the normalized values of the product of matrices $\boldsymbol{A}$ and $\boldsymbol{B^T}$

$$\boldsymbol{Q} = \begin{bmatrix} 0.0471 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.4438 & 0 & 0 & 0 & 0 & 0.0229 \\ 0 & 0 & 0.0148 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.0471 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.1345 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.0067 & 0 \\ 0 & 0.2690 & 0 & 0 & 0 & 0 & 0.0141 \end{bmatrix} \tag{6.10}$$

Summing over each of the rows of $\boldsymbol{Q}$ yields the desired centrality measure in equation (6.6) as follows:

$$N_i \begin{bmatrix} R_1 \\ R_4 \\ R_8 \\ R_{10} \\ R_{13} \\ R_{17} \\ R_{18} \end{bmatrix} = \begin{bmatrix} 0.0471 \\ 0.4667 \\ 0.0148 \\ 0.0471 \\ 0.1345 \\ 0.0067 \\ 0.2831 \end{bmatrix} \tag{6.11}$$

The centrality measure illuminates the cross-border financial flows network by providing a measure of node importance instead of relying on the human eye. For example, it is possible to identify nodes with large financial flows by inspecting the diagonal elements of the matrix Q. The measure also identified the nodes characterized by large financial flows, many connections, and large volumes, namely, R_4,R_{18} and R_{13} by allocating the highest centrality scores to such nodes.

The proposed centrality measure can assist authorities in planning their inspections of regulated entities for compliance with the Exchange Control laws in countries that maintain Exchange Controls. Authorities can identify residents who deliberately transact large volumes with low financial values from several authorized dealers to avoid detection. Upon identifying such nodes, regulatory authorities can investigate their apparent economic or visible lawful purpose under the FATF recommendations. The network structure facilitates identifying all the linked nodes to enable targeted investigations. The off-diagonal elements of the matrix **Q** increased the centrality for nodes that share connections.

6.5 Network characterization using degree distribution

Many often cite the vertex degree's frequency distribution as the most fundamental of network properties and the defining characteristic of the network structure. The directed bipartite networks comprise four vertex degree distributions, two for each vertex set (Newman M. J., 2010). The cross-border financial flows network comprises two sets of nodes; each node has two different degree distributions, resulting in four different degree distributions. The research uses the clustering procedure to reduce the cross-border financial flows network's degree distributions from four to two, thereby providing a distribution summary of the dual-weighted network.

This section characterizes the cross-border financial flows network using the resident node set's degree distribution. The resident node-set mainly acts as the gateway for both inward and outward financial flows of a country.

6.5.1 Description of the classification procedure

Many often cite the vertex degree's frequency distribution as the most fundamental network property and the network structure's defining characteristic. The directed bipartite networks comprise four vertex degree distributions, two for each vertex set (Newman M. E., 2003).

Ward's minimum variance procedure reduces the degree dimensions from four to two, thereby enabling cluster visualizations and simplifying large data sets analysis.

The rationale for choosing the Ward clustering method is that the hierarchical tree diagram computed with this method has a well-defined look in which clusters jump out at the eye. Anderberg provides a detailed mathematical treatment of the Ward clustering method (Anderberg, 1973). The early availability of a computer algorithm for the Ward clustering algorithm provided the stimulus to its wide use (Veldman, 1967; Romesburg, 2004).

Table 6.1 depicts the network data matrix, showing the input data's layout for the clustering procedure. The symmetric-key encryption algorithm computes the network data matrix entries, hence its use for constructing the cross-border financial flows network. The matrix entries represent the degrees per resident node as follows:

Let $x'_{ij} = \begin{cases} 1 \text{ if } b_{ij} > 0 \\ 0 \text{ otherwise} \end{cases}$ and $y'_{ij} = \begin{cases} 1 \text{ if } b'_{ij} > 0 \\ 0 \text{ otherwise} \end{cases}$ be the two indicator variables that denote the presence of cross-border flows from resident i to non-resident j and from non-resident j to resident i, respectively. The degrees are the connection counts in both directions.

Table 2: Network data matrix for the resident vertex set

Resident	r_1	r_2	...	r_k
Out-degree	$\sum_{j=1}^{l} x'_{1j}$	$\sum_{j=1}^{l} x'_{2j}$	...	$\sum_{j=1}^{l} x'_{kj}$
In-degree	$\sum_{j=1}^{l} y'_{1j}$	$\sum_{j}^{l} y'_{2j}$	...	$\sum_{j=1}^{l} y'_{kj}$

If there are g residents to be grouped using n variables then the clustering procedure follows a series of steps that begins with g clusters and reduces the number of clusters from g to $g-1$ in a manner that would minimize the total within group Error Sum of Squares (ESS) associated with each cluster formed and then, without modifying the clusters formed, repeats the process until the number of clusters is systematically reduced from g to 1 (Ward, 1963).

Alternatively, the problem could be formulated as a maximization problem where the proportion of variation explained by a grouping of vertices is maximized. Formally, let m_k denote the number of observations in the kth of h clusters and define the following quantities:

x_{ijk} = the value on the ith of n variables for the jth of m_k observations,

$\bar{x}_{ik} = \sum_{j=1}^{m_k} \frac{x_{ijk}}{m_k}$ is the mean on the ith variable for observations in the kth cluster,

$ESS_k = \sum_{i=1}^{n} \sum_{j=1}^{m_k} (x_{ijk} - \bar{x}_{ik})^2$ is the error sum of squares for cluster k,

$ESS = \sum_{k=1}^{h} E_k$ is the total within group error sum of squares for the collection of clusters,

$TSS = \sum_i \sum_j \sum_k (x_{ijk} - \bar{x}_{..k})^2$ is the total sum of squares.

The objective is to maximize r^2,

where

$$r^2 = \frac{TSS - ESS}{TSS} \tag{6.12}$$

The numbers of clusters to use in the classification must be determined to conclude the clustering procedure and interpret the classification results. Unfortunately, there is no prior knowledge of the number of clusters to use in a classification. It is also widely acceptable that there are no completely satisfactory methods for determining the number of clusters for any type of cluster analysis (Hartigan, 1975).

Milligan and Cooper compared 30 methods of estimating the number of clusters in a classification using four hierarchical clustering procedures and recommended that use of some automatic decision to eliminate the problems associated with human subjectivity (Milligan & Cooper, 1985). We use a combination of a cluster validity index plot called the $pseudo\ t^2$ statistic and a dendrogram to provide a graphical assessment of the cluster solution. The $pseudo\ t^2$ statistic is defined as

$$Pseudo\ t^2 = \frac{B_{JK}}{\left(\frac{W_J + W_K}{N_J + N_K - 2}\right)} \tag{6.13}$$

where N_J and N_K are the number of observations in clusters j and k, W_J and W_K are within cluster sum of squares of clusters j and k, B_{JK} is the between cluster sum of squares during a step in the hierarchical procedure to merge cluster K and cluster L. The $pseudo\ t^2$ statistic quantifies the difference between two clusters that are merged at a given clustering step. Therefore, if it has a distinct jump at step j of the hierarchical clustering, then the clustering step $j + 1$ is selected as the optimal cluster. It is closely related to Duda and Hart's index (Duda & Hart, 1973).

After determining the number of clusters, the plot of $p(k)$ versus cluster k, is constructed, where $p(k) = n(k)/n$ is the number of residents that belong to cluster k in the cross-border financial flows network with n residents. Each cluster is uniquely defined by the clustering procedure with the cluster name and attributes such as minimum outdegree, maximum outdegree, minimum indegree, maximum indegree, and frequency of residents.

The plot of $p(k)$ versus cluster k represents the approximate degree distribution of the cross-border financial flows network with reduced degree dimensions obtained using the clustering procedure.

6.5.2 Characterization results using the remittances data set -

The research uses a threshold of 15 out-degrees per resident to reduce the remittances data set for computational convenience, resulting in 34,002 residents' characterization.

Figure 6.4 shows a graphical depiction of the $pseudo\ t^2$ statistic used to determine the criteria for estimating the number of clusters. The plot shows a distinctive jump when the number of clusters is 6. Therefore, the $pseudo\ t^2$ statistic suggests that clusters may be reasonably well-separated when the number of clusters is 7.

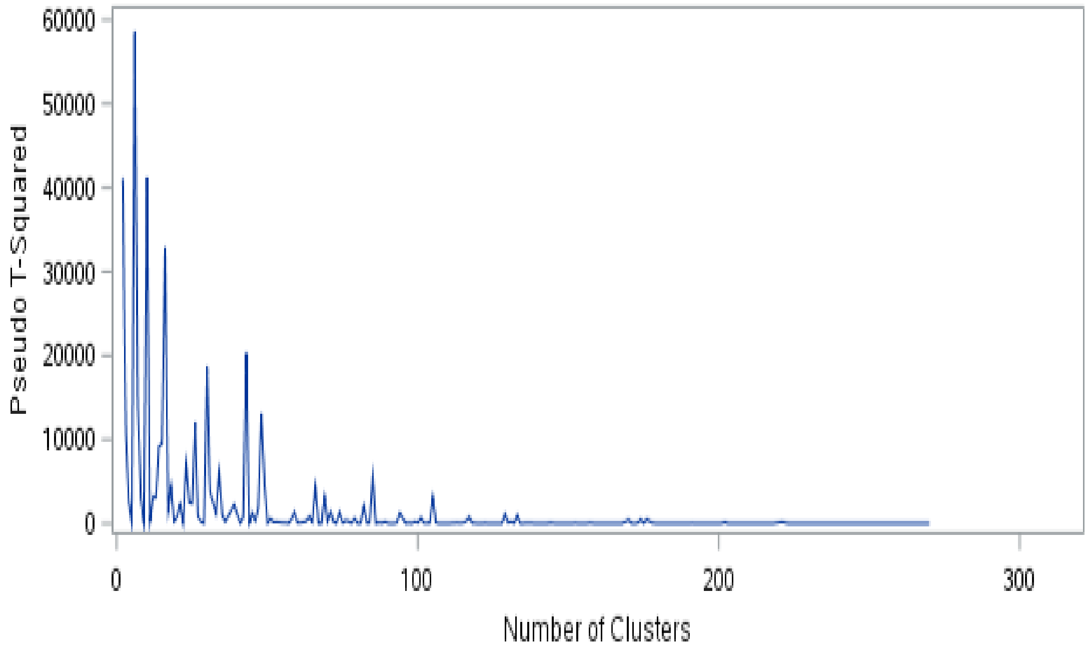


Figure 6.4: The pseudo t^2 statistics plot depicting the index at various number of clusters

Figure 6.5 shows the bar chart of the degree distribution of the resident vertex set of the cross-border financial flows network. Table A.3 (see Appendix A) indicates the vertex mergers starting with 34,002 clusters and ending with one cluster. The x-axis of the bar chart shows the 7-cluster partitions, while the y-axis displays the proportion of vertices in each cluster.

The distribution appears skewed to the right, with 18,390 residents (CL10 members) comprising 54% of the cross-border financial flows network. The maximum out-degree is 21, and the maximum indegree is 7 in this cluster. The last cluster (CL9) comprises five members, which resulted from two vertex mergers between CL13 (with four members) and resident r2434297. CL9 has the most connections, with an average out-degree of 615 and an average in-degree of 44.

Connections between clusters can potentially reveal interesting structural properties of the network, such as network nestedness. Nestedness indices mostly characterize bipartite networks such as this one (Csermely, London, Wu, & Uzzi, 2013). A visual assessment of the proposed degree distribution of residents shows that it is potentially more useful to analyze the one distribution graph instead of two distribution graphs (one for each degree type).

