Social Network Analysis for Credit Risk Modeling

By

Davis Bundi Ntwiga

A thesis submitted in fulfillment of the requirements for the award of the degree of Doctor of Philosophy in Mathematical Finance in the School of Mathematics, University of Nairobi, Kenya.

July, 2016

Declaration and Approval

I, the undersigned, declare that this thesis contains my own work. To the best of my knowledge, no portion of this work has been submitted in support of an application for another degree or qualification of this University or any other institution of learning.

Signature: .... Date: ....

Davis Bundi Ntwiga

I80/83841/2012

Approval

This thesis has been under our supervision and has our approval for submission

Signature: .... Date: ....

Professor Patrick Weke

School of Mathematics, University of Nairobi

Signature: .... Date: ....

Professor Moses Manene

School of Mathematics, University of Nairobi

Signature: .... Date: ....

Dr. Mwaniki Ivivi

School of Mathematics, University of Nairobi

Dedication

This thesis is dedicated to:

My **wife**, Benta Lenah Bundi

My **son**, Ethan Mwenda Bundi

My **parents**, Evasio Ntwiga and Harriet Kagendo

My **siblings**

for the endless love, support and encouragement.

Acknowledgments

My special gratitude goes to the University of Nairobi through the Director, School of Mathematics and Principal, College of Biological and Physical Sciences for the financial support towards this Doctorate degree.

I am grateful to the Director, School of Mathematics for the support in terms of reduced teaching duties.

My Supervisor Professor Patrick Weke for the time, effort and guidance through the journey of this research. I appreciate Professor Manene and Dr. Mwaniki for their support.

My colleagues, Dr. Ngare, Dr. Owour and Professor Pokhariyal for taking time to provide ideas, suggestions, comments or reading through my work, and not forgetting Mr. Miano for his inputs.

Special thanks to the School of Mathematics fraternity for providing a warm and friendly environment for research.

My friends, asante sana for the moral support and in believing in me. I cannot thank you enough.

Abstract

The modern society quest for credit has led to an increase in consumer credit uptake with proportionate increase in default rates. Some of the current credit risk models deteriorate over time and need frequent validation. This is due to use of historical data that is time dependent. The models lack the flexibility to take into account changes in economic and other extreme events. This research considers the use of social network data to offer an alternative approach in consumer credit scoring. Social data is widely available with vast amount of information due to increase in social network sites and technologies. This offers time dependent data that can be harnessed to develop time dependent models that do not need validation and will not deteriorate with time. Further, the poor and young consumers lack historical data and thus social data will cover that gap.

The set of agents who are part of a network are also obligors in a loan portfolio with a financial institution. The key aim is to estimate credit risk in a loan portfolio based on the agents' behavior at the network level. The agents interactions and cyclical inter dependencies in the social economic network are estimated to derive the social and economic factors. These factors are derived from the network matrices using singular value decomposition technique and scaled into $(0, 1]$. The scaled data forms the credit risk analysis factors that are used to learn and train the hidden Markov model. The model emits the credit quality levels, the dynamic threshold and the credit quality scores. These outputs are in turn used to estimate the model false rates and the obligors' delinquent cases, default rate, stopping time and survival rates.

The dynamic threshold is estimated at each time period to capture the dynamics of the credit quality of the obligors in the loan portfolio and emit the default and non-default rates. Obligors are classified into four credit quality levels; poor, average, good and excellent (PAGE).

Obligors with average and good credit quality levels ranges between 61.5% and 89.3%

while the excellent credit quality level was between 8% and 20.4%. The obligors classified in the false rate category ranges between 25% and 50.1%. The model performance is between 53% and 73% which is an accuracy rating of between medium and good accuracy. Sensitivity analysis and false rates in the model have a coefficient of determination of between 0.647 and 0.983.

The social network model offers an alternative approach to consumer credit scoring with time dependent data. Agents' interactions and cyclical interdependencies is an ideal approach to incorporate in consumer underwriting and capture the poor and the unbanked. The model has opened new frontiers in consumer credit scoring. Thus, the study contributes in opening up new frontiers and innovations in consumer credit scoring with social and economic data.

Abbreviations and Acronyms

BCBS	Basel Committee on Bank Supervision.
CRAF	Credit Risk Analysis Factors.
CQL	Credit Quality Level.
CQS	Credit Quality Scores.
EM	Expectation Maximization.
EMS	Empirical Martingale Simulation.
HCQS	Hybrid Credit Quality Score.
HMM	Hidden Markov Model.
HMP	Hidden Markov Process.
IRBA	Internal Ratings Based Approach.
KMV	Kealhofer, McQuown and Vasicek.
MCS	Monte Carlo Simulation.
PAGE	Poor, Average, Good and Excellent
RNG	Random Number Generator.
SVD	Singular Value Decomposition.
SEN	Socio-Economic Network.
SEN-HMM-CSD	Social Economic Network + Hidden Markov Model +Credit Scores and Default Model
SMD	Social Media Data.
SND	Social Network Data.

Keywords

Agent
Credit quality scores and levels
Default threshold
Hidden Markov Model
Interactions
Loan portfolio
Obligor
Simulation
Socio-economic network
Reputation
Trust

List of Symbols and Notations

$(\Omega, \mathcal{F}, \mathbb{P})$ The probability space

$(\Omega, \mathcal{F}, \mathcal{F}_t, \mathbb{P})$ A filtered probability space, $(\mathcal{F}_t)_{t\in[0,T]},\ \mathcal{F}_t \subseteq \mathcal{F}$

$\alpha_t^o(i)$ The forward variable for the sequence o at time t in state i

$\bar{\lambda}$ A hidden Markov model as defined by its $\bar{A}, \bar{B}, \bar{\pi}, \bar{O}, \bar{Q}$ for the default threshold

$\bar{\phi}^t$ The default threshold at time t

$\bar{A}$ $M \times M$ matrix of the average state transition probabilities (for all the agents)

$\bar{B}$ $M \times V$ matrix of the average observation probabilities (for all the agents)

$\beta_t^o(i)$ The backward variable for the sequence o at time t in state i

$\ddot{g}$ A false positive rating of an obligor based on the hybrid credit quality and the credit quality score

ΔX The changes in the private data of the agents

$\dot{\phi}^t$ The credit quality level of agents at time t

$\dot{\psi}$ Network reputation feedback an agent gains from the feedback of all other agents in the network, a summary of the reputation based on the outcome of other agents ratings

$\dot{g}$ A false negative rating of an obligor based on the hybrid credit quality and the credit quality score

$\dot{S}$ Accuracy notation

$\dot{X}$ Return on the private data generated by the agents in the network

$\hat{\lambda}$	A hidden Markov model as defined by its $\bar{A}, \bar{B}, \pi, O, Q$ for the hybrid credit quality
$\hat{\phi}_n^t$	The hybrid credit quality score of agent n at time t
$\hat{\Theta}$	The distrust levels of the agents in the network
$\hat{g}$	A negative negative rating of an obligor based on the hybrid credit quality and the credit quality score
$\hat{S}$	Sensitivity analysis parameter
$\hat{t}$	The time when the loan premiums are paid at the end of one month or 30 days period
$\kappa,\ i,\ j,\ m$	Index or counter to track changes in various variables in the model
λ	A hidden Markov model as defined by its A, B, π, O, Q for the credit quality
$\mathbf{A}$	Real valued matrix
$\mathbf{S}$	Diagonal matrix of the singular values of matrix $\mathbf{A}$
$\mathbf{U}$	Left eigenvectors of matrix $\mathbf{A}$ extracted with SVD
$\mathbf{V}$	Right eigenvectors of matrix $\mathbf{A}$ extracted with SVD
$\mathfrak{F}$	Sigma fields or a sequence of filtration which is a sequence of observations from the occurrences
$\mathfrak{C}$	Condition number in SVD technique
$\mathfrak{N}$	The maximum number of times an agent encounters another agent in the network at a given time period
Ω	The sample space
ϕ_n^t	The credit quality score of agent n at time t
π	Initial state probability
Ψ	The interaction experiences gained by an agent in the network from the interaction encounters

ψ A vector of the interaction experience that can be bad, neutral or good

τ The time interval within a period of time $[t-1, t]$ when interactions are estimated in the SEN

Θ The trust levels of the agents in the network

$\tilde{\Theta}$ The SEN risk factors of the agents in the network

$\tilde{R}$ Matrix of the peer to peer reputation ratings in the network

$\tilde{X}$ The ethical factors of the agents in the network

φ The vector of the age of the agents in the network

A $M \times M$ matrix of state transition probabilities

a_{ij} State transition probability from state i to j

B $M \times V$ matrix of observation probabilities

$b_j(\kappa)$ Observation symbol probability

g A positive positive rating of an obligor based on the hybrid credit quality and the credit quality score

K The number of observation symbols

M The number of state sequence symbols

N Set of agents or obligors in the network

O Sequence of observation symbols in a HMM

o_n^t An observed symbol for agent n at time t

P Probability symbol

Q Sequence of hidden states in a HMM

q_n^t A hidden state for agent n at time t

R Coefficient of determination

S	the set of hidden state space
T	The life or duration of the loan obligation
t	The time when the bank computes the credit quality of the obligors at the end of two months or after a period of 60 days
V	The set of the observation symbols
X	Private data of the agents which is vector $x_1, \ldots, x_n$
$y,\ z$	Random variables
L	The observation and state sequence length (with $L \geq 1000$)

List of Tables

List of Figures

Table of Contents

Chapter 1

Introduction

1.1 Motivation

The idea of modeling credit risk of obligors in the socio-economic network was triggered by the work of Eisenberg and Noe (2001) for a financial system. The study considers the properties of inter corporate cash flows that are assumed to have cyclical interdependence amongst the players and default rates are determined endogenously by use of a clearing vector in the network. Table 1.1 summarizes the existing research work of Eisenberg and Noe (2001) and is compared to the advances and our contributions to the field of consumer credit risk modeling using hidden Markov model with social economic network data.

Table 1.1: Current versus existing research framework

(Eisenberg and Noe , 2001)	Current Research
Corporate credit risk	Consumer credit risk
Financial network	Social economic network
Default 'waves' of firms measured systemic risk	Credit scores and threshold measures default rates
Cash flows of firms with cyclical inter dependencies	Agents SEN factors with cyclical inter dependencies
Simulation algorithm for individual firm	Simulation algorithm for SEN agents' variables
	Use hidden Markov model
	Use singular value decomposition

In the process of finalizing this study, three researchers with work on the similar area of credit scoring using social media data were spotted. Daniel and Grissen (2015) used the mobile phone usage data to predict loan repayment in a developing country; Masyutin (2015) used social data from one of Russia's popular social network to discriminate between solvent and delinquent debtors of credit organizations; while Wei et al. (2015) compares the accuracy of customer scoring obtained with and also without network data.

The motivations behind this research are; first, the experiences from the year 2007 financial crisis where the interwined nature of financial systems was brought to fore (Allen and Babus , 2008); and the existing credit risk models failure to capture the dynamics observable in the market (Capuano et al. , 2009); Second, the young and the poor households lack formal financial histories that can be used by the financial institutions for credit scoring thus widening the data set available and capturing of new markets (Daniel and Grissen , 2015; PWC , 2015). Social media data (SMD) or social network data (SND) can provide an alternative data to support decisions about an applicant's creditworthiness as this data provides vast amounts of information (PWC , 2015).

The other reasons that makes SMD (SND and SMD are used interchangeably in this research work) a good candidate for consumer credit scoring and estimation of default rates are based on the following facts:

(i) The dynamics and innovations observable in the consumer finance market calls for use of advanced analytics and big data from social cycles to gain insights in the demographic changes, borrowers needs as well as loop in households with minimal or no interactions with the formal financial institutions (Daniel and Grissen , 2015; PWC , 2015).

(ii) No *agent* lives in a vacuum as they must interact with other agents in the network to achieve its goals (Moe et al. , 2008)

(iii) Modeling trust in *complex dynamic environment* is non trivial; intelligent agents strategically change their behaviour to maximize the utility gained from the network (Liu and Datta , 2012; Moe et al. , 2008)).

(iv) Social networks continue to generate and accumulate huge amounts of data and information vital in providing insights on people's behaviour (Masyutin , 2015). This is the data that financial service industry can use effectively to make robust and informed decisions in the field of consumer credit risk.

(v) Modeling and analyzing *default risk* in a network using agents interaction effects is a key component of consumer credit risk

(vi) Modeling *default events* needs a dynamic process; the Hidden Markov Model (HMM), which is a stochastic process based model ideal for timing such defaults (Crowder et al. , 2005).

(vii) Minimum research exists on consumer credit and default prediction compared to corporate credit risk. Consumer credit is at the highest today but the default rates have risen (Horkko , 2010)

(viii) The 'true' default probability is elusive because default estimation depends on the information that is available. Accuracy of the default ratings is limited as default has no direct relationship to the observable quantities (David , 2004).

(ix) Credit scoring models performance deteriorates over time due to use of historical data. Periodic validation maintains the models' accuracy and completeness in order to generate time tested scores (Robert et al. , 1996)

(x) Economic conditions are not the only cause of change in credit risk as massive increase in defaults and bankruptcy have been observed even in good economic times (Thomas et al. , 2005).

The techniques of personal credit scoring are highlighted by Li and Zhong (2012) where they observe that credit scoring determines whether the applicant is qualified to receive credit. The duo offers an overview of the credit scoring techniques and discusses four current research problems, where we single out two of the four problems to be incorporated in this research. The research problems are:

(i) Type I and type II error in classification of customers. For type I error, we classify good customers as bad ones and reject their loans, and this reduces the bank's profit. For type II error, the bad customers are classified as good ones and provided with loans, which will bring loss to banks (Li and Zhong , 2012).

(ii) Incorporate economic conditions into credit scoring models. The assumption that the past economic condition are similar to the economic conditions today is misleading. The conditions are unpredictable and random, so an evolution criteria is important to estimate these changes and how they affect the credit scoring process (Li and Zhong , 2012).

The highlighted points and reasons form the main motivation in undertaking this research. We strongly believe that based on these facts, social networks are crucial in understanding the issue of consumer credit risk and it is the approach we take in this study.

1.2 Background

This research work on consumer credit risk, the credit quality scoring for both individual obligors at the portfolio level is based on a class of stochastic processes that have a finite set of states. They are considered as a special form of dynamic Bayesian networks which are based on Bayes theory. This class of stochastic process is known as the Hidden Markov Models (HMMs), a function of the Markov Chain and Markov process. The theory of Markov Process was the original work of a Russian Mathematician, Andrey Andreyevich Markov $(1856 - 1922)$. Markov main research work considered the theory of stochastic Markov processes which later came to be known as Markov process and Markov chains. HE introduced the Markov chains in 1906. In the year 1913, he used the Russian language to calculate the letter sequence of the Markov chain. Kolmogorov in 1931 generalized the results to countable infinite state spaces (Dymarski , 2011).

In the twentieth century, Markov chains were linked to Brownian motion and the ergodic hypothesis. This was extended from the law of large numbers. In general, Markov process in probability theory and statistics can be considered as time varying random phenomenon for which Markov properties are achieved. A stochastic process has a Markov property, that is memorylessness, where the present state of the system is dependent on preceding past but independent with the past (Dymarski , 2011).

A more powerful model for much more complicated stochastic process than the Markov model is the Hidden Markov Model (HMM). The HMM was introduced by Leonard Baum and others in 1966; and widely used in science, engineering and many other areas such as speech recognition, gesture recognition, language modeling, motion video analysis and tracking, protein and gene sequence alignment, finance, economics, among other areas.

Modeling credit risk has become an essential tool for modern risk management. The appetite for borrowing has truly become global in scope and created diversity in the portfolio risk. Credit scoring systems have continued to occupy research interest in the areas of financial engineering. The classical credit risk analysis core objective is to decide whether a loan should be granted, and after its granted, trying to assess the risk of default. As credit risk continue to concern commercial lenders, the financial industry representatives, academics worldwide and regulators, among others have continued to increase their effort towards improving the credit risk modeling process (Capuano et al. , 2009; Ching et al. , 2008).

Banks and other financial institutions are applying increasingly sophisticated methods to assess the risk of their loan portfolio. The methods for assessing risk try to find an answer to the question, what is the likelihood of the applicant defaulting at a given time in the future. That is why numerous methods continue to be developed for credit scoring. Li and Zhong (2012) mentions a number of credit modeling techniques from mathematical programing, expert systems, neural networks, discriminant analysis, genetic algorithms, logistic regression, HMM, partitioning trees and nearest neighbor concept, among others. The techniques can be classified into three groups (Li and Zhong , 2012) and the classification accuracy of the techniques are highlighted by (Thomas et al. , 2005).

Consumer credit dates back around $3,000$ years ago since the time of the Babylonians. Thomas et al. (2005) undertakes a historical and current times survey on consumer credit from time of pawn brokers and usurers (750 years ago) to the current consumer credit mass market. They give figures and facts that indicate the massive growth both in number and products on offer, plus output in research work. The authors further look at consumer credit modeling and current issues in the field. As these changes continue, the advent of Basel Capital Accord II has increased the need to model credit risk of a portfolio (interdependency), not just the risk of each loan defaulting independently (Thomas et al. , 2005). Assumption of independence between individual obligor's default rate leads to the underestimation of the portfolio's credit risk.

The increasing availability and widespread of credit in modern society has led to an increase in the rate of personal bankruptcy due to default on credit repayment. Financial institutions can cover small exposures using normal operating cash flow but multiple defaults simultaneously are a threat to the institutions. The concern is the concentration risk in a loan portfolio (Horkko , 2010). The quantitative methods used by banks to appraise ‘good’ and ‘bad’ obligors have been in use but in recent past, new interdisciplinary approaches to analyzing personal loan data have been in use from operations research, financial sciences, sociology, mathematical statistics to statistical methods of survival analysis (Capuano et al. , 2009). Innovations and research on techniques and methodologies to credit risk modeling has continued to take new dimensions.

Even with these advances, the recent global financial crisis of $2007 - 2008$ brought to fore the limitations of these innovative models. The financial systems revealed it’s intertwined nature and proved the limitations of the mathematical techniques as the models failed to measure credit risk. The inherent problem being the assumptions underlying these models that lack dynamical ways to capture the economic changes and extreme events (Capuano et al. , 2009). Even though the financial network provides fertile ground for increasing global integration of the markets, there are inherent risks.

A network describes a collection of nodes and links, with nodes as individuals, or firms, or countries (hereafter referred to as agents), and links could be friendship, tie, or free trade agreements, or interbank transactions. Agents are individuals who have the ability to act, and possibly to react to external stimuli and interact with the environment and other agents (Allen and Babus , 2008; Meyers , 2009). Social network theory studies the properties of networks and statistical generative models. Social influence is the ability of a node to manipulate other nodes to adopt or reject the transmission of information during the information propagation process (Madan and Pentland , 2009). The rise in online communities, population and the role of networks in our world, social networks continue to affect our social and economic lives (Jackson , 2008).

We use SEN as they permeate our social and economic lives playing a central role in the transmission of information. The term social network was first used in 1950 in sociometrics, the science that seeks to obtain data on social behaviour and to analyze it. Social networks create trust between agents because they allow for repeated interactions between the members and create room to learn about each other. The structural connectivity (how agents are connected to each other in the network) and the behavioral connectivity (how an individual agent actions affect other agents in the network) present high complexity in social networks (Pupazan , 2011). These social structures are ideal for effective monitoring and enforcement of risk sharing agreements as well as in gaining significant payoff advantages through the connecting of the disconnected agents (Jackson , 2008).

When peoples' interactions are captured over a period of time, the history of their past interactions forms a set of information that informs them about their abilities. A good reputation system collects, stores, distributes and aggregates feedback about the agents past behaviour (Resnick et al. , 2000). The reputation of an agent is an important factor in performing trust decisions as agents use it as a means to measure their own past experiences with the other agents and comparing their reputation in relation to the other agents (Ganesh and Sethi , 2013). In social networks, reputation quantifies the ratings from the underlying network and the agent's reputation is visible to all other agents. For example, in the online trading communities, the seller's reputation has an influence on the online auction process. Thus, trust levels of agents are extracted from the reputation ratings of each other in the network. Trust is crucial in formation of connections in social networks due to its influence on how information flows and in assessing the quality of information in the network (Ganesh and Sethi , 2013).

Social networks continue to gain popularity in describing social and scientific relations as they play a central role in the transmission of information (Chen et al. , 2008). The networks are dynamic, complex and with stochastically evolving agents who strategically change or reinforce their bahavior to improve their lot in the network (Meyers , 2009; Skyrms and Pemantle , 2009). The dynamics witnessed in the social media implies that financial institutions can leverage on advanced analytics and big data to gain insights on credit scores of consumers with no or very low finance histories (PWC , 2015). Young consumers and many households in developing countries lack formal interactions with financial institutions that generate the necessary data for credit scoring (Daniel and Grissen , 2015). The social media circles and social interactions provides vast amounts of information that can support decisions about the consumer credit scoring process. A few innovative lenders have pioneered in the use of social media data in credit underwriting process to supplement the traditional methods of credit scoring (PWC , 2015).

The social media users continue to generate content and interactions at unprecedented rate with data that is awaiting to be harnessed to improve the existing consumer credit risk (Dewing , 2012). The availability of powerful data mining techniques for internet content offers an opportunity for consumer credit scoring to single out credit worthy consumers and use that data to predict other consumers with scanty financial history (Dubois et al. , 2011; Tang et al. , 2014). This is possible as social networking allows people to indicate whom they trust and distrust creating links in the network.

Heterogeneity of the agents implies that they have different economic data, social data and the credit as well as behavioral probabilities. The stochastic changes generate structure changes on how agents behave and respond that can be captured by a Markov process. A hidden Markov Model which is a statistical Markov Model can be used. The system under study is assumed to be a Markov process with unobserved (hidden) states. The output dependents on the state and is visible which makes the model ideal for real world processes with observable outputs characterized as signals.

In credit scoring, models have been developed with default barrier being based on a random process (Koyluoglu and Hickman , 1998). Therefore, the HMM, can be used as they have the ability to output both the individual agent credit scoring and a dynamic threshold to estimate the changes in credit quality. This highlights a new frontier in developing accurate estimates of credit quality scores of obligors in a loan portfolio (Thomas et al. , 2005) who are part of a SEN. Further, the market has accepted the importance of risk based pricing of credit products. That is, customers with different risk profiles pay different amounts for the same product. The result is to develop a scoring model with accuracy and ability to rank the credit risk of individual consumers.

Simulation analysis is used to achieve this quest of using SND in consumer lending as we lack real life data for our analysis. We observe that, simulation technique has gained popularity due to its diverse applicability and versatility (Capuano et al. , 2009).

1.3 Credit Risk

Credit risk is likely or risk of loss resulting from failures of counterparties or borrowers to fulfill their obligations. It is the major source of risk for commercial banks. A key component of credit risk is the *default* event which occurs if the debtor is unable to meet its legal obligation as per the debt contract (David , 2004). Our main form of credit risk in this work is the repayment delinquency in retail loans (Baesens and Gestel , 2009). An obligor may be in default but has not defaulted on payment, in payment default but is not insolvent and in default but not declared bankrupt (Daniel and Grissen , 2015). Therefore, we have three scenarios. First, default - an obligation is not honored. Second, payment default - an obligor does not make a payment when it is due where we have; repudiation (refusal to accept a claim as valid), moratorium (stopped payment for a period of time), and credit default. Third, insolvency - inability to pay, and fourth, bankruptcy - the start of a formal legal procedure to ensure fair treatment of all creditors of a defaulted obligor

Credit risk is a common form of risk in almost all financial activities. The ability to measure, price and manage this risk is important for loan portfolios. Credit quality dynamics is important in credit risk measures and its application to pricing and portfolio risk management (Korolkiewicz , 2010). Many estimates are needed when performing default probabilities and its dynamics through time. A credit rating system serves to accurately assess the credit risk of the obligor. Credit ratings are expected to assess the likelihood of an obligor defaulting and accurately classify the obligor according to their credit quality (David , 2004). Most lenders calculate behavioral scores for all their borrowers on frequent basis. The scores form a basis to estimate default probability in a fixed time horizon. The scoring process uses recorded information about the individual and their loan requests to predict, in a quantifiable numerical value, the future performance regarding debt repayment (Robert et al. , 1996). Credit scores are superior to the subjective assessment of credit history. This allows underwriters to improve on the methods to better assess the strengths and weaknesses of the applicants. This increases the accuracy, consistency and speed of the credit evaluation process (Robert et al. , 1996).

In credit scoring systems, differences in application characteristics of the customers are observed. Stepanova and Thomas (2002) notes that these scoring systems are important to aid in the decision as whether to grant credit to an applicant or not. The concept of true default probability has been found to be elusive because it depends on the information that is available. This makes it hard to develop a methodology to judge the accuracy of the ratings as default has no direct relation to the observable quantities (Daniel and Grissen , 2015). Financial institutions should learn to segment obligors in terms of delinquency and default rates as well as the ease of recovery process. Delinquency occurs when an obligor fails to honor their obligation by making a payment as scheduled, but sometimes the term default and delinquency are used interchangeably (Robert et al. , 1996). The actual number of obligors who become delinquent on the loans is much greater that the number that actually default. The issue of segmenting bank's customers has been echoed by the Basel Accord recommended by the Basel Committee on Bank Supervision (BCBS). The first Basel was issued in the year 1988 and the second in the year 2004. Banks have the leeway of deciding what percentage of the loan amount to set aside to cover the possible defaults by using the Internal Ratings Based Approach (IRBA) (Thomas et al. , 2005).

The fact that these recommendations and other issues have been addressed, it is important to consider the psychological and economic factors influencing defaults. Default is a time dependent event, and these two factors should be measured and incorporated into new types of models for credit risk estimation of consumer loans (Thomas et al. , 2005). The time dependent events include consumer interaction with the environment which means taking actions with uncertain effects, even though the obligors are supposed to make informed decisions which account explicitly for the uncertainty with the world.

Agents interactions in a social network are important in understanding the social structures and how they affect the agents ability on their financial obligations. PWC (2015) asks the question, is it the time for consumer lending to go social? With the abundance of SMD, innovative approach and changes in the consumer trends requires better credit scoring and loans underwriting decision making process. The gap can be filled by the advanced analytics and big data from the social media (PWC , 2015). The use of the applicant's social media data has opportunities and benefits to the lenders. PWC (2015) highlights five key benefits of using the SMD in loan underwriting process;

(i) Capture new customer segments - an alternative set of underwriting data may help lenders assess creditworthiness of applicants with scanty or no financial history. This can expand and tap on the excluded customer base (unbanked and poor).

(ii) Provide a differentiated customer experience - SMD can establish a framework of analytical client knowledge that demonstrates an understanding of the customer needs. Customers will appreciate being treated like more than just a number and this could strengthen brand loyalty.

(iii) Strengthen existing underwriting processes - The credit scoring process is strengthened by availability of more data points, and hence helping to limit losses.

(iv) Prevent fraud - The available information on social media channels can be used to cross-check information provided in loan applications. This could help identify fraudulent activity early and prevent it from occurring before a loan is approved.

(v) Develop a competitive edge - The use of social media data may help lenders meet their strategic goals and gain from the potential benefits; better manage new clients with scanty financial history, expand their market share, better pricing of risk, minimize expected losses, and improve performance and credit scoring strategy.

But why now? PWC (2015) notes that data is everywhere and credit trends require a fresh approach. Some of the SMD for consumer lending are; basic personal and professional data, personal network, customer-provided data, and behavioral character data. It is on that line that Daniel and Grissen (2015) used mobile phone usage to predict loan repayment in a developing country. They quantified the so called soft information to complement the current methods in use. The method is promising even for the poor borrowers and those excluded from the main stream financial systems.

The issue of social media risk management arises. With growth comes opportunities and challenges. According to Ernst & Young survey, 67% of social media users say that social media influences their purchases (Ernst and Young , 2013). Thus, social media is a powerful tool to increase market share. Ernst and Young (2013) notes that social media compliance are consistent with those imposed on traditional and other electronic-based channels. The risks posed are similar to those of other electronic communication that include potential consumer compliance, operational, legal and reputation risk.

We highlight the strengths of using SMD in consumer lending process in the next section. SEN are powerful and rich sources of obligors interactions and behavior that are expected to change the way consumer lending is undertaken.

1.3.1 Social economic network

A socio-economic network (SEN) is where the primary action entails economic transactions, under the structure of the social capital. The behaviour of obligors is driven by both the social and economic factors. The relationships in the SEN arises from agent's strategies in investments which occur as individuals or collectively, with the aim of establishing social links and the utility derived from the links (Johnson , 2003). Social network analysis has provided a significant role in domains of security, sales, terrorism, biology, disease spread modeling, economy and marketing to secure higher profits, finance, etc. This is made possible by huge amount of social network data available with studies and simulation of different nature made possible. These contribute significantly to understanding the properties and the behaviour of social networks (Netrvalova and Safarik , 2011; Pupazan , 2011; Stanley , 2006).

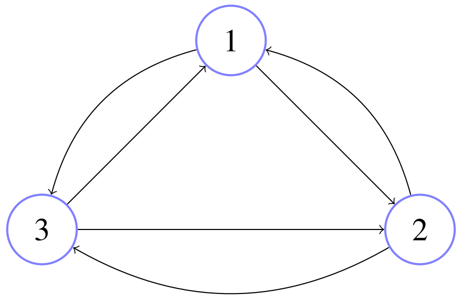

Figure 1.1: Connections of three agents in a SEN

A Socio-Economic Network depicting three agents who are interconnected to each other to show the interdependence between the agents.

Social networks aids in generating social capital which in turn generates resources to assist in accumulation of human capital. Social capital, as with an asset, depreciates over time. Social capital can contribute to economic situations as it represents an asset, although it is more difficult to measure (Johnson , 2003). But Pani (2008) observes that there is an association between social capital and the economic growth of a firm. Social capital is the sum of social obligations and can be converted into economic capital in some situations and conditions. Pani (2008) notes that the perspective on which social capital is built cannot be separated from an actor and its activities, including economic activities. Dynamic social networks are determined by stochastically evolving social network. These random and strategic interactions are a key application in financial risk management and our concern is how to capture them.

The data available in social networks is important as these networks continue to permeate our lives, playing a central role in the transmission of information. These networks are crucial in credit risk as a network embeds dynamic, complex and flexible agents activities and behaviour where the agents act and possibly react to external and internal environment influenced by other agents (Meyers , 2009). An agent is likely to encounter an agent generated content, some of which the agent uses to make decisions and develop context within a community with respect to whom they will continue to interact with (Dubois et al. , 2011). As agents in a social network interact, they form links that stochastically evolve over time. This evolution has history of past interactions that informs an agent about its abilities and dispositions (Resnick et al. , 2000). The social interactions between pairs of individuals with strong ties are more likely to exhibit greater similarity compared to those with weak ties (Xiang et al. , 2010).

The effectiveness of interactions in a SEN are guided by trust (Dubois et al. , 2011). Trust is derived from the agents reputation ratings in the network. In the next section, we consider HMM and why it is a powerful classification technique to use in this study.

1.3.2 Hidden Markov model

Hidden Markov Model has increasingly become popular for a wide range of applications due to its strong theoretical and mathematical structure. Hassan and Nath (2005) outlines four advantages of HMM; strong statistical foundation; ability to handle new data robustly; ability to predict similar patterns efficiently; and computationally efficient to develop and evaluate due to availability of training algorithms. Bilmes (2006) notes that there is no general theoretical limit on the capabilities of HMMs when we have enough hidden states, observation distributions, sufficient and adequate training data and the appropriate training algorithm, as it is more versatile than the normal Markov model. Hassan et al. (2006) provides a summary of the HMM strengths:

(a) Natural model structure

HMM is a stochastic process and the transition parameters model temporal variability (occurrence in time) and output parameters model spatial variability (quantity measured at different locations exhibit differences). Sequential data analysis and the HMM are a good match as the interactions of the agents in the SEN is hidden but the net value is observable. In sequential data (correlation among subsequent samples), where i.i.d assumption may no longer be a good approximation.

(b) Efficient and good modeling tool

Ideal for real world complex processes where the sequences have temporal constraints and spatial variability along the sequence.

(c) Proven technology

HMM theoretical basis forms a wide scope for different applications with high success rate, such as language modeling, speech recognition, motion video tracking, stock price prediction, protein sequence, gene sequence alignment, among other areas.

(d) Efficient evaluation, decoding and training algorithms

These algorithms are mathematically strong and computationally efficient. Transition probabilities and the observation generation probability density function are both adjustable. The flexibility of HMM to embedded another model, the threshold model and use of unsupervised and supervised learning technique to allow for new patterns in the models (Bilmes , 2006).

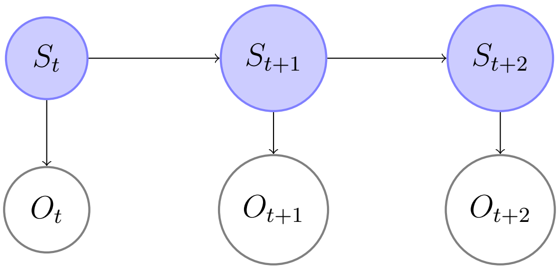

Figure 1.2: Standard (Single-Chain) Hidden Markov Model
The shaded circles are the hidden states and the empty circles are the observation nodes.

1.4 Problem Statement

Consumer credit dates back around 3000 years ago since the time of the Babylonians. With time, consumer credit moved to mass market contributing to increasing availability and widespread of consumer credit in our modern society. This increase in availability of consumer credit has led to increase in default rates. The likely reason being that the existing credit risk models performance deteriorating over time with a need for periodic validation. The validation is due to use of historical data that assumes credit quality to be time independent. Another reason is that most of these existing models lack the needed flexibility to take into account changes in economic or in other extreme events. For example, in the year 2007, the financial crisis proved that the credit models are inadequate in addressing the world dynamics due to the models underlying assumptions.

The exclusion of some consumers from credit facilities due to lack of data limits the efforts towards financial inclusion. The poor and young consumers have minimal or no financial history to apply and qualify for credit. The availability and increase in social network data is one of the innovative ways to deal with these challenges. The social network data is time dependent and this eliminates the need to frequently validate the existing credit risk model. The poor and young consumers tend to be active in the social network and the data generated by these networks can cover the existing gap in lack of financial history data.

Therefore, the limitations of credit risk models, the current status in consumer credit and exclusion of some consumers from credit facilities calls for alternative approaches in consumer credit scoring. This is the approach and solution being offered by this study. We estimate consumer credit scores with time dependent social and economic data using HMM. The obligors are part of a SEN with cyclical inter dependencies as well as of a loan portfolio in a financial institution. This, eliminates the credit risk models deteriorating, need of frequent validation and its inability to respond to changes in extreme events. The consumers excluded from credit facilities benefits from the SEN data and this leads to increase in financial inclusion.

1.5 Objectives of the Study

1.5.1 Overall objective

The overall objective of this study is to model the credit quality of obligors who are part of a bank's loan portfolio, and in a social economic network with the social media data using the hidden Markov model.

1.5.2 Specific objectives

The specific objectives of the study are to:

(i) Develop and simulate a SEN model for agents stochastic interactions and cyclical inter dependencies due to social and economic factors

(ii) Compute the CRAFs from the SEN

(iii) Estimate the HMM parameters for the model

(iv) Analyze the credit quality scores, dynamic threshold and default rates

(v) Estimate the delinquency, survival, false rates and stopping time of the obligors.

1.5.3 Limitations of the study

Lack of real life data in this area allows us to use simulation technique to generate the required data, as it enables us to capture the network complexities, versatility and dynamics. Real life interactions are much more complicated than is depicted in this study but simulation will offer great insight on the workings of SENs based on HMM.

The main data used to estimate the CQS of the agents is the SEN data only. Ordinarily, most credit risk analysis are done with the five C's of credit analysis (capital, collateral, conditions, character and capacity) which we lack in this work. We instead take a different dimension and look at some factors in social network analysis that can be used in consumer credit scoring.

No mathematical system model is perfect as these models only depicts those characteristics of direct interest to the modeller (Maybeck , 1979). The objective of this study is to represent the SEN data and apply it to model the consumer credit scoring process.

1.6 Significance of Study

The recent economic crisis has emphasized the need for new and fundamental understanding of the dynamics and structure of socio-economic networks. The networks are increasingly being built on inter dependencies, stressing the system complexity and reflect a dynamic interaction of a large number of different agents (Schweitzer et al. , 2009). Agents in the network have different behaviors and these evolving interactions are evident in a network dynamics, bound in space and time. This overcomes the shortcoming of historical data that assumes credit quality to be time independent. Decisions evolve under changing conditions, and HMMs have the capabilities to better assess new information on continues basis to capture the credit quality scores of the obligors.

The advantage of this model is its ability to dynamically track changes in the credit quality scores, incorporate other existing models on consumer credit risk to complement their usage, and develop a predictive model. Another factor is that the poor lack historical data on their financial obligations performance and SMD can cover that gap. The financial institutions benefits are in the ability to loop in new customer segment; strengthening and improving credit risk management processes; and expanding the market share with better pricing of risk for the institutions. The regulators of the financial institutions will be able to reduce the non performing loans and increase financial inclusion. In the academic front, there is going to be an increase in new and improvement in existing credit risk models; enhanced data mining techniques for social media data; and more research output in consumer credit risk.

1.7 Contributions of the Thesis

This thesis contribution was based on the problems being addressed by use of social and economic network data. First, there is an increase in modern credit that has led to rise in default rates. Second, some of the current credit scoring models deteriorates over time and needs frequent validation due to use of historical data. Third, some credit scoring models failed to capture the extreme events in the recent financial crisis of the year 2007. Fourth, the young, poor and unbanked fail to secure credit due to lack of financial histories. Fifth, the research by Eisenberg and Noe (2001) considered systemic risk in a corporate financial network. Sixth, there is an abundant social network data that has vast information and influences how we learn and interact. These form the basis of the thesis contributions.

(a) The increase in population and the growth in consumer credit mass market has provided new opportunities and challenges. More innovative methods are needed in credit underwriting. The model in this study is providing that opportunity to the credit risk market.

(b) The model in this thesis uses time dependent data from the social network and thus does not deteriorate or need any validation the life of a credit facility. This also means that the time dependent data incorporates change in time events during the period under study.

(c) As the young and unbanked lack financial histories, social media data provide vast amounts of information that can support consumer credit underwriting. This is possible as most young people have access to social networks at an early age compared to their access to financial institutions. The use of social network data in this study covers this gap to increase financial inclusion.

(d) The research by Eisenberg and Noe (2001) considers systemic risk in a financial network. The network is made of firms that have a common clearing vector to estimate default rates in the network. This study extends that research to consider agents in a social and economic network, affected by their cyclical interdependence and each agent has a loan obligation with a financial institution. Hidden Markov model is used to estimate the credit quality scores and levels of the obligors and a dynamic threshold that estimates the obligors default rates

(e) A new definition to suit the SEN is presented and a new theorem for reputation ratings as a stochastic process is developed and proved.

(f) We have estimated the transition matrix A and observation matrix B by supervised clustering using the CRAFs.

(g) We have modified the standard HMM to cater for the multiple agents HMM as the SEN has a set of heterogeneous agents interactions.

1.7.1 Thesis organization

This thesis comprises of seven chapters, list of references and two appendices (one for HMM and the other with list of publications with the manuscript attachments).

Chapter 2 has the literature review from other research work on the different areas applied in this study. The areas include SENs, trust and reputation, Hidden MArkov Model, consumer credit risk and simulation.

Chapter 3 outlines the mathematics preliminaries and basic tools from social networks, trust, stochastic processes, martingale, stopping time, singular value decomposition, simulation and Markov process and chains. The chapter is an introduction to known and accepted mathematical tools. This forms a basis for the chapters that follows, that is, chapter 4, chapter 5 and chapter 6.

Chapter 4 introduces the standard hidden Markov model specifications. This chapter further discuss the modifications of the standard HMM specifications to the multiple agents HMM, which forms part of the study methodology. The modifications forms the part of the HMM analysis in our new model that is developed to analyze the credit scores of the agents in the SEN.

Chapter 5 is the methodology of the proposed model that is composed of: the initial conditions, the social economic network dynamics, the hidden Markov model parameter estimation, the credit quality scores, default rates and false rates estimation. A new definition for SEN is developed and a new theorem for reputation as a stochastic process is developed and proved using the Martingale principle.

Chapter 6 is the analysis of the SEN-HMM-CSD model, its parameters and variables. The Monte Carlo simulation generates the data for the model analysis and estimation of the key variables in this study. Discussions from the results and findings are presented and reviewed based on the study objectives.

Chapter 7 has the summary of the discussions of each of the objective of the study and the conclusions from the study. Recommendations for further research areas based on the study are presented.

Chapter 2

Literature Review

The chapter highlights research work related to this study. We are using different techniques in this research from social networks, trust and reputation, Hidden Markov models, stopping time, stochastic processes, credit risk, default and simulation. We review the literature on these aforementioned areas.

2.1 Social and Economic Networks

Social networks have vast amount of social networking data which contributes significantly to understanding the properties and the behaviour of the networks (Meyers , 2009; Netrvalova and Safarik , 2011; Pupazan , 2011; Raghavan et al. , 2013). As agents in a social network interact, they form links that stochastically evolve over time and lead to network evolution (Meyers , 2009; Resnick et al. , 2000; Skyrms and Pemantle , 2009; Starnini et al. , 2013).

A social network is a social structure created by individuals that are bounded together on the basis of some particularity (Netrvalova and Safarik , 2011). Individuals in social network are called actors and social networks connects these individuals in the groups. Social network dynamics provides a platform to study agents and their collective behaviour on a large scale. Raghavan et al. (2013) notes that social interactions on networks affect agents activity and these activities should be incorporated in social networks to develop optimal models. They incorporated social effects of influence on a user's activity from the activity of a user's neighbour to increase the model explanatory and predictive power. Social networks are key in diffusion of information, ideas, opinions and the influence of individual nodes on the diffusion process (Madan and Pentland , 2009).

A socio-economic network is a relationship between a node and more than one node where at least one node enjoys a one-to-one relationship with one or more noded (Pani , 2008). The networks, whose primary action entails economic transaction embedded in social capital are called socio-economic network (SEN). A SEN is a set of nodes and their bidirectional relationships where for each one-to-one bi-directional relationship there exists two uni-directional relationships. SENs are stable but not static as the network relationships are dynamic and go through renovation, change or even disruption (Pani , 2008).

A study on a set of social network evolution and dynamics and how the structure is shaped by the incentives of the agents is the work of (Ehrhardt et al. , 2006). The models analyzes the idea of how the dynamics of the network struggle between volatility and the creation of new links. They observed that network dynamics exhibits three features, resilience, co-existence and sharp phase transition with positive feedback between link creation and inter-node similarity. A feedback mechanism in the network provides a powerful mechanism that effectively offsets the link decay as a result of volatility in the network (Ehrhardt et al. , 2006). Many social phenomena normally display an evolving inherent network dimension which provides an ideal platform for a range of social problems from spread of disease, to the establishment of research collaborations in both the scientific and industrial sectors, among other areas.

A dynamic social network model by Skyrms and Pemantle (2009) considers individual agents who interact at random, and those interactions are modeled as games. The payoff from the games determines which interactions are reinforced and the social network structure emerges from these stochastically evolving social network. They noted a difference between strategic dynamics by which individuals change their individual behaviors or strategies and the structural dynamics of the network. The social interaction structures that emerge tend to separate the agents into small interaction groups. Skyrms and Pemantle (2009) assumes that agents in the network make friends on asymmetric weights. Each agent goes out to visit some other agents and this choice is by chance that are determined by the relative weights each agent has assigned to the others. Let agent i assign weights $\{w_{i1}, \ldots, w_{im}\}$ to other agents with $w_{ii} = 0$. Then, the agent i visits agent j with probability

$$\text{Prob(agent } i \text{ visits} j) = \frac{w_{ij}}{\Sigma_k w_{ik}}$$

Even with this symmetry built at the starting point, different and varies types of structures emerge as as a consequence of the dynamics of the agents' learning behaviour. Choices made by each agent are independent of the choices made by each other agent.

Agent dynamics or social group drives the evolution of the social groups in a community. Chen et al. (2008) defines a social group as a collection of agents who share some common context with the dynamics being governed by the agents dynamics. The agents dynamic model developed has the ability to identify which parameters have a significant impact on the future evolution of the society. The agents are categorized as active, occasional or dormant. Stanley (2006) observes that the dynamics of societal cooperation have gradually incorporated more and more information about social network structures. For example, mathematical models on AIDS epidemic shows that the spread of HIV is sensitive to human behaviors including; the amount of risky behavior; the manner in which that risky behavior is distributed in the population; and the social network structures within which people practice those risky behaviors (Stanley , 2006).

A network approach by Allen and Babus (2008) to financial systems was to assess the financial stability and capture the externalities and risks associated with them. Mapping the financial institutions is an important step toward gaining a better understanding of modern financial systems. Here, the need for types of connections; the quality of the links; role of networks in mutual monitoring; gain an understanding of systemic risk; and how these dependencies stems from both the asset and liability side of the institutions balance sheet are discussed (Allen and Babus , 2008).

Social networks convey social capital which is a form of capital as with an asset, depreciates over time. Social capital represents as asset and might be an important determinant in some economic situations. It is more difficult to measure (Johnson , 2003). Social capital is also defined as social obligations which can be converted in certain conditions into economic capital. Thus, social connections or obligations formed through a social network can be maintained and used. Trust and civic co-operation is noted as a proxy for social capital and the former have significant impacts on aggregate economic activity.

