Modelo de identificação de churn rotacional nas comunicações móveis

por

Ana Luísa Lameira Ferreira

Dissertação de Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão

Orientada por

João Manuel Portela da Gama
Márcia Oliveira

Faculdade de Economia

Universidade do Porto

2015

Nota Biográfica

Ana Luísa Lameira Ferreira, nasceu em Aveiro a 4 de Fevereiro de 1992. Ingressou na Licenciatura em Economia pela Universidade de Aveiro em 2010 tendo-a completado em 2013. Ainda nesse ano, optou por seguir o seu percurso académico no Mestrado de Modelação, Análise de Dados e Sistemas de Apoio à Decisão, na Faculdade de Economia da Universidade do Porto procurando complementar a formação anterior. Em 2012 fez parte da equipa desportiva de *futsal* do curso de Economia, pela Universidade de Aveiro. Durante um ano, até 2013, pertenceu ao Núcleo de Estudantes de Economia da Associação Académica da Universidade de Aveiro (NEEC-AAUAv) desempenhando funções na secção de Comunicação e Imagem.

Agradecimentos

À minha família, pela educação e valores que me transmitiram desde sempre e pelo apoio constante ao longo da realização desta dissertação, bem como pela paciência e compreensão demonstrada. Aos meus colegas de mestrado, pela cooperação, partilha de conhecimento e apoio mútuo. Aos meus amigos, pelo apoio, paciência e motivação que me ofereceram sempre que necessário. Ao Professor João Gama e à Márcia Oliveira, pelo conhecimento que me transmitiram e pela orientação.

Este trabalho foi apoiado pela Comissão Europeia no âmbito do projeto MAESTRA (Grant number ICT-2013-612944), a quem cabe também o meu agradecimento.

Resumo

O principal objetivo desta dissertação é a criação de uma metodologia que permita a identificação dos clientes que abandonam uma operadora de telecomunicações móveis e regressam com nova subscrição sem que a empresa tenha conhecimento que se trata do mesmo indivíduo. Deste modo, é possível detetar casos em que o *churn* é acionado por conveniência. Isto é, em períodos complementares, pretendemos a identificação de indivíduos semelhantes, tendo por base as respetivas comunicações com outros clientes. Este é um problema real, que causa enviesamento tanto nos dados relativos ao *churn* como nas taxas de sucesso das campanhas para captação de novos clientes. Para além disso, provoca despesas desnecessárias à operadora, que poderá perder dinheiro a tentar recuperar um cliente que supostamente abandonou e que oferecerá regalias a um presumível novo cliente. Ainda não existem metodologias propostas na literatura para abordar o problema do *churn* rotacional, pelo que será criada uma que responda às necessidades da empresa e ao problema em questão. Temos como hipótese que um cliente suspeito de ser o mesmo indivíduo que determinado *churner*, terá um padrão de contactos semelhante, nos dois espaços temporais complementares. Esta hipótese é baseada no facto de as redes sociais serem relativamente estáveis ao longo do tempo e portanto, de existir uma tendência para que a comunidade de determinado indivíduo não sofra grandes alterações ao longo do tempo. A metodologia apresentada é pioneira, dado que este tema foi pouco estudado, e resulta no cálculo da semelhança entre dois indivíduos com base nos clientes com que contatam, em períodos de tempo complementares. Algumas simulações foram feitas à metodologia criada, sendo que esta é aplicada a dados reais e anonimizados cedidos por uma operadora de telecomunicações.
Palavras-Chave: Churn, Telecomunicações Móveis, Análise de Redes Sociais

Abstract

The main aim of this thesis is to create a methodology to identify clients that leave a mobile network company and return as new clients, while the network has no knowledge that they are the same person. With this methodology, we can detect cases when *churn* is actuated by convenience, *ie*, we want to identify similar subjects using their communications with other clients in complimentary time periods. This is a real life problem that causes bias in the *churn* related data, as well as in the success rates of the campaigns targeted to obtain new clients. Moreover, this causes unnecessary spending to the network, who might be losing money while trying to gain back a client that is supposed to have left them at the same time that it is giving benefits to a potential new client. There are no proposed methodologies in the literature to solve the problem of rotational *churn*, therefore this creates an answer for the needs of the company and to the problem described above. Our hypothesis is that a client that is suspected of being the same person as a certain *churner* will have a similar pattern of communications in the two complimentary time periods. This hypothesis is based in the fact that social networks are relatively stable in time and so there is a tendency that an individuals community does not change significantly in time. This is a novel methodology, since this is a subject that has not been studied in detail and it results on a formula to determine the similarity between two individuals that is based on the clients that they contact in complimentary time periods. Some simulations were performed with this methodology, where we applied this to real data which was anonymized and provided by a mobile network.
Keywords: Churn, Mobile Communications, Social Network Analysis

Índice

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução e Problema

1.1 Introdução

Neste primeiro capítulo da dissertação vamos introduzir o tema, motivando inicialmente o seu estudo. Serão também apresentados os objetivos do trabalho bem como o modo como este está organizado. De seguida, será descrito o problema.

1.1.1 Motivação

A motivação deste trabalho tem origem num problema de negócios real de uma operadora de telecomunicações nacional. Este tipo de empresas possui grandes bases de dados, com muito potencial e ainda muito por explorar. Com tal volume de dados a crescer tão rapidamente, a sua exploração pode-se apresentar como bastante lucrativa e como uma possibilidade de adquirir vantagem competitiva sobre as empresas concorrentes. Daí resulta que as telecomunicações sejam um dos mais avançados e interessados setores na era do *Big Data*, utilizando o conhecimento obtido da exploração dos dados para conhecer melhor os clientes. Deste modo, é possível criar estratégias orientadas para cada tipo de clientes, procurando maior eficácia das campanhas e a maximização dos ganhos potenciais a serem retirados de cada cliente. Os contatos efetuados de uns clientes para outros, que estão nos registos da empresa, permitem analisar as interações entre estes e identificar padrões. Torna-se assim possível para cada cliente, identificar por exemplo, quais são os seus contactos mais frequentes ou qual a densidade do seu entorno, isto é, com quantos clientes ele comunica. Pode-se tornar também interessante estudar a evolução do entorno de um cliente, utilizando esta informação para prever a sua possível insatisfação, por exemplo. Deste modo, a operadora pode atuar atempadamente, evitando potenciais danos nas suas receitas. Podemos também verificar o caso em que, o facto de um cliente comunicar com muitos outros clientes, possa ser indicativo da sua influência na rede. E que assim, o seu papel mais central na rede possa ser utilizado em benefício da operadora. Aspetos como a localização, tipo de serviços utilizados ou a

sua movimentação podem também ser utilizados para inferir conhecimento sobre os clientes.

Portanto, a possibilidade de extrair conhecimento que está subjacente aos dados através de técnicas de *data mining*, permite um avanço no *Customer Relationship Management*(CRM), isto é, na forma como a empresa se relaciona com cada cliente. O CRM permite personalizar a relação com cada cliente, promovendo ganhos quer para o cliente, quer para a operadora. Assume-se também como uma mais-valia na concorrência com outras operadoras, dado que, enquanto estas não obtiverem um avanço semelhante, os ganhos serão exclusivos da empresa que inovou primeiro.

Atualmente, em termos de práticas das operadoras de telecomunicações a predominância reside na previsão do abandono dos clientes e na construção do perfil destes. Através da sua segmentação, é possível a criação de estratégias e campanhas de marketing específicas para cada segmento. Dada a saturação do mercado de comunicação móvel em Portugal, torna-se cada vez mais difícil a captação de novos clientes. Deste modo, assiste-se a um desvio das atenções das operadoras de telecomunicações da captação de clientes para a sua retenção. Também aqui existe maior riqueza de dados em relação aos atuais clientes, ajudando também a que o custo da retenção seja menor do que o custo da angariação de clientes.

Partindo desta ideia, o *churn* assume-se como uma das maiores ameaças nas operadoras de telecomunicações. Prevenir o *churn*, identificando previamente os clientes com maior potencial para abandonarem a operadora, tem assim bastante relevância na retenção de clientes, dado que permite à operadora direcionar com eficácia a sua estratégia de retenção. No entanto, para que de cada estratégia seja retirado o maior valor possível é preciso que estas sejam bem direcionadas. Assim torna-se pertinente a segmentação do *churn* pelas diversas razões que o motivam e que serão apresentadas posteriormente.

A ideia deste trabalho surge neste seguimento. Será focado o *churn* rotacional, ainda pouco estudado, mas com relevância dentro do *churn*. O objetivo será identificar os clientes que, deliberadamente, subscrevem novo serviço junta da operadora, abandonando o cartão antigo sem que a operadora saiba que se trata do mesmo cliente. Deste modo, clientes deste tipo estarão a enviesar os dados relativos ao *churn*, sobrestimando-os. Também a taxa de sucesso relativa a campanhas de captação de clientes será sobrevalorizada. Abandonam assim a operadora por pura conveniência de usufruir das regalias apenas disponíveis a novos clientes. Dado que estaremos a lidar com tarifários pré-pagos, pode também acontecer que determinado cliente abandone um tarifário e o cartão a este associado e adira a um novo pré-pago junto da empresa, sem que esta tome conhecimento. A identificação destes clientes será feita com base na comparação do padrão de contactos entre clientes que abandonaram a empresa e clientes suspeitos de se tratarem da mesma pessoa, sendo que apenas serão utilizados os registos das chamadas entre clientes.

Fica assim visível a necessidade de segmentar os vários tipos de *churn* e de criar estratégias específicas para cada segmento. Para além disso, a identificação deste

tipo de clientes pode ser utilizada para extrair conhecimento sobre o seu padrão de contactos e utilizada para identificar futuramente os clientes com maior potencial de virem a cometer *churn* rotacional.

Os dados utilizados neste trabalho provêm de uma das maiores operadoras de telecomunicações nacionais e estão anonimizados. O estudo vai envolver uma incursão na área de análise de redes sociais e gestão de grandes bases de dados.

