<table>
<thead>
<tr>
<th>Title</th>
<th>On the application of analytical techniques to mobile network CDRs for the characterisation and modelling of subscriber behaviour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Wang, Han</td>
</tr>
<tr>
<td>Publication Date</td>
<td>2018-06-13</td>
</tr>
<tr>
<td>Publisher</td>
<td>NUI Galway</td>
</tr>
<tr>
<td>Item record</td>
<td><a href="http://hdl.handle.net/10379/7401">http://hdl.handle.net/10379/7401</a></td>
</tr>
</tbody>
</table>
On the Application of Analytical Techniques to Mobile Network CDRs for the Characterisation and Modelling of Subscriber Behaviour

A dissertation presented
by
Han Wang BE MSc

to
The College of Engineering and Informatics
in fulfilment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Electrical and Electronic Engineering
National University of Ireland Galway

January 2018

Supervisor: Liam Kilmartin
# Table of Contents

Table of Contents ..................................................................................................................i

List of Figures .........................................................................................................................vi

List of Tables ..........................................................................................................................xi

Acknowledgments ..................................................................................................................xii

Glossary of Terms .................................................................................................................... xiii

Statement of Originality .......................................................................................................... xv

Sponsors ....................................................................................................................................xvi

Abstract ................................................................................................................................... xvii

Chapter 1 Introduction ..............................................................................................................1

1.1 Introduction .........................................................................................................................1

1.2 Dynamically Pricing Mobile Phone Services ................................................................. 3

1.3 Motivations .........................................................................................................................5

1.4 Thesis structure .................................................................................................................6

1.5 Contributions .....................................................................................................................7

1.6 Publications .......................................................................................................................7

1.6.1 Journal Papers .............................................................................................................7

1.6.2 Conference Papers .......................................................................................................8

Chapter 2 Literature Review ................................................................................................... 9

2.1 Background .......................................................................................................................9
2.1.1 Graph Theory ................................................................. 10
2.1.2 Probability Density Functions ....................................... 16
2.1.3 Game Theory ............................................................... 17
2.2 Dynamic pricing .............................................................. 19
2.2.1 Dynamic pricing for mobile phone services ...................... 20
2.2.2 Dynamic pricing in other industries ................................. 25
2.3 Call detail record dataset analysis .................................... 27
2.3.1 Dataset pre-processing .................................................. 27
2.3.2 Traditional statistical analysis techniques ......................... 28
2.3.3 Graph-theory-based social network analysis ...................... 32
2.3.4 Applications of CDR dataset analysis ............................... 37
2.4 Agent-based modelling ..................................................... 41
2.4.1 ABM applied to the Mobile Phone Network Domain .......... 42
2.4.2 Applications of ABM in other domains ............................ 42

Chapter 3 Subscriber Behaviour in a Cellular Network implementing Dynamic Pricing ................................................................. 44

3.1 Introduction .................................................................. 44
3.1.1 Geographic, economic and social structure ....................... 44
3.1.2 Mobile phone services in Uganda .................................. 47
3.1.3 Related work .............................................................. 49
3.2 Dynamic Pricing Service CDR Dataset .............................. 51
3.2.1 Practical Operation of the DPS ..................................... 51
3.2.2 Overview of the CDR dataset ........................................ 52
3.3 Software Tools ............................................................... 55
3.4 Aims of initial phase of research .................................................................55

3.4.1 Removal of “Noisy” CDRs from the dataset ........................................56

3.5 Cell Site Location Estimation .......................................................................56

3.6 Aggregated Network Performance and Subscriber Behaviour ..................60

3.6.1 Cell Site Utilisation ................................................................................60

3.6.2 Daily Temporal Traffic Patterns ............................................................63

3.6.3 Call and subscriber mobility modelling .................................................66

3.6.4 Discount “Chasing” ..............................................................................75

3.6.5 Patterns in subscribers’ calling behaviour .............................................77

3.7 Graph analysis of subscriber behaviour .....................................................81

3.7.1 Subscriber mobility .............................................................................81

3.7.2 Subscriber mobility behaviour at different spatial scales ....................87

3.8 Summary and Conclusions ........................................................................91

Chapter 4 Insights into Social and Economic Behaviours in Uganda using CDR
analytics .........................................................................................................92

4.1 Introduction ................................................................................................92

4.2 Coarse grained regional analysis of subscriber behaviour .......................94

4.2.1 Does Tariff discounting have a bigger impact in poorer regions? ..........94

4.2.2 Is there regional insularity visible in social ties? ........................................95

4.3 Linking subscriber mobility and economic development .......................100

4.3.1 Identification of centres of economic activity ....................................100

4.3.2 Subscriber mobility in the different regions – A closer examination ....103

4.4 Fine grained analysis of subscriber mobility ............................................104
Chapter 4

4.4.1 Mobile travel graph analysis ......................................................... 104
4.4.2 Alternative method for classifying regional development .............. 109
4.5 Summary and Conclusions ................................................................. 110

Chapter 5

5.1 Introduction ....................................................................................... 112
5.2 Structure of the agent based model ..................................................... 113
  5.2.1 Call attempt model ....................................................................... 113
  5.2.2 Subscriber mobility model ............................................................. 117
  5.2.3 Social linkage/network model ........................................................ 118
5.3 Load based dynamic pricing model ....................................................... 122
5.4 Results ............................................................................................... 124
  5.4.1 Model parameter tuning ................................................................. 124
5.5 Summary and Conclusions ................................................................. 129

Chapter 6

6.1 Introduction ....................................................................................... 131
6.2 Alternative dynamic pricing algorithms ............................................. 131
  6.2.1 Random dynamic pricing ............................................................... 131
  6.2.2 Subscriber centric dynamic pricing .............................................. 132
6.3 Simulation Results ............................................................................. 135
  6.3.1 Off-peak discount strategy ............................................................ 135
  6.3.2 Subscriber centric discounting algorithm operation .................... 136
  6.3.3 Comparison of dynamic pricing strategies .................................... 139
6.4 Summary and Conclusions ............................................................... 143
Chapter 7 Conclusions and Future Work ................................................................. 145

7.1 Calling and mobility behaviour of subscribers ............................................... 146

7.2 Social and economic behavioural analysis ..................................................... 147

7.3 Agent-based modelling of subscriber behaviour ............................................ 147

7.4 Comparison of revenue generation capabilities of different pricing algorithms  ......................................................................................................................... 148

7.5 Future work ...................................................................................................... 150

Bibliography ........................................................................................................... 152

Appendix A Journal and conference papers relating to this work .................... 175
List of Figures

Figure 2.1: Examples of (a) an undirected graph and (b) a directed graph...........10

Figure 2.2: The cumulative distribution of the number of calls received on a single day by 51 million users of the AT&T long-distance telephone service in the United States [44].........................................................................................................................................17

Figure 2.3: Basic dynamic pricing system [16]..................................................20

Figure 2.4: LFC pricing strategy flowchart [52]..................................................24

Figure 2.5: Call holding time distribution [75]....................................................28

Figure 2.6: Distribution of call arrival rates averaged over four different days [86].29

Figure 2.7: Travel time distribution for a fictitious group of travellers. The travellers are clustered into three subgroups according to their travel times [96]..............31

Figure 2.8: (a) Degree distribution ($\alpha_k = 8.4$); (b) link weight distribution ($\alpha_w = 1.9$) [10].................................................................................................................................34

Figure 2.9: Definition of persistence [108].......................................................36

Figure 2.10: Examples of communities detected with different methods. The different methods are the Infomap method (IM, red), Louvain method (LV, blue) and Clique percolation method (CP, green). For each method, four examples are shown, with 5, 10, 20 and 30 nodes. The coloured links are part of the community, the grey nodes are the neighbours of the represented community [78, 109]................................................38

Figure 2.11: Community detection in Belgium (top) The communities of the Belgian network are coloured based on their linguistic composition: green for Flemish, red for French. Communities having a mixed composition are coloured with a mixed colour, based on the proportion of each language. (bottom) Most communities are almost monolingual [78].................................................................40

Figure 3.1: (a) Geography of Uganda; (b) administrative regions of Uganda and (c) ethnic diversity within Uganda (source: http://www.wikipedia.org)............45
Figure 3.2: Number of mobile cellular subscriptions in Uganda from 2000 to 2016 (in millions) [178].

Figure 3.3: (a) Number of call attempts through the DPS; (b) number of subscribers opted into the DPS; and (c) average number of calls per subscriber. All are displayed for a nine-month time period post service launch.

Figure 3.4: Location of cell sites in Uganda.

Figure 3.5: Distance error of cell location estimation algorithm.

Figure 3.6: Cumulative Distribution Function of error in location estimate.

Figure 3.7: Cell utilisation across 19 days.

Figure 3.8: Example of the daily call attempt distribution.

Figure 3.9: Call attempt intensity map across 18 weeks.

Figure 3.10: Distribution of call attempts in four regions in Uganda.

Figure 3.11: (a) Distribution of call attempts and (b) visited cells on a log-log scale.

Figure 3.12: Power law fit to (a) the distribution of call attempts and (b) visited cells on a log-log scale (discount range: 50%–60%).

Figure 3.13: Call attempt distribution with truncated power-law fitting.

Figure 3.14: Lognormal fit to the distribution of call attempts (discount range: 50%–60%).

Figure 3.15: Variation of mean and standard deviation of lognormal fit for distribution of call attempts and visited cells (19 days).

Figure 3.16: Variation of mean and standard deviation of lognormal fit for distribution of call attempts and of visited cells (12 days).

Figure 3.17: CDF distributions for the number of call attempts and the number of visited cells.

Figure 3.18: Joint calling and mobility pattern density contour map.

Figure 3.19: Distribution of (a) average discount versus access regularity and (b) average number of calls versus access regularity.
Figure 3.20: The distribution of the regularity of subscribers versus number of subscribers.................................................................79

Figure 3.21: Mean versus standard deviation based on days when subscribers called, for five different activity levels.................................................................81

Figure 3.22: (a) Trajectories of subscribers and (b) trajectories with 90% transparency ...............................................................................................82

Figure 3.23: Clustering of the mobile travel network ........................................83

Figure 3.24: Traffic generated between regions ........................................87

Figure 3.25: Subscribers’ movements at different spatial scales: (a) town level; (b) major town level; and (c) region level ........................................................89

Figure 3.26: Clusters of trajectories by similar routes in Kampala ...............90

Figure 4.1: Change in proportion of calls made from each analysis region during the analysis time period (a) to (e) illustrate calls in which the discount offered on the tariff was in the indicated band and (f) is the overall average across all calls ..........96

Figure 4.2: Change in proportion of callers making calls from each region during the analysis time period (a) to (e) illustrate calls in which the discount offered on the tariff was in the indicated band and (f) is the overall average across all calls ..........97

Figure 4.3: Inter-regional calls for (a) all calls, (b) calls receiving <20% discount, (c) calls receiving 20-40% discount, (d) calls receiving 40-60% discount, (e) calls receiving 60-80% discount and (f) calls receiving >80% discount .................100

Figure 4.4: Concentration of Firms in Uganda Source: World Bank\Uganda Bureau of Statistics (2009/10) ..................................................................................101

Figure 4.5: (a) Home location map, (b) Work location map..........................102

Figure 4.6: Home and work differential map..............................................103

Figure 4.7: Distribution of subscriber average travel distance ..................104

Figure 4.8: Undirected weighted Mobile Travel Graph, (a) North Region, (b) Kampala City, (c) West Region, (d) East Region.........................................................105

Figure 4.9: Undirected Weighted Mobile Travel Graph (strong links only), (a) North Region, (b) Kampala City, (c) West Region, (d) East Region.........................107
Figure 4.10: Neighbourhood connectivity ......................................................... 108
Figure 4.11: CDF of travel distance ................................................................. 109
Figure 5.1: Subscriber calling behaviour ......................................................... 115
Figure 5.2: Distribution of subscriber effect price on demand, $\lambda_i = 7$ .......... 116
Figure 5.3: Distribution of call duration ............................................................ 117
Figure 5.4: Discount factor, $p$ ...................................................................... 118
Figure 5.5: Estimation of subscriber calling location at 10a.m. (longitude and latitude divided into 200 bins) ................................................................. 119
Figure 5.6: Degree distribution ...................................................................... 120
Figure 5.7: Histogram of aggregated $p_g, \forall i$ ............................................. 121
Figure 5.8: Geometric distribution fitting for call frequency $p_g, i = 0.2$ in this case ................................................................................................. 121
Figure 5.9: The relationship between offered discount and cell utilisation ....... 122
Figure 5.10: The estimation of delta cell utilisation .......................................... 123
Figure 5.11: ABM simulation ........................................................................... 125
Figure 5.12: The simulation results of the total number of calls over a day based on different influential of discount factor $\alpha$. (a) $\alpha = 0.1$, (b) $\alpha = 0.5$, (c) $\alpha = 0.8$ .... 126
Figure 5.13: Normalised root mean square error ............................................ 127
Figure 5.14: Samples of the user demand function $\lambda_i \sim \tau(8, 0.7)$ .............. 128
Figure 5.15: Comparison between ABM simulation results and CDR dataset behaviour ......................................................................................... 129
Figure 6.1: The impact of revenue based on discounts (in the hours 2-6a.m.) .... 136
Figure 6.2: (a) Impact of mean discount offered to “normal” subscribers on revenue (b) impact of mean discount offered to high value subscribers on revenue ........ 137
Figure 6.3: Evolution of average $\gamma$ over the total of 50 days for (a) all subscribers (b) “mildly” influenced subscribers and (c) “significantly” influenced subscribers .... 138
Figure 6.4: Comparison of regional revenue generation, (a) North Region, (b) Kampala City, (c) West Region, (d) East Region ................................. 140

Figure 6.5: (a) Normalised revenue versus discount curves for various call duration model parameter ($k$) values and (b) associated mean call duration versus discount curves .............................................................................................................................................. 143
List of Tables

Table 2.1: Example of two-person zero-sum game ........................................... 18
Table 2.2: The Prisoner’s Dilemma game.......................................................... 19
Table 3.1: Regional population statistics ............................................................ 46
Table 3.2: Mobile network regional statistics (source: Uganda Communications Commission) ........................................................................................................ 48
Table 3.3: Example of an individual CDR........................................................... 52
Table 3.4: Statistics relating to the four regions analysed ................................. 54
Table 3.5: The pre-processed CDR data format................................................ 60
Table 3.6: Cell utilisation growth rate from group 1 to group 5 ....................... 60
Table 3.7: Adjacency matrix of the reduced weighted mobile travel graph network 85
Table 3.8: MTG statistics.................................................................................. 86
Table 4.1: Percentage of subscribers call to each region ................................. 98
Table 4.2: MTG network statistics (Strong links only)................................... 108
Table 6.1: Comparison of normalised revenue and equivalent mean random discount ........................................................................................................... 141
Completing this thesis would not have been possible without the help of a great number of people. First of all, I am extremely grateful to my supervisor Liam Kilmartin for his guidance and all the useful discussions and brainstorming sessions, especially during the difficult conceptual development stage. His deep insights helped me at various stages of my research. He gave me constant motivation and great direction to see this doctorate through to the end. My sincere thanks also goes to Dr. Damien Fay for providing his statistical expertise in this thesis and for guiding me to implement the agent based simulation model.

I would also like to thank my parents, Zhengjie Wang and Guangying Han for supporting me in every conceivable way during my two-decade trek through the education system. I can't thank them enough for everything they've done.

During my research I was lucky enough to have the assistance of my industry sponsors Tango Telecom. They provided a valuable dataset to support my research. In particular, I’m grateful to Kieran Kelly for his advice and knowledge.

Finally, I would like to thank the Irish Research Council (IRC) who funded this research.
## Glossary of Terms

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABM</td>
<td>Agent Based Model</td>
</tr>
<tr>
<td>ADSS</td>
<td>Agent Based Dynamic Spatial Simulation</td>
</tr>
<tr>
<td>CDF</td>
<td>Cumulative Distribution Function</td>
</tr>
<tr>
<td>CDR</td>
<td>Call Detail Record</td>
</tr>
<tr>
<td>DPS</td>
<td>Dynamic Pricing Service</td>
</tr>
<tr>
<td>DRC</td>
<td>Democratic Republic of Congo</td>
</tr>
<tr>
<td>EWMA</td>
<td>Exponentially Weighted Moving Average</td>
</tr>
<tr>
<td>FITS</td>
<td>Financial Inclusion Trackers</td>
</tr>
<tr>
<td>GIS</td>
<td>Geographic Information System</td>
</tr>
<tr>
<td>GoS</td>
<td>Grade of Service</td>
</tr>
<tr>
<td>GPS</td>
<td>Global Positioning System</td>
</tr>
<tr>
<td>HLR</td>
<td>Home Location Register</td>
</tr>
<tr>
<td>ICT</td>
<td>Information and Communication Technologies</td>
</tr>
<tr>
<td>IQR</td>
<td>Interquartile Range</td>
</tr>
<tr>
<td>LFC</td>
<td>Learning from Competitors</td>
</tr>
<tr>
<td>LUCC</td>
<td>Land Use and Cover Change</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>MCG</td>
<td>Mobile Call Graph</td>
</tr>
<tr>
<td>MSC</td>
<td>Mobile Switching Centres</td>
</tr>
<tr>
<td>MSISDN</td>
<td>Mobile Station International Subscriber Directory Number</td>
</tr>
<tr>
<td>MTG</td>
<td>Mobile Travel Graph</td>
</tr>
<tr>
<td>MTN</td>
<td>Mobile Telecom Networks</td>
</tr>
<tr>
<td>O&amp;M</td>
<td>Operation and Maintenance</td>
</tr>
<tr>
<td>OPTICS</td>
<td>Ordering Points to Identify the Clustering Structure</td>
</tr>
<tr>
<td>PDF</td>
<td>Probability Density Function</td>
</tr>
<tr>
<td>QoS</td>
<td>Quality of Service</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root Mean Square Error</td>
</tr>
<tr>
<td>RTP</td>
<td>Real Time Pricing</td>
</tr>
<tr>
<td>SIM</td>
<td>Subscriber Identity Module</td>
</tr>
<tr>
<td>SMS</td>
<td>Short Message Service</td>
</tr>
<tr>
<td>TOU</td>
<td>Time of Use</td>
</tr>
<tr>
<td>UBOS</td>
<td>Uganda Bureau of Statistics</td>
</tr>
<tr>
<td>UCC</td>
<td>Uganda Communications Commission</td>
</tr>
<tr>
<td>UPTC</td>
<td>Uganda Posts and Telecommunications Corporation</td>
</tr>
<tr>
<td>USSD</td>
<td>Unstructured Supplementary Service Data</td>
</tr>
<tr>
<td>VLR</td>
<td>Visitor Location Register</td>
</tr>
</tbody>
</table>
Statement of Originality

I hereby declare that the work contained in this thesis has not been submitted by me in the pursuance of any other degree.

Name:
Date:
Sponsors

This work was funded by the Irish Research Council (IRC) and Tango Telecom Limited.
Abstract

This thesis describes work completed on the characterisation and modelling of the behaviour of subscribers in mobile phone networks implementing dynamic pricing by means of Call Detail Records (CDR) datasets analysis. The initial focus of the work was the development of algorithms and visualisation techniques to facilitate an investigation of general subscriber behaviour in the Ugandan mobile phone network in which the CDR datasets was recorded. A cell site location estimation algorithm was developed which provided initial estimates for the location of cell sites which in 80% of all cell sites was within 20km of their true location. A joint calling and mobility analysis showed that the majority of subscribers visited fewer than six cells and that these subscribers usually made fewer than six calls per day. A coarse-graining strategy was also applied to identify the most significant communications corridors between the major urban centres in the country. The second component of this work was an investigation into what insights the CDR datasets could provide relating to aspects of social behaviour and economic activity in Uganda. A methodology for identifying centres of population concentration for residential and work activities was developed. This analysis highlighted a pattern where economic activity appeared to be concentrated around a small number of urban centres. Significant regional insularity in terms of population movement and communication was also observed along with regional behavioural homogeneity, particularly in the populations of Eastern and Western regions of the country compared to behaviour in the region around the capital and in the economically under-developed Northern region. The final component of this work focussed on the development of a novel Agent Based Model (ABM) for the simulation of subscriber behaviour based on patterns observed in the CDR datasets. Results showed that the simulated behaviour observed in the ABM bore strong similarities to that observed in the CDR datasets. The ABM was then utilised as a tool to evaluate the likely levels of revenue generation for a number of different dynamic pricing algorithms. Results obtained using the ABM suggested that all the dynamic pricing regimes investigated would likely result in revenue losses between 6% and 30% with a fixed subscriber base). Additional results also indicated that the scope for using optimisation techniques for revenue maximisation through real time control of dynamic pricing algorithms may be limited.
Chapter 1 Introduction

1.1 Introduction

In recent years, the use of mobile phones, and especially smartphones, has grown rapidly. According to statistics, overall worldwide mobile phone usage reached 4.43 billion in 2015 [1]. This number was predicted to grow to 4.61 billion in 2016 and 4.77 billion in 2017. Of these mobile cellular subscriptions, smartphone subscriptions reached 2.6 billion in 2015, and this figure is expected to grow to 6.1 billion by 2020 [2]. From 2010 onwards, telecommunication companies started to realise that the huge amount of data generated by network components every day as a result of mobile subscribers using the network could provide valuable information on customers’ behaviour, which could in turn help them to improve their services, grow their subscriber base and to increase their revenue streams. Researchers have also observed that these data sources can provide insights into the social and physical behaviour of subscribers at a level which is extremely difficult to identify from any other data source. Of course, the fundamental reason behind this is because most people carry a mobile phone with them wherever they go. Research from Statista Inc. shows that 87% of people always have their smartphone with them and 67% of Americans never even try to unplug from technology, meaning that these people always have either a smartphone, tablet or laptop with them [3]. Hence, data generated by the equipment within their network relating to the interaction of subscribers and their handsets/tablets with mobile networks are a valuable source of insights into people’s calling, mobility and social behaviour.

The particular form of data from telecommunication companies of interest in this work is usually referred to as *Call (or Charge) Detail Records* (CDRs) [4]. Traditionally, these records are generated by virtually every element within a mobile phone network (e.g. Mobile Switching Centres (MSC), databases such as the Home and Visitor Location Registers (HLR and VLR), base station infrastructure etc.) both when a subscriber is utilising a network service (e.g. making a voice call, sending or receiving a short message, sending or receiving mobile data). These records traditionally have
been used for test purposes within the network or to investigate complaints or queries raised by subscribers relating to their service or bills. However, within the last 5-10 years, there has been a significant growth in interest both from within the industry and also within the research community on the potential analysis of these large CDR data sets in order to gain insights on subscribers, their behaviour and on the network’s operation. In tandem with this, we have also seen the evolution of the more general fields of data analytics and techniques and technology to support the concept of “big data” [5]. Initial interest in the analysis of CDR datasets focussed on the application of “traditional” statistical analysis techniques in an attempt to identify high level characteristics relating to subscriber behaviour. For example, several early studies primarily focussed on the fitting of statistical distributions to certain characteristics associated with subscriber behaviours, such as degree distribution or call frequency with some success reported in approximating such distributions using power law distributions [6]. These results suggested, for example, that the majority of subscribers have a small number of contacts with a small but non-negligible number of subscribers having a very large number of contacts. Other early studies focussed on attempting to identify and analyse subscribers’ social networks, such as in [7], from CDR datasets from various countries. A common observation in these studies was the network of inter-subscriber connections determined from the CDR datasets appeared to have a “rich get richer” behaviour (i.e. highly connected nodes acquire more links than those that are less connected) or that such networks were “scale-free” [8] in nature. In another study, the representation of inter-subscriber calls by means of a mobile call graph (MCG) [9] illustrated that weak ties were more important for maintaining the network’s structural integrity whilst in other work it was noted that stronger ties play an important role in maintaining local communication [10].

CDR datasets also often contain geographical information whether coarse-grained (e.g. serving MSC identity) or fine-grained (e.g. serving cell site\base station). When present in a CDR, this provides an additional “dimension” to facilitate investigations which have a geographic or “physical mobility” component. For example, research studies have found a 93% potential predictability in user mobility across the whole user base of a CDR database, when using the entropy of each individual’s trajectory as a measure [11]. The analysis of human mobility patterns present in such CDRs can also help social planners to obtain a deeper understanding of existing or future city
Chapter 1 Introduction

dynamics [12]. Yet another area where the analysis of CDR datasets has been applied is in research relating to the monitoring, prediction and prevention of infectious diseases over time and across geographic areas. For example, several different models based on the mobility of people were developed to predict the spread of epidemic diseases in [13].

Using this brief overview of samples, it should be clear that there is significant interest and potential in the analysis of large mobile phone network CDR datasets in a wide variety of diverse inter-disciplinary fields.

1.2 Dynamically Pricing Mobile Phone Services

The last two decades has seen explosive growth in mobile phone penetration across the globe. In addition, the variety of services which can be accessed from a mobile phone has also diversified significantly with access to voice, messaging and data services from mobile smartphones now being commonplace particularly in economically “developed” countries. Whilst economically developing countries may have lagged in this trend, they are rapidly catching up in terms of smartphone versus feature phone penetration rates and voice versus non-voice service utilisation. In all cases, network operators have had huge capital and operational expenditure in order to establish and grow their network infrastructure in order to attract subscribers so that their subscriber base and revenue streams continue to grow. However, like the Internet, as the user base grows congestion during traffic periods is common issue for mobile networks. In order to address this increase in usage demand, operators typically need to provide additional “resources”. The most common way to solve this problem has been to increase the capacity of the mobile network. However, this results in significant expenditure (both initial capital expenditure and ongoing operational expenditure) and there is always the risk that once deployed there will be significant periods of time where this infrastructure is significantly under-utilised. Cell splitting and frequency re-use is often a second choice for increasing the network capacity. This approach also has capital expenditure implications and results in higher cell densities. In turn this can result in further complexities within the network operation due to an increased overhead relating to managing service hand-off [14] as subscribers move from one cell coverage area to another one.
Chapter 1 Introduction

Another potential way to address the congestion problem is to introduce the concept of dynamic pricing for subscriber access to certain services. In a dynamic pricing environment, subscribers would be charged a different price (or tariff) possibly at different times of the day or in different locations. The introduction of such a tariffing scheme could be used to encourage users to change their use pattern in order to reduce congestion. An additional motivation for using such approaches would be to attract new subscribers or indeed to maintain their existing subscriber base (i.e. prevent subscriber “churn”). Some basic implementation of a Dynamic Pricing Scheme (DPS) may involve implementing fixed but different charges based on “time-of-day” and/or location. However, more sophisticated techniques can also involve a “real time” element whereby the service tariff applicable to the subscriber is potentially continuously changing in time and location. This, for example, could be relatively easily achieved by linking the tariff offered to a subscriber of a DPS to the “real time” cell utilisation [15]. In such a scenario, service pricing is being used as a tool to attempt to adjust or alter the subscribers’ demand characteristics for a service offered by the mobile phone network [16].