The perspective on which social capital is built cannot be separated from an actor and its activities, including economic activities (Pani , 2008). The structural view of social capital supported by the social network analysis. Dynamic social networks are determined by stochastically evolving social network. Individuals interact at random, strategically changing positions to gain incentives.

When a network has new membership, the sum of social capital in the network exceeds the sum of individual investments. As occupational returns to social skills increase, social capital investment increases, indicating an association between social capital and social skills. This is also true between social capital and economic performance - as social capital accumulation patterns are consistent with the standard economic investment model (Glaeser et al. , 2002). The main predictions of this association are: the life-cycle effects which predict that social capital rises and then declines with age; individuals who work in occupations for which social skills are relatively important accumulate more social capital.

A relationship is observable between investment in social capital and economic growth of a firm (Pani , 2008). A SEN is where the primary action entails economic transactions, under the structure of the social capital. The network of relationships is induced by the investment strategies, individual or collective aimed at creating or establishing social relationships (Johnson , 2003). Social network analysis had provided a significant role in domains of biology, security, sales, terrorism, finance, and many more other areas. The availability of huge amount of social network data offers opportunities for studies and simulation of different nature of possibilities. These contribute to better understanding of the properties and the behaviour of social networks (Pupazan , 2011).

Agents in a network rationally form relationships based on the derived cost and benefits. There are three explanations offered as to why embedded resources in a network enhance the outcomes of agents actions; first, resources facilitate the flow of information. The social ties in strategic locations and positions provide individual agents with useful information, opportunities and choices that would otherwise be unavailable. Second, social ties can exert influence other agents in decision making process. Third, the resources tied up in the social network can be retrieved by individual relationships observable from the social credentials of an individual. This can create a social debt (Johnson , 2003; Nan , 1999).

The different characteristics of the agents in the system and uncertainty means that trust is a fundamental concern if effective interaction is to be achieved. HMM based trust model to focus on outcomes of the past interactions and interaction context that reflects on the dynamic behavior of an agent is modelled by (Liu and Datta , 2012). An investigation and utilization of the interaction contextual information as the observation is used to build a HMM. In order to achieve accurate prediction, information theory (that is, information gain) and machine learning technique (i.e. multiple discriminant analysis) are applied to select and process the contextual information. Liu and Datta (2012) observes that use of past transactions to model dynamic trust is appropriate in situations where an agent's behaviour is dynamic or changes frequently. However, if an agent changes in behaviour is relatively infrequent in changing patterns, then use of past observation sequence is not well suited to model dynamic trust.

Agents might not have had previous interactions and therefore, trust is a fundamental concern for effective interactions. Moe et al. (2008) used HMM for trust estimation, and developed a trust model that allows for dynamic behaviour based on HMM. That is, agents should decide how, when and with whom to interact with in a manner that will maximize their utility. Resnick et al. (2000) analyzes the eBay online auction site that has over four million auctions active at a time. They noted that the overall rate of successful transactions is high due to eBay use of a reputation system, called the feedback forum. This reputation system seeks to establish the shadow of the future to each transaction by creating an expectation that other people will look back on it. The seller reputation has a significant influence on online auction process and it is derived from the underlying network and visible to all the people in that system. A reputation system, called the feedback forum has been applied successfully by eBay system that has over four million auctions active at a time. The feedback forum assists in knowledge transfer and better understanding the behaviour of the individual in the network (Resnick et al. , 2000).

In a dynamic social network, reputation feedback systems can assist in knowledge transfer and better understand the properties and behavior of agents in these networks. The robustness of a social network assumes that social links are dynamic and allowed to change with and without constrains or restrictions.

2.1.1 Social media data

Social media has continued to gain widespread acceptance. For example, in the year 2012, Facebook had 1 billion users worldwide while in the same year, Twitter had an estimated 517 million users (Dewing , 2012). The social networks provide the social media data which refers to the wide range of internet based and mobile services data. The users of these services participate in online exchanges, join online communities or contribute user created content (Dewing , 2012). A number of factors has contributed to this rapid growth and embracement of social media services. These are; increased broadband availability, improvement of software tools, development of more powerful computers and mobile services (Dewing , 2012; Mustafa and Hamzah , 2011).

Table 2.1 shows the variables used in the studies of Daniel and Grissen (2015) for mobile phone data and Masyutin (2015) for the social network in credit scoring. The two studies offers promising insights on use of social media data in credit scoring process.

Table 2.1: Social media data variables used in credit scoring

Mobile Phone Data Variables	Social Network Variables
Age	Age
Gender	Gender
Top up and depletion patterns	Marital status
Mobility	Number of days since last visit
Patterns of handset use	Number of subscriptions
Strength and diversity of network connections	Number of days since the last post
Intensity and distribution over space and time	Number of user's posts with photos
Loan size	Number of user's posts with video
Loan term in days	Number of children
	Major things in life
	Major qualities in people

The availability of powerful and new innovations in data science tools for mining the internet content offers rich sets of data ideal for consumer credit lending. Evidence suggests that soft information can add value to financial institutions credit lending process by supplying the missing pieces that complement existing hard information used (Siva , 2010). There are over 2 billion people connected to the internet which implies an increase in online content and user interactions (Dubois et al. , 2011).

2.2 Trust and Reputation

A number of methods and models for computing trust and distrust have been developed (Dubois et al. , 2011; Guha et al. , 2004; Netrvalova and Safarik , 2011; Pupazan , 2011; Skyrms and Pemantle , 2009; Tang et al. , 2014). The works agree that trust levels of agents plays a central role in interactions of agents.

Modelling trust in complex dynamic environments is an important and challenging issue as intelligent agents strategically changes their behaviour for different reasons (Liu and Datta , 2012; Raghavan et al. , 2013; Skyrms and Pemantle , 2009). Trust is so fuzzy and personal that its not easy to compute it (Dubois et al. , 2011; Tang et al. , 2014). The set of incomplete information of each other in a network led to adoption of trust and reputation framework to maximize the security level by basing decision making on estimated trust values for network peers (Ehab and Sassone , 2013). The reputation model enhances the reliability and quality of trust judgment. Feedback information improves evaluation process about the trustee in the form of reputation reports (Ehab and Sassone , 2013). Reputation systems are easy to describe but the notion of trust and distrust is difficult to describe in a concise manner (Dubois et al. , 2011; Josang et al. , 2007).

Trust is the confidence in the ability of a person to be of benefit to trustworthy on something or someone at sometime in the future. Gambetta (1988) notes that trust is the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action both before he can monitor such action. Gambetta (1988) continues to note that reputation is a perception that an agent has of another's intentions and norms. Therefore, trust and distrust are estimated from the reputation ratings of the agents. (Netrvalova and Safarik , 2011) represent trust in the interval $(0, 1)$ where 0 represents complete distrust and value 1 blind trust. The model reflects members of social network and differentiates them according to their disposition to trusting somebody. A value of 1 indicates that the agent is highly trusted and hence blind trust (Netrvalova and Safarik , 2011). The work of Guha et al. (2004) highlights agents who optionally express some level of trust for the other agents in a network. The expressions become entries for a real valued matrix that is used to predict an known trust value between any two users. Dubois et al. (2011) computes trust using a path probability in a random graph. For each pair of users, (x, y), they placed an edge between them with some probability that depends on the direct trust between them denoted by t_{xy}. The rise of social networking has allowed people to indicate whom they trust and distrust creating links in the network. Trust assists the users to decide whom to accept information from and with whom to share information with (Dubois et al. , 2011).

Agent interactions require trust but Dubois et al. (2011) observes that knowing whom to distrust is equally important but is trickier to compute in a satisfying way. Guha et al. (2004) used a set of n users, each optionally expressing some level of trust and distrust for any other user. Tang et al. (2014) notes that distrust is a new dimension of trust. A difference between trust and distrust is that distrust information is publicly unavailable and social media services rarely implement distrust mechanism in their networks. Generally, trust and distrust are complex measures representing people's multi-dimensional utility function (Guha et al. , 2004; Josang et al. , 2007)

Reputation and trust systems produce a score that reflects the relying agent subjective view of the other agent trustworthiness. Transitivity is an explicit component in trust systems. The prediction of trust and distrust in social network, Dubois et al. (2011) refers to them as positive and negative trust with trust being transitive and distrust is not transitive. Trust is transitive while reputation is not. Josang et al. (2007) observes that the reputation score is seen by the whole community and take transitivity implicitly into account (Gambetta , 1988; Ganesh and Sethi , 2013; Josang et al. , 2007).

Social relationships changes continuously in a way correlated with the dynamical processes taking place during social interactions (Liu and Datta , 2012; Raghavan et al. , 2013; Skyrms and Pemantle , 2009; Zhao et al. , 2011). Interaction contextual information modeling is introduced by Liu and Datta (2012) to reflect an agents interactions dynamics better. Information theory is used to select features that generated a compact and effective feature vector for each agent interaction. Raghavan et al. (2013) incorporated social effects on a social interaction model using coupled HMM as the effects influence a user's activity based on the activity of a user's neighbour thus increasing the model explanatory and predictive power.

The static interactions of users in facebook using interaction graphs is the work of (Wilson et al. , 2009), while (Viswanath et al. , 2009) focuses on dynamics of user interactions in facebook. A decay in the amount of interactions between pair of users is noted due to network rapid changes over time. A model to infer relationship strength as the hidden cause of user interaction showed strong similarity between relationship and interactions (Xiang et al. , 2010). They noted that ancillary interaction information among the users can improve interaction modeling. The stronger the relationship, the higher the likelihood that a certain type of interaction will take place between the pair of agents (Zhao et al. , 2011).

2.2.1 Trust and economic performance

A study by Knack and Keefer (1997) noted that if the trust levels are high in an environment, the future actions of the agents in the network can be accomplished at lower cost. A fact that shows that social capital contributes immensely in measuring economic performance. Trust and civic cooperation were used to estimate economic performance with a strong association observed between economic performance, trust and civic norms. Government officials in societies with higher trust are perceived more trustworthy and this triggers greater investment and other economic activities.

Further, Knack and Keefer (1997) notes that trusting societies have stronger incentives to innovate and to accumulate physical capital with higher returns to accumulation of human capital. Social capital variables exhibit a strong and significant relationship to growth. The term *social capital* has no unified definition for its has been applied in varied fields from social sciences, to economics, political science, organizational sociology and so on (Sjoerd , 2004). At the individual level, social capital is formed through network participation and social interactions in groups. According to Sjoerd (2004), social capital is the features of social organization such as trust, norms, and networks that can improve the efficiency of society by facilitating co-ordinated actions.

2.3 Hidden Markov Model

A HMM tutorial by Rabiner (1989) opened a new frontier in this area of research with an analysis of HMM at depth. Different researchers have defined HMMs from different angles; HMM is the statistical tool for engineers and scientists to solve various problems (Bhusari and Patil , 2011); HMM is a state machine for a system adherent to a Markov process with unobserved states (Loni et al. , 2012). HMMS provide a flexible general purpose approach for modeling various dynamic systems that can be observed through univariate or multivariate time series (Lajos et al. , 2012).

HMMs have found a niche in different disciplines and have been applied with a lot of success (Bhusari and Patil , 2011; Chen et al. , 2008, 2007; Ching et al. , 2006; Crowder et al. , 2005; Davis et al. , 2005; Ehab and Sassone , 2013; Fonzo et al. , 2007; Liu and Datta , 2012; Mathew , 1997; Mhamanne and Lobo , 2012; Quirini and Vannucci , 2014; Srivastava et al. , 2008).

The popularity of HMM in bioinformatics with many software tools based on HMM is highlighted by (Fonzo et al. , 2007). A dependent hidden Markov model to analyze credit quality in discrete time with a Markov chain observed in martingale noise is proposed by (Korolkiewicz , 2010). The use of HMM to develop probabilistic model for social networking is presented in (Raghavan et al. , 2013). A proposed interactive hidden Markov model where the hidden states are affected by the observable states (Ching et al. , 2006). Credit card fraud is detected using HMM during transactions. Bhusari and Patil (2011) notes that the model helps to obtain a high fraud coverage combined with a low false alarm rate in the credit card transactions. (Lajos et al. , 2012) delves into the stock market with HMM to dynamically capture the behaviour of the various stock market equities and indices. Bilmes (2006) observes that HMM do not have a limitation for their applications. Netzer et al. (2008) uses HMM to model customer relationship dynamics. The effect of the encounters between the customer and the firm on customer-firm relationships and the customer's choice behavior are modeled.

HMM is dynamic in observing sudden downgrading of a customer credit worthiness (Quirini and Vannucci , 2014). The duo observed that HMM and related tools are essential to assess the credit risk in order to gain profitability in the complex and fluctuating credit market. Crowder et al. (2005) models the occurrence of defaults within a bond using HMM while Srivastava et al. (2008) uses HMM to model the credit card fraud detection system that is scalable for handling large volumes of transactions with high accuracy rate. Miller et al. (1999) used HMM for information retrieval by incorporating multiple word generation mechanism in the model.

The model by Korolkiewicz (2010) describes the inter-connected dynamics of user activity with the individual dynamics of each user being coupled to the aggregate activity profile of his neighbours in the network. A HMM to model trust in a dynamic environment that changes with time is the work of (Liu and Datta , 2012). They bypassed use of past interactions in modelling trust and introduced interaction contextual information of an agent in the system. This model is able to easily detect sudden changes in behaviour of agents in the network as trust in a social setting is dynamic. Static models are not able to dynamically capture these changes.

Agents dynamic behaviour is represented by a HMM as a trust and reputation model. The model enhances the trust evaluation using supplementary feedback reports about the trustee (Ehab and Sassone , 2013). Two components are key to the model in which there is a reputation reporting exchanged between the network peers; and a mixing scheme which uses multiple reputation reports about a trustee to evaluate the trust levels.

The work of Netzer et al. (2008) relaxes the assumption of homogeneity and uses a non homogeneous HMM for modeling customer relationship dynamics, where time varying covariates are investigated on its role in customer firm interactions. Ability to train HMM with multiple observers is the work of (Li et al. , 2000). Time varies with the actions of the multiple agents and observations of the emitted information. The independence-dependence property of the observations are characterized by the combinatorial weights, thus giving more freedom in making different assumptions in the study. HMM assumes that time is invariant but Chen et al. (2007) have relaxed that assumption in a multiple observer situation. Relaxation of the assumption makes the HMM more complicated but this work shows the versatility of the relaxed assumptions. Agent dynamics have multiple observations and this ideally works in our study. This relaxation was found not to satisfy the Markov property but proposed a method to find the maximum likelihood estimates by the Expectation Maximization algorithm.

Viterbi algorithm for HMM is based on the principles of dynamic programming. Karris (2007) explains that dynamic programming is based on Bellman‘s Principle of Optimality which states that: an optimum policy has the property that whatever the initial state and the initial decision are, the remaining decisions must constitute an optimum policy with regard to the state resulting from the first decision. A combination of stopping time and dynamic programming model has been applied to the valuation of American put options. These models check for optimality in exercising the options as they can be exercised any time until the expiry date. Equal spacing of the exercise dates for American option simplifies the notation and are less restrictive, with the price process of the option conditional on not having been exercised earlier (Haugh and Kogan , 2004).

In decision making, agents are faced with uncertainties due to lack of perfect information about the environment as noted in (Moe et al. , 2008). Each agent trust is needed in order to minimize the impact of the uncertainty and make optimal trust decisions over time. Agent's action choices is through the information gathered over time. Action choice being based on prediction of other agent's behaviour or directly on the reward received. Hassan and Nath (2005) applied HMM in stock market forecasting and showed a 100 percent accuracy in the prediction. The time series data was divided into two sets, one training set and one test (recall) set. Current price is linked to a smaller likelihood value, then this is used to predict the next day stock price.

In credit card fraud detection system, Srivastava et al. (2008) links the set of all possible types of purchase and line of business to be known by the bank in advance. A cardholder purchases depends on the need for procuring different types of items over a period of time, generating a sequence of transaction amounts. The transition in the type of purchase is considered as state transition in the model. Bhusari and Patil (2011) study on online banking fraud using HMM used the cardholder's spending habit to detect fraudlent transactions. HMM approach decreases the number of false positive transactions recognized as malicious by a fraud detection system even though they are really genuine. Spending are categorized into three profiles, namely, low, medium and high. They noted that the initial choice of parameters affects the performance of HMM algorithms. The model checks the upcoming transaction as fraudulent or not and decides to add new upcoming transaction to existing sequence or not, and a threshold level decides whether the transaction is genuine or fraudulent. (Mhamanne and Lobo , 2012) study online banking fraud and used HMM to detect and notify any fraud related online transactions.

A stochastic mesh method by Broadie and Glasserman (2004) solves a general optimal stopping problem of the American option with discrete exercise opportunity. The mesh method is flexible providing for lower, upper bounds and confidence intervals. The intervals are generated from combining high-biased and low-biased path estimator, and the algorithm converges with an increase in the number of mesh points and as computation effort increases. A forward algorithm to solve optimal stopping time is given by (Haugh and Kogan , 2004; Irle , 2006).

2.4 Consumer Credit Risk

Credit risk modeling continues to receive increased research interests through application of different techniques and approaches (Baesens and Gestel , 2009; Bhusari and Patil , 2011; Bucay and Rosen , 2000; Ching et al. , 2008; Crowder et al. , 2005; Denault et al. , 2009; Finger , 2000; Frey and Runggaldier , 2007; Giesecke , 2005; Gordy , 2000; Gurny and Gurny , 2013; Hodgman , 1960; Horkko , 2010; Madhur and Thomas , 2007; Robert et al. , 1996).

Consumer credit modeling has been considering each loan obligor in isolation. Lenders are more interested in the general characteristics of the portfolio for risk management purposes. The Basel Capital Accord, the bank's regulatory body emphasized this need of portfolio risk of retail loans. A need that has highlighted the importance of developing accurate estimates of default probability - accurate overall risk to minimize the capital required to cover these expected and unexpected default losses (Schweitzer et al. , 2009; Thomas et al. , 2005). The Basel regulators imposed a corporate credit risk model as consumer credit has not developed its own models of the default risk of a loan portfolio. Basel II Accord put the spotlight to focus on portfolio risk (Capuano et al. , 2009; Schweitzer et al. , 2009; Skyrms and Pemantle , 2009; Thomas et al. , 2005). Further, with increase in volume of securitization of retail loans, it means bundling together different loans and selling them at one price, which is misleading - there is no model to price these bundled portfolio (Thomas et al. , 2005).

Recent advances in credit risk modeling and recent economic crisis due to systemic complexity of economic networks is discussed (Capuano et al. , 2009; Schweitzer et al. , 2009). Under the prevailing credit risk models assumptions of independence of customers, socio-economic networks assumes that agents are interdependent. Socio-economic networks offers solution due to their complexity, dynamism and robustness. Agents interactions in the network are stochastic and bound in time and space, with varied effects which increases with increasing coupling strength between nodes.

A model by Eisenberg and Noe (2001) featuring a network among financial systems, where the properties of inter corporate cash flows has cyclical interdependence determined endogenously by clearing vectors is presented. The model has clearing vectors representing payment vector from nodes in the financial system to other nodes that satisfy the conditions of proportional repayments of liabilities in default (Schweitzer et al. , 2009). The clearing vector computation is through a fictitious sequential default algorithm, where there is a process of dynamic adjustment in which the set of defaulting firms at the start of each round is fixed by the adjustments of the system in the previous round. The system is cleared in each round by assuming that only nodes that defaulted in the last round default. If no new defaults occur, the algorithm terminates. Otherwise, the new wave of defaults is recorded and the process is iterated again, until all nodes are cleared. The algorithm yields the clearing vector, and this exposure in a given node indicates the likely of the node in the system to default given other firms. 'Waves' of default measure the systemic risk which can induce a firm in the system to fail (Eisenberg and Noe , 2001).

Survival analysis approach in credit risk modeling to assess aspects of profit and default is the work of (Stepanova and Thomas , 2002). A coarse classification approach to the characteristics of the customers was developed. Residuals extracted are then tested for fitness in the model. The competing risks approach of survival analysis to dealing with both defaults and early completion of loans is the model of (Banasik et al. , 1999). They tried to estimate how long until some events occurs even though in many cases the event will not occur. Competing risks approach assumes that several reasons exists as to why loan repayments are completed before the original intended term (Banasik et al. , 1999).

The work of Robert et al. (1996) is on credit risk and credit scoring with emphasis on delinquent and default cases. A borrower's credit history is summarized by a credit history score which is a strong predictor of loan delinquency. Earlier credit history assessment was subjective but introduction of a statistically derived measure of the credit risk increased the observable probability of default. (Bucay and Rosen , 2000) developed a simulation based framework to estimate the one period credit loss in a retail loan portfolio. The usefulness of the model was demonstrated by estimating one-year credit losses. The influence of economic cycles was introduced in the model. The results for this study showed that some refinement in the modeling of the portfolio may lead to the greater improvement. In conclusion, they observed that application of portfolio credit risk models to retail portfolios is in its infancy and much more research is required.

Traditional Microfinance has presented a solution to the repayment problem as the community members aid in monitoring the repayment process. This is one of the major challenges facing Microfinance institutions in their quest to provide credit facilities, as most of its clients have scarce financial information (Serrano-Cinca et al. , 2013). A study by the Centre for the Study of Financial Innovation states that Microfinance industry is faced by two main threats; credit risk which is worsened by the over-indebtedness of its clients; and the perception that the Microfinance industry has lost sight of its social purpose (Serrano-Cinca et al. , 2013). The credit scoring systems of microfinance institutions, if they exist, are strictly financial. They propose a decision support system to facilitate microcredit underwriting process and estimate the social impact of the microcredit. The system is partly based on the Social Net Present Value (SNPV) which is expressed as; SNPV $= \Sigma_{t=0}^{n} \frac{S_t}{(1+r)^t}$, where $S =$ social impact, $r =$ discount rate, $t =$ time and $n =$ number of periods. This component is to formalize the use of social aspects in the credit scoring process and capture the social impact the loan has on the borrower (Serrano-Cinca et al. , 2013).

The credit risk models in use face challenges of performance deterioration over time. This calls for periodic validation to arrest this shortcoming of the models to retain their accuracy, completeness and timelines of the information used to generate the scores (Robert et al. , 1996). The use of historical data contributes to this deterioration as it assumes that credit quality is time independent (Capuano et al. , 2009). A need remains to have models with generalization capability and applicability but this is becoming more elusive (Li and Zhong , 2012). Another factor is that no mathematical system model is perfect as these models only depict those characteristics of direct interest to the modeler. The traditional credit scoring has relied on the traditional underwriting factors which are character, capacity, collateral, capital and conditions. PWC (2015) outlines many of the benefits of using the big data from the social media to supplement the traditional credit scoring process. Recent trends in the consumer finance calls for innovative solutions due to shifting demographics and credit trends, calling for shift in credit underwriting strategies by use of social media (PWC , 2015). The young consumers and many households in developing countries lack financial histories creates a need for this innovative approach.

One of the major revolution on use of available data in the social circles is the work of (Daniel and Grissen , 2015). They used behavioral signatures in mobile phone data to predict default with accuracy than the approach of credit scoring using financial histories. The method was found to be promising even for the poor borrowers whose mobile phone usage is also very sparse (Daniel and Grissen , 2015). It was noted that a subscriber may be able to manipulate their score if they knew the algorithm, but the problem can be overcomed by combining with other credit scoring techniques. The algorithm reduces defaults by 41 percent while still accepting 75 percent of the borrowers.

The use of social data from Russia's most popular social network to discriminate between solvent and delinquent debtors of credit organizations is the work of (Masyutin , 2015). The social network data was found to better predict fraudulent cases rather than ordinary defaults, thus ideal to use in enriching the classical application scorecards. Wei et al. (2015) considers a number of models to compare the accuracy of customer scoring obtained with and without network data. They analyzed the benefits of collecting information from consumer's network where people with an above average chance of interacting with others with similar creditworthiness creating the social scoring. An increase in inclusion of population with limited personal financial history to be offered credit increased due to social scoring.

The social scoring methods are yet to gain popularity, but as they continue to do so, consumers may adapt their personal networks that may affect their scores. The benefits of using social media data in credit scoring increases when it involves networks of ties that exhibit great homophily (Wei et al. , 2015). The advantage is that as more and more people are getting connected to the internet every day, user interactions and online content continue to increase (Dewing , 2012; Dubois et al. , 2011).

Credit scoring systems have been built to answer the question of finding out how likely an applicant is going to default at a given time in future (Banasik et al. , 1999). As the credit status is dynamic, we refine the question to ask, if the applicant will default, when will the default occur? This question has various answers not just a single answer of yes/no. The use of historical data sometimes becomes a problem since the customer data is censored in that they are no longer in the bank's database (they have either paid bask the loan, or have died).

2.4.1 Default process

Let $Z(t)$ be the default process that takes the value of one if the default occurs and zero if no default occurs at time t. The ability to model the Knowledge of default path is equivalent to knowing the exact time of default. Since we have partial knowledge, the situation at time t is that either the obligor has defaulted or not defaulted (David , 2004).

The default process has numerous studies with different variations on how default is observed (Cetin et al. , 2004; Crowder et al. , 2005; Denault et al. , 2009; Eisenberg and Noe , 2001; Finger , 2000; Giesecke , 2005; Giesecke and Kim , 2010; Gurny and Gurny , 2013; Horkko , 2010; Iqbal and Ali , 2012; Kealhofer , 2003; Moffatt , 2005).

A clearing vector in a financial system represents the vector of payments from nodes to nodes in the system is the work of (Eisenberg and Noe , 2001). A fictitious sequential default algorithm is used to estimate the defaults at any given time in a network of financial firms that are part of a single clearing mechanism. The issue of cyclical interdependence is introduced to model the default process of the firms. Koyluoglu and Hickman (1998) considers the issue of default barrier based on stochastic and deterministic approach. A default threshold problem as a stochastic process with the inclusion of systemic risk factors is undertaken. Normality is assumed and it borrows widely from the Merton's model of a firm's capital structure. Joint-default situation among the obligors in a portfolio correlates to the extent of which the obligors asset value changes.

The idea of competing risks is employed when two possible outcomes are considered: default and early payoff (Stepanova and Thomas , 2002). Approach of how survival analysis proves useful in dealing with defaulters and early completion of loans is given (Banasik et al. , 1999). The approach is undertaken for repayer defaulting or paying off early by building survivor function models to estimate the distribution of each obligor. (Kealhofer , 2003) considers default as a binary event; it either occurs or it does not occur. The KMV model is used based on default-predictive power test that characterizes the relative ability of a default risk measure to correctly identify companies that subsequently default versus incorrectly identifying companies as likely defaulters that do not default.

The dependencies between different risks in a life insurance portfolio are modeled and analysis undertaken using survival probabilities (Dhaene and Goovaerts , 1996). Assumption is that if a person with lower probability survives, and then one with higher survival probability will survive. If a person dies, then all persons with lower survival probabilities will die too, as dependencies between individual risks being conditional. Customers' differences in application characteristics have different survival times. The customers‘ credit performance data normally recorded monthly means that several failures are observable at one time. If default probability is realized, then defaults are conditionally independent but high volatility observed if the defaults induce stronger correlations (Finger , 2000; Stepanova and Thomas , 2002). A high default rate would imply a generally decreased credit quality of other obligors who did not default. The impact would then be that the default rate for the second period would have a tendency to be high.

The estimate of the default probability and its dynamics through time is very important. Four approaches for estimating creditworthiness and default probabilities are given in (Denault et al. , 2009). The first one, default probability is estimated using the average frequency with which obligors of the same rating have defaulted. Moody's and Standard and Poor's collect data to perform these ratings estimates. The second approach uses statistical techniques and data from the balance sheet, current market conditions or past performance to estimate the probability that a firm will default. A third group uses the structural bond-pricing approach started by Merton in 1974. The fourth consists of models from the reduced form approach pioneered by Jarrow and Turnbull in 1995 and Duffie and Singleton in 1999 (Denault et al. , 2009).

The approach by Finger (2000) is to assign obligors a standard Wiener process and a minimum threshold, below which implies a default. Defaults are conditionally independent and the survival probabilities arise from expectations over the conditional probability, which evolves according to a stochastic differential equation. Two important general aspects of survival analysis which are connected to the use of stochastic processes are undertaken (Aalen and Gjessing , 2005). One is the issue of time. Decoupling statistical analysis from the development over time assumes that no changes take place when time passes. Time is relegated at the back instead of being the major parameter of survival data to capture and emphasize the changes over time. Second, models that allow for fruitful speculations on underlying mechanisms should be applied much more.

A HMM to calculate the likelihood threshold of human gestures is undertaken by (Lee and Kim , 1999). An ergodic model is developed where each state can be reached by all other states. Output observation probabilities and self transition probabilities are kept as in the gesture models. Lee and Kim (1999) noted that maintaining the self transition probability and the output probability distributions makes the states represent any sub pattern of reference patterns and the ergodic structure makes it match well with any patterns generated. Reduced forward transition probabilities makes the likelihood model, given a gesture pattern smaller than the dedicated gesture model. Thus, the likelihood can be used as an adaptive threshold (base-line) for selecting the proper gesture model.

The effects of threshold in credit card fraud detection system is investigated by (Alese et al. , 2012). They noted that different methods have been implemented to detect fraud but a threshold value add an extra advantage to detect anomalies in on-line transactions. They implemented a method of selecting adaptive/dynamic threshold values that are based on individual card holder spending profile. The process of detecting fraud in credit card on-line transaction involves training of data with HMM algorithms, namely, forward-backward for non-optimized and Baum-Welch for optimized HMM. Prediction of hidden states via Viterbi for non-optimized states and posterior-Viterbi for the optimized states.

2.5 Simulation

Simulation techniques continue to cover the gap where real data is missing. The advances in new methods and different applications continue to be evident (Capuano et al. , 2009; Duan and Simonato , 1998; Eisenberg and Noe , 2001; Haugh and Kogan , 2004; Johnson , 2003; Samik , 2008; Stanley , 2006; Terejanu , 2002).

The standard MCS procedure for computing the prices of the derivative securities is modified in (Duan and Simonato , 1998). This modification imposes the martingale property on the simulated sample paths of the underlying asset price, a procedure they referred to as the empirical martingale simulation (EMS). Simulation based on EMS yields results with substantial variance reduction, and the method dominates the conventional simulation methods in terms of computing time and price accuracy. MCS is generally numerically intensive if a high degree of accuracy is desired. The simulated paths for the underlying asset price almost always fail to posses the martingale property even though the theoretical model does. When EMS was used, even for small simulated samples, no statistically significant biases was observed. A note of concern was that EMS cannot provide a standard error estimate of its Monte Carlo price from using only one simulated sample (Duan and Simonato , 1998).

A simulation approach by Eisenberg and Noe (2001) for systemic risk in financial systems shows that there exists a unique clearing vector for a complex financial system. The algorithm clears both the financial system in a computationally efficient fashion and provides information on the systemic risk faced by the individual system firms.

Applications of simulation are taking ever new dimensions. Network finance analysis, financial systems contagion, economic networks, social capital formation, disease spread simulations, investment decisions, and other network applications (Capuano et al. , 2009; Eisenberg and Noe , 2001; Johnson , 2003; Stanley , 2006). The availability of powerful computing capabilities means that simulation techniques have an increasing use due to their diverse applicability and versatility. Stanley (2006) observes that with today's computer firepower, simulation is being used on individual agent based models and the results compared with those of differential equation models. But, the use of stochastic simulations that account for the social network of individual agents is crucial in providing a deeper understanding of the interactions between the individuals.

2.5.1 Conclusion

The agent dynamics in the social and economic network assumes that the actions of the agents emits a signal that is stochastic, non stationary and is corrupted with noise due to observations distortions and information asymmetry. Capturing this set of information increases the chances of effective modeling of the social and economic network and its effect on the credit quality of the obligors in the loan portfolio. Hidden Markov model classifies the agents into the different credit quality levels after emitting the CQS.

Chapter 3

Mathematical Preliminaries

This chapter is an highlight of the different mathematical techniques and methods that are used in this study. The areas considered are SVD, Markov process, stochastic process, MCS and reputation computation.

3.1 Singular Value Decomposition

We begin with an introduction of some of the basics of matrices that are important to keep in mind (Carla and Mason , 2012);

(i) A square matrix is symmetric if $\mathbf{A}^T = \mathbf{A}$

(ii) A square matrix $\mathbf{A}$ is orthogonally diagonalizable if there exist an orthogonal matrix $\mathbf{U}$ and a diagonal matrix $\mathbf{S}$ such that, $\mathbf{A} = \mathbf{USU}^T = \mathbf{USU}^{-1},$ where $\mathbf{U},$ $\mathbf{S}$ and $\mathbf{A}$ have all the same size

(iii) A vector norm is a function $||.|| : \mathbb{R} \to V$ that assigns a real-valued length to each vector in V and satisfies the following conditions

- $||y|| \geq 0$ and $||y|| = 0$ if and only if $y = 0$
- $||y + z|| \leq ||y|| + ||z||$
- $||\lambda y|| = |\lambda|||y||$

(iv) The matrix $2-$norm is the maximum stretch factor for the length of a vector after applying the matrix to it. We have $||\mathbf{A}||_2 = \sqrt{\sigma_{\max}(\mathbf{A})}$ and $||\mathbf{A}^{-1}||_2 = \frac{1}{\sqrt{\sigma_{\min}(\mathbf{A})}}$

The singular value decomposition is a matrix factorization method and has been used widely in different applications ever since an efficient algorithm for its computation was developed. The variety of applications are in engineering, chemistry, ecology, geology, geophysics, biomedical, scientific computing, automatic control and many other areas (Carla and Mason , 2012; Kalman , 1996; Lee et al. , 2013; musco , 2015; Sadek , 2012; Soman et al. , 2009). SVD is a matrix factorization technique that is stable and effective method to split the system into a set of linearly independent components, each of them bearing its own energy contribution (Sadek , 2012). SVD is used for optimal low rank approximation and a partial SVD can be used to construct a rank k approximation (Kalman , 1996).

SVD was discovered over 100 years ago independently by Eugenio Beltrami $(1835 - 1899)$ and Camille Jordan $(1838 - 1921)$. James Joseph Sylvester $(1814 - 1897)$, Erhard Schmidt $(1876 - 1959)$, and Hermann Weyl $(1885 - 1955)$ also (Carla and Mason , 2012) discovered the SVD using different methods . SVD is a powerful technique in matrix computations and analysis (Carla and Mason , 2012; Kalman , 1996; Leach , 1995);

(i) It reduces high dimensional, multidimensional and highly variable set of data to a lower dimensional space that exposes the substructure of the original data more clearly.

(ii) SVD has interesting and attractive algebraic properties and conveys important geometrical and theoretical insights about linear transformations. It exposes the geometric structure of a matrix.

(iii) SVD is closely related to the theory of diagonalizing a symmetric matrix

(iv) SVD is a numerically reliable estimate of the effective rank of a matrix. Even for linearly independent columns, the dependencies can be detected.

(v) Availability of algorithms to compute SVD with low computer resource utilization even with large matrices. A superhero in the fight against monstrous data that is available in every scientific discipline

(vi) Employed in a variety of applications such as least squares problems, noise signal filtering, time series, etc

SVD can be applied to any type of matrix from square, to rectangular, to Hermitian matrices (those that are identical with their conjugate transpose). Let $\mathbf{A}$ be an $m \times k$ matrix which can be represented as the product of two orthonormal matrices $\mathbf{U}$ and $\mathbf{V}$ and a diagonal matrix $\mathbf{S}$ (Carla and Mason , 2012);

$$\mathbf{A} = \mathbf{U}_{m\times m}\mathbf{S}_{m\times k}\mathbf{V}^T_{k\times k} \tag{3.1}$$

This is expressed in matrix form as;

$$\mathbf{A}_{m\times k} = \begin{pmatrix} | & & | \\ u_1 & \dots & u_m \\ | & & | \end{pmatrix} \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \end{pmatrix} \begin{pmatrix} - & v_1^T & - \\ & \vdots & \\ - & v_n^T & - \end{pmatrix}$$

The columns of $\mathbf{U}$, that is u_i are the eigenvectors of $\mathbf{A}\mathbf{A}^T$, and the columns of $\mathbf{V}$, that is v_i are the eigenvectors of $\mathbf{A}^T\mathbf{A}$. The singular values, σ_i on the diagonal of $\mathbf{S}$ are the square roots of the nonzero eigenvalues of both $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$, which are ordered decreasingly (Carla and Mason , 2012). We note that $\mathbf{U}_1 \dots \mathbf{U}_r$ spans the column space of $\mathbf{A}$ and $\mathbf{U}_{r+1} \dots \mathbf{U}_k$ spans the null space of matrix $\mathbf{A}^T$. The singular values are at most $\min(m, k)$ with $r \leq \min(m, k)$. For matrix $\mathbf{V}$, we have $\mathbf{V}_1 \dots \mathbf{V}_r$ spans the column space of matrix $\mathbf{A}^T$ and $\mathbf{V}_{r+1} \dots \mathbf{V}_m$ spanning the null space of matrix $\mathbf{A}$. The SVD expansion is (Leach , 1995)

$$\mathbf{A} = \mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^T + \mathbf{U}_2\mathbf{S}_2\mathbf{V}_2^T + \dots + \mathbf{U}_m\mathbf{S}_m\mathbf{V}_m^T \tag{3.2}$$

An important measure of the linear independence of the columns of the matrices in SVD is the condition number. Let $\mathfrak{C}(\mathbf{A})$ be the condition number which is a measure of linear independence between the column vectors of the matrix $\mathbf{A}$. then

$$\begin{aligned} \mathfrak{C}(\mathbf{A}) &= \frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}, \quad \text{with } \mathfrak{C}(\mathbf{A}) \geq 1 \\ &= \frac{||\mathbf{A}||_2}{||\mathbf{A}^{-1}||_2} \end{aligned} \tag{3.3}$$

If $\mathfrak{C}(\mathbf{A})$ is close to 1, the columns of $\mathbf{A}$ are very independent. If $\mathfrak{C}(\mathbf{A})$ is very large, the columns of matrix $\mathbf{A}$ are nearly dependent (Carla and Mason , 2012; Leach , 1995). We are interested in these three matrices, $\mathbf{U},\ \mathbf{S},$ and $\mathbf{V}$ in the estimation of what is referred to as the credit risk analysis factors. The factors are estimated from the reputation ratings of the agents, interactions, relationships, the private data and the demographic variables of the agents.

3.2 Stochastic Process

We lack sufficient data to exactly estimate the agent interactions and how the system behaves. The system is dynamic, complex and evolves with time to changing action choice of the agents.

3.2.1 Random variable

A HMM has a hidden Markov chain which is a stochastic process with a sequence of random variables. A random variable takes on values with certain probabilities, and might or might not have the ability to influence each other. If $\mathbb{Y}$ and Y are two random variables, and (Ephraim and Merhav , 2002; Ross , 2007),

$$P(\mathbb{Y} = \mathcal{Y},\ Y = y) = P(\mathbb{Y} = \mathcal{Y})P(Y = y)$$

then $\mathbb{Y}$ and Y are statistically independent - denoted as $\mathbb{Y} \perp Y$. If the two random variables are not independent of each other depending on the knowledge of a third random variable, then (Ephraim and Merhav , 2002)

$$P(\mathbf{Y} = \mathfrak{Y},\ Y = y|\mathbb{Y} = \mathcal{Y}) = P(\mathbf{Y} = \mathfrak{Y}|\mathbb{Y} = \mathcal{Y})P(Y = y|\mathbb{Y} = \mathcal{Y})$$

is the conditional independence and is different from unconditional (for marginal) independence, and it might be true that $\mathbf{Y} \perp \mathbb{Y}$ but not true that $\mathbf{Y} \perp Y|\mathbb{Y}$.

A discrete time stochastic process is a collection $\{Y_t\}$ for $t \in [0, T]$ of random variables ordered by the discrete time index t. The distribution for each of the variables Y_t can be arbitrary and different at each time t.

Definition 3.1 *A stochastic process $Y = \{Y_t, t \in [0, T]\}$ is a collection of random variables with index set I, where t is the time. A realization of Y is called a sample path.*

A discrete time stochastic process $\{Y_t\}$ is said to have independent increments if for all $t \in [0, T]$, the random variables (Ephraim and Merhav , 2002),

$$Y(t_1), Y(t_2), \ldots, Y(T))$$

are independent. It is said to possess stationary increments if $Y(t+\tau) - Y(t)$ has the same distribution $\forall t$ and the distribution depends only on τ.

Definition 3.2 Stationary Stochastic Process. *The stochastic process $\{Y_t : t \in [0, T]\}$ is said to be (strongly) stationary if the two collections of random variables (Ephraim and Merhav , 2002)*

$$\{Y_1, \ldots, Y_t\} \quad and \quad \{Y_{t+\tau}, \ldots, Y_T\}$$

have the same joint probability distributions for all t and τ

3.2.2 Martingale

Martingales are fundamental to the analysis of stochastic processes as they are random variables whose future variations are completely unpredictable given the current information set (Neftci , 2000). A martingale is always defined with respect to some information set and some probability measure. Changes in the probabilities associated with the process will make the process under consideration cease to be a martingale. The ability of the loan obligor is not fully known given the current information but we know the status at any given time t in the process of loan repayment. HMM can track the possible signs of difficulty of the agents in meeting their future obligations with the bank.

Let Y be a random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G}$ be a sub$-\sigma-$field of $\mathcal{F}$. The conditional mean $E\{Y|\mathcal{G}\}$ exists if $E\{|Y|\} < \infty$(Neftci , 2000)

Definition 3.3 *Let $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_T\}$ denote a sequence of sub$-\sigma-$fields of $\mathcal{F}$. The sequence $\mathcal{F}$ is called a filtration if $\mathcal{F}_t \subseteq \mathcal{F}_{t+\tau},\ t \in [0, T]$ and that $\mathcal{F}_T \subseteq \mathcal{F}$, (Neftci , 2000).*

Definition 3.4 *If $\mathcal{F} = \{\mathcal{F}_t,\ t \in [0, T]\}$ is a filtration, and $Y = \{Y_t,\ t \in [0, T]\}$ is a discrete time stochastic process, then Y is said to be adapted to $\mathcal{F}$ if Y_t is $\mathcal{F}_t$ measurable for each t, (Neftci , 2000).*

Let $Y = \{Y_t,\ t \in [0, T]\}$ denote a random process on the probability space. The process is said to be adapted to its natural filtration $\{\mathcal{F}_t,\ t \in [0, T]\}$ if Y_t is $\mathcal{F}_t$ measurable $\forall t$. Intuitively, $\mathcal{F}$ is the collection of events whose occurrence can be determined from observations of the process up to time t and an $\mathcal{F}-$measurable random variable is one whose value can be determined by time t. If Y is any random variable, then $E[Y|\mathcal{F}_t]$ is the 'best' estimate of Y based on observations of the process up to time t(Neftci , 2000).

Definition 3.5 *A discrete time stochastic process $Y = \{Y_t, t \in [0, T]\}$ is called a martingale with respect to the filtration $\mathcal{F} = \{\mathcal{F}_t, t \in [0, T]\}$, (Neftci , 2000).*

(i) Y is adapted to $\mathcal{F}$, that is, Y_t is $\mathcal{F}_t$ measurable

(ii) $E[Y_t] < \infty$

(iii) $E[Y_{t+1}|\mathcal{F}_0, \mathcal{F}_1, \ldots, \mathcal{F}_t] = Y_t,\ a.s,$ the best forecast of unobserved future value is the last observation on Y_t.

A stochastic process $Y = \{Y_t, t \in [0, T]\}$ is called submartingale if $E[Y_{t+1}|\mathcal{F}_t] \geq Y_t$ and is a supermartingale if $E[Y_{t+1}|\mathcal{F}_t] \leq Y_t$. Y_t is said to be a martingale with respect to $\mathcal{F}_t$.(Neftci , 2000). A Martingale, (1) makes the expected future value conditional on its present value or on the set of information that is known. (2) is not expected to drift upwards or downwards and thus it is a notion of a fair game. (3) is always defined with respect to some information set, and with respect to some probability measure (Bilmes , 2006; Neftci , 2000).

Martingale is invariant to certain operations that would destroy more classical relations like independence. Two states are considered given the number of individuals and a certain number of transitions from one state to the next in a given time interval. They observe that censoring can be incorporated easily. Martingale theory provides a fertile environment for discussing stochastic variables in a continuous time. Increments of a martingale should be totally unpredictable, no matter how small the time interval is (Neftci , 2000). Doob-Meyer decomposition implies that, under some general conditions, an arbitrary continuous-time process can be decomposed into a martingale process. The theory caters for time dynamics and this blends in well with stochastic environments. As we are working with discrete partition of a continuous time interval, the decomposition is important. For example, in the case of observed asset prices, occasional jumps and upward trends are observed at the same time, which can be converted into martingales. It is under the martingale principles that the stopping time is implemented (Neftci , 2000).

3.2.3 Stopping time

Optimal stopping times are generally obtained by using dynamic programming approach on which Viterbi algorithm of the hidden Markov model is based. The stopping times are special type of random variables that assume as outcomes random time periods, $\tilde{\tau}$. If $\tilde{\tau}$ is a stopping time, then $\tilde{\tau}$ is random and the range of its values is $[0, T]$. When the outcome is observed, $\tilde{\tau} = t$ (Neftci , 2000). This tracks the time a loan obligor pays off all the money or defaults in the payment process given a certain default level or the loan reaches its maturity time, T.