1.1.2 Objectivos

O principal objetivo deste trabalho é a criação de uma metodologia que permita a identificação dos clientes que abandonam a operadora e regressam, num momento posterior, com nova subscrição e sem que a empresa tenha conhecimento que se trata do mesmo indivíduo. Deste modo, é possível detetar casos em que o *churn* é acionado por conveniência. Isto é, em períodos complementares pretendemos a identificação de indivíduos semelhantes, com base nas suas comunicações com outros clientes.

Em relação a este problema em concreto (identificação de *churn* rotacional), ainda não existem metodologias propostas na literatura, pelo que será criada uma que responda às necessidades da empresa e ao problema em questão.

1.1.3 Organização

No capítulo 2 será feita uma revisão da literatura existente. Esta começa com um enquadramento histórico e teórico da Análise de Redes Sociais. De seguida são apresentadas as suas estatísticas e tipos de rede e complementaremos esta revisão com uma apresentação das aplicações da Análise de Redes Sociais.

As ego-redes também serão revistas, dado que representam um potencial *churner* e o seu entorno. Será explorado o seu conceito e quais as medidas utilizadas para o estudo deste tipo de redes e serão também referidos alguns casos de estudo que recorrem a ego-redes.

Posteriormente, surge uma revisão sobre as propriedades das redes reais bem como dos dados das telecomunicações, dado que lidaremos ao longo do estudo com dados deste tipo.

Por último, temos a revisão do tema que mais será focado neste estudo, o *churn*. Assim, começaremos por explorar o seu conceito e as suas motivações, avançando para algumas aplicações. Será também focada a pesquisa efetuada em *churn* rotacional.

1.2 Problema

O objetivo do trabalho, como já foi anteriormente referido, é a identificação dos clientes que abandonam a operadora mas que voltam a subscrever um serviço com

uma identificação diferente, num momento posterior. Assim, torna-se necessário explorar o conceito de *churn* e de um modo mais aprofundado o conceito de *churn* rotacional, aquele que é o tema central desta dissertação e para o qual será construída uma metodologia.

O *churn* é entendido por Mattison (2006) como uma situação em que, um cliente que habitualmente adquire um produto ou serviço a uma empresa, utiliza o seu direito de comprar esse serviço ou produto a uma empresa concorrente. Este autor aborda um tipo de *churn* que tem semelhanças com aquele que iremos explorar, nomeadamente o *churn* promocional. Segundo Mattison (2006), este diz respeito a uma situação em que o cliente abandona a empresa para poder usufruir de vantagens promocionais numa empresa concorrente. No caso do *churn* rotacional, esta pode ser uma das motivações para abandonar a operadora e regressar como novo cliente, ainda que se trate do mesmo. Utilizando tarifários pré-pagos, os *churners* são considerados por este autor como todos aqueles que não carregaram o seu telemóvel nos últimos três meses. Assim, um *churner* é um cliente que abandonou a operadora. Isto pode ter acontecido por decisão do próprio ou por decisão da operadora. A falta de carregamentos, tal como Mattison (2006) indica, ou a falta de utilização do equipamento durante algum tempo podem conduzir a que a operadora decida afastar o cliente, considerando-o *churner*. Isto é, como tendo deixado de usufruir do seu serviço. Neste estudo iremos considerar um cliente como *churner* quando este deixar de efetuar comunicações algum tempo antes do fim dos dados disponíveis. Existem seis meses de dados e a maior parte dos clientes da operadora irão, com naturalidade, realizar comunicações até ao fim do mês 6. Se a última comunicação encontrada for realizada no mês 4, por exemplo, isto significa cerca de dois meses sem comunicações. Logo, não será de esperar que este cliente continue na operadora, sendo por isso considerado *churner*, pela sua ausência de comunicações durante um período prolongado. É também conveniente referir que apenas serão utilizados os registos anonimizados das chamadas entre os clientes para quantificar a semelhança dos *churners* com outros clientes e o facto de serem apenas considerados clientes de alguns tarifários pré-pagos. Para a adesão a tarifários deste tipo não é exigida muita informação ao cliente, logo, a informação disponível sobre estes clientes é pouca.

Assim, neste trabalho introduzimos o conceito de *churn* rotacional, ainda pouco estudado e que marca a diferença do *churn* pela sua motivação e pelo comportamento seguido pelo cliente. Por isso, trata-se de *churn* rotacional quando um cliente utiliza o seu direito de abandonar a empresa mas volta à mesma com uma identificação diferente, num momento posterior. Para a empresa, tratar-se-á de um novo cliente e será recompensado como tal mas ao mesmo tempo será contabilizado como *churner*. Ou seja, o mesmo cliente, duas identificações distintas. Ao selecionarmos para este problema apenas alguns tarifários pré-pagos, pode também acontecer que um cliente abandone um tarifário deste tipo e mude para outro tarifário da empresa sem que esta tenha conhecimento disso. Assim, este trabalho permite às empresas identificar os clientes de tarifários pré-pagos que saem da operadora e que voltam para um

tarifário desse tipo com uma identificação diferente, ou que migram de tarifário não dando conhecimento do abandono do primeiro. Entende-se assim a motivação das empresas em detetar estes casos que não correspondem à realidade e que levam a que o cliente seja tratado de uma forma diferente do que deveria ser.

Tal como foi referido, tendo por base um tarifário pré-pago, a informação sobre o detentor de determinado cartão não é muita. Fica assim mais fácil trocar de cartão e aproveitar eventuais promoções para novos clientes. Deste modo, um cliente-tipo de *churn* rotacional, fará uma última comunicação em determinado dia e com determinada identificação e será considerado *churner* desse dia. Regressará, com nova identificação e como novo cliente, num momento posterior. Em termos analíticos, a base disponível é informação anonimizada de comunicações entre os clientes. Assim, entre outros, ficamos a saber quem ligou para quem.

Deste modo, propomos a seguinte hipótese:

- Um cliente suspeito de ser o mesmo indivíduo que determinado *churner*, terá um padrão de contactos semelhante, nos dois espaços temporais complementares.

Baseamos esta hipótese no facto de as redes sociais serem relativamente estáveis ao longo do tempo e portanto, de existir uma tendência para que a comunidade de determinado indivíduo não sofra grandes alterações ao longo do tempo.

Assim, mais à frente, e a nível individual, analisaremos com quem é que os *churners* contatam (entorno) e com que frequência, e procuraremos indivíduos que apresentem um entorno semelhante. Entorno será o termo utilizado para indicar o conjunto de indivíduos com quem um determinado cliente contata. Vamos supor a existência de um indivíduo que descarta o seu cartão para usufruir de uma promoção apenas disponível a novos clientes e que este cliente contata com determinadas pessoas. Ao utilizar o novo cartão e desfrutando do estatuto de novo cliente terá tendência a contatar com as mesmas pessoas. Deste modo, o desafio será encontrar pares de clientes em espaços temporais complementares que tenham um grau de semelhança elevado.

Este tipo de aproveitamento e abandono da operadora tem atualmente barreiras bastante reduzidas. Com tarifários pré-pagos, o abandono da operadora limita-se a deixar de utilizar o cartão, simplesmente descartando-o. Deste ponto de vista, torna-se aliciante a hipótese de trocar para uma operadora concorrente que ofereça melhores condições, ou mesmo manter-se na operadora mas como novo estatuto, que é exatamente aquilo que queremos identificar.

1.2.1 Sumário

Neste capítulo começamos por motivar o tema em estudo, partindo dos benefícios que as empresas podem obter ao utilizarem os dados que têm ao seu dispor. Evoluímos

para o tema que será focado, o *churn* rotacional, que terá apenas como base os registos anonimizados das chamadas entre os clientes e clientes de alguns tarifários pré-pagos. De seguida foi explorado o problema em estudo, que ainda é recente a nível de investigação, mas que no entanto tem contado com cada vez mais atenção. Alguns conceitos foram apresentados de forma a contextualizar a importância do *churn* rotacional. Isto é, a sua definição, a sua importância, bem como dos dados de que partiremos para alcançar resultados.

No próximo capítulo será apresentada uma revisão da literatura existente sobre alguns temas considerados relevantes no nosso estudo.

Capítulo 2

Revisão da Literatura

2.1 Análise de Redes Sociais

Uma rede social pode ser descrita como a representação visual das relações e interações entre um grupo de indivíduos, de acordo com Kempe et al. (2003). Assim, começaremos por estudar o modo como estas interações são visualizadas, recorrendo à utilização da teoria dos grafos e a sua ligação às redes.

Uma das bases da análise de redes sociais é a sociologia. Esta permite estudar o comportamento de um indivíduo ou de um grupo, ajudando a que se produzam resultados bastante úteis, não só para sociólogos ou psicólogos mas também a nível de marketing e de negócios. Existem mais disciplinas a basear esta área interdisciplinar. De acordo com Oliveira and Gama (2012) são consideradas como principais as seguintes: psicologia social, antropologia, física, matemática, ciência computacional.

O trabalho desenvolvido por Scott (1988), permite-nos perceber o enquadramento histórico do tópico de análise de redes sociais. O principal impulso foi dado por Jacob Moreno, ao criar o sociograma. Este é um gráfico com linhas e pontos cujo objetivo era representar as relações existentes no grupo que Moreno estava a estudar. Por volta de 1930 surgiram grupos de investigação na área da psicologia cognitiva e social que estudavam as dinâmicas sociais e possíveis métricas de fenómenos sociais, bem como estudos relacionados com as comunidades e a interdependência entre os indivíduos. Nesta altura, surgiram duas correntes de investigação distintas. A análise sociocêntrica foi a primeira e é oriunda da Universidade de Harvard incluindo a quantificação de relações e o estudo de padrões estruturais da rede. A segunda corrente, análise egocêntrica, surgiu na Universidade de Manchester. Aqui, o objetivo é estudar as relações de um indivíduo em específico, em vez de estudar toda a rede (Chung et al., 2005). Só em 1960 se começou a associar esta investigação a técnicas matemáticas, sendo esta a base da Análise de Redes Sociais que hoje temos, com uma vasta gama de estatísticas que ajudam a medir propriedades de determinada rede social.