Traditionally, dynamic pricing in various forms has been widely used in a variety of other industries. For example, the electricity supply sector has often used this approach to offer customers energy tariffs which vary depending on when the electricity is consumed (e.g. “day” versus “night” rates) [17]. Dynamic pricing is also common in the airline and hotel industry where ticket and room prices are adjusted over time depending on demand [18]. However, apart from some very basic forms (e.g. off- and on-peak tariffs), the use of more sophisticated forms of dynamic pricing is not particularly common in the telecommunication industry. However, since 2010, there has been a significant increase in commercial interest in the deployment of dynamic pricing algorithms for both voice and non-voice services, particular in mobile networks in developing countries of both Africa and Asia. The motivation behind this trend appears to be the dual interests of attempting to modify subscriber demand in order to better manage network capacity and as a marketing differentiator, with a view to minimising subscriber “out-churn” and maximising subscriber “in-churn”. Another interesting characteristic is that many of these deployments have taken place in countries with a very high level of post-paid subscriber penetration, where there is a significant residual penetration level of feature-phones and where the subscriber base
Chapter 1 Introduction

is extremely price-averse (due to general economic under-development). However, very little research has been carried out on the impact and dynamics which are likely to occur when a DPS is deployed in a mobile network, in terms of how it impacts subscriber access to the service and, most importantly from the operator’s perspective, what the impact on service revenue is likely to be.

1.3 Motivations

Whilst the area of CDR dataset analytics is one which is rapidly evolving, the particular dataset made available to this work appears to be unique in that it was captured in a network operating a dynamic pricing paradigm. One of the first motivations for this work was to investigate and develop new analysis and visualisation techniques which would factor in the dynamic pricing component of the dataset in terms of attempting to understand aspects of subscriber behaviour. The dataset also offered the opportunity to apply both more traditional analysis techniques and those developed to consider the impact of DPS in an attempt to gain insights into some economic and geographic characteristics in the country in which the data was captured, namely Uganda. In order to address any ethical issues associated with using CDR datasets, subscribers’ phone number were anonymised for data protection reason prior to any work on the dataset. Whilst similar forms of analysis have been applied in other African and non-African countries (and regions thereof), this dataset offered the first opportunity to apply this form of analysis to the country of Uganda.

The large scale deployment of dynamic pricing paradigms for mobile network services is in its infancy. If such a deployment is not implemented, calibrated and monitored carefully an operator may even lose significant revenues (within very short periods of time) when dynamic pricing is activated. Therefore, another significant motivator of this work was to develop a realistic simulation model which could allow an operator to investigate the potential impact of deploying a DPS in their mobile network. In the case of this work, the accuracy of any developed model would be grounded in the fact that it would be designed from utilising behaviour determined from the analysis of the available CDR dataset. Finally, once a robust model is available, the model could then be used to investigate and compare competing techniques which could be used to
implement different dynamic pricing algorithms from the perspective of their predicted impact on the operator’s revenue stream.

1.4 Thesis structure

Chapter 2 of this provides a detailed review of the relevant literature. It includes an overview of the techniques utilised in the analysis of large scale CDR datasets and the applications to which such analyses have been applied. In addition, it provides a detailed introduction to dynamic pricing and its applications in various fields. The chapter also contains an outline of the use of agent-based models in various settings.

Chapter 3 focusses on the initial set of investigative studies undertaken on the CDR datasets. As it is the country of origin of the CDR dataset used in this study, this chapter opens with a general review of Uganda, including an overview of its geographic, economic and social structure, along with a review of the country’s telecommunications environment, all of which are of relevance in providing context to subsequent work reported in this thesis. The bulk of this chapter reports on a variety of initial data processing algorithms and data visualisation techniques which were developed in order to facilitate initial investigations into the dataset, primarily at understanding both subscriber and network behaviour from a “top-down” (i.e. aggregated) perspective.

The focus of Chapter 4 is to present a study focussed on examining what insights the CDR dataset can provide on human social and economic behavioural patterns in Uganda. By examining the response of subscribers to a service incentivising higher mobile phone call rates through the offering of discounts, economically or socially motivated differences in subscriber behaviour in poorer versus wealthier regions of the country might be identified.

Chapter 5 describes the process of designing an agent-based model (ABM) capable of modelling a wide range of subscriber characteristics when operating in a dynamically priced network. The ABM includes components which simulate subscriber calling behaviour, mobility within the network and social linkages. Chapter 6 then builds on this work by utilising the model to investigate the efficacy of a number of alternative
dynamic pricing strategies and, in particular, their impact on predicted service revenue for the mobile network operator.

Chapter 7 concludes this work by presenting a summary of the completed work, proving a discussion on the main conclusions of the work and outlining some ideas for future work which could be completed in this area.

1.5 Contributions

The following are the major contributions of this work:

- A comprehensive review of the current state of the art in CDR dataset analysis, dynamic pricing technologies, including the use of statistical analysis, graph theory, social network analysis and agent-based modelling
- A set of CDR processing and visualisation techniques to investigate subscribers’ calling and mobility behaviours in the CDR dataset
- A study focussed on examining human social and economic behavioural patterns in Uganda based solely on insights provided from the CDR dataset
- The development of a novel ABM dynamic pricing mobile network simulation platform
- A comparative study, utilising the developed agent based model, of a number of alternative strategies for the dynamic pricing and their likely impact on network operator revenue

1.6 Publications

Several publications have resulted from this work, they include two conference publications and two journal publications.

1.6.1 Journal Papers

Chapter 1 Introduction


1.6.2 Conference Papers


Chapter 2 Literature Review

2.1 Background

In recent years, following the rapid development of the internet and mobile phone technology, the amount of data generated by elements within communication networks relating to the actions and behaviours of subscribers of these networks has grown exponentially. The potential insights which this data offers relating to subscriber behaviour is being recognised and many mobile phone networks are now realising that their most valuable asset are not necessarily their products or services but this data. There is growing interest in and challenges to be addressed when it comes to the processing of this data (often in real time) using techniques and technologies which are often categorised as big data processing/analysis [19-21]. In the mobile telecommunications industry, this data is typically in the form of subscribers’ Call Charge Detail Data Records (CDRs) [22, 23] which have historically been used primarily for billing or trouble-shooting purposes. However, with shrinking margins and increasing pressure to improve revenues, mobile operators have started looking closely at ways to use CDR datasets to their advantage. Basic statistical analyses often form the first step in gaining an overview on some aspect(s) of a CDR dataset. More advanced studies can utilise techniques from the realm of applied graph theory to aid in the analysis of the large complex networks representing social interactions between subscribers (e.g. represented in the CDR dataset by calls or messaging between subscribers). Other key components present in many CDR datasets are temporal and location/geographic information which provide information of “when” and “where” subscribers (or their phones) carried out some interaction with the mobile phone network. Such information, particularly when captured over extended periods of time, can provide valuable information which enable investigations using a whole spectrum of algorithmic and visualisation tools in a variety of fields e.g. subscriber mobility pattern analysis. A natural progression in many cases once the “investigative” phase of a CDR dataset is completed may be to move towards the development of models relating to one or more aspect of subscriber behaviour in order to facilitate modelling.
of past behaviour or prediction of future subscriber behaviour. In the next sections an introduction will be provided to some of the basic analysis concepts and tools in the areas of graph theory, probability density functions and game theory which have found use in prior research related to this work. Later sections in this chapter will examine more broadly related studies in the area of dynamic pricing and CDR dataset analysis in order to provide a detailed review of the state-of-the-art relating to this work.

2.1.1 Graph Theory

A graph can be defined as a collection of nodes connected by lines known as edges [24]. The nodes represent entities in the real world and the edges represent a relationship between the entities; for example, nodes may represent cities and the links the roads connecting them, or the nodes could represent individuals and the links some social linkage e.g. friendships between them. For example, in the realm of a mobile phone network CDR datasets, graphs may be used to represent relationships between individuals or geographic regions with the graph’s connections representing call or messaging activity between the nodes of the graph. Thus, a graph is an abstract representation of the real-world phenomenon and this representation has the ability to reveal relationships between entities which more traditional statistical techniques may not highlight.

Graphical networks can be divided into two types: undirected graphs and directed graphs. An undirected graph is a graph in which there is no directivity in the edges that link the vertices in the graph. Figure 2.1 (a) is an undirected graph with set of vertices \( V = \{v_1, v_2, v_3\} \). The set of edges can be expressed as \( V = \{(v_1, v_2), (v_2, v_3), (v_1, v_3)\} \). An undirected graph can be used to represent symmetric relationships between objects that are represented by vertices.

![Figure 2.1: Examples of (a) an undirected graph and (b) a directed graph](image-url)
Figure 2.1 (b) is a directed graph in which the edges in the graph that link the vertices have a direction. The set of edges in the above graph can be expressed as $V = \{(v_2, v_1), (v_3, v_1), (v_2, v_3)\}$. The edges in a directed graph are a list of ordered pairs which represents the directivity of the edge that links the two vertices.

In graph theory, an adjacency matrix (A) is the most basic form of representation for the graph and it can be used to describe the graph’s structure in a form suitable for computational analysis. For a graph $G$ with $n$ vertices, an $n \times n$ matrix, $A$, can be formed as follows:

$$a_{ij} = \begin{cases} 
1 & \text{if } n_i n_j \in (G) \\
0 & \text{if } n_i n_j \notin (G) 
\end{cases} \quad (2.1)$$

where:

- $a_{ij} = 1$ is when there is an edge from vertex $i$ to vertex $j$,
- $a_{ij} = 0$ is when there is no edge.

The adjacency matrices for the example graphs in Figure 2.1 (a) and (b) are therefore expressed as:

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \quad B = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \quad (2.2)$$

However, in practical applications, the “strength” of the connection between each pair of vertices can vary. The concept of a weighted graph was therefore introduced to quantify this strength, in which a weight $n_{ij}$ is associated with the link $n_i, n_j$. Weighted graphs are widely used in the analysis of complex networks [25]. For instance, in computer networking, the weights between two hosts might represent the amounts of data exchanged or, in the case of a mobile phone network, the weights might represent the number of calls or messages between two geographic regions of a country. In graph theory, a connected component of an undirected graph is a maximal set of nodes in which any two nodes are connected by a path.
Chapter 2 Literature Review

In complex networks, one of the most widely studied topological features is the degree distribution [26]. In undirected networks, the degree of a node is the number of connections which this node has. For example, in Figure 2.1 (a), node $v_1$ has a degree of two. Hence, for an undirected node, the degree of node $i$ from the adjacency matrix $A$ is:

$$d_i = \sum_j a_{i,j} \quad (2.3)$$

In directed graphs, the degree of a node is calculated separately as an in-degree $d_{in}(i)$ and an out-degree $d_{out}(i)$, based on the direction of the link. The degree distribution of an undirected network, $p_{deg}(k)$, can be defined as:

$$p_{deg}(k) = \frac{n_k}{n} \quad (2.4)$$

where:

- $n_k$ is the number of nodes in the network of degree $k$,
- $n$ is the size of the network.

Although the measure of degree distribution (i.e. a histogram formed by considering the degrees of all the nodes in the network) is relatively simple, it often can provide significant clarity on the network structure. In the real world, different types of networks generally have very different degree distributions. One of the more commonly observed forms of networks are scale-free networks, in which the degree distribution follows a power law distribution [27]:

$$p(k) \sim k^{-\alpha}, \quad \alpha > 1 \quad (2.5)$$

where:

- $p(k)$ is the probability of a node in the network having degree $k$, 

• \( \alpha \) is the degree exponent.

Network heterogeneity is a topological feature of complex networks. It is usually measured by the variance of the degree distribution.

A representation of the World Wide Web, where the nodes are the individual web pages and the links are URLs, is a commonly cited example of a network exhibiting scale-free characteristics. Another example of a scale-free network [8, 28], which is often quoted in the literature, is a representation of research collaboration where scientists are the nodes of the network and the co-authorship of scientific papers are the basis for the links. In the arena of mobile network CDR datasets, it has also been shown that scale-free network behaviour can be observed in graphs where individual subscribers form the nodes and the calls made between subscribers form the basis for the links [29].

For real large-scale graphs as might be commonly generated from a CDR dataset, the size of the network is usually large with the network having a complex structure which results in such networks being impractical to visualise using plots like that in Figure 2.1. Therefore, a set of summary statistics or quantitative measures are often used to describe and compare networks [30]. There is a huge range of such statistics that have been utilised in reported works in this area and the following paragraphs introduce some of the more commonly encountered measures.

The diameter of a graph is the “longest shortest path” between two vertices. In other words, the diameter of a network [31] is the length of the shortest path between the most distanced vertices. Clustering and centrality are another two common measures utilised in complex network analysis. In graph theory, a clustering coefficient is a measure of degree to which nodes in a graph tend to cluster together [32]. There are two versions of the clustering measure [33], namely global and local clustering coefficients. In a graph \((N, g)\), which consists of a set of nodes \(n = \{1, \ldots, n\}\) and an \(n \times n\) matrix \(g\), the global clustering coefficient \(Cl(g)\) is defined by
where a connected triplet is defined to be a connected subgraph consisting of three vertices and two edges. The local clustering coefficient is defined on an individual node basis. The individual clustering coefficient for a node $i$ is:

$$Cl_i(g) = \frac{\text{number of triangles connected to vertex } i}{\text{number of triples centered at } i} \quad (2.7)$$

As an alternative to the global clustering coefficient, the average clustering coefficient is the average of the local clustering coefficients of all vertices.

$$Cl^{Avg}(g) = \frac{1}{n} \sum_i Cl_i(g) \quad (2.8)$$

Within graph theory and social network analysis, the measurement of a vertex’s centrality is used to determine the relative importance of this vertex within the graph e.g. how well-used a road is within an urban network. There are many ways to measure the network centrality [34]. The most basic one is the degree centrality which is the node degree divided by $n - 1$ [35]. Therefore, the degree centrality for a node is between 0 and 1. Closeness is another basic measure in graph theory, where the closeness of a node tracks how close a given node is to any other node [36].

$$\frac{n - 1}{\sum_{j\neq i} l(i,j)} \quad (2.9)$$

where:

- $n$ is the number of nodes in the graph,
- $l(i,j)$ is the distance between $i$ and $j$.

Betweenness is another centrality measure of a vertex within a graph. The betweenness of a graph takes into account the connectivity of the vertex’s neighbours [37]. The
vertex will get a high value if it, for example, acts as a bridge between clusters. The betweenness of a vertex \( v \) in a graph \( G := (V, E) \) with \( V \) vertices are computed as follows:

\[
Betweenness(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}
\]  

(2.10)

where:

- \( \sigma_{st} \) is total number of shortest paths from node \( s \) to node \( t \),
- \( \sigma_{st}(v) \) is the number of those paths that pass through \( v \).

In graph theory, a graph can be categorised by density [38]. A dense graph is a graph in which the number of the edges is close to the maximal of edges whilst a sparse graph is a graph with relatively few edges. For an undirected graph, the graph density is computed as:

\[
D = \frac{2|E|}{|V|(|V| - 1)}
\]

(2.11)

For a directed graph, the graph density is defined as:

\[
D = \frac{|E|}{|V|(|V| - 1)}
\]

(2.12)

where:

- \( E \) is the number of edges in the graph,
- \( V \) is the number of vertices in the graph.

The density range is between 0 and 1.
2.1.2 Probability Density Functions

A key aspect of traditional statistical modelling is the choice of the correct probability density function (PDF) [39] to represent the behaviour observed in the captured data, for example in this work this may be some aspect of the CDR dataset. In prior work in this area [40-42], numerous studies have focussed on fitting PDFs to various characteristics, but most commonly call or message frequency or the degree distribution of some graph representation of the CDR based data. For example, it was observed that the network formed by inter-subscriber voice calls exhibited scale free behaviour and as such its degree distribution followed a power law, also often referred to as a heavy-tail distribution [40, 43]. In mathematical terms, a quantity \( x \) obeys a power law if it is drawn from a probability distribution of the form:

\[
p(x) = ax^{-\alpha}
\]

(2.13)

where:

- \( a \) is a constant parameter,
- \( \alpha \) is the power law exponent.

Another example, as illustrated in Figure 2.2, highlights the cumulative distribution (CDF) of voice telephony calls which was shown to follow a power law distribution in [44].

Similarly, a lognormal distribution has a natural connection with a power law distribution and it has been proposed as a possible alternative to power law distributions in many fields [45]. A random variable \( X \) has a lognormal distribution if the random variable \( Y = \ln(X) \) has a normal distribution.

Geometric distributions have also been used in other aspects of CDR dataset analysis. For example, the distribution of the number of calls to each of the subscribers’ contacts [39] can be well fitted using a geometric distribution of the form:

\[
p(r; p) = p(1 - p)^{r-1}
\]

(2.14)
where $0 < p < 1$. This equation expresses the probability of having to wait exactly $r$ trials before the first successful event if the probability of a success in a single trial is $p$ [39].

2.1.3 Game Theory

Game Theory is a form of strategy theory which is used for interactive decision making [46, 47]. In a game, several rational players are involved, under conditions of both conflict and cooperation. The approach underlying game theory is that each player will seek to optimise their own decisions at the expense of the other players. The strategies used by the players can be categorised into two types, namely pure or mixed strategies. In a pure strategy, the player specifies a particular move at every stage of the game with complete certainty [48]; in a mixed strategy, there is a certain amount of randomisation in at least one of the moves during the game. The result of the completion of one or more moves is the outcome; the payoff is the amount received for a given outcome. A payoff usually represents the profit each player receives for a specific move. In all cases, the payoff must reflect the motivations of the particular player. During the game, the players may cooperative with others in order to achieve
the maximum payoff and this is known as a cooperative game. In contrast, a non-cooperative game is one in which the players make decisions independently. A hybrid game is one which has elements of both cooperative and non-cooperative games.

2.1.3.1 Zero-sum games

In a zero-sum game, the total benefit to all players in the game, for every combination of strategies, always adds to zero. An example is given below of two competing companies, A and B; each has three strategies for promoting sales performance. Table 2.1 shows the payoff matrix for company A.

<table>
<thead>
<tr>
<th></th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>Row Minimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>A2</td>
<td>-1</td>
<td>4</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>A3</td>
<td>5</td>
<td>-2</td>
<td>0</td>
<td>-2</td>
</tr>
<tr>
<td>Column Maximum</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2.1: Example of two-person zero-sum game**

The best strategy for company A is to **maximin** the payoff, using strategy 1, which can guarantee a payoff of at least 2. For company B, the best strategy is to find the **minimax** payoff; this is strategy 3, which can guarantee a loss to A of at most 2. In this game, the maximin = minimax. This is the saddle-point of this payoff matrix. This pair of strategies (A1, B3) is also known as a solution of the game in a pure strategy approach and means that no player can benefit by unilaterally shifting from their saddle-point strategy.

2.1.3.2 Non-zero-sum games and the Nash equilibrium

The non-zero-sum game is different from the zero-sum game in that the total of gains and losses for all players is not equal to zero. “The Prisoner’s Dilemma” is a classic example of this type of game. Two prisoners are considered; there is no physical evidence to convict either one so the prosecutor seeks a confession. Each prisoner has two possible strategies: confess or do not confess. The consequences in terms of payoff are quantified in prison years and the payoff matrix is shown in Table 2.2.
Chapter 2 Literature Review

Table 2.2: The Prisoner’s Dilemma game

<table>
<thead>
<tr>
<th>Prisoner 1</th>
<th>Confess</th>
<th>Do Not Confess</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confess</td>
<td>-5, -5</td>
<td>-1, -15</td>
</tr>
<tr>
<td>Do Not Confess</td>
<td>-15, -1</td>
<td>-3, -3</td>
</tr>
</tbody>
</table>

In this game, the information about both strategies and payoff is complete; each player knows the available strategies and the payoffs from the intersection of all strategies. Strategies are chosen by the two prisoners simultaneously. For Prisoner 1, \(-5 > -15\) and \(-1 > -3\) and thus the optimum choice for Prisoner 1 is the dominant strategy to confess, regardless of the choice of Prisoner 2. Similarly, Prisoner 2 will also choose to confess. The solution of \((Confess, Confess)\) in this game is a Nash equilibrium\(^49\). A Nash equilibrium is a combination of strategies in which no player in the game can improve their payoff by picking a different strategy based on the assumption that the other player will adhere to the same strategy\(^50\).

Game theory and the concept of Nash equilibrium are now widely used in fields such as computer science, economics, political science and international relations\(^47, 49-51\). In recent years, game theory has also been applied to telecommunications systems, for example, with the aim of optimising the company’s strategy with respect to the strategy of its competitors in terms of maximising revenue\(^52-54\).

2.2 Dynamic pricing

As differentiated from “static” or flat-rate pricing, a dynamic pricing service (DPS) offers the possibility of re-setting pricing, typically to reflect levels of demand\(^55, 56\). This yield management system is used to balance supply and demand and to maximise profit margins. The travel industry and the electricity industry are the most widely known users of dynamic pricing; however, in recent years, this strategy has also been introduced into mobile network services.
2.2.1 Dynamic pricing for mobile phone services

In 2000, Fitkov-Norris and Khanifar [16, 57-59] proposed one of the first dynamic pricing algorithms for mobile network services. They proposed that an increase in price could shorten the duration of calls and reduce the numbers of call attempts by subscribers. This algorithm is a self-regulated model which regularly checks the load of the system to decide whether the network requires a price adjustment ($\Delta P$). $\Delta P$ is defined as the increment in the tariff which is calculated in real time and is positive when the load on the system is too high. A change to a higher price is therefore expected to discourage subscribers and to force them to wait for a short period before making further calls. Conversely, a negative $\Delta P$ means that network resources are under-utilised and results in a lower price in an attempt to encourage subscribers to make their calls. Figure 2.3 illustrates the basic self-regulated dynamic pricing algorithm which was proposed.

![Figure 2.3: Basic dynamic pricing system](image)

In such a dynamic pricing environment, the subscriber demand model is combined with a deterministic function $D(t)$ and a random demand $S(t)$. $D(t)$ is a “time of day”
function for the prediction of traffic and \( S(t) \) is a stochastic events simulator e.g. such as for the simulation of emergency calls.

The main objective of a dynamic pricing strategy is to use the price to adjust the demand from subscribers. This effect of the price on the demand model uses an exponential function [57]:

\[
Q_x = Ae^{-\beta p_{Q_x}}
\]  

(2.15)

where:

- \( Q_x \) is the quantity demand of goods \( X \),
- \( p_{Q_x} \) is the price of goods \( X \),
- \( A \) is the demand shift constant,
- \( \beta \) is the demand elasticity coefficient.

In this equation, parameter \( A \) is the demand shift constant which is determined from historical data. It reflects the way in which the demand changes due to the effect of the time of day and therefore can vary over time. However, historic data also introduces the concept of pricing bias. The model is corrected by taking the difference between the peak or off-peak price in the real system and the dynamic price in the modelled system. In addition, a classical problem is the question of how to set the price when the dynamic price drops below the historic off-peak price. Fitkov-Norris and Khanifar proposed that a fixed-line substitution effect should be taken into account using an exponential function. The modified version of Equation 2.15 then becomes:

\[
Q_x = A(t)e^{-\beta(p_{bias}-p_{dynamic})} + E(p_{bias} - p_{dynamic})^{-\beta}
\]  

(2.16)

where:

- \( E \) is the off-peak shift constant,
- \( \beta \) is the demand elasticity coefficient.
As mentioned above, the price cannot be lower than the off-peak price. Fishburn and Odlyzko [60] proposed a further solution to address this problem. If $\lambda$ is the current incoming call rate and $\lambda^*$ is the optimal incoming call rate, a constant price $p_0$ is charged if $\lambda < \lambda^*$; alternatively, if this is not the case, a user will be charged a dynamically calculated price $p(t)$. Based on this solution, these authors proposed the calculation of the ratio between the dynamic price $p(t)$ and the fixed price $p_0$ to determine the demand from each type of user. This results in the following demand function [60]:

$$D[p(t)] = e^{-\frac{(p(t)-p_0)^2}{p_0}} \text{ for } p(t) \geq p_0$$  \hspace{1cm} (2.17)

$D[p(t)]$ represents the percentage of users who will accept the current price. Since $D[p_0] = 1$, this means that $p_0$ is accepted by all users.

The dynamic pricing function $p(t)$ is dependent on the new call arrival rate, the retry call arrival rate and the optimal call arrival rate. The price at time $t$ is calculated by (2.18).

$$p(t) = D^{-1}(\min(\frac{\lambda_n^*}{\lambda_n(t) + \lambda_r(t)}, 1))$$  \hspace{1cm} (2.18)

These two user demand models described above are widely accepted; however, they do not take into account ways of utilising the price to control the quality of service. Viterbo and Chiasserini proposed the inclusion of this factor into an alternative user demand function [61, 62]. Their proposed demand function for QoS control is shown in (2.19):

$$D(p, Q) = e^{(-\alpha p + \beta Q)}$$  \hspace{1cm} (2.19)

where:

- $p$ is the price per unit time,
- $Q$ is the quality of service, which in this case is the call success rate,
Chapter 2 Literature Review

- $p_b$ is the call blocking rate.

Thus, the quality of service $Q$ is equal to $1 - p_b$. As in (2.19), $\alpha$ and $\beta$ are elasticity parameters and must be determined using appropriate market research.

Wallenius and Hamalainen have also presented a pricing model which includes QoS guarantees [63], as given by (2.20):

$$P_b = A_{ij}(e - e^{-Bx}).T_l$$  \hspace{1cm} (2.20)

where:

- $P_b$ is the unit of price of the call,
- $A_{ij}$ is the linear price factor for each traffic and subscriber class,
- $T_l$ is the traffic load,
- $B$ is a linearity parameter,
- $x$ is a QoS attribute, which can be either bandwidth, end-to-end delay or bit error rate.

The majority of DPS-related research has focussed on modelling the relationship between price and demand from users. However, such models do not take into account the fact that user behaviour may also be affected by the pricing offered by competitor networks operating in the same geographic area. Mansoori proposed an approach which integrates this effect using a game theory based “Learning From Competitors” (LFC) model which updates the price by learning from the competitors’ price [52]. LFC increases or decreases the price depending on the competitor’s price and the profit in a previous cycle. Figure 2.4 shows the flowchart used in LFC strategy.
In Figure 2.4 above,

- Profit represents the service provider’s (SP) profit,
- CProfit represents the competitor’s profit,
- Price represents the SP’s price,
- CPrice represents the competitor’s price.