If $(Y_t,\ t \in [0, T])$ is a stochastic process, then the non-negative integer valued random variable T is a stopping time for Y. If the event $\{T = t\}$ depends only on $(Y_0, Y_1, \ldots, Y_T)$; and does not depend on $\{Y_{t+k};\ k \geq 1\}$

Definition 3.6 *A non-negative integer-valued random variable T is called a stopping time with respect to $(X_r;\ r \geq 0)$ if, $\forall\ \ \tau$, the event $T = \tau$ may depend only on $\{Y_0, Y_1, \ldots, Y_\tau\}$ and does not depend on $Y_{\tau+m};\ m > 0$.*

Theorem 3.1 *Optional Stopping Theorem*
Let $\{Y_t : t \in [0, T]\}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which is a martingale sequence with respect to the filtration $(\Omega, \mathcal{F}_t, \mathbb{P})$ and $0 \leq \tilde{\tau}_1 \leq \tau_2 \leq T$ be two bounded stopping times. Then $E[Y_{\tilde{\tau}_2} | \mathcal{F}_{\tilde{\tau}_1}] = Y_{\tilde{\tau}_1}$
(Stirzaker , 2005).

Theorem 3.2 *Let $(Y_t;\ t \in [0, T])$ be a martingale, and let T be a stopping time for Y. Define*

$$Z_t = \begin{cases} Y_t, & \textit{if } t \leq T \\ Y_T, & \textit{otherwise} \end{cases} \tag{3.4}$$

So, Z_t is essentially Y_t stopped at T. Then, Z_t is a Martingale with respect to Y_t (Stirzaker , 2005).

3.3 Markov Process

Markov chain is named after Andrei A. Markov $(1856 - 1922)$ who first published his result in 1906. His research work on Markov chains launched the study of stochastic processes that led to a lot of applications.

A Markov process is a stochastic (random) process in which the probability distribution of the current value is conditionally independent of the series of past value. This characteristic is called the Markov property, that is, evolution occurs in a discrete time and the probability distribution of a state at a given time is explicitly dependent only on the previous state. A Markov process is the most useful and important class of stochastic models (Stirzaker , 2005).

A Markov has a property that the conditional probability distribution of future states of the process given the present state and all past states, depends only upon the present state and only on the recent past state. For the Markov chain, they have a well developed theory that allows us to do computations. Markov chains have successfully modeled a huge range of scientific and social phenomena such as in biology, economics and physical systems as they combine tractability with almost limitless complexity of behaviour (Stirzaker , 2005).

Definition 3.7 *A process $\{Y_t : \ t \in [0, T]\}$ is said to be a Markov process if it satisfies the Markov property,*

$$\begin{aligned} P(Y_t = y_t | Y_{t-1} = y_{t-1}, , \ldots, Y_1 = y_1) \\ = P(Y_t = y_t | Y_{t-1} = y_{t-1}), \ \ \forall t, \end{aligned} \tag{3.5}$$

If the state space $S = \{1, 2, \ldots, M\}$, then, it is a finite Markov chain and since we are dealing with chains, Y_t can take discrete values from a finite or a countable infinite set. This countable set S is called the state space of the chain (Stirzaker , 2005).

A Markov chain is a discrete time stochastic process with the Markov property. It is a first order Markov process in which the probability of the next 'future' state is dependent only on the present state and the 'past' states are irrelevant once the present state is given. Starting from a given initial state, the consecutive transitions from a state to the next one produce a time evolution of the chain that is completely represented by a sequence of states that a priori are to be considered random. The Markov process characterizes our

HMM where the stochastic process is transformed into the observed outcome through the stochastic process of choice made by the agents in the SEN (Volker , 2010).

Markov chain is viewed as a finite state automata with probabilistic state transitions that are either infinite or finite. Evolution of the chain is determined by the state transition probabilities that are assumed to be time-independent. The assumption of independence on time gives rise to time-homogeneous (or just homogeneous) Markov chain. This transitions form a stochastic transition matrix (Bilmes , 2006).

If $P(Y_t = i|Y_{t-1} = j) = P_{ji}$, then the chain is called homogeneous. The array $P = (P_{ji}),\ j, i \in S$ is called the matrix of transition probabilities; the chain Y is said to be the Markov (P). If $\sum P_{ji} = 1$, then the matrix P is said to be stochastic, else if $\sum P_{ji} = 1$, and $\sum P_{ij} = 1$, then it is said to be doubly stochastic.

3.3.1 Markov assumption

The h^{th}-order Markov chain can always be converted into an equivalent first-order Markov chain. Given a sequence of data $\hat{Y} = \{y_1, \ldots, y_T\}$ sampled from the random variable Y the likelihood can be written as

$$P(Y|\mathcal{M}) = \prod_{t=1}^{T} P(y_t)$$

Where $\mathcal{M}$ are the number of states in the Markov chain. When working with sequential data (correlation among subsequent samples) the identical independent distributed assumption is no longer a good approximation. There is time correlation between the different samples (Bilmes , 2006). Instead of assuming independence, we assume a casual dependence among the given samples;

$$\begin{aligned} P(Y|\mathcal{M}) &= P(y_t|y_T, \ldots y_1) \qquad (3.6) \\ &= P(y_t|y_{t-1}, \ldots, y_1)P(y_{t-1}|y_{t-2}, \ldots, y_1)P(y_2|y_1) \\ &= P(y_1)\prod_{t=2}^{T} P(y_t|y_1, \ldots, y_{t-1}) \\ &= P(y_1)\prod_{t=2}^{T} P(y_t|y_{t-1}) \end{aligned}$$

This is computationally hard and we simplify it by applying a Markov assumption

$$P(y_t|y_{t-1}, \ldots, y_1) \approx P(y_t|y_{t-1})$$

and is called the first order Markov assumption as the outcome of y_t is only dependent on the outcome at y_{t-1}. The observation variable Y in Markov models can have an observation sequence $y_1, \ldots, y_T$ where each of the variables y_t may take one of the M states $(S_1, \ldots, S_M)$. The likelihood of the discrete samples $Y = \{y_1, \ldots, y_T\}$can be calculated as

$$P(Y|\mathcal{M}) = P(y_1 = S_i) \prod_{t=1}^{T} P(y_t = S_j | y_{t-1} = S_i)$$

Therefore, a h^{th} order Markov chain may be transformed into a first order chain. Assuming a first-order Markov chain possess a sufficient states, there is no inherent fidelity loss when using a first-order as opposed to an h^{th} order Markov chain (Bilmes , 2006; Koubaa , 2008). We will use $P(y_t = j | y_{t-1} = i)$ and $P(y_t = S_j | y_{t-1} = S_i)$ interchangeably to imply a transition from state i (S_i) at time $t-1$ to state j (S_j) at time t (Volker , 2010)

3.3.2 Markov model

The conditional probabilities $P(y_t = S_j | y_{t-1} = S_i)$ are referred to as state transition probabilities or simply transition probabilities, where

$$a_{ij} = P(y_t = S_j | y_{t-1} = S_i)$$

We can assume that the transition probabilities are homogeneous, which means that the probabilities do not change over time, so

$$P(y_t = S_j | y_{t-1} = S_i) = P(y_{t+h} = S_j | y_{t-1+h} = S_i)$$

The transition probabilities can be written as a transition matrix, $A_{\mathcal{M} \times \mathcal{M}}$

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1\mathcal{M}} \\ a_{21} & a_{22} & \ldots & a_{2\mathcal{M}} \\ \vdots & \vdots & \ddots & \vdots \\ a_{\mathcal{M}1} & a_{\mathcal{M}2} & \ldots & a_{\mathcal{M}\mathcal{M}} \\ \end{pmatrix}$$

Each element in A is a probability of staying or jumping to another state, then,

(i) $a_{ij} \geq 0, \ \forall i, j$

(ii) $\sum_{j=1}^{M} a_{ij} = 1$ for $i = 1, \ldots, \mathcal{M}$

A full characterization of the Markov model is by including the initial state probability which is given as

$\pi_i(1) = P(x_1) = S_i;$ or $\pi(1) = [P(y_1 = S_1), P(y_1 = S_2) \ldots P(y_1 = S_M)]^T$.

This is the probability of being in one of the $\mathcal{M}$ states at the first state (Koubaa , 2008).

3.3.3 Homogeneous Markov chain

Time homogeneous Markov chain(s) with time homogeneous transition probabilities are processes where

$$P_{ij} = P(Y_{k+1} = j | Y_k = i) = P(Y_k = j | Y_{k-1} = i)$$

and P_{ij} is said to be stationary transition probability. The Y_k is a Markov chain of order one. Even though the one step transition is independent of k, this does not mean that the joint probability of Y_{k+1} and Y_k is also independent of k,

$$\begin{aligned} P(Y_{t+1} = j \text{ and } Y_t = i) &= P(Y_{t+1} = j | Y_t = i) P(Y_t = i) \\ &= P_{ij} P(Y_t = i) \end{aligned} \tag{3.7}$$

A time homogeneous Markov chain is stationary and there can be more than one stationary distribution for a given chain (as is the case with eigenvector of a matrix). The condition of stationarity for the chain depends on if the chain 'admits' a stationary distribution, and has positive probability only for positive-recurrent states (states that are re-visited). We note that the time homogeneous property of a Markov chain is distinct from the stationarity property. Stationarity implies time-homogeneity but on the other hand, a time homogeneous chain might not admit a stationary distribution and does not correspond to a stationary random process (Bilmes , 2006).

We note that if the process is stationary, then

$$P(Y_t = i, Y_{t-1} = j) = P(Y_{t-1} = i, Y_{t-2} = j)$$

$$\text{and } \; P(Y_t = i) = P(Y_{t-1} = i)$$

Therefore

$$a_{ij}(t) = \frac{P(Y_t = i, Y_{t-1} = j)}{P(Y_{t-1} = j)} = \frac{P(Y_{t-1} = i, Y_{t-2} = j)}{P(Y_{t-2} = j)} = a_{ij}(t-1)$$

So, by induction, $a_{ij}(t) = a_{ij}(t+\tau)$ for all τ, and the chain is time homogeneous (Bilmes , 2006). On the other hand, a time homogeneous Markov chain might not admit a stationary distribution and therefore never correspond to a stationary random process.

3.3.3.1 State probabilities

We are interested in the probability of finding the chain at various states, that is (Bilmes , 2006),

$$\pi_i(t) = P(Y_t = i)$$

By application of the total probability, we can write

$$\pi_i(t) = \sum_j P(Y_t = i|Y_{t-1} = j)P(Y_{t-1} = j) = \sum_j P_{ij}(t)\pi_j(t-1)$$

In vector form, for the homogeneous Markov chain, we write

$$\pi(t) = \pi(t-1)P$$

and for the non-homogeneous Markov Chain, we write

$$\pi(t) = \pi(t-1)P(k)$$

3.3.4 Classification of states

We consider a Markov chain Y_t with transition probability matrix A and a set of states S. A state j is said to be *accessible (or reachable)* from i to j if for some $h \geq 0$ the probability of going from i to j in h steps is positive, that is, $P_{ij}^{(h)}$. If $i \to j$ and $j \to i$, we say that i and j *communicate* and this is denoted as $i \leftrightarrow j$.

The equivalence relation $\leftrightarrow$ on the set of states indicates that the states communicate. The equivalence relation satisfies the following properties

(i) It is reflexive: $i \leftrightarrow j$ for all $i \in I$

(ii) It is symmetric: $i \leftrightarrow j$ if and only if $j \leftrightarrow i$

(iii) It is transitive: if $i \leftrightarrow j$ and $j \leftrightarrow k$ then $i \leftrightarrow k$

An equivalence relation on a set S decomposes the set into equivalence classes. If S is countable, then it can be partitioned into subsets (Ephraim and Merhav , 2002; Koubaa , 2008; Stirzaker , 2005);

(i) A stochastic matrix is called *irreducible* (ergodic) if the states consists of a single communicating class, that is, all states communicate with each other, $i \leftrightarrow j$.

(ii) A state i is said to be *transient* if after visiting the state, it is possible for it never to be visited again. We have two types of recurrent states, positive and null recurrent.

(iii) A state i is said to be *null-recurrent (less common)* if it is not transient but the expected return time is infinite. A state is *positive-recurrent* if it is not transient and the expected return time to that state is finite. For a Markov chain with a finite number of states, a state can only be either positive-recurrent or transient.

(iv) A state i is *periodic* with *period* $d > 1$, if d is the smallest number such that all paths leading from state i back to state i have a multiple of transitions. A state is *aperiodic* if it has period equal to one. A *path* is a sequence of states, where each transition has a positive probability of occurring .

3.3.5 Steady state

We recall that the state probability, the probability of finding the Markov chain at state i after the t^{th} step is given by

$$\pi_i \equiv P(Y_t = i)$$

We are interested in what happens in the long run, that is, $\pi_i = \lim_{t\to\infty} \pi_i(t)$. This is referred to as steady state or equilibrium or stationary state probability. The existence of a steady state implies that $\pi(t+1) \approx \pi(t)$. Therefore, the steady state probabilities are given the solution to the equations (Koubaa , 2008)

$$\pi = \pi P \quad \text{and} \quad \sum_i \pi_i = 1$$

The presence of periodic states in irreducible Markov chain prevents the existence of a steady state probabilities.

3.4 Reputation Computation

Trust is easy to recognize because we experience and rely on it everyday. It can manifest itself in many different forms, making it quite challenging to define. Reputation is what is generally said or believed about a person's or thing's character or standing. Reputation is a collective measure of trustworthiness from a social network based on the ratings from members in that community (Josang et al. , 2007). Therefore, reputation and trustworthiness are closely linked with trust being a complex social relationship Dubois et al. (2011).

We rely on trust everyday for every social transaction. There is a chance that one can defect against one's opponent so as to increase one's personal gain (Mui , 2002). There has been a rise in virtual communities such as online electronic markets. These markets use an online reputation rating system which attempts to provide a summary of a user's reputation history. With so much interactions and the content created online, the question of whom and what to trust has become an increasingly complex and important challenge (Dubois et al. , 2011). Reputation systems are expected to posses the following properties (Josang et al. , 2007)

(i) Entities must be long lived, so that with every interaction there is always an expectation of future interactions. This means that it should be impossible or difficult for an agent to change identity and re-enter as a new agent.

(ii) Ratings about current interactions are captured and distributed. Agents should be willing to provide ratings.

(iii) Ratings about past interactions must guide decisions about current interactions. Agents must respond to reputation system ratings and this is reflected in their activities and interactions

Trust and trustworthiness are positively correlated across societies (Knack and Keefer , 1997). Reputation and trust rating systems have a wide applications with many different types of mechanisms. No single solution is known to exist that is suitable in all contexts and applications. A reputation system design depends on constraints available and type of information that can be used as input ratings. Basic criteria for judging the quality and soundness of reputation computation engines (Josang et al. , 2007)

(i) Accuracy for long-term performance: Must have the capability to distinguish between unknown quality and poor long term performance

(ii) Weighting toward current behaviour: Tdhis is to detect sudden changes in the agent trustworthiness

(iii) Robustness against attacks: Resist attempts by the agents to manipulate the reputation ratings scores.

(iv) Smoothness: A new single score added should not influence the score significantly

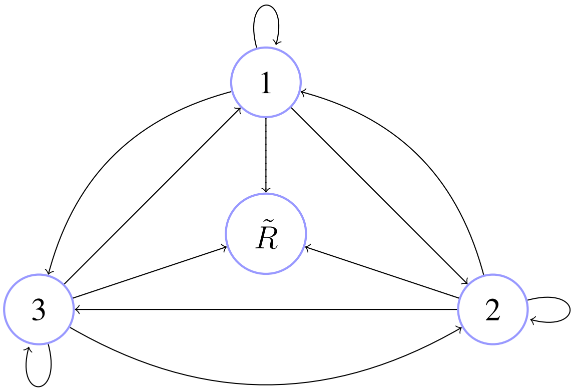

Figure 3.1: Three agents in a network

A reputation computation system showing a connection of three agents and the reputation engine $(\tilde{R})$ used to estimate their reputation ratings.

Reputation systems are typically based on agents information on each other resulting from personal experiences. A number of reputation computational engines exists and we have summarized a five of them according to (Josang et al. , 2007);

(i) Simple summation or average of ratings: The method applies the summation principle with the sum of positive ratings done separately with those of the negative ratings. A principle applied by the eBay's website reputation forum.

(ii) Bayesian system: They take positive and negative ratings as input and by use of statistical updating of beta probability density functions, they output reputation scores.

(iii) Discrete trust models: They use discrete measures ; an agent can be referred to as having high trust level, medium trust level, fair trust level or poor trust level. The limitation is that they are not easy in computations algorithms.

(iv) Belief models: Based on probability theory but with the assumption that the sum of probabilities over all possible outcomes is not equal to 1. The remaining probability is the uncertainty.

(v) Flow models: Reputation is computed by transitive iteration through loops or long chains. Some of these models assume a constant reputation weight for the whole community; while others do not always require the sum of the trust scores to be constant.

We have used the flow model in this research work to compute the reputation ratings of the agents in the network and thereafter use SVD to estimate the trust levels of the agents in the network.

3.5 Monte Carlo Simulation

Simulation is a numerical technique applied in conducting experiments by imitating a real life situation using logical and mathematics models. This estimates the likelihood of different possible outcomes probable over a given period of time. Simulation can be used in a number of situations with a lot of success (Rubinstein , 1981):

(i) If the process to obtain data is impossible or extremely expensive. The simulated data can then be used to formulate and test an hypothesis about the system under study.

(ii) When it is impossible or very costly to validate the mathematical model that describes the system

(iii) If the system cannot be modeled with a tractable numerical model to offer a analytical solution, then simulation can cover that gap.

Monte Carlo simulation (MCS) is a simulation technique that uses repeated random sampling and statistical analysis to compute the desired results. In most cases, an analytical solution is not known (Terejanu , 2002). A mathematical model is used where input parameters are input in the model, then processed through the mathematical formulas and the result is one or more outputs. MCS is a methodical way of doing the so called what-if-analysis (Samik , 2008).

Normally, we identify a statistical distribution and draw random samples from the distribution, which represents the values of the input variables. Each simulation run has output parameter. The output values are collected to perform statistical analysis on the values of the output parameters to make decisions. A statistical distribution or probability distribution describe the outcomes of varying a random variable and the probability of occurrence of those outcomes. Simulation is only as good as the estimates you make as it represents probabilities or uncertainties and not certainty (Samik , 2008). MCS has varied application areas from finance, real options analysis, portfolio analysis, to personal financial planning among other areas (Samik , 2008).

Input Data
Math Model
Output Analysis

Figure 3.2: A diagram of the simulation process in mathematical models

The diagram shows the process used in mathematical models, where input parameters are processed through mathematical formulas in the model and it results in one or more outputs.

Simulation efficiency can be increased by: first, developing good simulation algorithm; second, minimizing the storage requirement; third, minimizing the execution time; fourth, decreasing the variability of the simulation output, and the techniques used to reduce variability are called variance reduction techniques (Haugh , 2010). A complete simulation process has a large number of outputs derived from the input values. The output from the model describe the probability of achieving the results based on the input values in the model (Haugh , 2010). The standard simulation algorithm is;

(i) Generate $\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_m$

(ii) Estimate $\mathcal{O}$ with $\hat{\mathcal{O}}_m = \frac{1}{m}\Sigma_{i=1}^{m}\mathcal{W}_i,$ where $\mathcal{W}_i = h(W_i)$

(iii) Approximate $100(1-\alpha)\%$ for the confidence interval

$$[\hat{\mathcal{O}}_m - Z_{1-\alpha/2}\frac{\hat{\sigma}_m}{\sqrt{m}}, \quad \hat{\mathcal{O}}_m + Z_{1-\alpha/2}\frac{\hat{\sigma}_m}{\sqrt{m}}]$$

where $\hat{\sigma}_m$ is the usual estimate of var$(\mathcal{W})$ based on $\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_m$

A number of steps are normally performed for the MCS as outlined by (Samik , 2008);

(i) Static model generation - a deterministic model is used that closely resembles the real scenario where mathematical relationships are used with the input parameters and transformed with the desired output.

(ii) Input distribution identification - If the deterministic model working as expected, we add the risk components of the model which is from the stochastic nature of the input variables, governed by the underlying distribution of the variables

(iii) Random variable generation -We generate a set of random numbers or random samples from a random number generator (RNG). A RNG is a computational device designed to generate a sequence of numbers that appear to be independent draws from a population. We repeat the process of generating more random numbers for each input distribution and collect the different sets of possible output values.

(iv) Analysis and decision making - We perform statistical analysis on the output values. The statistical confidence is key for the decisions after running the simulations.

To increase the accuracy in MCS, one way is by increasing the number of samples but the convergence is very slow. A better way is to decrease the variance in the samples simulated. This is achieved through the variance reduction techniques (Boyle , 1977; Terejanu , 2002).

3.5.1 Variance reduction techniques

The variance reduction technique refines and improves the efficiency of the simulation. A number of methods are available; stratified sampling, importance sampling, control variates, antithetic variates, acceptance rejection sampling, partial sampling, among others (Terejanu , 2002). We briefly mention three of these techniques.

3.5.1.1 Antithetic Variable

In this technique, a simulation trial involves generating two random numbers from the specific probability distribution. The first value f_1 is generated in the usual way. The second number f_2 is generated by changing the sign of the number generated. So, if f_1 and f_2 are drawn from $p(y)$, then, the estimate of the sample drawn $\hat{f}$ is (Boyle , 1977);

$$\hat{f} = \frac{f_1 + f_2}{2} \tag{3.8}$$

and the variance of the two samples drawn to estimate one sample is;

$$\text{Var}(\hat{f}) = \text{Var}[\frac{1}{2}(f_1 + f_2)] = \frac{1}{4}\text{Var}[f_1] + \frac{1}{4}\text{Var}[f_2] + \frac{1}{2}\text{Cov}[f_1, f_2] \tag{3.9}$$

The above variance analysis indicates that

$$\text{Cov}[f_1, f_2] \approx \begin{cases} = 0, & \text{If estimate remains the same due to samples independence} \\ < 0 & \text{If the estimate is improved} \\ > 0 & \text{If the actual performance is worse} \end{cases} \tag{3.10}$$

The confidence interval is computed by estimating the standard error using the sample standard deviation of the samples from $\bar{f}$. Thus, the antithetic variate exploits the existence of the negative correlation between two estimates.

3.5.1.2 Acceptance rejection sampling

In this technique, we have an upper bound for the underlying probability distribution function $p(y)$ and we use a proposal distribution $q(y)$ also called the importance density. Then, there is $c < \infty$ such that $p(x) < cq(x)$. The algorithm for the rejection sampling is (Terejanu , 2002)

(a) Select a uniform random variable $u \sim U(0, 1)$

(b) Draw a sample $y \sim q(y)$

(c) if $u < \frac{p(y)}{cq(y)}$ then

(d) Accept y

(e) else

(f) reject it and repeat the process

(g) end if

The choice of c has a profound effect on the results as a too small c has a low rejection rate and a too big c has low acceptance rate.

3.5.1.3 Control variates

In this technique, the evaluation of an unknown expectation is replaced with the evaluation of the difference between the unknown quantity and a related quantity, whose expectation is known (Boyle , 1977). We carry out two simulations and let f_1 and f_2 be the respective values of the two random numbers. Then, we can write $f_1 = \mathrm{E}[f_1^*]$ and $f_2 = \mathrm{E}[f_2^*]$, where f_1^* and f_2^* are estimate values of the two random numbers respectively. A random variate f_2 is a control variate for f_1 if it is correlated with f_1. Then,

$$\hat{f}_1 = f_1^* + (f_2 - f_2^*) \tag{3.11}$$

where f_2 is the known value of the second random number and the known error $(f_2 - f_2^*)$ is used as a control in the estimation of f_1. The value $\hat{f}_1$ adjusts the estimator f_1 according to the difference between the known value f_2 and the observed value f_2^*. We reduce the variance by comparing the values of the two random numbers, with (Terejanu , 2002);

$$\mathrm{Var}[\hat{f}_1] = \mathrm{Var}[f_1^*] + \mathrm{Var}[f_2] + \mathrm{Var}[f_2^*] - 2\mathrm{Cov}[F_1^*, f_2^*] \tag{3.12}$$

and $\mathrm{Var}[f_2] = 0$ since f_2 is the known value of the second random number and thus not a random variable. This control variate technique is effective if the covariance between f_1^* and f_2^* is large, that is, if $2\mathrm{Cov}[f_1^*, f_2^*] > \mathrm{Var}[f_1^*] + \mathrm{Var}[f_2^*]$. The variance reduction is achieved.

We have observed that MCS is a very useful mathematical technique for analyzing uncertain scenarios and providing probabilistic analysis of different situations and applications. In real-life situations, it is not practical to investigate many factors but in simulation experiments, we can have hundreds of factors (Kleijnen , 2009).

3.5.2 Conclusion

The mathematical preliminaries introduced in this chapter are captured in different parts of the methodology and data analysis that are found in the following chapters. SVD technique is ideal for the estimation of the variables from the matrices generated in the SEN. The HMM are based on the Markov process, and has a stochastic component to capture the agents dynamics. Stopping time is used to estimate defaults or non defaults in the loan portfolio of the obligors who are agents in the SEN.

Chapter 4

Multiple Agents HMM

This chapter is part of the methodology in which we modify the standard HMM to multiple agents HMM. The agents in the SEN are heterogeneous and one key requirement is to have a HMM with the capabilities to model the dynamics of each agent individually and as a group. This brings out the individuality and group dynamics in the SEN interactions and cyclical inter dependencies. Therefore, the modifications in this chapter forms part of the contributions brought forward in this study.

4.1 Introduction

The standard hidden Markov model specifications are introduced by outlining the five parameters of HMM. A modification of the standard HMM is undertaken by introducing multiple variables in the parameters to cater for the expected observations in the multiple agent SEN. The modifications in this chapter forms a part of the multiple agents HMM and one of the major contributions made in this study.

4.2 Standard HMM

The work of Lawrence Rabiner in the year 1989 opened up a whole new frontier in the field of HMM in his seminal paper (Rabiner , 1989). A hidden Markov model is a double embedded stochastic process with two hierarchy levels in which the system being modeled is assumed to be a Markov process with unobserved state. In a regular Markov model, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In HMM, the state is not directly visible 'hidden', but output, dependent on the state, is visible. Each state has a probability distribution over the possible emissions. This sequence of emissions gives some information about the hidden states (Mhamanne and Lobo , 2012).

Hidden Markov models (HMMs) are capable statistically to characterize and estimate the signal in a precise and well defined manner (Rabiner , 1989). These models are inexpensive, intuitive and versatile for modeling stochastic processes, to estimate and track activities based on noisy information. States involved are finite and state space is known but the current state of the process is not known with certainty and has to be estimated from whatever evidence is available.

HMM makes no marginal independence assumptions, that is, $\mathbf{Y} \perp Y$ but the only assumptions of conditional independence exist in an HMM of the form $\mathbf{Y} \perp Y | \mathbb{Y}$. Conditional independence has a power to make a statistical model undergo enormous simplifications - this implies that some factorization of the joint distribution exists (Bilmes , 2006).

4.2.1 Specifications of standard HMM

To fully specify a hidden Markov model, we have five parameters (Rabiner , 1989)

(a) M is the number of states in the model. We denote the set of all possible states as

$S = \{S_1, S_2, \ldots, S_M\}$

(b) K, the number of distinct observation symbols per state, that is, the discrete alphabet size of the output set. We denote the set of all possible output symbols as

$V = \{v_1, v_2, \ldots, v_K\}$, the output symbol at time t as O_t. The sequence of observed symbols is denoted as $O = O_1 O_2 \ldots O_T$.

(c) The state transition probability distribution

$A = \{a_{ij}\}$, where $a_{ij} = P[q_{t+1} = S_j | q_t = S_i],\ 1 \leq i, j \leq M$

(d) The observation symbol probability distribution in state j,

$B = \{b_j(k)\}$, where $b_j(k) = P[O_t = v_k | q_t = S_j],\ \ 1 \leq j \leq M,\ 1 \leq k \leq K.$

(e) The initial state distributions

$\pi = \{\pi_i\}$, where $\pi_i = P[q_1 = S_i], \;\; 1 \leq i \leq M$

4.2.2 Multiple agents HMM parameters modification

The modification of the standard hidden Markov model to cater for the multiple agents in the SEN is undertaken in this section. This modification is important as part of the contributions of the study and to offer a HMM for the multiple agents in the network as the model emits the credit quality scores and levels of the agents under study.

1. **The set of hidden states**

   The set of hidden states are given by;

   $$S = (S_1, S_2, \ldots, S_M) \tag{4.1}$$

   Where M is the number of state transitions. If an agent n is connected to all the agents at time period t, then,

   $$S_t^{(n)} = S_t^{ij}, \;\; 1 \leq i, \; j \leq M, \;\; t = 1:T$$

2. **State transition probability distribution**

   Let the transition matrix be given by

   $$A_t = \{a_{ij}\}_{M \times M}, \;\; 1 \leq i, \; j \leq M$$

   $$a_{i,j} = P(q_{t+1} = j | q_t = i) \tag{4.2}$$

   Where $a_{i,j}$ is the transition probability from hidden state i to hidden state j, while q_t denotes the hidden state at time t. The sequence of hidden state in Markov model is $Q = (q_1, \ldots, q_T)$

   The set of all the hidden states for the N agents at time t is given in the matrix

   $$Q = \begin{pmatrix} q_1^1 & q_2^1 & \cdots & q_T^1 \\ q_1^2 & q_2^2 & \cdots & q_T^2 \\ \vdots & \vdots & \ddots & \vdots \\ q_1^N & q_2^N & \cdots & q_T^N \\ & & & \end{pmatrix}$$

Where q_t^n is the hidden state of agent n at time $t,\ 1 \leq n \leq N,\ t \in [1, T]$. The structure of this stochastic matrix defines the connection structure of the model, and the transition probabilities should satisfy the normal stochastic constraints with

$$a_{ij} \geq 0,\ \ 1 \leq i,\ j \leq M,\ \ \sum a_{ij} = 1,$$

3. **Set of observation symbols**

The observation of all the N agents at each time t is a vector

$$O_t = (O_t^{(1)}, O_t^{(2)}, \ldots, O_t^{(N)}) \tag{4.3}$$

The sequence of T observations made on agent n is

$$O^{(n)} = (O_1^{(n)}, O_2^{(n)}, \ldots, O_T^{(n)})$$

The set of all the observations for the N agents at time t is given in the matrix

$$O = \begin{pmatrix} O_1^1 & O_1^1 & \ldots & O_T^1 \\ O_1^2 & O_2^2 & \ldots & O_T^2 \\ \vdots & \vdots & \ddots & \vdots \\ O_1^N & O_2^N & \ldots & O_T^N \\ \\ \end{pmatrix}$$

4. **Observation symbol probability distribution**

Let the observation or emission matrix be given as

$$B = \{b_j(\nu_\kappa)\},\ \ 1 \leq i,\ j \leq M,\ \ 1 \leq \kappa \leq K$$

$$b_j(\kappa) = P[O_t = \nu_\kappa | q_t = j],\ 1 \leq j \leq M,\ \ 1 \leq \kappa \leq K \tag{4.4}$$

Matrix B is also known as the emission matrix, $b_j(\nu_\kappa)$ is the probability that symbol ν_κ is emitted in state j, ν_κ denotes the κ^{th} observation symbol and O_t is the current parameter vector. The following stochastic constraints must be satisfied

$$\begin{gathered} b_j(\kappa) \geq 0,\ \ 1 \leq j \leq M,\ \ 1 \leq \kappa \leq K, \qquad (4.5) \\ \sum b_j(\kappa) = 1,\ \ 1 \leq j \leq M \end{gathered}$$

There are K set of possible observations of the agents

$$V = \{\nu_1, \nu_2, \ldots, \nu_K\},\ 1 \leq \kappa \leq K \tag{4.6}$$

We have K observation symbols per state and each agent net value changes at each observation time due to their dynamics and connections in the network.

5. **Initial state probability distribution**

   The initial state distribution $\pi = \{\pi_i^{(n)}\},$ where π_i is the probability that the model is in state i at the time $t = 0$ with

$$\pi_i^{(n)} = P(q_{n1} = i), \tag{4.7}$$
$$\sum_{i=1}^{M} \pi_i^{(n)} = 1,\ \ 1 \leq i \leq M$$

   We denote an HMM as a triplet $\lambda = (A, B, \pi)$, the model training parameters

4.3 Basic Problems for HMM

The three main problems in HMM are probability calculation, state estimation and parameter estimation, which were solved by Lawrence Rabiner (1989).

4.3.1 Evaluation problem

In this problem, we compute the probability that a given model generates a given sequence of observations, $O = (O^{(1)}, \ldots, O^{(N)})$. That is, given the model $\lambda = (A, B, \pi),$ compute $P(O|\lambda)$. The most used algorithms to solve the problem are (Rabiner , 1989)

(a) **Forward algorithm**: find the probability of emission distribution (given a model) starting from the beginning of the sequence

(b) **Backward algorithm**: find the probability of emission distribution (given a model) starting from the end of the sequence

A direct calculation is by enumerating every possible state sequence of length T (observation times). Consider a fixed state sequence

$$Q_n = q_{nt},\ \ 1 \leq n \leq N,\ \ 1 \leq t \leq T \tag{4.8}$$

Likelihood of an observation sequence given a state sequence, or likelihood of an observation sequence along a single path: the probability of the observation sequence $O_t^{(n)}$ for the state sequence Q_t^n of the same length, determined from a HMM with parameters λ, the likelihood of O along the path Q for the n^{th} agent is equal to:

$$P(O^{(n)}|Q^{(n)},\lambda) = \prod_{t=1}^{T} P(O^{(n)}|q_t^{(n)},\lambda) \tag{4.9}$$

Joint likelihood of an observation sequence O and a path Q : it is the probability that O and Q occur simultaneously, $P(O,Q|\lambda)$, and decomposes into a product of the two quantities. The joint probability of O and Q is derived from Bayes given as:

$$P(O^{(n)},Q^{(n)}|\lambda) = P(O^{(n)}|Q^{(n)},\lambda)\ P(Q^{(n)}|\lambda) \tag{4.10}$$

The probability of a state sequence $Q = q_1, q_2, \ldots, q_n$ coming from a HMM with parameters λ corresponds to the product of the transition probabilities from one state to the next, and they are expressed as

$$P(Q^{(n)}|\lambda) = \pi_{q_1^{(n)}} a_{q_1^{(n)} q_2^{(n)}} a_{q_2^{(n)} q_3^{(n)}} \ldots a_{q_{T-1}^{(n)} q_T^{(n)}} \tag{4.11}$$

while the observed probabilities given the model is

$$P(O^{(n)}|Q^{(n)},\lambda) = \prod_{t=1}^{T} a_{q_t^n, q_{t+1}^n} = b_{q_1^{(n)}}(O_1^{(n)}) b_{q_2^{(n)}}(O_2^{(n)}) \ldots b_{q_T^{(n)}}(O_T^{(n)}) \tag{4.12}$$

The likelihood of an observation sequence O with respect to a HMM with parameter λ given the model is obtained by summing the joint probabilities over all possible state sequences

$$\begin{aligned} P(O^{(n)}|\lambda) &= \sum_{\text{all } Q} P(O^{(n)}|Q^{(n)},\ \lambda)\ P(Q^{(n)}|\lambda) \\ &= \sum_{q_1,\ldots,q_T} \pi_{q_1^{(n)}} b_{q_1^{(n)}}(O_1^{(n)}) a_{q_1^{(n)} q_2^{(n)}} b_{q_2^{(n)}}(O_2^{(n)}) \ldots \\ &\quad \ldots a_{q_{T-1}^{(n)} q_T^{(n)}} b_{q_T^{(n)}}(O_T^{(n)}) \end{aligned} \tag{4.13}$$

At $t = 1$, the n^{th} agent is in state $q_1^{(n)}$ with probability $\pi_{q_1^{(n)}}$ and generate the symbol $O_1^{(n)}$ with probability $b_{q_1^{(n)}}(O_1^{(n)})$. We continue in this manner until $t = T$ from state $q_{T-1}^{(n)}$ to state $q_T^{(n)}$ with probability $a_{q_{T-1}^{(n)} q_T^{(n)}}$ and generate symbol $O_T^{(n)}$ with probability $b_{q_T^{(n)}}(O_T^{(n)})$. The equation (4.14) involves $2^T N^T$ calculations and a more efficient procedure is required. The Forward-Backward procedure is used.

4.3.1.1 Forward variable

The forward variable $\alpha_t^{(n)}(i)$ is defined as

$$\alpha_t^{(n)}(i) = P(O_1^{(n)}, \ldots, O_t^{(n)}, q_t^{(n)} = i|\lambda) \tag{4.14}$$

That is, the probability of the partial observation sequence until time t and state i at time t given the model λ. $\alpha_t^{(n)}(i)$ can be obtained inductively, see Appendix A.1.1

4.3.1.2 Backward variable

We consider a backward variable $\beta_t^{(n)}(i)$ defined as

$$\beta_t^{(n)}(i) = P(O_{t+1}^{(n)}, \ldots, O_T^{(n)}|q_t^{(n)} = i|\lambda), \;\; 1 \leq t \leq T \tag{4.15}$$

That is, the probability of the partial observation sequence from $t+1$ to the end, given state i at time t and the model λ. We can solve $\beta_t^{(n)}(i)$ inductively, see appendix A.1.2.

4.3.1.3 Forward-backward variable

The Forward-Backward procedure is based on the technique known as dynamic programming (Lyengar , 2005). To apply the procedure, we find a recursive property that allows us to do calculations for the next instance based on the current one. We can see that

$$\alpha_t^{(n)}(i)\beta_t^{(n)}(i) = P(O_t^{(n)}, q_t^{(n)} = i|\lambda), \tag{4.16}$$
$$1 \leq i, \; j \leq M, \; 1 \leq n \leq N$$

The evaluation problem can be solved by both forward and backward algorithm

$$P(O^{(n)}|\lambda) = \sum_{i=1}^{M} P(O^{(n)}, q_t^{(n)} = i|\lambda) \tag{4.17}$$
$$= \sum_{i=1}^{M} \alpha_t^{(n)}(i)\beta_t^{(n)}(i) = \sum_{i=1}^{M} \alpha_T^{(n)}(i)$$

4.3.2 Decoding problem

Given a model, λ, and a sequence of observations, i, induce the most likely hidden states (that best explains the observations) more specifically (Rabiner , 1989)

(i) Find the sequence of internal states that has the highest probability. We use the Viterbi algorithm

(ii) Find for each position the internal state that has the highest probability. Mostly used algorithm is the posterior decoding algorithm

This involves finding the ‘optimal’ state sequence associated with the given observation sequence. This optimality criterion maximizes the expected number of correct individual states. The Viterbi algorithm is used to find the single best state sequence and is based on dynamic programming method.

4.3.2.1 Viterbi algorithm

This technique is similar to forward algorithm and traces the most likely hidden states while reproducing the output sequence. It differs from forward algorithm in that the transition probabilities are maximized at each step, instead of summation. Define an auxillary variable (Galassi , 2008)

$$\delta_t^{(n)}(i) = \max_{q_1,\cdots,q_{T-1}} P(q_1^{(n)},\ldots,q_t^{(n)},\ O_1^{(n)},\ldots,O_{T-1}^{(n)}|\lambda) \tag{4.18}$$

That is, $\delta_t^{(n)}(i)$ is the best score (highest probability) along a single path, at time t, which accounts for the first t observations and ends in state i. By induction

$$\begin{gathered}\delta_{t+1}^{(n)}(j) = b_j(O_{t+1}^{(n)})\left[\max_{1\le i\le M} \delta_t^{(n)}(i)a_{ij}\right] \\ \text{with }\ \delta_1^{(n)}(j) = \pi_j^{(n)} b_j(O_1^{(n)}),\ 1\le j\le M\end{gathered} \tag{4.19}$$

To retrieve the state sequence, we also need to keep track of the state that maximizes $\delta_t^{(n)}(i)$ at each time, t. This is done by constructing an array

$$\psi_{t+1}^{(n)}(j) = \text{argmax}_{1\le i\le M}\left[\delta_t^{(n)}(i)a_{ij}\right] \tag{4.20}$$

$\psi_{t+1}^{(n)}(j)$ is the state at time t from which a transition to state j maximizes the probability $\delta_{t+1}^{(n)}(j)$. The Viterbi algorithm for finding the optimal state sequence is given in the appendix A.1.3.

4.3.3 Learning problem

Given the observation sequence O and the model parameters A, B and π, how do we find the model that best explains the observed data. According to (Rabiner , 1989);

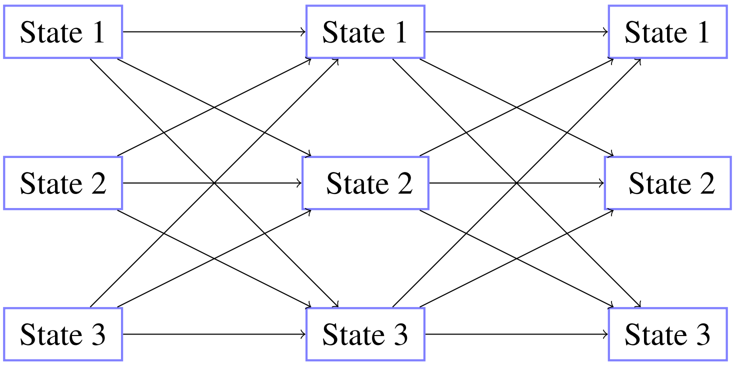

Figure 4.1: Trellis diagram for a three states HMM

The diagram shows the three possible states the model can take at any given time period interval. Only one state is reached at each time period.

(i) Find the optimal model based on the most probable sequences. Most used algorithm is the Viterbi training (that uses recursively the Viterbi algorithm)

(ii) Find the optimal model based on the sequences of the most probable internal states. Most used algorithm is the Baum Welch algorithm (that uses recursively the posterior decoding algorithm)

How can we adjust the HMM parameters in way that a given set of observations (the training set) is represented by the model in the best way for the intended application? We choose $\lambda = (A, B, \pi)$ such that $P(O^{(n)}|\lambda)$ is locally maximized. The iterative procedure used is the Baum-Welch method.

4.3.3.1 Baum-Welch method

It is also equivalently called the expectation modification or maximization (EM) method. This method can be derived by maximization of the auxiliary quantity (Rabiner , 1989)

$$Q(\lambda, \tilde{\lambda}) = \sum_q P\left(q_t^{(n)}|O^{(n)}, \lambda\right) \log\left(P[O^{(n)}, q_t^{(n)}, \tilde{\lambda}]\right)$$

over $\tilde{\lambda}$. We define two more auxiliary variables (Dymarski , 2011)

(i) The first one is the *joint event*

$$\xi_t^{(n)}(i,j) = P(q_t^{(n)} = i, q_{t+1}^{(n)} = j|O^{(n)}, \lambda)$$

which can be expressed as

$$\xi_t^{(n)}(i,j) = \frac{P(q_t^{(n)} = i, q_{t+1}^{(n)} = j, O^{(n)}, \lambda)}{P(O^{(n)}|\lambda)} \tag{4.21}$$

We use forward and backward variables in (4.21) to get

$$\xi_t^{(n)}(i,j) = \frac{\alpha_t^{(n)}(i) a_{ij} \beta_{t+1}^{(n)}(j) b_j(O_{t+1}^{(n)})}{\sum_{i=1}^{M} \sum_{j=1}^{M} \alpha_t^{(n)}(i) a_{ij} \beta_{t+1}^{(n)}(j) b_j(O_{t+1}^{(n)})} \tag{4.22}$$

This is the probability of being in state i at time t and state j at time $t+1$, given the model and the observation sequence.

(ii) The second variable is the a posterior probability - state variable

$$\gamma_t^{(n)}(i) = P(q_t^{(n)} = i|O^{(n)},\ \lambda) \tag{4.23}$$

That is, the probability of being in state i at time t given the observation sequence $O^{(n)}$ and the model λ. In forward and backward variables, we can express it as

$$\gamma_t^{(n)}(i) = \frac{\alpha_t^{(n)}(i)\beta_t^{(n)}(i)}{\sum_{i=1}^{M} \alpha_t^{(n)}(i)\beta_t^{(n)}(i)} \tag{4.24}$$

$\alpha_t^{(n)}(i)$ accounts for the partial observation sequence $O_1^{(n)}, \ldots, O_t^{(n)}$ and state i at time t. $\beta_t^{(n)}(i)$ accounts for the remainder of the observation sequence $O_{t+1}^{(n)}, \ldots, O_T^{(n)}$ given the state i at time t.

The relationship between $\gamma_t^{(n)}(i)$ and $\xi_t^{(n)}(i,j)$ is given by

$$\gamma_t^{(n)}(i) = \sum \xi_t^{(n)}(i,j), \quad 1 \leq i, \quad j \leq M, \quad 1 \leq t \leq T \tag{4.25}$$

The normalization factor

$$P(O_n|\lambda) = \sum_{i=1}^{M} \alpha_t^{(n)}(i)\beta_t^{(n)}(i)$$

makes $\gamma_t^{(n)}(i)$ a probability measure so that

$$\sum_{t=1}^{T} \gamma_t^{(n)}(i) = 1$$

Therefore

$$\sum_{t=1}^{T-1} \gamma_t^{(n)}(i) = \text{Expected number of transitions from } i$$

$$\sum_{t=1}^{T-1} \xi_t^{(n)}(i,j) = \text{Expected number of transitions from } i \text{ to } j$$

4.3.3.2 Baum-Welch Learning Process

This is the parameter updating equations. We want to maximize the quantity $P(O^{(n)}|\lambda)$. Assume a starting model $\lambda = (A, B, \pi)$ and calculate the α and β values. Then calculate the γ and ξ values. We now have equations known as *re-estimation formulas* and are used to update the HMM parameters.