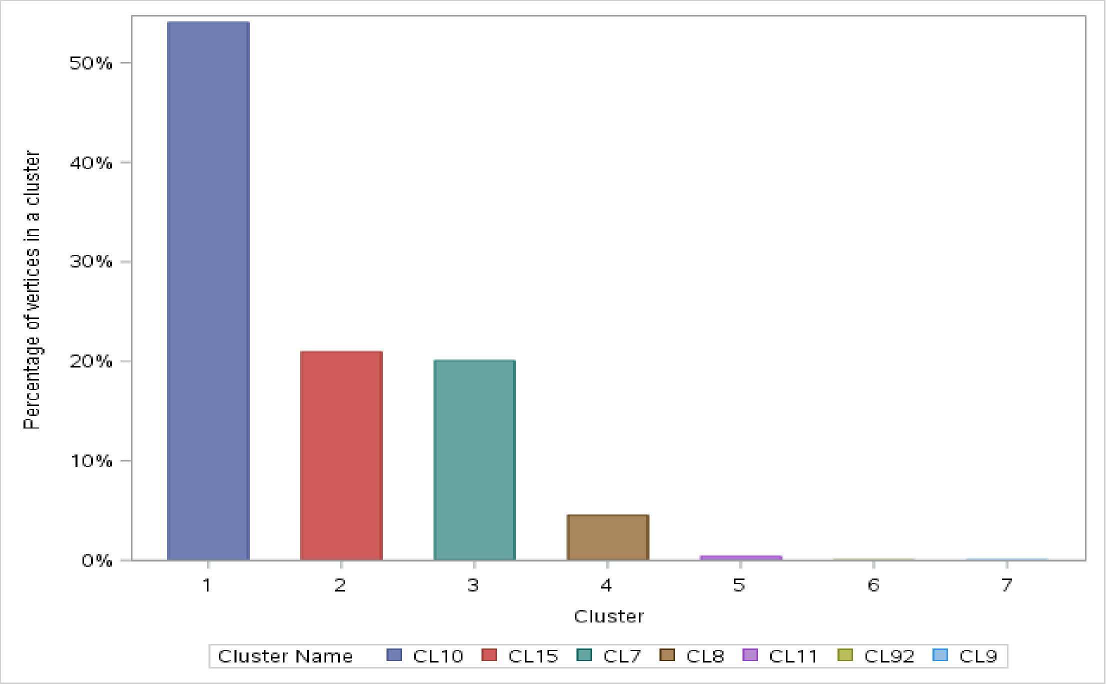


Figure 6.5: Bar chart of the degree distribution of the resident vertex set of the cross-border financial flows network

It is not feasible to depict the clusters' hierarchical formation using a dendrogram when the network size is enormous (34, 002 vertices). However, it is possible to use the last 50 clusters from the cluster history table to plot the dendrogram's top end, as illustrated in Figure 6.6. The dendrogram provides a visual assessment of the cluster solution.

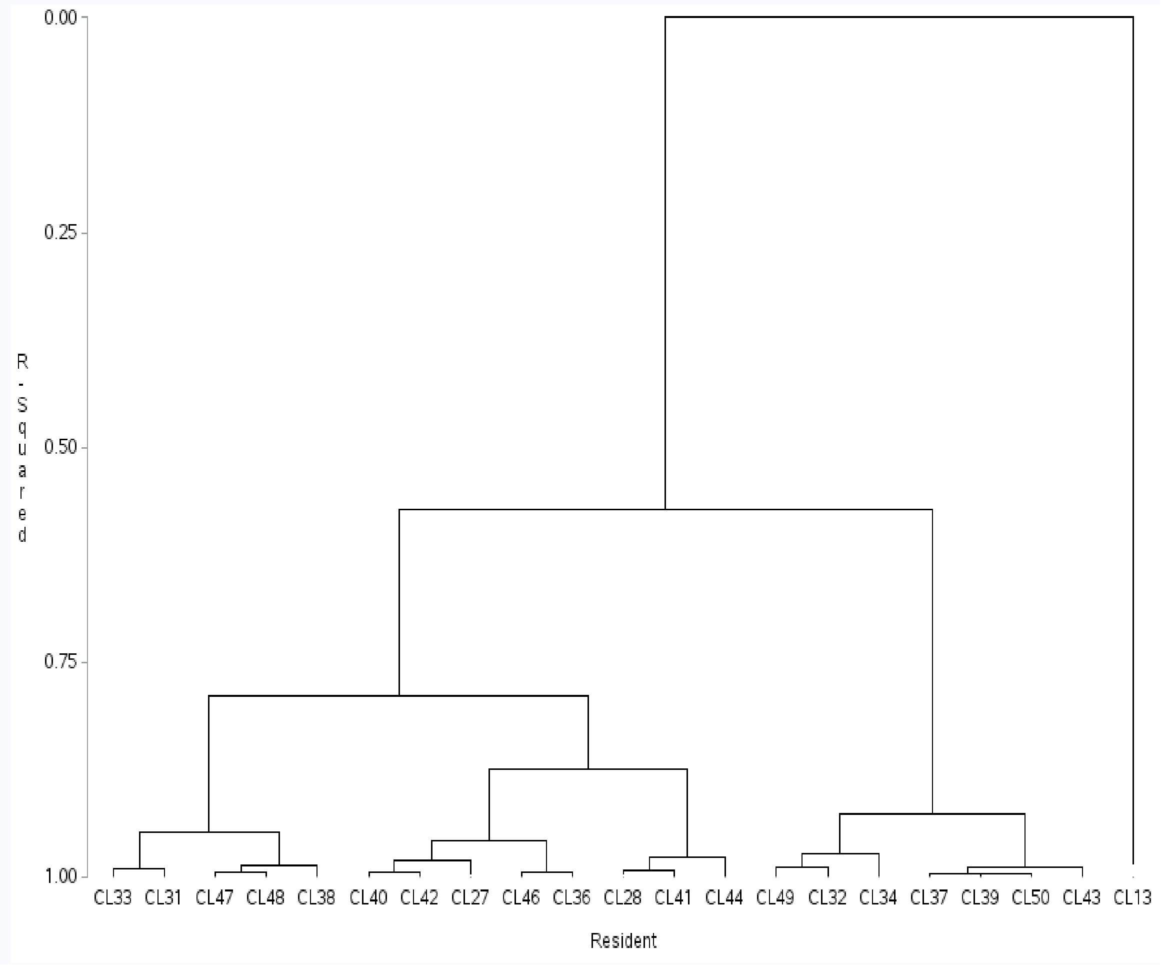


Figure 6.6: A dendrogram depicting the last 50 clusters of the cross-border financial flows network

The next step replicates residents' allocation to clusters in the data set, which creates a new variable. The computer code in Appendix B.4 creates a text file with embedded programming logic, generating multiple "if statements" to allocate residents to clusters. The resulting data set enables the computation of statistical measures such as sums and averages of inflows and outflows per resident cluster to provide further insight into the transaction patterns between residents and non-residents

6.5.3 Significance of the results

The characterization results availed additional variables in node clusters, enabling the computation of traditional statistics such as sums and averages of inflows and outflows per cluster to provide further insights into the transaction patterns between the resident and non-resident nodes. The significance of the clustering algorithm's variables is that they provide guidelines for financial institutions and regulatory organizations to design a sampling rule for inspections planning as part of AML strategies.

A representative sample of cross-border financial transactions for inspection can include transactions from each of the clusters instead of randomly selecting transactions as often done in practice. Most importantly, clusters that pose the highest risks, such as clusters 5, 6, and 7, can be afforded the most resources during the inspections.

6.6 Summary

In this chapter, the network centrality measure was derived based on the method of matrix multiplication to answer the question, “Which are the most important resident/non-resident nodes in the cross-border financial flows network?” The research used the hypothetical data set cross-border financial flows to demonstrate the proposed centrality measure. The results showed that the centrality measure effectively identified the resident nodes responsible for the most cross-border flows and those responsible low financial value but high transaction volumes.

In addition, the chapter characterized the cross-border financial flows network using node degrees. A hierarchical clustering procedure was used to derive the approximate degree distribution of the resident node set. The results were presented in a dendrogram and a distribution bar chart.

Chapter 7: Discussion

7.1 Introduction

This research proposed data mining of cross-border financial flows using a network theoretic approach to solve the money laundering problem. Networks science provides useful tools such as visualization and statistical properties such as degree distributions, assortativity, and centrality measures to help understand real-world phenomena in various fields. The research focused primarily on the illegal transfer of funds across country borders, central problem to regulatory organizations, financial institutions, and law enforcement agencies.

The research developed several analytical tools, leveraging advances in technology to enable data processing and analysis of large multi-dimensional data sets comprising cross-border financial flows while circumventing information privacy concerns.

This chapter summarizes the research objectives, discusses the findings, the contributions to knowledge, the limitations, and the areas of further research.

7.2 Summary of the research objectives and a discussion of findings

Money laundering is a global problem, which devastates economies around the world. Criminal activities such as illegal transfer of funds across country borders, exploitation of mineral resources, organized crime, terrorist financing, drugs counterfeiting, corruption, and fraud in international trade have devastating impacts on the affected communities.

The money laundering process starts with the criminal activity that gives rise to crime proceeds, such as bribery, drug trafficking, tax evasion, and corrupt business practices. The money launder seeks to obscure the origins of illegally obtained money by passing it through the banking system or business transactions. This research focused on analyzing international business transactions posing international money laundering risks for financial institutions, regulatory organizations, and law enforcement agencies. The research leveraged advanced technology to fortify the regulatory restraints of AML.

In Chapter 2, the research outlined two main channels for measuring illicit financial flows, i.e., deliberate manipulation of customs invoices on external trade and leakages from the balance of payments (GFI, 2015), also known as the World Bank Residual Method (World Bank, 1985). Trade misinvoicing is trade-based money laundering, widely thought to be the largest component of illicit financial flows. However, the two channels used to estimate illicit financial flows still fall short of measuring all the unrecorded flows due to the lack of bilateral trade data on services and the secretive nature of such flows.

Chapter 2 also discussed other illicit financial flows channels that are not easy to measure using the available economic data due to their hidden nature. This research focused on analyzing transaction patterns between residents and non-residents to enhance the surveillance of cross-border financial transactions for regulatory organizations, financial institutions, and law enforcement agencies.

Chapter 3 provided the background on the fundamental concepts and definitions from mathematical graph theory and an overview of some of the essential properties of graph structures, mostly attributable to the work done by König (König, 1936). Hence, it laid a solid mathematical foundation for this research, which enabled the introduction of the directed and weighted bipartite as a model for cross-border financial flows and the network's characterization using degree distributions in subsequent chapters.

The research developed the symmetric-key encryption algorithm to circumvent information privacy concerns when analyzing the multi-dimensional data sets comprising cross-border financial flows. Chapter 4 presented the proposed encryption algorithm and discussed its advantages and disadvantages. In addition to encryption, the algorithm computed the necessary statistical measures of networks such as node degree and weights discussed in Chapter 3, thereby accomplishing the research’s first stated objective.

Chapter 5 provided a formal definition of the network structure of cross-border financial flows using a directed and weighted bipartite graph with dual weights, representing the monetary value and volume of transactions. The study leveraged advances in technology to construct the network structure. Presentation of the transactions in a network form enabled efficient use of business intelligence tools and statistical software packages to compute basic statistics and visualizations at resident and non-resident node level.

The proposed network structure enables the regulatory organizations to plan inspections of both authorized dealers and resident/non-resident transactions optimally by focusing on suspicious transactions using the FATF's notions instead of relying on tip-offs and random sampling transactions. The network weights provide transaction aggregates from various authorized dealers for each resident/non-resident node. Hence, the defined network structure strengthens the implementation and enforcement of AML by financial institutions and regulatory organizations. Chapter 2 discussed the compliance and risk management environment encompassing the risk-based approach, which is the appropriate deployment area for the proposed cross-border financial flows model.

Chapter 6 proposed a network measure to identify suspicious activity and criminal conduct per FATF recommendations, advocating using advanced technology for regulatory purposes. The proposed centrality measure for directed and weighted bipartite networks expands the available scientific tools for mining cross-border financial transactions data. The centrality measure addresses the question, "Which are the most important or central nodes in the cross-border financial flows network?".

Chapter 6 further proposed the characterization of the cross-border financial flows network using the distribution of node degrees, often cited as the most fundamental of network properties and the defining characteristic of network structure. The cross-border financial flows network developed in this research comprises four-degree distributions, two for each node-set. Characterization reduces the degree distributions from four to two and avails additional variables that can facilitate the design of a sampling rule for inspecting transactions, as discussed in Chapter 6.

The research used a real data set comprising remittances transactions to illustrate the symmetric-key encryption algorithm and cross-border financial flows' network structure. The hypothetical data set also provided a mechanism for illustrating the symmetric-key encryption algorithm step by step and the computation of the centrality measure.

The research used SAS® Visual Analytics software to visualize the network structure and a SAS® computer program to develop the encryption code. The results showed that the proposed centrality standard could illuminate the network by providing a quantitative measure of node importance, which is a significant finding in the battle against the illegal cross-border transfer of funds.