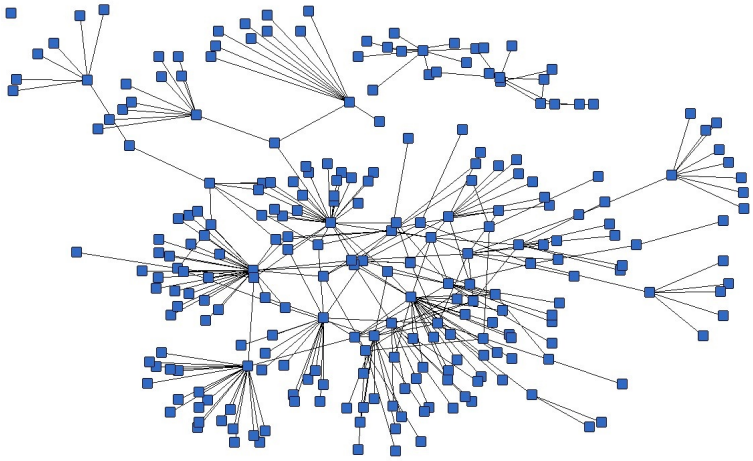

Figura 2.1: Exemplo de uma rede social

Uma rede pode ser definida como um conjunto de objetos (nós ou vértices) com ligações (arestas) entre si e o seu estudo na forma matemática provém da teoria de grafos. Assim, um grafo trata-se de um conjunto de vértices e arestas, no qual é possível aplicar técnicas matemáticas e em que ao mesmo tempo é possível a visualização das relações entre as entidades. A teoria dos grafos foi primeiramente aplicada por Euler na resolução do problema das pontes de Königsberg (Newman, 2003).

Na Figura 2.1 temos um exemplo de uma rede social, representada através de um grafo.

As redes sociais podem ser representadas de distintos modos. O mais comum é a representação através de grafos matemáticos, que são compostos por arestas e vértices. Cada aresta é definida por um par de vértices e estes podem corresponder a várias entidades: pessoas, países, organizações, plantas, entre outros. De um modo semelhante, uma aresta une dois vértices e assim, pode também representar vários tipos de relações: amizade, troca, comunicação, cooperação, entre outros.

Formalmente, um grafo G é composto por um conjunto não-vazio V(G) de vértices e por um conjunto de arestas E(G), sendo definido por G=(V(G),E(G)).

Segundo Oliveira and Gama (2012) existem na literatura duas estruturas de dados apropriadas para o armazenamento e posterior análise e representação das redes: estruturas de lista e estruturas de matriz. A primeira possibilidade é ideal para grafos esparsos. Pelo contrário, as estruturas de matriz, que incluem por exemplo as matrizes de adjacência, são úteis para a representação de matrizes densas.

Nas Figuras 2.2 e 2.3 apresentamos exemplos da representação de grafos direcionados e não direcionados em listas ou matrizes de adjacência.

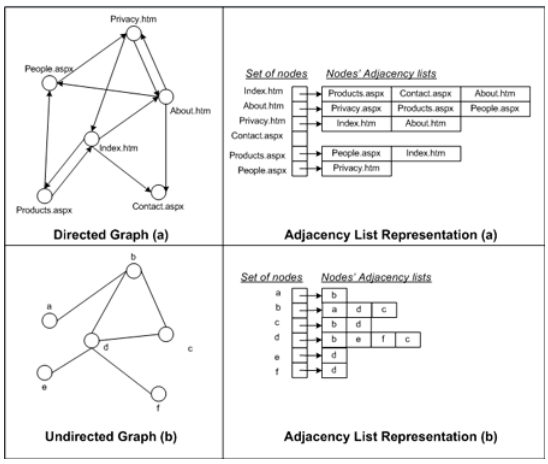

Figura 2.2: Grafos direcionados e não-direcionados representados através de uma lista de adjacência.

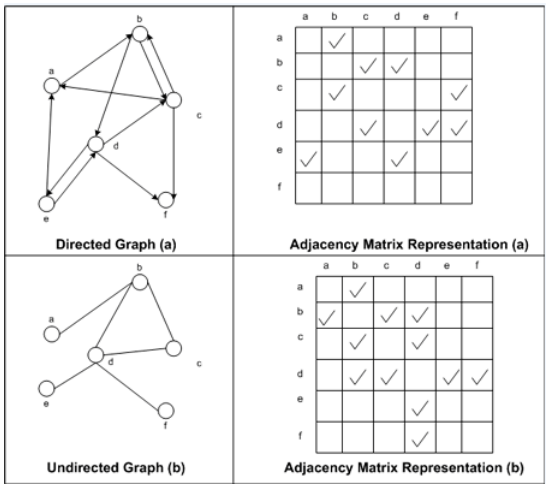

Figura 2.3: Grafos direcionados e não-direcionados representados através de uma matriz de adjacência.

2.1.1 Tipos de redes e Estatísticas

Tipos de redes

Uma das principais propriedades das redes é a direção das suas ligações. Assim, estas são direcionadas se partirem de um nó para outro, ou não direcionadas, caso as ligações sejam recíprocas. Podem também ser ponderadas, sendo que neste caso a relação entre os nós é quantificada. Por exemplo, para representar a frequência da chamada entre diversos indivíduos devem-se usar pesos nas ligações, originando possivelmente ligações com maior valor do que outras. Caso as ligações da rede não sejam ponderadas, isto significa que nenhuma terá mais valor do que outra, assumindo todas a mesma importância.

Estatísticas

Existem diversas estatísticas que permitem o estudo de uma rede e a extração de conhecimento das mesmas. Este ponto ganha relevância com as redes de tamanho elevado que temos hoje em dia, dado que sem estas estatísticas seria complicado perceber a sua topologia e a forma como estão organizadas. Assim, as estatísticas de redes podem ser calculadas ao nível dos nós ou ao nível da rede.

As medidas clássicas de centralidade que permitem quantificar a centralidade dos nós na rede, podem ser relativas aos nós ou à posição destes na rede. No primeiro caso, podem incidir sobre as suas ligações diretas, no caso de ser estudado o seu grau de centralidade (do inglês *degree*), o grau de entrada ou o grau de saída. Mas também podem ter como objetivo estudar a posição de um nó na rede, através das seguintes estatísticas: grau de intermediação, grau de proximidade, grau do vetor próprio e coeficiente de coesão local.

O grau é medido através da quantificação do número de ligações diretas de um indivíduo. Caso a rede seja direcionada, temos o grau de entrada, que são as ligações cujo destino é o indivíduo em análise, e o grau de saída, ligações que têm como origem o indivíduo em análise. O valor do grau pode ser utilizado para interpretar o prestígio de determinado indivíduo, que pode ser dividido em suporte ou influência caso tenhamos redes direcionadas. O grau de entrada estará associado ao suporte e o grau de saída à influência.

A intermediação verifica, tal como o nome indica, quantas vezes determinado indivíduo está em posição de intermediário. Estar nesta posição significa que está no caminho mais curto entre outros dois indivíduos. Assim, o valor de intermediação será tanto maior, quanto maior for o número de pares de quem certo nó é intermediário. Esta é uma posição crítica e que se pode apresentar como muito vantajosa, dado que muitas vezes estarão a controlar a informação que circula entre grupos, por exemplo.

O grau de proximidade é calculado para cada nó e tem como objetivo identificar a posição do indivíduo na rede e a sua proximidade a outros nós. Para isso, utiliza

não só as suas ligações diretas mas também a distância mais curta deste nó para cada um dos restantes elementos da rede. Deste modo, o grau de proximidade de um nó traduz-se no valor médio de todos caminhos mais curtos entre este e os outros nós da rede. Permite-nos assim saber, em média, quantos ligações são necessárias para ligar um nó a qualquer outro elemento da sua rede.

Estas três medidas referenciadas (grau de centralidade, grau de intermediação e grau de proximidade) foram propostas por Freeman (1979).

O grau do vetor próprio valoriza não só se o indivíduo está bem conectado mas se está ligado a outros indivíduos também bem conectados. Isto significa que, um indivíduo bem conectado e com uma vizinhança bem conectada terá mais poder do que um indivíduo com ligações fortes mas em que essas ligações não correspondam com outras ligações de qualidade. Assim, esta estatística valoriza mais a qualidade das ligações, diminuindo a relevância da sua quantidade e serviu também como base à construção do algoritmo PageRank, da Google, que assenta nesta ideia.

A última estatística relativa aos nós é o grau de coesão local. Este surge da transitividade existente nas redes, que significa que se A é amigo de B e B é amigo de C, então é provável que A também seja amigo de C. Esta propriedade é quantificada pelo coeficiente de agrupamento, que pode ser global, caso seja calculado ao nível da rede, ou local, caso seja calculado para cada nó. Neste caso, a transitividade será tratada como uma propriedade local da vizinhança de um nó e dar-nos-á informação acerca do nível de coesão entre os vizinhos de um determinado indivíduo.

As estatísticas utilizadas ao nível da rede permitem descrever a sua estrutura. As medidas mais populares são as seguintes: diâmetro e raio, distância geodésica média, grau médio, reciprocidade, densidade e coeficiente de coesão global.

Tanto o diâmetro de uma rede como o raio se baseiam no conceito do caminho mais curto entre dois nós. Assim, o diâmetro corresponde ao comprimento máximo de todos os caminhos mais curtos encontrados, enquanto que o raio é dado pelo comprimento mínimo dessa distância. Estas medidas são importantes para analisar a proximidade dos indivíduos verificando, por exemplo, qual a maior distância geodésica possível entre quaisquer dois nós na rede.

A distância geodésica média corresponde à média do comprimento de todos os caminhos mais curtos entre todos os pares de nós da rede. Deste modo é possível saber, em média, quantos passos são necessários para alcançar qualquer nó da rede.

O grau médio é a média do grau de todos os nós da rede. Diz-nos portanto, em média, quantas ligações tem cada nó.

A reciprocidade é uma estatística apenas calculada em redes direcionadas e quantifica a tendência para existirem ligações mútuas entre os nós. Isto é, qual a probabilidade de dois nós na rede terem ligação direta recíproca.