(i) Initial state probability

$$\tilde{\pi}_i^{(n)} = \gamma_1^{(n)}(i), \;\; 1 \leq i \leq M$$

(ii) State transition probabilities

$$\tilde{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t^{(n)}(i,j)}{\sum_{t=1}^{T-1} \gamma_t^{(n)}(i)}$$

(iii) Symbol emission probability

$$\tilde{b}_j(O_t^{(n)}) = \frac{\sum_{t=1}^{T-1} \gamma_t^{(n)}(j)}{\sum_{t=1}^{T-1} \gamma_t^{(n)}(j)}$$

This adjustment is reasonable as our model parameters are updated by calculating corresponding ratios of proportions. The set of new model parameters is $\tilde{\lambda} = (\tilde{A}, \tilde{B}, \tilde{\pi})$. We have the new parameters, which proves that, either, (1) the initial model λ defines a critical point of the likelihood function, in which case $\tilde{\lambda} = \lambda$, or, (2) model $\tilde{\lambda}$ is more likely than model λ in the sense that

$$P(O^{(n)}|\tilde{\lambda}) > P(O^{(n)}|\lambda)$$

The iterative use of $\tilde{\lambda}$ instead of λ improves the probability of $O^{(n)}$ being observed from the model until some limiting point is reached. The final result of this procedure is called a maximum likelihood estimate of the HMM. The solution to training of the problem is in appendix A.1.4

4.3.3.3 Baum's auxilliary function

Maximizing Baum's auxilliary function over $\tilde{\lambda}$

$$Q(\lambda, \tilde{\lambda}) = \sum_{Q} P(Q^{(n)}, O^{(n)}, \lambda) \log(P(O^{(n)}, Q^{(n)}|\tilde{\lambda})) \tag{4.26}$$

where $\tilde{\lambda}$ is the auxilliary variable that corresponds to λ. The maximization of $Q(\lambda, \tilde{\lambda})$ leads to increased likelihood, that is,

$$\max(Q(\lambda, \tilde{\lambda}))$$

which implies

$$P(O^{(n)}|\tilde{\lambda}) \geq P(O^{(n)}|\lambda)$$

the likelihood function converges to a critical point.

4.4 HMM Topologies

A HMM is classified into one of the following types depending on its state transition

4.4.1 Ergodic model

Ergodic model is also known as fully connected or full state transition HMM. Every state of the model could be reached in a single step from every other state of the model (Rabiner , 1989).

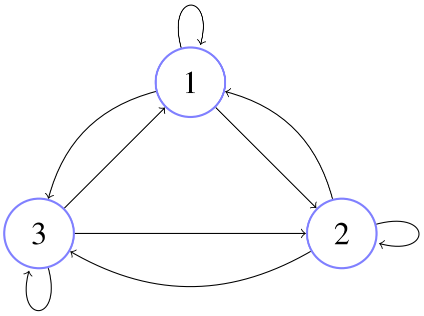

Figure 4.2: A three states Ergodic Model

The diagram depicts a three state transition model. Each of the three state is connected to the other states in the model.

This is the type of HMM topology that is utilized in this study to understand agents dynamics in a SEN and how they affect them in their loan obligations with a financial institution. A ergodic transition matrix A with M transition states is expressed as

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1M} \\ a_{21} & a_{22} & \dots & a_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \dots & a_{MM} \\ \end{pmatrix}$$

4.4.2 Left-to-right model

The underlying state sequence associated with the model has the property that as time increases, the state index increases (or stays the same), that is, the states proceed from left to right. This model has the desirable property that it can model situations that change over time. They have the state transition coefficient that $a_{ij} = 0,$ for all $j < i$, that is, no transitions allowed to states whose indices are lower than the current state. .

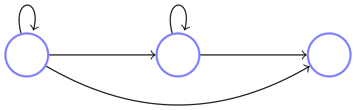

Figure 4.3: A three states Left to Right Model

The diagram shows that the model only moves from left to right without any possibility of transiting from right to left. That is why it is referred to as the left to right model, the only direction the transition processes can take.

This model is good for modeling order constrained time series whose properties change over time (Couvreur , 1996). The transition matrix A is upper triangular, that is

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1M} \\ 0 & a_{22} & \dots & a_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{MM} \\ \end{pmatrix}$$

and the initial distribution is the unit vector $\pi = (1, 0, \dots, 0)'$. The Markov chain evolves along the states in increasing order and the last state (M) will be reached infinite time with probability one (Couvreur , 1996).

4.4.3 Stationary model

If the initial state distribution π of an ergodic HMM is the unique stationary distribution π^*, then, $\pi^* = \pi^* A$

The assumption is valid since the state distribution of an ergodic Markov chain always converges toward the stationary distribution. $\lambda = (A, B, \pi^*)$ is redundant since π^* can be computed from A by solving $\pi^* = \pi^* A$. For stationary ergodic HMMs, $\lambda = (A, B)$. For non-ergodic HMM, the solution of $\pi^* \neq \pi^* A$ need not be unique. If π is a stationary distribution, the Markov Chain $\{Y_t\}$ is stationary (Couvreur , 1996).

4.4.4 HMM architectures

The easiest way to increase an HMM's accuracy is by increase the number of hidden states and the capacity of the observation distributions (Bilmes , 2006). As the number of states $\mathcal{M}$, increases, the computations associated with HMM grow quadratically $O(T\mathcal{M}^2)$, but there is an appreciable associated computational cost. Bilmes (2006) notes that HMM can accurately model any real world probability distribution, given enough hidden states and a sufficient rich class of observation distributions. HMM is a very powerful class of probabilistic model families with no limit to its ability to model any distribution.

The conventional HMM has deficiencies and many architectures have been proposed to handle the deficiencies. Some of the HMM architectures that exist are interactive (Ching et al. , 2006); factorial and poisson (Bilmes , 2006); auto-regressive (Rabiner , 1989); layered (Bilmes , 2006); linked (Mathew , 1997); hierarchical (Galassi , 2008; Ueda , 2004); input output (Bilmes , 2006); and coupled hidden Markov (Mathew , 1997; Zhong and Ghosh , 2001), among other architectures.

4.4.5 Conclusion

The modifications of the HMM parameters caters for the multiple agents in the SEN. Each agent has a set of CRAFs that are used in learning and training the HMM. Thus, each agent has a set of the five HMM parameters, namely $\lambda = \{A_n, B_n, \pi_n, O_n, Q_n\}$, which are applied in the HMM model to emit the CQS and CQL. The modifications are also part of the contributions made by this study in the arena of consumer credit using the HMM as the technique to classify the agents into different CQL and in emitting the CQS.

Chapter 5

SEN-HMM-CSD Model

This chapter introduces the mathematical techniques, simulation and analysis guidelines as it forms the methodology of this study. The proposed model has five components which are discussed in detail to expound on its applicability to consumer credit scoring.

5.1 Introduction

Humans tend to exhibit specific behavioristic profiles that are more individual based but with some similarities with each other. It is on this basis that HMM are ideal tools to model this behavior as the patterns in the SEN are dynamic and interdependent amongst the agents in the network. A real world process like human behaviour produces observable output which is characterized as a signal. This signal can be modelled to help us learn as much as possible by use of simulations and other techniques. Hidden Markov models have been found ideal to characterize the parametric random process of a signal. The stochastic process generated can be estimated in a precise and well defined manner (Rabiner , 1989).

In this study, HMM has the ability to dynamically segment agents or the bank customers in a loan portfolio into credit quality states, and estimate the evolution of customer relationship with the bank over time. As credit quality changes from time to time, with certain correlation structure, this induces individual unpredictable credit risk making individual analysis of risk dynamic process. This in turn affect the quality of a loan portfolio through default probabilities. Interaction effects of agents due to interdependency are an important component of portfolio credit risk. A simple description of the interaction process justified by economic and/ or empirical ground is needed. Our model can be used for the purposes of credit risk management for consumer loans. The model incorporates both the economic data (or private data) and social data (or social capital data) from the network.

The data, both social and economic for the agents is enhanced by the interdependency of the agents in the SEN. These sets of data from the SEN are used to estimate different components of credit risk analysis factors (CRAF) for each agent, and extract the observation and states of the HMM. In turn, the HMM classifies the agents into the respective credit scores or levels, estimation of the dynamic default threshold and probabilities of credit for each agent. These credit score estimates are undertaken at each time $t \in [0, T]$. Therefore, the model captures the every changing economic and social conditions as well as the dynamism observed in this consumer credit scoring model.

We observe that the N agents are in the SEN while the N obligors are part of the loan portfolio with the financial institution (N = agents = obligors). We therefore use the terms agents and obligors interchangeably in some of the sections of this study.

5.2 The Model

We propose a model that is referred to as Socio-Economic Network + Hidden Markov Model + Credit Scores and Default (SEN-HMM-CSD) model which has five levels outlined below;

1. Initial conditions of the agents
2. Social economic network dynamics
3. Extraction of the credit analysis risk factors
4. The Hidden Markov learning and training
5. Credit scoring and default rates

Mathematical models can be used to gain insights into many aspects of the world around us but there is no mathematical model that is perfect as these models only depict those characteristics of direct interest to the modeler (Stanley , 2006). The SEN-HMM-CSD model enables us to study the dynamics of the social network, the parameter estimation in the HMM and the credit quality dynamics as the output.

Figure 5.1 shows the flow of activities in the SEN-HMM-CSD model for each of the five levels. In figure 5.2, level 1 and level 2, the initial conditions and the SEN dynamics forms the first part of SEN-HMM-CSD model. For figure 5.3, it depicts the activities in levels $3,\ 4$ and 5 which forms the clustering of the CRAF to estimate matrices A and B for the HMM and CSD components of the SEN-HMM-CSD model. Figure 5.2 and figure 5.3 depicts the flow of activities for the five levels of the model.

We expound on each level in the next section of our work. Before then, the key assumptions are outlined to simplify the model.

5.2.1 SEN-HMM-CSD model assumptions

Assumptions are an important component of a model as it plays the role of bridging the 'real world' to the 'mathematical world'. The main assumptions in this study are:

(a) No population drift during the life of the loan

(b) Default rates are based on the CQS of the agent and the default threshold

(c) No interest rates and no early repayment

(d) The obligor loans are from the same financial institution

(e) Obligors have similar loan repayment duration and amount

(f) The network is fully connected

(g) Agents affected by social and economic factors

(h) Agents reputation ratings estimates trust and distrust

The assumptions in the model increases the tractability when complexity is involved but this leads to model limitations which are listed in the next section.

5.2.2 SEN-HMM-CSD model limitations

We highlight some of the limitations as a result of the model assumptions and in reducing its complexity

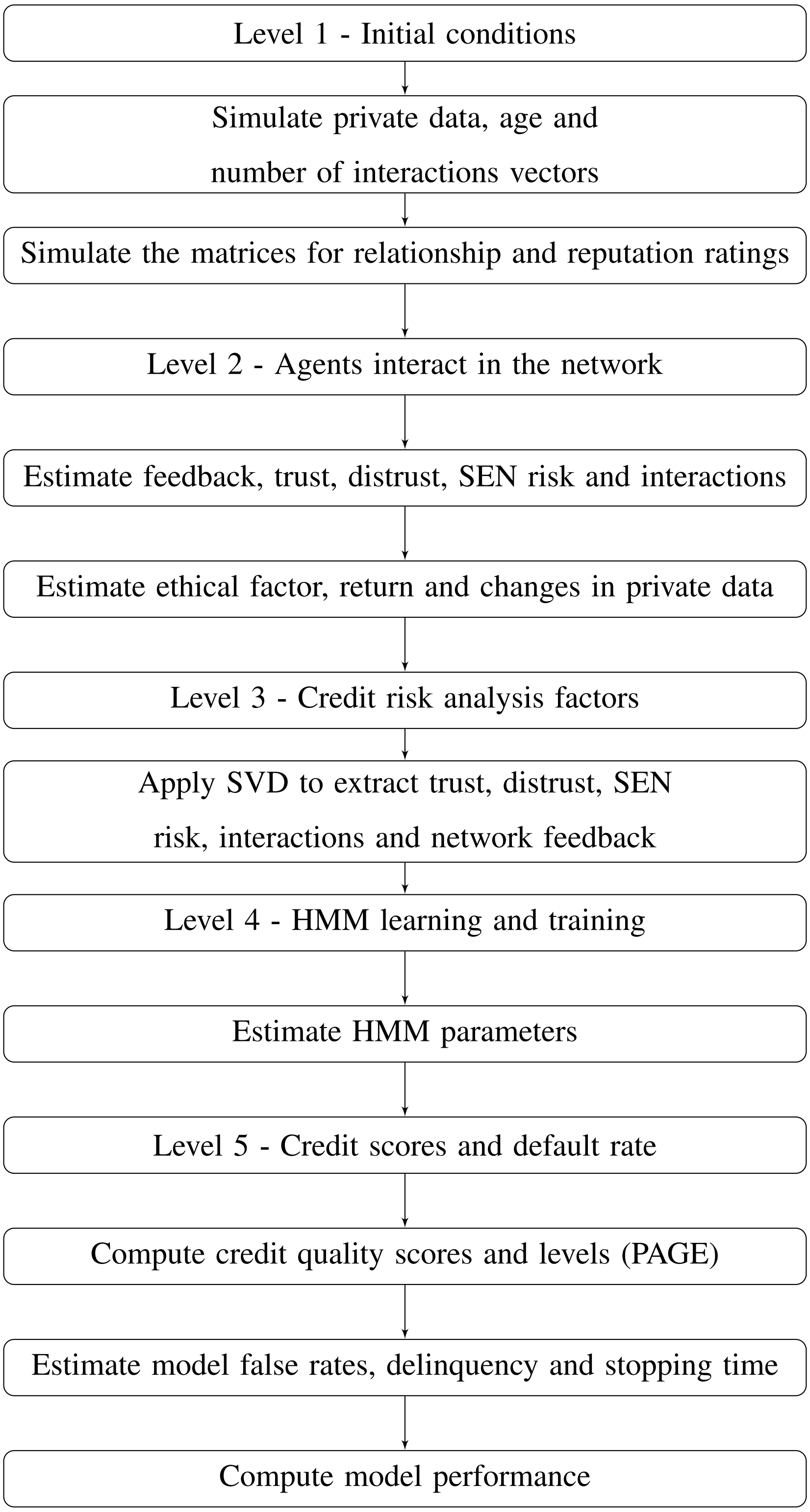
Level 1 - Initial conditions
Simulate private data, age and number of interactions vectors
Simulate the matrices for relationship and reputation ratings
Level 2 - Agents interact in the network
Estimate feedback, trust, distrust, SEN risk and interactions
Estimate ethical factor, return and changes in private data
Level 3 - Credit risk analysis factors
Apply SVD to extract trust, distrust, SEN risk, interactions and network feedback
Level 4 - HMM learning and training
Estimate HMM parameters
Level 5 - Credit scores and default rate
Compute credit quality scores and levels (PAGE)
Estimate model false rates, delinquency and stopping time
Compute model performance

Figure 5.1: Flow of activities in the SEN-HMM-CSD model

The flow sequence of activities in the model, from first level to the fifth level.

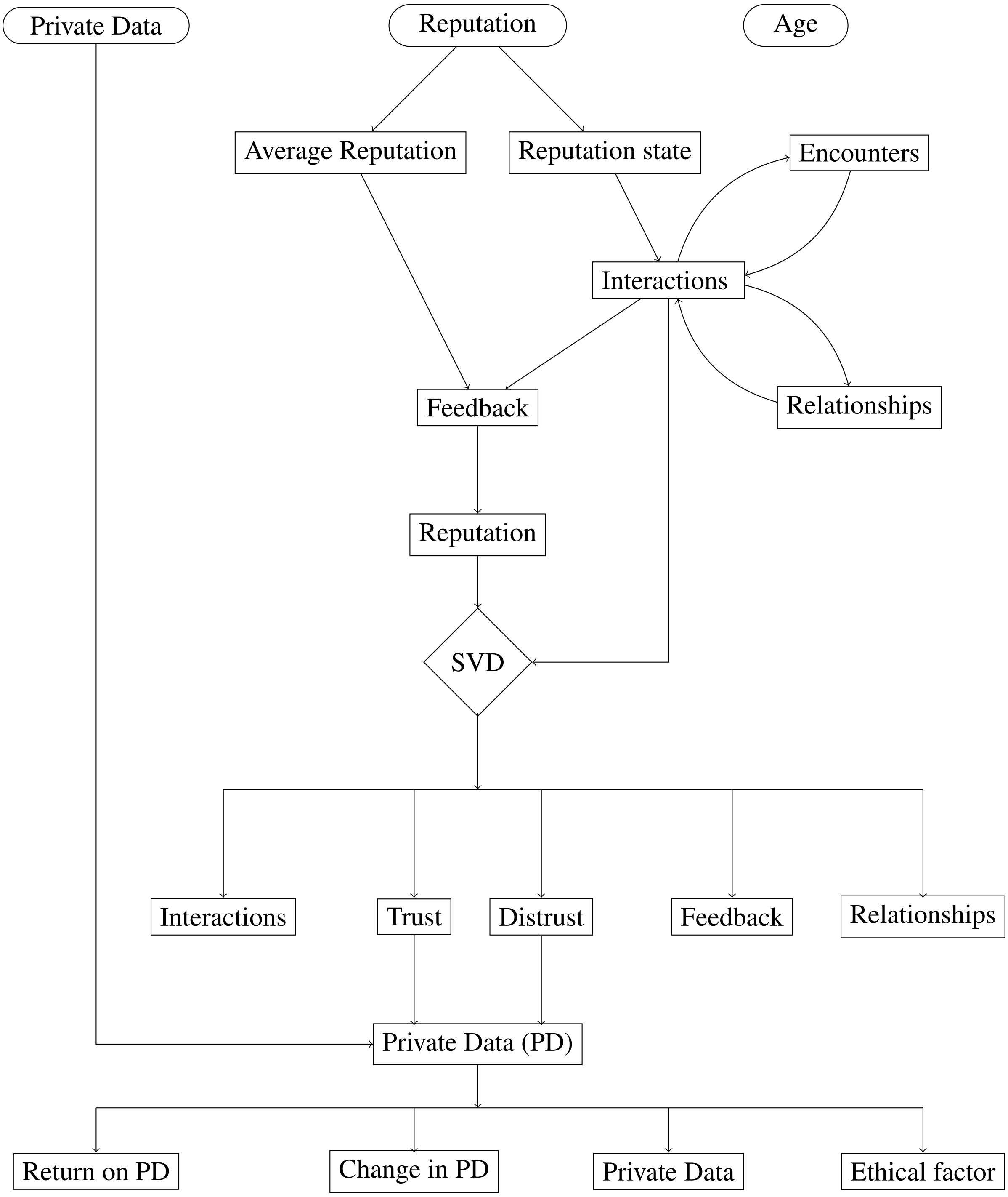

Figure 5.2: Flow chart of the activities in levels 1 and 2 of the network.

The flow chart shows the sequence of activities required in developing the algorithm for this model with the first two levels of the initial conditions and agents interactions in the social network. The chart highlights the connectivity evident in these first two levels of the model.

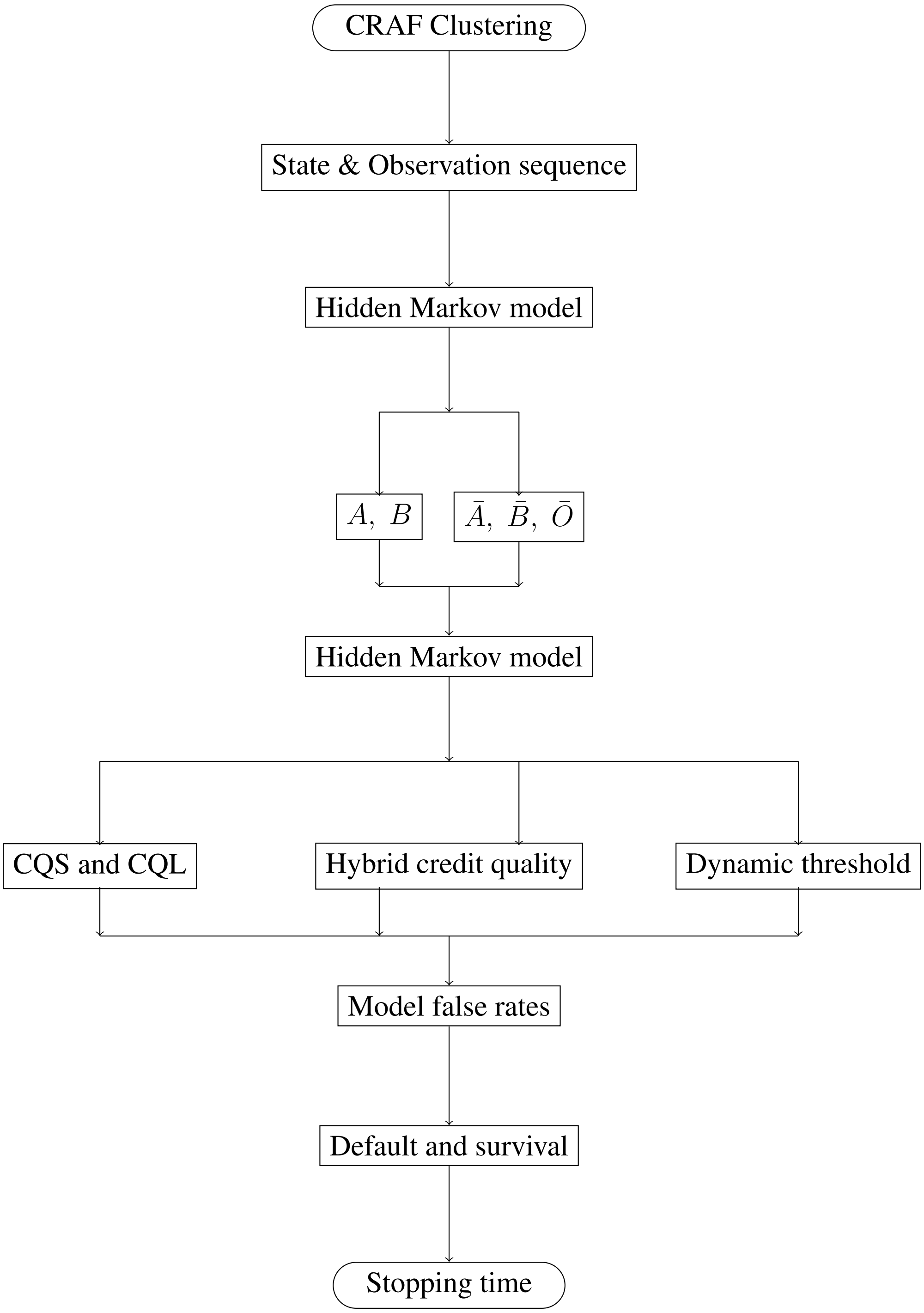

Figure 5.3: Flowchart of the activities in levels $3,\ 4$ and 5 of the network.
The flow chart shows the sequence of activities from credit risk analysis factors, the HMM estimation and training and the estimation of the credit quality as well as the default rates and stopping times. That is, from level three to level five of the model.

(a) All the agents are not likely to have a loan obligation with the same financial institution. In an ideal situation, this is not always true.

(b) Not all agents are connected to each other.

(c) Peers have the ability to review the reputation ratings of each other in the SEN

(d) The only social and economic data has a bearing on the CQS of the obligors

(e) All the credit facilities for all the agents have same premium and duration

These limitations emphasizes the fact that mathematical models only capture the key objectives of the modeler.

5.3 Model Simulation

We lack real data set of SEN interactions. There is no benchmark data set available for this study. A large scale simulation is implemented to test the system where a mixture of reputation ratings, private data, number of encounters and relationships at the onset are simulated. The computations, simulations, algorithms, graphs and all the analysis are based on Matlab version 7.0.1.

5.3.1 Uniform distribution

The simulation is based on the uniform distribution, that is, $\cup(0, 1)$. We have assumed that each agent has an equal chance in the network to interact with other agents in the system. This creates a situation where each and every agent is given an equal chance to exhibit the characteristics exhibited in the SEN. This equal probability mimics the dynamics in the different SEN variables and its influence in the model output.

5.3.2 Summary of the simulation procedure

This section provides a summary of our overall simulation approach to credit quality model from the uniform distribution. The HMM software package is accessible in Appendix (A.2). The numerical formulations for this part are explained in sections that follows. The main simulation steps for the five levels are:

(a) **Initial conditions**

For each variable, draw a sample of size N

(i) Private data, $0 < X_o < 1$

(ii) Age, $18 \leq \varphi \leq 70$

(iii) Number of encounters, $\mathfrak{N} \geq 1$

Draw a sample of $N \times N$ matrix for

(i) Reputation ratings, $1 \leq \tilde{R} \leq 5$

(ii) Relationships, $0 < \dot{R} \leq 1$

(iii) Reputation state, $-1 \leq \hat{R} \leq 1$

(iv) Average reputation, $\bar{R}$ (use reputation ratings)

A variance reduction technique is introduced at this starting stage of drawing samples

(b) **Social and economic network dynamics**

The simulation of the parameters in this level of the model are based on the results in (a) above

(i) Estimate the interaction experiences (Ψ) as [bad, neutral or good]$= [-1, 0, 1]$

(ii) Estimate network feedback $(\dot{\Psi})$

(iii) Apply SVD to estimate trust (Θ) using reputation ratings

(iv) Apply SVD to estimate distrust $(\hat{\Theta})$ and SEN risk

(v) Estimate the ethical factor $(\tilde{X})$, return of private data $(\dot{X})$, change in private data (ΔX) and the re-estimated private data (X)

(c) **Credit risk analysis factors**

The CRAFs is the collection of the data from levels (a) and (b) above. They form a crucial part of the data set to estimate the credit quality scores and default rates.

(i) The ten CRAF at time $t = 0$ are trust, distrust, SEN risk, interactions, network feedback, private data, return on private data, changes in private data, ethical factor and age

(ii) From $t \geq 1$, we include credit quality score (ϕ) and distance to default $\hat{\Delta}$ to be part of the CRAFs. This increases the factors to twelve.

(iii) CRAFs are scaled in the range of $(0, 1]$ for each of the N agents

(d) **HMM learning and training**

Matlab has inbuilt function to estimate part of the parameters in this section

(i) Estimate matrix A_n^{est} and B_n^{est} using CRAFs with supervised clustering

(ii) Compute O_n and Q_n with a sequence of length, $L \geq 1,000$

(iii) Calculate the maximum likelihoods of A_n and B_n for each agent

(iv) Calculate $\bar{A}$, $\bar{B}$ and $\bar{O}$

(e) **Credit scores and defaults**

This part is the last level of the SEN-HMM-CSD model that computes the main parameters, the CQS and the dynamic threshold among other variables.

(i) Compute the credit quality scores and credit quality level (PAGE)

(ii) Compute the dynamic threshold and hybrid credit quality score

(iii) Estimate the default rates

(iv) Estimate the model false rates, delinquent cases, defaults, survival rates and stopping time

(v) Test the model performance using accuracy rate and sensitivity analysis

Repeat the procedures from step (b) to step (e) for $t \in [1, T]$

5.4 Modeling SEN Dynamics

According to Allen and Babus (2008), a network is a collection of nodes and the links between them. The node may be individuals or firms or countries or even a collection of such entities. A link between two nodes represents a direct relation between them. A social network theory is the study of the apparently universal properties of natural networks and statistical generative models that explain such properties.

The definition given below is a modification of the definition for the graph theory to fit and suit the type of network modeled in this study.

5.4.1 Definition of social economic network

This definition was developed to cater for the social and economic factors in a network. Since a network is a graph, which is represented as $G = (\mathcal{N}, \dot{R})$.

Definition 5.1 *A SEN, $\mathcal{S} = (\mathcal{N}, \dot{R}, \mho_t, X_t)$ is a quadruplet with $\mathcal{N}$ nodes (agents or customers), a set of edges (relationships) $\dot{R}$, the set of private data, X_t, and the social dynamics $\mho_t$. The agents state and interactions evolve in discrete time and the conditions satisfied are:*

(i) $\mathcal{N}, \dot{R} \neq \emptyset$, *SEN has agents with relationships*

(ii) $\mathcal{N} = \{1, 2, \ldots, N\}$ *agents in the SEN*

(iii) $\dot{R} = \{\dot{r}_1, \ldots, \dot{r}_N\}$ *and* $\dot{R} = N(N-1)$, *set of relationships*

(iv) $\mho_t = \{\mho_t^1, \ldots, \mho_t^N\}$, *the social factors*

(v) $X_t = \{x_t^1, \ldots, x_t^N\}$, *the set of private data (economic factor)*

Further to the new definition above, let time T be the duration of the loan obligation allocated to the agents, and $\hat{t},\ (\hat{t} \in [1, T])$ be the time when the premiums are payable to the bank at the end of the month (after 30 days period) by the agents. Let $t = 2, 4, \ldots, T$ or $t = 2\hat{t}$ be the time when the credit quality of agents are estimated which is after every two months (60 day period). Let $\tau \in [t, t+1]$ be the time intervals within the 60 days period when the agents interact. The number of interactions occur at $t < \tau < t + 1$. We note that $\hat{t} < t < T$ for the premiums are paid every end month $\hat{t}$ and the credit quality of the obligors is computed every two months period at t.

We have introduced a definition of SEN and the next part introduces a theorem to prove that reputation ratings are a stochastic process.

5.4.2 Theorem of reputation ratings as a stochastic process

The following theorem proves that the reputation ratings of the agents behaves like a stochastic process. It has been coined from this study which indicates the filtration process which is a Martingale is important in the SEN-HMM-CSD model dynamics.

Theorem 5.1 *Let $\tilde{R}$ be the reputation ratings that depend on stochastic changes in the behaviour of the agents in the social and economic network. Then*

Table 5.1: Time cycles in the model

Status	Time, t	Time Intervals
Initial condition	$t = 0$	$\tau = 0$
Initial model state	$t = 0$	$\tau = 1, 2, 3$
Loan life	$t \geq 1$	$\tau = 1, 2, 3$
Loan duration	T	$t \in [1, T]$

The table 5.1 is a summary of the time cycles at different stages of the model. Each of the initial model state and the loan duration have an interval of three recent time periods as three sets of the credit risk analysis factors are used in each stage of the model.

(i) $\sigma \subset \mathfrak{F}$ *the sigma-field generated by* $\tilde{r}_i,\ i = 1, \ldots, N$

(ii) $w_i^t = \tilde{r}_i^{t+1} - \tilde{r}_i^t$ *is a function of* $\mathfrak{F}_n^t$

(iii) $E[w_i^{t+1} | \mathfrak{F}_i^t] = 0$ *for every* $t \geq 0$

Proof

Let w_n^t denote the net gain or loss in reputation at the t^{th} time period, and let $\tilde{r}_{nn'} = \tilde{r}_n$. Then, $w_n^t = \tilde{r}_n^{t+1} - \tilde{r}_n^t$ for agent n and $w_n^t,\quad t = 1, 2, \ldots, T,\quad i = 1, 2, \ldots, N$, are independent random variables with

$$P(w_n^t = \text{Decrease}) = p_d \text{ and } P(w_i^t = \text{Increase}) = q_I.$$

The total net gain or loss is expressed as

$$W_n^t = \begin{cases} \text{Increase} & \text{if, } w_n^{t+1} \geq w_n^t; \\ \text{Decrease} & \text{if, } w_n^{t+1} < w_n^t \end{cases} \tag{5.1}$$

At time T, we have

$$W_n^T = w_n^1 + w_n^2 + \ldots + w_n^T = \sum_{t=1}^{T} w_n^t.$$

Let $\mathfrak{F}_n^t = \sigma(\tilde{r}_n^1, \tilde{r}_n^2, \ldots \tilde{r}_n^T)$ denote the σ field generated by $(\tilde{r}_n^1, \tilde{r}_n^2, \ldots \tilde{r}_n^T)$. The $\mathfrak{F}_n^t \subset \mathfrak{F}_n^{t+1}$ and $\mathfrak{F}_n^t$ can be regarded as the history of the increase and decrease up to time T.

The average gain in ratings after the $(T+1)^{\text{th}}$ time given the history up to time T is:

$$\begin{aligned} E[W_n^{T+1}|\mathcal{F}_n^T] &= E[W_n^T + w_n^{T+1}|\mathcal{F}_n^T] \\ &= E[W_n^T|\mathcal{F}_n^T] + E[w_n^{T+1}|\mathcal{F}_n^T] \\ &= W_n^T + E[w_n^{T+1}] \end{aligned} \tag{5.2}$$

where $W_n^T = w_n^1 + w_n^2 + \ldots + w_n^T = \sum_{t=1}^T w_n^t$ is determined by $\mathcal{F}_n^T$ and w_n^{T+1} is independent of $\mathcal{F}_n^T$. Thus

$$\begin{aligned} E[W_n^{T+1}|\mathcal{F}_n^T] &= W_n^T + E[w_n^{T+1}|\mathcal{F}_n^T] \\ &= W_n^T + p_d + q_I \\ &= \begin{cases} W_n^T, & \text{if } p_d = q_I; \\ > W_n^T, & \text{if } p_d < q_I; \\ < W_n^T, & \text{if } p_d > q_I. \end{cases} \end{aligned} \tag{5.3}$$

In the first case, W_n^T is called a martingale. It is a submartingale in the second case and a supermartingale in the third case. The fact that the reputation ratings is a martingale shows that changes in the dynamics and interactions are stochastic.

5.5 Network Initial Conditions

We refer to this data as the initial conditions as it forms the basis for the extraction of any other data that follows in the model. Five of the parameters to form the initial conditions are; age of the agents, private data, reputation ratings (and the mean of the reputation ratings for each agent, the maximum reputation level, reputation state), the relationship matrix of the agents and the number of interactions of the agents.

Table 5.2 has the initial condition variables and the accompanying notations. These are simulated or estimated at time $t = \hat{t} = 0$ as τ changes during this time period.

5.5.1 Private data

At $t = 0$, each agent has private or the economic data before they even visit the financial institution to sign up for a loan obligation. This data is referred to as the economic data

Table 5.2: Initial Conditions in the Social Network

Variable	Conditions	Function of
Age	$18 \leq \varphi \leq 70$	$\varphi \sim \cup(0,1)$
Private data	$0 < X_0 < 1$	$X \sim \cup(0,1)$
Reputation ratings	$1 \leq \tilde{R} \leq 5$	$\tilde{R} \sim \cup(0,1)$
Relationship factors	$0 < \dot{R} \leq 1$	$X \sim \cup(0,1)$
Reputation state	$-1 \leq \hat{R} \leq 1$	$f(\tilde{R})$ $\hat{R} = [-1,0,1] =$ [**L**, **M**, **H**]
Interaction encounter frequency	$\mathfrak{N} \in [1,8]$	$\mathfrak{N} \sim \cup(0,1)$
Average reputation	$0 < \bar{R} \leq 5$	$\bar{R} = \frac{1}{N}\sum_{j=1}^{N} \tilde{R}_{ij}$
Interactions experience	$-1 \leq \Psi \leq 1$	$f(\hat{R}, \dot{R}, \mathfrak{N})$ $\hat{R} = [-1,0,1] =$ [**B**, **N**, **G**]
Network feedback	$0 < \dot{\Psi} < 1$	$f(\bar{R}, \Psi)$

The variables highlighted in this table are explained further. The numerical formulations and computations are explained in detail.

of the agents since we are dealing with a socio-economic network. This set of the private data for the $\mathcal{N}$ agents is given by the vector

$$X_0 = \{x_0^1, x_0^2, \ldots, x_0^N\}, \text{ and } 0 < x_n < 1.$$

The vector X_0 is drawn at time $t = 0$. This set of information is available to the bank but limited in the sense that the bank is not aware of the actual interactions that take place before the first premium is paid at the end of the month at $\hat{t} = 1$.

We know that the interactions affect the obligors behaviour that in turn impact their credit quality levels. X_0 limits the net worth or private data of the agents, $(0 < X_0 \leq 1)$, ideally to enhance our computations. This initial vector is generated at time $t = 0$ using simulation but the minimum private data is placed at $0.3 \leq x_0^n < 1$. The 0.3 is assumed to be the acceptable minimum in terms of the private data for agents to gain access to a financial institution consumer loan facilities. We note that private data is the socio-economic value gained through social capital or an economic activity like investment, etc, that benefits the agent. This data is generally observed to a limited extent by other agents and the institution that issues the loan.

5.5.2 Reputation ratings

Each agent knows the reputation rating of each other in the SEN. Let $R_i = \{r_{1i}, r_{2i}, \ldots, r_{ni}\}$ be the reputation rating agent $i = 1, 2, \ldots, n$ receives from the other $N-1$ agents in the SEN. Agents have the reputation rating information even before joining the bank loan portfolio. Agents are aware of whom they have been interacting with even before approaching the bank for a loan. Therefore, they have information on the reputation levels of each other.

$$\tilde{R}_t = \begin{cases} \tilde{r}_{ij} = 5 & \text{if } i = j; \\ 1 \leq \tilde{r}_{ij} \leq 5 & \text{if } i \neq j; \end{cases} \tag{5.4}$$

The peer to peer reputation rating is based on the five star scale: $1-$lowest, $2-$low, $3-$medium, $4-$good and $5-$high. That is, $\tilde{R} \in [1, 2, 3, 4, 5]$. Therefore, each agent is expected to rate the other $N-1$ agents. As would be ideal in life situations, if we are to rate ourselves, we would likely give a maximum score of 5. Therefore, the matrix $\tilde{R}$ has $\tilde{r}_{ii} = 5,\ \forall i$.

5.5.3 Average reputation rating

Agents feedback system is important in aiding agents make informed decisions based on the information from other agents, how they rate each other. So, apart from an individual reputation rating, an agent will also use the rating from other agents to make an informed decision. Let $\bar{R}$ be the average reputation rating extracted from the matrix $\tilde{R}$, the reputation ratings matrix using the SVD for each agent. This represents the average value an agent receives from the peers in the network. The average reputation rating is computed with;

$$\bar{R} = \frac{1}{N} \sum_{j=1}^{N} \tilde{R}_{ij} \tag{5.5}$$

5.5.4 Reputation state

The reputation state shows the overall rating an agent receives from the other $N-1$, which is a perception level in the network. Let $\hat{R}$ be the reputation state of the agents. We let $\hat{R} \in [-1, 0, 1] =$ [low, Medium, High]. An agent is then classified as follows;

$$\hat{R}_t = \begin{cases} -1 \text{ if } & \hat{r} < 2.33 & \text{Low;} \\ 0 \;\text{ if } & 2.33 \leq \hat{r} \leq 3.66 & \text{Medium;} \\ 1 \;\text{ if, } & \hat{r} > 3.66 & \text{High;} \end{cases} \tag{5.6}$$

We have partitioned the reputation scores $[1, 5]$ into three divisions.

5.5.5 Relationship matrix

Let $\dot{R}$ be a relationship matrix and $\dot{r}_{nn}$ be the relationship intensity between agent n and agent n'. We set $0 < \dot{R} \leq 1$. This matrix defines the mutual links between the agents irrespective of the agent reputation but linked to the interaction experiences each agent gains from the other. Human beings tend to link up to others based on the mutual connection irrespective of the reputation of the agent. That is why we have people ganging up to engage in good and evil acts in the society.

$$\dot{R}_t = \begin{cases} \dot{r}_{ij} = 1 & \text{if } i = j; \\ 0 < \dot{r}_{ij} < 1 & \text{if } i \neq j; \end{cases} \tag{5.7}$$

In sociology, homophily is a principle which states that people tend to form ties with other people who have similar characteristics (like associating with like). This indicates that strong ties shows a high similarity amongst the agents.

5.5.6 Interactions frequency

The number of times an agent interacts with another agent is also referred to as the interaction frequency or the interaction encounter frequency. Let $\mathfrak{N}_{nn'}$ be the number of encounters between agent n and n'. Let agents interact on frequent intervals with a minimum of 1 interaction at any given time period of two months and a maximum of 8 interactions in the same period. We assume a maximum of 1 interaction per week and this translates to 8 interactions for a period of two months, during the time period $[t, t+1]$. The number of interactions can increase or decrease depending on the experiences they accumulate from the SEN dynamics. These number of interactions of agent n to agent n' will then change depending on the interaction experiences.

5.5.7 Age of the agents

The age of the agents is the only demographic variable we have included in this study. Let $18 \leq \varphi \leq 70$ be the vector for the age of the agents. The age is expected to change annually and the new variable reflected in the model. Thomas (2000) observes that there are social issues in using social variables as credit analysis tools. It is illegal to use some characteristics like race, sex and religion.

5.5.8 Interaction experiences

Interaction experiences are based on the agents interactions encounters that have a specific outcomes. The experiences are classified in three levels, namely,

$$\begin{aligned} [-1, 0, 1] &= [\text{bad, neutral, good}] = [B, N, G] \qquad (5.8) \\ &= [\psi_1, \psi_2, \psi_3] \qquad (5.9) \end{aligned}$$

The Markov process in figure 5.4 offers insights on the possible transitions an agent can make. This is from the interaction experiences. We have three scenarios based on figure 5.4, the agents relationships and the reputation state. Three scenarios are evident in the interactions. First, an agent starts with low reputation state with either high or low relationship with other agents. Second, an agents starts at a medium reputation state coupled with low, medium or high relationship levels with other agents. Third, an agent starts at a high reputation state and has a low or high relationship levels with other agents. We express the formulation for the interaction experiences using the relationships and reputation ratings as:

$$\dot{F}_t = \begin{cases} -1 & \text{if } \Psi = 1 \,\&\, -1 < \bar{R} < \frac{5}{3}; \\ 0 & \text{if } \Psi = 0 \,\&\, \frac{5}{3} \leq \bar{R} \leq \frac{10}{3}; \\ 1 & \text{if } \Psi = -1 \,\&\, \bar{R} > \frac{10}{3}; \end{cases} \qquad (5.10)$$

Matrix $\dot{F}$ entries are then used to extract the estimated interaction experiences using the singular value decomposition.

5.5.8.1 Interaction encounter

Social interaction is defined as participating in social networks (Sjoerd , 2004). We define an interaction encounter as a situation where an agent interacts with another agent. Such

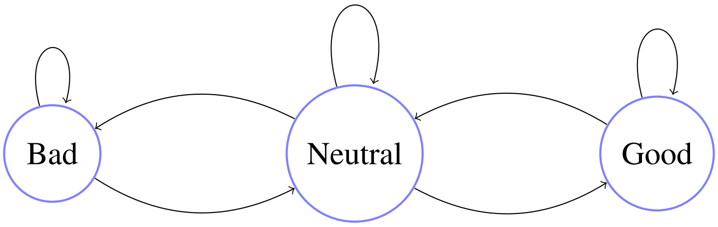

Figure 5.4: A three state Markov Model for the agents relationship experiences
A Markov Model showing the transitions between bad (B), neutral (N) and good (G) experiences of the agents. The only limitation is that an agent cannot move from bad to good experience level, but has to first pass through the neutral level and vice versa.

interactions might include involvement in an investment project, financial advice, being members of the same social group, communication via email, telephone, among other activities. We can simply define it as an interaction where individuals generate social capital and/ or engage in economic activities, enhancing the outcome of their actions. Connections facilitate timely access to important information on a timely manner.

Social capital investment increases with the occupational returns to social skills. Communities with more aggregate social capital investment generally have an increase in social capital as individuals are able to benefit from the resources embedded in social networks (Nan , 1999). Therefore, an interaction encounter occurs in a dynamic environment where agents try to optimize their decisions and interactions bound in space and time.

In psychology and behavior science, an agents behavior change is associated with interaction contextual information (Liu and Datta , 2012). They observed that inclusion of interaction context information captures agent behavior changes that are relatively infrequent. Further, they observed that a past good record can conceal any sudden change in behavior. A careful investigation of interaction contextual information can reflect the agents dynamic behavior in a better way.

5.6 Social and Economic Factors

A SEN is where the primary action entails economic transactions under the structure of the social capital. Agents and their actions are viewed as interdependent rather than independent autonomous units. As the SEN is dynamic, trust and reputation levels are bound to change over time and in turn affect the net worth of agents. linkages between agents are channels for transfer or 'flow of resources'. We normally lose or gain more wealth depending on the agents we are interacting with in a network. The initial reputation ratings changes are guided by the outcome of interactions or encounters the agents have with each other.

Table 5.3: Credit risk analysis factors classification

Factor	Category	Symbol
Interactions experience	Social	Ψ
Trust level	Social	Θ
Distrust level	Social	$\hat{\Theta}$
SEN risk factor	Social	$\tilde{\Theta}$
Network feedback	Social	$\dot{\Psi}$
Private data	Economic	X
Return on private data	Economic	$\bar{X}$
Ethical factor	Economic	$\tilde{X}$
Changes in private data	Economic	ΔX
Age of agents	Demographics	φ

The table shows the social and economic factors derived from the SEN dynamics and are outlined in this section.