The network theoretic approach adopted in this research complements the regulatory guidelines by enriching the data environment with non-classical analytical tools such as network structure, centrality measure, and characterization variables for identifying suspicious transactions. Hence, providing an answer to the following research question:

Can the use of networks' statistical properties to analyze transactional patterns between residents and non-residents in cross-border financial flows data enable identifying suspicious activity and criminal conduct to detect and impede the illegal transfer of funds across country borders?

7.3 Contributions and significance of research

The algorithm's application areas are vast due to the rapid increase in digital services, which enable people to use banking, transportation, payments, healthcare, navigation, shopping, and healthcare services. The digital services create sizeable multi-dimensional data sets comprising private and confidential information. The analysis of such data sets often triggers information privacy concerns. Like most symmetric-key encryption techniques, the technique is simple to implement and fast in execution.

The distinguishing factor between the proposed algorithm and other existing encryption algorithms discussed in Chapter 4 is the use of temporary variables to indicate By-groups' start and end. The second part of the algorithm leveraged the temporary variables to compute the network weights, thereby enabling the constructing of the cross-border financial flows network structure.

While the study used the algorithm to encrypt only two variables, the technique is applicable in cases involving several variables. The minimum requirement for applying the technique is that it must be possible to sort and group the multi-dimensional data sets to enable BY-group processing. Like in any other encryption algorithm, data quality challenges can potentially limit its effectiveness.

The research defined the cross-border financial flows network using a directed and weighted bipartite graph with dual weights. The two network weights represented the total monetary amount and the total number of transactions between each resident and non-resident node pairs. The network model enabled non-classical statistical measures to study transaction patterns between the resident and non-resident nodes. The network was represented on a computer using the adjacency matrix and visualized using SAS ® software.

The research proposed a measure of network centrality based on matrix multiplication, allocating centrality points based on each node's weights and degrees. Like other measures of centrality and assortativity, it does not rely on the centrality's absolute magnitude, only about which nodes have high or low centrality values. Hence, it is possible to compare its performance against eigenvector centrality or other measures such as the Katz centrality, which addresses eigenvector centrality's limitations (Katz, 1953).

The significance of the proposed centrality measure and network characterization using node degrees is that they both provide the ability to measure the importance of a node based on the node's contribution in the network in addition to the node's degrees. Therefore, the cross-border financial flows network structure and its statistical properties obtained in this research can enhance regulatory compliance and risk management practices in regulated entities, thereby promoting these entities' safety and soundness.

7.4 Limitations of the study and areas of further research

This research proposed using network science to understand the workings of cross-border financial flows and, in many ways, given rise to more questions that require further research. This section presents a list of topics for future investigation.

7.4.1 Performance of the proposed symmetric-key encryption algorithm

Chapter 4 discussed several privacy-preserving approaches and proposed the symmetric-key encryption algorithm, which made efficient use of computer memory by using temporary variables and the multi-dimensional data set's group structure. The chapter outlined both advantages and disadvantages of the proposed algorithm. However, the algorithm’s performance comparisons against the other encryption techniques discussed in this research is a subject for further investigation. The algorithm's use appears limited due to the requirement that the data set comprises multiple dimensions.

7.4.2 Limitations of the cross-border financial flows network

The directed and weighted bipartite graphs have been used recently as models for complex networks. This research introduced the directed and weighted bipartite graph model with two weights representing transaction volumes and amounts. Therefore, further research must develop an understanding of network measures for directed and dual weighted networks. Areas such as community detection and network assortativity are subjects of further investigation in directed and dual weighted networks.

The research used the Ward clustering technique to characterize the cross-border financial flows network. The chosen hierarchical clustering method performs well in small networks but underperforms with increasing data size. Using other classification methods suitable for large data sets can benefit the research.

7.4.3 Limitations of the proposed network centrality measure

The proposed network centrality measure was significant in identifying the transaction patterns between residents and non-resident nodes in the cross-border financial flows network. The highly connected nodes identified using the measure require further analysis and inspection. However, the measure was applied to a small data set, making it challenging to observe the computational inefficiencies and complexities of execution. Performing the same computations involving much larger data sets can shed some light on the proposed centrality measure's performance and provide insights into its limitations.

The proposed centrality measure's statistical significance can be evaluated based on the discrepancy between the random directed and weighted bipartite network and the cross-border financial flows network. The research discussed the Configuration model and the Curveball algorithm for generating the random network with the same order and size, and the same degree distribution or sequence as the observed cross-border financial flows network. The evaluation of the centrality measure's statistical significance is an area requiring further research.

Further research on comparative performance studies of the proposed centrality measure against other available centrality measures for directed and weighted networks will undoubtedly provide beneficial information.

7.4.4 Policy response to research findings

The research emphasized the development of network science tools for financial institutions, regulatory organizations, and law enforcement agencies to combat money laundering. However, the study did not provide policy recommendations suitable for integrating the proposed methodology into the current regulatory landscape. The use of the FATF's recommendations served to indicate the relevant application area of this study. Implementation of the proposed network theoretic approach requires further analysis from a regulatory policy front.

7.5 Summary

Chapter seven provided a summary of the research objectives, discussed the significance of the research findings and the contributions to knowledge. Also, the chapter discussed the limitations of the research along with the areas of further research.

Unexpectedly, the network structure and the centrality enriched the regulatory compliance and risk management environments in financial institutions with additional data to enable RegTech and SupTech. SupTech is useful for regulatory organizations to plan their inspection of regulated entities optimally and to identify regulatory non-compliance in advance. RegTech enables the regulated entities to effectively monitor risk factors, as discussed in Chapter 2 of this research.

Chapter 8: Concluding remarks

8.1 Introduction

This chapter concludes the thesis. The next section presents the concluding remarks, which notes the main research findings. The summary excludes the first three introductory chapters of the thesis, which do not contain new research material. The last section concludes with summarized suggestions for future research.

8.2 Summary of the research findings

This research primarily focused on developing network tools for analyzing cross-border financial transactions to assist financial institutions, regulatory organizations, and law enforcement agencies in detecting and impeding the illegal transfer of funds across country borders.

Chapter 2 discussed three different approaches used by criminal networks to launder their criminal proceeds. The research focused on the first method, which involves transferring value through the global financial system using wire transfers and commercial transactions. The study demonstrated the effectiveness of the proposed network theoretic approach for modeling cross-border financial transactions using a sample of real data set drawn from the South African database of international financial transactions. This section summarizes the main conclusions.

8.2.1 Privacy-preserving symmetric-key encryption algorithm

Chapter 4 provided an overview of both cryptographic and non-cryptographic techniques for preserving the privacy of personally identifiable information in statistical databases. Most importantly, the chapter proposed and developed the symmetric-key encryption algorithm.

The research presented the two components of the algorithm. The first part performed algebraic operations on the multi-dimensional data set for encryption purposes. The second part computed the edge weights to facilitate constructing the cross-border financial flows network from international financial records.

The algorithm proved to execute fast due to its simplicity, thereby making it possible to encrypt large data sets. It used decryption operations like the Permutation Cipher, which is a lookup table. More massive data sets with multiple encryption variables complicate the algorithm's symmetric-key derivation, thereby improving its safety. The encrypted data set derived using the proposed encryption algorithm is less susceptible to linkage attacks since the algorithm does not provide statistical (demographic) data linked to individuals.

In contrast to the non-cryptographic techniques discussed in Chapter 4, the algorithm underperforms since it cannot encrypt a live statistical database. The algorithm's use is limited to multi-dimensional data sets only due to the data set's desired group structure. Also, the algorithm's safety depends on the database's security used to store the decryption key. Chapter 4 discussed the advantages and disadvantages of the proposed symmetric-key encryption algorithm.

In addition to preserving privacy in financial transactions, the algorithm's application areas can extend to patient records held by the healthcare system, salary records held by employers, investigation records held by the criminal justice system, and public institutions' motor vehicle registration information.

8.2.2 Network structure of cross-border financial flows

The study of directed graphs is a well-researched are called Network flow theory, introduced in Chapter 3 of this research. Network flow problems arise in many graph theory applications such as internet traffic, transportation systems, communication systems, road traffic flow, and power supply networks. In contrast, this research studied directed and weighted graphs often classified as financial networks, with a primary focus on the transaction patterns between residents and non-residents for regulatory compliance and risk management purposes.

The research developed the cross-border financial flows network using a directed and weighted bipartite graph with dual weights and leveraged advanced technology to construct and visualize its structure. The two network weights represented the total monetary amount and the total number of transactions between each resident and non-resident node pairs.

Chapter 5 provided a mathematical definition of the cross-border financial flows network. The research used SAS® Visual Analytics software for network visualization, making it easier to identify the network's highly connected nodes. Network visualization revealed that most nodes have one or two connections, while few nodes form part of largely connected sub-networks. However, due to the limitations of using the human eye to eyeball the network, the research suggested using the cross-border financial network's statistical properties to understand the structure further. Transaction sampling strategies focused on highly connected nodes can add value to the financial institutions' regulatory compliance departments and risk management functions instead of relying on transaction-specific triggers.

The network model enabled non-classical statistical measures to study transaction patterns between the resident and non-resident nodes in the database of international financial transactions. The proposed model fits both the SupTech and RegTech environment defined in Figure 1, as part of the AML systems, business intelligence, analytical systems. Also, the network structure can complement the built-in transaction-specific triggers in AML systems discussed in chapter 2, which enable active tracking of transactions under the risk-based approach guidelines and recommendations provided by the FATF for regulatory compliance purposes.

8.2.3 Centrality measure and network characterization

The research proposed a centrality measure based on the network's adjacency matrix multiplication to answer the question, "Which are the most important resident/non-resident nodes in the cross-border financial flows network?". The study used the hypothetical data set, comprising cross-border financial flows to demonstrate the proposed centrality measure. The results showed that the centrality measure effectively identified the resident nodes responsible for the most cross-border flows and those responsible for low financial value but high transaction volumes. These are significant results, which answer the research question.

The study used the hierarchical clustering procedure to derive the resident node-set's approximate degree distribution, thereby characterizing the cross-border financial flows network. The SAS® computer program presented the clustering procedure results using a dendrogram and a distribution bar chart. Both the dendrogram and the distribution chart provided additional data that enriched the cross-border transactional database.

The area of regulatory concern in the remittances industry is the detection and prevention of money laundering activity. The visualization depicted in Figure 5.4 showed residents who remitted funds to multiple non-residents, while the non-residents remitted funds to other residents. Identifying such transactions for money laundering investigations is critical under the risk-based approach. Other suspicious activity arises when multiple residents remit money to the same non-resident. Such cases often occur when criminals attempt to disguise the beneficial owner of the transactions.

After network characterization, the enriched data landscape and the derived centrality measure can enable classical statistical measures to understand the network structure further. For example, it can be possible to study transactional patterns for businesses and/or individuals by focusing directly on their activities as cluster members and using the normalized matrix derived in Chapter 6. Hence, the proposed solution to the research problem can enable RegTech and SupTech in regulatory organizations and financial institutions' compliance and risk management environments.

8.3 Directions for future research

Uses of the proposed symmetric-key encryption algorithm appear limited because the data set must comprise multiple dimensions to enable By-group processing, which is the cornerstone of the algorithm's design. Further research can extend the algorithm's uses to single-dimensional data sets to ascertain this fact. The algorithm's secrecy and performance comparisons against the other encryption techniques are also a subject for further investigation.

The research discussed the Configuration model and the Curveball algorithm for generating the random network with the same order and size, and the same degree distribution or sequence as the observed cross-border financial flows network. Future research can be directed at generating the random directed and weighted bipartite network using the Configuration model or the Curveball algorithm to evaluate the proposed centrality measure's statistical significance based on the discrepancy between the randomly generated network and the cross-border financial flows network.