A densidade permite-nos avaliar a conetividade geral da rede. É calculada como a proporção de ligações existentes na rede em relação ao número total de ligações possíveis. Se não existirem ligações entre nós na rede o valor da densidade será zero. Caso o valor da densidade seja igual a um, a rede apresenta conetividade perfeita.

Isto é, todos os nós estão ligados entre si.

O coeficiente de coesão já foi abordado nas estatísticas relativas aos nós, mas num contexto local. Agora, o objetivo é estudar a transitividade em toda a rede. Existem vários métodos para calcular este valor, sendo que um dos possíveis é através da média dos valores do coeficiente de coesão local.

2.1.2 Aplicações

A área da análise de redes sociais tem sido alvo de uma crescente investigação nos últimos anos. A oportunidade de estudar o comportamento dos indivíduos e a sua relação com outros indivíduos e a posição que estes assumem em determinada comunidade aparece como bastante relevante, seja em termos de marketing, saúde, mercado das telecomunicações, entre outros. Assim, várias aplicações são possíveis com análise de redes sociais, apresentamos algumas: previsão de *churn* , gestão de campanhas (para aumentar o valor da marca ou as vendas de determinado produto, por exemplo), deteção de clientes influentes, difusão da informação, publicidade online, entre outras.

Um dos setores de investigação da análise de redes sociais está relacionado com o estudo de como a informação se difunde pela rede. Como exemplo, temos os trabalhos de Kempe et al. (2003) e Kiss and Bichler (2008). O primeiro tem como objetivo responder à seguinte pergunta: se queremos convencer alguns indivíduos a adotar um novo produto e depois desencadear uma cascata de novas adoções, quais são os indivíduos que devemos selecionar inicialmente? A adoção de um produto é apenas um exemplo de motivação para a determinação dos clientes influentes. Este autor utilizou dois modelos distintos em comparação com a seleção aleatória, aplicando-os numa base de dados acerca da coautoria de trabalhos em publicações da área da física.

O estudo de Kiss and Bichler (2008) incorpora a simulação de processos de difusão em redes sociais e outros tipos de redes e conclui que, em redes de clientes, as medidas de centralidade provaram ser bastante melhores que a seleção aleatória, com um *lift* comparativamente muito elevado. O *lift* é o rácio do número de consumidores alcançados pelo número de consumidores selecionados.

Bonchi et al. (2011) apresenta também algumas aplicações da ARS nos negócios, nomeadamente no setor das telecomunicações. Deste modo, considera o *churn* como sendo, provavelmente, a mais importante aplicação de negócios da ARS e no setor das telecomunicações, dado que o serviço oferecido está fortemente ligado à rede social do cliente. Aponta também a deteção de clientes influentes, identificação de comunidades e o modo como a informação se propaga como determinantes na previsão do *churn*. Este autor refere também a deteção de fraude, a atribuição de valor à reputação, confiança e lealdade como práticas comuns nas empresas utilizando como base os dados sociais disponíveis. Para além disso, a deteção de comunidades e o estudo de como estas evoluem ao longo do tempo são utilizadas para a criação

de estratégias de marketing mais direcionadas a cada tipo de cliente, bem como para melhorar a oferta de produtos e serviços nas redes sociais *online*. Na prática, esta informação pode ser utilizada para recomendações sociais *online*, para avaliar a lealdade do consumidor e até por questões de segurança, como seja a prevenção do terrorismo, entre outros.

2.2 Ego-Redes

Por volta de 1930, como já foi referido, surgiram duas abordagens distintas sobre as redes. A primeira, sociocêntrica, engloba o estudo de toda a rede, das suas características estruturais, dos padrões observados. Deste modo, procuram-se generalizar estes padrões estruturais tornando possível a criação de modelos que expliquem os diversos comportamentos sociais. Esta abordagem permite, por exemplo, estudar como é que as características estruturais da rede explicam a concentração de poder ou de outro recurso, dentro de um grupo (Chung et al., 2005).

A abordagem egocêntrica centra-se em apenas um indivíduo. Assim, temos o nosso indivíduo alvo, o ego, e aqueles com quem ele detém alguma relação, os alters. A análise inclui estudar a relação do ego com os seus alters e a relação dos alters entre si (Chung et al., 2005).

2.2.1 Conceito de Ego-Rede

Friedman and Aral (2001) avançam que os dados provenientes de ego-redes podem ser analisados de duas formas: pelos atributos individuais de cada indivíduo ou como um conjunto de relações.

Como já referimos uma ego-rede será composta pelo ego, o elemento central, e pelos seus alters, isto é, os indivíduos que com ele está relacionado diretamente. Poderá ser analisada a relação do ego com os seus alters e as relações entre alters.

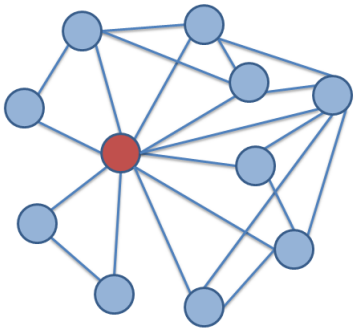

Figura 2.4: Exemplo de uma ego-rede

Na Figura 2.4 temos um exemplo de uma ego-rede.

Por exemplo, se o ego for intermediário de comunicação da maior parte dos pares de alters, este fator aumenta o seu poder, dado que provavelmente conseguirá controlar uma grande parte da informação que circula na rede. Por outro lado, se os alters conseguem comunicar com facilidade entre si, então o poder do ego é diminuído, dado que a dependência dos alters face ao ego é menor.

Hanneman and Riddle (2005) introduzem o seu estudo com a referência à utilidade das ego-redes para o estudo do comportamento de um indivíduo e não de um grupo, bem como o seu foco nas relações entre os indivíduos, em detrimento das suas posições na rede.

2.2.2 Medidas de Análise das Ego-Redes

No decorrer do seu estudo, Hanneman and Riddle (2005) utilizam o software *Ucinet* (Borgatti et al., 2002) e descrevem as medidas providenciadas pelo mesmo. Assim, as medidas consideradas relevantes para a caracterização da vizinhança dos egos em análise são as seguintes: Tamanho, Ligações, Número de Pares, Densidade, Distância Geodésica Média e Número de componentes fracas.

O tamanho da rede é indicado pelo ego mais os nós que estão diretamente ligados a este.

O número de ligações é, tal como o nome indica, o número de conexões existentes entre todos os elementos da rede

O número de pares representa o número de ligações diretas possíveis entre os nós.

Densidade é o resultado do número de ligações dividido pelo número de pares. Mostra-nos assim a percentagem de ligações possíveis que estão presentes na rede. Deste modo, quantas mais ligações existirem, mais conexão existirá entre os nós e mais densa a rede será. Este índice poderá afetar a criação de oportunidades e constrangimentos para os elementos da rede.

A distância geodésica média corresponde à média do comprimento dos caminhos mais curtos. Isto é, dados dois nós e as várias hipóteses de se conectarem, será utilizado aquele que for mais curto e posteriormente calculada a média de todos esses valores. Deste modo saberemos, em média, qual a distância de um nó a outro.

Uma componente fraca é composta pelo maior número de elementos da rede ligados entre si, não considerando a direção das ligações. Suponhamos que A,B,C e D são nós de uma rede egocêntrica e que A está diretamente ligado a B tal como C está diretamente ligado a D. No entanto, A e B não têm qualquer via de contato com C e D, à exceção de todos terem uma conexão com o ego. Deste modo, nesta rede, estamos perante duas componentes fracas, dada a existência de duas comunidades distintas apenas ligadas pelo ego. Esta medida pode ser normalizada, tornando possível a interpretação independente do tamanho da rede.

Na Figura 2.5 podemos verificar a existência de uma componente fraca, composta pelos nós A,B,C e D e de uma componente forte (que considera a direção das ligações) que inclui apenas os nós B,C e D.

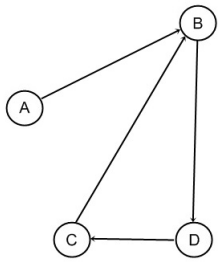

Figura 2.5: Componentes fraca e forte numa rede

As falhas estruturais (*structural holes*) resultam da estrutura de ligações entre os nós que podem colocar um indivíduo em posição mais vantajosa ou desvantajosa no contexto da rede. Este termo foi popularizado e formalizado por Burt (2009) que considera que duas pessoas serão equivalentes a nível de estrutura se possuírem os mesmos contatos.

Consideremos a figura 2.6 onde temos três nós: A, B e C. Cada um está diretamente ligado aos outros dois, o que coloca todos os indivíduos, em termos de estrutura, em posição de igualdade, dado que não há nenhum em posição mais vantajosa que o outro. Todos os elementos têm duas alternativas para comunicar. Supondo todos os indivíduos no mesmo nível, a posição de A para negociar com C não é superior, dado que C terá sempre a alternativa de negociar com B, diminuindo assim o poder de A. Nesta figura não existe falha estrutural.

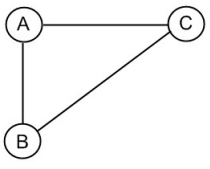

Figura 2.6: Exemplo de uma rede sem falha estrutural

Agora vejamos a figura 2.7, onde foi criada uma falha estrutural entre B e C. Deste modo, estes indivíduos não poderão comunicar diretamente e terão de o efetuar através de A. Nesta rede, A já estará numa posição vantajosa, devido à falha

estrutural entre B e C que impede a sua comunicação, sem ser por intermédio de A. Enquanto A tem duas alternativas para comunicar, B e C só têm uma.

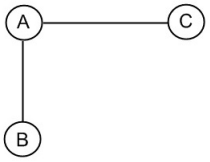

Figura 2.7: Exemplo de uma rede com falha estrutural

Quanto maior for a dimensão da rede, menor será a sua densidade e mais falhas estruturais potencialmente terá, aumentando também assim as fontes de desigualdade.