5.6.1 Reputation ratings

The dynamic interactions of the agents enables them to collect almost accurate information on trust and reputation levels of each other. An independent rating of each other increases transparency and reduces the rate of collusion amongst the agents. For an effective rating process, the agents have to interact. As the SEN comprises of N agents, peer to peer review of each other is crucial in estimating the reputation of each agent. The changes is a combination of the previous reputation rating, the feedback from the network and the individual interaction experience expressed as:

$$\tilde{R}_{t+1} = \breve{\alpha}[\tilde{R}_t + \dot{\Psi}_t + \Psi_t] \tag{5.11}$$

Where R_t is the reputation at time t, $\dot{\Psi}_t$ is the network feedback, Ψ_t is the interactions experiences of the agents and $\breve{\alpha}$ is a constant to limit $1 \leq \tilde{R}_{t+1} \leq 5$. Modeling trust is through a review of one agent by the other $N - 1$ agents with the assigned reputation ratings. The rating scale is in the range of $[1, 2, 3, 4, 5]$, where $1-$lowest, $2-$low, $3-$medium, $4-$good and $5-$high . Thus, we let the individual scale of ratings be denoted as $\tilde{R} = \{\tilde{r}_1, \tilde{r}_2, \tilde{r}_3, \tilde{r}_4, \tilde{r}_5\}$. Agent n is rated in the scale of $1, 2, 3, 4, 5$ by the other $N - 1$ agents and they rate themselves with a score of 5 because we humans have a tendency to deceive ourselves and to justify our actions even if others deem them to be fraud or deceit. At any instance $t = 1, 2, \ldots, T$, a matrix $\tilde{R}_t$ is generated

$$\tilde{R}_t = \begin{cases} \tilde{r} = 5 & \text{if } \; n = n'; \\ 1 \leq \tilde{r} \leq 5 & \text{if } \; n \neq n'; \end{cases} \tag{5.12}$$

The reputation ratings depends on the stochastic changes in the behaviour of agents in the SEN and the availability of the set of information.

5.6.2 Social factors estimation

The five social factors are interaction experience, trust, distrust, SEN risk and network feedback with age as the demographic factor. SVD technique extracted all the five social factors as the technique is a data reduction method that involves taking sets of high dimensionality data and reduces it to a lower dimensional space (Kalman , 1996).

5.6.2.1 Trust estimation

Trust levels for the agents are extracted from the reputation ratings R_t with the aid of the SVD. The SVD of the matrix R_t gives rise to the left eigenvectors, singular values and the right eigenvectors. Similarity in the trust levels of the agents is then estimated and the data scaled. The best rank one approximation by the SVD is used

$$\Theta = ||\tilde{R} - VV^T\tilde{R}^T|| \tag{5.13}$$

5.6.2.2 Distrust level

We subtract the reputation rating matrix with $\check{R}_t = 5 - \tilde{R}_t$ to give us the residue of what remains from the trust, which is the distrust rating of the agents. The matrix $\check{R}_t$ is extracted with SVD for the three components as the case in trust levels. Similarity in the agents distrust levels are estimated and the data is scaled. The rank one approximation for the distrust levels in the SEN is computed with

$$\hat{\Theta} = ||\check{R} - VV^T \check{R}^T|| \tag{5.14}$$

5.6.2.3 SEN risk factor

The SEN risk factor is the risk introduced in the network by the agents as a relative factor of the distrust and trust levels. The matrix for the values of the risk factor before extraction with SVD is estimated as

$$\acute{R}_t = \frac{\check{R}_t}{R_t} \tag{5.15}$$

The rank one SVD approximation for distrust level in the SEN is estimated with

$$\tilde{\Theta} = ||\acute{R} - VV^T \acute{R}^T|| \tag{5.16}$$

5.6.2.4 Network Feedback

The interactions experiences and the mean reputation rating for each agent in the network is used to estimate the network reputation feedback or the SEN feedback mechanism. Let $\dot{\Psi}$ be the SEN feedback. We use SVD to extract its matrices and estimate the feedback ratings for the agents in the network using the procedure applied for the trust and distrust levels. The rank approximation of using SVD for the network feedback is

$$\dot{\Psi} = ||\hat{F} - VV^T \hat{F}^T|| \tag{5.17}$$

Where the matrix $\hat{F}$ is estimated from $\bar{R}$, the average reputation ratings, and Ψ, the interactions experiences. Matrix $\hat{F} = f(\bar{R}_t, \Psi_t)$ with $\Psi = [-1, 0, 1]$ and $0 < \bar{R}_t \leq 1$ is expressed as

$$\hat{F}_t = \begin{cases} 1.0 & \text{if } \Psi_t = 1 \;\; \& \;\; \bar{R}_t > \frac{2}{3}; \\ 0.5 & \text{if } \Psi_t \geq 0 \;\; \& \;\; \frac{1}{3} \leq \bar{R}_t \frac{2}{3}; \\ 0 & \text{if } \Psi_t = 0 \;\; \& \;\; \frac{1}{3} \leq \bar{R}_t \leq \frac{2}{3}; \\ -0.5 & \text{if } \Psi_t = 0 \;\; \& \;\; \frac{1}{3} \leq \bar{R}_t \leq \frac{2}{3} \\ -1.0 & \text{if } \Psi_t = -1 \;\; \& \;\; \bar{R}_t < \frac{2}{3} \end{cases} \tag{5.18}$$

5.6.2.5 Agents' age

The agents' age is the only demographic variable that has been used in this work. Let $\varphi_0 \in [18, 70]$ be the age of the agents at time $t = 0$. Then, the parameter φ is expected to change in each time period after one year.

5.6.2.6 Encounter experience

This is a social factor extracted from the gains or losses an agent makes by interacting with other agents in the network. They are the cumulative values of the bad, neutral or good outcome of the dynamic interactions. SVD aids in extracting the experiences of each agent from the matrix Ψ (interactions experience).

$$\Psi = ||\hat{R} - VV^T\hat{R}^T|| \tag{5.19}$$

5.6.3 Economic factors estimation

The four economic data emanates from the private data, namely, private data changes, private data return and ethical factor.

5.6.3.1 Private data

We now turn our attention to the private data, $X_t(\Theta_t,\ \hat{\Theta}_t, \epsilon_x)$, that is a function of the trust levels, distrust and an error component. Since the vector X_0 exists, changes in X_t at a time interval $[t,\ t+1]$ is based on the trust level, distrust level and noise in the network that cannot be fully eliminated by the SVD dimensionality reduction. The singular values of the trust matrix are used to compute the condition number which is a ratio of the largest to the smallest singular value. This number measures the sensitivity of the eigenvalues.

The number acts as a discount factor in estimating the new private data. The condition number is;

$$\mathfrak{C} = \frac{\sigma_{\text{Max}}(\Theta)}{\sigma_{\text{Min}}(\Theta)} \tag{5.20}$$

The private data is estimated using

$$\begin{aligned} \Delta X_{t+1} &= \left[\frac{1}{\mathfrak{C}}(\Theta - \hat{\Theta})X_t + \epsilon_x\right] \\ X_{t+1} &= X_t + \left[\frac{1}{\mathfrak{C}}(\Theta - \hat{\Theta})X_t + \epsilon_x\right], \ \text{with } \epsilon_x = \frac{1}{\max(\text{eigenvalue}(\tilde{R}))} \end{aligned} \tag{5.21}$$

where ΔX_{t+1} is the change in the private data, ϵ_x is a random variable which is the error term or noise in the network, Θ_t is the trust level, $\hat{\Theta}$ is the distrust level and $\hat{c}$ is the condition number of the trust matrix.

5.6.3.2 Private data changes

The changes in private data is the differences in the private data between two time periods. Let ΔX be the changes in private data. Then,

$$\Delta X_{t+1} = X_{t+1} - X_t$$

.

5.6.3.3 Private data return

At each time period, the agent private data return is estimated using,

$$\dot{X} = \frac{X_{t+1}}{X_t} - 1 \tag{5.22}$$

This is the return an agent gains or losses from the changes in the private data at any given time period.

5.6.4 Ethical factor

We let $\tilde{X}_0 = X_0$. Then, the ethical factor is computed as;

$$\tilde{X}_{t+1} = \tilde{X}_t - \frac{1}{\tilde{\mathfrak{C}}}\tilde{\Theta}_t\tilde{X}_t + \epsilon_x, \ \text{ with } \tilde{\mathfrak{C}} = \frac{\sigma_{\text{Max}}(\tilde{\Theta})}{\sigma_{\text{Min}}(\tilde{\Theta})} \tag{5.23}$$

where $\tilde{X}_{t+1}$ is the ethical value of the agent, ϵ_x is a random variable which is the error term or noise in the network, $\tilde{\Theta}_t$ is the SEN risk factor and $\tilde{c}$ is the condition number of the SEN risk factor matrix.

5.7 Credit Risk Analysis Factors

The variables are categorized into three, namely, social, economic and demographic factors. The symbols for each of the variable and the full description is in table 5.3. We can express the credit risk analysis factors in a vector form

$$\begin{aligned} \Gamma &= [\varphi,\ X,\ \tilde{X},\ \dot{X},\ \Delta X,\ \Theta,\ \dot{\Psi},\ \Psi,\ \hat{\Theta},\ \tilde{\Theta}] \quad &(5.24)\\ &= [\text{Age, Trust, Distrust, SEN risk, Interactions, Feedback,} \quad &(5.25)\\ &\quad \text{Private data, Ethical, Return on private data, Changes in private data}] \quad &(5.26)\\ & &(5.27) \end{aligned}$$

5.7.1 Credit score and dynamic threshold value

Once the first credit score values and dynamic threshold valu are computed at the end of time $t = 0$, then from time $t \geq 1$, we have a set our CRAFs to twelve factors. This is the set of the ten social and economic factors plus the two factors of the credit scores and dynamic threshold values. The new set of CRAFs is,

$$\Gamma = [\varphi,\ X,\ \tilde{X},\ \bar{X},\ \Delta X,\ \Theta,\ \dot{\Psi},\ \Psi,\ \hat{\Theta},\ \tilde{\Theta},\ \hat{\Delta},\ \phi] \tag{5.28}$$

Table 5.4 shows the two new factors after the first computation of the credit quality score and the default threshold.

Table 5.4: Emitted credit quality scores

Factor	Category	Symbol	Computation process
Credit quality score	Credit risk score	ϕ	$\phi = P(A, B, \pi, O, Q\|\lambda)$
Distance to default	Credit risk score	$\hat{\Delta}$	$\hat{\Delta} = \phi - \bar{\phi}$

Each of the value in the table is obtained from the hidden Markov model after the credit scores are computed. The two variables are scaled as the rest of the ten factors and included in the CRAF at time $t \geq 1$ for the next time period credit quality analysis. It is only at time $t = 0$ that these two variables are not included in the next computation of the credit scores as this is the starting point of our computations.

5.7.2 Credit factors set

At time $t = 0$, we have ten CRAFs for clustering into state sequences and observation sequences for each agent. Therefore each agent has a set of ten state and ten observation sequences at time t which are estimated at each time period τ for we have $\tau = 1, 2, 3$ at $t = 0$. At $t \geq 1$, we increase the CRAF from ten to twelve by adding the credit quality and distance to default for each agent. The distribution of the credit risk analysis factors ranges between 30 and 36 values of any given data set for each of the N agents as indicated in equation 5.29.

$$\Gamma_{N\times 1} = \begin{cases} \text{30 credit factors at} & t = 0, \quad \text{use } \tau_1, \tau_2, \tau_3; \\ \text{32 credit factors at} & t = 1, \quad \text{use } t = 0(\tau_2, \tau_3), t = 1; \\ \text{34 credit factors at} & t = 2, \quad \text{use } (t = 0, \tau_3), t = 1, 2; \\ \text{36 credit factors at} & t \geq i, \quad \text{use } t = i-2, i-1, i, \quad \forall i \in [3, T] \end{cases} \tag{5.29}$$

This is based on the fact that $\Gamma_{t=0} = 10$ and $\Gamma_{t\geq 1} = 12$ and that for each HMM training, we use a three τ period. We observe that $t = 0$ is the period before the agents become part of the loan portfolio and covers a period of six months. Estimation of CRAFs are done three times in this time period t at τ_1, τ_2 and τ_3.

5.7.3 Data scaling

To ease the computation process, we introduce uniformity in the CRAFs by scaling all the ten variables at time, $t = 0$ in the range of $(0, 1]$ and also for the twelve variables from time $t \geq 1$ in the same range. Let $\mathbb{y}$ be the data set to be scaled. The commonly used method is

$$\hat{\mathbb{y}} = \frac{\mathbb{y} - \text{min}(\mathbb{y})}{\text{max}(\mathbb{y}) - \text{min}(\mathbb{y})}$$

But we modified the above method to

$$\hat{\mathbb{y}} = \frac{\mathbb{y} - \text{min}(\mathbb{y}) + \epsilon_y}{21\epsilon_y}, \quad \text{with } \epsilon_y = 0.05(\text{max}(y) - \text{min}(y)) \tag{5.30}$$

Where $\hat{\mathbb{y}}$ is the scaled data ($\hat{\mathbb{y}} \in (0, 1]$) and $\mathbb{y}$ is the original data set that is to be scaled into the interval $(0, 1]$. The data scaling eliminates the zero value to transform the data for the clustering for estimating HMM parameters.

5.8 HMM Parameters Estimation

The HMM we utilize are ergodic in nature with the Markov chain underlying the HMM being irreducible and aperiodic. An HMM is irreducible if each state is reachable with non-zero probability from every other one. It is aperiodic if it has at least one aperiodic state. A state i is aperiodic if it does not recur with a cyclic period, that is if the greatest common divisor of times $t > 0$ such that $P(q_{t+1} = i|q_t = 1) > 0$ is 1. A periodicity is guaranteed if one state has a self transition with non-zero probability (Ehab and Sassone , 2013).

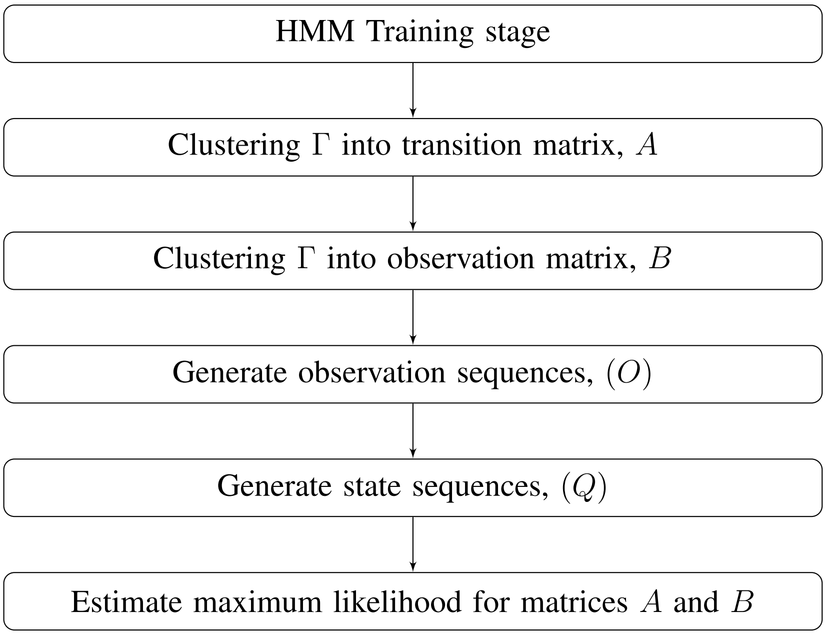
HMM Training stage
Clustering Γ into transition matrix, A
Clustering Γ into observation matrix, B
Generate observation sequences, (O)
Generate state sequences, (Q)
Estimate maximum likelihood for matrices A and B

Figure 5.5: The flow of the sequence of events in HMM learning and training

The flow chart outlines the clustering process estimation of the state transitions and observation matrices for each of the agents and for the default threshold. Hybrid credit quality, default threshold and credit quality probabilities are also estimated at this stage.

5.8.1 Transition and observation matrices

A two state Markov model for the transitions of the agents in the network is depicted in figure 5.6.

Each agent has two state transitions, $S = \{S_1, S_2\} = [\text{Low score}, \text{High sCore}] =$

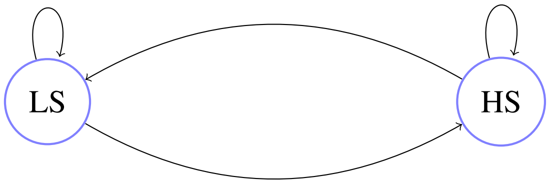

Figure 5.6: A two state Markov Model for the agents Interactions
A Markov process showing the ergodic transitions between low score (LS) and high score (HS) for the state transitions of the agents in the network.

[L, H]. For the observations, the symbols are generated by :

$$b_j(\kappa) = P(O_t = \kappa | q_t = j), \quad \kappa = 1, 2, 3, 4 \;\; j = 1, 2 \tag{5.31}$$

where q_t is the states and O_t are the four observation symbols, namely, poor, average, good and excellent credit quality levels. The observation profile is heterogeneous to all the agents as they have different observation probability given by the matrix $B_n^t = \{b_{j\kappa}\}$. By default, the Hidden Markov Model functions in Matlab begin with the model in state 1 at step 0. We want to assign different probabilities to the initial states. This makes our Markov chain time independent and thus the ability to change the transition and observation matrices at each time $t \in [1, T]$ is simplified. Again, see Matlab (2003) on how to change the probabilities of the initial states.

The transition and observation matrix for estimating the dynamic default threshold has its roots on the properties of matrix addition and subtraction. A mean of the observation and transition matrices of all the agents is obtained to estimate the default threshold observation and transition matrix set.

5.8.2 Clustering

Clustering is applied to the set of data Γ to emit the transition matrix and observation matrix of each agent. We are interested in using the CRAFs to estimate the transition matrix A and the observation matrix B. For the transition matrix A, we have $(\gamma \in \Gamma)$, with $A_{2\times 2} = A_{est}$

$$A_{2\times 2} = \begin{cases} LL & \text{if} \quad \gamma_{i,j} < \gamma_{i,j+1} \;\; \& \;\; \gamma_{i,j+1} < \gamma_{i,j+2}; \\ LH & \text{if} \quad \gamma_{i,j} < \gamma_{i,j+1} \;\; \& \;\; \gamma_{i,j+1} > \gamma_{i,j+2}; \\ HL & \text{if} \quad \gamma_{i,j} > \gamma_{i,j+1} \;\; \& \;\; \gamma_{i,j+1} < \gamma_{i,j+2}; \\ HH & \text{if} \quad \gamma_{i,j} > \gamma_{i,j+1} \;\; \& \;\; \gamma_{i,j+1} > \gamma_{i,j+2} \end{cases} \tag{5.32}$$

Where LL indicates that a Low score is followed by a Low score; LH indicates that a Low score is followed by a High score; HL indicates that a High score is followed by a Low score; and HH indicates that a High score is followed by a High score. The clustering for matrix A can also be expressed in a simple form as;

$$A = \begin{array}{c c} & \begin{array}{cc} L & H \end{array} \\ \begin{array}{c} L \\ H \end{array} & \begin{pmatrix} LL & LH \\ HL & HH \end{pmatrix} \end{array}$$

For the observation matrix B, we have $(\hat{\gamma} = \gamma - \text{mean}(\gamma))$ and $\frac{\max(\hat{\gamma})}{4} = u$, $\frac{\min(\hat{\gamma})}{4} = l$ with $B_{2\times 4} = B_{est}$

$$B_{2\times 4} = \begin{cases} LP & \text{if} \quad \hat{\gamma}_{i,j} < l_{i,j}; \\ LA & \text{if} \quad l_{i,j} \leq \hat{\gamma}_{i,j} < 2l_{i,j}; \\ LG & \text{if} \quad 2l_{i,j} \leq \hat{\gamma}_{i,j} < 3l_{i,j}; \\ LE & \text{if} \quad 3l_{i,j} \leq \hat{\gamma}_{i,j} < 0; \\ HP & \text{if} \quad 0 \leq u_{i,j} < u_{i,j}; \\ HA & \text{if} \quad u_{i,j} \leq \hat{\gamma}_{i,j} < 2u_{i,j}; \\ HG & \text{if} \quad 2u_{i,j} \leq \hat{\gamma}_{i,j} < 3u_{i,j}; \\ HE & \text{if} \quad \hat{\gamma}_{i,j} \geq 3u_{i,j}; \end{cases} \tag{5.33}$$

Where LP indicates that a Low score is connected with a Poor credit quality; LA indicates that a Low score is connected with an Average credit quality; LG indicates that a Low score is connected with good credit quality; LE indicates that a Low score is connected with an Excellent credit quality; HP indicates that a High score is connected with Poor credit quality; HA indicates that a High score is connected with an Average credit quality; HG indicates that a High score is connected with Good credit quality and HE indicates that a High score is connected with an Excellent credit quality. This information can also be expressed in a short form as a matrix B as;

$$B = \begin{array}{c} \\ L \\ H \end{array} \begin{array}{c} \begin{array}{cccc} P & A & G & E \end{array} \\ \begin{pmatrix} LP & LA & LG & LE \\ HP & HA & HG & HE \end{pmatrix} \end{array}$$

5.8.2.1 HMM parameters

Given the transition matrix A_{est} and observation matrix B_{est}, the other parameters can be estimated.

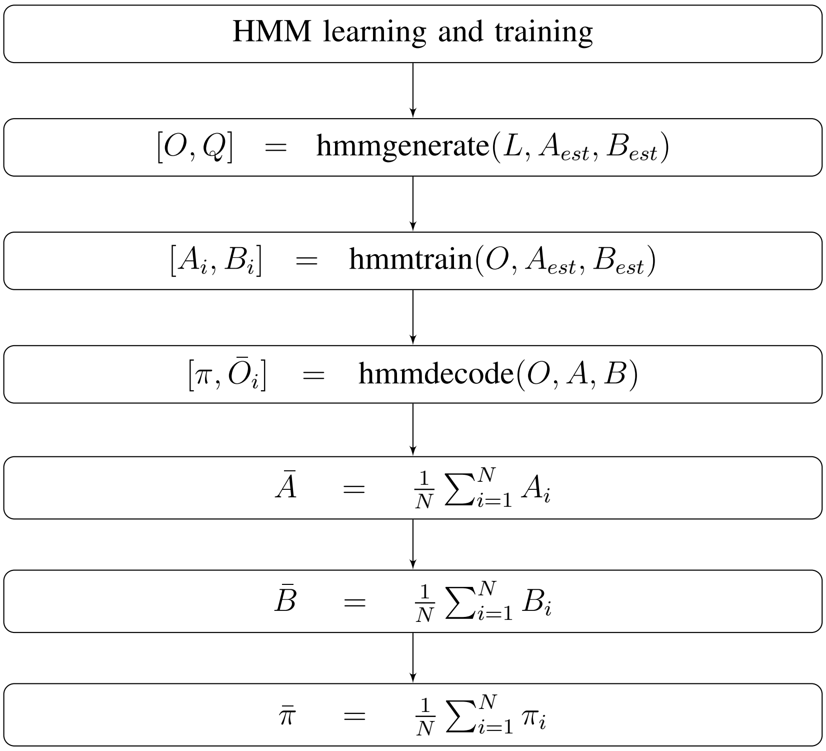
HMM learning and training
$[O, Q] \quad = \quad \text{hmmgenerate}(L, A_{est}, B_{est})$
$[A_i, B_i] \quad = \quad \text{hmmtrain}(O, A_{est}, B_{est})$
$[\pi, \bar{O}_i] \quad = \quad \text{hmmdecode}(O, A, B)$
$\bar{A} \quad = \quad \frac{1}{N}\sum_{i=1}^{N} A_i$
$\bar{B} \quad = \quad \frac{1}{N}\sum_{i=1}^{N} B_i$
$\bar{\pi} \quad = \quad \frac{1}{N}\sum_{i=1}^{N} \pi_i$

Figure 5.7: The HMM training and learning procedure

The flow diagram outlines the HMM training an learning procedure for the local and global HMM parameters in the model.

5.8.3 Multiple observations

The emissions from the HMM corresponding to individual agents are as a result of the CRAFs. We shall assume that the observations of the different agents are independent, as dependence would weaken the need for privacy. The observation vector at time t is $O_t = \{O_t^{(1)}, O_t^{(2)}, \ldots, O_t^{(N)}\}$ and they form a set to estimate the credit quality of an

Table 5.5: Local and global estimated parameters

Parameter	Local	Global
Transition matrix	A	$\bar{A}$
Observation matrix	B	$\bar{B}$
Initial state probability	π	$\bar{\pi}$
Observation sequence	O	$\bar{O}$
State sequence	Q	$\bar{Q}$

The HMM learning and training parameters use either the local or global matrices and the observation sequence or the state sequence. These parameters estimate the CQL, CQS and HCQS for each obligor and also the overall dynamic threshold that measures the system likely default rates.

obligor. Each agent has an observation which is derived from the CRAFs based on the dynamic strategies and connections in the network.

Each agent has a set of possible observations V_t and the observation probability matrix B is unique for each agent. As the observations are made at time $t \in [1, T]$ each agent makes its own observations of the SEN and these observations are secret.

The observation symbol emitted by HMM for each agent are combined to estimate the dynamic default threshold of all the agents. We note that the observation and transition matrices were derived from the mean of the individual agent observation and transition matrix together. We denote the default threshold transition matrix as $\bar{A}$ and the observation matrix as $\bar{B}$ with the observation symbols from all the agents after training as $\bar{O}$.

5.9 Credit scores and dynamic threshold

Credit quality analysis is key to the success of this model and we outline its estimation in this section. Table 5.6 shows the credit quality, default threshold, the hybrid credit quality tuples and the delinquent estimation for the model training parameters and for the estimation of the respective probabilities. The four variables are intertwined where at any given time t we have one default threshold while we have N values for the hybrid credit quality, credit quality scores, credit quality levels and delinquent cases.

Table 5.6: Model training and credit quality scoring parameters

Variable	Symbol	Parameters
Credit quality tuple	ϕ	$\lambda = (A, B, \pi, O, Q)$
Default Threshold tuple	$\bar{\phi}$	$\bar{\lambda} = (\bar{A}, \bar{B}, \bar{\pi}, \bar{O}, \bar{Q})$
Hybrid credit quality tuple	$\hat{\phi}$	$\hat{\lambda} = (\bar{A}, \bar{B}, \bar{\pi}, O, Q)$
Credit Quality	ϕ_i^t	$\phi_i^t = P(O_1^t, \ldots, O_n^t \vert \lambda_i)$
Hybrid credit quality	$\hat{\phi}_i^t$	$\hat{\phi}_i^t = P(O_1^t, \ldots, O_n^t \vert \hat{\lambda})$
Default threshold	$\bar{\phi}^t$	$\bar{\phi}_i^t = P(\bar{O}_1^t, \ldots, \bar{O}_n^t \vert \bar{\lambda})$
Delinquent	$\tilde{\phi}_t$	$\tilde{\phi}_t = 0.85\bar{\phi}_t$
Credit Quality Level	$\dot{\phi}$	$\dot{\phi}_i^t =$ [Poor, Average, Good, Excellent]

We observed that each agent has a unique tuple represented by $\lambda_i,\ i = 1, 2, \ldots, n$ as shown in table 5.6. The next section expounds more on each of the measures for the consumer credit scoring in this study

5.9.1 Credit quality scores

The credit quality scores of the obligors is estimated by the individual agent matrices A and B, and the state and observation sequence. The credit scores are based on the local parameters in table 5.5. We express them as

$$\lambda = (A, B, \pi, O, Q), \text{ and} \tag{5.34}$$

$$\phi_n^t = P(O_1, \ldots, O_L \vert \lambda) \tag{5.35}$$

Where L is the length or size of the observation sequence for each agent; and $0 < \phi < 1$ the interval of the CQS for each agent at each time period. This is made easy by use of the Matlab inbuilt function

$$[\phi] = \text{hmmdecode}(O_j, A_i, B_i),\ i = 1, \ldots, N,\ j = 1, 2, \ldots, L \tag{5.36}$$

The individual agent has a transition matrix, observation matrix and the observation sequence that emits a score for each obligor. The parameter ϕ_i^t, the credit quality, is dynamic and changes at every time t for each agent which is an indicator of the credit quality level of the agent. It can remain constant, increase or decrease. To fully cater for the dynamics, the dynamic default threshold is used to estimate the credit quality level as being poor, average, good or excellent (the PAGE credit quality levels of the obligors in the loan portfolio).

5.9.2 Credit quality level

The obligors credit quality scores classifies the obligors into four credit quality levels. Matlab function hmmdecode emits an observation which is the CQL of the agent and is estimated with:

$$[\bar{O}_i] = n\text{hmmdecode}(O_j, A_i, B_i),\ i = 1, \ldots, N,\ j = 1, \ldots, L \tag{5.37}$$

The value $\bar{O}_i$ = [P, A, G, E] $= [1, 2, 3, 4]$ is emitted by HMM and that is what classifies the obligor into a specific CQL. The symbols are represented in table 5.7

Table 5.7: The symbols for the key obligor credit quality levels

Variable	Levels	Poor	Average	Good	Excellent	(PAGE)
Credit quality	Symbol	ν_1	ν_2	ν_3	ν_4	
	Points	1	2	3	4	

The data in this table 5.7 has the summary of the different credit quality levels of the obligors.

5.9.3 Hybrid credit quality

This is estimated using the same model parameters for credit quality and some from the default threshold as shown in 5.5. The purpose is to test how the individual agent observations and state sequence performs when compared to the threshold matrices $\bar{A},\ \bar{B}$ and $\bar{\pi}$ against the individual agent observation and state sequence. In table 5.5, we use the three global and two local model training parameters, that is,

$$\hat{\lambda} = (\bar{A}, \bar{B}, \bar{\pi}, \bar{O}, \bar{Q}),\ \text{and} \tag{5.38}$$

$$\hat{\phi} = P(\bar{O}_1, \ldots, \bar{O}_N | \hat{\lambda}) \tag{5.39}$$

The Matlab function is

$$\hat{\phi} = \text{hmmdecode}(O, \bar{A}, \bar{B}) \tag{5.40}$$

5.9.4 Dynamic threshold

A dynamic threshold classifies the obligors in the respective credit quality levels, and as defaulters and non defaulters. The dynamic threshold depends on the average obligor transition matrix and individual agent observation value at any given time period. The estimations are based on

$$\bar{\lambda} = (\bar{A}, \bar{B}, \bar{\pi}, O, Q), \text{ and} \tag{5.41}$$

$$\bar{\phi} = P(\bar{O}_1, \ldots, \bar{O}_N | \bar{\lambda}) \tag{5.42}$$

The matlab inbuilt function for this estimation is:

$$\bar{\phi} = \text{hmmdecode}(\bar{O}, \bar{A}, \bar{B}) \tag{5.43}$$

A dynamic default threshold value is necessary to help detect the obligors below or above the threshold for classification purposes and detect default rates. We expect the threshold to be a good indicator of the credit quality level of the obligors not to increase the number of false rates in the model but perform as expected.

5.9.5 Model false rates

The estimated numbers of the false positive and false negative are indicators of the quality of our model in estimating the default rates. We use ϕ_i^t, $\bar{\phi}^t$ and $\hat{\phi}_i^t$ which are the credit quality, default threshold and the hybrid credit quality respectively to estimate the false rates in the model. Four rates to be estimated are as shown in table 5.8

1. False Positive - an obligor is estimated by the model as above the threshold level but in actual sense, it is below the threshold level. A probability of accepting a ‘bad’ obligor.

2. False Negative - an obligor is estimated by the model as below the threshold level but in actual sense it is above the threshold level. A probability of rejecting a ‘good’ obligor.

Table 5.8: The model quality rates and their estimation

Ratings	Points	Symbol	Estimation
Positive Rates	2	g	If $\phi_i^t \geq \bar{\phi}^t$ & $\hat{\phi}_i^t \geq \bar{\phi}^t$
False Negative	1	$\dot{g}$	If $\phi_i^t \leq \bar{\phi}^t$ & $\hat{\phi}_i^t \geq \bar{\phi}^t$
False positive	-1	$\ddot{g}$	If $\phi_i^t \geq \bar{\phi}^t$ & $\hat{\phi}_i^t \leq \bar{\phi}^t$
Negative Rates	-2	$\hat{g}$	If $\phi_i^t < \bar{\phi}^t$ & $\hat{\phi}_i^t < \bar{\phi}^t$

We have $G = \{g, \dot{g}, \ddot{g}, \hat{g}\}$ which is the matrix for the rates in the model. Our model should have very low false positive rates and false negative rates. The use of credit quality and hybrid credit quality to estimate these rates enhances the quality of the results in computing the false rates. For positive rates, both the credit quality and the hybrid credit quality are above the default threshold. In the false negative cases, credit quality is less than the default threshold but hybrid credit quality is above the threshold. For the false positives, credit quality is greater than the default threshold but hybrid credit quality is less than the default threshold. For negative, both credit quality and hybrid credit quality are below the default threshold.

5.9.6 Default and survival rates

The algorithm for the false rates in table 5.8 is combined with the algorithm in table 5.9 to aid in estimation of the defaults, survival, delinquent cases, stopping time and credit quality levels. All the analysis for the obligors in this last section of our model are well outlined in table 5.9 where the instances when an obligor experiences a certain situation in the loan portfolio is given. A summary of the symbols for the default rates, survival, delinquent cases, stopping times and the credit quality levels are given in table 5.10.

5.9.7 Stopping time

The stopping time instances in the loan portfolio for the obligors is summarized in table 5.9 using the credit quality and the model quality estimation.

We track the number of optimal and sub-optimal stopping time and when they occur in the life of the loan. These sub-optimal stopping time are also the default rate detection parameters of the model. We know that if the obligor does not default, then, $P(t = T) =$

Table 5.9: Estimation summary for the key obligor credit and default variables

	Estimator of an event	Default	Delinquent	Survival	Stopping time
If	$G = g \ \& \ \dot{\phi} \geq \bar{\phi}$	0	0	1	0
	$G = \dot{g} \ \& \ \dot{\phi} \geq \bar{\phi}$	0	0	1	0
	$G = \dot{g} \ \& \ \dot{\phi} \geq \tilde{\phi} \ \& \ \dot{\phi} \leq \bar{\phi}$	0	1	1	0
	$G = \dot{g} \ \& \ \dot{\phi} < \tilde{\phi}$	1	0	0	1
	$G = \ddot{g} \ \& \ \dot{\phi} \geq \bar{\phi}$	0	0	1	0
	$G = \ddot{g} \ \& \ \dot{\phi} \geq \tilde{\phi} \ \& \ \dot{\phi} \leq \bar{\phi}$	0	1	1	0
	$G = \ddot{g} \ \& \ \dot{\phi} < \tilde{\phi}$	1	0	0	1
	$G = \hat{g} \ \& \ \dot{\phi} \geq \tilde{\phi} \ \& \ \dot{\phi} \leq \bar{\phi}$	0	1	1	0
	$G = \hat{g} \ \& \ \dot{\phi} < \tilde{\phi}$	1	0	0	1

We have combined the model quality, default rates, the survival and delinquent rates, stopping time (both optimal and non-optimal) and the credit quality levels of the obligors. They are estimated using the credit quality, hybrid credit quality and the average of both the credit quality and the hybrid credit quality. Definitions of the notations are in table 5.6

Table 5.10: The symbols for the key obligor credit and default variables

Parameter	Default rates	Survival	Stopping time	Delinquent
Symbol	$\mathcal{D}$	$\mathcal{S}$	$\dot{t}$	$\mathbb{D}$
Outcome	0 or 1	0 or 1	0 or 1	0 or 1

The outcome 0 indicates that the event has not occurred while 1 is an indicator of the occurrence of the specific event. The mathematical formulations for these outcomes are highlighted in table 5.9.

1. Let $\dot{t}$ denote the sub-optimal stopping time. The default rates are random and thus $\dot{t}$ is also random. We are more concerned with sub-optimal stopping time, $\dot{t}$ in our loan portfolio as it is an indicator of the default points in the life of the loan.

5.10 Accuracy and Sensitivity Analysis

We test the model performance using the sensitivity analysis by varying the input variables and comparing them with output variables. Rags (2001) observes that sensitivity analysis can be derived from the simulated sample to quantify the influence of the inputs and identify the key contributors. The main assumption being that the input and output of the model varies in a monotonic and linear manner. We estimate sensitivity through use of Pearson correlation coefficient and coefficient of determination to compare how the changes in the input variable influences the output variable. The false rates estimates the accuracy of the model.

5.10.1 Sensitivity analysis

Sensitivity analysis is the process of variation in output of a model with respect to changes in the values of the model's input (s) (Rags , 2001). The purpose is to provide a ranking of the model inputs based on their relative contributions to model output variability and uncertainty. Sensitivity analysis in simulation models estimates the change in the simulation output as the simulation input changes, either for discrete event, continuous or hybrid models (Kleijnen , 2009). In real life experimentation, it is hard to vary a factor over many values, but in simulation experiments this restriction does not apply.

As simulation continues to receive increasingly use in solving problems and to aid in decision making, there is a need to continue to develop models that are accurate and give results that are "correct" (Sargent , 2011). A model is valid for a set of experimental conditions if the model's accuracy is within its acceptable range, that is, the amount of accuracy required for the model's intended purpose (Sargent , 2011). The commonly used metrics to test for sensitivity are the Pearson correlation coefficient, sensitivity ratio (also called elasticity) and sensitivity score among other metrics (Rags , 2001). The sensitivity ration (S.R) is expressed as

$$\hat{S} = \frac{\Delta\text{Output} \times 100}{\Delta\text{Input} \times 100} \tag{5.44}$$

The interpretation of the sensitivity analysis is based on;

$$SR = \begin{cases} \hat{S} < 1, & \text{if sensitivity is low;} \\ \hat{S} = 1, & \text{if sensitivity is zero;} \\ \hat{S} > 1, & \text{if sensitivity is high;} \end{cases} \tag{5.45}$$

Sensitivity equal to zero indicates that the model output does not change with changes in the input. If $\hat{S} > 1$, a small change in input reflects a higher change in model output. If $\hat{S} < 1$, a small change in model input leads to an even lower change in model output.

5.10.2 Accuracy

According to Loso and Koski (2014), the measure used to test for the performance of the model is through its accuracy. The term accuracy is a measurement of how often the model predicts the correct value. The accuracy is calculated as

$$\dot{S} = \frac{\text{Number of true predicted states}}{\text{Total number of tested objects}} \times 100 \tag{5.46}$$

The study based its model performance on $\dot{S} \in [0, 100]$. This metric system was developed in this study to measure the model accuracy as an indicator of the model's performance;

$$\dot{S} = \begin{cases} < 25 & \text{Low accuracy;} \\ 25 \text{ to } 50 & \text{Medium accuracy;} \\ 51 \text{ to } 75 & \text{Good accuracy;} \\ > 75 & \text{High accuracy;} \end{cases} \tag{5.47}$$

The accuracy levels are important to measure how the model fairs in terms of estimating the false rates. High false positive rates means that the model accepts 'bad' agents as good and this increases the likely default rate in the loan portfolio. If the model has high false negatives, it means 'good' agents are rejected as bad and this high default rates though in actual sense it would be lower. Therefore, any increase in false rates is not ideal for the model, but decrease in false rates is what would be expected in a good model.

5.10.3 Conclusion

The discussions, numerical techniques and models in this chapter have outlined the main methodology procedures. We have five levels: network initial conditions; the SEN dynamics that captures the agents cyclical inter dependencies and changes in the social and economic factors; the CRAFs which is a set of the five social factors, five economic factors and two credit quality factors; the HMM applied the CRAFs for learning and training the model which in turn emits the CQS, CQL and HCQS using the dynamic threshold; and the CQS and default rates dynamics. The last part of the methodology discussed how to estimate the model performance through use of accuracy rate and sensitivity analysis.

Chapter 6

Credit Scoring with Social Network Data

This chapter undertakes the data analysis process based on the methodology in chapter 5 and using the Matlab software inbuilt functions appendix A.2. The analysis is presented in form of charts, graphs, tables and descriptive statistics from the SEN model, the HMM emissions and credit quality analysis. We have used different random generator seed in this analysis to depict the varied scenarios when using simulation and in the analysis using this type of model.

6.1 Network Initial Conditions Analysis

The initial conditions and the social network analysis are presented in this section. We have the initial conditions for the reputation ratings, age, relationships, private data and the frequency or number of encounters in the network. In table 5.1, we have the different time cycles at the different stages in the model. At time $t = 0$, we have $\tau = 0$ which indicates the initial conditions of the model. At time $t = 0$ and with $\tau = 1, 2, 3$, we have the three interactions undertaken by the agents which indicates the initial model state. These interactions are taking place before the agents approach the financial institution for a loan. We assume that these events take place six months before the loan period is effective. The assumption offers us a rich set of data for HMM training

We recall that the computations for the credit quality and other analysis are computed after a period of two months.

6.1.1 Reputation ratings

The initial condition matrix is depicted in the first plot at $\tau = 0$ in figure 6.1 and the other three plots $\tau = 1, 2, 3$ show the changes observed at time $t = 0$. Evidently, changes in reputation ratings are observable at each instance in time, $\tau = 0, 1, 2, 3$ from the different patterns of the plots. such changes are expected due to the stochastic nature of the SEN. The main diagonal entries of the reputation rating matrix are equal to five. We observe that the diagram of the plots does not show any changes due to this set condition for all $\tilde{R}_t,\ t \in [0, T]$. We note that $1 \leq \tilde{R}_t \leq 5$ for $t \in [0, T]$ with $\tilde{r}_{ii} = 5$ for all $i = 1, 2, \ldots, N$.

The initial reputation ratings become more sharply divided in each time period, τ. These changes are due to the initial state of the network as agents learn each others reputation.

6.1.2 Relationships

The changes in the relationship coefficients of the agents as part of the initial conditions are depicted in figure 6.2. The dynamics are evident on how agents changed their relationships levels and value to each other as observed in the figure with ten agents in the network. We note that $0 < \dot{R} \leq 1$ and the agents exhibit different levels of relationships during this time period. A close look at the figures in figure 6.2 shows changes at each time period.

The plots shows the discernible changes in the agents relationships along the time period $t = 0$ for different τ values. The increases and decreases over time are noticeable due to the SEN dynamics as expected.

6.1.3 Frequency of interactions

The frequency of encounters is how often an agent interacts with another agent at any given time interval, τ. Figure 6.3 shows the frequency of encounters of the agents at time $t = 0$ which are the initial conditions of the agents. Changes in the number of these encounters are observable with both an increase and decrease in the frequency of the encounters. The x-axis shows the number of agents in the network while the y-axis has the frequency of encounters of each agent respective to the other network agents. Variations in the frequency of interactions between the agents is observable from figure

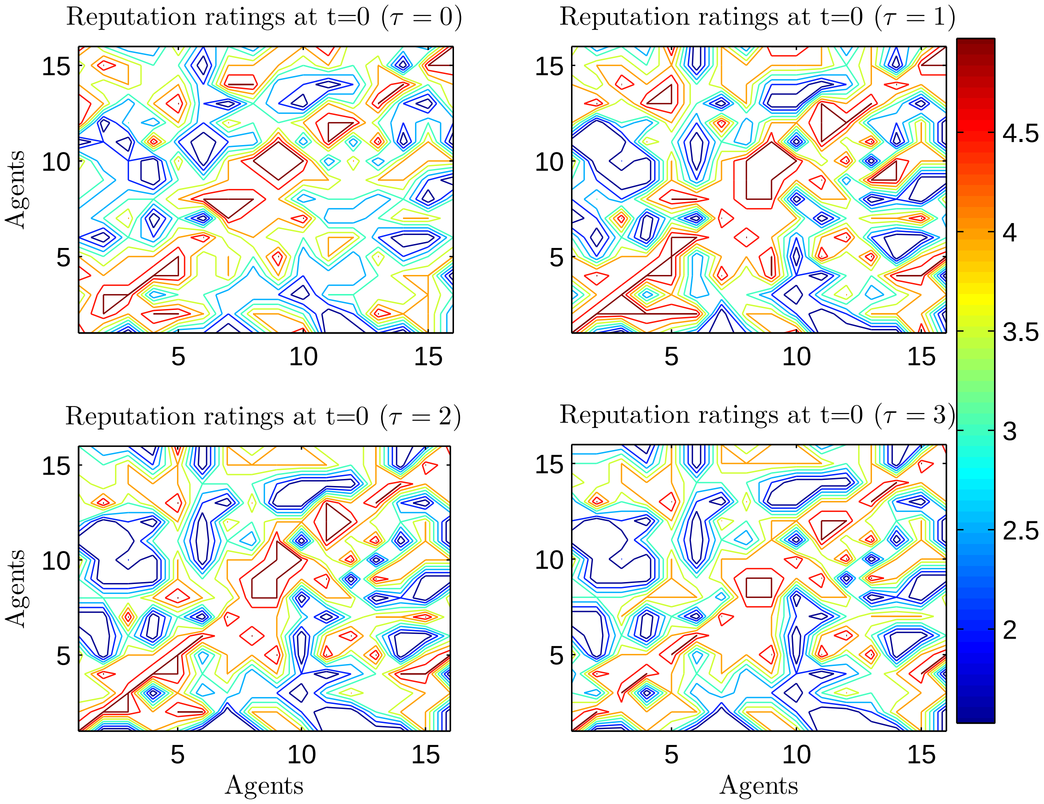

Figure 6.1: The reputation ratings of the agents at $t = 0$

The diagram shows the dynamics observed in the reputation ratings of the agents at time $t = 0$ with 15 agents in the network. This is the period before the agents become part of the loan portfolio. Changes in the reputation ratings are observable in each plot where the minor diagonal do not change as this is the individual ratings of $\tilde{R} = 5$. Some agents have reputation ratings of as low as $\tilde{R} = 1$ to the maximum possible of $\tilde{R} = 5$.