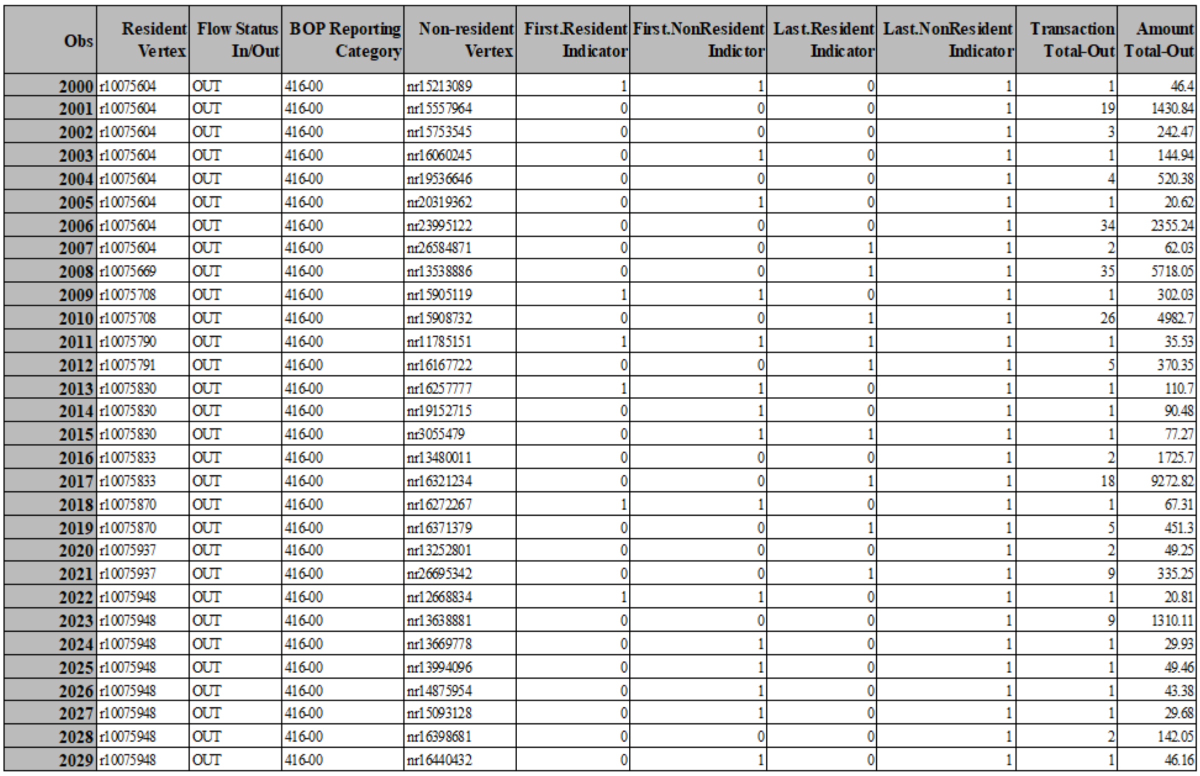
The research emphasized the development of network science tools for financial institutions, regulatory organizations, and law enforcement agencies to combat money laundering and terrorism financing. Hence, the research fell short of providing policy recommendations suitable for integrating the proposed methodology into the current regulatory and compliance landscape. The FATF's recommendations only served to indicate the application area of this study. Implementation of the proposed network theoretic approach requires further analysis from a regulatory policy front.

Appendix A:**Tables**

Table A.1: Remittances data set – variables description

Alphabetic List of Variables and Attributes						
#	**Variable**	**Type**	**Len**	**Format**	**Informat**	**Label**
9	CATEGORY	Char	6	$6.	$6.	BoP reporting category
13	DOLLARAMOUNT	Num	8			Amount of flow in United States Dollars
3	FLOW	Char	3	$3.	$3.	Outflow or Inflow
11	FLOWDATE	Num	8			Transaction date
10	LOCATIONCOUNTRY	Char	66	$66.	$66.	Destination country for flows
6	NR_COUNTRY	Char	66	$66.	$66.	Country of non-resident
5	NR_SURNAME_OR_LEGALNAME	Char	70	$70.	$70.	Non-resident name
1	REPORT_QUAL_NAME	Char	6	$6.	$6.	Report qualifier name - BoP customer
12	RESIDENTNAME	Char	200			Resident name and surname joined
8	R_NAME_OR_TRADINGNAME	Char	70	$70.	$70.	Resident name
7	R_SURNAME_OR_LEGALENTITYNAME	Char	70	$70.	$70.	Resident surname
2	STATUS	Char	2	$2.	$2.	Either original or cancelled transaction
4	TRNREFERENCE	Char	34	$34.	$34.	Transaction reference number

Table A.2: Encrypted sample of remittances data set using the symmetric-key encryption algorithm



Obs	Resident Vertex	Flow Status In/Out	BOP Reporting Category	Non-resident Vertex	First.Resident Indicator	First.NonResident Indictor	Last.Resident Indicator	Last.NonResident Indicator	Transaction Total-Out	Amount Total-Out
2000	r10075604	OUT	416-00	nr15213089	1	1	0	1	1	46.4
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
2010	r10075708	OUT	416-00	nr15908732	0	0	1	1	26	4982.7
2011	r10075790	OUT	416-00	nr11785151	1	1	1	1	1	35.53
2012	r10075791	OUT	416-00	nr16167722	0	0	1	1	5	370.35
2013	r10075830	OUT	416-00	nr16257777	1	1	0	1	1	110.7
2014	r10075830	OUT	416-00	nr19152715	0	1	0	1	1	90.48
2015	r10075830	OUT	416-00	nr3055479	0	1	1	1	1	77.27
2016	r10075833	OUT	416-00	nr13480011	0	0	0	1	2	1725.7
2017	r10075833	OUT	416-00	nr16321234	0	0	1	1	18	9272.82
2018	r10075870	OUT	416-00	nr16272267	1	1	0	1	1	67.31
2019	r10075870	OUT	416-00	nr16371379	0	0	1	1	5	451.3
2020	r10075937	OUT	416-00	nr13252801	0	0	0	1	2	49.25
2021	r10075937	OUT	416-00	nr26695342	0	0	1	1	9	335.25
2022	r10075948	OUT	416-00	nr12668834	1	1	0	1	1	20.81
2023	r10075948	OUT	416-00	nr13638881	0	0	0	1	9	1310.11
2024	r10075948	OUT	416-00	nr13669778	0	1	0	1	1	29.93
2025	r10075948	OUT	416-00	nr13994096	0	1	0	1	1	49.46
2026	r10075948	OUT	416-00	nr14875954	0	1	0	1	1	43.38
2027	r10075948	OUT	416-00	nr15093128	0	1	0	1	1	29.68
2028	r10075948	OUT	416-00	nr16398681	0	0	0	1	2	142.05
2029	r10075948	OUT	416-00	nr16440432	0	1	0	1	1	46.16

Table A.3: Cluster history table showing the last 50 vertex mergers

Cluster History							
Number of Clusters	Clusters Joined		Frequency	Semipartial R-Squared	R-Squared	Pseudo F Statistic	Pseudo t-Squared
50	r2417378	CL57	6	0	0.999	810,000	4.9
49	CL324	CL72	2,843	0	0.999	790,000	5,192.0
48	CL346	CL105	1,397	0	0.999	770,000	13,000.0
47	CL382	CL75	970	0	0.999	750,000	1,844.0
46	CL77	CL81	136	0	0.999	730,000	290.0
45	CL47	CL364	1,752	0.0001	0.999	710,000	1,254.0
44	CL80	CL95	7	0.0001	0.999	690,000	19.2
43	CL373	CL85	2,266	0.0001	0.999	670,000	20,000.0
42	CL76	CL63	518	0.0001	0.999	650,000	755.0
41	CL68	CL102	25	0.0001	0.999	630,000	53.9
40	CL64	CL101	408	0.0001	0.999	610,000	1,086.0
39	CL371	CL56	1,980	0.0001	0.998	590,000	2,236.0
38	CL71	CL73	621	0.0001	0.998	570,000	1,524.0
37	CL384	CL403	2,869	0.0001	0.998	560,000	
36	CL84	CL61	106	0.0001	0.998	550,000	210.0
35	CL55	CL38	1,532	0.0001	0.998	530,000	988.0
34	CL362	CL53	3,846	0.0001	0.998	520,000	6,063.0
33	CL74	CL51	825	0.0001	0.998	510,000	1,112.0
32	CL355	CL347	4,348	0.0001	0.998	490,000	
31	CL59	CL66	1,317	0.0002	0.998	480,000	3,739.0
30	CL320	CL49	5,987	0.0002	0.997	460,000	19,000.0
29	CL44	r1324529	8	0.0002	0.997	440,000	19.6
28	CL54	CL60	74	0.0002	0.997	420,000	231.0
27	CL52	CL58	370	0.0002	0.997	400,000	838.0
26	CL34	CL319	8,055	0.0002	0.997	390,000	12,000.0
25	CL39	CL50	1,986	0.0003	0.996	380,000	2,393.0
24	CL37	CL25	4,855	0.0003	0.996	360,000	2,534.0
23	CL45	CL48	3,149	0.0004	0.996	350,000	7,074.0
22	CL29	CL67	21	0.0004	0.995	330,000	24.7
21	CL40	CL42	926	0.0006	0.995	310,000	2,324.0
20	CL46	CL36	242	0.0007	0.994	290,000	834.0
19	CL28	CL41	99	0.0009	0.993	270,000	231.0
18	CL33	r2459199	826	0.001	0.992	250,000	4,491.0
17	CL18	CL31	2,143	0.001	0.991	230,000	1,541.0
16	CL30	CL32	10,335	0.0012	0.99	220,000	33,000.0
15	CL24	CL43	7,121	0.0013	0.988	210,000	9,501.0
14	CL23	CL35	4,681	0.0016	0.987	200,000	9,190.0
13	CL142	r243436	4	0.0024	0.984	180,000	3,121.0
12	CL21	CL27	1,296	0.0027	0.982	170,000	3,123.0
11	CL19	CL22	120	0.0044	0.977	150,000	255.0
10	CL16	CL26	18,390	0.0045	0.973	130,000	41,000.0
9	CL13	r2434297	5	0.0063	0.966	120,000	7.8
8	CL12	CL20	1,538	0.009	0.957	110,000	2,895.0
7	CL17	CL14	6,824	0.0098	0.948	100,000	14,000.0
6	CL10	CL15	25,511	0.0202	0.927	87,000	59,000.0
5	CL92	CL9	9	0.022	0.905	81,000	17.9
4	CL8	CL11	1,658	0.0313	0.874	78,000	2,562.0
3	CL7	CL4	8,482	0.0839	0.79	64,000	11,000.0
2	CL3	CL6	33,993	0.2165	0.573		41,000.0
1	CL2	CL5	34,002	0.5733	0		46,000.0

Table A.4: Cross-border financial transactions – the hypothetical data set

Resident_name	Transaction_date	Flow	NonResident_name	$_amount
Lynn	2018-01-16	Out	Joaquin	3,500
Elizabeth	2018-01-25	Out	Victoria	1,000
Christo	2018-02-07	Out	Mavis	500
Elizabeth	2018-02-14	In	Victoria	250
Lynn	2018-03-04	In	Victoria	100
Rosalia	2018-03-12	In	Sara	200
Christo	2018-03-12	In	Benjamin	2,500
Martina	2018-03-30	In	Benjamin	1,000
Martina	2018-03-31	Out	Benjamin	2,250
Rosalia	2018-04-18	Out	Mariana	250
Martina	2018-04-23	In	Benjamin	1,250
Rosalia	2018-05-05	In	Benjamin	1,000
Christo	2018-06-05	Out	Catalina	1,000
Michael	2018-06-10	Out	Matias	500
Elizabeth	2018-06-15	Out	Mariana	1,000
Elizabeth	2018-06-18	Out	Mariana	5,000
Rosalia	2018-07-02	In	Victoria	150
Linda	2018-07-11	Out	Diego	275
Lynn	2018-07-14	In	Mariana	750
Linda	2018-08-08	Out	Diego	275
Martina	2018-08-17	Out	Lucas	500
Rosalia	2018-09-02	Out	Victoria	150