In this scheme, a competitor is first selected randomly from the other service providers. Then, the model compares the service provider’s profit (Profit) with competitor’s profit (CProfit). If the service provider’s profit is greater than the competitor’s profit then the price difference from the competitor’s price is increased; if the opposite is true, then the price difference from competitor’s price is reduced. The amount by which price is changed is defined as $\alpha \times Price$, where the default $\alpha = 0.1$. In general, an increase in the factor $\alpha$ results in more fluctuations in price but a faster convergence towards the final price, which corresponds to the equilibrium point in game theory.
Chapter 2 Literature Review

2.2.2 Dynamic pricing in other industries

Although the deployment of dynamic pricing services is a relatively recent occurrence in the area of mobile telecommunications, it has been commonly used in many other industries and hence there exists a body of research related to its application in these domains.

One of the most well-known of these is the use of dynamic pricing in the airline industry [18]. Airline companies rely heavily on computerised systems to adjust ticket prices in real time. The key criteria for price changes are based on the number of seats remaining and the amount of time before the flight departs [64-66]. Hence, a passenger can usually get a cheaper deal on flight tickets in the off-season. In addition, a passenger can often buy a flight ticket extremely cheaply at the very “last minute”, since there is little time left before the flight departs, and in this case the airline company wishes to sell this seat as soon as possible. The benefits of using dynamic pricing in the airline industry are obvious; the use of dynamic pricing to constantly update ticket prices is likely to maximise the company’s potential profits. This system can also clear slow-moving inventory using reductions in ticket prices. Similarly, the hotel industry also changes room prices dynamically based on the occupancy rates of a hotel and the time of year [67] and this is generally considered part of the hotel’s revenue management system [68, 69].

Dynamic pricing has also become a common strategy in the retail industry; the clothing retail market often uses a time-based pricing strategy. For example, during the off-season discount periods or clearance sales, the price for a piece of clothing may fall to a third of the “in-season” price. Some customers have changed their buying behaviour accordingly, in order to accommodate this dynamic pricing [70]. During the 2008 fall/winter season, the fast-fashion retailer Zara introduced a new dynamic pricing system in all its Belgian and Irish stores. This new DPS service increased clearance revenues for Zara by about 6% [71]. In food retailing, DPS is also used for perishable foods, since consumer purchasing demand may change based on the quality of the food, that is, its shelf-life [72].

Another industry which has traditionally used DPS widely is the electricity industry. A smart meter is an electricity usage monitoring device that records electricity
consumption and automatically sends usage data to the power supplier companies. This new device forms a part of a smart grid and it can accurately measure customer power consumption in real time, in terms of both time and amount. Based on these records, a dynamic price is charged for each customer depending on the time of the day the electricity is used. This DPS is known as time-of-use (TOU) pricing [51]. TOU is a long-term pricing strategy and generally updates the price on a daily basis. There is also a real-time pricing (RTP) strategy, which defines the price over the short term in order to directly reflect the wholesale price of electricity [73].

A dynamic pricing strategy has also been considered for use as a congestion control system for packet-based networks such as the Internet. There are two main DPS models for packet-based networks: the smart market and shadow pricing models. The smart market approach was first introduced by Varian and MacKie-Mason [74]. This approach adjusts the price for each packet the user sends, based on the congestion in the network. Each packet has a ‘bid’ field, which is populated according to the significance level of the packet. For example, less important packets, such as email packets, are given a low bid price, whereas real-time voice or video packets are assigned a higher bid price. At any given time, the network only accepts and charges for packets with a bid price greater than the current cut-off price. The cut-off price is calculated dynamically based on congestion measures within the network. Any packet containing a bid that does not meet the required price will be rejected or queued in buffers. In packet-based networks, shadow pricing is another well-known dynamic pricing concept. This system was proposed by Gibbens and Kelly [53]. In a shadow pricing framework, users are given sufficient information to enable them to decide whether or not to submit packets. This price information is sent using “marks”. A “mark” represents a certain fixed price that the user will have to pay for the packet, and it is only sent when the network is congested. Thus, when a “mark” is sent, this alerts the user that the network is congested and hence payment will be required for the transmission of subsequent packets; the decision is then made by the sender, according to their requirements. If the communication is not very urgent, the user may wait until the network becomes less congested and the cost is less. On average, the more congested the network is, the more expensive it is to send a packet [14].
2.3 Call detail record dataset analysis

There is a growing body of research which has focussed on the data processing algorithms used to complete the analysis of the large datasets of CDRs generated by mobile networks. Research in recent years in this field highlights that the analysis of CDR datasets has primarily focussed on three major areas. The first of these is the application of traditional statistical analysis techniques where the primary aim of the analysis is an attempt to model or understand subscribers’ calling and/or mobility behaviours, for example [23]. A second popular area of research area is social and spatial network analysis based on graph theory; this approach has been used for applications such as churn prediction and urban planning [75-77]. The third area of research is the modelling of subscriber behaviour when interacting with mobile networks. This technique has played an increasingly important role as, for example, telecommunication companies are keen to use this type of tool to test new price plans before deploying these to their real mobile network in order to avoid risk [78-80].

2.3.1 Dataset pre-processing

In any form of analysis of mobile network CDRs, some form of pre-processing of the raw CDRs is required, for example, in order to separate different categories of activities or subscribers, or to remove records which are errored or are in some way corrupted such that they correspond to “noise” in a dataset. However, the actual nature of the pre-processing to be applied to a dataset depends greatly on the form of analysis being undertaken. For example, Onnela et al. [81] proposed the filtering out of calls involving other network operators in order to minimise any bias introduced due to only having access to CDRs from a single operator’s network. Other filtering is more straightforward and is used to remove invalid records (“noise”), which can often be clearly distinguished, for example by the fact they contain invalid numbers (MSISDN). Certain types of filtering, such as that proposed by Dasgupta et al., is based on heuristics; for example, calls made between the same parties within five seconds of each other are filtered out since these are likely to be simply a single call which was interrupted by being accidentally dropped by the network [75].
2.3.2 Traditional statistical analysis techniques

To understand the calling behaviour of subscribers, an initial step will often be to analyse the distribution of call attempts. The probability distribution function (PDF) of call attempts in a mobile network typically appear to exhibit characteristics of a power-law-like distribution, meaning that most subscribers make a few calls per day while a small number of subscribers make very large numbers (e.g. hundreds) of calls [41].

The modelling of call holding time can give operators a better understanding of the characterisation of the traffic in a mobile network [82-84]. Figure 2.5 shows a typical distribution for call durations. The plot shows that most of the calls in a mobile network are short-lived and that only a small number of calls last for an hour or more. The peak in this plot highlights the fact that callers generally finish a phone call within 60 seconds [75].

Figure 2.5: Call holding time distribution [75]

The inter-event time is defined as the interval between two mobile phone calls made by the same user. Candia et al. demonstrated that the distribution of inter-event times of consecutive calls is heavy-tailed, which is often referred to as the dynamics of spreading phenomena [85]. The calling activity pattern is highly heterogeneous in the
same way as many other human activities. For example, some users only make one phone call per week while others may make hundreds of calls in a single day.

The call arrival rate is another important distribution in mobile networks. This captures the traffic at each hour of the day and this distribution is often used when building simulation models of network operation or subscribers’ behaviour. Figure 2.6 shows the normalised average call arrival rates calculated using five-minute slots [86]. This dataset was collected from hundreds of cell sectors of a US CDMA-based cellular operator which consists of tens of millions of calls and billions of minutes of talk time.

Figure 2.6: Distribution of call arrival rates averaged over four different days [86]

Not unexpectedly, Figure 2.6 clearly shows that there are two distinct periods in terms of calling behaviour, namely “day” and “night”, which have higher and lower call arrival rates respectively. The arrival rates also exhibit a spike in the morning, between around 6 am and 8 am, and late in the evening, around 9 pm to 11 pm. These are the transition periods between the “day” and “night” periods. As the call arrival rates are calculated using a five-minute slot, a relatively short time interval, Figure 2.6 indicates that the stationary timescale for the system is about one hour. This is shown even during the transition hours since the mean arrival rates do not vary significantly during each hour. In addition, the plot indicates that the weekday and weekend call arrival
Chapter 2 Literature Review

rates show different trends. This is not surprising, since the majority of subscribers will not only have different underlying behaviours at the weekend but also will have special tariffs which incentivise calls at the weekend [86].

Understanding the mobility patterns of subscribers is one of the most important parts of CDR dataset analysis since this can be applied to various application spaces e.g. traffic forecasting [87] and the spread of biological [88-90] and mobile viruses [91]. A research study by Barabási shows that human trajectories have a very high degree of temporal and spatial regularity. Subscribers have a very high probability of returning to locations they frequently visit [92]; other research has shown that human mobility behaviour is 93% predictable [11]. Zang and Blot examined mobility data in cellular networks [93], and showed that 96% of users visit fewer than 40 cells in a month. Moreover, those users make or receive 75% of the total number of calls in the network. This means that the other 4% of users make or receive 25% of the total number of calls indicating that the highly mobile users tend to make and receive more phone calls than static users. Space and time are two other essential factors that can reflect the mobility of users since every activity takes place at a particular time and place. Geographical distributions can also be used to interpret users’ daily movements, as shown in [94]. The distribution of the typical distance travelled by a user against the number of subscribers who travelled those distances indicates that the population initially grows quickly as the distance increases. However, the population begins to decrease when the travel distance becomes greater than 2 km. It then maintains a steady low-level population with further increases in travel distance. The CDF of the distance versus the population shows more than 95% of users have a typical travel distance of less than 10 km. Hence, it can be deduced that most people travel a modest distance every day. Data for subscriber travel distances can also generate additional useful information in that this reflects information about the size of the city in which the data was recorded and also can contain correlations with the complexity of the city’s transportation systems [94]. The correlation between phone usage and movement radius has also be investigated by using linear regression. Yuan and Raubal showed that phone usage and average radius of movement had a significant positive correlation relationship. Conversely, a negative correlation between phone usage and maximum movement radius was observed [95]. Analysis of subscribers’ mobility can also be used as a transportation management tool. As discussed above, it is difficult to
get information on the location of users from their mobile devices unless they are willing to install an application on their mobile phones to report their GPS information on a frequent basis. However, users generally turn off GPS modules to save power and, with current technology, it is not practical to leave GPS modules active in handsets all day. As an alternative, however, large CDR datasets allow for other approaches to the collection of users’ location information (e.g. based on serving cell site or some form of network location area identifier). For example, as shown in [96], these coarser grained location indicators can help in identifying how many travellers moved from a given starting point to the same end point with the mean travel time also being available. Information on the number of travellers and their travel times was then used to infer the modes of transportation of travellers. Figure 2.7 shows the distribution of the travel time of a fictitious group of travellers. The travel times have various ranges and the number of travellers is different within each range; it should therefore be possible to divide the travellers into three subgroups based on density. Each subgroup can then be associated with a different mode of transportation (such as car, public transit or walking) [96].

![Travel time distribution for a fictitious group of travellers. The travellers are clustered into three subgroups according to their travel times [96]](image)

In addition, analysing subscribers’ calls and mobility behaviours, there are many other different statistical analyses possible by means of investigations based on CDR dataset analysis. Zang and Bolot proposed the classification of users into a dominant type of...
activity based on which service they use for more than half of the total service in a day. If a user does not have a dominant service type, they are classified as a “balanced user” [93]. The results show that most subscribers were “voice-dominant” users with the second and the third most prevalent classification being data-dominant users and SMS-dominant users respectively. Zang and Bolt further examined location entropy and demonstrated that users’ current location could be determined by examining their previous location entropy. Location entropy can provide information on whether or not a user has visited a specific cell in the past. This approach can be implemented by computing the entropy of the locations (cells) visited by each user and the conditional entropy of a user’s location given their previous N locations [93].

2.3.3 Graph-theory-based social network analysis

Applied graph theory and associated visualisation techniques have also been applied to the analysis of large CDR datasets. Complex network analysis of this nature has historically been used in a wide range of other fields such as the analysis of biological neural networks, social networks and communication networks [7, 97-99]. When analysing CDR datasets, the network nodes can be either a subscriber or a location and they may occupy a defined position in a Euclidian space. For example, in some CDR datasets the subscriber’s location when they make a phone call or the GPS location for each cell site in that mobile network will be known. In such situations, geographic space is important as the network topology alone does not contain all the system information. Therefore, an important consequence of spatial networks is that there is a cost associated with the length of edges [100-103]. This spatial property is very useful for regional analysis because the distance between two subscribers or two cell sites reflects the connectivity level between these two entities.

The most widely used graph-theory-based representation applied to mobile phone network CDR datasets is Mobile Call Graph (MCG) analysis [42, 104]. A Mobile Call Graph, G, consists of a pair (V, E) where V is the set of mobile subscribers and E is a set of “linked” vertex-pairs from V. The edge between any two nodes indicates that there exists a relationship between the two subscribers represented by the nodes, as evidenced by CDRs relating to voice calls or messaging between these two subscribers. There are many ways for such a graph to be constructed from the data within a CDR dataset. The resultant graph representation can be either directed or
undirected, based on the research assumption, as discussed in the following paragraphs.

The most common approach to constructing a MCG is to connect two subscribers with an undirected link if they have communicated at least once, that is, either user $i$ called user $j$, or $j$ called $i$ [81]. However, many of these calls are one-way, which means user $i$ called user $j$ but not vice versa. Such cases often reflect specific interaction scenarios between subscribers, for example when a customer service centre attempts to reach a person. Thus, in many cases, these types of links could be viewed in many ways as “meaningless”; a more useful graph can filter out one-off events by taking only the pairs of users who have made at least one reciprocated pair of calls, that is, $i$ has called $j$ and $j$ has called $i$ [10, 75, 81, 105]. An undirected graph is a simplification of a directed graph. In a directed graph, the initiator of the call is taken into account. A link $e = (i,j)$ is considered to be directed from $i$ to $j$ if a phone call has been made from $i$ to $j$. The degree of a mobile call graph usually has a skewed distribution, as shown in Figure 2.8(a), illustrating that the majority of the users communicate only with a few individuals although a small number of users communicate with large groups of individuals [10]. This distribution would initially appear to follow a power law distribution; in particular, this power law distribution (where the estimated exponent value $\alpha_k$ is equal to 8.4) appears to fit the data substantially better than an exponential distribution [97]. This exponent value is much higher than the value measured for a similar analysis carried out on landline networks [106]. This fast “decay” may reflect the fact that many people are not generally always co-located with their landline telephone and hence are often unable to answer calls in time. This is typically not the case for calls to a mobile handset which is generally co-located with an individual. This distribution is also different to the power law distributions observed in email networks; one probably reason for this dissimilarity may be due to the fact that a single email can be sent to more than one recipient. Therefore, this fast-decaying power law distribution reflects the fact that the mobile phone network is a well-connected, “person-to-person” communication network. A weighted mobile call graph is a generalisation of either the basic directed or undirected graph discussed above. In this case, the weights between nodes quantify the strength of the relationship between two subscribers; for example, this could be based on the call frequency or call volume [75].
Figure 2.8(b) shows an example of a link weight distribution (i.e. tie strength distribution) which when fitted with a power law distribution exhibits an exponent of $\alpha_w = 1.9$. The strength of the tie is measured based on the number of reciprocated pair of phone calls between two users. This indicates that most users make calls lasting only a few minutes and that a small number of users spend hours talking to each other.

However, graphs of this nature are naturally dynamic representations of social networks which change relatively rapidly with time. This characteristic is distinctly different from other social network graphs such as Internet or email networks. It is therefore often necessary to determine a suitable timescale over which data should be aggregated when constructing the graph. Willkomm et al. demonstrated that an hour is typically a good timescale [86] with the link weight between subscribers $a$ and $b$ at time $t$ computed as follows:

$$W_{a,b}(t) = \sum_i w_i \exp(-\lambda |t - t_i|/w_i)$$

(2.21)

where:

- $w_i$ is the weight for event $i$,
- $t_i$ is the time when event happens,
- $\lambda$ is the elasticity coefficient.
Chapter 2 Literature Review

The summation here is for all events that $a$ and $b$ are involved in, for example a phone call between $a$ and $b$, with $w_i$ being the weight of event $i$ at time $t_i$. The link will be deleted if the weight of the link falls below a certain threshold $w^*$. 

As mentioned above, the graph representations of these social networks evolve very rapidly with time. It is therefore often useful to construct a smoothly changing graph structure that can emphasise the recent behaviour of subscribers. An exponentially weighted moving average (EWMA) is one such approach as proposed in [107]:

$$G_t = \theta G_{t-1} + (1 - \theta)g_t$$  \hspace{1cm} (2.22)

where:

- $G_t$ is the graph representation on day $t$,
- $G_{t-1}$ is the graph representation on day $t - 1$,
- $g_t$ is the graph representation formed by data on day $t$,
- $\theta$ is a scalar decay parameter in the range of $(0,1)$.

The influence of the previous graph structure can be controlled by the $\theta$ value with values close to 1 emphasising historical data and values close to 0 placing emphasis on more recent graph structures.

In fact, the weights between two subscribers reflect the strength of the relationship between them. For the analysis of social networks, Hidalgo and Rodriguez-Sickert [108] have proposed a concept termed social ties. A social tie is very similar to the weighted edge between two subscribers in a mobile call graph and it can act as a surrogate for a measure of social connection between people. As described previously, the weight of the link between two subscribers could be used to detect strong or weak social ties. In addition, the persistence (stability) of the ties can be measured [108]. In this work, CDR data from a block of 40 days was analysed and separated into four panels, each summarising 10 days of mobile phone call data. It was then possible to examine whether there was a link between two subscribers over all four panels. The persistence measurement proposed was:
Chapter 2 Literature Review

\[ P_{ij} = \frac{\sum_{T} A_{ij}(T)}{M} \]  

(2.23)

where:

- \( M \) is the total number of panels,
- \( A_{ij}(T) \) is 1 if nodes \( i \) and \( j \) communicated during panel \( T \), and 0 otherwise.

Figure 2.9 illustrates an example with five subscribers in four different panels, indicating that the link between subscribers 2 and 4 appears in all panels. The link between subscribers 1 and 2 is only apparent in some of the panels. The persistence of the link between subscribers 2 and 4 can therefore be expressed as \( 4/4 \) and that between subscribers 1 and 2 as \( 2/4 \). It is thus possible to view several panels as a single weighted network. Figure 2.9(b) illustrates the resultant weighted network with persistence representation. Links are observed whenever they appear in \( N > 2 \) panels, and the degree of stability is defined as \( 1/N \leq P \leq 1 \).

Persistence is an attribute of a tie. The average of persistence for a particular node for all its ties is described as perseverance, and is defined as [108]:

\[ P_i = \frac{1}{K_i} \sum_j P_{ij} \]  

(2.24)

\( K_i \) is the degree (number of connections) of the \( i^{th} \) node. This metric can be used to investigate which types of nodes have persistent ties.

![Figure 2.9: Definition of persistence [108]](image-url)
Chapter 2 Literature Review

2.3.4 Applications of CDR dataset analysis

Statistical and graph theory based analysis has been widely used in the application of CDR dataset analysis in many different application fields. Previous sections have touched on some of these potential applications and, in this section, a broader review is presented on some of the potential applications to which CDR dataset analysis has been applied in the literature.

In social network analysis, CDR datasets have been used as a data source to investigate community structure (at the large scale of national mobile phone networks) [109]. Louvain method [110], Infomap [111] and the Clique Percolation method [112] are three well-known methods for community detection based on such analysis. The Louvain method and Infomap method find almost tree-like structures as every node belongs to one community. The Clique Percolation finds the dense subparts of the network. Figure 2.10 are examples of the comparison between these three community detection methods. The Louvain method, for example, has been used to study the linguistic distribution of communities in a mobile call graph from a Belgian mobile network [110]. Figure 2.11, relating to this work, highlights a well-known linguistic based split in communities in Belgium as also identified through the CDR dataset analysis. Walsh and Pozdnoukhov [113] also applied Blondel’s method to explore the temporal evolution and the spatial organisation of urban communities.

In addition, a number of studies have focussed on the identification of communities or comparisons of the behaviour of two or more communities through the analysis of mobile phone network CDR datasets. A common theme in these studies is that they utilised location information such as the GPS location of the cell tower and were focussed on spatial network analysis. Expert et al. [114] highlighted the issue that many previous studies utilised standard metrics extracted from networks of subscribers after all spatial properties had been removed. As a result, the authors focussed on the problem of community detection and proposed a modularity function adapted to spatial networks. They showed that the inclusion of spatial information can reveal hidden structural similarities between nodes. Eagle et al. [115] provided a comparison of the behaviour of different communities i.e. between rural and urban societies. They demonstrated that individuals change their patterns of communication to increase similarity with their new social environment.
Figure 2.10: Examples of communities detected with different methods. The different methods are the Infomap method (IM, red), Louvain method (LV, blue) and Clique percolation method (CP, green). For each method, four examples are shown, with 5, 10, 20 and 30 nodes. The coloured links are part of the community, the grey nodes are the neighbours of the represented community [78, 109].

In addition to community detection, detection of important locations is considered to be another useful focus in the analysis of CDR datasets, particularly for urban planning and emergent event detection. Isaacman et al. [116] used clustering and regression techniques on a mobile phone dataset in order to identify important locations such as subscribers’ home and work locations. Vieira et al. [117] adopted a more general point of view to characterise dense urban areas in order to study social dynamics. Becker et al. [12] presented several ways in which CDRs can be used to provide important information about city dynamics to urban planners, such as the ability to automatically identify residential areas. Calabrese et al. [118] analysed 1 million cell-phone traces and associated their destinations with social events. They found that the behaviour of people attending an event are strongly correlated to the type of event and that people who live close to an event are preferentially attracted by it. This information is very useful for city management functions such as events management and congestion mitigation. From another perspective, Soto and Frías-Martínez [119] used a fuzzy c-means clustering algorithm to identify the land use in urban areas.
In mobile phone networks, each phone call is connected through a “nearest” cell station. This facilitates an approximation for subscribers’ geographic locations to be made thus facilitating various forms of spatial or geographic analysis to be completed on different types of networks [120]. Deville et al estimated the population density in France and Portugal based on the number of subscribers who are calling from cell sites [121]. A similar study was also reported by Sterly et al in estimating the density of population in the Ivory Coast in [122]. Doyle proposed a methodology for distance measurements which enables the identification of mobile subscriber travel paths and a methodology for population density estimation based on significant mobile subscriber regions of interest [229]. Mobile data was also used for traffic analysis. Furno et al proposed an Exploratory Factor Analysis (EFA) techniques for network profiling and land use detection [230]. Derrmann et al investigated if handovers can be used as a proxy metric for flows in the underlying road network [231]. The results show that characteristic profiles of handovers within and between clusters of mobile network cells exit which can be used for traffic state estimation.

As CDRs general contain a timestamp, combining space and time information can be used to form subscriber mobility traces or trajectories. Analysis of such traces has suggested that individual mobility is not random. Instead, human trajectories show a high degree of temporal and spatial regularity [92]. For example, Song et al [11] investigated whether it was possible to predict the subsequent location of a subscriber based on the sequence of their most recently visited locations. The results show that up to 93% of a user’s location on average are predictable using such knowledge.

Another application area where CDR dataset analysis has been utilised is examining the response of populations to extreme events [85]. For example, Bagrow et al analysed the reaction of populations to different emergency situations, such as a bombing, a plane crash or an earthquake [123]. The results show the response of the population to the event manifested itself most obviously in the calling pattern of subscribers who typically would not be particularly active in making calls at that time (rather than via an increase in the calling rate of subscribers who would be typically active during that time of day).
The communities of the Belgian network are coloured based on their linguistic composition: green for Flemish, red for French. Communities having a mixed composition are coloured with a mixed colour, based on the proportion of each language. (bottom) Most communities are almost monolingual [78].

A significant focus of research relating to CDR dataset analysis resides in the potential use of mobility patterns (e.g. aggregated subscriber trajectories) determined from the dataset to aid in the processes of urban planning and development [124, 125]. For example, [126] reported on leveraging the result of CDR dataset analysis for transportation planning in an urban setting, with the aim of reducing the road traffic. Another example of such an application was outlined by Toole et al [127] where the flow of residents between each pair of intersections in a city’s road network was estimated. The main results of this work showed that the model could help to estimate
Chapter 2 Literature Review

congestion and detect local bottlenecks in the city. In spatial network analysis, the gravity model has been widely used to model flows such as the road and airline networks between cities [128, 129]. For example, Krings et al. [130] found that inter-city communication intensity is well characterised by a gravity model. The gravity model is based on Newton’s universal law of gravitation, which measured the attraction of two objects based off their mass and distance. They associated users with locations and then formed the links between the subscribers’ locations rather than between the users themselves. Apart from spatial network analysis, there are other areas of research based on CDRs such as identifying usage groups [131], understanding traffic dynamics in cellular data networks [132], analysing urban human mobility [133] and identifying information diffusion in mobile networks [134]. Yet another domain which have used subscriber mobility patterns derived from CDRs is in the modelling of how a disease could travel and spread across a country [135]. Tizzoni et al analysed three European countries using agent-based model to simulate the epidemics spreading across the country. The results showed that the spatio-temporal evolution of epidemics using the CDR based models for human mobility provided similar results to models which solely utilised census data as the basis for human mobility modelling.

2.4 Agent-based modelling

Agent-based model (ABM) is a discrete dynamic modelling paradigm [136]. This individual-level modelling approach has recently been used to model complex systems in a wide variety of fields. With increases in computing power and storage capacity, ABM has grown in popularity as it can model and simulate the spectrum from basic linear situations to discontinuous non-linear situations [137-140].

An agent-based model is composed of three types of elements:

1) The agents and their attributes and behaviours
2) The methods of interaction and relationships between agents
3) The environment of the agents and the interaction between the agents and the environment

The three elements above are the essential features for the creation of an agent-based model. It is emphasised that the agents may execute various behaviours appropriate
Chapter 2 Literature Review

for the system they represent. Agent-based models have generally been used in the development of new theories and formalising existing ones [141, 142].

2.4.1 ABM applied to the Mobile Phone Network Domain

Agent-based modelling has been widely used in the many fields but particularly in the social science arena. However, little research is reported in the literature on the application of agent based modelling of subscriber behaviour in mobile phone networks. Mohammed proposed the use of an ABM approach to investigate customer retention in the UK mobile market [143]; simulation results from this study showed that mixed customer retention strategies can target not only high-value customers but also customers with large personal networks. Fras-Martinez et al used a CDR dataset to develop an ABM model to simulate the spread of an epidemic based on data relating to the H1N1 outbreak in Mexico in 2009 [144]. These simulations demonstrated that the peak number of individuals infected by the virus was reduced by 10% due to government-mandated restricted mobility and that this also postponed the peak of the pandemic by two days. Twomey and Cadman have also presented a business application which utilised the concept of agent-based modelling in telecommunication and media markets [145].