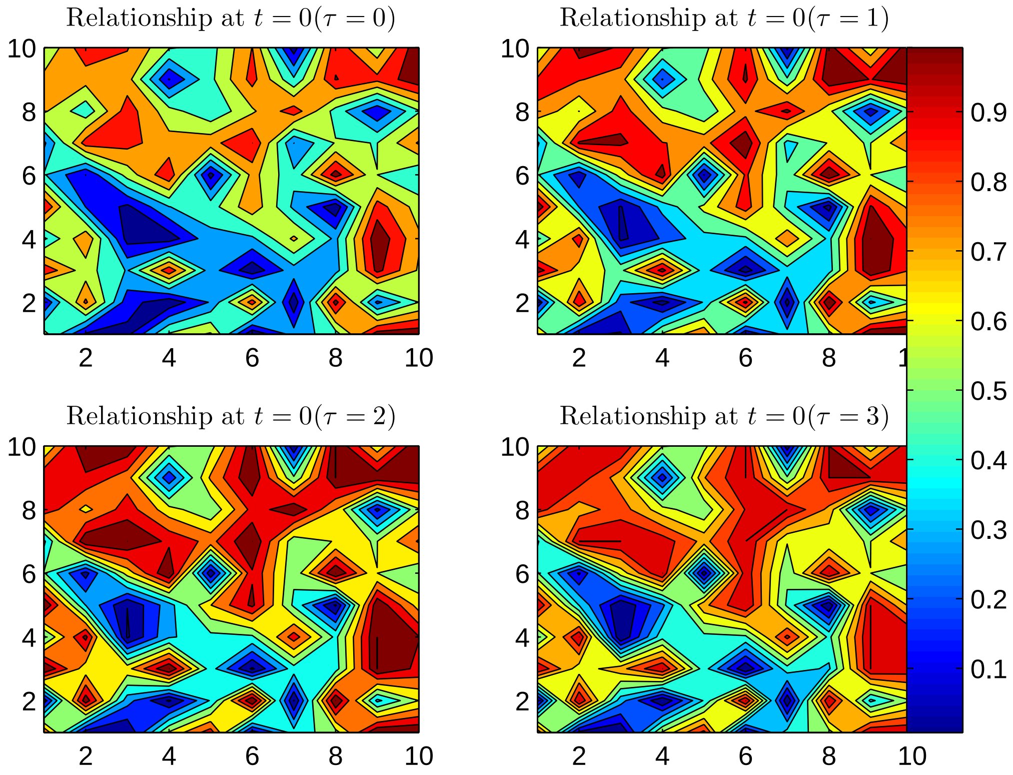

Figure 6.2: The relationship matrix at $t = 0$ of the agents.

Evidently, agents changed the relationship levels at each time period τ. The dynamics in the relationships of the agents shown in the plots indicates that the SEN induced dynamic relationships. However, we observe that the dynamics were not extreme but we can say they were in acceptable range though that was not measured. Relationship levels are vital to keep the agents interactions active and form part of the cyclical inter dependencies being modeled in the social and economic network.

6.3 at the indicated different time periods.

The frequency at which agents interact also changes over time. Figure 6.3 shows the dynamics evident with agents interactions in the network due to change in the amount and flow of information of each other.

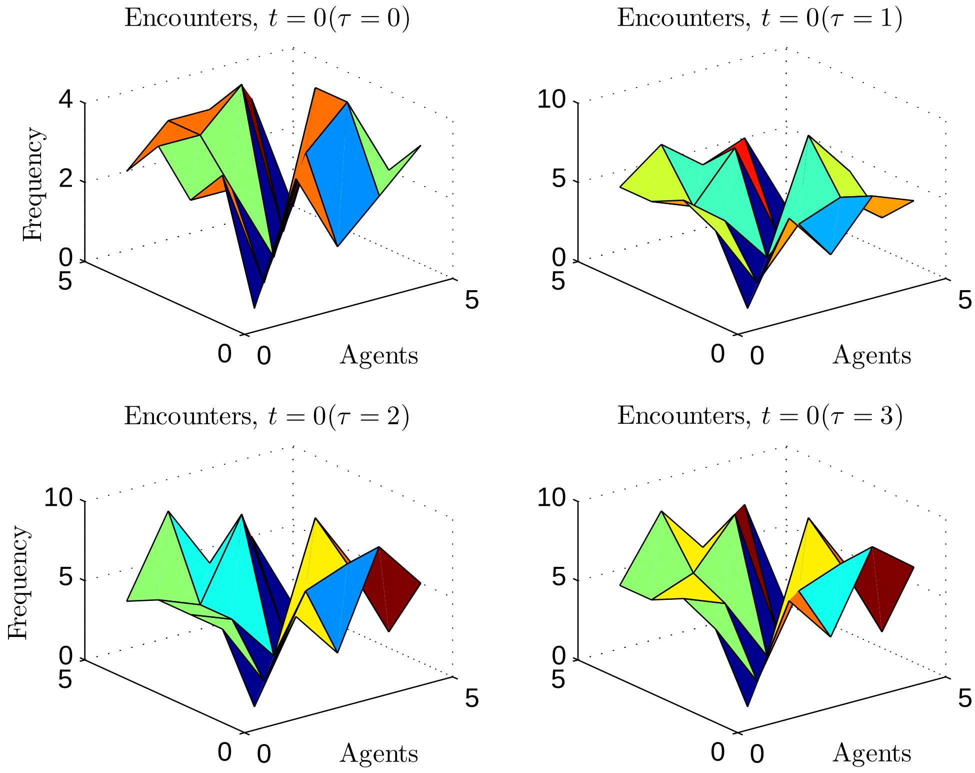

Figure 6.3: Frequency of encounters of the agents in the network

We have set the number of encounters $1 \leq \mathfrak{N} \leq 8$ for all time period $t \in [0, T]$. We have assumed a maximum of four encounters per month and since our time t is two months, then $\mathfrak{N} \leq 8$. The plots shows the combination of the number of agents, frequency of interactions and changes in time period. We observe changes in the patterns of the plots that clearly indicates that the agents have changing levels of interaction together with the frequency of interaction as with changes in time. The time period is at $t = 0$ with different interaction times, $\tau = 0, 1, 2, 3$.

6.1.4 Private data

Figure 6.4 shows the changes in the private data of the agents at time $t = 0$ with $\tau = 0, 1, 2, 3$ where $t = 0$ and $\tau = 0$ is the initial condition with 20 agents in the network.

There are no major changes observable in the private data as it changes slowly. We noted that the private data is a function of the trust and distrust levels in the network and the accumulated private data at the previous period of time. Major spikes in private data are kept in check by use of the SEN ratings of trust and distrust levels as the private data forms a key component in this model.

The private data component of the SEN-HMM-CSD model shows high levels of stability in terms of dynamics of the accumulation or loss of the factors. In general, economic factors do not exhibit stochastic behaviour, but has an element of deterministic behaviour.

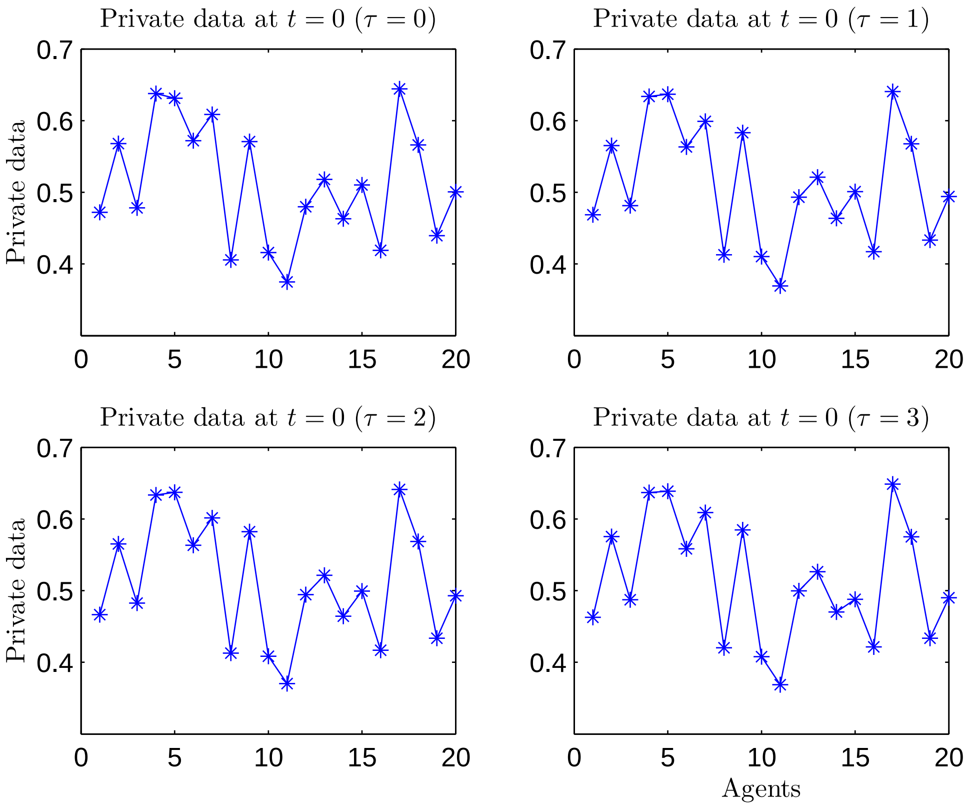

Figure 6.4: The initial private data of the agents at time $t = 0$

The twenty agents show minimal changes in the private data during this time period. The changes in private data observable among the agents. Looking at the plots do not seem to show any observable changes but in actual sense, the plots exhibit changes. We remember that private data cannot change very frequently as is the case with social factors. Again, the plots are for one period only, $t = 0$ with different interactions at $\tau = 0, \ldots, 3$.

6.1.5 Age of the agents

The set of data for the agents age forms the demographic component of the SEN. Age increament is observed after every twelve months. The mean for this group is 44.3 years with a standard deviation of 12.74 and skewness of -0.211. The majority of the agents are aged between 23 years and 45 years. Age was the only social-demographic variable used in this study.

6.2 Network Dynamics Analysis

The SEN dynamics are observable from the different strategies and interactions agents makes with each other. These dynamics are in reputation changes, relationships levels, number of times the agents interact with each other, the interaction experiences and the feedback from all the agents and individual agents. We analyze the dynamics from $t \in [1, T]$ with varying number of agents in the network for the different variables in the SEN using the available different plots and graphs to give us a blend and mix of the possible outcomes in the analysis.

We compare the relationship intensity, encounter experiences and the reputation ratings of 8 agents in the network at time $t = 6$ as depicted in figure 6.5. Obligors with high reputation ratings, and high relationship intensity had high encounter experiences and vice versa. The colorbar in figure 6.5 represents the reputation ratings of the obligors. This highlights the link between these variables in the SEN.

Figure 6.5 shows the link between interaction experiences, relationship levels and the reputation ratings of eight obligors at $t = 6$. Dynamics are noted with obligors having high levels of relationship intensity and reputation ratings showing high levels of interaction experiences. The dynamics are not showing a causation between the relationship levels, encounter experiences and the reputation ratings. Each of these variables has either a direct or indirect link with the other variables for the obligors in the network.

6.2.1 Reputation ratings

Figure 6.6 shows the agents when they are issued with loans by the financial institution and the changes in the reputation ratings are tracked for a period of eight months, that is, $\hat{t} = 8$. This translates to four periods, that is, $t = 4$ because we have a time interval of

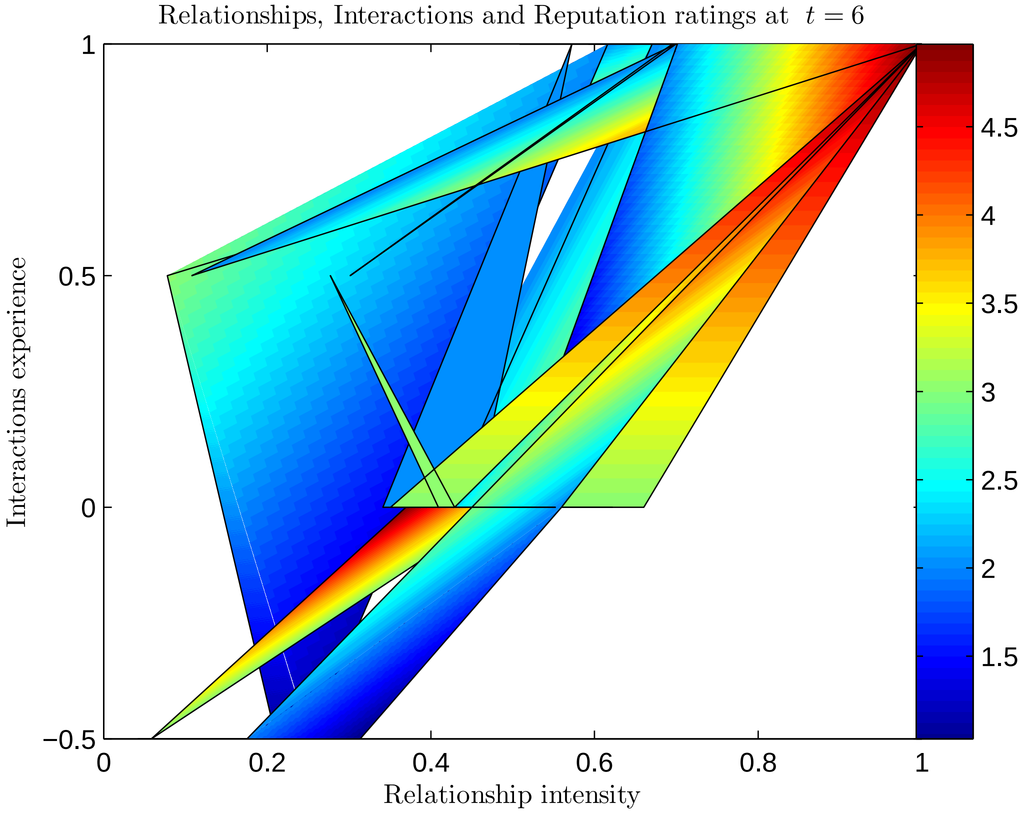

Figure 6.5: Relationships, interaction experiences and the reputation ratings

The interdependency of relationships, interaction experiences and reputation ratings are depicted in the plot. The colorbar represents the reputation ratings of the agents and are compared to the two other variables. High levels of reputation have high levels of interaction experiences as well as good relationship levels amongst the agents. This is within the expectations of the SEN dynamics due to the inter dependencies of the agents.

2 months for the computation of the reputation ratings and other variables in the model. Figure 6.6 depicts the dynamics observed in the reputation rating changes in those eight months. Clearly, dynamics in the reputation ratings are observed even after the loan is issued by the financial institution. The SEN offers dynamics of the changes in the reputation ratings of the peers in the network as shown in figure 6.1 and figure 6.6. These ratings enables us to gain insights on the behavior of the agents at each time period.

Reputation ratings are key part of the SEN-HMM-CSD model as they are used to estimate trust, distrust, SEN risk and in the computation of the private data.

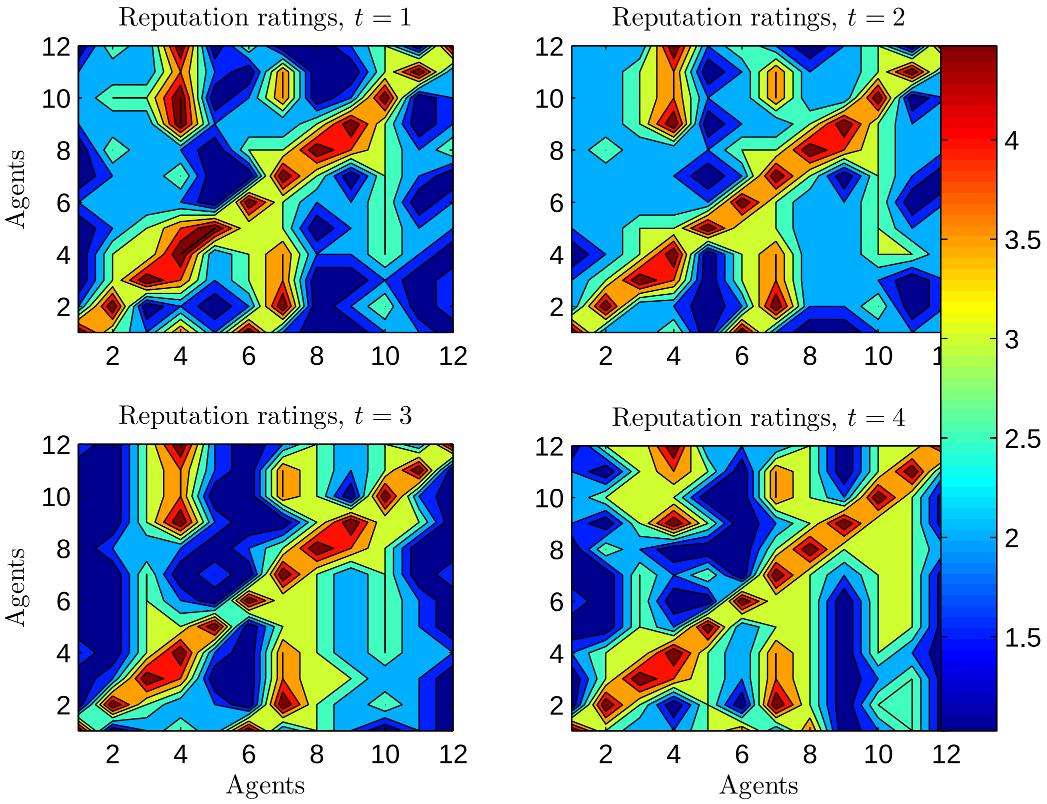

Figure 6.6: Reputation ratings at $t = 1, 2, 3, 4$ with 12 obligors in the loan portfolio

The patterns in the diagonal observed in these plots are similar to those in figure 6.1 which remains constant at $\tilde{r}_{ii} = 5,\ \forall i$. The changes in time led to changes in reputation ratings of the agents in the network as is depicted in the plots.

6.2.2 Agents relationships

The figure 6.7 shows the changes in the relationships levels after the loan was issued by the financial institution for the period of the first eight months at $t = 1, 2, 3, 4$. Marked changes in the strength of the relationships and the SEN dynamics in those relationships are noticeable. The evolution of the relationships are made independently at each time period by each agent. Agents choices are rational and change the relationship levels to maximize gains from the network interactions and links. We noted that relationships have costs and benefits that lower or increase the outcome of the agents actions.

As noted in the plots in figure 6.7, the changes are not sharply different with changes in time periods. Relationships are crucial as they define which agent will relate to which agent. This is the connecting factor which varies with experiences and outcomes of the recent relationships.

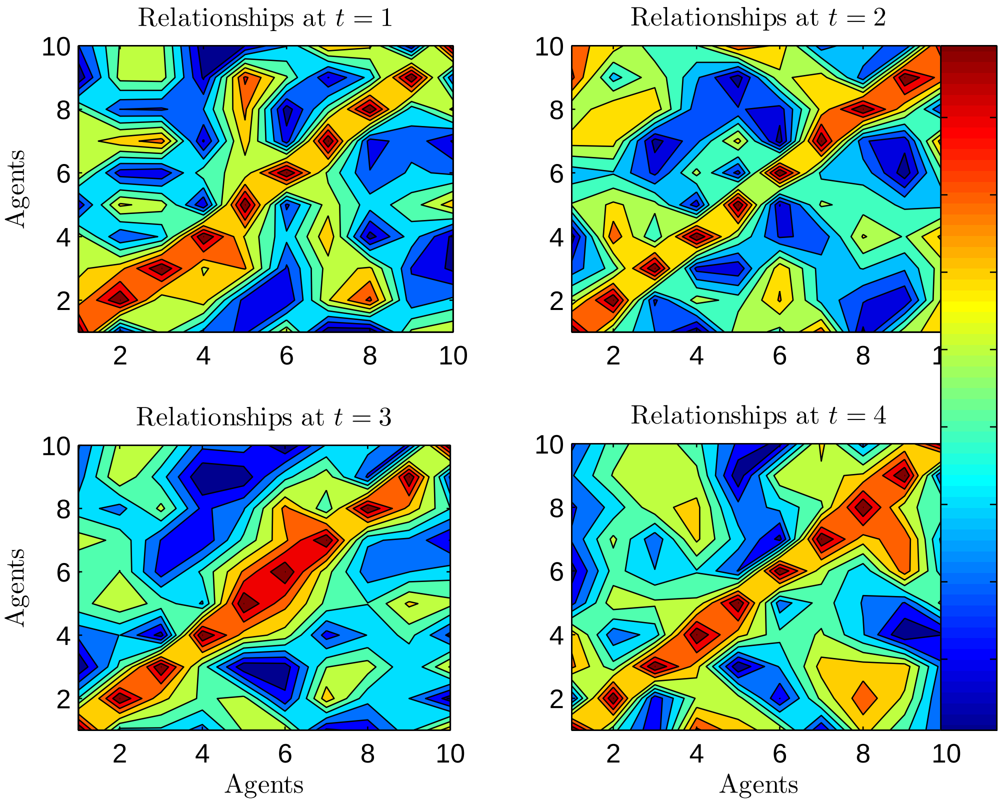

Figure 6.7: The relationship matrix during the life of the loan.

Relationship levels at $t = 1, 2, 3$ with ten obligors in the loan portfolio. The subtle changes are noted in the four plots indicating that relationship changes are dynamic with changes in time periods. This is the expectations we have observed from the agents in the network.

6.2.3 Interaction experiences

The interaction or encounter experiences are possibly enhanced by the different characteristics of the agents and the uncertainties there in. Interaction information enables the agents to decide on how frequently to interact with another agent, the relationship level and the reputation rating to award the respective agent. We therefore note that the flow in the SEN is cyclical as one parameter has a link with another as well as one agent with another. Figure 6.8 shows the encounter experiences of the agents at time $t = 1, 2, 3, 4$ and how the dynamics of these experiences for the ten agents in the SEN changes over a period of time.

The variations in the plots indicates that agents experienced varied interaction satisfaction with changes in time. Therefore, agents interaction experiences forms part of the main variables in the SEN dynamics.

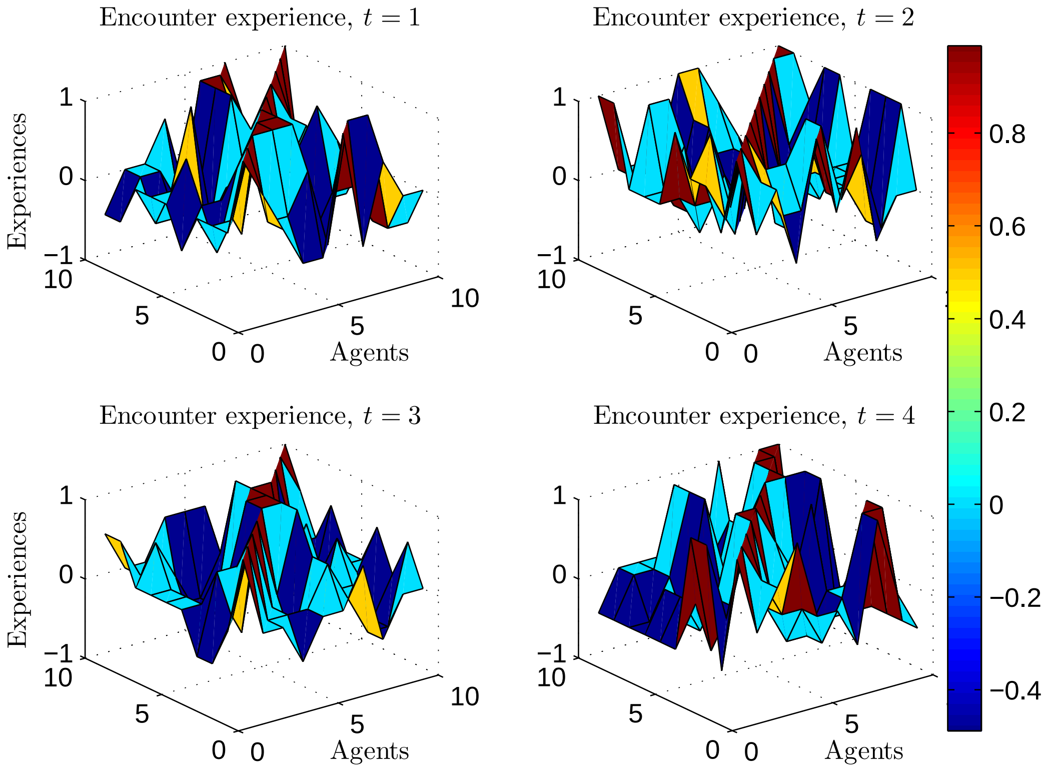

Figure 6.8: The encounter experience matrix during the period of loan obligations Encounter experiences at $t = 1, 2, 3$ with ten agents in the network during the life of the loan obligations with the financial institution.

6.2.4 Reputation feedback

Figure 6.9 is the contour plots for the reputation feedback at time $t = 1, 2, 3, 4$ with 14 agents in the network. The feedback mechanism is to ensure that the different parameters offers us the general status of an agent. Social network model shows the desirable and undesirable outcomes of the agents and the feedback from the agents reputation ratings.

Feedback systems in a network are important as they show how the agents interact, react and rate each other in the networking process. Changes are observable in the plots in figure 6.9 at each time period. Evidently, reputation feedback is dynamic over time. A SEN characteristic is dynamism and stochastic nature from the agents interactions.

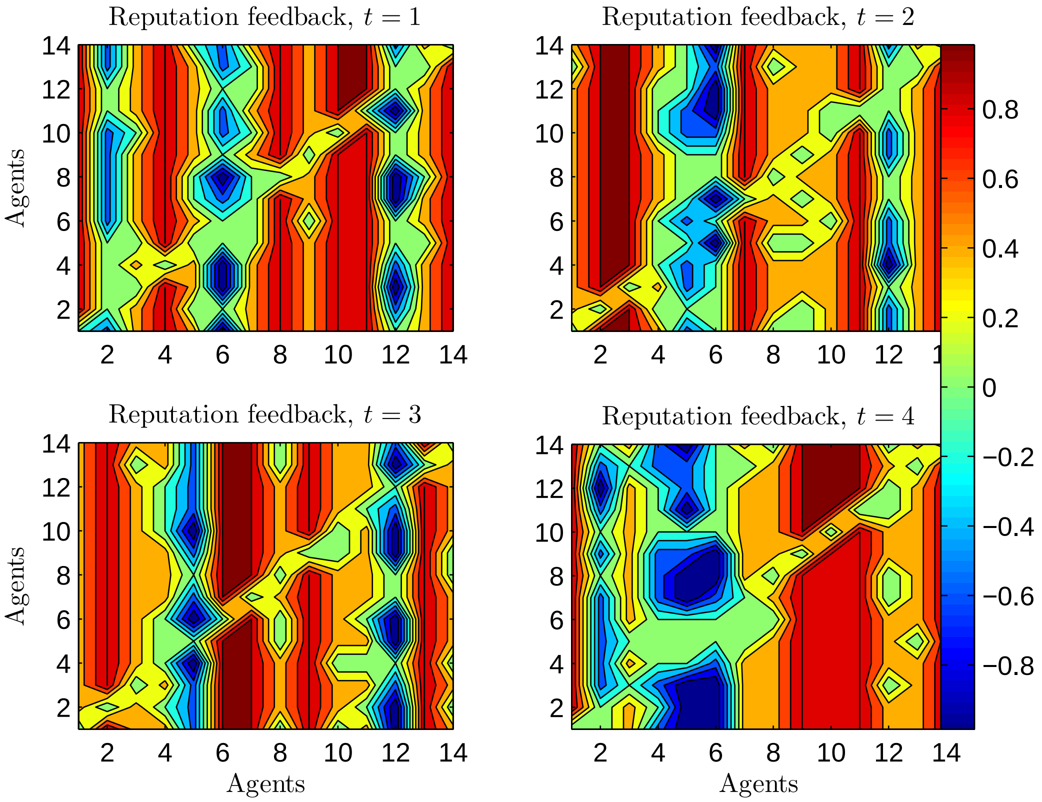

Figure 6.9: The reputation feedback of the agents after the loan issuance.

The reputation feedback of fourteen obligors in the loan portfolio at time $t = 1, 2, 3,$ and 4. As we have noted earlier, the plots exhibits the dynamics of the agents inter dependencies in the network and the social factors that changes with time.

6.3 Credit Risk Analysis Factors Data

Equation (5.29) shows the number of variables at each time period $t \in [0, T]$. This in turn eases the HMM learning and training process. The credit risk analysis factors are scaled at $(0, 1]$ to set the uniformity for ease of data clustering. The distribution of the CRAF for ten agents at time $t = 0$ is depicted in figure 6.10. The plots shows that the CRAF had varied dynamics based on individual agent. No specific distribution is observable for the CRAFs of these agents but the 7^{th} agent CRAFs show a curve that depicts the normal distribution.

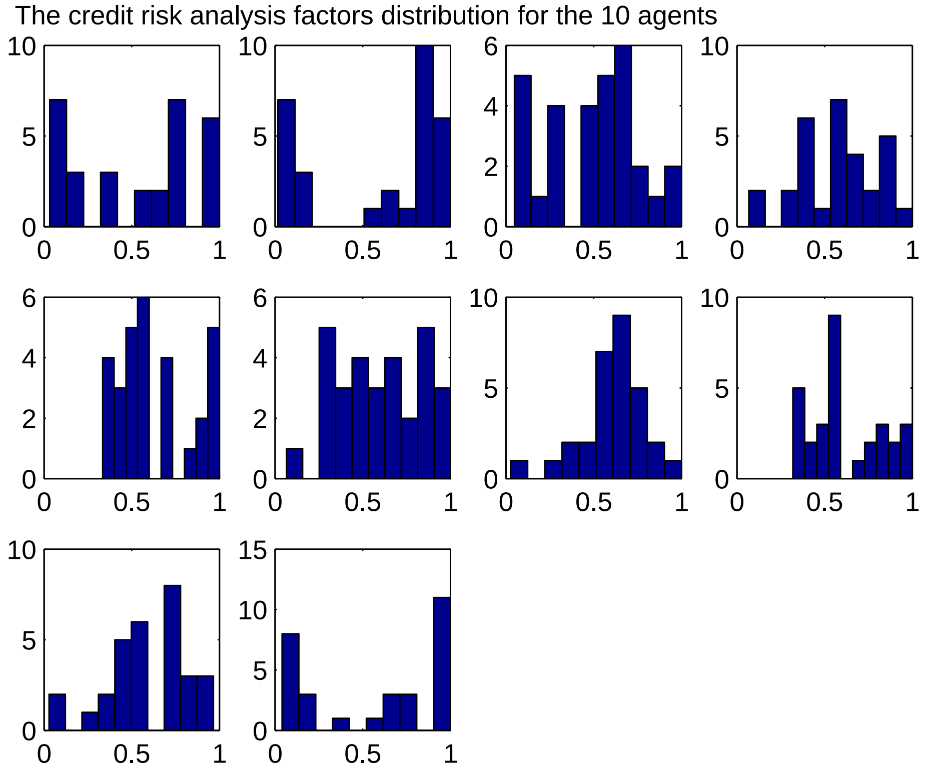

Figure 6.10: The credit risk analysis factors scaled to $(0, 1]$ for ten agents

Credit risk analysis factors are used to cluster the agents states sequence and observation sequences which are in turn used for HMM learning and training.

The table 6.1 is a summary of the descriptive statistics for the CRAF at time $t = 0$ and $t = 1$ for five agents in the network. Variability is high from the coefficients of variation (CV) and this translates to high spread of the data a fact that is supported by the dynamics expected from the network. We expect similar patterns to be observed in the credit quality dynamics of the obligors later in the chapter. The asymmetry in the CRAF data is observed from the skewness values in the two time periods. Average values of these factors is $[0.4435, 0.7112]$ for the highlighted period in table 6.1. The coefficient of variation ranges between 46.9% and 85.6% for the time period $t = 0$. For $t = 1$, coefficient of variation ranges between 36.3% and 85.1%. The second time period variations were lower than those of the first time period. This scenario could be due to initial conditions as the initial data was gathered from the agents for the first time. Therefore, in this scenario, the five agents had lower CQS and CQL at time $t = 1$ compared to time $t = 0$.

Table 6.1: Descriptive statistics for the credit risk factors

Time	Agent	Mean	Standard Error	Standard deviation	Skewness	C.V (in %)
$t = 0$	1	0.5223	0.0757	0.4147	0.1000	79.40
	2	0.7112	0.0610	0.3338	-0.7408	46.90
	3	0.5458	0.0559	0.3059	0.0276	56.00
	4	0.5500	0.0507	0.2777	0.4036	50.50
	5	0.4490	0.0702	0.3844	0.2444	85.60
$t = 1$	1	0.5224	0.0952	0.4448	-0.0094	85.10
	2	0.6371	0.0668	0.3131	-0.0949	49.10
	3	0.6936	0.0635	0.2981	-0.7308	43.00
	4	0.5648	0.0437	0.2050	0.3569	36.30
	5	0.4435	0.0787	0.3691	0.5701	83.20

Figure 6.11 shows the changes in the credit risk analysis factors at $t = 0, 1, 2, 3$. The data is from the initial conditions to the sixth month of the loan premiums being in operation. The factors are clustered to estimate the state sequences and observation sequences for HMM training process. A look at the plots shows that the factors had changes over time as is also depicted in table 6.1. The coefficient of variation (CV) measures the variations in the agents CRAFs at two different time periods.

6.3.1 Singular value decomposition analysis

The SVD analysis in the study is used to extract the social factors which form part of the credit risk analysis factors of the agents. Figure 6.12 is the plots for the SVD extraction

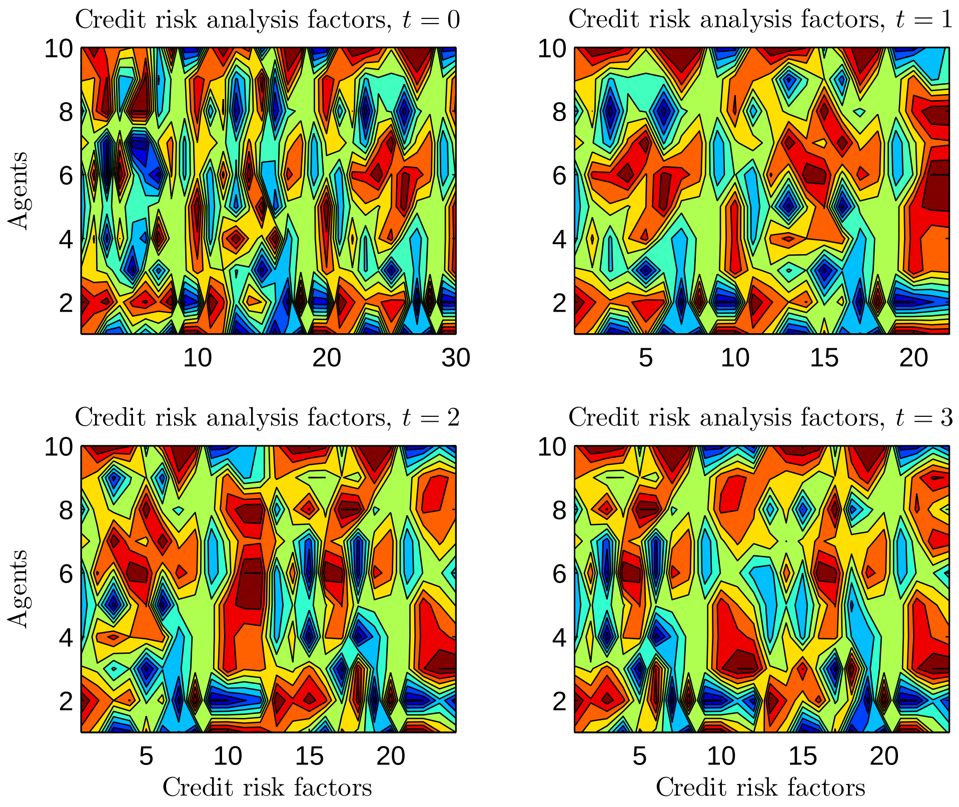

Figure 6.11: The credit risk analysis factors for hidden Markov model training. The credit risk analysis factors for ten agents at time $t = 0, 1, 2, 3$. The data used in these plots were scaled into the range $(0, 1]$ to ease HMM training and learning process. It is evident that the CRAFs changed with changes in time because of the dynamics of the agents in the SEN.

of the trust levels based on the reputation ratings of the peers in the network. The right eigenvectors and the reputation ratings are combined to estimate the trust levels in the network at time $t = 1$ with ten agents in the network. Figure 6.13 is the plots for the SVD extraction of the distrust eigenvectors and singular values based on the distrust ratings. The distrust ratings are extracted from the reputation ratings of the agents, which are based on peer to peer levels in the network. The right eigenvectors and the distrust ratings are combined to estimate the distrust levels in the network at time $t = 1$ with ten agents in the network.

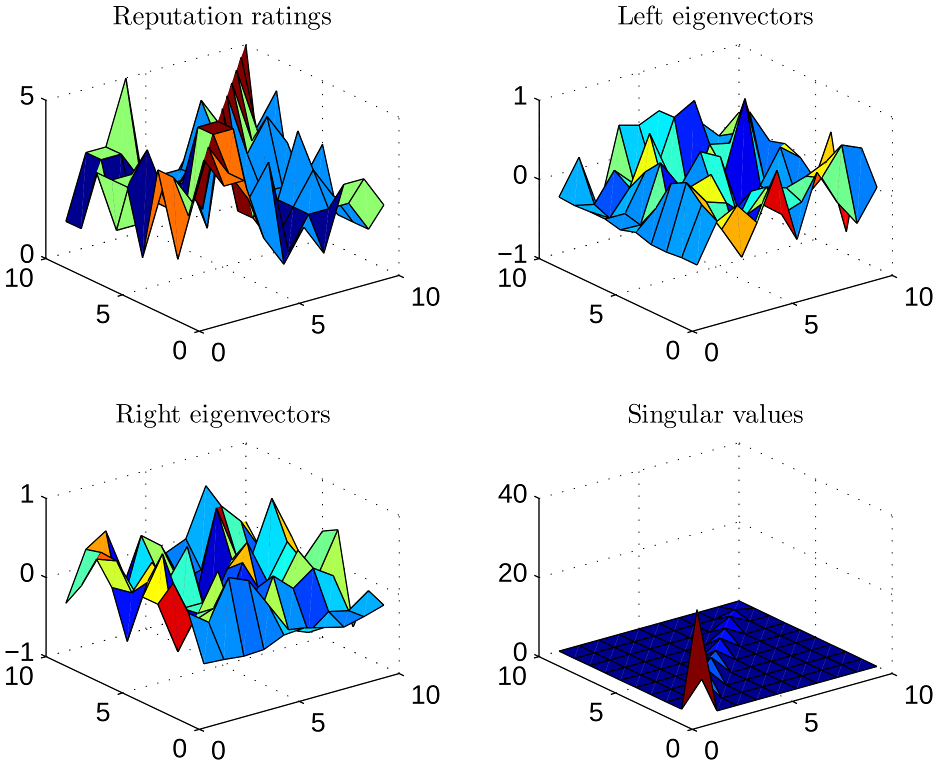

Figure 6.12: The SVD of the reputation ratings to extract the eigenvectors and singular values for the trust levels

The plots shows the reputation ratings and the corresponding extractions using the SVD. These are the right eigenvectors, the singular values, and left eigenvectors. The right eigenvectors are frequently used to estimate the trust levels as these eigenvectors are the columns of the reputation ratings matrix extractions.

We use the SVD to analyze agents by extracting the reputation ratings to estimate the trust levels of the agents. This gives rise to three matrices with the left eigenvectors,

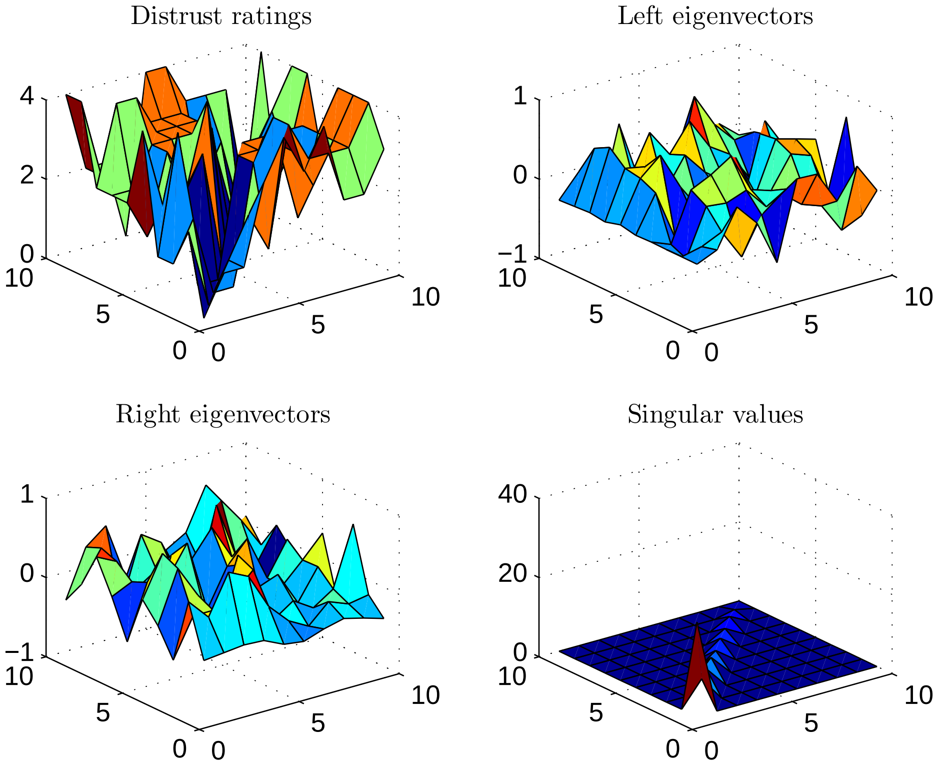

Figure 6.13: The SVD of the distrust ratings extracted from the reputation ratings for the eigenvectors and singular values for distrust levels

The distrust levels of the agents are estimated with the right eigenvectors as depicted in the third plot. Columns of the distrust ratings matrix shows the levels of perception an agent receives from the other agents.

singular values and the right eigenvectors. We have selected a reputation ratings matrix with five agents to show an example of a SVD analysis. The condition number for the system is, $\mathfrak{C}(\tilde{R}) = 9.7023$. This information is combined with information in table 6.2 to explicitly show how the reputation ratings are used to extract the trust levels, among the variables of the agents required for the CRAFs.

$$\tilde{R} = \mathbf{U}_{m\times m}\mathbf{S}_{m\times k}\mathbf{V}^{T}_{k\times k} \tag{6.1}$$

$$\begin{pmatrix} 5 & 3 & 2 & 4 & 2 \\ 2 & 5 & 2 & 2 & 1 \\ 2 & 2 & 5 & 2 & 3 \\ 2 & 2 & 3 & 5 & 2 \\ 1 & 1 & 4 & 3 & 5 \\ \end{pmatrix} = \begin{pmatrix} -0.496 & 0.481 & -0.447 & 0.549 & 0.146 \\ -0.368 & 0.505 & 0.677 & -0.291 & 0.258 \\ -0.450 & -0.323 & 0.396 & 0.368 & -0.633 \\ -0.458 & 0.027 & -0.429 & -0.691 & -0.358 \\ -0.454 & -0.640 & -0.020 & -0.032 & 0.619 \\ \end{pmatrix} \times$$

$$\begin{pmatrix} 14.160 & 0 & 0 & 0 & 0 \\ 0 & 4.898 & 0 & 0 & 0 \\ 0 & 0 & 3.056 & 0 & 0 \\ 0 & 0 & 0 & 2.011 & 0 \\ 0 & 0 & 0 & 0 & 1.459 \\ \end{pmatrix} \begin{pmatrix} -0.387 & 0.445 & -0.317 & 0.739 & -0.078 \\ -0.395 & 0.558 & 0.641 & -0.241 & 0.253 \\ -0.506 & -0.434 & 0.352 & 0.077 & -0.653 \\ -0.514 & 0.102 & -0.604 & -0.597 & -0.065 \\ -0.417 & -0.541 & 0.005 & 0.184 & 0.707 \\ \end{pmatrix}$$

We extract the trust levels of the five agents and the levels are shown in table 6.2 where the agents with high average reputation ratings had high trust levels. Even though some agents have the same average reputation ratings as in table 6.2, the trust levels differs as the SVD analysis takes into account the other matrix entries in the reputation ratings matrix in its analysis.

In figure 6.14, the graph shows the dynamics observed in the ethical factors, return on private data and changes in private data at time $t = 1$ for ten agents in the network. Obligors with high ethical factors had low return and change in private data and those with low ethical factor had high return on private data and high changes in private data at any given time period. We do not have an actual explanation as to why this is observable. But, from real life situations, agents who are unethical can sometimes make high returns

Table 6.2: Estimated trust levels for five agents using SVD

Agents	1	2	3	4	5
Trust levels	0.2439	0.546	1.0000	0.8849	0.6390
Average reputation ratings	2.4	2.6	3.2	3.2	2.6

on their investments. Though, this differs from case to case basis. An investment can be done on ethical or unethical ways and still produce the desired results of high returns.