Table A.5: Encryption key generated by DATA step processing

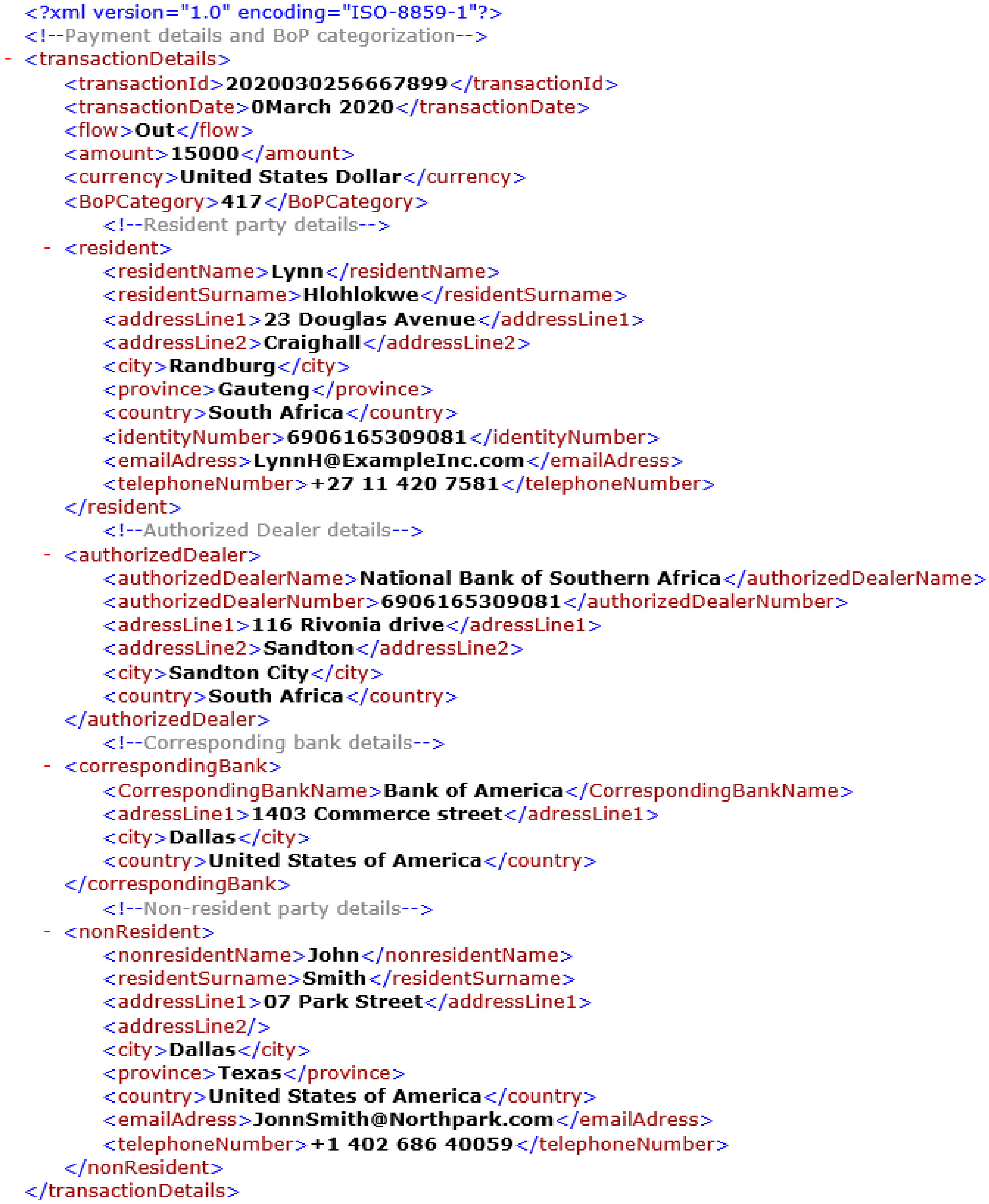
Resident name	Transaction date	Flow	Non-resident name	$_amount	Resident label	Non-resident label
Christo	12Mar2018	In	Benjamin	2,500	R1	NR1
Martina	23Apr2018	In	Benjamin	1,250	R13	NR1
Martina	31Mar2018	Out	Benjamin	2,250	R13	NR1
Martina	30Mar2018	In	Benjamin	1,000	R13	NR1
Rosalia	05May2018	In	Benjamin	1,000	R18	NR1
Christo	05Jun2018	Out	Catalina	1,000	R1	NR6
Linda	11Jul2018	Out	Diego	275	R8	NR7
Linda	08Aug2018	Out	Diego	275	R8	NR7
Lynn	16Jan2018	Out	Joaquin	3,500	R10	NR9
Martina	17Aug2018	Out	Lucas	500	R13	NR10
Elizabeth	15Jun2018	Out	Mariana	1,000	R4	NR11
Elizabeth	18Jun2018	Out	Mariana	5,000	R4	NR11
Lynn	14Jul2018	In	Mariana	750	R10	NR11
Rosalia	18Apr2018	Out	Mariana	250	R18	NR11
Michael	10Jun2018	Out	Matias	500	R17	NR15
Christo	07Feb2018	Out	Mavis	500	R1	NR16
Rosalia	12Mar2018	In	Sara	200	R18	NR17
Elizabeth	25Jan2018	Out	Victoria	1,000	R4	NR18
Elizabeth	14Feb2018	In	Victoria	250	R4	NR18
Lynn	04Mar2018	In	Victoria	100	R10	NR18
Rosalia	02Sep2018	Out	Victoria	150	R18	NR18
Rosalia	02Jul2018	In	Victoria	150	R18	NR18

Table A.6: Adjacency matrix representation of the cross-border financial flows network with dual weights

	R1	**R4**	**R8**	**R10**	**R13**	**R17**	**R18**	**NR1**	**NR6**	**NR7**	**NR9**	**NR10**	**NR11**	**NR15**	**NR16**	**NR17**	**NR18**
R1	0	0	0	0	0	0	0		(2,1500)						(1,500)		
R4	0	0	0	0	0	0	0						(2,6000)				(3,7000)
R8	0	0	0	0	0	0	0			(2,550)							
R10	0	0	0	0	0	0	0				(1,3500)						
R13	0	0	0	0	0	0	0	(2,2250)				(2,2750)					
R17	0	0	0	0	0	0	0							(1,500)			
R18	0	0	0	0	0	0	0						(1,250)				(2,400)
NR1	(1,2500)				(2,2250)		(1,1000)	0	0	0	0	0	0	0	0	0	0
NR6								0	0	0	0	0	0	0	0	0	0
NR7								0	0	0	0	0	0	0	0	0	0
NR9								0	0	0	0	0	0	0	0	0	0
NR10								0	0	0	0	0	0	0	0	0	0
NR11				(1,750)				0	0	0	0	0	0	0	0	0	0
NR15								0	0	0	0	0	0	0	0	0	0
NR16								0	0	0	0	0	0	0	0	0	0
NR17							(2,1200)	0	0	0	0	0	0	0	0	0	0
NR18		(1,250)		(2,850)			(3,1350)	0	0	0	0	0	0	0	0	0	0

Appendix B: **Computer Code**

B.1 : Example of an XML document for reporting cross-border financial transactions



```
  <?xml version="1.0" encoding="ISO-8859-1"?>
  <!--Payment details and BoP categorization-->
- <transactionDetails>
      <transactionId>2020030256667899</transactionId>
      <transactionDate>0March 2020</transactionDate>
      <flow>Out</flow>
      <amount>15000</amount>
      <currency>United States Dollar</currency>
      <BoPCategory>417</BoPCategory>
          <!--Resident party details-->
    - <resident>
          <residentName>Lynn</residentName>
          <residentSurname>Hlohlokwe</residentSurname>
          <addressLine1>23 Douglas Avenue</addressLine1>
          <addressLine2>Craighall</addressLine2>
          <city>Randburg</city>
          <province>Gauteng</province>
          <country>South Africa</country>
          <identityNumber>6906165309081</identityNumber>
          <emailAdress>LynnH@ExampleInc.com</emailAdress>
          <telephoneNumber>+27 11 420 7581</telephoneNumber>
      </resident>
          <!--Authorized Dealer details-->
    - <authorizedDealer>
          <authorizedDealerName>National Bank of Southern Africa</authorizedDealerName>
          <authorizedDealerNumber>6906165309081</authorizedDealerNumber>
          <adressLine1>116 Rivonia drive</adressLine1>
          <addressLine2>Sandton</addressLine2>
          <city>Sandton City</city>
          <country>South Africa</country>
      </authorizedDealer>
          <!--Corresponding bank details-->
    - <correspondingBank>
          <CorrespondingBankName>Bank of America</CorrespondingBankName>
          <adressLine1>1403 Commerce street</adressLine1>
          <city>Dallas</city>
          <country>United States of America</country>
      </correspondingBank>
          <!--Non-resident party details-->
    - <nonResident>
          <nonresidentName>John</nonresidentName>
          <residentSurname>Smith</residentSurname>
          <addressLine1>07 Park Street</addressLine1>
          <addressLine2/>
          <city>Dallas</city>
          <province>Texas</province>
          <country>United States of America</country>
          <emailAdress>JonnSmith@Northpark.com</emailAdress>
          <telephoneNumber>+1 402 686 40059</telephoneNumber>
      </nonResident>
  </transactionDetails>
```

B.2 : SAS® code for encryption of the hypothetical data set.

```
/* the encryption phase of the algorithm*/

proc sort data = ExampleDataset;
    By Resident_name NonResident_name;
run;

data CrossBorderData;
  set ExampleDataset;
    by Resident_name;
    retain resident_label;
      if FIRST.Resident_name then
      Resident_label = CAT(‘R’,_N_);
run;

proc sort data = CrossBorderData;
    By NonResident_name Resident_name;
run;

data CrossBorderDataFinal;
  set CrossBorderData;
    by NonResident_name;
    retain NonResident_label;
      if FIRST.NonResident_name then
      NonResident_label = CAT(‘NR’,_N_);
run;

data CrossBorderDataFinal;
  set CrossBorderData;
    by NonResident_name;
    retain NonResident_label;
      if FIRST.NonResident_name then
      NonResident_label = CAT(‘NR’,_N_);

/* display the encrypted data set*/

data CrossBorderAnalysis;
    retain Resident_label Flow_Amount NonResident_label;
      set CrossBorderDataFinal (drop = Transaction_date Resident_name NonResident_name);
 run;

/* the data reorganization phase and computation of descriptive statistics*/

proc sort data = CrossBorderAnalysis;
    by Resident_label NonResident_label;
  run;

  data NetworkOUTdegrees;
```

```
    set CrossBorderAnalysis;
      where Flow = 'Out';
           by Resident_label NonResident_label;
             if First.Resident_label = 1 then do;
                trans_count = 0;
                total_amount = 0;
             end;
             trans_count + 1;
             total_amount + amount;
             if Last.NonResident_label = 1 then output;
run;

data NetworkINdegrees;
    set CrossBorderAnalysis;
      where Flow = 'In';
         by Resident_label NonResident_label;
           if First.Resident_label = 1 then do;
           trans_count = 0;
           total_amount = 0;
        end;
        trans_count + 1;
        total_amount + amount;
        if Last.NonResident_label = 1 then output;
run;

data NetworkOUTdegrees;
    set CrossBorderAnalysis;
      where Flow = 'Out';
         by Resident_label NonResident_label;
           if First.Resident_label = 1 then do;
           trans_count = 0;
           total_amount = 0;
        end;
        trans_count + 1;
        total_amount + amount;
        if Last.NonResident_label = 1 then output;
run;
```