2.4.2 Applications of ABM in other domains

Although the use of agent-based modelling in mobile networks is relatively uncommon, it has been widely used in the analysis of a wide variety of complex system. Complex systems generally contain numerous interacting entities, among which the interactions are non-linear. Since agent-based modelling is particularly suited for simulations of multiple non-linear entities, this lends to their potential use as a tool in modelling complex networks. As agent-based models can be computed in parallel, they can also be applied to large-scale simulations using super-computing if required [139]. One common application area of ABMs is in the area of large-scale road traffic simulations [146, 147]. Nagel and Rickert proposed the use of agent-based techniques to model individual travellers in order to simulate traffic across a whole city for route planning [146]. Recently, applications have been developed using agent-based models for crowd simulation in emergency and evacuation situations [148, 149]. These models have then been used to gain a deeper understanding of human and
geophysical characteristics in order to develop improve emergency evacuation plans [150, 151]. Agent-based models have been applied in financial analysis [152, 153]. An agent-based approach was shown to successfully demonstrate fat tails for the distribution of returns in [154, 155]; in addition, it has been used to simulate financial crises in order to understand the vulnerability of liquidity to crisis-like shocks [156]. Prior to agent-based models being applied to larger-scale social and financial studies, the approach was used in residential segregation simulations [157-159]. Crooks used geometric data and socioeconomic attributes to construct an agent-based model to study residential segregation (99) and demonstrated how this approach can be used in applications such as vector-based GIS. In recent years, interest in the application of agent-based models to the study of land use and land-cover (LUCC) change has seen rapid growth [160]. A cellular model is used to represent the landscape of interest; a combination of this cellular model with an agent-based model provides a way to study the relevant spatial processes, spatial interactions and multi-scale phenomena. For instance, Manson constructed an agent-based dynamic spatial simulation (ADSS) model to assist with policy development and the understanding of human decision making in environmental contexts [161]. This would allow policy-makers to test the efficacy or impact of alternative decisions. In a similar way to LUCC simulations, agent-based models have been also used for urban and regional development. Raney and Nagel developed an agent-based model for transportation simulation, with travellers as individual agents in this model. The eventual goal for this work was to construct a multi-agent traffic simulation of the whole of Switzerland [162]. Unlike transport simulations, in which agents are the individual travellers driving cars, crowd simulations treat each individual person as the agent in order to simulate the movements of individuals so that understandings can be achieved on how spatial activities are used and developed. Batty has also demonstrated agent-based pedestrian modelling for spatial scales, which can help in urban policy-making [163].
Chapter 3 Subscriber Behaviour in a Cellular Network implementing Dynamic Pricing

3.1 Introduction

The call detail records which were the basis for this research was gathered from a prepaid mobile network in Uganda, which implemented a dynamic pricing service. Hence, before proceeding with a discussion on the initial analysis and processing which was completed, it may be helpful to provide a brief overview of material relating to the geography and economy of Uganda and an overview of the mobile phone services in that country.

3.1.1 Geographic, economic and social structure

The Republic of Uganda is located in East Africa and borders a number of countries, including the Democratic Republic of Congo (DRC) to the west, South Sudan to the north, Kenya to the east and Rwanda and Tanzania to the south, as shown in Figure 3.1(a). Uganda has a land surface of 241,038 square kilometres [164]; it is approximately the same size of the United Kingdom. Kampala is the capital city, and is by far the largest metropolitan area in Uganda. Other large urban areas in Uganda are shown in Figure 3.1(a) and include Gulu, Lira, Mbarara, Jinja, Mbale and Masaka.

In addition to the capital Kampala, there are four political administrative regions in the country, as shown in Figure 3.1(b); these are the central, western, eastern and northern administrative regions. Although most of the population of Uganda live in rural areas, there are significant regional variations. The northern region has the lowest population density, while the eastern region has the highest [165]. More than 40 ethnic groups are indigenous to Uganda, and these ethnic groups typically live in well-defined regions of the country.
Figure 3.1: (a) Geography of Uganda; (b) administrative regions of Uganda and (c) ethnic diversity within Uganda (source: http://www.wikipedia.org)
Chapter 3 Subscriber Behaviour in a Cellular Network implementing Dynamic Pricing

Figure 3.1(c) shows the distribution of the four main language families present in Uganda with their associated ethnic sub-groups. Baganda, a sub-group of the Bantu people, is the largest ethnic group, with almost 20% of the total population in Uganda, mainly in the central region. It is clear from Figure 3.1(b) and (c) that the central and western administrative regions are ethnically rather homogenous, with a higher degree of ethnic diversity seen in the eastern region. Although the northern region also has a high degree of ethnic homogeneity, there are distinct ethnic enclaves in the north-western corner of the country and also near the Kenyan border in the east of Uganda. It is also of note that a significant number of refugee
displaced persons camps exist in a number of urban centres, particularly near the South Sudanese and DRC borders [166]. These refugee camps contain large populations of both refugees from these neighbouring states and internally displaced people. Table 3.1 provides a summary of regional population statistics for the administrative regions of Uganda [165].

<table>
<thead>
<tr>
<th></th>
<th>Northern</th>
<th>Central</th>
<th>Western</th>
<th>Eastern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>7,283,300</td>
<td>8,220,900</td>
<td>7,978,600</td>
<td>8,301,900</td>
</tr>
<tr>
<td>Area (km²)</td>
<td>85,392</td>
<td>61,403</td>
<td>55,277</td>
<td>39,479</td>
</tr>
<tr>
<td>Population Density (km²)</td>
<td>85.29</td>
<td>133.9</td>
<td>144.3</td>
<td>210.3</td>
</tr>
</tbody>
</table>

Table 3.1: Regional population statistics

Uganda is an interesting “laboratory” for investigating the use of CDR dataset analysis particularly when applied to the area of social and economic geographical analysis for a number of reasons. Economically Uganda is classified as a developing country but one which is generally viewed as having significant economic potential [167]. Despite significant political challenges, recent decades have seen impressive economic growth in the country particularly driven by industries related to its agricultural sector [168]. More recently, it has been argued that the country, whilst still displaying growth, has economically under-achieved. However recent oil and gas discoveries in the east of the country may offer a catalyst for returning the country to and beyond its historical economic growth levels. Like many of its neighbouring countries, Uganda remains primarily an agriculture-based economy, and this area generates 90% of its export earnings. Industrial activity is primarily located in Kampala and its environs, with
smaller clusters of industrial activity near the other main urban centres. In general, however, Ugandan economic growth has been regionally quite uneven. The northern region is the poorest, with the lowest income [169], followed by the eastern, central and western regions [170].

From a social perspective, the country also displays several interesting characteristics from a research perspective. Whilst improving through developmental and educational initiatives in recent years, poverty levels in Uganda remain high, particularly in rural areas of the country [171]. Like many of its neighbours, internal migration motivated primarily by economic reasons towards urban centres is significant. From an ethnic perspective, the country displays significant regional ethnic homogeneity which, in recent years, has been challenged particularly in some northern and eastern regions by influxes of displaced peoples from conflict regions in neighbouring states and, up until very recently, significant internal conflicts particularly in the Northern region. [172] and [173] offer more detailed discussions on the topics of ethnicity and migration in Uganda, respectively.

3.1.2 Mobile phone services in Uganda

ICT, and more specifically mobile phone\cellular technology, have also played a very significant role in the social and economic life of Uganda, particularly in recent years. Before 1993, there was only a single state-owned monopoly telecommunications provider in Uganda, the Uganda Posts and Telecommunications Corporation (UPTC) [174]. Subscribers had no choice in terms of a plan or switching to other providers. Established in 1998, the Uganda Communications Commission (UCC) is the regulator for the communications sector in Uganda and since 2007 this sector was fully liberalised in Uganda, to allow growth in competition and particularly in order to end a long-established duopoly [175, 176]. As new private telecommunication companies have joined the market in recent years, new pricing plans and services have been provided for subscribers. As a result of this policy, a significant number of mobile operator licences have since been awarded, although not all of these licences have resulted in the launch of commercial networks. An examination of mobile phone penetration rates for Uganda [177], as shown in Figure 3.2, illustrate the very significant growth in mobile phone technology in Uganda since 2000.
A significant majority (approximately 60%) of subscribers are still contracted to the largest operator, the Ugandan national subsidiary of Mobile Telecom Networks (MTN) [179, 180]; Airtel (formerly Zain) is the next largest operator, with nearly 20% of the subscriber base. Based on the statistics from UCC, the number of mobile phone subscribers had reached close to 14 million by June 2011, about 42% of Uganda’s population. Table 3.2 presents information on mobile service coverage and traffic in the five different regions in Uganda in 2012. The coverage figure of 4G is not listed in the table as 4G service is primarily restricted to Kampala.

<table>
<thead>
<tr>
<th>Region</th>
<th>GSM coverage</th>
<th>3G coverage</th>
<th>Share of traffic (GSM)</th>
<th>Share of traffic (3G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kampala</td>
<td>100%</td>
<td>91%</td>
<td>35%</td>
<td>82%</td>
</tr>
<tr>
<td>Central</td>
<td>87%</td>
<td>2%</td>
<td>14%</td>
<td>6%</td>
</tr>
<tr>
<td>Eastern</td>
<td>82%</td>
<td>1%</td>
<td>16%</td>
<td>6%</td>
</tr>
<tr>
<td>Northern</td>
<td>70%</td>
<td>1%</td>
<td>10%</td>
<td>3%</td>
</tr>
<tr>
<td>Western</td>
<td>64%</td>
<td>1%</td>
<td>22%</td>
<td>4%</td>
</tr>
</tbody>
</table>

Table 3.2: Mobile network regional statistics (source: Uganda Communications Commission)
Some of the most recently available reports [181-183] relating to aspects of the mobile phone market in Uganda also highlight some additional interesting features which highlight the importance of mobile phone technologies to large proportions of the population. These reports highlight the overwhelming popularity of prepaid mobile phone services (with over 99% of mobile phone subscriptions being prepaid), the enduring importance of “basic” mobile phone services (i.e. voice calls and short messaging) in the market and the significant levels of coverage provided by the mobile operators in Uganda. Also highlighted are the importance of a number niche services, particularly in the area of m-money services [184], across the economic and social fabric of Ugandan society. Mobile phones have become increasingly important in Uganda as a tool for increasing commercial activity in the agricultural area. For example, they have been used to help farmers to understand the changing requirements of the markets for their crops and to check the daily variations in the prices offered for their crops [185]. Several projects including the VillagePhone project and the various Smartphone application projects run by the Grameen Foundation [186] illustrate the efforts being made to increase mobile phone usage amongst the rural population. Indeed, there are many other innovative initiatives relating to encouraging mobile phone based service delivery in fields such as mHealth [187-189], education [190] and civic engagement [191] under full scale deployment or being pilot tested throughout Uganda.

3.1.3 Related work

The use of quantitative techniques in geographical studies is well established with many arguing that such approaches have their roots in the seminal work of Schaefer [192]. Throughout the “quantitative revolution” [193] from the late 1950s to the 1970s, the application of quantitative techniques, which made use of statistical analysis and related modelling techniques, became common in many fields of geography. In the field of economic geography, Barnes [194] provides an excellent historical perspective on this era and its impact on that particular scholarly field. Whilst Sheppard [195] also places quantitative techniques in a historical perspective, the author also addresses the conflict which often existed between quantitative and qualitative techniques and the impact which the then evolving technology of Geographic Information Systems (GIS)
could potentially have on research practices, particularly the fusion of qualitative and quantitative based analysis.

Much contemporary geographic research utilises statistical analysis or modelling approaches (often in conjunction with qualitative analysis) applied to data which has been collected through small and large-scale surveys, case studies, focus groups or “one-on-one” interviews. In the context of studies specifically relating to Uganda, a significant source of raw data which has been utilised in studies are national census and household surveys compiled by the Uganda Bureau of Statistics [196-198]. Mukwaya et al. [199, 200] used data gathered in this way as the basis for several analyses on topics of interest to our work including the identification and distribution of urban centres within Uganda, the structure and evolution of road transportation links in the country as well as focusing on aspects of long term migration and economic development both at a national and regional level. Gollin and Rogerson [201] also made use of raw data and statistics generated by the UBOS in the development of an economic model which captures the impact of effects such as transportation links, national and regional development and productivity levels. The use of alternative data sets generated through technology is illustrated in [199]. In this work, the authors specifically focussed on examining the expansion of the Ugandan capital of Kampala and leveraged satellite imagery datasets to propose a mathematical model to predict the future expansion of the urban area and to investigate how public policies could be used to mediate this expansion.

The impact and linkages between the penetration and use of Information and Communication Technologies (ICT) and more specifically mobile phone technology on population behaviour and economic activities in Uganda and neighbouring countries has also been investigated in a number of works. Buys et al. [202] utilised data from national census, survey data, economic and infrastructural data, satellite imagery and relief topography data to identify determinants for mobile phone coverage across sub-Saharan Africa. A study presented in [203] concerned itself with the question of whether mobile phone expansion levels in Uganda had an impact on a number of agriculture related economic activities. By means of panel data analysis, the authors posited that mobile phone expansion during the period analysed encouraged farmers to participate in market activity, particularly amongst farmers who
were located in rural areas and who grew perishable crops (i.e. banana compared to maize). Most recently, InterMedia [181] provide an extremely detailed survey based analysis utilising data gathered as part of their Financial Inclusion Trackers (FITS) survey project focusing on “m-money” services offered by various Ugandan mobile network operators. These “m-money” services are operated by mobile operators and they allow users to transfer money to each other, to save money and to pay bills by means of PIN-secured SMS. The report provides a wealth of information, based on household surveys completed in 2012, on the success of and user experience with “m-money” services. The report also provides insights into mobile phone penetration and usage in Uganda, regional variation in the use of mobile phones and usage variations across gender, urban/rural and wealth/poverty divides.

3.2 Dynamic Pricing Service CDR Dataset

3.2.1 Practical Operation of the DPS

In the Ugandan DPS deployment on which the CDR dataset was recorded, subscribers who opted into the service were offered a discount on the default tariff applied to each call. This discount varied continuously during the day and was also dependent upon the network cell in which the subscriber was currently located. All subscribers in a cell were aware of the discount currently on offer to users of the service (by means of a cell broadcast service notification which was used to advertise the current discount rate in the cell). The system on which the CDRs were captured processed all call attempts made by subscribers using the service and the DPS would also confirm the discount being applied to each individual call by means of a USSD notification sent to the caller’s handset during the call set-up. The rationale from a network operator’s perspective of using this type of real-time DPS deployment is to maximise cell utilisation or revenue by offering a discount in order to encourage larger volumes of calls from the subscriber base. Hence, it is common, as in this case, that the discount on offer would vary throughout the day in each individual cell based on the current traffic load experienced in that cell.
3.2.2 Overview of the CDR dataset

The raw data (CDRs) analysed in this research were downloaded from the DPS platform for a period of 19 weeks. The records for each Wednesday were downloaded, starting with 28/04/2010 and ending with 22/09/2010. Typically, the daily CDRs generated by the system involved approximately 6.5 million call attempts involving 2 million unique participants. A subset of the CDRs were analysed, namely call attempts to other prepaid subscribers of the same network (i.e. on-net calls). This typically consisted of CDRs representing an average of 3.5 million call attempts between approximately 800,000 unique participants each day. An example of the format of one of these anonymised CDR is shown in Table 3.3.

<table>
<thead>
<tr>
<th>Time Stamp</th>
<th>Caller ID</th>
<th>Called ID</th>
<th>Cell ID</th>
<th>Discount</th>
<th>Cell Utilisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>734263</td>
<td>1</td>
<td>2</td>
<td>265</td>
<td>70</td>
<td>0.2365</td>
</tr>
</tbody>
</table>

Table 3.3: Example of an individual CDR

Figure 3.3 illustrates the subscriber behaviour within the DPS, beginning from the service launch and for a period of approximately nine months thereafter. In particular, the shaded area highlights the time period during which the analysed CDRs were generated. Figure 3.3(b) shows that during the shaded time period, the total number of subscribers who had opted into the DPS was relatively stable and exhibited only rather modest growth. In addition, and most importantly from the perspective of the network operators, the total number of calls (Figure 3.3(a)) and the average number of calls per subscriber (Figure 3.3(b)) exhibited growth during the period of analysis. This growth, from slightly less than 12 calls per subscriber at the start of the analysis window to approximately 14 calls per subscriber at the end, is clearly visible in Figure 3.3(c).

For the purpose of some initial analysis, the 1279 cell sites identified in the CDR dataset were grouped into four different analysis groups based on their geographic location, namely:
1. Kampala – this analysis region primarily consisted of the urban centre formed by the capital and some adjacent urban areas to the north and south of the city.
2. North – this analysis region mainly consisted of the northern administrative region and a small section of the central administrative region, bordering on the north of Kampala and located south of Lake Kyoga.
3. East – this analysis region corresponded to the eastern administrative region.
4. West – this was a combination of the western administrative region and most of the central administrative region.

This classification of cell sites into these four “regions” was also the basis of how the cell sites were managed in the Operations and Maintenance function within the Ugandan mobile network.

![Graph showing number of call attempts through the DPS, number of subscribers opted into the DPS, and average number of calls per subscriber.](image)

**Figure 3.3**: (a) Number of call attempts through the DPS; (b) number of subscribers opted into the DPS; and (c) average number of calls per subscriber. All are displayed for a nine-month time period post service launch

Table 3.4 summarises the characteristics of these four regions and demonstrates that in terms of basic population, area and mobile usage, they are representative of regions which display some relatively distinct characteristics (as summarised from [204]). The diversity between these regions is also visible in terms of the ethnic make-up of these
areas, as illustrated in Figure 3.1. It is not unexpected to note the significantly higher penetration level of DPS within the urban region of Kampala (probably due to the increased exposure of the general populace to marketing relating to the service and a higher level of subscriber educational attainment and prosperity) compared to the other, more rural (and less prosperous) regions. The very low level of penetration of the service within the northern region (particularly when compared to the regional access to mobile handsets and SIMs reported in [181]) is interesting, since a DPS might intuitively be viewed as being more attractive to low/limited income subscribers who are more prevalent in this region due to the lower tariffs offered by the discounting paradigm.

<table>
<thead>
<tr>
<th></th>
<th>North</th>
<th>Kampala</th>
<th>West</th>
<th>East</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of cells</td>
<td>196</td>
<td>291</td>
<td>490</td>
<td>302</td>
</tr>
<tr>
<td>Average number of subscribers</td>
<td>42,879</td>
<td>268,654</td>
<td>303,715</td>
<td>215,546</td>
</tr>
<tr>
<td>Average number of calls per day</td>
<td>96,129</td>
<td>911,292</td>
<td>1,443,277</td>
<td>877,687</td>
</tr>
<tr>
<td>Average number of calls per subscriber per day</td>
<td>2.24</td>
<td>3.39</td>
<td>4.75</td>
<td>4.07</td>
</tr>
<tr>
<td>Estimated population</td>
<td>8,360,579</td>
<td>2,021,100</td>
<td>11,214,068</td>
<td>10,188,954</td>
</tr>
<tr>
<td>Estimated area (km²)</td>
<td>110,396</td>
<td>1,908</td>
<td>81,938</td>
<td>47,309</td>
</tr>
<tr>
<td>Estimated population density (inh./km²)</td>
<td>75.73</td>
<td>1059.28</td>
<td>136.86</td>
<td>215.37</td>
</tr>
<tr>
<td>Estimated DPS penetration %</td>
<td>0.5</td>
<td>13.3</td>
<td>2.7</td>
<td>2.1</td>
</tr>
</tbody>
</table>


Table 3.4: Statistics relating to the four regions analysed
3.3 Software Tools

All analysis and modelling described in this work was completed using custom scripts written and executed within the scientific and engineering computational programming package, MATLAB® (from Mathworks Inc.) [232]. The plots and graphs contained hereafter were in the majority of cases generated using the standard plotting and graphing functions within MATLAB® (including some image display functions available within the MATLAB® Image Processing Toolkit). In particular, Figure 3.4 and Figure 3.5 were generated using map images captured from GoogleMaps [233] as a backdrop. Also, Figure 3.22, Figure 3.25 and Figure 3.26, which contain movement trajectories overlaid on a geographic representation of Uganda, were generated using the Geospatial Visual Analytics Toolkit [207] developed by the Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS).

3.4 Aims of initial phase of research

The subsequent sections of this chapter will focus on an initial phase of research which was completed. The aims of this initial phase were to complete analysis of the CDR dataset in order to:

- Estimate (in terms of GPS co-ordinates) the location of the individual cell sites reported in the CDR dataset. This level of granular location information was not available as part of the CDR dataset provided (for network operation security reasons). The only information provided was a short name for the site and which of the four operational regions it was located in.
- Obtain a general understanding of the nature of the subscribers’ behaviour in the network during the time period of the study.
- Analyse whether caller behaviour differs significantly depending on the level of discount on offer to the subscriber.
- Determine whether there is a significant level of “discount chasing” visible in the network. Anecdotal reports from the network operator suggested the growth of “micro-businesses” which involved individuals moving from cell
to cell during the day, “chasing” the largest discount on offer and then “selling on” calls from their handsets to others in their physical vicinity at that time.

- Obtain a general understanding of the movement behaviour of subscribers.

3.4.1 Removal of “Noisy” CDRs from the dataset

An initial review of the CDR dataset used in this research was implemented in order to “sanity check” the parameters of each CDR. As a result of this review, it was observed that some CDRs contained unexpected values in some of the fields. In particular it was observed that some CDRs contained indications of invalid discounts (i.e. outside the expected range of 0 to 99%) or invalid cell utilisation factors (i.e. outside the expected range of 0 to 1.0). The source of these errors was unknown but likely to have been related to software bugs either on the DPS platform or other platforms which it interfaced to in order to access data relating to cell utilisation. Whilst the percentage of CDRs containing such issues was small (i.e. 0.5% of the overall dataset), it was decided to remove these CDRs so that they would not introduce a source of “noise” into any subsequent analysis. A second set of CDRs which were also removed were those in which it was observed that the subscriber had dialled their own numbers. Again, the percentage of CDRs of this form was very limited (i.e. 0.1% of the overall dataset) so these too were removed from subsequent analysis. After this “noise removal” or pre-processing of the CDR dataset, the remaining CDRs formed the basis for all subsequent analysis reported here.

3.5 Cell Site Location Estimation

An initial challenge encountered in analysing this dataset was that there was no access to information on the longitudinal and latitudinal positions of the 1279 cell sites in the mobile network. Instead, each cell site was identified by a simple (single word) location name and an indication of the (network O&M based) region in which the cell site was located. In order to facilitate any fine grained spatial analysis, it would be necessary to determine an estimate in terms of latitude and longitude of the location of these cell sites. An initial estimate of most of the cell locations was determined manually by utilising the available cell site name and administrative region through searching a variety of online sources (e.g. Google Maps, the Ugandan Ministry of
Water and Development and the Uganda Bureau of Statistics). However, it was soon apparent that an automated algorithm was required which would provide a means of (a) validating the cell site locations determined by the “manual search” process above and (b) providing an estimate of the geographic location of the small number of cell sites for which a reliable location was not found using the “manual search” method. The basic premise of the developed algorithm is that two cell sites are likely to be close to each other if a subscriber has visited both of the cell sites within a “short” time window (after some initial tests, 60 minutes was selected as the length of the time window used). The steps in the algorithm were:

a. Determine all unique subscribers who visit Cell A.
b. For each subscriber:
   i. Determine a list of the cell sites that the subscriber visited within a 60-minute time window before and after they visited Cell A.
   ii. Calculate the centroid location for this list of “visited” cell sites and assign it as the estimate of the location of Cell A for this subscriber.
c. Average the estimate location of Cell A for all the subscribers who visited it.

In order to determine the accuracy of this estimation algorithm, the estimated location was computed for all the cell sites for which initial location estimates had been obtained through the manual search of the public data sources (which represented 1273 of the 1279 cells); these are referred to here as the “original” cell site locations. An error distance was then calculated between the “original” cell site location and the estimate of the cell site location provided by our algorithm. We then visually reviewed all cell sites which had error distances which appeared unusually large (i.e. > 30 km). Several cases were identified where an error had been made in the “original” cell site location, typically due to the fact that the location name of the cell site was not particularly unique in terms of location names in Uganda, or that the cell site name referred to a building or organisation headquarters rather than a location name. In these cases, the location estimated by the algorithm provided a “seed” location for a further manual search of public data sources. A number of iterations of this overall process were completed until it was clear that reasonable estimates had been obtained for all cell site locations. As regards the very small number of cells for which no “original” location estimates were found using online sources, since the algorithm performed
well in matching locations to a high degree of accuracy when the location of the sites were known, it was assumed to be reasonable to proceed using the cell site estimate provided by the algorithm for these “unknown” cells. Figure 3.4 illustrates the resultant locations of the 1279 cells; these are colour-coded based on the analysis regions in which they are located.

The performance of the location estimation algorithm is summarised in Figure 3.5 and Figure 3.6. In particular, the latter illustrates that 80% of the error distance values are lower than 20 km, suggesting that the overall process for estimating locations for cell sites is reasonable. However, there is a small percentage of cells which have an error distance in the range from 30 km to approximately 65 km. Figure 3.5 provides a visualisation of the location of the cells based on a colour-coding of the error distance parameter. Many of these cell sites with large error distances (marked in red) are located in very isolated areas, where there are few neighbouring cells which could reasonably be travelled to within the time window used in the location estimation algorithm, and/or are located very near the national borders of Uganda.

Figure 3.4: Location of cell sites in Uganda

Both of these factors would result in the introduction of a bias into the centroid-based location estimation algorithm, due to the lack of neighbouring cells in the bordering
area of the candidate cell site. Thus, there are valid reasons for the larger error distance values in all these cases and they do not appear to indicate inaccuracies in our location estimation algorithm.

Figure 3.5: Distance error of cell location estimation algorithm

Figure 3.6: Cumulative Distribution Function of error in location estimate
Chapter 3 Subscriber Behaviour in a Cellular Network implementing Dynamic Pricing

After the estimation of cell site locations, the final CDR dataset structure is shown in Table 3.5. As a result of this process, it includes all the available information for each call attempt.

<table>
<thead>
<tr>
<th>Time Stamp</th>
<th>Caller ID</th>
<th>Called ID</th>
<th>Cell ID</th>
<th>Discount</th>
<th>Cell Utilisation</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>734263</td>
<td>1</td>
<td>2</td>
<td>265</td>
<td>70</td>
<td>0.2365</td>
<td>0.282663</td>
<td>32.56622</td>
</tr>
</tbody>
</table>

Table 3.5: The pre-processed CDR data format

3.6 Aggregated Network Performance and Subscriber Behaviour

3.6.1 Cell Site Utilisation

The CDR dataset was generated during 19 weeks after the deployment of the DPS system. An examination of the changes in usage for all cell stations during this period was carried out in order to get an overview of the uptake of the service and its impact on the cell site utilisation. As described in Section 3.2.2, there are 1279 cell stations in this mobile network. In order to examine whether the low- and high-usage cell stations present different usage growth behaviour, the cell stations were separated into five groups based on the average number of calls for each cell station across the total of 19 days. The number of calls was normalised by the average number of calls for each day. The results in Figure 3.7 show a growth in the number of calls for each group (which is well modelled by a linear growth using the average utilisation factor). However, the low-usage cell stations, Group 1, show a much higher increase, with a 2.4% growth rate compared to the highest-usage cell station, Group 5, with only 1.3% growth rate. Table 3.6 shows the cell utilisation growth rate from the low-usage cell group 1 to high-usage cell group 5.