2.2.3 Aplicações

O estudo de Chung et al. (2005) aborda um exemplo da aplicação das ego-redes na área da saúde. Assim, o objetivo passa por perceber os processos sociais subjacentes aos médicos e entre estes que, afetam a sua taxa de adoção de um medicamento. Pelos dados de cada médico estes foram divididos em duas categorias: orientados para a profissão ou para o paciente. Foi descoberto que os médicos mais orientados para a profissão foram mais rápidos a prescrever o medicamento e os dados relacionais sugeriram que os médicos que estão mais integrados com os seus colegas são mais rápidos a adotar a prescrição do medicamento do que aqueles que estão mais isolados na rede. Este trabalho foi desenvolvido por Coleman et al. (1957).

Temos também o estudo de Fisher (2005) que refere dois projetos: *Soylent* e *The roles*. O primeiro baseia-se em emails e nos padrões verificados nas interações sociais. Este projeto tem início com o ego que enviou um email para um determinado número de pessoas diferentes e o foco está no estudo da vizinhança deste, tentando perceber as suas conexões uns com os outros e no contexto de toda a rede. O segundo projeto tem como foco cada indivíduo de cada fórum de discussão de modo a perceber qual é o seu papel na rede através dos padrões e características estruturais da rede.

Por fim temos mais um estudo relacionado com a área da saúde, da autoria de Argeseanu et al. (2013). Este foi baseado na evidência de recentes investigações da área, que sugerem a importância das conexões sociais para o risco de doenças crónicas e incidência das mesmas. Assim, o primeiro objetivo do trabalho é quantificar e caracterizar as ego-redes de uma amostra representativa da comunidade urbana de Delhi. O segundo objetivo passa por analisar os padrões e a extensão do risco de

doenças cardiovasculares dentro das redes observadas. Foi concluído que o tamanho da rede era útil para previsão, isto é, pequenas redes são associadas a eventos negativos relacionados com a saúde. Também a maior heterogeneidade da rede está relacionada com o menor risco de mortalidade no geral, entre outras associações.

2.3 Propriedades das Redes Reais e Dados de Telecomunicações

2.3.1 Propriedades das Redes Reais

As redes reais são caracterizadas por padrões não triviais. Ou seja, não são totalmente aleatórias, nem totalmente regulares. Este tema foi sumarizado e estudado por Newman (2003), que apresenta como principais caraterísticas comuns entre as redes, as seguintes: efeito "pequeno mundo", transitividade ou *clustering*, distribuição da probabilidade do grau, resiliência da rede, homofilia (padrões de mistura), correlações do grau, estruturas de comunidades, navegação de rede.

O efeito do pequeno mundo, foi inicialmente estudado por Milgram (1967) e pretendia responder à seguinte pergunta: Dados dois indivíduos aleatoriamente escolhidos e sem relevar a distância que os separa, quantos intermediários serão necessários para os unir? Segundo este estudo, o número de intermediários entre o indivíduo inicial e o final rondava os seis na maior parte dos casos. Este valor foi popularizado depois por Guare (1990) que reforçou o poder das conexões e de que seriam suficientes, em média, seis ligações para conectar quaisquer dois indivíduos. O efeito do pequeno mundo oferece-nos mais uma propriedade interessante e peculiar das redes. Milgram (1967) mostrou que eram necessários, em média, poucos intermediários para ligar quaisquer dois indivíduos da rede. A isto, Kleinberg acrescenta, segundo Newman (2003) que esta ligação entre os indivíduos não é feita aleatoriamente, mas que os indivíduos são bons a identificar qual o melhor vizinho para atingir determinado alvo. Isto é, as pessoas não têm noção de toda a estrutura da rede para decidir o caminho ótimo a seguir mas sim conhecimento dos seus vizinhos e de quais poderão ser mais eficientes para atingir o alvo, o que permite que a ligação entre dois indivíduos possa ser mais curta mesmo com grandes distâncias a separá-los.

Newman (2000) referiu a elevada população mundial, mas também que a estrutura das redes sociais é de tal forma, que todos estamos muito próximos uns dos outros. Esta propriedade é interessante, dado que menoriza a influência do número de nós na distância entre dois indivíduos aleatoriamente escolhidos na rede.

A transitividade demonstra a ideia de que: se A está ligado a B e B está ligado a C então a probabilidade de que A esteja ligado a C é fortificada. Advém da ideia de "o amigo do teu amigo, é provável que também seja teu amigo". Este fenómeno da transitividade promove a formação de triângulos na rede, aumentando a probabilidade de estes ocorrerem. Supondo uma rede com três nós em que, num

momento inicial, apenas dois estão ligados ao terceiro nó, é então provável que num momento posterior todos se liguem entre si formando então um triângulo.

A distribuição de probabilidade do grau na maior parte das redes reais mostra que existem muitos indivíduos com baixo grau de centralidade e poucos indivíduos com valor alto neste índice, conforme referido por Oliveira and Gama (2012). Esta propriedade é demonstrativa do efeito cumulativo existente na realidade. Do mesmo modo que um rico tem tendência a ficar mais rico, alguém com muito amigos tem maior probabilidade de esse número aumentar, do que alguém com um número de ligações mais reduzido.

A resiliência da rede aborda como a remoção de nós da rede afeta a distância média entre os nós, ao interferir na conetividade entre estes. Se a remoção for realizada de um modo aleatório a rede mostra-se resistente, não tendo um impacto muito grande na distância entre os nós. Este fato justifica-se pela existência, em redes reais, de mais indivíduos com poucas ligações e menos indivíduos com muitas ligações, logo a probabilidade de selecionar nós que provoquem menor impacto é maior. No entanto, se a remoção de nós incidir sobre aqueles com mais ligações então o efeito pode ser devastador. Com poucos nós removidos será possível provocar um grande impacto na rede e na sua comunicação.

Zafarani et al. (2014) classificam de homofilia quando indivíduos semelhantes se tornam amigos. Newman (2003) reforça esta ideia ao afirmar que os indivíduos tendem a associar-se socialmente a indivíduos que partilhem algo em comum, seja faixa etária, gostos comuns, educação, entre outros.

Como acabámos de ver, as redes tendem a apresentar homofilia o que promove a formação de sub-comunidades, apresentando estruturas de comunidade na rede. Estas são compostas por grupos de nós com elevada densidade de ligações entre eles, ou seja, grande interação entre os membros do grupo.

2.3.2 Dados de Telecomunicações

Este tipo de dados têm provado ser um bom *proxy* para as interações sociais, conforme afirmam Candia et al. (2008). Acrescentam também a qualidade da informação passível de ser retirada através da comunicação móvel, como o volume e padrões de comunicação e a localização do dispositivo, bem como a abrangência de espaço e de tempo destes dados, impulsionando assim o estudo da evolução dos mesmos. Eagle et al. (2009) também apontam a vantagem de utilização dos dados de telecomunicações, dado que permitem o estudo de maiores bases de dados, ao invés das recolhidas através de inquéritos, que são usualmente de dimensão reduzida e também a possibilidade de avaliar a evolução ao longo do tempo.

2.4 Churn

2.4.1 Conceito

O conceito de *churn* é entendido por Kazienko et al. (2009) como um processo pelo qual uma empresa perde um cliente para uma empresa concorrente, visão reforçada por Karnstedt et al. (2010). A capacidade de prever quais os clientes com maior potencial de abandonar a operadora torna-se uma mais-valia e potencia um aumento do esforço na sua retenção. Assim, assiste-se a um desvio de atenção da captação de clientes para a sua retenção, o que está diretamente ligado aos custos implícitos a cada alternativa (Dasgupta et al., 2008) (Richter et al., 2010). Richter et al. (2010) apontam a fluidez do mercado como mais uma ameaça aos lucros das empresas, dado que a facilidade dos clientes em se movimentarem de uma operadora para outra lhes fornece maior poder de decisão, aumentando o risco para as operadoras. Neste artigo, é enfatizada a importância da influência e das comunidades existentes, dado que a informação será provavelmente mais confiável dentro destas comunidades e proveniente de indivíduos influentes e de referência. Assim, se o líder de uma comunidade exercer influência suficiente sobre os restantes elementos, a saída desse membro pode ter como consequência a saída de mais elementos, aumentando assim o impacto da decisão.

Mattison (2006) propõe uma classificação do *churn* baseada na motivação dos indivíduos para abandonar a operadora. Começa por definir os *churners* em tarifários pré-pagos, como todos aqueles que não carregaram o seu telemóvel nos últimos três meses. Depois, inicia a divisão em dois grupos: *churn* voluntário e *churn* involuntário. Este último caracteriza-se por situações em que seja a empresa a classificar o cliente de *churner* seja por falta de pagamento ou falta de utilização do serviço. Sempre que o *churn* ocorrer quando o cliente dá início ao fim do seu serviço com a operadora, será considerado como voluntário. Dentro do *churn* voluntário, podemos dividir em acidental ou deliberado. O primeiro caso é incomum e pode-se dever a alterações na vida dos clientes, como por exemplo: condição financeira, morada, entre outros. Isto é, não foi por vontade do cliente, mas porque algo externo o obrigou a abandonar a operadora. Os restantes casos estarão inseridos no *churn* voluntário e deliberado, que acontece quando, por imensas razões possíveis, o cliente usa o direito de abandonar a operadora. Os motivos podem ser tecnológicos, quando o cliente procura numa operadora concorrente a oferta ou disponibilização de melhores equipamentos, melhores condições e mais inovadoras, como sejam novas funções. Aqui pode-se referir o exemplo da internet, mais concretamente da fibra. Aqueles que primeiro avançaram com a sua comercialização, aumentaram a captação de clientes fruto de terem algo que o concorrente não tinha. Outro motivo, e talvez um dos mais comuns, será o preço. A não ser que o cliente tenha elevada lealdade à operadora, considerará o abandono sempre que encontrar uma oferta economicamente mais vantajosa. A qualidade do serviço também assume elevada importância. Aqui

podemos incluir a qualidade e cobertura da internet ou a qualidade da rede disponível para efetuar chamadas, como fatores que podem influenciar a troca de operadora. Depois temos os fatores sociais e psicológicos, que podem incluir a mudança para outra operadora, influenciada por familiares ou amigos. Pode também acontecer que o cliente se associe mais à imagem de uma operadora do que outra, e, não havendo lealdade suficiente a uma, troca para aquela em que se sente mais confortável. Por último, temos o *churn* por conveniência, sendo este difícil de detetar. Pode acontecer através dos agentes vendedores da operadora que também comercializem as operadoras concorrentes. Nestes casos, findo o período contratual e não tendo nada a ganhar com a continuação do cliente na operadora, pode encaminhá-lo para outra, ganhando de novo a comissão de um novo contracto. Este tipo de *churn* é denominado de *churn* de canal, relativo aos canais de distribuição e comercialização da marca. Temos também o *churn* promocional que ocorre quando um cliente toma a decisão de abandonar a operadora com vista a usufruir do estatuto de novo cliente noutra operadora concorrente.