In figure 6.14, the third obligor has very low ethical factor and high return and changes in the private data. For the second obligor, the ethical factor, return on private data and return on private data are almost the same at around 0.7 in the scale of 1.0. These dynamics mark the agents behaviour in the SEN.

6.4 HMM Output

The availability of the credit risk analysis factors propels us to the clustering stage to estimate the state sequence and the observation sequences. Supervised clustering is applied to cluster the credit risk factors into 1 and 2 for the state sequence and $1, 2, 3$ and 4 for the observation sequence. The state and observation sequences are used for HMM training to estimate the transition and observation matrices for the agents and the dynamic threshold levels.

Figure 6.15 shows the plots for the state sequence of an obligor from $t = 0$ to $t = 3$. The dynamics of the state symbols are noted in each plot and the variations in the state symbol length is observable. The state sequence form the transition matrix for each sequence using the hmmestimate inbuilt function in Matlab. The two sequences are estimated after supervised clustering was done to estimate matrices A and B. The agent has either a state sequence as one or two depending on the set of CRAF data. There is an equal chance that an agent can exhibit either a low score or high score.

Figure 6.16 shows the plots for the observation sequence of an obligor from $t = 0$ to $t = 3$. The symbols shows the changes at each time period. The data in figure 6.15 and figure 6.16 are trained to emit the transition and observation matrices for each individual agent. For example, the transition and observation matrices for an agent at time $t = 0$

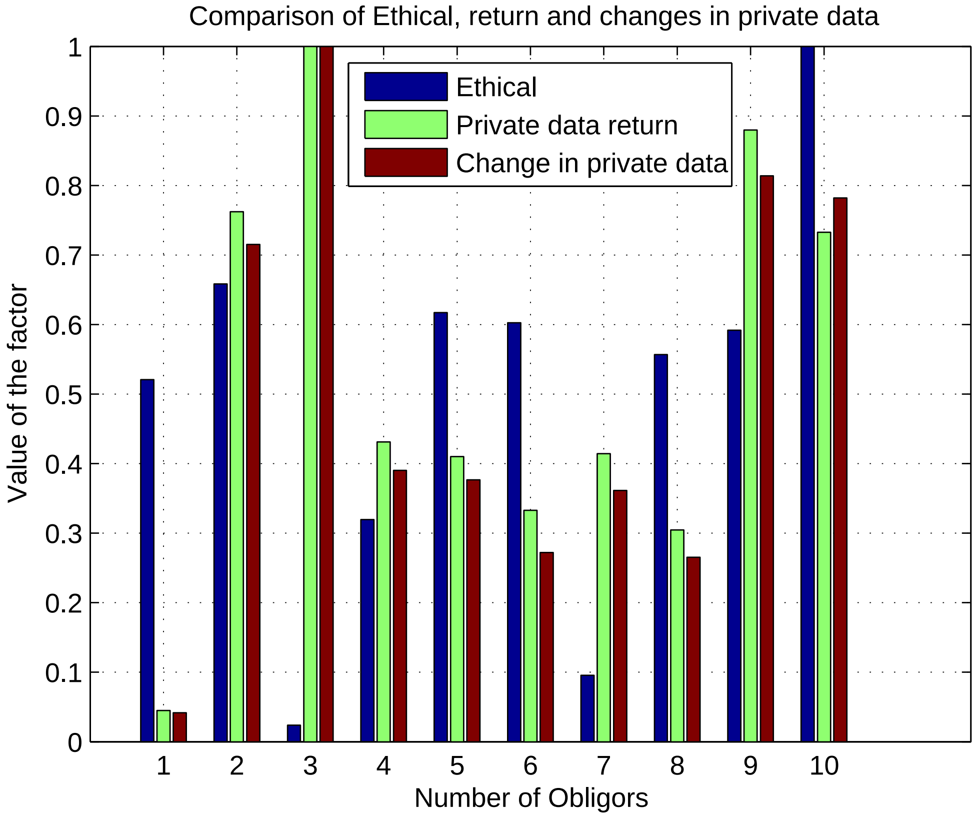

Figure 6.14: Comparison of the ethical factors, return on private data and changes in private data

The ethical factor was used in the model to capture that component of the agents that affects the private data. An agent can be ethical or non-ethical in their investment decisions. That also means they can decide to act in a manner that compromises the returns of other agents in the network.

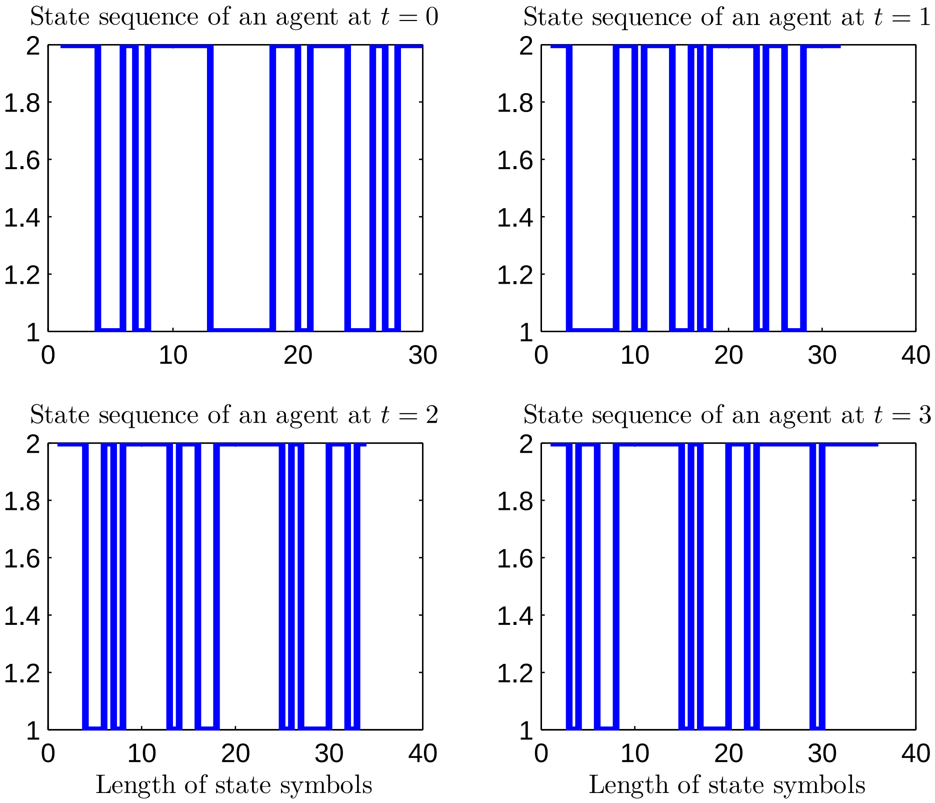

Figure 6.15: The state sequence estimated from the credit risk analysis factors

The diagrams shows the state sequences of one agent at different time periods. The sequences are either 1 for low change or 2 for high change in the transition of the states.

is respectively given as matrices $\hat{A}$ and $\hat{B}$. The observation sequence plots are also the credit quality levels of the obligors. The changes in the observation symbols are evidently observable from the obligors dynamics in the model analysis.

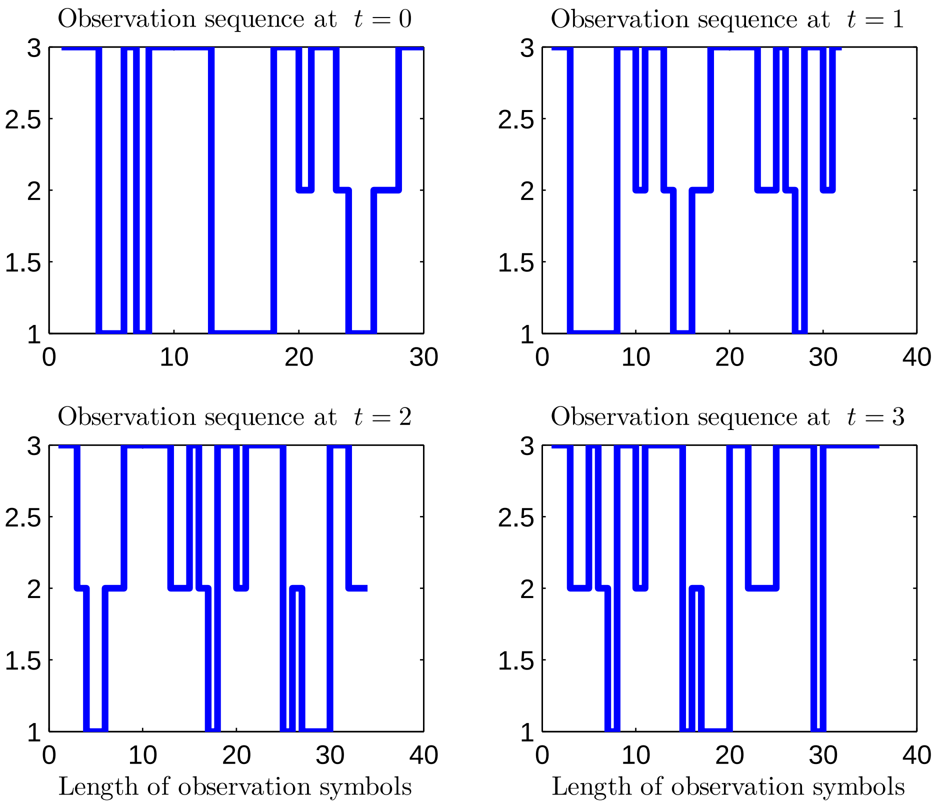

Figure 6.16: The observation sequence estimated from the credit risk analysis factors The diagram shows the observation sequences of an agent with 1 indicating a poor observation, 2 is average, 3 is good and 4 is excellent observation of the agent credit quality.

The transition matrix $\hat{A}$ is embedded with the initial state probabilities (in the first row of the matrix) for the estimation of the dynamic threshold for 12 obligors at time $t = 3$ is given as;

$$\hat{A} = \begin{pmatrix} 0 & 0.5139 & 0.4861 \\ 0 & 0.7457 & 0.2543 \\ 0 & 0.3068 & 0.6932 \\ \end{pmatrix}$$

The transition matrix is fully connected or is an ergodic model. The first row of matrix $\hat{A}$ is the initial state probability distribution. The second and third rows of $\hat{A}$ are the components of the transition state probabilities. The dynamic threshold observation matrix $\hat{B}$ is embedded with the initial state probabilities for 12 obligors at time $t = 3$, is given as;

$$\hat{B} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0.3859 & 0.2782 & 0.1701 & 0.1658 \\ 0.2064 & 0.1477 & 0.3315 & 0.3144 \\ \end{pmatrix}$$

With the estimated transition matrix $\hat{A}_n$ and the observation matrix $\hat{B}_n$ for each agent at each time period, we are now in a position to estimate the credit quality levels of the agents.

6.5 Credit Quality Analysis

The key objective of a credit rating system is to accurately and in a timely manner assess the credit risk of an obligor. Characteristics of the obligors are inferred from the dynamics of these credit quality scores and credit quality levels. We consider the analysis for the CQS, CQL, the dynamic threshold, false rates and the stopping time estimations.

6.5.1 Credit quality scores classification

The dynamics of the credit quality scores of ten agents for a period of six months from $t = 0, 1, 2,$ and 3 is depicted in figure 6.17. Agents with high or low credit quality tend to maintain the trend of the levels for the four periods estimated since entry in the financial institution loan books but with some variations noted for some of the obligors.

Figure 6.18 shows the CQS dynamics of 35 agents against the dynamic threshold level as an indicator of the barrier for the possible defaults and non defaults in the loan portfolio. This is the sixth month since the loan inception. The stochastic nature of the agents credit quality when observed against the dynamic threshold is clearly shown. The fact that an agent is below the threshold does not automatically imply a default but signals a likely default. Some of the values of the CQS of the obligors are very close to the dynamic threshold value while others are far. In fact, the third obligor has its CQS equal to the dynamic threshold value. The sixteenth obligor had the best CQS value and the first obligor had the lowest CQS. If any defaults are observed in this scenario, then the first obligor is among the defaulters.

Figure 6.19 shows the credit quality level and the hybrid credit quality levels of the obligors against the dynamic threshold at $t = 3$ for 20 agents. The different transitions in the credit quality level of each obligor indicates the level of dynamics of the model and the general observation in the credit risk modelling world. When the obligors CRAF are estimated in HMM using the threshold HMM parameters, the hybrid credit quality is emitted. We are interested in observing the role of standardization of the credit scoring and how it affects the obligors credit scores and default dynamics.

In some instances, the HCQS were higher than the CQS. This is due to the fact that an obligor scores highly when compared to the peers average dynamics but poorly when on their own.

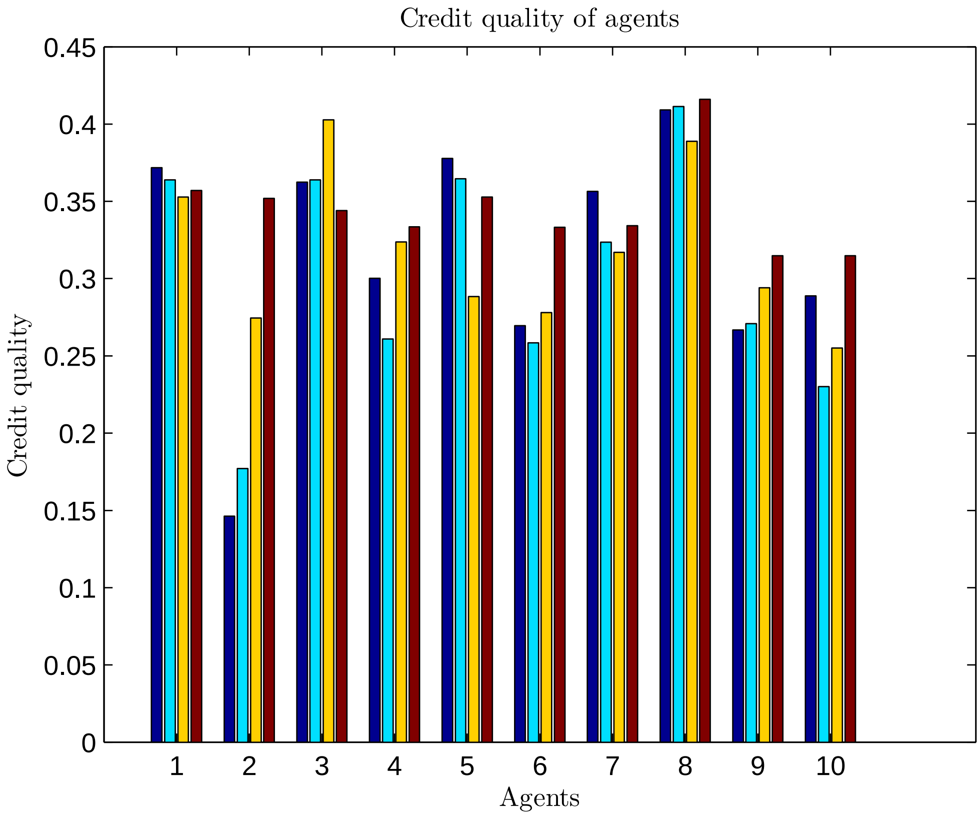

Figure 6.17: The credit quality scores of the obligors for a period of six months. The second agent had its credit quality scores increases at each time period. For the first agent, the score was almost constant for these four time periods. The dynamics through time and uncertainties are observed. The six months indicates that we have four values of the credit quality for each agent at time $t = 0, 1, 2,$ and 3.

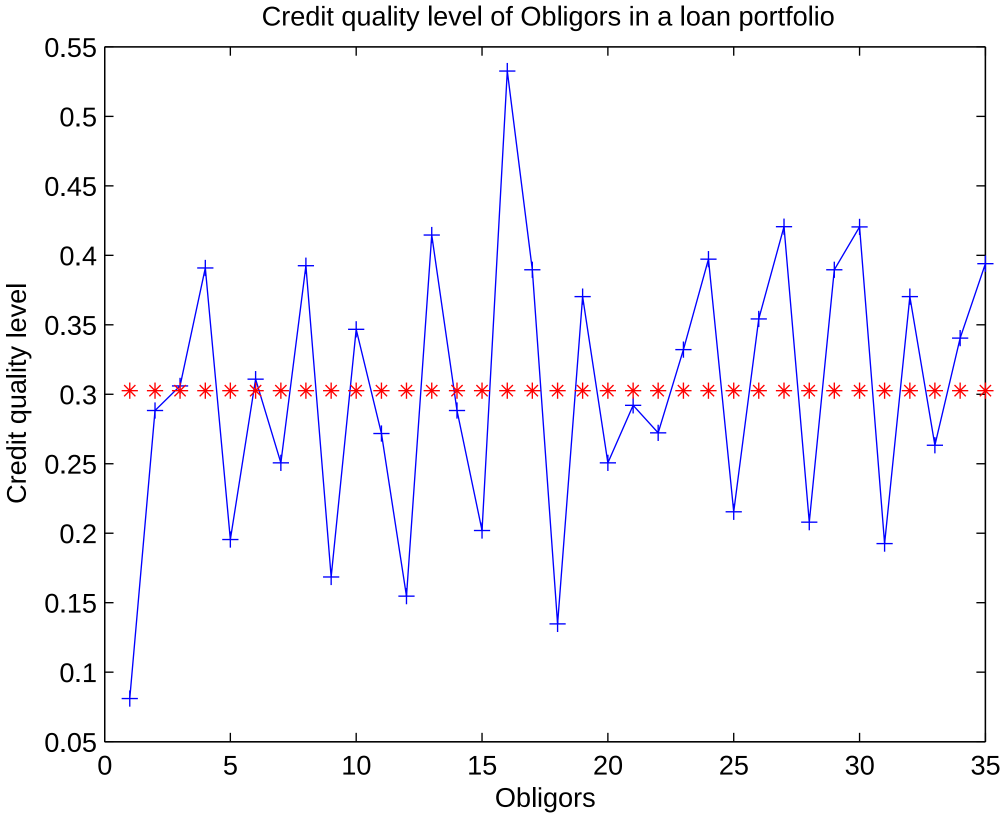

Figure 6.18: Distribution of the obligors credit quality scores when compared against the dynamic threshold.

We observe that 17 of the agents were above the threshold level, 17 below the threshold level and one agent on the threshold level. The different obligor characteristics are seen from the different credit quality scores.

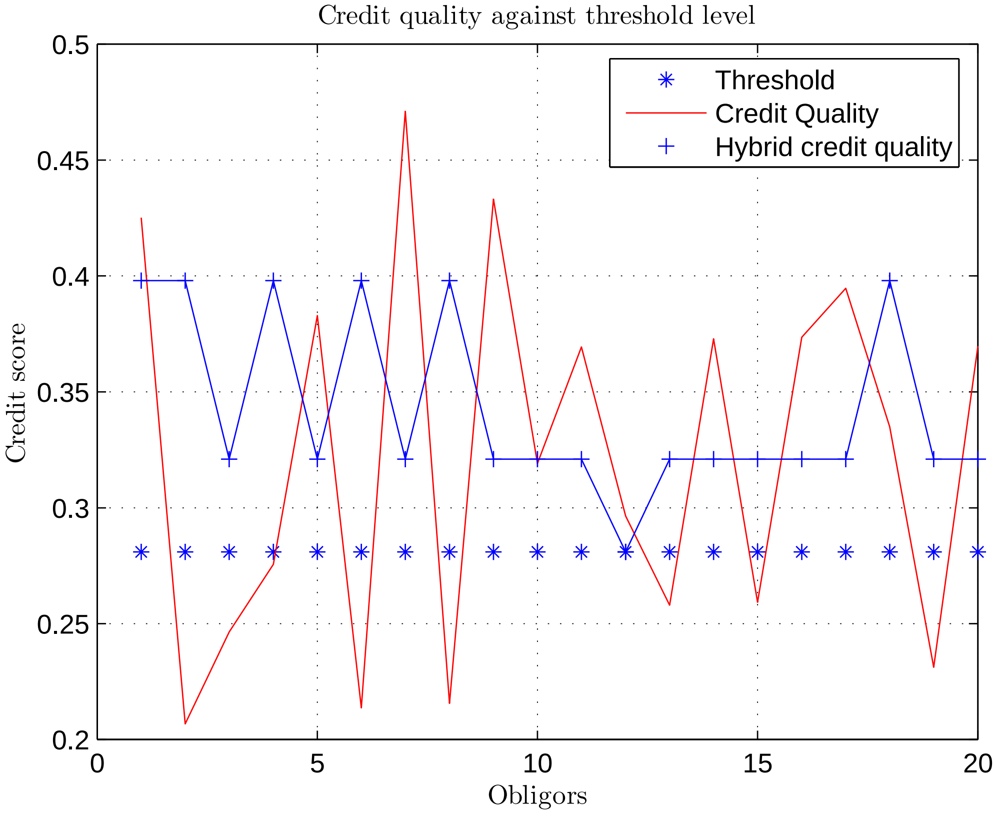

Figure 6.19: The Dynamics of the credit quality levels against the threshold level

The dynamics in the hybrid credit quality and the credit quality levels are compared against the threshold level at time $t = 3$ with twenty obligors in the loan portfolio.

Table 6.3: 95$\%$ confidence interval and the credit quality of obligors

Time, t		Dynamic	Agents	indexing			
months	Confidence interval	threshold	1	2	3	4	5
0	(0.3223, 0.5269)	0.3611	0.3835	0.4180	0.3667	0.4665	0.4883
1	(0.3488, 0.5113)	0.3842	0.4708	0.4504	0.3853	0.3854	0.4583
2	(0.3538, 0.5074)	0.3757	0.4310	0.4690	0.4016	0.3822	0.4691
3	(0.3101, 0.5149)	0.3557	0.4476	0.4608	0.3877	0.3331	0.4332
4	(0.3313, 0.4717)	0.3391	0.4098	0.3993	0.4231	0.3418	0.4334
5	(0.3099, 0.5137)	0.3455	0.4501	0.4055	0.4736	0.3417	0.3879
6	(0.3360, 0.5471)	0.3651	0.4894	0.4203	0.5060	0.3793	0.4126

The obligors credit quality score lies within the 95% confidence interval as shown in table 6.3. In this scenario, it is only the 4^{th} agent at time $t = 3$ whose credit quality score was below the threshold level, that is, $0.3331 < 0.3557$. If we apply the $0.85\bar{\theta}$ rule for the delinquent cases, then, this 4^{th} agent at $t = 3$ was a delinquent case. Thus, no agent defaulted in this scenario.

Figure 6.20 shows a display of the credit quality dynamics of the credit quality scores of the agents for a period of one year. The display is a reflection of the expectations in the long run on the shape exhibited by the credit scores of the agents. The normality observable in figure 6.20 is in agreement with observations from other studies and researchers (David , 2004). We note some outliers in the figure for agents with extreme CQS.

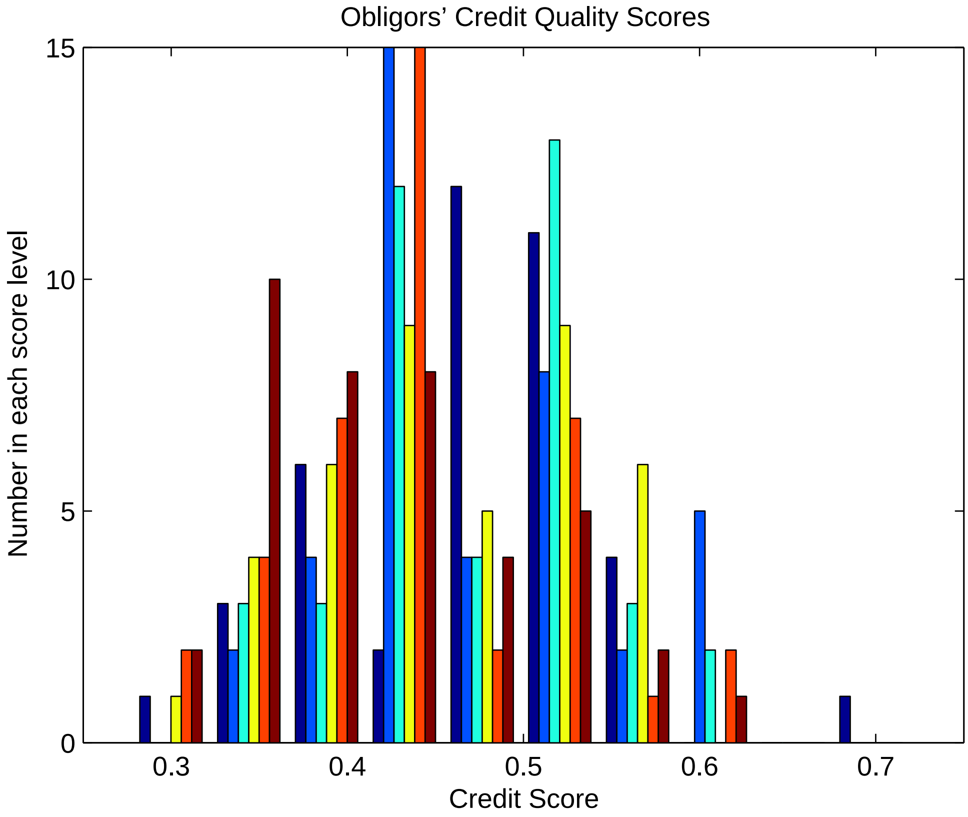

Figure 6.20: Obligors' credit quality scores

The plot shows a histogram of the credit quality scores of 40 obligors for a period of one year from $t = 1, 2, \ldots, 6$. A normal like curve is displayed by the data.

Table 6.4 shows the correlation between the credit quality levels at the different time periods during an interval of one year. There is a high positive correlation between time $t = 1$ and $t = 2$ at 0.622. This is also observed between time $t = 5$ and $t = 6$ at 0.632. We observe negative correlation at time period $t = 1$ and $t = 4$ with a correlation of -0.231. Low association is observed between time $t = 5$ and the time periods $1,\ 2$ and 3 with no general trend observable.

Time periods preceding each other have higher positive associations. This could be because of using current preceding data in estimating CRAFs. This is in order as the

SEN-HMM-CSD model estimates are time dependent.

Table 6.4: Correlation of the credit quality scores

Time, t	1	2	3	4	5	6
1	1.000					
2	0.622	1.000				
3	0.011	0.608	1.000			
4	-0.231	0.319	0.460	1.000		
5	0.030	-0.013	0.013	0.498	1.000	
6	0.247	0.177	-0.165	0.276	0.632	1.000

The correlation of the credit quality scores of 40 obligors for a period of one year after the agents are incorporated in the loan portfolio.

6.5.2 Credit quality levels classification

The obligors in the loan portfolio are classified as being in the following credit levels, poor, average, good and excellent (PAGE).

Table 6.5 shows the dynamics in the classification of the obligors into the four CQLs namely, poor, average, good and excellent. The majority of the obligors falls into the average and good levels but with some percentages in the poor and excellent levels. The model proves that its possible to use SEN data for consumer credit scoring process.

Table 6.5: Percentage of agents $(N = 40)$ in each credit quality level

Time (months)	Poor	Average	Good	Excellent
1.0	5.0	17.5	65.0	12.5
2.0	5.0	20.0	65.0	10.0
3.0	5.0	10.0	57.0	27.5
4.0	2.5	22.5	55.0	20.0

Between 65% and 82.5% of the agents credit quality is classified as is either average or good.

Table 6.6 shows the mean and standard deviation of obligors credit quality and hybrid credit quality. This is compared to the actual values of the dynamic threshold values for a period of one year. The mean credit quality is higher than the hybrid credit quality and the threshold level values. At $t = 2$ and $t = 3$, the mean credit quality is slightly lower than the threshold level. The HCQS have higher coefficient of variation compared to the CQS. This indicates that the HCQS had higher variability compared to the CQS at any given time during this period of one year with 40 obligors in the loan portfolio. The CQS had a higher mean value when compared to the HCQS mean value.

Table 6.6: Descriptive statistics for the credit scores

		Time, t					
Credit Score	Statistic	1	2	3	4	5	6
Threshold	Actual score	0.2397	0.4696	0.4700	0.2952	0.2548	0.3145
CQS	Mean	0.4724	0.4695	0.4699	0.4581	0.4424	0.4218
	Standard deviation	0.0783	0.0747	0.0665	0.0691	0.0739	0.0737
	C.V (%)	16.60	15.90	14.20	15.10	16.70	17.50
HCQS	Mean	0.3752	0.3743	0.3539	0.3438	0.3771	0.3569
	Standard deviation	0.1050	0.1028	0.1027	0.1009	0.0784	0.0714
	C.V (%)	28.00	27.50	29.00	29.30	20.80	20.00

The data is the descriptive statistics for the credit quality scores of 40 obligors for a period of one year during the life of the loan. The coefficient of variation (CV) measures the variations in the CQS and HCQS for the time period.

In table 6.7, the number of agents increased but there was also an increase in the number of agents with credit levels of average and good. This second set of agents, as compared in table 6.5 shows that the credit scoring model with social and economic data is a reliable model when it comes to consumer underwriting process. Our model, SEN-HMM-CSD offers promising prospects in use of social network data to increase the accuracy of default rates estimation.

We therefore note that the SEN-HMM-CSD model emits credit quality scores that classifies the agents with poor, average, good and excellent credit quality levels.

Table 6.7: Percentage of agents ($N = 65$) in each credit quality level

Time (months)	Poor	Average	Good	Excellent
1.0	3.1	12.3	75.4	9.2
2.0	4.6	23.1	56.9	15.4
3.0	3.1	20.0	60.0	16.9
4.0	1.5	21.5	58.5	18.5

From the table, between 80% and 87.7% of the agents credit quality is either average or good when the SEN had 65 obligors

6.5.3 Default and survival rate

The default and survival rates are one of the key components in this model. Figure 6.21 shows the obligors credit quality levels against the dynamic threshold value and the delinquent level in the model. These are the values at time $t = 3$ with 20 obligors in the loan portfolio. Any score below the lower delinquent level indicates a default while any obligor score above the lower delinquent level indicates a survival at that time period. We are more concerned with the obligors below the lower delinquent level as they are the likely default cases. As noted earlier, different random generator seed have been used in different section to depict at least some of the possible scenarios in this study.

The default and survival rates in the SEN-HMM-CSD model are as a result of CQS and the dynamic threshold. Agents CQS falling below the dynamic threshold are assumed to be delinquent cases. Those below the delinquent threshold are defaulters. The default rates are the main concern for any financial institution. They always want to understand if a default is likely to occur and when it occurs.

Figure 6.22 indicates instances in which an obligor credit score was below both the dynamics threshold level and the lower delinquent level. The marked sections are where default was realized in a scenario of 12 obligors for a period of one year where $t = 0, \ldots, 7$. Evidently, some of the obligors would have defaulted more than once assuming that they were allowed to continue with the loan even after it was noted that they had defaulted. Two obligors have instances of more than one default period while two other obligors would have defaulted only once during this period.

The survival rates for the 12 obligors for a period of one year, where $t = 0, \ldots, 7$ is

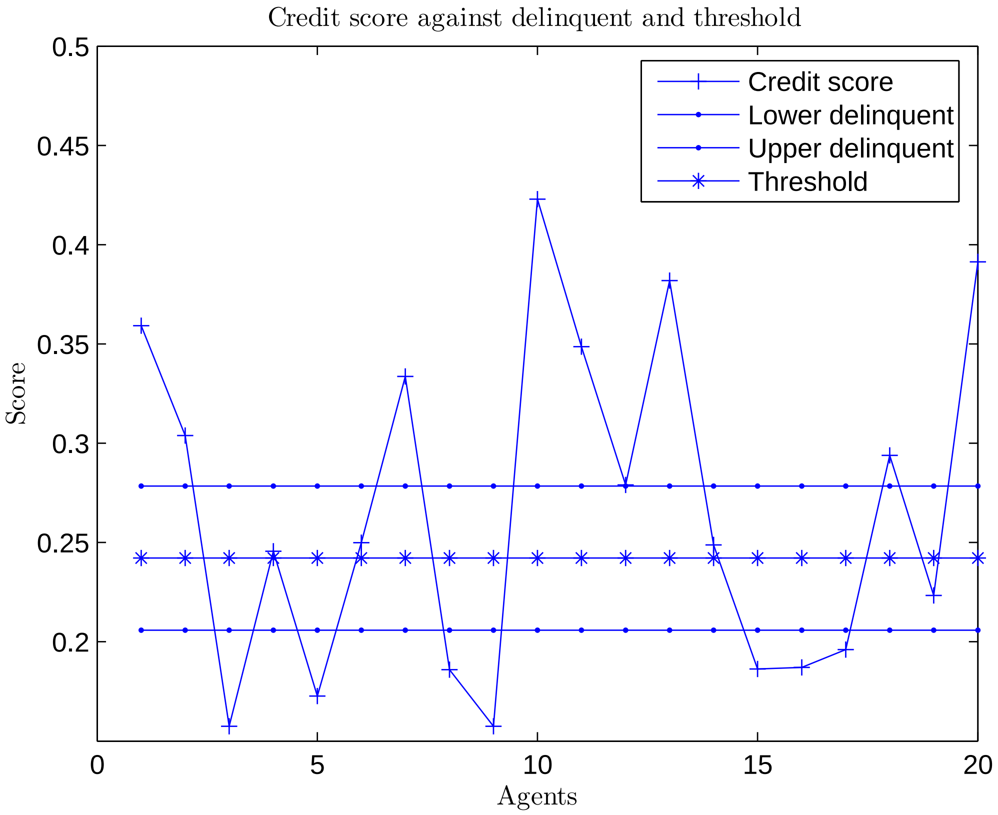

Figure 6.21: The credit score against the delinquent level and threshold value

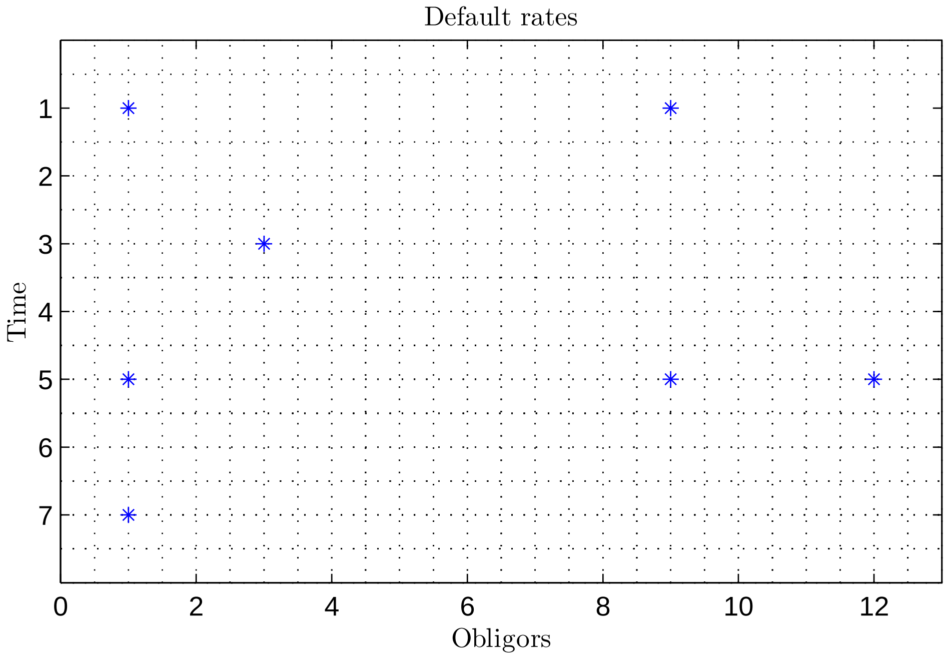

Figure 6.22: Default rates

The marks on the plot shows the instances when the obligor would have defaulted on the loan obligation. For the third and twelfth obligors, the plot indicates a default but it shows that if they were allowed to continue paying the loan premiums, they possibly would have been able to meet the obligations at the end of the time period.

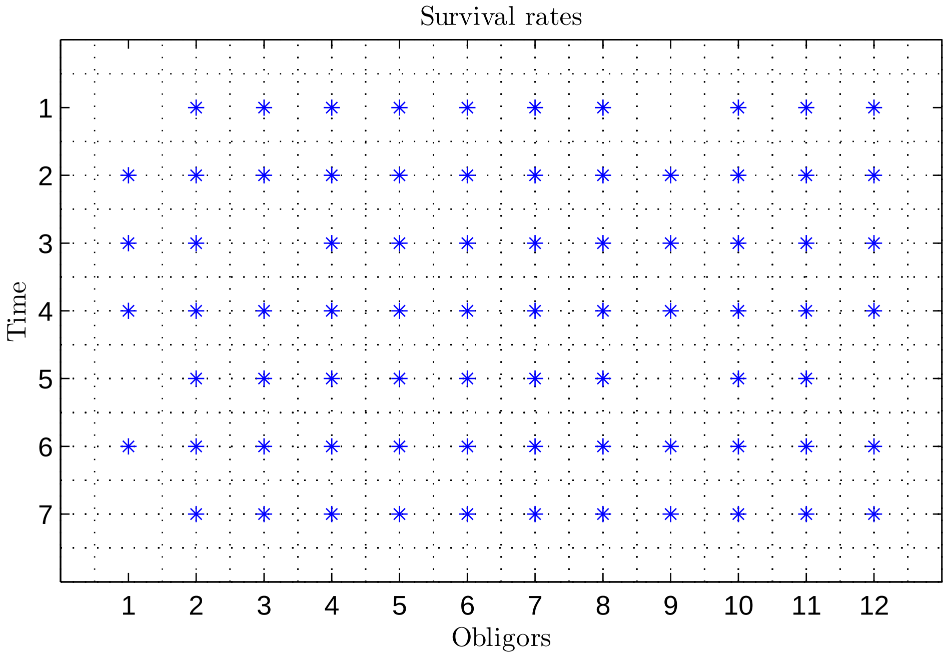

Figure 6.23: Survival rates

The blank spaces are the default rates while the marked spots are instances of survival. This compares to figure 6.22 which is a mirror image of this figure. On average as shown in figure 6.23, if we estimate the total expected marks in the figure and compare to the empty pockets which are the default rates, the defaults are very few.

depicted in figure 6.23. When figure 6.23 is compared to figure 6.22, the pockets missing in the former is the default rates while those of the later are the survival rates. We generally know that an obligor survives if they do not default. The set of the obligors in the SEN and the loan portfolio is the union of the defaults and survival rates.

6.5.4 False rates analysis

We note that high false rates in a credit scoring model reduces the accuracy of the model. This is detrimental to the actual expectations of such a model. We consider the cases of false positives and false negatives observed in the model. Figure 6.24 shows the false rates clustering of six obligors in the network for a period of one year, $t = 1, 2, \ldots, 7$.

The combination of the credit quality and hybrid credit quality of an obligor results into the classification below,

$$\text{G} = \begin{cases} g = 2 & \text{for Positive Rates} \quad \text{If } \phi \geq \bar{\phi} \text{ \& } \hat{\phi} \geq \bar{\phi}; \\ \dot{g} = 1 & \text{for False Negative} \quad \text{If } \phi \leq \bar{\phi} \text{ \& } \hat{\phi} \geq \bar{\phi}; \\ \ddot{g} = -1 & \text{for False Positive} \quad \text{If } \phi \geq \bar{\phi} \text{ \& } \hat{\phi} \leq \bar{\phi}; \\ \hat{g} = -2 & \text{for Negative Rates} \quad \text{If } \phi \leq \bar{\phi} \text{ \& } \hat{\phi} \leq \bar{\phi}; \end{cases} \tag{6.2}$$

Figure 6.24 depicts the false rates for each of the six agents for a period of one year. Obligor two had four positive ratings with obligor one, three, five and six having three positive ratings. Obligor one did not show any signs of defaulting because there is no negative ratings in that one year period. For obligor two, four, five and six, each had two negative ratings during this one year period. Obligor three had four false positive rating during this period.

The general trend according to the six plots in figure 6.24 for the six obligors is that they all exhibited different types of patterns of the false rates at a given time period of one year. However, in the last three periods of the year, the obligors had a positive rate showing that the SEN-HMM-CSD model rated them as good clients. Low false rates are ideal in real life situations to eliminate chances of denying credit facility to a ‘good’ client and giving credit to a ‘bad’ client.

Table 6.8 shows the number of false rates in the model for ten agents for a period one year. The obligor with the highest chance of not defaulting (positive rating)is the obligor number 6 with a 85.7 percent of not defaulting with a 14.3 percent rate of false positive which are instances when the obligor was declared to have defaulted while in real sense they had not defaulted. The actual default probability is 0 for the first obligor. For the third obligor, default probability was 14.3 percent, with a 28.6 percent rate of false negatives, that is, assumed to have defaulted but had not defaulted, a 14.3 percent false positive rate, that is, assumed to have not defaulted but in actual sense defaulted and a 42.8 percent rate of not defaulting.

Each of the ten obligors is counted seven times in 6.8 which is a period of one year. Only the fifth and eigth obligors who had high false negative rates of 57.1% and 42.8% respectively. This is relatively high but chances are that they defaulted during this time period. The other set of agents had a relatively low false rates during this time period.

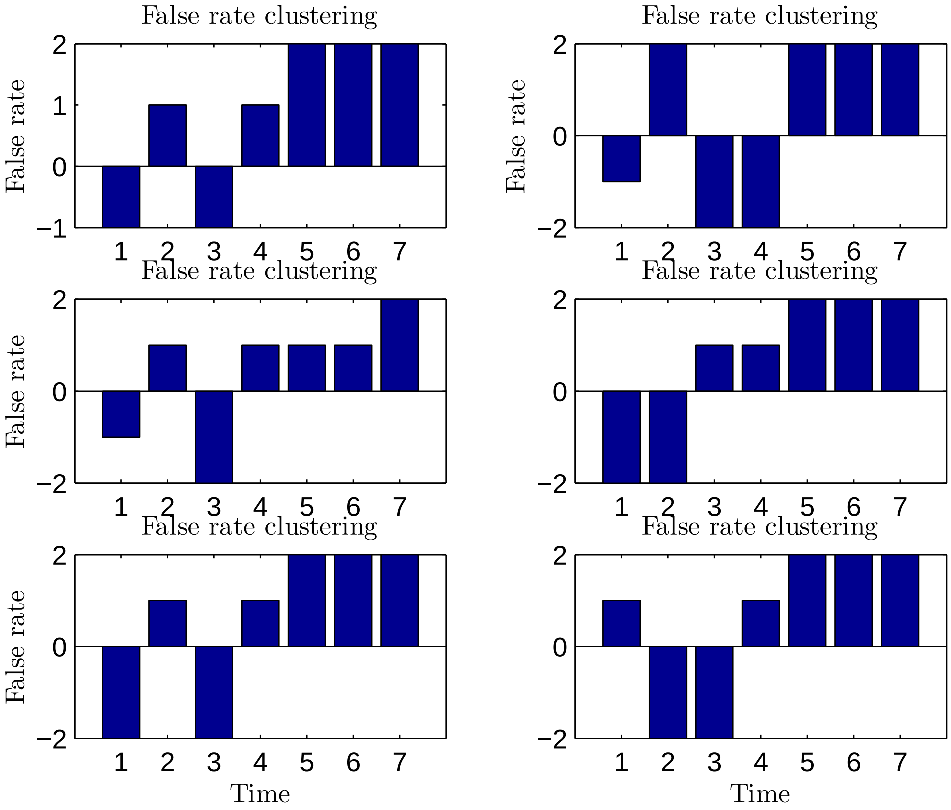

Figure 6.24: The false rates estimations

The false negative rates are represented by the value 1 while the false positive rates by the value -1 in the model. This is depicted in the six plots. They are six different obligors for a period of one year.

Table 6.8: False rates clustering

	Agents									
False Rates	1	2	3	4	5	6	7	8	9	10
Positives	5	5	3	5	2	6	5	4	5	6
Percent	71.4	71.4	42.8	71.4	28.6	85.7	71.4	57.1	71.4	85.7
False positives	1	1	1	0	0	1	0	0	2	1
Percent	14.3	14.3	14.3	0	0	14.3	0	0	28.6	14.3
False negatives	1	1	2	1	4	0	2	3	0	0
Percent	14.3	14.3	28.6	14.3	57.1	0	28.6	42.8	0	0
Negatives	0	0	1	1	1	0	0	0	0	0
Percent	0	0	14.3	14.3	14.3	0	0	0	0	0
Total	7	7	7	7	7	7	7	7	7	7
Total (percent)	100	100	100	100	100	100	100	100	100	100

The false rates for each agent are for a period of one year where $t = 1, 2, \ldots, 7$. Overall, 65.7 percent had a positive rating, 10 percent showing false positives in one year, 20 percent with false negatives and 4.3 percent with negative ratings in the same period. The negative rated obligors are the ones with poor credit quality levels.

6.5.5 Stopping time

Figure 6.25 shows the non optimal stopping time which is also similar to the default rates spots in the model. Thus, the non-optimal stopping time is the default rates while the optimal stopping time of an obligor is the ability of the agent to survive until the end of the life of the loan. We track a discrete stopping time process ti indicate the time at which an obligor defaults. The event occurs when the CQS of an obligor falls below both the dynamic threshold and the delinquent threshold levels.