<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
<th>Group 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Growth rate</td>
<td>2.4%</td>
<td>1.9%</td>
<td>1.7%</td>
<td>1.5%</td>
<td>1.3%</td>
</tr>
</tbody>
</table>

Table 3.6: Cell utilisation growth rate from group 1 to group 5
Chapter 3 Subscriber Behaviour in a Cellular Network implementing Dynamic Pricing

Average number of calls: 1-400

Average number of calls: 401-800

Average number of calls: 801-1200

Group 3
Figure 3.7: Cell utilisation across 19 days

The above would appear to suggest that the higher discounts on offer in cells with low utilisation did appear to attract an increase in calling rates resulting in the observed higher growth rate in the cell utilisation.
3.6.2 Daily Temporal Traffic Patterns

The question of whether discounting of the tariff has an impact on the distribution of traffic peaks during the days was also examined. In this analysis, call attempts were categorised into one of ten different discount ranges (i.e. 0–10%, 10–20%, etc.) depending on the discounting factor applied to the call. The number of call attempts for each discount range was determined in each of 48 thirty minute periods during a day. Figure 3.8 shows an example of the resultant call attempt intensity map for all cell sites on one such day (namely 5th May 2010).

![Figure 3.8: Example of the daily call attempt distribution](image)

The same type of analysis was completed for the remaining days in the dataset which is shown in Figure 3.9. The results show that the daily modal discount range varied slightly each day. However, the periods of high traffic intensity were consistent, and these took place in the morning between 7 am and 8 am and at night between 8 pm and 9 pm.
Chapter 3 Subscriber Behaviour in a Cellular Network implementing Dynamic Pricing

Discount
Time (30 mins interval)
04/28/2010

Discount
Time (30 mins interval)
05/12/2010

Discount
Time (30 mins interval)
05/19/2010

Discount
Time (30 mins interval)
05/26/2010

Discount
Time (30 mins interval)
06/02/2010

Discount
Time (30 mins interval)
06/09/2010

Discount
Time (30 mins interval)
06/16/2010

Discount
Time (30 mins interval)
06/23/2010
Chapter 3 Subscriber Behaviour in a Cellular Network implementing Dynamic Pricing

![Discount Time (30 mins interval) Graphs for Different Dates]

- **07/14/2010**
- **07/21/2010**
- **07/28/2010**
- **08/04/2010**
- **08/11/2010**
- **08/25/2010**
- **09/01/2010**
- **09/15/2010**

These graphs represent the subscriber behaviour in terms of discount and time intervals over different dates (07/14/2010 to 09/15/2010). The x-axis represents the discount percentage, while the y-axis represents time (in 30 mins intervals). The graphs show the distribution of subscriber activities under varying discount scenarios.
3.6.3 Call and subscriber mobility modelling

As has been reported in other studies [10, 92] with large CDR datasets, the probability distribution function (PDF) of call attempts in this dataset does appear to exhibit characteristics of a power-law-like distribution. In addition, the mobility pattern of subscribers (which is estimated here by computing the probability of number of cell sites subscribers’ call through during the 24-hour period) also shows a similar power-law-like distribution.
Figure 3.10 shows the distribution of the number of calls in the four different regions (i.e. North, Kampala, West and East) of Uganda. The northern region of Uganda appears to exhibit a distinctive distribution compared to the other three regions, a result which is similar to others as examined. The potential reasons for this difference are discussed in more detail in Chapter 4.

Figure 3.11 shows the distributions of call attempts and visited cells on a log-log scale.

Figure 3.11: (a) Distribution of call attempts and (b) visited cells on a log-log scale
Since each voice call is assigned a dynamic discount rate, which is calculated based on the current cell utilisation to which the call is connected in real time, this analysis can be further extended by considering the discount rate offered to the calls. This inclusion of the pricing information provides a way to examine the call distribution at different discount levels. The CDRs were separated into ten sub-groups based on the average discount received by subscribers over one day. Figure 3.12 illustrates the distribution of call attempts and visited cells for subscribers with an average discount of between 50% and 60% with a traditional power law fitted to these distributions.

Figure 3.12: Power law fit to (a) the distribution of call attempts and (b) visited cells on a log-log scale (discount range: 50%-60%)
As shown in Figure 3.12, a linear regression does not fit the power-law-like distribution very well over the complete data range. Hence, an alternative model which was then examined was a truncated power law fitting [40], as shown in Figure 3.13.

Figure 3.13: Call attempt distribution with truncated power-law fitting

Although the truncated power law distribution shows a better fit compared to a linear fit, it lacks an effective model for low rate callers within the distribution. Due to the fact that the majority of the subscribers usually make few calls per day (and hence fall into this category of subscribers) a more comprehensive model was desirable. Therefore, a lognormal distribution was examined as a potential better fitting algorithm for the distribution of mobile call attempts [41]. The probability density function (PDF) of the lognormal distribution can be expressed by the equation:

$$f(x|\mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{\frac{-(\ln(x) - \mu)^2}{2\sigma^2}\right\}, \quad x > 0$$  \hspace{1cm} (3.1)

The optimal lognormal fit to the distribution of call attempts (in the 50-60% discount range) is shown in Figure 3.14 which displays a substantially better fit across the complete discount range with $\mu = 1.4262$ and $\sigma = 1.0052$. 
Figure 3.14: Lognormal fit to the distribution of call attempts (discount range: 50%–60\%)

With a lognormal fitting, the fitted parameters, namely the mean and standard deviation, are examined based on different average discount ranges for total of 19 days. Figure 3.15 illustrates the variation of the $\mu$ and $\sigma$ for the distributions of call attempts and visited cells.

Figure 3.15: Variation of mean and standard deviation of lognormal fit for distribution of call attempts and visited cells (19 days)
As temporal variations are not captured in a box plot, the same analysis was applied to each day. The results indicate that two distinctive behaviours are observed, namely between weeks 1–7 and between weeks 8–19. In other words, the behaviour of subscribers appears to have undergone a noticeable change around the eighth week. The possibility of adjusting the DPS system’s operational parameters after deployment in the first seven weeks was therefore considered as the possible cause for this sudden change in behaviour. Data from the first seven weeks was removed from the lognormal fitting analysis, and the updated results show that the variance falls significantly, as can be seen in Figure 3.16.
The results in Figure 3.16 show a difference from Figure 3.15 as the interquartile range (IQR) dropped significantly which suggest a relatively well-behaved and predictable variation in the lognormal mean and standard deviation versus the average dynamic pricing discount.
Figure 3.17: CDF distributions for the number of call attempts and the number of visited cells

A further extension of this analysis allows us to examine how this aspect of subscriber behaviour appears to have evolved over the time periods covered by the dataset. In Figure 3.17, the Cumulative Distribution Functions (CDFs) representing subscriber call attempts and mobility are plotted based on subscriber behaviour on the first day of three time periods during the full 19 week period. The results appear to suggest that
subscribers made more calls and tended to be more active (i.e. visited more cell sites) during the earlier phase of the period covered by the CDRs.

### 3.6.4 Discount “Chasing”

As a result of the actual service deployment in Uganda, some anecdotal evidence existed that the introduction of the service resulted in significant cohort of subscribers “chasing” discount. This, allegedly, involved the subscribers moving throughout the day from one cell to another in an attempt to make calls for the low possible charge (i.e. moving from one cell to another nearby one when they know that a larger discount is available solely for the purpose of making calls). These subscribers were also allegedly offering to other members of the public to make calls on their behalf (for a fee), thus profiting from the discount on offer. It was decided to analyse the CDR dataset in order to determine if there was any evidence of such wide-spread behaviour of “discount chasing”. In this analysis, the joint calling and mobility pattern was examined (i.e. an investigation of whether highly mobile subscribers make more or fewer calls than static subscribers). Figure 3.18 shows sample density contour maps for three different days, which reflects the number of subscribers making X calls from Y cells (locations) during a single specific day.

Figure 3.18 illustrates that most of the subscribers visited fewer than six cells and that they usually made fewer than six calls per day. The higher-mobility subscribers tended to make more phone calls, reaching a peak of about 32 calls with 17 cells visited. Above this level of usage, subscribers tended to be more static; the subscribers who made 140 calls per day, for example, typically only visited fewer than six cells. This is not surprising, since highly mobile subscribers tend to have occupations requiring contact with multiple individuals every day (e.g. trades or business people). However, if we compare these three plots (covering days at the start, middle and end of the analysis period) it is clear that although the behaviour of the majority of subscribers does not change over the 19-week time period, there are some subscribers who made increasing numbers of phone calls (which would result in the contours being “stretched” in the direction of increasing numbers of calls) and visited more cells on a daily basis (which would result in the contours being “stretched” in the direction of increasing numbers of cells).
Chapter 3 Subscriber Behaviour in a Cellular Network implementing Dynamic Pricing

Figure 3.18: Joint calling and mobility pattern density contour map
Overall, in examining this behaviour over the 19 days under study, it was found that this joint mobility/calling pattern analysis seemed to reflect the general increase in calls made as the service was rolled out but also some evidence of higher mobility and hence possible “discount chasing” (but without any evidence that this was for the purpose of making very large volumes of calls).

3.6.5 Patterns in subscribers’ calling behaviour

As the data analysed here consists of 19 different days of CDRs, these can be used to categorise the subscribers into 19 groups, based on the number of days a subscriber appears within the total 19 days. This measure provides a surrogate for the regularity with which a subscriber used the network’s voice services. Figure 3.19(a) shows the average discount that the subscribers obtained versus the number of days during which they made calls. The average discount obtained by subscribers decreases as the regularity of their access increases. In other words, it would appear that subscribers who make phone calls every day are less concerned with the discount offered by the operator. On a related note, Figure 3.19(b) shows that subscribers who regularly make phone calls (i.e. one or more calls on a given day) also tend to make more calls per day, with a very distinctive change in behaviour in both plots once subscribers access the network on 10 or more of the 19 days (i.e. subscribers who access the network on more than ~53% of days).
Chapter 3 Subscriber Behaviour in a Cellular Network implementing Dynamic Pricing

Figure 3.19: Distribution of (a) average discount versus access regularity and (b) average number of calls versus access regularity

Figure 3.20 shows how the subscriber base is spread between very infrequent callers and high volume regular callers. This spread appears to be very predictable using a power law fit with exponent $\alpha=0.66$. 

From 28/04/2010 to 22/09/2010 (19 days)
Figure 3.20: The distribution of the regularity of subscribers versus number of subscribers

The regularity of subscribers can be also quantified based on the proportion of days on which they are active within the total of 19 days. Figure 3.21 shows the average number of calls made by subscribers versus its standard deviation for five levels of activity. The results show that the subscribers with a medium level of activity (between 0.4 and 0.8) have a more stable calling behaviour. Conversely, the subscribers with low and high levels of activity behave more randomly, sometimes making more calls than usual.
3.7 Graph analysis of subscriber behaviour

3.7.1 Subscriber mobility

Previously in this chapter, the subscribers’ mobility behaviour was examined based on the analysis of the distribution of the cell sites visited. However, geographic information relating to the sites’ locations was not considered in that statistical analysis. In order to include this information, a graph-based network analysis was applied in order to gain further insights into the mobility of subscribers. As each voice call is connected through a cell site, each subscriber’s geo-location at that time can be represented as the serving cell site location. Hence, for each individual subscriber, the path between the cell sites serving two consecutive calls can act as a relatively fine-grained surrogate of a subscriber’s physical movement. For a given day, the full set of connections between the serving cell sites for each subscriber forms a representation of the trajectory of their movement during that day.

Figure 3.22 shows the trajectories of all subscribers on 05/05/2010 within the CDR dataset.
Figure 3.22: (a) Trajectories of subscribers and (b) trajectories with 90% transparency

In Figure 3.22(b), the opacity of the plotted links ($0 \leq o_i \leq 1$) provides a clearer illustration of geographic distribution of the key links in the network (i.e. those links
which are present in larger numbers of trajectories). For each edge between two urban centres in the figure, the opacity of the plotted link is based equation 3.2:

\[ o_{i\infty} \propto \frac{w_i}{\max(W)} \]  

(3.2)

where:

- \( w_i \) is the number of trajectories which contain edge \( i \)
- \( \max(W) \) is the maximum number of trajectories which pass between urban centres in the graph.

Figure 3.23: Clustering of the mobile travel network

In order to quantify the level of communications between major cities and/or towns, the complexity of this representation, which will be referred to as a Mobile Travel Graph (MTG), needs to be reduced. A coarse-graining strategy was selected to complete this task since it allows the approximation of large networks by smaller ones whilst maintaining most of the relevant spectral properties [205, 206] of the graph.
Chapter 3 Subscriber Behaviour in a Cellular Network implementing Dynamic Pricing

The strategy of coarse-graining the MTG network is to reduce the number of nodes and edges used to represent the network with \( N \) nodes and \( E \) edges into a smaller network with \( \tilde{N} \) nodes and \( \tilde{E} \) edges. The new \( \tilde{N} \) and \( \tilde{E} \) should be small enough to facilitate visualisation of the graph and to reduce the computational complexity of analysing the reduced graph. A random walk based coarse-graining algorithm was selected for this network reduction process [206].

Figure 3.23 shows the 20 clusters for the reduced network with the “crossed circles” indicating the centre of each cluster. Table 3.7 shows the reduced weighted adjacency matrix for these 20 clusters.

These weights are a representation of the traffic generated within and outside of the “regions” represented by each cluster. Based on this adjacency matrix, the results of the traffic generated in each cluster (i.e. ‘In’), the total traffic into a region including within it (i.e. ‘Total Move In’), the total traffic into a region excluding within it (i.e. ‘Move In’), the total traffic out of a region (i.e. ‘Total Move Out’) and the total traffic out of a region excluding within it (i.e. ‘Move Out’) are computed and shown in Table 3.8.

Figure 3.24 is the reduced network graph, in which the red points are the “super-nodes” (i.e. centres of the clusters) in the network. In order to offer an interpretation of the role of these super-nodes, a constrained network was defined for the geography layout. In Figure 3.24, the values associated with each super-node is the traffic generated in that region and the thickness of the links between nodes represents the traffic level between the two nodes. The blue points are the locations of the 15 most populous urban centres in Uganda. A comparison of the location of these super-nodes with these top 15 towns shows that the most significant urban centres are located very close to the position of these super-nodes. The capital city of Kampala is the most active region with a degree of 318, followed by Masaka, Jinja, Mbale, Mbarara and Arua. The thickness of the links shows that Kampala has strong communication with the neighbouring towns in both the western and eastern regions and reduced levels of communication with the northern town of Masindi. Also highlighted are that Jinja, Mbale and Masaka also appear to have strong communication with their neighbouring towns.
### Table 3.7: Adjacency matrix of the reduced weighted mobile network graph

|   | 0  | 2 | 0 | 7 | 3 | 2 | 5 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
Table 3.8: MTG statistics

<table>
<thead>
<tr>
<th>Region</th>
<th>Clusters</th>
<th>Kampala</th>
<th>Jinja</th>
<th>Mbale</th>
<th>Masaka</th>
<th>Masindi</th>
<th>Mbarara</th>
<th>Nyanya</th>
<th>Kitgum</th>
<th>Gulu</th>
<th>Lira</th>
<th>Arua</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>30.76%</td>
<td>24.35%</td>
<td>21.52%</td>
<td>59.89%</td>
<td>13.53%</td>
<td>43.01%</td>
<td>32.66%</td>
<td>20.89%</td>
<td>2</td>
<td>76.62</td>
<td>33.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>27.26%</td>
<td>22.02%</td>
<td>25.44%</td>
<td>26.86%</td>
<td>17.06%</td>
<td>42.45%</td>
<td>37.38%</td>
<td>17.16%</td>
<td>2</td>
<td>8.49</td>
<td>6.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11.05%</td>
<td>3.73</td>
<td>3.22</td>
<td>34.06%</td>
<td>2.74</td>
<td>39.3</td>
<td>39.3</td>
<td>25.79%</td>
<td>7.84</td>
<td>16.75</td>
<td>7.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15.21%</td>
<td>0</td>
<td>1.74</td>
<td>2.02</td>
<td>34.46%</td>
<td>14.3</td>
<td>34.52%</td>
<td>27.51</td>
<td>13.53</td>
<td>36.53</td>
<td>36.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td>39.3%</td>
<td>3.73</td>
<td>34.46%</td>
<td>14.3</td>
<td>34.06%</td>
<td>2.02</td>
<td>34.52%</td>
<td>27.51</td>
<td>13.53</td>
<td>36.53</td>
<td>36.53</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>3.36%</td>
<td>3.36%</td>
<td>3.36%</td>
<td>3.36%</td>
<td>3.36%</td>
<td>3.36%</td>
<td>3.36%</td>
<td>3.36%</td>
<td>3.36</td>
<td>3.36</td>
<td>3.36</td>
</tr>
<tr>
<td></td>
<td>Move in</td>
<td>3.36%</td>
<td>3.36%</td>
<td>3.36%</td>
<td>3.36%</td>
<td>3.36%</td>
<td>3.36%</td>
<td>3.36%</td>
<td>3.36%</td>
<td>3.36</td>
<td>3.36</td>
<td>3.36</td>
</tr>
<tr>
<td></td>
<td>Move out</td>
<td>3.36%</td>
<td>3.36%</td>
<td>3.36%</td>
<td>3.36%</td>
<td>3.36%</td>
<td>3.36%</td>
<td>3.36%</td>
<td>3.36%</td>
<td>3.36</td>
<td>3.36</td>
<td>3.36</td>
</tr>
</tbody>
</table>
However, the results suggest that there is little communication with the towns in the northern region. The graph also shows there is an isolated node at the northeast region of Uganda. These identified patterns of inter-regional communications is further examined in more detail in Chapter 4, within the context of what it suggests in terms of social and economic activity within Uganda.

Figure 3.24: Traffic generated between regions

3.7.2 Subscriber mobility behaviour at different spatial scales

In order to distinguish frequently travelled paths from less frequently travelled ones, the movement data was aggregated using a trip summarisation algorithm proposed in [207]. This summarisation algorithm has three major steps. The first step is to extract the characteristic points of the trajectory which, in our case, are the starting cell site, ending cell site, number of visited cells and the minimum angle which is to be characterised as a “turn” in the trajectory. Then, a generalisation from points to areas is used to form circles around the cell sites. The final step is then to summarise moves between these circles. After several initial tests, the parameters for the aggregation algorithm were selected to be 30° minimum angle for a change in direction, 300 seconds minimum duration stay at a cell site, 1000 m minimum distance to the next cell site and 5000 m maximum distance to the next cell site. Figure 3.25 shows the resultant subscribers’ aggregated movements on three different spatial scales with a
radius of point clusters of 20 km, 90 km and 200 km; these roughly correspond to the levels of smaller towns, major towns and regions in Uganda.
Figure 3.25: Subscribers’ movements at different spatial scales: (a) town level; (b) major town level; and (c) region level

In Figure 3.25, the radius of a circle is proportional to the magnitude of subscriber movements within each city. Similarly, the size of the arrow is proportional to the number of subscriber movements between each city. The major town level representation, Figure 3.25(b), illustrates that most movements appear to occur within the same region or between neighbouring towns. As expected, since Kampala is the capital of Uganda, it always has the largest X value (and hence is represented by the largest circle regardless of spatial scale); In addition, unlike other large urban centres, some frequent long-distance movement are shown between Kampala and other major towns such as Masaka and Jinja.

Since subscriber traffic generated by subscribers located in the Kampala area makes up the majority of traffic within the CDR dataset, it would also be beneficial to understand the mobility behaviour of subscribers in more detail within Kampala through further analysis of their trajectories. In order to identify meaningful trips, the OPTICS [208] clustering algorithm was used to identify similar trips. The “route similarity” distance functions proposed in [207] were used as the distance function.
(i.e. similarity measurement) for this study. The “route similarity” distance function is used to find the “common routes”. Each route is constructed with a number of positions. These routes are then applied to the “common routes” clustering algorithm. This algorithm scans two routes to find the closest pair of positions between the routes. During the scanning process, two derivative distances are computed. One distance is the mean value between the position pairs; the other distance is a penalty value. The penalty distance increases when a position is skipped and decreases when corresponding positions are found. A maximum and minimum distance (m) between a pair of positions are also specified for normalising the clusters. The sum of the two derivative distances is the final result.

Figure 3.26 shows the result of the “common routes” for the city of Kampala. A maximum distance of 306 m and a minimum distance of 3 m were selected based on the size of the city.

Figure 3.26: Clusters of trajectories by similar routes in Kampala

Figure 3.26 shows the top 7 clusters of trips from the total of 15 clusters. These trajectories demonstrates that the majority of people work and live in the city centre.
Chapter 3 Subscriber Behaviour in a Cellular Network implementing Dynamic Pricing

However, also visible is that a significant number of individuals travelled a long distance between the city centre and Mukono and a shorter distance between the city centre and Makindye every day. These two locations are likely to be the principal living areas in Kampala for significant daily “commuting” cohorts within the population. A certain proportion of subscribers also travelled a longer distance to the area around Entebbe located to the south of Kampala. Uganda’s transportation information indicates that Entebbe has Uganda’s only international airport which suggests that the identified route is that utilised by subscribers who are working in, transporting others to or from or travelling through the airport (or with someone or meeting someone who is travelling through the airport).

3.8 Summary and Conclusions

In this chapter, a comprehensive review of geographic, economic and mobile phone services in Uganda is presented before proceeding with a discussion on the initial analysis and processing of the CDR dataset. An overview of the raw dynamic pricing CDR dataset was provided followed by details of a number of steps used to remove records from the dataset due to noise and other reasons. A cell site location estimation algorithm was proposed and implemented to determine the geographic location for each cell site based on the cell site location’s name. A series of aggregated network performance and subscriber behaviour analyses was presented whose aim was to gain some insights about subscribers’ calling and mobility behaviours in the DPS based mobile network. The result of joint mobility/calling pattern analysis reflects the general increase in calls made as the service was rolled out but also provided some evidence of higher mobility and hence possible “discount chasing” behaviour in a proportion of the subscriber base. Subsequently, a graph theory based coarse-graining strategy was selected to quantify the level of communications between major cities and/or towns.
Chapter 4 Insights into Social and Economic Behaviours in Uganda using CDR analytics

4.1 Introduction

The previous chapter focussed on a more generalised analysis of subscriber behaviour by applying a number of traditional analysis techniques to the CDR dataset. This analysis highlighted that the dynamic behaviour of humans varies across a range of temporal and spatial scales. In this chapter, the focus is on applying the CDR dataset to a study focussed on examining human social and economic behavioural patterns in Uganda. By examining the response of subscribers to a service incentivising higher mobile phone call rates through the offering of discounts, economically motivated differences in subscriber behaviour in poorer versus wealthier regions of the country are identified. Whilst CDR dataset based analysis have been applied to the analysis of social or economic behaviour in other countries and regions [117, 209, 210], this is, to the best of our knowledge, the first application of this form of analysis applied to the country of Uganda.

This chapter also presents an analysis which suggests a high degree of social insularity within the regions of Uganda which is most likely related to regional economic development levels in addition to the high levels of ethnic homogeneity within those regions. A methodology for identifying centres of economic activity using the dataset alone is also presented and the accuracy and implications of the resultant regional patterns are discussed. Finally, measures of human mobility, and their relationship with economic and social regional characteristics, are examined through the use of graph theoretic based analysis techniques.

Traditionally, studies examining topics such as economic development and social behaviour often focus on the nation as the basic unit of analysis. In many cases, particularly for example in developing countries, the logistics and costs associated with the use of large scale surveys as the basis for gathering detailed raw data dictate
the need to focus on national rather than regional behaviour. However, it is well accepted that in most, if not all nations, significant regional variation can be expected to be observed particularly when it comes to the analysis of economic development levels within the nation and when investigating characteristics of social behaviour within a national population. In the field of economic geography, a significant body of work exists on the application of theoretical analysis, case studies and surveys on what influences regional economic development and on the relative performance of difference regions within nations [211]. In particular, works such as those by Scott [212], Clark et al. [213] and Hanson [214] all provide excellent reviews of the field of economic geography and in particular highlight the data gathering and analysis techniques which have typically been applied in these fields. These data capture techniques have also commonly been utilised in many studies related to topics in the arena of human geography. However, a number of limitations can readily be identified with the common approach of using surveys, in some form or another, as a base data gathering paradigm. With the notable exception of national level census (which typically only take place every few years and hence lacks temporal granularity), survey based data tends to exhibit a lack of temporal and/or spatial granularity and diversity. Because of the practical difficulties involved in the completion of surveying activities, the time period over which the data is gathered can be quite extended (ranging from a number of days to perhaps several months). Thus, aggregation or filtering effects will be intrinsically present in the data thus making it extremely challenging to identify behaviour which is temporally short lived or spatially localised.

Even when one considers the most complete forms of surveys, such as national level census, whilst these offers what is effectively a detailed survey of a complete population, the economic costs of completing such activities ensure that they only offer a temporal snapshot and in practice cannot be repeated at a sufficient temporal regularity to be useful in identifying shorter term effects which may be taking place either within social or regional sub-groups of a population.

In recent years, research communities in the areas of social geography and economics have begun to recognise that CDR datasets potentially represent a rich and detailed dataset source for providing insights on human activity and behaviour. They represent a dataset which, because of the degree of mobile phone service penetration in the general population, captures a form of behavioural data for a very significant
proportion of any national or regional population; they also represent a dataset which is re-generated on each and every day of the year, thus potentially capturing behaviours on many scales, both spatially and temporally. It should however be acknowledged that there are clear limitations with the CDR dataset analysis paradigm. This form of analysis will obviously be only appropriate in attempting to identify or analyse certain forms of human behaviours or activities, whilst there will always be certain forms of activity which can only be reliably analysed through detailed surveying of population samples. It is often the case that, when interpreting the results of analysing CDR datasets, that one will have to resort to unsupervised classification techniques [215] due to the lack of any additional information (e.g. age, gender, physical address etc.) concerning individual subscribers represented within the dataset. In some cases, limited demographic information concerning subscribers may be available but this is typically the exception rather than the norm. Thus, while there are a multitude of algorithms and techniques available for the analysis of CDR based datasets, it is commonly the final step of interpreting the results of the analysis in the context of the research questions being investigated which is most challenging.