B.3: SAS® code for construction of the directed and weighted bipartite network

```
Libname Articles 'C:\Research\Input_Data';
Libname Outicles 'C:\Research\Output_Data';

/* Prepare the input data files for processing */

%Let CrossborderData = Articles.File1_2015 Articles.File2_2015 Articles.File3_2015
     Articles.File4_2015 Articles.File5_2015 Articles.File6_2015
     Articles.File7_2015 Articles.File8_2015 Articles.File9_2015
     Articles.File10_2015 Articles.File11_2015 Articles.File12_2015
     Articles.File13_2015 Articles.File14_2015 Articles.File15_2015 ;

/* Read the input data files into the model, create the resident name variable,
   format the valuedate of the flow and convert all the ZAR amounts to USD amounts */

DATA Articles.DollarizeFlowAmounts;
  SET &CrossborderData;
    FLOWDATE = INPUT(VALUEDATE,yymmdd10.);
    RESIDENTNAME                    =                    CATX('
',R_NAME_OR_TRADINGNAME,R_SURNAME_OR_LEGALENTITYNAME);
    DROP VALUEDATE RANDVALUE;
       IF Year (FLOWDATE) = 2014 and Month (FLOWDATE) = 01 THEN DollarAmount =
RandValue * 0.089936;
       IF Year (FLOWDATE) = 2014 and Month (FLOWDATE) = 02 THEN DollarAmount =
RandValue * 0.092980;
       IF Year (FLOWDATE) = 2014 and Month (FLOWDATE) = 03 THEN DollarAmount =
RandValue * 0.095065;
       IF Year (FLOWDATE) = 2014 and Month (FLOWDATE) = 04 THEN DollarAmount =
RandValue * 0.094986;
       IF Year (FLOWDATE) = 2014 and Month (FLOWDATE) = 05 THEN DollarAmount =
RandValue * 0.094571;
       IF Year (FLOWDATE) = 2014 and Month (FLOWDATE) = 06 THEN DollarAmount =
RandValue * 0.094092;
       IF Year (FLOWDATE) = 2014 and Month (FLOWDATE) = 07 THEN DollarAmount =
RandValue * 0.093435;
       IF Year (FLOWDATE) = 2014 and Month (FLOWDATE) = 08 THEN DollarAmount =
RandValue * 0.093736;
       IF Year (FLOWDATE) = 2014 and Month (FLOWDATE) = 09 THEN DollarAmount =
RandValue * 0.088519;
       IF Year (FLOWDATE) = 2014 and Month (FLOWDATE) = 10 THEN DollarAmount =
RandValue * 0.090702;
       IF Year (FLOWDATE) = 2014 and Month (FLOWDATE) = 11 THEN DollarAmount =
RandValue * 0.090394;
       IF Year (FLOWDATE) = 2014 and Month (FLOWDATE) = 12 THEN DollarAmount =
RandValue * 0.086575;
       IF Year (FLOWDATE) = 2015 and Month (FLOWDATE) = 01 THEN DollarAmount =
RandValue * 0.085848;
```

```
       IF Year (FLOWDATE) = 2015 and Month (FLOWDATE) = 02 THEN DollarAmount =
RandValue * 0.085755;
       IF Year (FLOWDATE) = 2015 and Month (FLOWDATE) = 03 THEN DollarAmount =
RandValue * 0.082355;
       IF Year (FLOWDATE) = 2015 and Month (FLOWDATE) = 04 THEN DollarAmount =
RandValue * 0.083930;
       IF Year (FLOWDATE) = 2015 and Month (FLOWDATE) = 05 THEN DollarAmount =
RandValue * 0.082320;
       IF Year (FLOWDATE) = 2015 and Month (FLOWDATE) = 06 THEN DollarAmount =
RandValue * 0.082305;
       IF Year (FLOWDATE) = 2015 and Month (FLOWDATE) = 07 THEN DollarAmount =
RandValue * 0.078869;
       IF Year (FLOWDATE) = 2015 and Month (FLOWDATE) = 08 THEN DollarAmount =
RandValue * 0.075403;
       IF Year (FLOWDATE) = 2015 and Month (FLOWDATE) = 09 THEN DollarAmount =
RandValue * 0.072166;
       IF Year (FLOWDATE) = 2015 and Month (FLOWDATE) = 10 THEN DollarAmount =
RandValue * 0.072291;
       IF Year (FLOWDATE) = 2015 and Month (FLOWDATE) = 11 THEN DollarAmount =
RandValue * 0.069009;
       IF Year (FLOWDATE) = 2015 and Month (FLOWDATE) = 12 THEN DollarAmount =
RandValue * 0.064601;
RUN;

DATA Articles.CrossBorder_Out ;
  SET Articles.DollarizeFlowAmounts;
          WHERE STATUS = 'OT' AND FLOW = 'OUT';
RUN;

DATA Articles.CrossBorder_In  ;
  SET Articles.DollarizeFlowAmounts;
          WHERE STATUS = 'OT' AND FLOW = 'IN';
RUN;

/* Invoking automatic variables for outward transactions aggregation */

PROC SORT DATA = Articles.CrossBorder_Out;
     BY RESIDENTNAME NR_SURNAME_OR_LEGALNAME;
RUN;

DATA Articles.OutDegree_Parametarization;
  SET Articles.CrossBorder_Out (WHERE = (Category IN ('416-00' '417-00')));
    BY RESIDENTNAME NR_SURNAME_OR_LEGALNAME;
      IF First.RESIDENTNAME = 1 THEN First_Res_Indicator = 1;
        ELSE First_Res_Indicator = 0;
      IF First.NR_SURNAME_OR_LEGALNAME = 1 THEN First_NonRes_Indicator = 1;
        ELSE First_NonRes_Indicator = 0;
      IF Last.RESIDENTNAME = 1 THEN Last_Res_Indicator = 1;
        ELSE Last_Res_Indicator = 0;
      IF Last.NR_SURNAME_OR_LEGALNAME = 1 THEN Last_NonRes_Indicator = 1;
```

```
        ELSE Last_NonRes_Indicator = 0;
RUN;

PROC SORT data = Articles.OutDegree_Parametarization;
   BY RESIDENTNAME NR_SURNAME_OR_LEGALNAME;
RUN;

/* First level aggregation using indicator variables */

DATA Articles.OutDegree_Comp;
  SET Articles.OutDegree_Parametarization;
    BY RESIDENTNAME NR_SURNAME_OR_LEGALNAME;
      IF First_NonRes_Indicator = 1 THEN DO;
        Trans_CountOut = 0;
        Total_DollarAmountOut = 0;
      END;
        Trans_CountOut + 1;
        Total_DollarAmountOut + DollarAmount;
      IF Last_NonRes_Indicator = 1 THEN OUTPUT;
RUN;

/* Computation of vertex out-degree for the resident vertex */

PROC SORT DATA = Articles.OutDegree_Comp;
     BY RESIDENTNAME NR_SURNAME_OR_LEGALNAME;
RUN;

DATA Outicles.Out_DegreeFinal;
  SET Articles.OutDegree_Comp;
    BY RESIDENTNAME NR_SURNAME_OR_LEGALNAME;
      IF First_Res_Indicator = 1 THEN DO;
        Out_degree = 0;
        Tot_TransCountOut = 0;
        Tot_DollarAmountOut = 0;
      END;
        Tot_TransCountOut + Trans_CountOut;
        Tot_DollarAmountOut + Total_DollarAmountOut;
        Out_degree + 1;
      IF Last_Res_Indicator = 1 THEN OUTPUT;
RUN;

/* Invoking automatic variables for inward transactions aggregation */

PROC SORT DATA = Articles.CrossBorder_In;
     BY RESIDENTNAME NR_SURNAME_OR_LEGALNAME;
RUN;

DATA Articles.InDegree_Parametarization;
  SET Articles.CrossBorder_In (WHERE =(Category IN ('416-00' '417-00')));
    BY RESIDENTNAME NR_SURNAME_OR_LEGALNAME;
   IF First.RESIDENTNAME = 1 THEN First_Res_Indicator = 1;
    ELSE First_Res_Indicator = 0;
   IF First.NR_SURNAME_OR_LEGALNAME = 1 THEN First_NonRes_Indicator = 1;
    ELSE First_NonRes_Indicator = 0;
   IF Last.RESIDENTNAME = 1 THEN Last_Res_Indicator = 1;
    ELSE Last_Res_Indicator = 0;
   IF Last.NR_SURNAME_OR_LEGALNAME = 1 THEN Last_NonRes_Indicator = 1;
    ELSE Last_NonRes_Indicator = 0;
RUN;

PROC SORT DATA = Articles.InDegree_Parametarization;
  BY RESIDENTNAME NR_SURNAME_OR_LEGALNAME;
RUN;

DATA Articles.InDegree_Comp;
 SET Articles.InDegree_Parametarization;
  BY RESIDENTNAME NR_SURNAME_OR_LEGALNAME;
   IF First_NonRes_Indicator = 1 THEN DO;
    Trans_CountIn = 0;
    Total_DollarAmountIn = 0;
   END;
    Trans_CountIn + 1;
    Total_DollarAmountIn + DollarAmount;
   IF Last_NonRes_Indicator = 1 THEN OUTPUT;
RUN;

PROC SORT data = Articles.InDegree_Comp;
  BY RESIDENTNAME NR_SURNAME_OR_LEGALNAME;
RUN;

DATA Outicles.In_DegreeFinal;
 SET Articles.InDegree_Comp;
  BY RESIDENTNAME NR_SURNAME_OR_LEGALNAME;
   IF First_Res_Indicator = 1 THEN DO;
    In_degree = 0;
    Tot_TransCountIn = 0;
    Tot_DollarAmountIn = 0;
   END;
    Tot_TransCountIn + Trans_CountIn;
    Tot_DollarAmountIn + Total_DollarAmountIn;
    In_degree + 1;
   IF Last_Res_Indicator = 1 THEN OUTPUT;
RUN;


/* Prepare output data for match-merging by sorting*/

DATA Outicles.Bipart_NetworkSubsetOut ;
 SET Outicles.Out_DegreeFinal (WHERE =(RESIDENTNAME ne " "));
RUN;
```

```
PROC SORT data = Outicles.Bipart_NetworkSubsetOut;
   BY RESIDENTNAME NR_SURNAME_OR_LEGALNAME;
RUN;


DATA Outicles.Bipart_NetworkSubsetIn ;
  SET Outicles.In_DegreeFinal (WHERE =(RESIDENTNAME ne " "));
RUN;

PROC SORT DATA = Outicles.Bipart_NetworkSubsetIn;
   BY RESIDENTNAME NR_SURNAME_OR_LEGALNAME;
RUN;

/* Perform match-merging by sorting*/

DATA Outicles.Bipart_Network (Keep = Nr_Surname_or_Legalname Residentname
Out_degree In_degree Tot_TransCountOut Tot_DollarAmountOut Tot_TransCountIn
Tot_DollarAmountIn) ;
  MERGE Outicles.Bipart_NetworkSubsetOut Outicles.Bipart_NetworkSubsetIn;
    BY RESIDENTNAME;
      IF In_degree = " " then Do;
        Tot_TransCountIn = 0;
        Tot_DollarAmountIn = 0;
        In_degree + 0;
      END;
      IF Out_degree = " " then Do;
        Tot_TransCountOut = 0;
        Tot_DollarAmountOut = 0;
        Out_degree + 0;
      END;
RUN;

/* Relabelling the Resident Vertex Set for reporting results and producing the final Bipartite
Network */

DATA Outicles.Bipart_Network_Final;
  RETAIN Res_Vertex Residentname Out_degree In_degree Tot_DollarAmountOut
Tot_DollarAmountIn NetFlow Tot_TransCountOut Tot_TransCountIn ;
    SET Outicles.Bipart_Network (Drop = Nr_Surname_or_Legalname);
      BY RESIDENTNAME ;
        Netflow = Tot_DollarAmountOut - Tot_DollarAmountIn;
        IDENTIFIER = _N_;
        Res_Vertex = Cat('r',Identifier);
        DROP IDENTIFIER ;
RUN;


Proc Print data = Outicles.Bipart_Network_Final (firstobs = 250000 obs = 250050);
```

```
Format Tot_DollarAmountOut: Comma10.2 Tot_DollarAmountIn: Comma10.2 NetFlow:
Comma10.2;
Run;
```

B.4: SAS® code for degree distribution and dendrogram based on remittances data set

```
ods graphics on;
ods html style = journal2;
ods escapechar='^';

/* Sub setting the bipartite network by out-degree for clustering purpose */

DATA Outicles.Network_data_Matrix;
 SET Outicles.Bipart_Network_Final;
 RUN;

/* Invoking the Ward clustering procedure to produce the cluster history and dendrogram */

PROC CLUSTER DATA = Outicles.Network_data_Matrix Method = ward Pseudo
outtree=Dendrogram noeigen noprint
      plots = (dendrogram (horizontal height=rsq));
  VAR Out_Degree In_Degree;
  ID Res_Vertex;
RUN;

goptions vsize=6in hsize=9.4in htext=10pt htitle=6pct;
axis1 order=(0 to 1 by 0.2);

/* Invoking the Tree procedure to produce the full cluster history without printing the
dendrogram */

PROC TREE DATA = Dendrogram out=Outicles.clust nclusters=12 noprint
  HAXIS = axis1 horizontal;
  HEIGHT _rsq_;
  COPY Out_Degree In_Degree ;
  ID Res_Vertex;
RUN;

/* Scatter plot of the clusters */

PROC SGPLOT DATA= Outicles.clust;
  SCATTER x = Out_Degree y = In_Degree / group=cluster ;
RUN;

PROC SORT DATA = Outicles.clust;
  BY clusname;
RUN;

PROC PRINT DATA = Outicles.clust;
  BY clusname;
  ID clusname;
RUN;
```

B.5: SAS® code for automated creation of text file for allocating nodes to clusters

```
/* Create a text file containing rules to execute the if-then-else statement for clusters */

DATA _NULL_;
  SET Outicles.clust end=last;
  FILE 'C:\Research\Output_Data\RVertexClus.txt';
    IF _n_ = 1 THEN put "Select (Res_Vertex);";
    put " when ('" Res_Vertex +(-1) "') Res_Vertex_clus = '" cluster +(-1) "';";
    IF last THEN do;
    put " otherwise Res_Vertex_clus = 'U';" / "end;";
    END;
RUN;


PROC MEANS DATA = Outicles.clust min max maxdec=0 nway;
   CLASS clusname;
     LABEL clusname = "Cluster Name";
     VAR cluster Out_Degree In_Degree;
RUN;

/* Produce the vertex degree distribution (bar chart) */

PROC SGPLOT DATA = Outicles.clust;
   VBAR  Cluster/ group=clusname groupdisplay = cluster barwidth = 0.7 stat = percent;
     XAXIS label ="Cluster";
             YAXIS label ="Percentage of vertices in a cluster ";
             LABEL clusname = "Cluster Name";
RUN;

ods graphics off;
```

Bibliography

Adam, N. R., & Wortmann, J. C. (1989, December). Security-Control Methods for Statistical Databases: A Comparative Study. *ACM Computing Surveys, 21*(4).