2.4.2 Aplicações

Em termos de investigação e modelos atualmente em uso apresentamos duas visões encontradas: o estudo baseado nas características pessoais dos indivíduos ou na posição destes na rede e na estrutura e dinâmica da própria rede. Dasgupta et al. (2008) calculam a propensão de um indivíduo ser *churner*, dependendo do número dos seus amigos (ligações) que já estão classificados como *churners*, e analisando estes e a sua rede. O *churn* é previsto através de modelos de difusão de influência e o resultado apresentado através do conceito de *lift*, que permite uma aproximação com a realidade de negócios. Assim, este trabalho, valoriza as relações sociais e a influência das ligações.

Kawale et al. (2009) utilizam uma rede de um *Multiplayer Online Role-Playing Game* (*MMORPG*), que se trata de um jogo online passível de ser jogado por milhares de utilizadores. Aqui, o *churn* também assume grande importância, dado que existem jogadores que vão deixando de jogar e isso consiste numa ameaça ao retorno da empresa detentora do jogo. Assim, estes autores realçam a importância da análise de redes sociais (a força das ligações e a estrutura e dinâmica da rede) na previsão de *churn*. A influência existente entre os jogadores e o seu envolvimento no jogo são utilizados para criar um modelo de previsão que detete aqueles que provavelmente se tornarão *churners*. Para isso, recorrem a um modelo de difusão para propagar a influência na rede, analisando assim o modo como as ligações de determinado indivíduo o influenciam.

2.4.3 *Churn Rotacional*

O *churn* rotacional, que nos propomos a identificar neste trabalho tem ainda muito pouca informação científica disponível, sendo que a nível de consultoras já existe um maior avanço e disponibilidade de informação. Ainda assim, Babu et al. (2015) refere o tema, reforçando a necessidade de identificar e prevenir subscritores móveis que se desconectam, e conectam novamente o seu serviço com identificação diferente, de forma a tirar proveito de promoções apenas dirigidas a novos clientes. Este resultado da pesquisa científica, mostra que o problema do *churn* rotacional ainda está numa fase prematura de estudo e portanto ainda não existem metodologias propostas para a sua identificação, nem muitos testes efetuados.

2.5 Sumário

Neste capítulo foi realizada uma revisão da literatura existente sobre alguns temas. Assim, começamos por um enquadramento histórico e teórico da Análise de Redes Sociais. De seguida são apresentadas as suas estatísticas e tipos de rede e complementaremos esta revisão com uma apresentação das aplicações da Análise de Redes Sociais. As ego-redes também foram revistas, dado que representam um potencial *churner* e o seu entorno. Foi explorado o seu conceito e quais as medidas utilizadas para o estudo deste tipo de redes e foram também referidos alguns casos de estudo que recorrem a ego-redes. Posteriormente, surgiu uma revisão sobre as propriedades das redes reais bem como dos dados das telecomunicações, dado que lidaremos ao longo do estudo com dados deste tipo. Por último, temos a revisão do tema que mais será focado neste estudo, o *churn*. Assim, começamos por explorar o seu conceito e as suas motivações, avançando para algumas aplicações. Foi também focada a pesquisa efetuada em *churn* rotacional, que ainda é muito reduzida.

No próximo capítulo será apresentada a metodologia que visa identificar os suspeitos de *churn* rotacional para cada churner, baseando-se para isso apenas nos registos das chamadas entre os clientes e utilizando apenas clientes de alguns tarifários pré-pagos.

Capítulo 3

Metodologia

3.1 Introdução

Este capítulo tem como objetivo a identificação de clientes que abandonam a operadora e regressam com identificação diferente, num momento posterior. Assim, propomos uma metodologia que tem como propósito a identificação de *churn* rotacional tendo como base apenas os registos das chamadas entre os clientes e utilizando apenas tarifários pré-pagos, que não obrigam o cliente a disponibilizar muita informação.

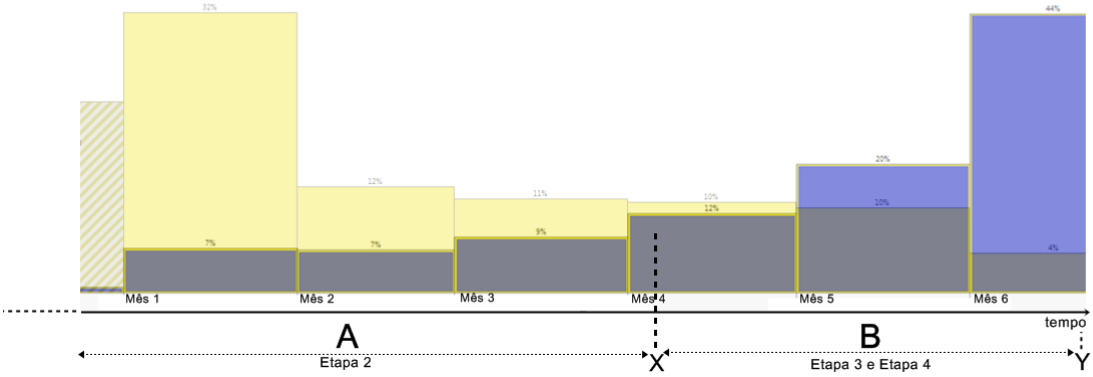

Figura 3.1: Esboço do funcionamento da metodologia

Na Figura 3.1 ilustramos a ideia subjacente à metodologia proposta. O ponto X representa o instante do tempo onde serão identificados os potenciais *churners*, sendo que a atividade desse potencial *churner* será estudada na área A. O ponto Y representa o momento de tempo atual, isto é, onde estamos. Na área B é estudada a atividade dos suspeitos de *churn* rotacional que entraram após o ponto X.

Esta metodologia requer a existência de registos de chamadas antes e depois do

abandono do cliente. Deste modo, não é possível identificar *churners* rotacionais que tenham saído no momento Y, isto é, agora. Assim, procuramos identificar os potenciais *churners* rotacionais que tenham supostamente abandonado a operadora no momento X. Deste modo, garantimos que dispomos de informação antes e depois do seu abandono, o que nos permite a construção do entorno em ambos os períodos. Os clientes que supostamente abandonam a operadora no momento X serão considerados potenciais *churners* e será pesquisado o seu entorno no período A, para posteriormente poder ser comparado com o entorno dos respetivos suspeitos de *churn* rotacional, construídos no período B. Podemos também verificar nas Figuras 3.2 e 3.3, respetivamente, as distribuições da primeira e última chamada de todos os clientes considerados no estudo. Estas validam o facto de que a maior parte dos clientes efetua a primeira comunicação logo no início do ano e que a última comunicação tem tendência a ser efetuada junto do fim dos dados disponíveis.

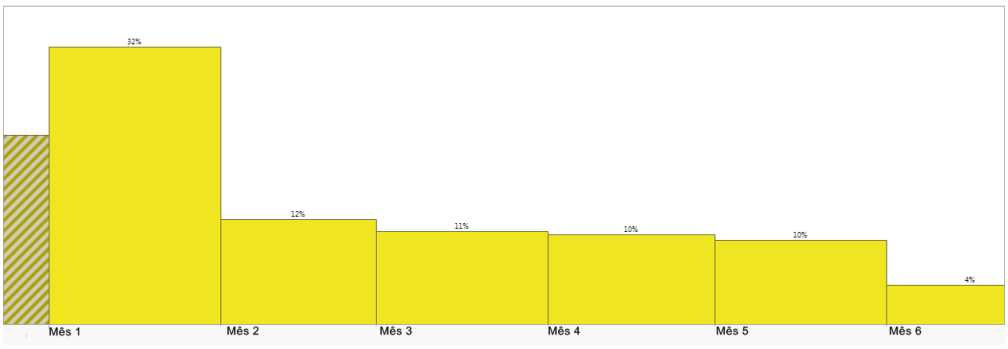

Figura 3.2: Distribuição da 1ª comunicação dos clientes-base considerados no estudo

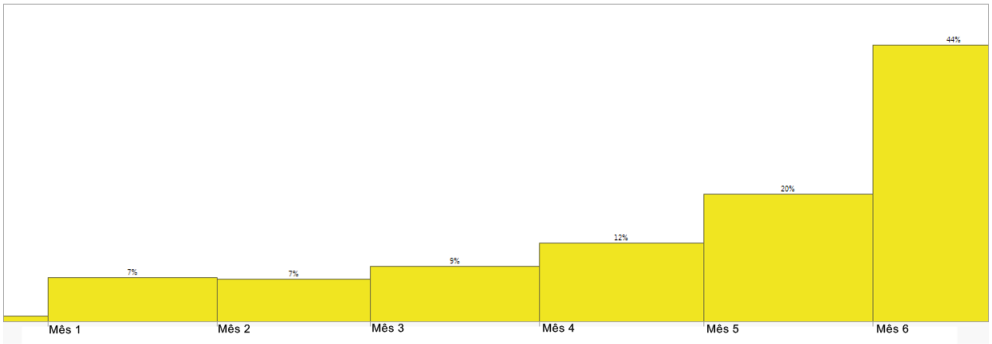

Figura 3.3: Distribuição da última comunicação dos clientes-base considerados no estudo

Na Figura 3.1, ilustrada no início do capítulo, a distribuição da primeira chamada está representada da esquerda para a direita, em tons de amarelo e a distribuição da última chamada está representada da direita para a esquerda, em tons de azul.