4.2 Coarse grained regional analysis of subscriber behaviour

4.2.1 Does Tariff discounting have a bigger impact in poorer regions?

Our initial investigations focus on the calling behaviour of subscribers using the DPS to determine whether there are significant variations in behaviour across the four geographic analysis regions examined. The relative regional usage of the service during the analysis time window as shown in Figure 4.1 and Figure 4.2 (in terms of the distribution of the geographic origins of calls and the location of unique callers) is examined. These figures illustrate how the relative proportion of calls and callers in the service vary over the 19 weeks analysis period. In order to reduce noise, a simple low pass filter implemented by amalgamating CDRs for $k = 4$ consecutive weeks before calculating each time point in these plots was applied.

These plots highlight the fact that a very small proportion of calls and callers were from the rural and economically under-developed North region and the majority of activity took place in the other three regions. By considering all calls (as in Figure 4.1 (f)), the use of the DPS does not appear to have had any significant impact on the
relative proportions of calls made in the four regions. This would appear to suggest that the growth in the average number of calls appears to have been relatively evenly distributed across all four regions. However, Figure 4.2 (f) on the other hand clearly highlights that there has been a noteworthy change in the overall proportion of callers using the DPS in the four regions over the 19 weeks time period analysed. This shows that the proportion of callers located in cells in the Kampala (urban) region has grown quite significantly (almost 10%) over the 19 weeks primarily at the expense of the West analysis region. This may be attributable to the likely higher levels of relative affluence and DPS related marketing activities in the main urban centre of the country compared to the other three analysis regions, all of which are primarily rural in their nature.

A more granular analysis of both the regional source of calls and the location of callers is possible when the discounting feature of the DPS is considered. Figure 4.1 (a) to (e) show the proportion of calls made from each of the four regions when the discount in the call tariff fell into one of five discount ranges. In terms of the source of calls which had relatively low discounts applied (i.e. <40% discount), Figure 4.1 (a) and (b) indicate a small increase in the proportion of calls made from the East region and a reduction in the proportion of calls made from the Kampala and West regions. Correspondingly, Figure 4.1 (c) and (d) highlight a significant change in the proportion of calls made from the West and East regions and a relative increase in the proportion of calls made from Kampala and North regions when larger discounts were offered to subscribers. This would seem to suggest that there is some evidence that callers from the North and Kampala regions appear to have been taken advantage of the discounting of call tariffs to increase their calling behaviour when a very significant discount was offered while reducing it at times when less discount was offered.

4.2.2 Is there regional insularity visible in social ties?

Having investigated variations in the relative usage of the mobile network in the analysis regions, it is now examined whether inferences could be drawn about inter-regional communication patterns and, in particular, whether there was evidence of social or ethnic “insularity” amongst the population. An initial step in this analysis was to assign each subscriber to one of the four analysis regions.
Figure 4.1: Change in proportion of calls made from each analysis region during the analysis time period (a) to (e) illustrate calls in which the discount offered on the tariff was in the indicated band and (f) is the overall average across all calls.
Figure 4.2: Change in proportion of callers making calls from each region during the analysis time period (a) to (e) illustrate calls in which the discount offered on the tariff was in the indicated band and (f) is the overall average across all calls.

This assignment was completed by analysing all cells used by each subscriber to make calls and then associating each subscriber with the cell (and hence region) in which they had made the majority of call attempts. Figure 4.3 (a) and Table 4.1 summarises how the mobile network was being used for intra- and inter-regional communications. It is also examined whether the call tariff discounting has an impact on these inter-regional connections as shown in Figure 4.3 (b) to (f).
A number of noteworthy observations can be made based on the results of this analysis:

1. The East and West region exhibit evidence that the population in these regions do not communicate significantly with other subscribers from outside their region. In both cases, >87% of calls made from cells in the regions are to subscribers who are also very likely to be located in the region. The next most likely location of the called party is in Kampala in both cases (which are not unexpected as both regions border onto the capital). This similarity in behaviour is somewhat unexpected since the East and West analysis regions are distinctly different in terms of size, geography and ethnicity. On the other hand, the level of economic development and poverty levels in these two regions are comparable as is the general nature of the urban\rural mixture of the population (i.e. primarily rural with a small number of urban centres). Hence, it is likely that the similarities in this type of calling behaviour in these two regions are likely to be driven by economic development levels and a resultant concentration of social linkages within the region.
(2) The calling pattern of the North region is distinctly different. As noted previously, the North analysis region is the economically least developed of the four regions and it also has significant challenges in terms of geography, limited road and rail communication linkages and significant populations of displaced people (both from neighbouring countries and internally displaced due to security issues in some areas in the Northern administrative region). While still having about two thirds of calls remaining within the region, a much higher proportion of calls made from the region terminate in the other three regions. This could be explained by a combination of calls to neighbouring “border” regions and significant calls to the capital region possibly associated with commercial activity, political administration and population migration.

(3) The impact of internal migration towards the capital (Kampala) also appears to be highlighted by examining the calling pattern from the Kampala region in Table 4.1. The majority of calls remain within the region but a significant proportion of calls terminate in the other regions. Some of these calls will be related to economic activity but a proportion are likely related to migrants (either daily or much longer term) in Kampala remaining in contact with their home areas.

(4) It is also noteworthy that the discount implemented by the DPS only appears to have had a minimal impact on these inter-regional calling patterns (as illustrated by Figure 4.3 (b) to Figure 4.3 (f)). The one significant difference is in the calling pattern from the North region when discounts are low; This appears to result in a lower proportion of calls being made to subscribers in the other regions when the discount is low, perhaps indicating that subscribers are strategically deferring when they are making such calls until time periods when larger discounts are offered. Thus, while the DPS did experience a significant level of growth in terms of both calls and callers during the analysis period, the subscribers did not, for example, appear to leverage the increased discounts on offer by the DPS to significantly alter their normal calling patterns between regions.
Chapter 4 Insights into Social and Economic Behaviours in Uganda using CDR analytics

4.3 Linking subscriber mobility and economic development

4.3.1 Identification of centres of economic activity

In [185] a pattern of concentration of both industrial activity and agricultural activity into particular locations was identified in Uganda, as shown in Figure 4.4. It is intuitive that daily patterns of population movement may be used as a means of identifying areas where significant economic activity is taking place. Hence, the CDR dataset for the cells in each analysis region was examined in order to identify if similar patterns of “concentration” could be identified.

The first step in this analysis was to identify what is categorised as the “home” and “work” locations of subscribers in the four analysis regions. To this end, two time periods in the day are defined, namely the “home” time (i.e. non-working hours - from 8pm to 7 am) and the “work” time (i.e. working hours – 9am to 5pm). The assumption being that subscribers were more likely to make phone calls in their home/work locations during the corresponding home/work time periods.
The next step was, for each unique subscriber, to determine the most frequent cell site from which they called during these two time periods and hence to assign those cells as their “home” and “work” cells. Figure 4.5 (a) and Figure 4.5 (b) are the resultant “home” and “work” location maps. The size of the circles in these figures reflect the number of subscriber having either their “home” or “work” location assigned in a cell.

A cursory examination of Figure 4.5 (a) and Figure 4.5 (b) highlights that the pattern of subscribers during the “home” time period is more diffuse compared to the pattern in the “work” time period. This is particularly evident in the North analysis region. The majority of significant subscriber activity in the North region during the “work” period is concentrated in a small number of towns (particularly Arua on the North-West border with the DRC and Gulu, Lira and Nimbule) which likely indicates that these are significant centres of commercial and administrative activity. The tendency of commercial activity to concentrate in urban centres is also visible in these plots for the East and West analysis regions, particularly around the urban centres of Masaka and Mbarara in the south, Mityana and Mbuende to the west of Kampala, Jinja and
Mukono to the east of Kampala and around a number of Eastern border towns including Tororo and Mbale.

Figure 4.5: (a) Home location map, (b) Work location map

However, unlike in the North analysis region, there is a certain amount of diffusion of subscribers to more rural areas during the “work” time period in the East and West regions. Figure 4.6 is a differential plot between the “home” and “work” maps and it highlights cells in which there is significant increases in subscriber activity during either the “home” (blue) or “work” (yellow) time periods. This plot identifies that apart from Arua in the far north of the country, the most significant daily migratory patterns are concentrated around Kampala and its urban environs, Jinja in the East, and Masaka and Mbarara in the West analysis region. Figure 4.6 also highlights a pattern where a
number of centres which have significant increases in subscriber activity during “work” hours, have neighbouring centres where there is significant increases in subscriber activity during “home” hours. One example of this is in the East analysis region where subscribers likely reside in large numbers in the Bugembe area and each day migrate into nearby Jinja (a dominant urban centre in the area) most likely for commercial, educational and administrative reasons.

![Home and work differential map](image)

Figure 4.6: Home and work differential map

### 4.3.2 Subscriber mobility in the different regions – A closer examination

Some basic statistics relating to individual subscribers’ mobility patterns (estimated using only the cell site location through which calls were made by the subscriber) were examined for each of the analysis regions. Figure 4.7 is the distribution of the average distance travelled by a subscriber between consecutive calls.

The urban nature of the Kampala region is quite distinctly highlighted by this plot and occurs because the cell site density would be very high in this region. Hence the likely subscriber travel distance between calls would be quite low. It is interesting once again to note that the behaviour of subscribers in both the East and West analysis group is almost identical, as had been noted previously in other analysis completed. However, the rural, poorly developed and poverty stricken North region, also exhibits behaviour which is very similar to West and East in terms of this calling mobility factor. This
might appear to suggest that subscriber mobility in the North region (despite its relatively under-developed communication infrastructure and other challenges (i.e. economic and security)) may not be that distinct from that of subscribers located in the other mainly rural West and East areas. In order to investigate this further a graph theory based analysis of subscriber mobility behaviour in these regions was utilised, as discussed in the following section.

Figure 4.7: Distribution of subscriber average travel distance

4.4 Fine grained analysis of subscriber mobility

4.4.1 Mobile travel graph analysis

Graph theory based Mobile Call Graph (MCG) analysis has been widely used in research relating to subscriber behaviour in mobile networks. The most common construction of an MCG is to connect two subscribers with an undirected link if they have communicated at least once [10]. MCG has been shown to be useful for determining that social ties are the basis of information-carrying connections between individuals [108].

Based on the concept of the MCG, it is proposed to construct a variant on this theme to investigate subscriber mobility, namely a Mobile Travel Graph (MTG). This graph
is based on connections between cell sites rather than connections between subscribers. A link is defined as existing between two cell sites if subscribers consecutively use the two cells to make consecutive calls. Currently, only an undirected weighted MTG is considered which means that the transition between the two cell sites could be either from cell A to cell B or from cell B to cell A. The weight of the link is a function of the number of subscribers which consecutively utilised two cells for calls in a day. In order to visualise the subscribers’ mobility behaviour, only those subscribers in the dataset for whom it has been previously determined that they have both a “home” and “work” cell site location are selected. The subscribers are separated into four groups based on their “home” cell site location. Figure 4.8 (a) to (d) are the resultant undirected weighted MTG for the four analysis region subscriber groups. In these plots, two different “strengths” of links between a given pair of connected cell sites are defined, namely “weak” and “strong” - reflecting the number of subscribers whose data link the two connected cell sites.

Figure 4.8: Undirected weighted Mobile Travel Graph, (a) North Region, (b) Kampala City, (c) West Region, (d) East Region
The MTGs shown in Figure 4.8 illustrate a number of interesting points. Firstly, the city of Kampala appears to play a key hub role in terms of mobility of individuals travelling between the other three regions. Secondly, the MTG for the West and East regions highlights that individuals from these regions do not significantly travel outside their own region into the two other regions. Only a small number of weak links extend across the “border” regions between East, West and North regions. Thirdly, the MTG for the North region is distinctly different and much more localised in its nature. The majority of links in the graph are weak particularly those which extend the graph into the other rural regions (and in most cases this occurs through Kampala). Lastly, the Kampala region MTG is primarily focused on the capital but does have strong links extending along the main roadways to the east (towards Jinja) and to the south (towards Masaka) indicating that these routes are of significant commercial importance.

A more focussed examination can be done by only considering the MTG formed by “strong” links between cell sites, as is shown in Figure 4.9. These MTG plots reinforce the comments made above relating to the hub nature of Kampala in subscriber travel patterns from the other regions, the apparent “insularity” of individuals in the East and West region (particularly in regions where these two regions bordered onto the North region) and the extension of travel patterns out of Kampala mainly towards the east and south along the main road routes. Of particular note is the disconnected nature of the MTG for the North region (in Figure 4.9 (a)). This graph suggests that subscribers in the North region have a “localised” behaviour in terms of their mobility (most likely as a result of poverty, security concerns in some parts of the region and poor road communication links). The largest component in this MTG relates to subscribers with their “home” cells located in a region north of Kampala who travel into the capital (most likely for economic reasons). The only other significant components in this travel graph are (i) centered around the urban centres of Gulu and Arua, (ii) formed by a small cluster of cells at the eastern end of Lake Kyoga (which most likely is related to subscriber mobility on the lake relating to fishing and transport activity) and (iii) another cluster in the mountainous region near the Kenyan border to the south of Moroto (this component may in particular be associated with tourist activity in this region and possibly also may be an artefact caused by the use of enhanced range cell coverage in this mountainous and sparsely populated area).
Figure 4.9: Undirected Weighted Mobile Travel Graph (strong links only), (a) North Region, (b) Kampala City, (c) West Region, (d) East Region

Table 4.2 presents a statistical analysis of the four regional MTG (strong links only) and more specifically their main component. These results once again provide evidence of very strong similarities in the behaviour (in this case mobility behaviour) of subscribers in the East and West analysis regions. The statistics of the Kampala MTG reflect its urban nature whereby subscribers would transition often between cells with smaller coverage areas in the centre of the city (resulting in large average number of neighbours). However, the “tentacles” of the Kampala MTG also extend out towards less dense parts of the graph (particularly its extension towards Jinja and Masaka) resulting in a higher heterogeneity compared to the relatively more homogenous East and West MTG. The statistics of the main component of the North MTG are presented for completeness sake. However, these essentially represent the statistics of a small subset of the northern edge of the Kampala MTG (as is reflected in the similarity of some of the graph statistics).
Chapter 4 Insights into Social and Economic Behaviours in Uganda using CDR analytics

<table>
<thead>
<tr>
<th></th>
<th>North</th>
<th>Kampala</th>
<th>West</th>
<th>East</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connected components</td>
<td>16</td>
<td>37</td>
<td>20</td>
<td>18</td>
</tr>
<tr>
<td>Number of nodes</td>
<td>91</td>
<td>554</td>
<td>626</td>
<td>431</td>
</tr>
<tr>
<td>Largest connected component size</td>
<td>32</td>
<td>451</td>
<td>574</td>
<td>390</td>
</tr>
</tbody>
</table>

**Largest Connected Component Network**

<table>
<thead>
<tr>
<th></th>
<th>North</th>
<th>Kampala</th>
<th>West</th>
<th>East</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network diameter</td>
<td>7</td>
<td>17</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>Clustering coefficient</td>
<td>0.338</td>
<td>0.443</td>
<td>0.404</td>
<td>0.415</td>
</tr>
<tr>
<td>Network centralisation</td>
<td>0.282</td>
<td>0.257</td>
<td>0.086</td>
<td>0.096</td>
</tr>
<tr>
<td>Characteristic (Average) path length</td>
<td>3.02</td>
<td>3.869</td>
<td>5.184</td>
<td>4.609</td>
</tr>
<tr>
<td>Avg. number of neighbours</td>
<td>3.8125</td>
<td>22.905</td>
<td>10.16</td>
<td>10.031</td>
</tr>
<tr>
<td>Network density</td>
<td>0.123</td>
<td>0.051</td>
<td>0.018</td>
<td>0.026</td>
</tr>
<tr>
<td>Network heterogeneity</td>
<td>0.746</td>
<td>1.021</td>
<td>0.847</td>
<td>0.839</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>North</th>
<th>Kampala</th>
<th>West</th>
<th>East</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. number of neighbours</td>
<td>3.8125</td>
<td>22.905</td>
<td>10.16</td>
<td>10.031</td>
</tr>
<tr>
<td>Network density</td>
<td>0.123</td>
<td>0.051</td>
<td>0.018</td>
<td>0.026</td>
</tr>
<tr>
<td>Network heterogeneity</td>
<td>0.746</td>
<td>1.021</td>
<td>0.847</td>
<td>0.839</td>
</tr>
</tbody>
</table>

Table 4.2: MTG network statistics (Strong links only)

Figure 4.10: Neighbourhood connectivity

Figure 4.10 shows a final comparison between subscriber behaviour in the four regions. In this the average of the neighbourhood connectivity of all nodes is plotted against the number of neighbours for each node in the MTG. The neighbourhood connectivity for a node is the average connectivity of all neighbours of that node. As with the other analysis, almost identical characteristics for the East and West regions are visible whilst North and Kampala regions are distinctly different. These plots
highlight the potential of using this form of representation as a surrogate for economic development levels, as it clearly illustrates examples ranging from poorly developed (North), moderate levels of development (East and West) and urban levels of development (Kampala).

4.4.2 Alternative method for classifying regional development

![CDF of travel distance](image)

**Figure 4.11: CDF of travel distance**

Whilst an analysis of the Mobile Travel Graphs does provide insights into the level of economic development and urbanisation in the four different regions in Uganda, the use of a much simpler measure based on subscribers’ travel distances as a surrogate for a region’s level of urbanisation is proposed. The travel distance is the summation of the distance between each two cell sites that a subscriber consecutively utilises to make calls. Figure 4.11 contains the Cumulative Distribution Function (CDF) plot of the subscriber daily travel distance for the different regions. The behaviour illustrated for the North region suggests subscribers who remain "localised" on a daily basis. It shows that about 82% of subscribers in the north region travelled less than 10 km in a day. This conclusion also has been confirmed in Figure 4.8 (a) which supports the premise that infrastructure and economic activity is underdeveloped in the North region. The trend of travel distance for subscribers in the West and East is very similar using this parameterisation – with a gradual growth in the CDF as travel distance
increases. This is likely due to the presence of a larger number of medium to large size towns in both the West and East of Uganda. These towns are reasonably well linked (i.e. as reflected by Figure 4.8 (c-d)) to each other which means that it is feasible for larger segments of the population to travel larger distances to complete their daily activities and for significant commerce to take place between these urban centres. In addition, a certain proportion of individuals in the West and East region also appear to travel to the capital, Kampala, every day. The plot representing behaviour in Kampala exhibits distinctly different behaviour compared to the plots relating to the other three regions. It reflects the fact that it is a significant urban centre where the majority of the population travel short distances (but with a well-ordered PDF that appears to closely follow a power-law like distribution). Thus, once again a sequence of graphs is available which can be parameterised and which reflect the range of economic development levels present in the regions of Uganda.

4.5 Summary and Conclusions

This analysis has focussed on addressing a number of questions. Firstly, an investigation has been completed as to whether there were obvious regional variations in the country in subscriber uptake and usage of this new service. A number of changes in how subscribers from different regions of the country utilised the service during the analysis time period have been identified. On examining the inter-regional calling behaviour, the quite “insular” calling behaviour for subscribers from the West and East of the country, whilst subscribers from the poorer Northern part of the country were more likely to make a higher proportion of calls to other regions (and appeared to be slightly more likely to utilise higher tariff discounts when making these calls) was observed. The next focus of the analysis was to develop a methodology for identifying centres of subscriber concentration for residential and work activities. This analysis highlighted a pattern where economic activity (as measured by subscribers calling during “working” hours) appeared to be concentrated around the main urban centres, at the expense of rural areas in all of the regions (admittedly to a larger degree in some and to a lesser degree in others). Finally, the concept of a Mobile Travel Graph (MTG) as a means of representing subscriber mobility patterns in the four analysis regions was introduced and analysed. Once again distinctive patterns highlighting the similarity between subscriber behaviour in the East and West regions of the country
and highlighting distinctly different behaviour in the North and the Kampala regions were highlighted through this analysis.
Chapter 5 An Agent Based Model for Subscriber Behaviour Simulation

5.1 Introduction

In this chapter, the development and operation of an Agent Based Model (ABM) for subscriber behaviour in a dynamically priced mobile telephony network is presented. An ABM is an individual-level modelling system which describes and simulates a system which is designed to model the behaviour and interaction of large groups of real-world entities. In recent years, the increase in computing power and storage capacity has facilitated a growth in interest in the research community on the use of ABMs [216-218]. Compared to traditional modelling approaches, an agent-based approach offers more flexibility and can be used to model and simulate discontinuous and non-linear situations [219].

Traditional statistical models which might be applied to this problem have limitations in terms of the identification of individual subscribers’ calling patterns and the evolution of subscribers’ social and mobility patterns [92, 108, 220]. Whilst some previous work [14, 57, 59] did examine the issue of the impact of dynamic pricing in voice networks, to our knowledge this is the first paper to examine the issue of modelling subscriber behaviour by utilising a data driven approach (i.e. the model will be based, where possible, on behaviour observed in the CDR dataset) rather than from a theoretical simulation approach.

The ABM includes components which simulate subscriber calling behaviour, mobility within the network and social linkages. An overview of the process through which the Agent Based Model (ABM) developed in this research was designed is presented and a detailed description of the various component sub-models within the ABM and how these were designed based on observations from the CDR dataset analysis is outlined.
5.2 Structure of the agent based model

In ABM, agents are autonomous decision-making entities with diverse characteristics. Therefore, in our simulation of a mobile telephony network, the agents represent the individual subscribers and the agent’s characteristics should simulate the real subscribers’ calling, mobility and social behaviour. In order to develop a realistic ABM, the behaviour of the agents must replicate that of the real subscribers in terms of initiating voice calls at different time and locations to their social connections, in a manner which may be dependent on the DPS discount on offer at a given time. Therefore, the agent behaviour contains three major sub-models:

1. Call attempt model (when they call).
2. Subscriber mobility model (where they call from).
3. Subscriber social linkage/network model (who they call).

In order to produce a simulated behaviour which is similar to that of the real subscribers, each of these sub-models was designed based on analysing the real subscriber behaviour hidden in the CDR dataset. The following sections provide details of the design and operation of each of these agent sub-models of the overall model.

5.2.1 Call attempt model

The call attempt model is used to generate a probability of the agent “making a call” at each moment in time during a simulated day. In addition, this sub-model is also responsible for generating the duration of calls which the agent makes.

The call attempt probability will have two underlying drivers which mimic the fact that some calls in a DPS will be driven by the fact that a discount is on offer, whilst other calls are not discount driven and hence they will be initiated regardless of the discount on offer from the DPS. Such a “non-discount” call might for example be related to the subscribers’ employment or some urgent call which a subscriber would make no matter how much or little discount is offered. Alternatively, it is assumed that there are some subscribers who are particularly sensitive to the discount offered to them and therefore their calling behaviour may change based on the tariff that they are
Chapter 5 An Agent Based Model for Subscriber Behaviour Simulation

offered in real time. Such calls which are driven by the discount on offer from the DPS will be referred to as a “discount” call. Hence, since most subscribers should exhibit a mixture of these two types of behaviour, our model for the subscriber calling probability factors the call probability as the probability that a user will make a call independent of the discount, \( p_{nd} \), (which reflects the probability changing diurnally) and the probability that they will make a call influenced by the discount, \( p_{nd} \) and the time of day, i.e. \( p_{nd}, p_d \) in equation (5.1):

\[
p_{ci} = \alpha p_{nd,i} + (1 - \alpha) p_{nd,i} p_{d,i}
\]  \hspace{1cm} (5.1)

where:

- \( \alpha \) is a constant factor to adjust the influence of discount on the calling behaviour of subscribers,
- \( p_{nd} \) is a probability distribution function for the “non-discount” calls,
- \( p_d \) is a probability distribution function for the discount driver calls.

The probability distribution function which was used for modelling \( p_{nd} \) was determined by considering the calling behaviour of subscribers in the CDR dataset when very little discount was offered by the DPS. For each subscriber, a kernel density estimation function [221] was used as a fit to the subscriber’s calling histogram during the day. Figure 5.1 illustrate examples of this PDF for three subscribers as computed from the CDR data. When executing a simulation of a day’s calling activity by the ABM, the \( p_{nd} \) associated with an agent was simply sampled from the fitted PDFs for the real subscribers from the CDR dataset.

The second component in equation (5.1) relates to calls which are driven by the discount on offer from the DPS at a given time. In order to model the subscribers’ calling demand based on the dynamically varying tariff, a discount-demand function (5.2) which was proposed by Fitkov-Norris [57] to quantify the subscribers’ response to the tariff offered was used.
$D(y) = Ae^{-\lambda y}$ \hspace{1cm} (5.2)

where:

- $y$ is the real time calling tariff (which is dependent on the offered discount),
- $A$ is a demand constant shift with defaults to 1,
- $\lambda_i$ is the user demand parameter for user/agent $i$,
- $D$ is the user demand based on the real time calling tariff $y$.

As the discount demand function is an exponential, the conjugate prior is used to model $\lambda_i$ for individual subscribers. The conjugate prior for an exponential is a Gamma distribution with parameters $a, b$. Thus, for a given agent $\lambda_i \sim \tau(a, b)$. Figure 5.2 illustrates an example of the resultant distribution for a subscriber’s calling demand based on price.
The final element of the call attempt model relates to the duration of simulated calls. Unfortunately, no data was available in the CDR dataset relating to call durations and its dependency on discount offered for real calls made on the Ugandan network. As a result, and given the absence of any significant work in the literature that has examined this issue, it was decided to revert initially to a model for call duration (in a static tariffing environment) proposed by Pattavina and Parini [82] whereby the random call duration was sampled from a lognormal distribution with $\mu = 3.758$ and $\sigma = 1.129$ (equivalent to a mean call duration of approximately 81 seconds), as illustrated in Figure 5.3.

However, this model was then modified on the assumption that subscribers will tend to elongate the call duration of discounted calls when the discounting factor offered by the DPS is large. After investigating a number of other alternative models, equation (5.3) is used to alter the average call holding time (and hence the value $\mu$ in the lognormal distribution) in an exponential fashion which increases with square of discount. This particular model was selected compared to the alternatives which were examined as it appeared to offer a more realistic behaviour in terms of how the average call holding time would increase as the discount was increased.

Figure 5.2: Distribution of subscriber effect price on demand, $\lambda_i = 7$
Chapter 5 An Agent Based Model for Subscriber Behaviour Simulation

\[ \tau(y) = \tau_0 e^{kd^2} \]  

(5.3)

where:

- \( \tau_0 \) is the average call duration generated from the lognormal distribution,
- \( d \) is the discount of the call,
- \( k \) is a constant factor which can be used to adjust the impact of the discount multiplier effect.

In this work, a value of \( k = 0.75 \) was utilised as this resulted only in a potential doubling in the average call holding time in scenarios when a very large discount was offered, as illustrated in Figure 5.4.