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

Albert, R., & Barabási, A. L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics, 4*.

Amini, H. (2010). Bootstrap Percolation in Living Neural Networks. *Journal of Statistical Physics, 141*(3), 459-475.

Amini, H., & Minca, A. (2013). Mathematical Modeling of Systemic Risk. *Advances in Network Analysis and Its Applications*, 3-26.

Anderberg, M. R. (1973). *Cluster analysis for applications.* New York: Academic Press.

Anderberg, M. R. (1973). *Cluster Analysis for applications.* New York: Academic Press.

Aronson, B., Yang, K.-C., Odabas, M., Ahn, Y. Y., & Perry, B. L. (2020). Comparing Measures of Centrality in Bipartite Social Networks: A Study of Drug Seeking for Opioid Analgesics.

Asratian, A. S., Denley, T. M., & Häggkvist, R. (1998). *Bipartite Graphs and their Applications.* Cambridge: Cambridge University Press.

Ayushi. (2010). A Symmetric Key Cryptographic Algorithm. *International Journal of Computer Applications (0975-887), 1*(15), 1-4.

Ba, H., & Huynh, T. (2018, June). Money laundering risk from emerging markets: the case of Vietnam. *Journal of Money Laundering Control, 21*(2), 385-401.

Babus, A. (2016). The formation of financial networks. *The RAND Journal of Economics, 47*(2), 239-272.

Bani Baker, S. I., & Al-Hamami, A. H. (2017). Novel Algorithm in Symmetric Encryption (NASE): Based on Feistel Cipher. *Proceedings of the 2017 International Conference on New Trends in Computing Sciences (ICTCS). 1*, pp. 191–196. Amman: IEEE. doi:10.1109/ICTCS.2017.54

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*.

Barbaro, M., & Zeller Jr, T. (2006, August 9). *The New York Times*. Retrieved March 27, 2020, from The New York Times Technology: https://www.nytimes.com/2006/08/09/technology/09aol.html

Berge, C. (1957). Two theorems in graph theory. *Proceedings of the National Academy of Science, 43*(9), 842-844.

Bhagwati, T. N. (1964). On the Underinvoicing Imports. *Bulletin of the Oxford University, Institute of Economics and Statistics, 26*, 389-397.

Bondy, J. A., & Murty, U. S. (2008). *Graph Theory* (1st ed.). London: Springer-Verlag .

Broeders, D., & Prenio, J. (2018). *Innovative technology in financial supervision (suptech) - the experience of early users.* Basel: Bank of International Settlements.

Busacker, R. G., & Saaty, T. L. (1965). *Finite graphs and networks: an introduction with applications.* New York: McGraw-Hill Book Company.

Buyya, R., Broberg, J., & Goscinski, A. (2010). *Cloud computing: Principles and paradigms.* New Jersey: Wiley.

Carreta, A., Farina, V., & Schwizer, P. (2017). Risk culture and banking supervision. *Journal of Financial Regulation and Compliance*, 209-226.

Carretta, A., Vincenzo, F., & Schwezer, P. (2017). Risk culture and banking supervision. *Journal of Financial Regulation and Compliance, 25*(2), 209-226.

Cassandras, C. G., & Lafortune, S. (2009). *Introduction to Discrete Event Systems.* New York: Springer.

Casterns, C. J. (2015). Proof of uniform sampling of binary matrices with fixed row sums and column sums for the fast Curveball algorithm. *Physical Review E, 91*(4).

Chang, P. H., Claessen, S., & Cumby, R. E. (1997). Conceptual and methodological issues in the measurement of capital flight. *International journal of Finance and Economics, 2*, 101-119.

Cios, K. J., & Moore, W. G. (2002, September). Uniqueness of medical data mining. *Artificial Intelligence in Medicine, 26*(1-2).

Claessens, S., & Naude, D. (1993). *Recent Estimates of Capital Flight, Policy, Research Working Papers; no. WPS 1186.* Washington D.C: World Bank.

Clauset, A., Newman, M. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70(6):066111.

Clifton, C., & Vaidya, J. (2004). Privacy-Preserving Data Mining: Why, How, and When. *IEEE Seurity & Privacy*, 19-26.

Cohen, J. (1987). *The Flow of Funds in Theory and Practice* (1 ed.). Dordrecht: Springer Netherlands.

Cox, D. (2011). *An introduction to money laundering deterrence.* United Kingdom: John Wiley & Sons Ltd.

Cox, D. (2014). *Handbook of Anti-Money Laundering.* West Sussex: John Wiley & Sons Ltd.

Csermely, P., London, A., Wu, L.-Y., & Uzzi, B. (2013). Structure and dynamics of core/periphery networks. *Journal of Complex Networks, 1*, 93-123.

Dehmer, M., Pickl, S., & Wang, Z. (2015). Recent developments in network analysis and their applications. *Systemics, cybernetics and informatics, 13*(4).

Dierckx, S. (2011). The IMF and capital controls: towards postneoliberalism? Reykjavik: Paper presented at 6th European Consortium for Political Research.

Diffie, W., & Hellman, M. E. (1976). New directions in cryptography. *IEEE Transactions on Information Theory*, 644-654.

Dinur, I., & Nissim, K. (2003). Revealing information while preserving privacy. *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 202-210). San Diego California: Association for Computing Machinery, New York, United States.

Dorogovtsev, S. N., & Mendes, J. F. (2003). *Evolution of Networks: From Biological Nets to the Internet and WWW.* Oxford: Oxford University Press.

Duda, R. O., & Hart, P. E. (1973). *Pattern Classification and Scene Analysis.* New York: Wiley.

Dwork, C. (2011). A firm foundation for private data analysis. *Communication of the ACM, 54*(1), 86-95. doi:Doi:10.1145/1866739.1866758

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science, 9*(3-4), 211-407.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. *Theory of Cryptography Conference* (pp. 265-284). New York: Springer.

Ebadi, H., Sands, D., & Schneider, G. (2015). Differebtial Privacy: Now it's Getting Personal. *ACM Sigplan Notces*, 69-81.

Edgar, N. G. (1959). Random Graphs. *Annals of Mathematical Statistics*, 1141-1144.

Edmonds, J., & Johnson, E. L. (1995). Arc Routing Problems, Part I: The Chinese Postman Problem. *Operations Research*, 231-242.

Edmonds, J., & Karp, R. M. (1972). Theoretical improvements in algorithmic efficiency for network flow problems. *JACM, 19,* 248-264.

Erdős, P., & Rényi, A. (1959). On Random Graphs I. *Publicationes Mathematicae*, 290-297.

Erlingsson, U., Pihur, V., & Korolova, A. (2014). RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. *ACM SIGSAC Conference on Computer and Communications* (pp. 1054–1067). Scottsdale, Arizona: ACM.

Ermann, L., & Shepelyansky, D. L. (2013). Ecological analysis of world trade. *Physics Letters A*(377), 250-256.

Estrada, E. (2012). *The Structure of Complex Networks: Theory and Applications.* New York: Oxford University Press Inc.

Euler, L. (1736). Solutio Prolematis ad Geometriam Situs Pertinentis. *Commentarii Academiae Scientarum Imperialis Petropolitane*, 128-140.

FATF. ((2012-2020)). *International Standards on Combating Money Laundering and the Financing of Terrorism & Proliferatio.* Paris: FATF. Retrieved from http://www.fatf-gafi.org/media/fatf/documents/recommendations/pdfs/fatf%20recommendations%202012.pdf

FATF. (2006, June 23). *Trade Based Money Laundering.* Retrieved from https://www.fatf-gafi.org: https://www.fatf-gafi.org/media/fatf/documents/reports/Trade%20Based%20Money%20Laundering.pdf

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

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

Financial Action Task Force. (2014). *Guidance for a risk-based approach: The Banking Sector.* Paris: FATF/OECD. Retrieved from https://protect-za.mimecast.com/s/crgICElv6pC5GRwiNofra?domain=fatf-gafi.org

Fleischner, H. (1991). *Eulerian Graphs and Related Topics (Part 1, Volume 1).* Amsterdam: North-Holland.

Ford, L. R., & Fulkerson, D. R. (1962). *Flows in Networks.* Princeton, N.J: Princeton University Press.

Fratzscher, M. (2012, February). *Capital Controls and Foreign Exchange Policy.* Frankfurt: European Central Bank. Retrieved April 9, 2017, from ecb.europa.eu

Freeman, L. C. (1977). A set of measures of centrality based upon betweenness. *Sociometry, 40*, 35-41.

Freeman, L. C. (1978). Centrality in Social Networks. *Social Networks*, 215-239.

Gallagher, K. P. (2011, February 20). *Regaining Control? Capital Controls and the Global Financial Crisis.* Retrieved April 9, 2017, from ase.tufts.edu

GFI. (2015, December 9). *Illicit Financial Flows from Developing Countries: 2004-2013.* Retrieved April 10, 2016, from http://www.gfintegrity.org/wp-content/uploads/2015/12/IFF-Update_2015-Final-1.pdf

GFI. (2017, April 13). *Global Financial Integrity.* Retrieved April 13, 2017, from http://www.gfintegrity.org/issue/trade-misinvoicing/

Gribkovskaia, I., Halskau, Ø., & Laporte, G. (2007). The bridges of Königsberg - A historical perspective. *Networks, 49*(3), 199-203.

Gross, J. L., & Yellen, J. (2006). *Graph Theory and ITS APPLICATIONS* (2nd ed.). Boca Raton, Florida: Chapman & Hall/CRC.

Guillaume, J.-L., & Latapy, M. (2006). Bipartite graphs as models of complex networks. *Physica A , 371*, 795-813.

Gulati, S. K. (1987). "A Note on Trade Misinvoicing". In D. R. Lessard, & J. Williamson (Eds.), *Capital Flight and Third World Debt* (pp. 68-78). Washington D.C: Institute for International Economics.

Hall, P. (1935). On representatives of subsets. *Journal of the London Mathematical Society*(10), 26-30.

Harary, F. (1979). The explosive growth of graph theory. *Annals of the New York Academy of Sciences, 328*(number 1), 5-11.

Hartigan, J. A. (1975). Statistical theory in clustering. *Journal of classification*(1), 63-76.

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*, 333-346.

Hierholzer, C. (1873). Uber die Moglichkeit, einen Linienzug ohne Widerholung und ohne Unterbrechnung zu unfahren. *Matematische Annalen VI*, 30-32.

Hosseini-Pozveh, M., Zamanifar, K., & Naghsh-Nilchi, A. R. (2017). A community-based approach to identify the most influential nodes in social networks. *Journal of Information Science, 43*(2), 204-220.

IMF. (2011). *Recent Experiences in Managing Capital Inflows - Cross-Cuting Themes and Possible Policy Framework.* Washington: International Monetary Fund.

IMF. (2013). *Sixth Edition of the IMF's Balance of Payments and International Investment Position Manual (BPM6).* Washington: International Monetary Fund.

Kaltenbrunner, A. (2016). Stemming the Tide: Capital Account Regulations in Developing and Emerging Countries. In P. Arestis, & M. Sawyer (Eds.), *Financial Liberalisation* (pp. 265-308). Leeds: Springer International Publishing.

KANG, S. (2018). Rethinking the global anti-money laundering regulations to deter corruption. *INTERNATIONAL AND COMPARATIVE LAW QUARTERLY, 67*(3), 695-720.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 39-43.