O período A corresponde ao tempo de atividade do potencial *churner* enquanto que o período B estará relacionado com a atividade do potencial suspeito de *churn* rotacional. Como já foi anteriormente referido, esperamos que um cliente suspeito de ser o mesmo indivíduo que determinado *churner*, tenha um padrão de contactos semelhante, nos dois espaços temporais complementares. Deste modo, procuraremos a construção do entorno do potencial *churner*no período A e do potencial suspeito no período B, tendo como objetivo a sua comparação e atribuição de probabilidade de se tratar do mesmo indivíduo.

Na Figura 3.4 é apresentado um esquema que ilustra as etapas da metodologia proposta, que têm como objetivo identificar pares de indivíduos semelhantes e atribuir-lhes uma probabilidade de se tratarem do mesmo cliente.

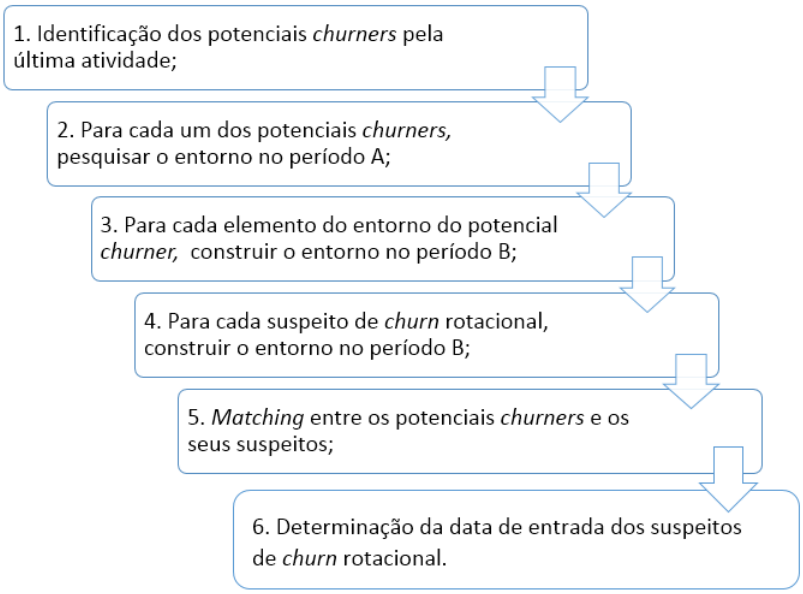

Figura 3.4: Esquema ilustrativo da metodologia

3.2 Identificação dos suspeitos de *churn* rotacional

De seguida vamos apresentar os vários passos da metodologia que visam a identificação dos suspeitos de *churn* rotacional.

1. **Identificação dos potenciais *churners* pela última atividade**

   Para identificar os clientes que abandonam a operadora e tendo como base os registos de chamadas entre um universo de clientes, vamos utilizar a data da

última comunicação do cliente como uma aproximação para determinar *churners* em determinado dia. Isto é, vamos identificar como potenciais *churners*, aqueles que efetuaram a última comunicação prematuramente, estando cerca de dois meses sem utilizar o seu serviço.

Assim, é feita uma pesquisa pelos registos de chamadas existentes, dos quais será relevante o último registo verificado e a data em que ocorreu, bem como a qual cliente se refere. Logo, vamos determinar quais são os clientes cuja última chamada é no dia referente ao ponto X, da figura 3.1. O valor de X tem de garantir a existência de dados suficientes tanto antes como depois.

2. **Para cada um dos potenciais *churners*, pesquisar o entorno no período A (até à última chamada do potencial *churner*)**

   Depois de sabermos quais os clientes que abandonaram a operadora em determinado dia, vamos identificar, para o período A, aquele que antecede a data de potencial saída, com que elementos esse potencial *churner* comunica. Essa lista será o seu entorno e será representada por uma ego-rede como a que temos na Figura 3.5.

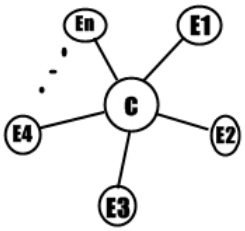

Figura 3.5: Ego-rede de um potencial *churner* obtida no período A

Nesta ego-rede temos o potencial *churner* como ego, representado por C e os seus vizinhos de nível 1, representados por $E_1, ..., E_n$, que correspondem aos elementos do entorno, isto é, temos o ego e com quem este comunica.

3. **Para cada elemento do entorno do potencial *churner*, construir o entorno no período B**

Neste passo, para cada $E_{1,c}, .., E_{n,c}$ de cada potencial *churner*, construiremos o entorno no período B, isto é, após a data de saída verificada no Passo 1. Aqui obteremos todos os elementos que comuniquem pelo menos uma vez com um dos elementos com que o potencial *churner* comunicava até à data da sua saída.

De seguida, na Figura 3.6, é apresentada uma ego-rede como exemplo, de um elemento do entorno de um potencial *churner*. Aqui, temos o elemento n do entorno do potencial *churner* c, e a sua vizinhança, composta pelos clientes com o qual comunica. Também aqui estamos perante uma ego-rede de nível 1.

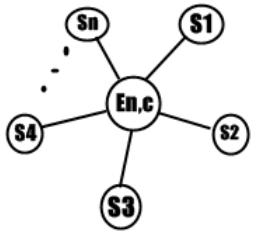

Figura 3.6: Ego-rede de um elemento do entorno de um potencial *churner* obtida no período B

Em trabalhos futuros poderiam ser utilizadas as ego-redes de nível 1 obtidas, para o estudo dos casos que suscitassem mais dúvidas. Assim, poderíamos quantificar a equivalência estrutural existente entre duas redes. Outra hipótese seria a inclusão de mais um nível nas ego-redes, isto é, para além de estudar a relação entre o ego e os seus alters, estudaríamos também o modo como os alters se relacionam.

4. **Para cada suspeito de *churn* rotacional, construir o entorno no período B**

   Da lista recolhida no passo anterior, vamos selecionar apenas os clientes que comuniquem com pelo menos $\frac{2}{3}$ do entorno do potencial *churner* no período A e em que o entorno desse potencial *churner* seja constituído por mais do

que um elemento. Assim, serão considerados suspeitos de *churn rotacional* todos os clientes do passo anterior (para cada potencial *churner*) cujo entorno no período B tenha pelo menos $\frac{2}{3}$ de semelhança com o entorno do potencial *churner* no período A e em que este último seja composto por mais do que 1 elemento.

Apenas os clientes que respeitem estas condições serão considerados suspeitos de *churn rotacional*, e apenas para estes será construído o entorno no período B. Deste modo, de uma lista com todos elementos que contactam pelo menos uma vez com um dos elementos do entorno do potencial *churner* iremos apenas selecionar aqueles com maior grau de semelhança, isto é, pelo menos $\frac{2}{3}$.

5. ***Matching* entre os potenciais *churners* e os seus suspeitos**

   Neste último passo é calculada a probabilidade de os indivíduos do Passo 1 serem os mesmos que aqueles identificados no Passo 4. Isto é, para cada suspeito referente a cada potencial *churner* vamos atribuir uma probabilidade de se tratar do mesmo indivíduo. Esta equação que nos dará o valor da probabilidade foi teorizada por Jaccard (1901). A equação de Jaccard (1901) é utilizada para comparar, estatisticamente, a semelhança e diferença entre conjuntos. Seja,

   - C= Potencial *Churner*
   - S= Suspeito
   - E(C)= Entorno do Potencial *Churner* (Período A)
   - E(S)= Entorno do Suspeito (Período B)

   $$P(S_i|C) = \frac{|E(S_i) \bigcap E(C)|}{|E(C) \bigcup E(S_i)|} \text{ (Jaccard, 1901)}$$

   Isto é, a probabilidade de determinado suspeito e potencial *churner* serem o mesmo cliente é dada pela intersecção entre eles (em espaços temporais complementares, A e B) a dividir pela sua união. Caso o resultado da equação seja 1 podemos concluir que os conjuntos são iguais e portanto a sua semelhança será de 100%. Por outro lado, uma intersecção nula entre os conjuntos originará uma probabilidade de semelhança nula. Neste caso, os conjuntos são totalmente diferentes, não tendo nada em comum. Caso tanto a intersecção como a união sejam iguais a zero então o valor da equação será igual a um, dado que teremos dois conjuntos iguais. Esta equação terá valores entre zero e um, o que nos leva a probabilidades de semelhança entre os indivíduos a poderem ocorrer entre 0% e 100%. Um valor igual a 80% mostra que os indivíduos têm 80% de probabilidade de serem o mesmo indivíduo.

6. **Determinação da data de entrada dos suspeitos de *churn* rotacional**

Já referimos anteriormente que o objetivo deste estudo é a identificação de clientes que abandonam a operadora e regressam com identificação diferente, num momento posterior. No entanto, até este momento da metodologia, não ficou assegurado que a primeira comunicação dos suspeitos de *churn* rotacional fosse realizada após o dia referente ao ponto X. Assim, neste passo, serão apenas validados como suspeitos de *churn* rotacional os suspeitos cuja primeira comunicação na rede tenha sido realizada após o dia referente ao ponto X da Figura 3.1. Apenas estes clientes serão considerados suspeitos de *churn* rotacional. No entanto, os restantes elementos que foram inicialmente considerados suspeitos e posteriormente afastados por a sua data de primeira comunicação ser antes do momento X, na Figura 3.1, podem ser relevantes. Isto porque a semelhança do entorno ocorrerá num período em que o potencial *churner* já não estará presente na rede, após o momento X, apesar de terem entrado na rede antes deste momento.

O resultado da aplicação desta metodologia será uma lista dos suspeitos para cada potencial *churner* e a respetiva probabilidade de semelhança, conforme exemplificado na seguinte figura:

Churner	Suspeito	Probabilidade
123	456	93%
123	789	82%
123	1011	67%

Figura 3.7: Exemplo de resultado obtido para um potencial *churner*, com os seus suspeitos e respetivas probabilidades

3.3 Sumário

Neste capítulo foi apresentada a metodologia a ser utilizada, e que tem como base apenas os registos anonimizados das chamadas entre os clientes e a utilização de tarifários pré-pagos. Como resultado, esperamos obter uma lista dos suspeitos para cada potencial *churner* e as respetivas probabilidades. No próximo capítulo iremos aplicar esta metodologia a dados reais e anonimizados cedidos por uma operadora de telecomunicações nacional, bem como analisar os resultados obtidos.