5.2.2 Subscriber mobility model

The ABM contains a subscriber mobility model in order to simulate the actual behaviour of subscribers in terms of their physical location in the network throughout the day and movement patterns between cells. The inclusion of this module within the model is important as the discount on offer to subscribers (and hence the revenue generated for the network operator) is dependent on the serving cell when the call attempt is made.
In order to develop this model, 1000 subscribers (who made fewer than 50 calls and visited fewer than 20 cells per day) were randomly selected from the CDR dataset. The location (i.e. cell site) and time of calls made by these subscribers were used to generate a two-dimensional histogram (with bins based on time of day and cell site). A multivariate kernel density estimation [82] algorithm was then applied to this histogram in order to estimate the probability of subscribers’ calling locations at each moment during 2 hour intervals throughout the day. An example of one of the fitted PDFs is shown in Figure 5.5, representing the PDF for subscribers calling locations at 10:00 am.

5.2.3 Social linkage/network model

The inclusion of a model of social connections between subscribers was important in order to accurately simulate the likely calling parties (and their geographic locations) in each simulated call between agents. In addition, in Chapter 6 a number of alternative dynamic pricing strategies are investigated, a number of which are based around how often and who a subscriber is calling. For these reasons, the development of a suitable model based on the social connections within the real subscriber base was important for the ABM.
In order to develop a suitable model, once again the behaviour present in the CDR dataset was examined. In this dataset, social interaction/linkages between subscribers are assumed to be reflected in the number of calls made between each other. Using a simple threshold on the number of inter-subscriber calls in the CDR dataset as an indicator of a social linkage, the degree distribution of the resultant social network was examined. Figure 5.6 is the subscribers’ degree distribution where the number of connections represent the number of distinct subscribers that each subscriber has called.

This degree distribution appears to follow a power-law distribution and, hence, there is evidence that the underlying social network is scale-free network [28] properties. As a result, a Barabási Albert (BA) model was selected as the means of simulating an equivalent un-weighted scale-free social network [7] for use within the ABM. When simulating subscriber behaviour, the ABM selects a candidate “called subscriber” for each call based on the calling subscriber’s contact list. However, the call probability to each contact is not uniform; individuals exhibit a large variability in the frequency of calls to their contacts.
Therefore, a weight is associated with each subscriber connection in the network. The weight is determined from the distribution of the number of calls to each of the subscriber’s contacts. Based on the aggregated actual data, a Geometric distribution $w_{i,k} = p_{g,i}(1 - p_{g,i})^{k-1}$ was found to give an excellent fit to the ratio of calls made to contacts.

The value $w_{i,k}$ is used as a weight for the connections, where $i$ is the $i^{th}$ agent and $k$ is the $k^{th}$ contact ordered by popularity. To account for individual differences, the conjugate prior for a Geometric distribution, namely the Beta distribution with parameters $\alpha, \beta$ is used; thus $p_{g,i} \sim Beta(\alpha, \beta)$. The geometric distribution parameter ($p$) is obtained from a Beta distribution given a prior $Beta(\alpha, \beta)$. These two parameters were estimated based on the distribution of the $p_g$ value for all users over the 19 days of data in the CDRs, as illustrated in Figure 5.7 (where values of $\alpha = 2.3854$ and $\beta = 4.6924$ were used). Figure 5.8 also shows an example of this geometric fit for one particular subscriber in the CDR data.
Chapter 5 An Agent Based Model for Subscriber Behaviour Simulation

Figure 5.7: Histogram of aggregated $p_{g,i}$, $\forall i$

Figure 5.8: Geometric distribution fitting for call frequency $p_{g,i} = 0.2$ in this case
Chapter 5 An Agent Based Model for Subscriber Behaviour Simulation

5.3 Load based dynamic pricing model

![Image: The relationship between offered discount and cell utilisation]

Figure 5.9: The relationship between offered discount and cell utilisation

The initial pricing algorithm which was investigated using the ABM was one which modelled the algorithm used in the real mobile network from which the CDR dataset was gathered. This pricing algorithm operated by offering a discount to a caller which was a function of the cell utilisation factor at the time at which the call was made. An analysis of the call records in the dataset, as shown in Figure 5.9, resulted in a mapping between a cell utilisation factor and the discount offered to the caller which was used when modelling this pricing algorithm. One challenge however which was encountered in terms of the implementation of this algorithm within the ABM was the issue of linking the cell utilisation factor to the (channel) capacity of individual cells within the model. Whilst in the real network there would likely be significant differences in the cell capacity in different cells in the network, it was decided that within the ABM framework a cell’s capacity would be modelled using a per call cell utilisation change factor. In short, this would be the amount by which the cell utilisation factor for a cell would increase when a new call was made in that cell (up to a maximum utilisation factor of 1). Figure 5.10 illustrates how this factor was calculated for each cell (using an analysis of the call records from the dataset). The figure shows, for a particular cell in the real mobile network, how a linear
approximation (whose slope of 0.043 is the per call cell utilisation change factor) was fitted to data points plotted from the call records.

![Delta Cell Utilization 0.0429](image)

**Figure 5.10: The estimation of delta cell utilisation**

One issue which was quickly identified during initial simulations of this form of discounting algorithm related to the impact which this algorithm had on revenue related to call attempts made at very low load periods (e.g. in the middle of the night). During these hours, the cell utilisation factor for most cells was very low because there was little or no demand from subscribers to make calls at these times (e.g. between 2AM and 6AM). However, because the cell utilisation factor would be very low at these times, the associated discount on offer would be very large. As a result, the revenue for what few calls were being made during this time period was being hugely diluted due to the un-necessarily large discount being offered. As a result, an initial variation on the cell load based discount algorithm was to examine the use of a random discount in all cells throughout this “middle of the night” time period.

From a practical perspective, the deployment of such a form of discounting algorithm would require significant infrastructure (e.g. CDR data warehousing and real time data analysis systems) to be deployed to allow access to traffic load information (whether real time or averaged over a number of recent days) for all cells in the network. Clearly
the scalability of such an approach in very large networks with potentially tens of thousands of cells would be practically and commercially challenging.

5.4 Results

In this section, the results obtained during the process of tuning certain model parameters in order to achieve comparable results with the subscriber behaviour observed in the underlying CDR dataset used in this work are presented. Figure 5.11 provides an illustration of the ABM simulation process used in generating these results. In general, there are four major steps in this simulation:

1. Agent initialisation.
2. Voice call simulation.
3. Discount offering and.

The agent based model was implemented and running in the mathematical computing software MATLAB® executing on a Windows desktop computer. Unless otherwise stated, all results were generated after multiple simulation runs of the ABM using a given configuration. Each of these simulation runs were based on a simulation scenario consisting of 1000 simulated agents, whose operational characteristics matched those of 1000 subscribers randomly selected from the CDR dataset, operating in the simulated DPS environment for a period of 19 days (as per the underlying CDRs). The decision to limit the simulation to 1000 agents was taken with consideration of the computational speed limitation of using MATLAB® (given that each simulation run of this type required 3 days to complete execution).

5.4.1 Model parameter tuning

One of the key parameters in the model which could not directly be estimated from analysis of the CDRs is the \( \alpha \) value in the call attempt model as shown in (5.1). This parameter effectively controls the relative number of call attempts which are discount driven (as distinct to calls which would occur regardless of the discount level on offer). In order to determine a reasonable value for this parameter, simulations were carried out using a series of values for \( \alpha \) ranging from 0 (i.e. all subscriber calls are motivated
by the offered discount) to 1 (i.e. no subscriber calls are motivated by the offered discount). Figure 5.12 shows a comparison between the total number of calls (in bins of 15 minute duration) per simulated day generated by the ABM and the equivalent data calculated from an analysis of the CDR dataset for settings of $\alpha = 0.1$, $\alpha = 0.5$ and $\alpha = 0.8$ respectively.

Figure 5.11: ABM simulation
Figure 5.12: The simulation results of the total number of calls over a day based on different influential of discount factor $\alpha$. (a) $\alpha = 0.1$, (b) $\alpha = 0.5$, (c) $\alpha = 0.8$
These clearly show that the behaviour of the model with a setting of $\alpha = 0.8$ is qualitatively comparable to the behaviour observed for the real subscribers. Furthermore, Figure 5.13 contains a plot of a normalised root mean squared error (RMSE) between the simulated and real call rate (derived from the CDRs) for the range of investigated $\alpha$ values. As a result, a value of $\alpha = 0.8$, which resulted in a normalised RMSE value of $\approx 0.1$, was selected as offering a good match with the behavior present in the CDRs (whilst not having an unreasonable dependence on offered discount within the call attempt model).

The user demand factor $\lambda$ in equation (5.2) is sampled from a Gamma distribution with parameters $a, b$ as noted in section 5.2.1. The parameters of this distribution were set to $a = 8$ and $b = 0.7$ such that a larger proportion of agents in the simulation would more likely exhibit a behaviour where discount motivated calls would only be made when larger rather than smaller discounts were offered. Figure 5.14 provides an illustration of a set of resultant user demand versus discount curves generated from this distribution. In terms of the social network/linkage model, a BA network generator with its parameters were set to $p = 0.2$, $q = 0.3$ and $m = 2$ was used to simulate the inter-subscriber network structure as discussed in section 5.2.3.
Figure 5.14: Samples of the user demand function $\lambda_i \sim \tau(8, 0.7)$

Utilising these various parameter settings, Figure 5.15 provides a more comprehensive comparison of the behaviour of the ABM with the subscriber behaviour exhibited in the CDR dataset. Figure 5.15 (a) shows the average discount the subscribers obtained over a day (using the load based discounting algorithm in the ABM as was used in the real DPS deployment). Figure 5.15 (b) illustrates the average cell utilisation over a day and Figure 5.15 (c) highlights the distribution of the number of calls per subscriber. The results illustrate that the results obtain from the ABM simulation are quite close to the behaviour observed from the real CDR dataset across all three of these metrics.
Chapter 5 An Agent Based Model for Subscriber Behaviour Simulation

5.5 Summary and Conclusions

In this chapter, an agent-based model for the simulation of the subscribers’ calling behaviour in a dynamically priced mobile network offering voice telephony services is presented. The behaviour of the agents in the model was developed through the analysis of the CDRs generated from the mobile network in Uganda in which dynamic pricing was deployed. After completion of a process of parameter tuning, the results...
obtained using the model was compared across a number of different performance parameters with those deduced from the real network CDR dataset. These results illustrate that agent based modelling appears to offer a means of accurately simulating the behaviour of subscribers in accessing voice services in a dynamic pricing paradigm. The developed model was subsequently utilised in a study investigating the potential revenue generation impact of utilising different dynamic pricing algorithms, as will be discussed in Chapter 6.
Chapter 6 Comparative Study of Dynamic Pricing Algorithms for Voice Services

6.1 Introduction

In the previous chapter, an agent based model which has been developed to model subscriber behaviour in a dynamically priced mobile telephony network was introduced. This chapter will focus on investigations which have been carried out using this model into the likely impact of a variety of different algorithms for controlling the discount offered to subscribers when making voice calls within the DPS environment. The purpose of these investigations is to determine how the revenue generation capabilities of the various algorithms compare when applied to a simulated population of subscribers. The results of these investigations provide insights into the ability of dynamic pricing services to actually deliver on their aims to network operators who deploy them (in terms of at least maintaining if not increasing their revenue for voice services when a DPS is deployed). The estimated revenue generating capabilities of these various discounting algorithms will be compared to a “base line” revenue figure associated with the load based discounting model (i.e. where discount offered to subscribers is inversely proportional to the contemporary load in a cell) detailed in the previous chapter.

6.2 Alternative dynamic pricing algorithms

6.2.1 Random dynamic pricing

Whilst the use of a real time load based pricing algorithm (as discussed in the previous chapter) has some attractive features (e.g. encouraging and discouraging service use during low and high load periods, respectively), there are significant technical and cost challenges in its deployment given the need to measure and process potentially large volumes of data concerning the loads in individual cells. An alternative dynamic tariffing strategy which addresses these challenges to some degree would be to utilise a random pricing algorithm. In practice this would result in the subscriber being
offered a discount whose value would be drawn from some form of probability distribution function (e.g. discrete uniform distribution over the range of discount values) (PDF). However, there still exists a spectrum of variants on this form of discounting which could be considered. At one extreme is the scenario where a single PDF is used for all cells in the network throughout the complete day. While this approach would be quite attractive from an Operations and Maintenance (O&M) perspective for the network operator, it is highly unlikely to result in a maximisation of revenue generated from service usage. At the opposite extreme of the spectrum would be a deployment where each cell in the network uses its own PDF and that this PDF is changed regularly (e.g. every 15 minutes) throughout the day. Intuitively this form of a deployment is far more likely to result in higher revenues for the operator but its deployment would be extremely challenging. Such an approach would require significant and ongoing processing of “per-cell” revenue generation to determine the optimal PDF for each cell in the network (which could be of the order of tens of thousands of cells in larger networks) and in each analysis time period during the day.

Hence, in this study, an intermediate solution between these two extremes was selected in order to investigate the potential of this form of discounting algorithm. The proposed algorithm was based on the utilisation of only two time periods during the day (i.e. “off-peak” and “on-peak”) and cells were grouped into four groups (with each region applying its own PDF applied to all cells in that region). The four cell groupings were based on the regional classification of cells (i.e. Northern region, Eastern region, Western/Central region and Kampala region) which appears appropriate given the relative homogeneity in subscriber behaviour within these regions, as highlighted in Chapter 4. This approach would also be a very reasonable approach for the practical management of such a deployment given that a network O&M centre would only have to monitor, configure and control the performance of the system (and its revenue generation characteristics) using the manageable combination of four regions and two time periods.

6.2.2 Subscriber centric dynamic pricing

A second form of dynamic pricing algorithm which was investigated was one which was “subscriber centric” in offering “high value” subscribers of the network enhanced discounts. The algorithm would operate by offering all subscribers to the service a
baseline discount (in the form of a random discount implementing a relatively low mean discount value similar to that outlined in section 6.2.1). However, the approach requires a certain percentage of subscribers who exhibited some key characteristics to be offered a significantly enhanced discount (again implemented in the form of a random discount but with a far more substantial mean discount value). Two different methodologies for identifying these high value subscribers were investigated namely, (i) the subscribers with the largest average number of call attempts per day and (ii) the subscribers with the largest number of contacts in their social network structure (i.e. largest degree value). This form of discounting algorithm would likely be quite attractive to a network operator in terms of the practicalities for deployment. Since the vast majority of subscribers simply receive the same (low) random discount, there is little need for significant high volume data processing in terms of discount calculation (for example as is required in the load based discounting algorithm). The process of selecting the key subscribers would require significant analysis of subscribers’ calling patterns but this could be done in an off-line manner (using a short term historical analysis of CDRs at some regular time interval). The nature of the algorithm could also introduce, from a service marketing perspective, the opportunity for gamification [222, 223] in an attempt to encourage subscribers who are not in receipt of the larger discounts to increase their service usage pattern.

This phenomenon is further investigated within our model by inclusion of an additional component to the call attempt model introduced in Chapter 5.2.1. When investigating this form of subscriber centric discounting, a component is included whereby subscribers, who receive calls from a subscriber in receipt of the large (“high value” subscriber) discount levels, may increase their call generation rate. The role of this enhancement is to model the likely effect by which the calling party or a network element might “inform” the called party that the caller was in receipt of enhanced discounts because of their “high value” status. This in turn, in some instances, could influence the calling party to make more calls in an attempt to also achieve a similar “status” in the network of subscribers. The modelling approach which was utilised to simulate this effect is based on an adaptation of the epidemic spreading model proposed by Barrett et al. [224], and outlined in equation (6.1).
\[ \Delta p_{d,i} = 1 - e^{\sum_{r \in R} \ln(1 - ks_i)} \]  

(6.1)

where:

- \( \Delta p_i \) is an increase in the discount sensitive calling probability of subscriber \( i \),
- \( s_i \) reflects the susceptibility of normal subscriber \( i \) to “advertising” encouraging them to increase their use of the service (i.e. in the form of the high value subscribers calling them and telling them that they can get more discount by making more calls),
- \( k \) is a constant which reflects the strength of high value subscribers as “advertisers” of the service,
- \( R \) is the set of “advertisers” (i.e. “advertising” subscribers) and \( r \) is the individual “advertiser”.

In our simulation, this equation was calculated at the end of each simulated day in order to determine the increase of \( \Delta p_{d,i} \) for the next simulated day of network operation. Similar to epidemic spreading in social networks as described in [224], it is also assumed that these “normal” subscribers will continue to increase their calling probability if they keep receiving calls from “high value” subscribers. On the contrary, they will reduce their calling probability by a factor of \( 2/3 \) if they do not receive any calls during the preceding day from “high value” subscribers. Hence, the increase in the subscriber calling rate for each agent was computed using equation (6.2)

\[
\gamma^t_i = \begin{cases} 
\Delta p_{d,i}^{t-1} + \Delta p_{d,i}^t (1 + \Delta p_{d,i}^{t-1}), & \text{from high value subscribers} \\
2/3 \Delta p_{d,i}^{t-1}, & \text{otherwise}
\end{cases}
\]

(6.2)

The resultant discount sensitive probability \( (p_{d,i}) \) determined in equation (5.2) is modified using equation (6.3) to provide \( p^*_{d,i} \) which is used in the simulation.

\[ p^*_{d,i} = p_{d,i} (1 + \gamma^t_i) \]  

(6.3)
6.3 Simulation Results

6.3.1 Off-peak discount strategy

As mentioned in chapter 5.3, a significant issue which was initially investigated related to the impact which discounting would have on revenue generation during the significant “off-peak” time period (i.e. during “middle of the night” hours). As can be seen in the plot of call activity extracted from the CDR dataset in Figure 5.14, this period (when averaged across CDRs from the real network) extends from approximately 2 AM and 6 AM. For marketing/commercial reasons, it is likely that some form of discounting strategy would have to be offered during this time period even though there is little commercial motivation for doing this. There was no evidence present in the CDRs that offering very significant discounts (as was the case in the real network where the load based discounting algorithm offered huge discounts during this low load off peak time period) had any impact on encouraging significant call traffic during this time period. Hence, it is proposed that a random discount be offered during this off-peak time period (regardless of what strategy would be utilised during the rest of the day when traffic is far more significant). In order to evaluate the impact on revenue of possible discount levels offered during this period, the ABM was used to simulate the scenario where a random discount drawn from a uniform distribution between $X - 2\%$ and $X + 2\%$ (where $X$ ranged from 2\% to 98\%) was offered to “off-peak” calls. Figure 6.1 illustrates the results of this simulation (after $N = 10$ simulation runs of the ABM) with the normalised (i.e. relative to the revenue when no discount being offered) revenue generated for the operator being plotted against various values of $X$ (with a polynomial spline fitting). These results suggest that significant amounts of night-time revenue losses would be incurred if (mean) discounts greater than 5\% are offered during this time period. However, as noted in Figure 6.1, the revenue generated during this night-time period is quite small (i.e. approximately 1.75\% of overall daily revenue) and hence larger discounts in practice might only be considered as a gambit in a marketing sense. However, a random discount with a mean value of 3\% during the 2-6AM time period was used in all subsequent investigations of dynamic pricing algorithms.
6.3.2 Subscriber centric discounting algorithm operation

Section 6.2.2 outlined a proposed paradigm for delivering subscriber centric discounts. Using this approach, it is proposed to offer during daytime hours a larger discount to a cohort of “high value” key subscribers (i.e. selected based either on node degree or call volume). In practice the size of the cohort of subscribers to which these larger discounts would be offered would likely be driven by marketing rather than technical forces. In our simulations, it was decided to investigate the case where 25% of the subscriber base were offered these enhanced discounts. The impact which the discount had on revenue was investigated for both the “high value” subscriber base and the remaining “normal” subscriber base. Initially, a uniform random discount with a mean value of 5% (with 2% standard deviation) was offered to the “normal” subscriber base with the (mean) discount offered to the key subscribers being varied to investigate the impact on overall revenue. Figure 6.2 (a) shows the revenue generated when the (mean) discount offered to the “high value” subscriber cohort had mean values between 20% and 50%. These plot shows that the revenue reduces in a manner which is reasonably well modelled by a linear fitting for both the case where the “high value” cohort is selected based on call volume and for case where these subscribers are selected based on their connection degree. It also highlights that this decrease in

Figure 6.1: The impact of revenue based on discounts (in the hours 2-6a.m.)
revenue appears to be slightly less for the latter means of selecting the high value subscriber group compared to the former selection criterion.

![Graph](image)

**Figure 6.2:** (a) Impact of mean discount offered to “normal” subscribers on revenue (b) impact of mean discount offered to high value subscribers on revenue.

In order to visualise the impact which the discount offered to the “normal” subscriber cohort has on revenue, a similar analysis was completed with the discount offered to the “high value” subscriber group being fixed (at 25%). Figure 6.2 (b) illustrates the results of this simulation and highlights the need to keep the mean discount offered to this “normal” subscriber cohort small in order to avoid significant revenue losses. Clearly the actual mean discounts offered to these two subscriber groups in practice would have to be set at “reasonable” value for marketing and possibly for regulatory reasons. Hence, in subsequent comparisons, mean discounts of 20% and 5% were utilised for the “high value” and “normal” subscriber groups respectively.

### 6.3.2.1 Inclusion of epidemic spreading model

Section 6.2.2 further outlined the inclusion of an epidemic spreading model between agents which could potentially result in an increase in the likelihood of discount sensitive calls being generated by individual agents. The impact of the inclusion of this behaviour into the agent model was further investigated through simulation runs of the ABM. In an initial simulation, each agent was initialised with a $S_i$ value from equation (6.1) which was drawn from a lognormal distribution with $\mu = 0.6$ and $\sigma = 0.05$. Figure 6.3 presents the average increase in the discount motivated call generation probability ($\gamma$) over a 50 day simulation run for scenarios where “high
value” subscribers are selected based on call volume and based on social connections (i.e. node degree).

Figure 6.3: Evolution of average $\gamma$ over the total of 50 days for (a) all subscribers (b) “mildly” influenced subscribers and (c) “significantly” influenced subscribers
The results indicate that approximately 45% of “normal” subscribers have experienced some level of influence on the basis of calls received from the “high value” subscriber group during the simulation period of 50 days. By splitting the influenced subscriber group into two, based on the level of influence (using a threshold of 0.05 for the average $\gamma$ value), Figure 6.3 (b) highlights that 37% of influenced subscribers were only mildly influenced over the simulation period. Figure 6.3 (c) illustrates the temporal evolution of the $\gamma$ value for the remainder of the group who experienced more significant influence (i.e. these subscribers’ probability of generating a discount sensitive call would have increased on average by 22% by the end of the 50 day simulation period). These results would suggest that significant changes could occur in the membership of the “high value” group over time. The impact of the inclusion of the epidemic model on the estimated revenue generated for the network operator was far less pronounced. In fact, the revenue estimates from the ABM indicates that inclusion of the epidemic model has a negligible effect on revenue (i.e. 0.53% more revenue compared to the non-epidemic model).

6.3.3 Comparison of dynamic pricing strategies

In this section, a comparison is completed of the overall revenue generation capabilities of the three types of dynamic pricing strategy that were introduced in section 6.2. The analysis of the impact on revenue for different strategies was examined based on a regional approach (using the four geographic based regions of Uganda, as discussed in [225]). For each region, firstly the operation of a random discount algorithm operating during daytime hours at a regional level is investigated. In this simulation, the revenue generated when subscribers were offered mean discounts at different levels between 0% and 100% was modelled. Subsequent simulations using the ABM focussed on estimating revenue generated when (a) a load based discounting strategy (as outlined in chapter 5.3) was used and (b) when the subscriber centric discounting algorithm outlined in section 6.2.2 and 6.3.2, were active. Figure 6.4 provides a graphical comparison of the impact on revenue of the discounting algorithms when simulated independently within the four geographical regions of the country. The red fitted spline in this figure was the mean random discount offered from 0% to 100%. The coloured horizontal lines in this figure are provided to facilitate a comparison of the revenue generated using the other DPS
Chapter 6 Comparative Study of Dynamic Pricing Algorithms for Voice Services

pricing strategies and to allow the reader to determine the random discount level which would result in the same revenue as these alternative discounting strategies.

Figure 6.4: Comparison of regional revenue generation, (a) North Region, (b) Kampala City, (c) West Region, (d) East Region

Table 6.1 provides a summary comparison of normalised (relative to the non-discounted case) revenue when load based discounting and subscriber centric based discounting was simulated. In addition, the table also indicates the equivalent mean random discount which would have to be offered to achieve a similar level of revenue.

The relationship between the level of random discount offered and revenue generated, as shown in Figure 6.4 appears to be very similar in three of the four geographical regions. However, the characteristics of this relationship for the simulated North region of the country is somewhat different in that, initially, as the level of discount being offered is increased the fall in revenue is less pronounced.

140
Table 6.1: Comparison of normalised revenue and equivalent mean random discount

For example, in the North region, a random discount of approximately 65% would result in a normalised revenue value of 0.7, whereas in the other three regions this would occur when discounts of only 45 – 50% were offered. Whilst in absolute terms, these all represent a revenue loss for the operator (compared to the case of offering no discount), it does highlight the fact that subscriber agents in the North region are more attracted to this form of tariff in that it encourages them to increase their level of usage of the service when larger discounts are offered compared to the agents in the other regions. This would provide some evidence that random discounting may be more attractive in regions where subscribers are particularly price sensitive (as would likely be the case in the mostly rural and economically underdeveloped Northern region of Uganda). Clearly, as noted above, based on the simulation results such a pricing algorithm would still result in an absolute loss in revenue for the operator which would have to be addressed by other mechanisms.

Table 6.1 clearly illustrates that load based discounting results in very poor revenue generation for the operator. At a national level, the results show that an approximately
30% drop in network operator revenue would result were a discounting algorithm of this nature to be deployed. When the revenue at a regional level was investigated, it is clear that a pricing paradigm of this nature would perform particularly poorly in the North region of our simulation with a nearly 40% drop in revenue being observed. The root cause of this very poor regional performance is the small subscriber base present in this geographic region resulting in there being consistent over-capacity in the cell sites of this region. Hence, some significant adaptation of the load based algorithm would have to be undertaken due to the fact that, for the majority of the time, most cell sites are very lightly loaded (and, hence, unnecessarily large discounts are offered in these cells).

In terms of maximising revenue for the network operator, the results suggest that there are really only two sensible options. Firstly, operators could use random discount algorithms but with mean values which are very small. However, in practice this approach would not be feasible from a marketing perspective and unlikely to maintain customer satisfaction, given a likely expectation amongst subscribers of getting more than low single digit percentage discounts. The second alternative is the deployment of subscriber centric discounting. The simulation results indicate that such algorithms would result in < 6% revenue loss to the operator. From a marketing and subscriber retention perspective, this form of algorithm is also far more palatable given that all subscribers would receive discounts and the “high value” subscriber group in particular would receive significant discounts. In addition, the “competitive” nature of the algorithm (in terms of which subscribers are selected for inclusion in the “high value” group) would likely be particularly attractive from a marketing perspective. A potential downside of this form of algorithm, particularly compared to random discounting, is the additional infrastructure (and associated operations and maintenance overhead) associated with the process of how very large volumes of CDRs are collated and processed in order to regularly update membership of the “high value” subscriber cohort.