König, D. (1936). Theorie der Endlichen und Unendlichen Graphen. (Birkhauser, Trans.) *english translation,Birkhäuser, Boston 1990*.

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*.

Lancichinetti, A., & Fortunato, S. (2011). Limits of modularity maximization in community detection. *PHYSICAL REVIEW E*. doi:10.1103/PhysRevE.84.066122

Lazen, V. (2018, June 12). *RegTech, SupTech and risk-based supervision*. Retrieved from RegFin: http://www.regfin.cl/98710-2/

Lehmann, S., Schwartz, M., & Hansen, L. K. (2008). Biclique communities. *Physical Review E, 78*(1), 016108. doi:https://doi.org/10.1103/PhysRevE.78.016108

Li, H. (2018). Attack Vulnerability of Online Social Networks. *2018 37th Chinese Control Conference (CCC)* (pp. 1051-1056). Wuhan, China: IEEE.

Li, L., Alderson, D., Doyle, C. J., & Willinger, W. (2006). Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications. *Internet Mathematics, 2, Number 4,* 431-523.

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). New York: AAAI.

Lindert, P. (2002). *International Economics.* New Delhi: All India Traveller Bookseller.

Magnusson, D. (2009). The costs of implementing the anti-money laundering regulations in Sweden. *Journal of Money Laundering Control*, 101-112.

Marino, S., Zhou, N., Zhao, Y., Wang, L., Wu, Q., & Dino, I. (2019). HDDA: DataSifter: statistical obfuscation of electronic health records and other sensitive datasets. *Journal of Statistical Computation and Simulation*, 249-271.

Martin, J., Anders, L., Huseby, R. B., Geir, Å., & Johannes, L. (2020). Detecting money laundering transactions with machine learning. *Journal of Money Laundering Control, 23*(1), 173-186. doi:10.1108/JMLC-07-2019-0055

Menon, V. G. (2017). A Review on IoT based m-Health Systems for Diabetes. *International Journal of Computer Science and Telecommunications, 8*, 13-18.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*(2), 159-179.

Morris, C. H. (2019). *The Law of Financial Services Groups.* New York: Oxford University Press.

Nadakuditi, R. R., & Newman, M. J. (2012). Graph spectra and the detectability of community structure in networks. *Physical Review Letters*, 108(18):188701.

Nagurney, A., & Hughes, M. (1992). Financial flow of funds networks. *Networks, 22*(2), 141-161.

Nagurney, A., & Ke, K. (2001). Financial networks with intermediation. *Quantitative Finance, 1*, 441-451.

Narayanan, A., & Shmatikov, V. (2008). Robust De-anonymization of Large Sparse Datasets. *2008 IEEE Symposium on Security and Privacy* (pp. 111-125). Oakland, CA, USA: IEEE.

National Academy of Science. (2013). *Frontiers in Massive Data Analysis.* Washington, D.C: National Academies Press.

Newman, J. R. (1953). Leonhard Euler. *Scientific American, 189*, 66-70.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM Review, 45*, 167-256.

Newman, M. E. (2010). *Networks: An introduction.* Oxford: Oxford University Press.

Newman, M. J. (2003). Mixing patterns in networks. *Physical Review E,* 67(2):026126.

Newman, M. J. (2003). The Structure and Function of Complex Networks. *Society for Industrial and Applied Mathematics, 45*(2), 167-256.

Newman, M. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E,* 74.

Newman, M. J. (2010). *Networks: An introduction.* Oxford: Oxford University Press.

Newman, M. J. (2013). Spectral methods for community detection and graph partitioning. *Physical Review* , 88(1):010801.

Newman, M. J., & Peixoto, T. P. (2015). Generalized communities in networks. *Physical Review E,* 115(8):088701.

Newman, M., Barabasi, A. -L., & Watts, D. (2006). *The structure and Dynamics of Networks.* Princeton: Princeton University Press.

OECD. (2020). *OECD Code of Liberalisation of Capital Movements,*. Retrieved from www.oecd.org/investment/codes.htm.

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.

Pathak, M. A., Rane, S., & Raj, B. (2010). Multiparty differential privacy via aggregation of locally trained classifiers. *International Conference on Neural Information Processing Systems (NIPS) - Volume 2* (pp. 1876-1884). Vancouver: 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). Prague: IEEE.

Patnaik, I., Gupta, A. S., & Shah, A. (2012). Determinants of Trade Misinvoicing. *Open Economic Review, 23,* 891-910.

Peixoto, T. P. (2013). Eigenvalue Spectra of Modular Networks. *Physical Review E,* 111(9):098701.

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

Ravasz, E., & Barabási, A. L. (2003). Hirarchical organization in complex networks. *Physical Review E 67, 026112, 67.*

Rebecca, O. (2019, 11 12). *Record High Remittances Sent Globally in 2018.* Retrieved from Worldbank.org: https://www.worldbank.org/en/news/press-release/2019/04/08/record-high-remittances-sent-globally-in-2018

Rebeiro, C., Nguyen, P. H., Mukhopadhyay, D., & Poschmann, A. (2013). Formalizing the Effect of Feistel Cipher Structures on Differential Cache Attacks. *IEEE Transactions on Information Forensics and Security, 8*(8), 1274 - 1279. doi:10.1109/TIFS.2013.2267733

Rivest, R., Shamir, A., & Adleman, L. (1978). A Method for Obtaining Digital Signatures and Public Key Cryptosystems. *Communications of the ACM, 21*(2), 120-126.

Romesburg, C. H. (2004). *Cluster analysis for researchers.* North Carolina: Lulu Press.

Saha Ray, S. (2013). *Graph Theory with Algorithms and its Applications* (2 ed.). Berlin: Springer.

SAS Institute Inc. (2017). *SAS® Visual Analytics 7.4: User's Guide.* Cary, NC, USA: SAS Institute Inc.

SAS Institute Inc. (2001). *Step-by-step programming with Base SAS® software.* Cary, North Carolina, USA: SAS Institute Inc. Retrieved May 23, 2019, from https://support.sas.com/documentation/cdl/en/basess/58133/PDF/default/basess.pdf

SAS Institute Inc. (2010, May 10). *SAS® 9.2 Language Reference Concepts Second Edition.* Cary, North Carolina: SAS Institute Inc. Retrieved from Support.sas.com: https://support.sas.com/documentation/cdl/en/lrcon/62955/HTML/default/viewer.htm #a000961108.htm

Sciarra, C., Chiarotti, G., Laio, F., & Ridolfi, L. (2018). A change of perspective in network centrality. *Scientific Reports, 8*(1), 15269. doi:10.1038/s41598-018-33336-8

Sekgoka, P., & Adetunji, O. (2019). Cryptosystem for protecting personal information and data visualization using SAS Visual Analytics. *SAS Gobal Forum Proceedings.* Dallas-Texas: SAS Institute Inc. Retrieved from https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2019/3398-2019.pdf

Shannon, C. E. (1949, Oct). Communication theory of secrecy systems. *Bell Systems Technical Journal, 28*(4), 656-715.

Shokri, R., & Shmatikov, V. (2015). Privacy-preserving deep learning. *Proceedings of he 22nd ACM SIGSAC conference on computer and communications security* (pp. 1310-1321). Denver: ACM.

Silva, T. C., de Souza, S. R., & Tabak, B. M. (2016). Structure and dynamics of the global financial network. *Chaos, Solitons and Fractals, 88*, 218-234.

Silverman, B. S. (1986). *Density Estimation.* London: Chapman and Hall.

Smid, M. E., & Branstad, D. K. (1988). The data encryption standard: past and future. *Proceedings of the IEEE*, 550-559.

South African Reserve Bank. (2020). *Exchange Control Manual.* Retrieved from https://exchange4free.co.za/exchange-control-manual.html: https://exchange4free.co.za/exchange-control-manual.html

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. doi: https://doi.org/10.3390/sym11020293

Stallings, W. (1999). *Cryptography and Network Security: Principles and Practice* (2nd edition ed.). Prentice-Hall, Inc.

Stinson, D. R., & Paterson, M. (2018). *Cryptography theory and practice.* New York: Chapman and Hall/CRC; 4 edition.

Strona, G., Nappo, D., Boccacci, F., Fattorini, S., & San-Miguel-Ayanz, J. (2014). A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nature communications, 5*, 4114. doi:https:doi.org/10.1038/ncomms5114

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). Houston, TX.: IEEE Computer Society.

Sweeney, L. (2000). *Uniqueness of Simple Demographics in the U.S. Population.* Pittsburgh: Carnegie Mellon University, Laboratory for International Data Privacy.

Takabi, H., Joshi, J. B., & Ahn, G.-J. (2010). Security and privacy challenges in cloud computing environments. *IEEE Security and Privacy*, 24-31.

Takemoto, K., & Oosawa, C. (2012). Introduction to complex networks: measures, statistical properties, and models. In M. Dehmer, & S. C. Basak (Eds.), *Statistical and Machine Learning Approaches for Network Analysis* (pp. 45-75). New Jersey: John Wiley & Sons.

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. 6*, pp. 3453–3457. Guangzhou: IEEE. doi:10.1109/ICMLC.2005.1527539

Tang, J., Korolova, A., Bai, X., Wang, X., & Wang, X. (2017, September 11). *Privacy Loss in Apple's Implementation of Differential Privacy on MacOS 10.12.* Retrieved from arXiv.org: https://arxiv.org/abs/1709.02753

Teichmann, F. M. (2019). Recent trends in money laundering and terrorism financing. *Journal of Financial Regulation and Compliance*, 2-12.

Thore, S. (1969). Credit Networks. *Economica, 36*(141), 42-57.

Turner, J. E. (2011). *Money Laundering Prevention: Deterring, Detecting, and Resolving Financial Fraud.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Unger, B., & van der Linde, D. (2013). *Research handbook on money laundering.* Cheltenham: Edward Elgar Publishing.

United Nations. (2015, January 26). *United Nations Economic Commission for Africa.* Retrieved October 3rd, 2016, from www.uneca.org

United States. Bureau of Economic Analysis. (1990). *The Balance of Payments of The United States: Concepts, Data Sources, and Estimating Procedures.* Washington: US Government Printing Office.

Vaidya, J., & Clifton, C. (2004). Privacy-preserving data mining: why, how, when. *IEEE security and privacy*, 19-27.

van Duyne, P. (1994). Money-Laundering: Estimates in Fog. *Journal of Financial Crime*, 58-74.

Veldman, D. J. (1967). *Fortran programming for the behavioral sciences.* New York: Holt, Rinehart and Winston, 1967.

Walker, J. (1999). How Big is Global Money Laundering? *Journal of Money Laundering Control*, 25-37.

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). Switzerland: Springer, Cham.

Wang, S.-N., & Yang, J.-G. (2007). A Money Laundering Risk Evaluation Method Based on Decision Tree. *International Conference on Machine Learning and Cybernetics* (pp. 283–286.). Hong Kong: IEEE. doi:10.1109/ICMLC.2007.4370155

Ward, J. J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the Americal Statistical Association, 58*(301), 236-244.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature, 393*, 440-442.

Wheeler, D. J., & Needham, R. M. (2004). TEA, a tiny encryption algorithm. *Fast Software Encryption: Second International Workshop* (pp. 363–366). Leeuven: doi:10.1007/3-540-60590-8_29.

Wikipedia. (2016, September 02). *Wikipedia.* Retrieved September 02, 2016, from https://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg

Wilson, R. J. (1996). *Introduction to Graph Theory* (Fourth ed.). Edinburgh: Addison Wesley Longman Limited.

World Bank. (1985). *World Development Report.* Washington, DC: Oxford University Press.

World Bank. (2017). *Migration and Remittances: Recent Developments and Outlook.* Washington D.C.: World Bank. Retrieved November 07 , 2019, from http://pubdocs.worldbank.org/en/992371492706371662/MigrationandDevelopmentBrief27.pdf

Xiong H, L. Z., & 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). Sydney: IEEE.

Yang, K.-C., Aronson, B., & Ahn, Y.-Y. (2020). BiRank: Fast and Flexible Ranking on Bipartite. *Journal of Open Source Software, 5*(51), 2315.

Zhiguo, Z., Jingqin, S., & Liping, K. (2015). Measuring influence in online social nework based on the user-content bipartite graph. *Computers in Human Behavior, 52*, 184-189.