Assuming that a default occurs once an obligor is not able to meet its obligations, then we have four defaults as shown in figure 6.25, that is, four non-optimal stopping time periods If it is assumed that default occurs once an obligor defaults two obligors defaulting.

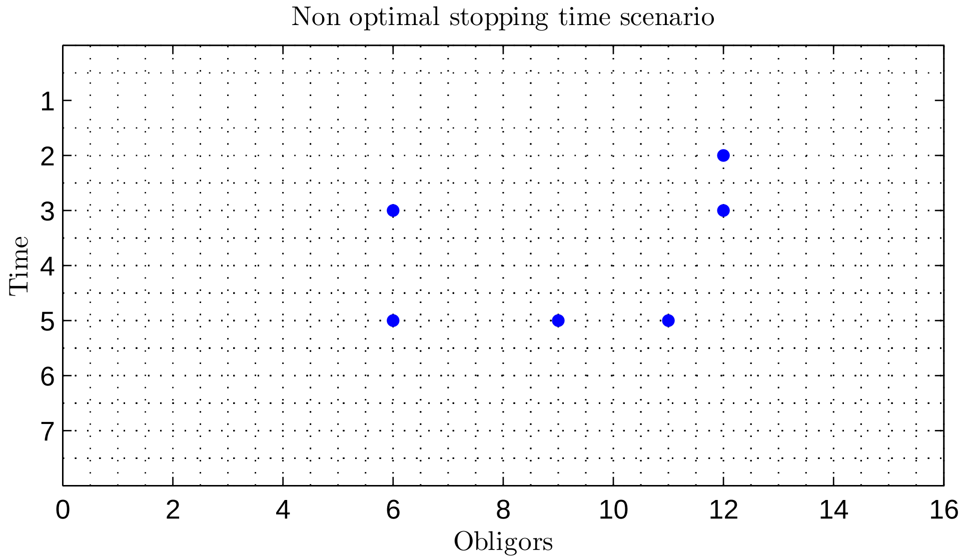

Figure 6.25: Non optimal stopping time

We have sixteen obligors in the loan portfolio and this is for a duration of one year. The dots shows instances in time the specific obligors would have defaulted, thus having non-optimal stopping time.

6.6 Model Performance Analysis

The sensitivity analysis is tested using the Pearson correlation coefficient and the results are in table 6.9. The coefficient of determination is higher for the analysis with 30 obligors compared to the situation where we have 20 obligors. The model is thus more sensitive in cases where there are fewer agents as compared to the SEN with many agents.

We refer to equation (5.47) which highlights the accuracy measure used in this study. According to table 6.9 and equation 5.47, the SEN-HMM-CSD model accuracy is between good and high accuracy rates.

Table 6.9: Sensitivity analysis with Pearson correlation coefficient

Obligors (N)	Input 1	Input 2	Correlation $(t = 1)$	Correlation $(t = 2)$
20	5	6	0.8043 $(R^2 = 0.647)$	0.8072 $(R^2 = 0.652)$
30	5	6	0.9245 $(R^2 = 0.855)$	0.9914 $(R^2 = 0.983)$

We have 20 and 30 obligors whose credit quality sensitivity analysis is undertaken when the reputation ratings are increased from 5 to 6

We test the accuracy of this simulation model by comparing the negative and positive ratings against the false negative and false positive ratings. Table 6.10 shows the percentage accuracy rates for SEN with different number of obligors. We vary the number of the obligors in the SEN and the reputation rating inputs in the model. A comparison is made on the accuracy rates in percentages at two time periods. The accuracy levels range between 53% and 73% which we feel are satisfactory for this model.

6.6.1 Credit quality and false rates

The quality of the model output in terms of the default rates is dependent on the random generator used in the simulation process. We are able to depict different scenarios that can occur in a real life situation. This study has shown that default rates can be captured based on credit scores of the obligors and the dynamic threshold which is a function of the CQS of all the obligors in the loan portfolio.

Table 6.11 is the summary of repeated simulation replications with different number of obligors in the loan portfolio who form the SEN and the time period expressed in years.

Table 6.10: Model accuracy in percentages based on the false rates

	$t = 2$		$t = 3$	
	Reputation ratings		Reputation ratings	
Obligors (N)	$R = 5$	$R = 6$	$R = 5$	$R = 6$
15	73.3	53.3	60	53.3
25	60	64	60	60
35	60	57.1	57.1	62.9
40	60	60	60	60
50	54	66	54	62

The data shows the dynamics in the credit quality levels of the obligors and the false rates in the model at different periods of time. False positives were few with the highest being at 15 percent and the lowest at 3.1 percent. The replication with the highest number of obligors with poor credit quality level had the highest number of false positives in the model.

6.6.2 Conclusion

The analysis and findings for the SEN-HMM-CSD model has been presented. Each of the model's five levels has been discussed and the findings presented in the form of tables, charts and graphs. The model performance was also presented and the accuracy noted to be within good and high accuracy rates which is acceptable for a model.

Table 6.11: Repeated replications using the model

	Replications	1	2	3	4	5
	Number of obligors	40	30	15	20	25
Time	In years	2.0	1.5	1.0	0.5	0.5
Credit Quality	Poor	23.1	9.3	10.7	20.0	1.3
(Percent)	Average	19.0	14.0	20.0	15.0	9.3
	Good	42.5	56.3	69.3	52.5	74.7
	Excellent	15.4	20.4	8.0	12.5	14.7
	Total	100	100	100	100	100
False Rates	Negatives	23.9	9.3	10.7	20.0	1.3
(Percent)	False Positives	3.1	7.4	12.0	15.0	6.7
	False Negatives	47.0	33.7	37.3	10.0	36.0
	Positives	26.0	49.6	40.0	55.0	56.0
	Total	100	100	100	100	100

The table is a summary of the credit quality levels and the false rates in the model using different simulation replications with different random generators.

Chapter 7

Discussions, Conclusions and Recommendations

The chapter presents the discussions from the study, the conclusions made from the objectives and the analysis and recommends areas of future research.

7.1 Discussions

7.1.1 SEN-HMM-CSD Model

We have developed a dynamic model that captures the cyclical and inter dependencies of the agents affected by the social and economic factors in the network. The dynamics of these SEN factors have been analyzed and depicted in figures, graphs and tables which have shown changes observable with time. Each social factor was estimated using the SVD method after the interactions and ratings at each time period. The economic factor (private data) was based on the trust levels and initial private data of the agents. The other economic factors, ethical factor, changes in private data and return on private data were also estimated. The analysis showed the stochastic nature and the dynamics in SEN.

A new definition for a social and economic network was formulated as part of the contributions brought forward in this study. It was modified from the existing definition of the social networks and graph theory. A theorem to prove that reputation ratings was developed to show that these ratings are a stochastic process. The theorem derives its proof from the filtration process which is a Martingale.

7.1.2 Credit risk factors

The SEN-HMM-CSD model has a set of values for each agent at each time period. At time, t, we have $\Gamma_{N\times\gamma}$ which are scaled to $(0,1]$ to ease the supervised clustering in estimating the transition and observation matrices. The CRAFs were estimated from the social and economic factors that were noted to be stochastic and this in turn induced the same dynamics in CRAFs.

The credit risk analysis factors were used in this study for the first time. No other research or study has ever introduced these factors in the analysis of consumer credit scores using the social and economic data.

7.1.3 HMM parameters

The multiple agent HMM was introduced by the modification of the existing standard HMM. This caters for the heterogenety in the SEN dynamics and in learning and training the HMM.

The CRAF data set formed an excellent basis for the learning and training of HMM. Credit risk analysis factors were used to estimate the transition and observation matrices for each agent. These matrices were in turn used to estimate the observation sequences and state sequences for each agent at each time period. Maximum likelihood was then used alongside the observation and state sequences to obtain optimal transition and observation matrices. The Matlab inbuilt functions for HMM were applied to estimate all the parameters necessary for the computation of the credit scores and default rates.

The supervised clustering used in the study to estimate matrices A and B were also used the first time in this study. No known research work exists that has estimated the transition and observation matrix with such as a technique. In chapter four, the parameters of the HMM were modified to suit the multiple agents and multiple observations for the study. Thus, this is a new contribution that the study was able to offer in the field of credit scoring using the HMM.

7.1.4 Credit scores and default rates

The HMM emissions were the credit scores and the credit levels based on the poor, average, good and excellent. The CQS had a range of between zero and one, which is a probability of survival and is weighed against the dynamic threshold. Default rates (from dynamic threshold) were observed to vary with time and conditions induced in the SEN model. The HCQS was used to offer a buffer as a form of global credit score to develop a benchmark of how obligors would perform when a comparison was made against one set of HMM parameters.

The CQLs analysis showed that between 61.5% and 89.3% of the obligors had average or good credit quality levels. The excellent credit quality levels were in the range of between 8% and 20.4%. The rest would then belong to the poor credit quality level which is the group that we would not expect to receive any credit facility from a financial institution. The dynamic threshold was able to mimic the stochastic and the dynamics of the social and economic network. This in turn showed the changes in the possible default rates from the credit scores and the dynamic threshold values.

7.1.5 False rates

The false rates in the model showed that between 25% and 50.1% of the obligors were classified as either having false positive or false negative rating by the model. We observe that the accurate estimation were between 49.9% and 75%. The SEN-HMM-CSD model had good estimates of the delinquent cases, survival rates and the stopping time, both optimal and non-optimal. When the accurate estimations are considered, the model developed has good accuracy rating.

7.1.6 Model performance

The model performance was based on the accuracy level using the false rates of the model and the sensitivity analysis with the coefficient of determination. The model sensitivity with changes in the input and how it affects output had a coefficient of determination between 64.7% and 98.3%. For the accuracy with false rates, they range between 57.3% and 70.0%. A combination between false rates and sensitivity analysis showed a model performance of between 54% and 66%. These ranges are within our set accuracy levels of good and high accuracy rate.

We are confident that the SEN-HMM-CSD model has proved to be accurate and within acceptable performance rate in estimating consumer credit scores using the SEN data.

7.2 Conclusions

Social networks are a source of social media data which provide more insights on how social links are formed, evolve over time and how this affects people and the interactions choices they make. The availability of powerful data mining tools can harden the soft information to increase the information levels crucial in consumer credit underwriting. This offers an opportunity to develop innovative products to widen the market share for credit lending to the poor and young people who lack financial histories; and increase the data set for improved credit scoring.

The SEN-HMM-CSD model does not deteriorate over time and generates time dependent data for consumer credit scoring. The model is dynamic as it captures the agents cyclical inter dependencies and the market events as agents do not live in vacuum. The model does not rely on historical data which is one of the key limitations of the current credit scoring models.

The SEN-HMM-CSD model exhaustively captured the study objectives and forms a basis for consumer credit scoring research. The three components in this model: the SEN estimated agent dynamics and captured the data necessary for credit scoring; the HMM as a stochastic model with ability to classify agents according to the richness of data available emitted the credit scores, credit levels and the dynamic threshold value; the CQS estimated all the scoring variables to successfully estimate the obligors credit risk.

Therefore, the SEN-HMM-CSD model is stochastic, dynamic and an alternative approach to consumer credit scoring using social and economic data. The model is reliable and offers results that are accurate to estimate consumer credit scores using SEN data.

7.3 Recommendations

The study brought out areas that require further and advanced research work. First, use of real life data from the social media life twitter, facebook, and other networks. Second, develop a numerically tractable model that can be used as a standalone or alongside the existing credit scoring models. Third, relax some of the assumptions in the SEN-HMM-CSD model like use of uniform distribution, use of ergodic transitions and having a fully connected SEN network, among other assumptions. Fourth, validate the SEN-HMM-CSD model with actual market data.

References

Aalen, O. O., and Gjessing, H., K. (2005). Stochastic processes in survival analysis. Unpublished manuscript, University of Oslo, Norway [38]

Alese, B.K., Adewale. O.S., Aderounmu, G.A., Ismaila, W.O., and Omidiora, E.O. (2012). Investigating the effects of threshold in credit card fraud detection system. International Journal of Engineering and Technology, 2(7) [39]

Allen, F., and Babus, A. (2008). Networks in Finance. Unpublished manuscript, Partnership Between University of Pennsylvania and Erasmus University [3, 7, 23, 83]

Baesens, B., and Gestel, T. (2009). Credit Risk Management Basic Concepts: Financial Risk Components, Rating Analysis, Models, Economic and Regulatory Capital. Oxford University Press, United Kingdom [9, 33]

Banasik, J., Crook, J. N., and Thomas, L.C. (1999). Not if but when will borrowers default. The Journal of the Operational Research Society, 50 (12), 1185-1190 [34, 36, 37]

Bilmes, J.A (2006). What hidden Markov models can do. IEICE Transactions, Information and Systems (Special Issue on IEICE Transactions), E89-D (3) [14, 15, 30, 46, 49, 50, 51, 52, 62, 74]

Bhusari, V., and Patil, S. (2011). Study of hidden Markov model in credit card fraudulent detection. International Journal of Computer Applications, 20 (5) [30, 32, 33]

Boyle, P. (1977). A Monte Carlo approach. Journal of Financial Economics, 4 (3) , 323-338 [58, 60]

Broadie, M., and Glasserman, P. (2004). A stochastic mesh method for pricing high-dimensional American options. Journal of Computational Finance, 7 (4), 35-72 [32]

Bucay, N., and Rosen, D. (2000). Applying portfolio credit risk models to retail portfolios. The Journal of Risk Finance, 2 (3), 35-61 [33, 34]

Carla, D.M., and Mason, A.P. (2012). The extraordinary SVD. Unpublished manuscript, arVix:1103.2338v5 [41, 42, 43]

Capuano, C., Chan-Lau, J., Gasha, G., Medeiros, C., Santos, A., and Souto, M. (2009). Recent advances in credit risk modeling. International Monetary Fund. Unpublished manuscript, WP/09/162 [3, 6, 7, 9, 33, 35, 39, 40]

Cetin, U., and Jarrow, R. and Protter, P. and Yildirim, Y. (2004). Modeling credit risk with partial information. The Annals of Applied Probability, 14 (3), 1167-1178 [37]

Chen, H.J., Goldberg, M., Malik, M., and Wallace, W. (2008). Reverse engineering an agent-based hidden Markov model for complex social systems. International Journal of neural Systems, 1-24 [8, 23, 30]

Chen, H.,Geng, Z., and Jia, J. (2007). Hidden Markov models with multiple observers. Springer-Verlag Berlin Heidelberg, 427-435 [30, 31]

Ching, W.K., Fung, E., Ng, M., Siu, T.K., and Li, W.K. (2006). Interactive hidden Markov models and their applications. Unpublished manuscript, Department of Mathematics, The University of Hong Kong, Hong Kong [30, 74]

Ching, W., Leung, H., Wu, Z., and Jiang, H. (2008). Modeling global risk via a hidden Markov model of multiple sequences. Unpublished manuscript, The Department of Mathematics, The University of Hong Kong, Hong Kong [6, 33]

Crowder, M., Davis, M., and Giampieri, G. (2005). A hidden Markov model of default interaction. Quantitative Finance, 5, 27-34 [4, 30, 33, 37]

Couvreur, C. (1996). Hidden Markov models and their mixtures. Unpublished PhD Thesis, Faculty of Science, Department of Mathematics, University Catholic of Louvain, Belgium [73, 74]

Daniel, B., and Grissen, D. (2015). Behavior revealed in mobile phone usage predict loan repayment. Unpublished manuscript, Department of Economics, Brown University, United States of America [2, 3, 8, 9, 10, 12, 26, 35, 36]

David, R.K. (2004) Professional Risk Managers Handbook. A Comprehensive Guide to Current Theory and Best Practices, Volume III: Risk Managment Practices, Edited by Carol Alexander and Elizabeth Sheedy, PRMIA, United States of America [4, 9, 10, 37, 140]

Davis, M., Crowder, M., and Giampieri, G. (2005). A Hidden Markov model of default interaction. Unpublished manuscript, Department of Mathematics, Imperial College, London [30]

Denault, M., Gauthier, G., and Simonato, J. (2009). Estimation of physical intensity models for default risk. The Journal of Futures Markets, 29 (2), 95-113 [33, 37, 38]

Dewing, M. (2012). Social media: An introduction. In Brief, Library of Parliament, Social Affairs Division, Parliamentary Information and Research Services, Publication No. 2010-03-E, Ottawa, Canada [8, 26, 36]

Dhaene, J., and Goovaerts, M. J. (1996). On the dependency of risks in the individual life model. Insurance: Mathematics and Economics, 19, 243-253 [37]

Duan, J. and Simonato, J. (1998). Empirical martingale simulation for asset prices. Management Science, 44 (9), 1218-1233 [39]

Dubois, T., Golbeck, J., and Srinivasan, A. (2011). Predicting trust and distrust in social networks. Unpublished manuscript, University of Maryland, College Park, United states of America [8, 13, 14, 27, 28, 36, 54]

Dymarski, P. (2011). Hidden Markov models, theory and applications. Intech Open Access Publisher. Online Version: [http://www.intechopen.com] [5, 69]

Ehab, E., and Sassone, V. (2013). A HMM-based reputation model. Advances in Security of Information and Communication Networks 381, 111-121 [27, 30, 31, 99]

Ehrhardt, G. C.M.A., Marsili, M., and Fernando, V. (2006). Phenomenological models of socio-economic network dynamics. Unpublished manuscript, arXiv: physics/0604036 [22]

Eisenberg, L., and Noe, H. T. (2001). Systemic risk in financial systems. Management Science, 47 (2), 236-249 [2, 18, 19, 33, 34, 37, 39, 40]

Ephraim, Y., Merhav, N. (2002). Hidden Markov processes. IEEE Transactions on Information Theory, 48 (6) [44, 45, 52]

Ernst & Young. (2013). Social Media Risk Management and FFIEC Proposed Guidance. Financial Services Alert. Score No. CK0680, [www.ey.com] [12]

Finger, C. C. (2000). A comparison of stochastic default rate models. The Risk Metrics Group, Working Paper No. 00-02, 44 Wall Street, New York, NY 10005 [33, 37, 38]

Fonzo, V., Filippo, A., and Parisi, V. (2007). Hidden Markov models in bioinformatics. Bentham Science Publishers Limited, 2, 49-61 [30]

Frey, R., and Runggaldier, W. (2007). Credit risk and incomplete information: A nonlinear filtering approach. Unpublished work, University of Leipzig. [33]

Galassi , U. (2008). Structure hidden Markov model: A general tool for modeling process behaviour. Unpublished PhD Thesis, Department of Informatics, Universita degli studi di Torino [68, 74]

Gambetta, D. (1988). Trust: Making and breaking cooperative relations. Oxford: Basil Blackwell [27, 28]

Ganesh J., and Sethi, P. (2013). Reputation and trust in social networks: Empirical results from a Facebook reputation system. Proceedings of the Nineteenth Americas Conference on Information Systems, Chicago, Illinois, August 15-17. [8, 28]

Giesecke, K. (2005). Default and information. Journal of Economic Dynamics and Control (School of Operations Research and Industrial Engineering, Cornell University, United States of America) [33, 37]

Giesecke, K., and Kim B. (2010). Systemic risk: What defaults are telling us. Working Paper, Department of Management Science and Engineering, Stanford University, United States of America [37]

Glaeser, E. L., Laibson, D., and Sacerdote, B. (2002). An economic approach to social capital. The Economic Journal, 112 (483), F437-F458 [24]

Gordy, B. M. (2000). A comparative anatomy of credit risk models. Journal of Banking and Finance, 24, 119-149 [33]

Guha, R., Kumar, R., Raghavan, P., and Tomkins, A. (2004). Propagation of trust and distrust. ACM 1-58113-844-X/04/0005. WWW2004, New York, United States of America [27, 28]

Gurny,P., and Gurny, M. (2013). Comparison of credit scoring models on probability of default estimation for US banks. Prague Economic Papers, 2, 163-181 [33, 37]

Hassan, R.M., Nath, B., and Kirley, M. (2006). A data clustering algorithm based on single hidden Markov model. Proceedings of the International Multiconference on computer Science and Information Technology, pp. 57-66 [14]

Hassan, R., and Nath, B. (2005). Stock market forecasting using hidden Markov model: A new approach. Proceedings of the 2005 5th International Conference on Intelligent Systems Design and Applications (ISDA) [14, 32]

Haugh, M. B., and Kogan, L. (2004). Pricing American options: A duality approach. Operations Research, 52 (2), 258-270 [32, 33, 39]

Haugh, M. (2010). Simulation efficiency and an introduction to variance reduction methods. Monte Carlo Simulation: IEOR E4703 [57]

Hodgman, D. R. (1960). Credit risk and credit rationing. The Quarterly Journal of Economics, 74 (2), 258-278 [33]

Horkko, M. (2010). The determinants of default in consumer credit market. Unpublished Master's Thesis, Department of Accounting and Finance, School of Economics, Aalto University [4, 6, 33, 37]

Iqbal, N., and Ali, S.A. (2012). Estimation of probability of default for low default portfolios: An actuarial approach. Research paper presented at the 2012 ERM Symposium [37]

Irle, A. (2006). A forward algorithm for solving optimal stopping problems. Journal of Applied probability, 43 (1), 102-113 [33]

Jackson, M. O. (2008). Social and Economic Networks. Princeton University Press, USA [7]

Johnson, C. (2003). A model of social capital formation. Social Research Demonstration Corporation Working Paper Series 03-01 [12, 13, 23, 24, 39, 40]

Josang, A., Ismail, R., and Boyd, C. (2007). A survey of trust and reputation systems for online service provision. Decision Support Systems, 43, 618-644 [27, 28, 54, 55]

Kalman, D. (1996). A singular value decomposition: The SVD of a matrix. The College Mathematics Journal, 27 (1) [42, 93]

Karris, S. T. (2007). Numerical Analysis: Using Matlab and Excel. Orchard Publications, United States of America [31]

Kealhofer, S. (2003). Quantifying credit risk 1: Default prediction. Financial Analysts Journal, 59 (1), 30-44 [37]

Knack, S., and Keefer, P. (1997). Does social capital have an economic payoff? The Quarterly Journal of Economics, 112 (4), 1251-1288 [29, 54]

Koubaa, A. (2008). Modeling and simulation lecture notes. Unpublished manuscript, College of Computer Science and Information Systems. Al-Imam Mohammad Ibn Saud University [50, 51, 52, 53]

Korolkiewicz, M.W. (2010). A dependent hidden Markov model of credit quality. Hindawi Publishing Corporation. International Journal of Stochastic Analysis, Article ID 719237 [10, 30, 31]

Kleijnen, J.P.C. (2009). Sensitivity analysis of simulation models. Discussion Paper No. 2009-11, Department of Information Management, Tilburg University [60, 109]

Koyluoglu, H. and Hickman, A. (1998). Reconcilable differences. Risk, October [9, 37]

Lajos, J., George, K.M., and Park, N. (2012). A six state HMM for the S& P 500 Stock Market Index. Unpublished manuscript, Mathematical Sciences, Oklahoma State University, USA [30]

Leach, S. (1995). A singular value decomposition- A primer. Unpublished manuscript, Department of Computer Science, Brown University, Providence, RI, 02912 [42, 43]

Lee, H., and Kim, J.H. (1999). An hidden Markov model-based threshold model approach for gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21 (10) [38]

Lee, J., Kim, S., Lebanon, G.,and Singer, Y. (2013). Local low-rank matrix approximation. Proceedings of the 30th International Conference on Machine Learning, JMLR: W&CP Vol. 28, Atlanta, Gorgia, United States of America [42]

Li, X., Parizeau, M., and Plamondon, R. (2000). Training hidden Markov models with multiple observations-A combinatorial method. IEEE Transactions on PAMI, 22 (4), 371-377 [31]

Li, A. and Zhong, Y. (2012). An overview of personal credit scoring: Techniques and future work. International Journal of Intelligence Science. 2, 181-189 [4, 5, 6, 35]

Liu, X. and Datta, A. (2012). Modeling context aware dynamic trust using hidden Markov model. Proceedings of the Twenty-Sixth Association for the Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence [3, 25, 27, 28, 30, 31, 91]

Loni, B., Joseph, C.E., and Nobakht, B. (2012). Stock market analysis and prediction using hidden Markov models. Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS) [30]

Loso, J., and Koski, T. (2014). Forecasting of self-rated health using hidden Markov algorithm. KTH Royal institute of Technology, Mathematical Statistics [110]

Lyengar, G. N. (2005). Robust dynamic programming. Mathematics of Operations Research, 30 (2), 257-280 [67]

Madan, A., and Pentland, A.S. (2009). Modeling social diffusion phenomena using reality mining. AAAI Spring Symposium on Human Behavior Modeling [7, 22]

Madhur, M. and Thomas, L. (2007). Modelling credit risk in portfolios of consumer loans: Transition matrix model for consumer credit ratings. School of Management, University of Southampton, United Kingdom [33]

Masyutin, A.A. (2015). Credit scoring based on social network data. Business Informatics, 3 (33), 15-23 [3, 26, 36]

Mathew, B. (1997). Coupled hidden Markov models for modeling interacting processes. MIT Media Lab Perceptual Computing/ Learning and Common Sense Technical Report 405 [30, 74]

The MathWorks. (2003). Statistics Toolbox: For use with Matlab, User's Guide, Version 4. Retrieved from [https://www.mathworks.com] [100, 173]

Maybeck, P. S. (1979). Stochastic models, estimation and control, Volume 1. Academic Press, New York, United States of America [17]

Meyers, R. A. (2009). Complex systems in finance and economics (Selected Entries from the Encyclopedia of Complexity and Systems Science). Springer, New York, United States of America [7, 8, 13, 21]

Mhamanne, S., and Lobo, L.M.R.J. (2012). Fraud detection in online banking Using HMM. 2012 International Conference on Information and Network Technology, 37 [30, 32, 62]

Miller, D.R.H., Leek, T., and Schwartz, R.M. (1999). A hidden Markov model information retrieval system. ACM 1-58113-096-1/99/0007 [30]

Moe, M.E.G., Tavakolifard, M., and Knapskog, S.J. (2008). Learning trust in dynamic multiagent environments using HMMS. In Proceedings of the 13th Nordic Workshop on Secure IT Systems (NordSec 2008) [3, 25, 32]

Moffatt, P. G. (2005). Hurdle models of loan default. The Journal of the Operational Research Society, 56 (9): 1063-1071 [37]

Mui, L. (2002). Computational models of trust and reputation: Agents, evolutionary games and social networks. Unpublished PhD Thesis, MIT, USA [54]

Musco, C., and Musco, C. (2015). Stronger and faster approximate singular value decomposition via the block Lanczos method. CoRR, abs/1504.05477 [42]

Mustafa, S.E., and Hamzah, A. (2011). Online social networking: A new form of social interaction. International Journal of Social Science and Humanity, 1 (2), 96-103 [26]

Nan, L. (1999). Building a Network Theory of Social Capital. Department of Sociology, Duke University, Connections 22 (1): 28-51 [24, 91]

Neftci, S. N. (2000). An Introduction to the Mathematics of Financial Derivatives. Academic Press, London, United Kingdom [45, 46, 47]

Netrvalova, A., and Safarik, J. (2011). Trust model for social network. Department of Computer Science and Engineering. University of West Bohemia, Czech Republic [12, 21, 27]

Netzer, O., Lattin, J.M., and Srinivasan, V. (2008). A hidden Markov model of customer relationship dynamics. Marketing Science, 27 (2) [30, 31]

Pani, S. K. (2008). Organizational Social Capital (OSC): A Theoretical and Mathematical Treatment (Understanding the Structural and Reputational Aspects of OSC). Working Paper, Indian Institute of Management, Bangalore, India [13, 22, 24]

Pupazan, E. (2011). Social Networking Analytics. BMI Paper, VU University Amsterdam [7, 12, 21, 24, 27]

PWC. (2015). Is it Time for Consumer Lending to go Social? How to strengthen Underwriting and grow your customer base with social media data. Consumer Finance Group. [www.pwc.com/consumerfinance] [3, 8, 11, 12, 35]

Quirini, L., and Vannucci, L. (2014). Creditworthiness dynamics and hidden Markov models. Journal of Operational Research Society, 65, 323-330 [30]

Rabiner, L.R. (1989). A tutorial on hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE, 77 (2) [29, 61, 62, 65, 67, 68, 69, 72, 74, 75, 174]

Raghavan, V., Steeg, G., Galstyan, A., and Tartakovsky, A.G. (2013). Modeling temporal activity patterns in dynamic social networks. Department of Mathematics, University of Southern California, United States of America [21, 27, 28, 30]

Risk Assessment Guidance for Superfund (RAGS). (2001). Sensitivity analysis: How do we know what's important? Process for conducting probabilistic risk assessment, Volume 3, Part A [109]

Resnick, P., Zeckhouse, R., Friedman, E., and Kuwabara, K. (2000). Reputation system. Communications of the ACM, 43 (12) [8, 13, 21, 25]

Robert, B.A., Raphael, W.B., Paul, S.C., and Glenn, B. C. (1996). Credit risk, credit scoring, and the performance of home mortgages. Federal Reserve Bulletin [4, 10, 33, 34, 35]

Ross, S.M. (2007). Introduction to Probability Models, 9th Edition. Academic Press, United States of America [44]

Rubinstein, R.Y. (1981). Simulation and the Monte Carlo Method. John Wiley, New York, United States of America [56]

Sadek, R.A. (2012). SVD Based image processing applications: State of the art, contributions and research challenges. International Journal of advanced computer science and applications, 3 (7) [42]

Samik, R. (2008). Introduction to Monte Carlo simulation. Proceedings of the 2008 Winter Simulation Conference [39, 56, 57]

Sargent, R. A. (2011). Verification and validation of simulation models. Proceedings of the 2011 Winter Simulation Conference (S. Jain, R.R. Creasey, J. Himmelsspach, K.P. White, and M. Fu, eds.) [109]

Schweitzer, Fagiolo, Sornettel, Vega-Redondo, Vespignani and White. (2009). Economic networks: The New Challenges. Science, 325: 422-425 [17, 33, 34]

Serrano-Cinca, C., Gutierrez-Nieto, B., and Reyes, N.M. (2013). A social approach to micro-finance credit scoring. Centre Emile Bernheim. Research Institute in Management Science, Unpublished manuscript, CEB Working Paper No. 13/013 [35]

Shen, D. (2008). Some mathematics for HMM. Retrieved from [http://courses.media.mit.edu/2010fall/mas622j/ProblemSets/ps4/tutorial.pdf] [174]

Siva, W. (2010). Business intelligence and predictive analytics for financial services. The untapped potential of soft information. Center for digital innovation, technology and strategy. School of Business, University of Maryland, United States of America [27]

Sjoerd, B., and Smulders , S. (2004). Social capital and economic growth. Faculty of Economics, Tilburg University, Warandelaan 2, Tilburg [29, 90]

Skyrms, B., and Pemantle, R. (2009). A dynamic model of social network formation. Proceedings of the National Academy of Sciences of the United States of America, 97 (16), 9340-9346 [8, 21, 22, 27, 28, 33]

Soman, S.T., Soumya, V.J., and Soman, K.P. (2009). Singular value decomposition. International Journal of Recent Trends in Engineering, Academy Publisher, 1 (2) [42]

Srivastava, A., Kundu, A., Sural, S., and Majumdar, A.K. (2008). Credit card fraud detection using hidden Markov model. IEEE Transactions on Dependable and Secure Computing, 5 (1) [30, 32]

Stanley, E.A. (2006). Social networks and mathematical modeling. Connections, 27 (1) , 39-45 [12, 23, 39, 40, 76]

Starnini, M., Baronchelli, A., and Romualdo, P. (2013). Modeling Human dynamics of face to face interaction networks. Physical Review Letters, 110, 168701 [21]

Stirzaker, D. (2005). Stochastic Processes and Models. Saint John's College, Oxford University Press, Oxford. [47, 48, 52]

Stepanova, M., and Thomas, L. (2002). Survival analysis methods for personal loan data. Operations Research, 50 (2), 227-289 [10, 34, 37, 38]

Tang, J., Hu, X., and Liu, H. (2014). Is distrust the negation of trust? The value of distrust in social media. Association of Computing Machinery, 978-1-4503-2954-5/14/09, Santiago, Chile [8, 27, 28]

Terejanu, A. G. (2002). Tutorial on Monte Carlo techniques. Department of Computer Science and Engineering, University of Buffalo, New York [39, 56, 58, 59, 60]

Thomas, L. C., Oliver, R. W, and Hand D. J. (2005). A survey of the issues on consumer credit modeling research. The Journal of the Operational Research Society, 56 (9), 1006-1015 [4, 6, 9, 11, 33]

Thomas, L.C. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. International Journal of Forecasting, 16, 149-172 [90]

Ueda, N. and Sato, T. (2004). Simplified training algorithm for hierarchical hidden Markov models. Electronics and Communications in Japan, Part 3, 87 (5) (Translated from Denshi Joho Tsushin Gakkar Ronbunshi, J85-D-I (6), 538-548 (June 2002)) [74]

Viswanath, B., Mislove, A., Cha, M., and Gummadi, K.P. (2009). On the evolution of user interaction in facebook. ACM 978-1-60558-445-4/09/08 [29]

Volker, S. (2010). Markov chains and Monte Carlo simulation. Lecture notes, Institute of Stochastic, ULM University [49, 50]

Wei, Y., Yildrim, P., Bulte, C., and Dellarocas, C. (2015). Credit scoring with social network data. Marketing Science, Articles in Advance, pp. 1-25 [3, 36]

Wilson, C., Boe, B., Sala, A., Puttaswamy, K.P.N., and Zhao, B.Y. (2009). User interactions in social networks and their implications. ACM, 978-1-60558-482-9/09/04 [29]

Xiang, R., Neville, J., and Rogati, M. (2010). Modeling relationship strength in online social networks. WWW2010: ACM 978-1-60558-799-8/10/04, North Carolina, United States of America [14, 29]

Zhao, K., Stehle, J., Bianconi, G., and Barrat, A. (2011). Social network dynamics of face to face interactions. Physical Review E 85, 056109 [28, 29]

Zhong, S., and Ghosh, J. (2001). A new formulation of coupled hidden Markov models. Department of Electrical and Computer Engineering, The University of Texas at Austin [74]

Appendix A

Appendix

A.1 HMM Basic Problems Derivations

The basic problem in hidden Markov model for the multiple agents are modified and outlined in this section of the appendix

A.1.1 Forward Variable

$\alpha_t^{(n)}(i)$ can be obtained inductively as follows

(1) Initialization

$\alpha_1^{(n)}(i) = \pi_i^{(n)} b_i(O_1^{(n)}),\ 1 \leq n \leq N,\ t = 1$

$$\begin{pmatrix} \alpha_1^{(1)}(i) \\ \alpha_1^{(2)}(i) \\ \vdots \\ \alpha_1^{(N)}(i) \end{pmatrix} = \begin{pmatrix} \pi_i^{(1)} b_i(O_1^{(1)}) \\ \pi_i^{(2)} b_i(O_1^{(2)}) \\ \vdots \\ \pi_i^{(N)} b_i(O_1^{(N)}) \end{pmatrix}$$

Initialize the forward probability as the joint probability of state i and initial observation $O_1^{(n)}$ for each agent.

(2) Induction

$\alpha_t^{(n)}(i) = [\sum \alpha_t^{(n)}(i) a_{ij}] b_j(O_{t+1}^{(n)})$

$$\begin{pmatrix} \alpha_{t+1}^{(1)}(j) \\ \alpha_{t+1}^{(2)}(j) \\ \vdots \\ \alpha_{t+1}^{(N)}(j) \end{pmatrix} = \begin{pmatrix} (\sum \alpha_t^{(1)}(i) a_{ij}) b_j(O_{t+1}^{(1)}) \\ (\sum \alpha_t^{(2)}(i) a_{ij}) b_j(O_{t+1}^{(2)}) \\ \vdots \\ (\sum \alpha_t^{(N)}(i) a_{ij}) b_j(O_{t+1}^{(N)}) \end{pmatrix}$$

$\alpha_t^{(n)}(i) a_{ij}$ is the probability of the joint event that $O_1^{(n)}, O_2^{(n)}, \ldots, O_t^{(n)}$ are observed and state j is reached at time t via state i at time $t-1$. By multiplying the summed quantity by the probability $b_j(O_{nt+1})$, $\alpha_t^{(n)}(j)$ is obtained by accounting for observation $O_t^{(n)}$ in state j.

(3) Termination

$P(O^{(n)}|\lambda) = \sum_{i=1}^{M} \alpha_T^{(n)}(i), \;\; 1 \le n \le N, \;\; 1 \le i \le M$

$$P(O^{(n)}|\lambda) = \begin{pmatrix} \sum \alpha_T^{(1)}(i) \\ \sum \alpha_T^{(2)}(i) \\ \vdots \\ \sum \alpha_T^{(N)}(i) \end{pmatrix}$$

We obtain the desired calculation of $P(O^{(n)}|\lambda)$. The calculations are reduced to the order of N^2T.

A.1.2 Backward Variable

(1) Initialization

$\beta_T^{(n)}(i) = 1, \; 1 \le n \le N, \; 1 \le i \le M$

(2) Induction

$\beta_t^{(n)}(i) = \sum_{j=1}^{M} a_{ij} b_j(O_{t+1}^{(n)}) \beta_{t+1}^{(n)}(j), \quad t = T-1, \ldots, 1, \quad 1 \le i, \; j \le M$

This step is less obvious than that for the forward variable

$$\begin{aligned}
\beta_t^{(n)}(i) &= P(O_{t+1}^{(n)},\dots,O_T^{(n)}|q_t^{(n)}=i,\ \lambda) \qquad \text{(A.1)}\\
&= \sum P(O_{t+1}^{(n)},\dots,O_T^{(n)},q_{t+1}^{(n)}=j|q_t^{(n)}=i,\ \lambda)\\
&= \sum P(O_{t+1}^{(n)},\dots,O_T^{(n)}|q_{t+1}^{(n)}=j,q_t^{(n)}=i,\ \lambda)P(q_{t+1}^{(n)}=j|q_t^{(n)}=i,\ \lambda)\\
&= \sum P(O_{t+2}^{(n)},\dots,O_T^{(n)}|q_{t+1}^{(n)}=j,\ \lambda)P(O_{t+1}^{(n)}|q_{t+1}^{(n)}=j,\ \lambda)P(q_{t+1}^{(n)}=j|q_t^{(n)}=i,\ \lambda)\\
&= \sum a_{ij}b_j(O_{t+1}^{(n)})\beta_{t+1}^{(n)}(j)
\end{aligned}$$

where

$$\begin{aligned}
\beta_{t+1}^{(n)}(j) &= P(O_{t+2}^{(n)},\dots,O_T^{(n)}|q_{t+1}^{(n)}=j,\ \lambda) \qquad \text{(A.2)}\\
b_j(O_{t+1}^{(n)}) &= P(O_{t+1}^{(n)}|q_{t+1}^{(n)}=j,\ \lambda)\\
a_{ij} &= P(q_{t+1}^{(n)}=j|q_t^{(n)}=i,\ \lambda)
\end{aligned}$$

(3) Termination
$P(O^{(n)}|\lambda)=\prod_{i=1}^{M}\pi_i b_i(O_1^{(n)})\beta_1^{(n)}(i)$

A.1.3 Viterbi Algorithm

This is the Viterbi algorithm for finding the optimal state sequence

(1) Initialization

$$\begin{aligned}
\delta_t^{(n)}(i) &= \pi_i b_i(O_1^{(n)}),\ 1\le i\le M \qquad \text{(A.3)}\\
\psi_1^{(n)}(i) &= 0,\ \text{that is, no previous state}
\end{aligned}$$

(2) Recursion

$$\begin{aligned}
\delta_t^{(n)}(j) &= \text{max}[\delta_{t-1}^{(n)}(i)a_{ij}]b_j(O_t^{(n)}) \qquad \text{(A.4)}\\
\psi_t^{(n)}(j) &= \text{argmax}[\delta_{t-1}^{(n)}(i)a_{ij}]\\
& 1\le j\le M,\ 2\le t\le T
\end{aligned}$$

(3) Termination

$$\tilde{P}_n = \max[\delta_T^{(n)}(i)] \qquad \text{(A.5)}$$

$\tilde{P}$ gives the state optimized probability

$\tilde{q}_T^{(n)} = \text{argmax}[\delta_T^{(n)}(i)]$

$\tilde{Q^{(n)}} = \{\tilde{q}_1^{(n)}, \ldots, \tilde{q}_T^{(n)}\}$is the optimal state sequence

(4) Path (state sequence) backtracking

$$\tilde{q}_t^{(n)} = \psi_{t+1}^{(n)}(\tilde{q}_{t+1}^{(n)}), \quad t = T-1, \ldots, 1$$

The array $\psi_t^{(n)}(j)$ keeps track of the argument which maximizes $\delta_t^{(n)}(j)$. A lattice or trellis structure efficiently implements the computation of the Viterbi procedure.

A.1.4 Baum-Welch Learning Process algorithm

1) initialize $\lambda = (A, B, \pi)$

2) Compute $\alpha_t^{(n)}(i), \ \beta_t^{(n)}(i), \ \xi_t^{(n)}(i,j),$ and $\gamma_t^{(n)}(i)$

3) Re-estimate the model $\lambda = (A, B, \pi)$

4) If $P(O^{(n)}|\lambda)$ increases, to go 2

It is desirable to stop if $P(O^{(n)}|\lambda)$ does not increase by at least some predetermined threshold and/or to set a maximum number of iterations

A.2 HMM Matlab Software Package

The Matlab software has inbuilt functions for the Hidden Markov model analysis. MATLAB implements the HMM via five functions described below. Where Q represents the sequence of states, O is the observation sequence, A is the transition matrix and B is the observation matrix (Matlab , 2003)

1) $[O, Q] = \text{hmmgenerate}(\text{length}, A, B)$: Generates a sequence Q of states and a sequence O of observations.

2) $[A_{est}, B_{est}] = \text{hmmestimate}(O, Q)$: Calculates maximum likelihood estimates of transition and observation probabilities.

3) $[A, B] = \text{hmmtrain}(O, A_{est}, B_{est})$: Calculates the maximum likelihood estimates of transition and observation probabilities from the given initial estimates A_{est} of transition matrix and B_{est} of the observation matrix using the Baum Welch algorithm.

4) $Q = \text{hmmviterbi}(O, A, B)$: calculates the most probable sequence Q of states for a given sequence O of observations.

5) $P = \text{hmmdecode}(O, A, B)$: Calculates the sequence P of posterior state probabilities.

A.3 HMM Scaling

The three HMM problems computations involve products of probabilities. For example $\alpha_t^{(n)}(i) \to 0$ as $T \to \infty$. Any attempt to implement the formulae as given will inevitably result in underflow. The solution to this underflow problem is to scale the numbers. Care must be taken to ensure that the re-estimation formulae remain valid. Recall that (Shen , 2008)

$$\alpha_t^{(n)}(i) = P(O_1^{(n)}, \dots, O_t^{(n)}, q_t^{(n)} = i|\lambda) \tag{A.6}$$

$$= \sum_{q_1^{(n)}, \dots, q_{t-1}^{(n)}} P(O_1^{(n)}, \dots, O_t^{(n)}, q_t^{(n)} = i|q_1^{(n)}, \dots, q_{t-1}^{(n)}, \lambda) \cdot P(q_1^{(n)}, \dots, q_{t-1}^{(n)}|\lambda)$$

$$= \sum_{q_1^{(n)}, \dots, q_{t-1}^{(n)}} \left(\prod_{t=1}^{T-1} b_{q_t^{(n)}}(O_t^{(n)}) \prod_{t=1}^{T-1} a_{q_t^{(n)} q_{t+1}^{(n)}} \right)$$

This summation goes to zero quickly as t becomes sufficiently large. The solution is by scaling all $\alpha_t^{(n)}(i)'s$ appropriately. The Baum-Welch algorithm changes by using the modified forward and backward variables (Rabiner , 1989; Shen , 2008). Matlab functions for training the HMM have inbuilt capabilities in the functions. Therefore, no need for further scaling of the variables as matlab functions are utilized in the HMM training in this study.

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \hat{\alpha}_t^{(n)}(i) \cdot a_{ij} b_j(O_{t+1}^{(n)}) \cdot \hat{\beta}_{t+1}^{(n)}(j)}{\sum_{t=1}^{T-1} \hat{\alpha}_t^{(n)}(i) \cdot \hat{\beta}_t^{(n)}(i) / c_t^{(n)}} \tag{A.7}$$

$$\hat{b}_j(O_t^{(n)}) = \frac{\sum_{t=1}^{T} \hat{\alpha}_t^{(n)}(j) \cdot \hat{\beta}_t^{(n)}(j) / c_t^{(n)}}{\sum_{t=1}^{T} \hat{\alpha}_t^{(n)}(j) \cdot \hat{\beta}_t^{(n)}(j) / c_t^{(n)}} \tag{A.8}$$

Appendix B

List of Publications

1. Credit scoring for M-Shwari using hidden Markov model. European Scientific Journal, 12 (15): 176-188, 2016
2. Modeling trust in social network. International Journal of Mathematical Archive, 7(2): 64-68, 2016
3. Trust and distrust: A reputation ratings approach. International Advanced Research Journal in Science, Engineering and Technology, 3(2): 111-114, 2016
4. Consumer lending using social media data. International Journal of Scientific Research and Innovative Technology, 3(2): 1-8, 2016
5. Trust model for social network using singular value decomposition. Interdisciplinary Description of Complex Systems, 14(3): 296-302, 2016