It is clear from the simulation results in Figure 6.4 that in absolute terms revenue is maximised when no discount is offered to subscribers. Also it is interesting to note that, using the model which has been outlined, there does not appear to be a secondary maximum in the revenue-versus-discount relationship. This is noteworthy as it would
suggest that there is little role to be played by optimisation algorithms (e.g. genetic algorithms [226], particle swarm optimisation [227], and variants on gradient descent [228]) in attempting to optimise revenue using on-line adaptation of random discounting.

Figure 6.5: (a) Normalised revenue versus discount curves for various call duration model parameter \((k)\) values and (b) associated mean call duration versus discount curves

Figure 6.5 (a) examines the conditions under which our model would result in a scenario where optimisation might be a desirable approach by considering the dependence on network wide revenue on the \((k)\) parameter in the call duration model equation (5.3). Whilst a secondary maximum appears in the revenue curve for values of \(k\) greater than approximately \(k = 1.5\), these do not actually result in revenue gain for the operator until \(k\) values greater than approximately \(k = 1.8\) were considered. Using Figure 6.5 (b) it can be seen that values of \(k\) greater than 1.8 result in a call duration model where the mean call duration is likely to have very large values (i.e. greater than 5 to 6 times the mean call duration of the “no discount”) when very large discounts are offered to subscribers. It is difficult to argue a case that this level of revenue elasticity is very likely to be realistic.

6.4 Summary and Conclusions

This chapter has presented a study investigating the revenue generation potential of a number of different paradigms for offering discounted tariffs for voice calls using the developed ABM. The most significant conclusion drawn from the results in this study
is that, short of offering no or very little discount to subscribers, the model suggests that all the forms of discounting considered will result in revenue loss to the network operator. This loss ranges from approximately 6% in the case of a subscriber centric discounting model where “high value” subscribers are offered enhanced random discounts to approximately 30% in the case of a basic load based discounting algorithm. In addition, the ABM provides a revenue-versus-mean discount approximation for the case where random discounting is deployed which allows estimation of likely revenue loss were such an algorithm deployed (operating at a specific mean discount value).

The potential for deploying optimisation algorithms within the constraints of the model have been examined and the results indicate that these only potentially deliver a revenue growth scenario when the call duration model operates with parameters which are likely to be unrealistic. The results obtained clearly suggest that operators can only minimise revenue loss, within the constraints of the model, by offering little or no discount. Clearly, in practice, this would make such a tariff virtually impossible to market and/or likely to result in significant subscriber dissatisfaction (and resultant outward subscriber churn with associated loss of revenue).
Chapter 7 Conclusions and Future Work

The work presented in this thesis was focussed on leveraging a CDR based dataset gathered on a platform implementing a dynamic pricing algorithm for voice calls within a mobile network in Uganda from both an analysis and modelling perspective. The main contributions of this work are as follows:

- A number of novel algorithms and visualisation techniques applicable to CDR data sets were developed. These facilitated a high level analysis of the behaviour of subscribers and mobile network performance when dynamic pricing is in use in a mobile network. These techniques were applied in our study which, as far as we are aware, is the first published work relating to the application of CDR data set analysis to the country of Uganda.
- A methodology was developed to identify centres of subscriber concentration for residential and work activities. This analysis highlighted a pattern in which economic activity appeared to be concentrated around the main urban centres at the expense of rural areas in all of the regions.
- The concept of a mobile travel graph (MTG) was introduced as a means of representing and analysing subscriber mobility patterns in the four geographic regions of Uganda under analysis.
- A novel application of an agent-based model (ABM) was presented for the simulation of subscribers’ behaviour in a dynamically priced mobile network offering voice telephony services. The operation of the components within this model was based on the results of various analyses of the CDR dataset. To the author’s knowledge, this methodology has never been previously applied to the process of modelling subscriber behaviour in mobile networks.
- The utilisation of the developed ABM to facilitate investigations into revenue generation using a variety of different dynamic pricing algorithms. The results of these simulations suggest limited scope for revenue enhancement using dynamic pricing of voice services given the subscriber behaviour captured by the analysed CDR data set.
7.1 Calling and mobility behaviour of subscribers

Chapter 3 of this thesis outlined the steps involved in the pre-processing of the raw CDR dataset, a novel semi-automated algorithm for cell site location estimation and a number of initial analyses carried out to investigate subscriber behaviour in the network and the impact of the dynamic pricing service on network performance. It was found that higher discounts attracted more calls resulting in a sudden increase in cell utilisation for low-usage cell stations. The results also show that periods of high-intensity call traffic were consistent and that these took place in the morning between 7 am and 8 am, and at night between 8 pm and 9 pm. A lognormal distribution was shown to offer a better fit for the distribution of call attempts, across the complete discount range, rather than a power-law fit. Using a lognormal fitting for the entire dataset, the results indicate that subscribers made more calls and tended to be more mobile (i.e. visited more cell sites) during the early (growth) phase of the DPS deployment. A joint calling and mobility analysis indicated that most of the subscribers visited fewer than six cells and that they usually made fewer than six calls per day. The higher-mobility subscribers tended to make more phone calls, reaching a peak of about 32 calls with 17 cells visited. Above this level of usage, subscribers tended to be more static; subscribers who made 140 calls per day, for example, typically only visited fewer than six cells. In terms of how the calling behaviour of subscribers was linked to the discount on offer, it was observed that the average discount obtained by subscribers decreased as the regularity of the subscribers’ access increased. In other words, subscribers who made phone calls every day appear to be less concerned with the discount offered by the operator. A graphical analysis of subscribers’ behaviour, the mobile travel graph (MTG), was developed. A coarse-graining strategy was selected and this was used to identify 20 major cities/towns in Uganda. The mobility behaviour of subscribers was further examined by aggregating this data based on a trip summarisation algorithm which identified Kampala, the capital of Uganda, and the other three major cities as key nodes relating to the physical mobility trajectories of subscribers. The analysis of the trips taken by subscribers in Kampala using the OPTICS clustering algorithm offered a great deal of insight into subscribers’ daily activities.
Chapter 7 Conclusions and Future Work

7.2 Social and economic behavioural analysis

The focus of Chapter 4 of this thesis was on the examination of social and economic behaviours in rural and urban regions within Uganda using evidence obtained from the CDR dataset. It was observed that subscribers primarily located in the northern and Kampala regions appeared to take advantage of the discounting of call tariffs to increase their calling behaviour when a very significant discount was offered whilst reducing their use of voice calls at times when lower discounts were offered. Having investigated variations in the relative usage of the mobile network in the regions analysed, it was found that inferences could be drawn regarding inter-regional communication patterns; in particular, there was evidence of a certain social or ethnic insularity amongst the regional populations. The study also identified centres of economic activity by identifying the likely geographic locations of subscribers’ homes and workplaces and daily movements between these, for the four geographic regions within Uganda. It was concluded that subscribers were more likely to make phone calls within the regions of their home or workplace during the corresponding home or work time periods. A closer examination of the mobility of subscribers within the different regions suggests that the urban nature of the Kampala region makes it quite distinctive compared to the other three regions that were examined. A fine-grained analysis of subscriber mobility was carried out by means of a graph theoretic approach which was termed as a mobile travel graph (MTG). This graphical representation once again identified distinctive patterns which highlighted the similarity between subscribers’ behaviours in the eastern and western regions of the country, while distinctly different behaviour was observed in the northern and Kampala regions.

7.3 Agent-based modelling of subscriber behaviour

The development of an agent-based model of subscriber behaviour within a dynamically priced mobile telephony network was presented in Chapter 5. The main sub-components of the ABM attempted to model subscribers’ calling behaviour, their mobility and their social linkages (in terms of which subscribers were likely to call each other) and the design of these sub-components was based on behavioural insights extracted from the CDR dataset from the real dynamically priced Ugandan voice call service. These sub-components within the ABM were implemented using a number of
advanced statistical and graph theoretic techniques which were grounded in evidence unearthed from the CDR dataset.

The results of simulations using the developed agent-based model were observed to well match the equivalent behaviour of the real behaviours of subscribers observed from the CDR dataset, across all measurement metrics. As a result, it was concluded that the agent-based model could be used as an effective simulation tool for examining subscriber behaviour in a dynamic pricing environment. The developed ABM subsequently provided a flexible framework for evaluating a number of dynamic pricing strategy by adjusting both the pricing model and the subscribers’ reaction model as detailed in Chapter 6.

7.4 Comparison of revenue generation capabilities of different pricing algorithms

The core focus of Chapter 6 was on the utilisation of the developed ABM, whose design was outlined in Chapter 5, to examine the revenue generating capabilities of a number of different dynamic pricing algorithms. Initially, a dynamic pricing model utilising randomly generated discounting was examined with simulation results highlighting that significant amounts of night-time revenue losses would be incurred if (mean) discounts of greater than 5% were offered during this time period. An alternative pricing model proposed in this work was a subscriber-centric discounting algorithm. This approach offers a larger discount during daytime hours to a smaller cohort of “high-value\key” subscribers compared to the discounts offer to the larger cohort of “normal” subscribers. The results indicate that such a pricing strategy would result in revenue being reduced in a manner which is reasonably well modelled by a linear fit, both for the case where the “high value” cohort is selected based on call volume and for the case where these subscribers are selected based on their level of social connectivity. Furthermore, an epidemic spreading model was applied between agents, in order to model the effect of users influencing each other resulting in an increase in the discount-sensitive calls being generated by individual agents. The results suggested that approximately 45% of normal subscribers would likely experience some level of influence on the basis of calls received from members of the “high-value” subscriber group. Overall, this comparison of dynamic pricing strategies
suggested that operators can minimise revenue loss, within the constraints of the model, only by offering little or no discount.

A final significant conclusion which can be drawn based on the results of this investigation is that the integration of optimisation techniques into a real time or near real time discount adaptation algorithm is unlikely to result in a commercially attractive offering from the network operator’s perspective (within the constraints of the fixed subscriber based population which have been modelled). The deployment of such algorithms in practice would also be extremely challenging (both technically and financially) given the requirements to implement “big data” capture and processing systems in order to access and process the huge volumes of near real time traffic, cell load and revenue generation information from across the network which would be required by the optimisation algorithm. Additionally, such semi- or fully automated algorithms would introduce significant risks from an operator’s perspective in terms of the potential for large revenue loss (or large subscriber dissatisfaction) should the algorithm malfunction in some manner. From a purely algorithmic perspective, these revenue optimisation algorithms would be attempting to carry out their revenue maximisation remit in an extremely noisy feature space which might result in (i) elongated convergence times (during which revenue generation would be sub-optimal) or (ii) the algorithm converging to a local revenue generation maximum rather than a global one.

There are other effects, which would impact on revenue generation, which have not been considered in our model and which could offset these apparent revenue losses from the modelled fixed subscriber base. Firstly, it is likely that innovative and well marketed dynamic pricing tariff offerings will result in an increase in the subscriber base which commits to the dynamic pricing service through the process of inward subscriber “churn” [76] (i.e. where subscribers from other networks change to the network offering dynamic pricing). Another option to address the potential revenue loss is to make the base (or “no-discount”) tariff within the dynamic priced service contract higher than the tariff for subscribers who remain outside the dynamic priced service. Through careful marketing, particularly in countries where price sensitivity is a key driver amongst the general population base and perhaps where there is a cultural attraction towards risk taking, subscribers would still likely be attracted into such a
service in the hope that they will “gain” rather than “lose out” through the discounts which they encounter when making calls. From the network operator’s perspective, the premium which might be added to the “no discount” tariff might simply be selected to offset the predicted revenue loss associated with the discounting algorithm, with revenue growth being achieved by the additional inward subscriber churn due to the dynamically priced service.

7.5 Future work

There is significant scope to build on the work described in this thesis. A suggested initial focus of such future work would be to investigate enhancements to the current ABM in order to address some of its limitations. Specific areas in which the model could be improved are in the area of how the current model simulates (i) call handover, (ii) the dependence of duration on the offered discount, (iii) the impact of discount on subscriber satisfaction (in terms of Grade of Service or GoS) and (iv) inbound and outbound subscriber “churn”. However, in order to undertake such improvements, it would be necessary to access, amongst other things, a greatly enhanced CDR dataset containing data relating to each of the above. Another focus of future work would be to investigate whether the model development approach used in this work can be generalised to CDR datasets captured in other economically developing and developed countries. This could facilitate comparisons in terms of the accuracy of the modelling technique and in terms of subscriber response to dynamic pricing across markets representing countries at different levels of economic development.

In the current work, a dataset gathered in a country where the subscriber base is almost completely prepaid and virtually all prepaid subscribers opted into the DPS has been utilised. In other networks, particularly those located in countries which are regarded as being economically developed, a significant portion of the subscriber base (and, more importantly, the revenue generation base) will be made up of post-paid subscribers (who tend, in general, to be far less price sensitive). In some cases, significant numbers of the post-paid subscriber base may not opt into a DPS (if access to such a service is even offered to post-paid subscribers). However, such subscribers would have an expectation that their GoS will not be negatively impacted by any traffic load increase associated with the deployment of a DPS. Hence, future work
Chapter 7 Conclusions and Future Work

should investigate and integrate methodologies for modelling such post-paid subscribers' behaviour and the impact of constraints enforced on any dynamic pricing algorithm as a result of network operators’ desire to maintain the GoS delivered to non-DPS subscribers.

While the work presented here has identified concerns over the suitability of using semi- or fully automated revenue optimisation algorithms in a DPS (given the subscriber response to the service as captured by the specific CDR dataset utilised in this work), it is felt that there is still merit in future work which investigates whether it is possible to deliver revenue growth optimisation to network operators using such algorithms while still offering subscribers attractive discount levels.

Finally, since the time at which the CDR dataset used in this work was captured (i.e. 2010), the revenue mix between voice and non-voice services for network operators has seen significant change with non-voice service (e.g. data services) typically now being in the ascendency. As a result, there is now significant interest from mobile operators on the potential use of dynamic pricing for data services. Hence there is significant scope for future work to investigate the applicability of the techniques and models reported in this work when applied to the area of dynamic pricing for mobile data services.
Bibliography


166


160. Parker, D.C., Berger, T. and Manson, S.M., Agent-based models of land-use and land-cover change, 2001, LUCC International Project Office.


Appendix A Journal and conference papers relating to this work

Journal Papers


Conference Papers


Han Wang and Liam Kilmartin

ABSTRACT The analysis of Call Detail Record (CDR) data sets generated by mobile telephony networks has generated much interest in recent years, particularly as an easily accessed source of large volumes of data capable of reflecting the dynamic behavior of humans across a range of temporal and spatial scales. This paper presents a study focused on examining human social and economic behavioral patterns in Uganda through the analysis of a CDR data set generated in a Uganda mobile telephone network in 2010. By examining the response of subscribers to a service incentivizing higher mobile phone call rates through the offering of discounts, economically motivated differences in subscriber behavior in poorer versus wealthier regions of the country are identified. The paper also presents an analysis which suggests a high degree of social insularity within the regions of Uganda which is most likely related to regionally economic development levels in addition to the high levels of ethnic homogeneity within those regions. A methodology for identifying centers of economic activity using the data set alone is also presented and the accuracy and implications of the resultant regional patterns are discussed. Finally, measures of human mobility, and its relationship with economic and social regional characteristics, are examined through the use of graph theoretic based analysis techniques.

KEYWORDS CDR data set; big data; graph theory; Uganda; economic development; regional geography; economic geography

Introduction

Traditionally, studies examining topics such as economic development and social behavior often focus on the nation as the basic unit of analysis. In many cases, particularly for example in developing countries, the logistics and costs associated with the use of large-scale surveys as the basis for gathering detailed raw data dictate the need to focus on national rather than regional behavior. However, it is well accepted that in most, if not all nations, significant regional variation can be

Correspondence Address: Liam Kilmartin, Electrical & Electronic Engineering, College of Engineering and Informatics, National University of Ireland, Galway, Ireland. E-mail: liam.kilmartin@nuigalway.ie

# 2014 The Society of Urban Technology
Modelling revenue generation in a dynamically priced mobile telephony service

Han Wang¹ · Damien Fay² · Kenneth N. Brown³ · Liam Kilmartin¹

© Springer Science+Business Media New York 2015

Abstract Dynamic pricing has been used extensively in specific markets for many years but recent years have seen an interest in the utilization of this approach for the deployment of novel and attractive tariff structures for mobile communication services. This paper describes the development and operation of an agent based model (ABM) for subscriber behavior in a dynamically priced mobile telephony network. The design of the ABM was based on an analysis of real call detail records recorded in a Uganda mobile telephony network in which dynamic pricing was deployed. The ABM includes components which simulate subscriber calling behavior, mobility within the network and social link-ages. Using this model, this paper reports on an investigation of a number of alternative strategies for the dynamic pricing algorithm which indicate that the network operator will likely experience revenue losses ranging from a 5 %, when the pricing algorithm is based on offering high value subscriber cohort enhanced random discounts compared to a lower value subscriber cohort, to 30 %, when the priding algorithm results in the discount on offer in a cell being inversely proportional to the contemporary cell load. Additionally, the model appears to suggest that the use of optimization algorithms to control the level of discount offered in cells would likely result in discount simply converging to a “no-discount” scenario. Finally, commentary is offered on additional factors which need to be considered when interpreting the results of this work such as the impact of subscriber churn on the size of the subscriber base and the technical and marketing challenges of deploying the various dynamic pricing algorithms which have been investigated.

Keywords Agent-based model · Revenue optimization · Dynamic pricing · Mobile network services

1 Introduction

Mobile phone penetration levels have experienced exponential growth over the last decade, growing from 34 % in 2005 to over 96 % in February 2013 according to ITU-T statistics [48]. This growth has been particularly noteworthy in regions which are categorized as “developing” by the ITU-T, displaying a growth from 1.2 billion subscriptions (23 % penetration level) to over 5.2 billion subscriptions (89 % penetration) in the same time period. As a result, there has been an increased commercial focus for mobile phone networking companies on such developing markets given its potential for revenue generation into the future. Africa in particular is a region which clearly illustrates these characteristics with the ITU-T estimating that mobile phone subscriptions having grown from approximately 90 million (12 % penetration rate) in 2005 to nearly 550 million (63 % pen
Subscriber Behaviour in a Cellular Network implementing Dynamic Pricing for Voice Calls

Han Wang, Liam Kilmartin

Electrical & Electronic Engineering,
National University of Ireland, Galway, Ireland
h.wang1@nuigalway.ie, liam.kilmartin@nuigalway.ie

Abstract—In this paper we present an analysis of a large Call Detail Record (CDR) data set gathered from a cellular network implementing a novel Dynamic Pricing Service (DPS). In the DPS, the tariff applied to voice calls is discounted depending on time of day and cell location in a directed manner to typically increase or decrease subscriber calls for revenue or traffic load reasons. We have examined the CDRs to determine the subscriber calling and mobility characteristics with a view towards modelling this behaviour and in order to examine whether there is significant evidence of subscriber’s attempt to exploit the discounting algorithm by means of a behaviour we have termed discount chasing.

I. INTRODUCTION

Recent years have seen significant research interest in the analysis of records relating to the activity of subscribers of cellular networks as a means of providing insight into subscribers’ daily behaviour, their mobility patterns and their social interactions. A typical mobile phone network will process several million users’ phone calls every day and, with such high volumes of phone calls and other subscriber activity such as SMS, multimedia messaging and data usage, the daily datasets generated by these individual subscriber interactions with the network are extremely large. In this study we are specifically focussed on the analysis of Call Detail Records (CDRs) generated by a specific entity in a mobile network in an African country which is responsible for implementing an opt-in real-time Dynamic Pricing Service [1-3] for voice calls made by prepaid subscribers in that network.

In the Dynamic Pricing Service (DPS), subscribers who opt into the service are offered a discount on the tariff applied to each call. This discount varies continuously during the day and it is also dependent upon in which cell in the network the subscriber is currently located. All subscribers in a cell are aware of the discount currently on offer to users of the service by means of a Cell Broadcast service notification which is used to advertise the current discount rate in the cell. The system from which our CDRs are taken processes all call attempts made by subscribers using the service and confirms the discount being applied to each individual call by means of a USSD notification sent to the caller’s handset during the call set-up. The rationale from a network operator’s perspective of using such a real time DPS is to maximise cell utilisation or revenue by offering discount in order to encourage larger volumes of calls from the subscriber base. Hence, it is common, as in this case, that the discount on offer will be varied throughout the day in each individual cell based on current traffic load being experienced in that cell. Another significant motivator for the utilisation of a DPS is as a flexible admission control strategy which attempts to discourage users from accessing the network under heavy load conditions by offering little or no discount and encouraging them to utilise the network services during light load periods by offering significant discounts at those times.

The data set which has been analysed in this paper consists of a record of every call attempt made in the network by “opted in” prepaid subscribers on a given week day (i.e. Wednesday) over a period of several months during which time the DPS was active. On any given day there would typically be several million such call attempts in the data set which was analysed. There are several aims to this initial study:

1. To obtain a general understanding of the nature of the subscriber behaviour in the network during the study time period.
2. To analyse whether there is significantly different caller behaviour depending on the level of discount on offer to the subscriber.
3. To determine whether there is a significant level of “discount chasing” visible in the network. Anecdotal reports from the network operator...
Appendix A.4

Inter-regional Calling Patterns in a Dynamically Priced Mobile Network

Han Wang, Liam Kilmartin

Electrical & Electronic Engineering, National University of Ireland, Galway, Ireland h.wang1@nuigalway.ie, liam.kilmartin@nuigalway.ie

ARTICLE INFO

Keywords:
Mobile network Dynamic pricing service Calling patterns Mobile travel graph Uganda

ABSTRACT

In this paper we examine a dataset of Call Detail Records (CDRs) for voice call attempts gathered by a node in a prepaid network in Uganda. This node was implementing a Dynamic Pricing Service (DPS) which resulted in subscribers receiving a discount (i.e. calculated using a cell utilisation factor) for each call attempt made by prepaid subscribers who opt into the service. We present an analysis to understand how subscribers are behaving when using this service and particularly we identify regional variations which appear to be related to levels of economic activity. We also investigate insights which the dataset provide us with concerning centralisation of daily economic activity in the country. We apply a graph theoretic analysis to the analysis of regional subscriber mobility patterns and isolate some distinct patterns in these Mobile Travel Graphs which differentiate the regions of Uganda and which appear to correlate to levels of regional development in the country.

1. Introduction

The analysis of data sets generated by entities within mobile phone networks which capture the behaviour of very large numbers of individual subscribers has generated significant interest amongst researchers in a number of fields in recent years. The modelling of subscriber behaviour in particular is viewed as being a fundamental starting point in attempting to understand subscriber behaviour in mobile networks (Gonzalez, Hidalgo, & Barabasi, 2008; Hidalgo & Rodriguez-Sickert, 2008; Hohwald, et al., 2010; Seshadri, et al., 2008; Wang & Kilmartin, 2011). However, due to the sheer size of the data sets involved (particularly those which span significant periods of time and which represent the behaviour of a significant proportion of the network’s subscriber base) it is quite challenging to identify and parameterise the behaviour of such a large body of subscriber in a robust and meaningful way. As a result, most research of this type focuses on obtaining a macroscopic perspective of the dataset, with an emphasis on goals such as community detection, determination of important locations or individuals, urban planning etc.

In this study we have focussed on the analysis of Call Detail Records (CDRs) generated in a mobile network in Uganda over a significant time period. The network element which generated these CDRs was responsible for the implementation of a real-time Dynamic Pricing Service (DPS) for voice calls made by prepaid subscribers in that network. The data set which has been analysed in this paper consists of a record of every call attempt made in the network by “opted in” prepaid subscribers on a given week day (i.e. Wednesday) over a period of several months during which time the DPS was active. This analysis time period commenced shortly after the introduction of the DPS in the network and hence captures subscriber behaviour during a part of the “growth” phase of subscriber interest in the service.

In the Dynamic Pricing Service (DPS), subscribers who opt into the service are offered a discount on the tariff applied to each voice call which they initiate. This discount varies during the day and it is also dependent upon in which cell in the network the subscriber is currently located. All subscribers in a cell are aware of the discount currently “on offer” to users of the service by means of a Cell Broadcast Service notification which is used to advertise the current discount rate in the cell. The system from which our CDRs are taken processes all call attempts made by subscribers using the service and it confirms the discount being applied to each individual call by means of a USSD notification sent to the caller’s handset during the call set-up. The rationale from a network operator’s perspective is that the real time DPS is to maximise cell utilisation or revenue by offering discounts in order to encourage larger volumes of calls from the subscriber base. Hence, it is common, as in this case, that the discount on offer will be varied throughout the day in each individual cell based on the current traffic load being experienced in that cell. Another significant motivator for the utilisation of a DPS from the network operator’s perspective is as a flexible admission control strategy which attempts to discourage users from accessing the network under heavy load conditions by offering little or no discount and encouraging them to utilise the network services during light load periods by offering significant discounts at those times. From a commercial perspective, the DPS also offers the network operator a powerful means of incentivising subscribers to use their services and hence it is potentially a powerful tool to combat subscriber churn.

In this paper we present the results of our analysis which attempts to address a number of questions. Firstly, we investigate whether the discounting feature of the DPS appears to significantly change subscriber behaviour. This analysis leads to an additional interesting question of whether subscriber behaviour to the DPS is homogenous across the complete country or whether there are distinctive regional behavioural patterns visible. We examine the question of whether we can determine centres of economic activity/development in the regions of the country and whether this development appears to be concentrated in certain areas or more widely spread based on analysis of the CDR set. We then present a graph theoretic based analysis to examine subscriber mobility patterns and geographic variations in these characteristics.

Section 2 in this paper provides background material to our research containing both an overview of Uganda and relevant socio-geographic information about the country and also related research of relevance. Section 3 describes the CDR dataset which was used in this analysis and some important analysis steps which we carried out to enable the spatial analysis of the CDR dataset. Section 4 introduces the analysis which has been carried out to examine subscriber’s behaviour in the DPS environment and in particular to intra- and inter-regional behaviour. Section 5 in this paper is focussed on examining subscriber mobility in a number of ways. We examine daily mobility of individuals and examine how this can be used to identify centres of economic activity and to investigate whether such activity is happening equally across Uganda. We also introduce our utilisation of a graph theory based representation of subscriber mobility. The final section concludes the paper with a summary of the results which we have observed and we outline how these results are to be utilised in future work relating to the simulation of subscriber behaviour.

2. Background

2.1 Overview of Uganda

2.1.1 Geographic, economic and social structure

The Republic of Uganda is a country located in East Africa and borders a number