Capítulo 4

Caso de Estudo e Análise de Resultados

4.1 Caso de Estudo

4.1.1 Introdução

A metodologia proposta será aplicada a um caso real com dados reais e anonimizados cedidos à Faculdade de Economia do Porto para o seu estudo. Neste capítulo serão descritas todas as etapas relativas ao tratamento dos dados. Começaremos pela sua apresentação, avançando para o pré-processamento necessário bem como os filtros aplicados. Depois, é apresentada a aplicação da metodologia, que começará pela identificação dos potenciais *churners* e explicará detalhadamente cada etapa da metodologia proposta. O objetivo final é, para cada potencial *churner*, ter uma lista dos suspeitos de *churn* rotacional e uma probabilidade de se tratarem do mesmo indivíduo. Para a aplicação desta metodologia não estão disponíveis grupos de teste.

4.1.2 Dados

Tal como referido, os dados disponibilizado por uma operadora de telecomunicações são reais e anonimizados. São compostos pela informação das comunicações efetuadas entre os clientes durante 24 semanas de dados, isto é, cerca de 6 meses, que vão desde um pouco antes do mês 1 até ao mês 6. Para além da informação anonimizada de comunicações tivemos também acesso a alguma informação sobre os tarifários. Dado tratar-se de um enorme volume de dados, foi necessária uma primeira exploração da mesma, de forma a considerar aquilo que seria considerado ruído ou inútil para a metodologia que propomos aplicar, e aquilo que poderíamos utilizar para construir o modelo.

Com uma base de dados desta dimensão tornou-se necessário encontrar um software adequado. Assim, a análise dos dados deste estudo foi realizada utilizando o software SAS, versão 9.4, da versão Sistema SAS para *Windows*. Este foi utilizado para executar todos os passos deste estudo pela sua capacidade em trabalhar com grandes bases de dados e pelo elevado número de funcionalidades disponíveis, sendo que o trabalho foi desenvolvido em linguagem SQL.

Pré-Processamento dos Dados

Numa fase inicial não foi realizado qualquer pré-processamento. As interações de e para números de apoio, comerciais ou de marketing foram inicialmente consideradas como potencial ruído. No entanto, dado que aquilo que será relevante é o padrão de contactos de cada cliente considerámos útil manter esta informação. Outro ponto considerado foram as comunicações com duração inferior a quatro segundos. O primeiro passo será determinar a data da primeira e última comunicação para cada cliente. Para este ponto foram consideradas todas as comunicações, mesmo aquelas que tenham duração inferior a quatro segundos, dado que, por exemplo, uma comunicação com duração de um segundo será indicativa da sua presença na rede, apesar da inexistência de potencial em termos de conteúdo. Mais à frente, para a construção do entorno de cada cliente não serão consideradas as chamadas com duração inferior a quatro segundos, dado que serão irrelevantes do ponto de vista da construção do padrão de contatos de determinado cliente.

Filtros aplicados à base de dados

Referimos anteriormente que a metodologia iria incidir sobre os tarifários pré-pagos. Assim, reunimos junto da operadora de telecomunicações um conjunto de tarifários deste tipo. Deste modo, através do cruzamento de informação proveniente das tabelas de dados disponibilizadas foi possível manter apenas os clientes que tinham subscrito um tarifário pré-pago.

É neste tipo de tarifários que se torna mais difícil a identificação do *churn* rotacional, devido à escassa informação disponível sobre cada cliente, exigindo maior esforço analítico.

Como já foi referido anteriormente, a data da última comunicação de um cliente será utilizada como aproximação da sua potencial saída da rede. Para além disso, referimos no Capítulo 4 que o ponto X teria de garantir dados antes e depois e portanto teria de estar perto do centro do espaço temporal desejado. Deste modo, foi selecionado um dia no início do mês 4 para determinar um conjunto de potenciais *churners*, que daqui para a frente será denominado de dia ou momento X do mês 4, fazendo alusão à Figura 3.1 . Teria de ser uma data central no intervalo de dados, de modo a que tivéssemos dados à nossa disposição, tanto antes como depois do momento X. A classificação de potenciais *churners* advém do facto de utilizarmos

uma aproximação para os definirmos e de o espaço temporal dos dados não ser de tamanho suficiente para garantir que estes estão, de facto, a abandonar a operadora.

4.1.3 Identificação dos suspeitos de *churn* rotacional

1. **Identificação dos potenciais *churners***

   De forma a obtermos a lista dos potenciais *churners* precisámos primeiro de construir uma tabela que nos informasse para cada cliente qual a data da sua primeira e última comunicação.

   (a) **Filtro dos tarifários**

   Na informação anonimizada de comunicações de que dispomos, estão incluídos não só os clientes da operadora em estudo bem como os clientes de outras operadoras com que estes contactam. Por esse motivo, começamos por selecionar os clientes que utilizam um tarifário pré-pago. Este filtro não é feito a partir dos registos de comunicações, mas sim de outra tabela com informação sobre os tarifários, que posteriormente cruzamos com a tabela que contém os registos anonimizados das comunicações entre os clientes. Assim, para cada mês, desde um pouco antes do mês 1 até ao mês 6, procuramos todos os clientes que utilizem um dos tarifários pretendidos. Como resultado, teremos uma tabela com todos os clientes já filtrados pelo tarifário e qual o tarifário que estes utilizam em cada mês. Obtivemos daqui cerca de dois milhões de clientes que utilizam um dos tarifários pré-pagos pretendidos.

   (b) **Construção da base de partida**

   Daqui, iremos procurar em cada semana de dados da informação anonimizada de comunicações entre os clientes, a data mínima e máxima em que cada cliente foi emissor da chamada. Posteriormente faremos o mesmo para o caso em que o cliente seja recetor. Unindo todas as tabelas aqui obtidas, e procurando novamente a data mínima e máxima para cada cliente obteremos aquela que consideramos a nossa base de partida. Isto é, uma tabela, composta pelos clientes utilizadores dos tarifários alvo e a respetiva data de primeira e última chamada.

   Esta base de partida é composta por quase dois milhões de clientes.

   (c) **Identificar potenciais *churners* do momento X do mês** 4

   Tendo a base de partida, precisamos apenas de pesquisar quais os clientes cuja última comunicação ocorreu no momento X do mês 4. Daqui, obteremos uma tabela com informação sobre a identificação do cliente, a data da primeira comunicação e a data da última comunicação, que para todos os casos será realizada no momento X do mês 4.

Identificámos quase 3000 clientes cuja última comunicação foi no dia X do mês 4. Isto é, obtivemos o conjunto de potenciais *churners*. Dado que a nível computacional este número de potenciais *churners* exigiria um elevado esforço, optámos por seguir apenas com uma amostra destes. Assim, através de uma amostragem simples aleatória selecionámos 100 potenciais *churners*, sendo que será sobre esses que os próximos passos vão incidir.

2. **Construção do entorno dos potenciais *churners* antes de dia X do mês** 4

   Neste passo, para cada potencial *churner* construímos o seu entorno desde o início dos dados disponíveis até ao dia X do mês 4. Assim, começamos por pesquisar entre estas datas já referidas todas as comunicações em que este cliente é emissor ou recetor. Partindo desse conjunto de comunicações, vamos criar uma tabela, ignorando a partir daqui a diferença entre o cliente ser emissor ou recetor. Esta tabela terá inicialmente o total da duração das chamadas entre o potencial *churner* e cada um dos elementos com quem este contacta, bem como o total das comunicações efetuadas para esses elementos. Deste modo, podemos calcular qual o valor percentual das comunicações entre o potencial *churner* e cada um dos elementos do seu entorno. Esta última variável será importante dado que decidimos apenas considerar os elementos que acrescentem pelo menos 5% ao total das comunicações, sendo esta a definição utilizada de entorno *core*. Isto porque consideramos que um elemento que comunique menos de 5% com o potencial *churner* não assumirá relevância no seu entorno e como tal não deve ser considerado parte do seu padrão de contactos. Por isso, como resultado da aplicação deste passo a todos os clientes considerados potenciais *churners* obteremos a sua tabela de entorno composta pelos elementos que acrescentem pelo menos 5% ao seu total de comunicações. Na Tabela 4.1 temos o exemplo do resultado da aplicação deste passo ao potencial *churner* com a identificação 151340.

	MSISDN_REF	MSISDN_ENTORNO	TOTAL_DURATION	DURATION_REF	PCT_DURATION	CUM_PCT
1	151340	39507673	1656	2169	0.763	0.763
2	151340	3723472	372	2169	0.172	0.935
3	151340	8049553	137	2169	0.063	0.998

Tabela 4.1: Exemplo do entorno do potencial *churner* com a identificação 151340 até ao dia X do mês 4

3. **Construção do entorno após o dia X do mês** 4 **dos elementos do entorno (até ao dia X do mês** 4**) do potencial *churner***

   Neste passo procedemos à construção do entorno após o dia X do mês 4 e até ao último dia disponível do mês 6, para todos os elementos identificados

no passo anterior. A ideia é que, um suspeito de *churn rotacional*, tem de obrigatoriamente comunicar com pelo menos um dos elementos com que o potencial *churner* comunica, mas em períodos complementares. Assim, para cada potencial *churner* e partindo dos elementos com quem este contacta até abandonar potencialmente a operadora construímos o entorno destes elementos após o dia X do mês 4. Posteriormente, e para cada potencial *churner*, fazemos a união do entorno dos seus elementos, que acabámos de calcular. Deste modo obtemos uma lista de clientes que têm, no seu entorno (após o dia X do mês 4), pelo menos um cliente que também está presente no entorno do potencial *churner*.

De seguida, e dado que a lista obtida neste passo é geralmente extensa, aplicamos um novo filtro. Assim, de todos os elementos desta lista, escolhemos apenas aqueles que comuniquem com pelo menos $\frac{2}{3}$ do entorno do potencial *churner* no período complementar. Isto porque acreditamos que, uma percentagem inferior a $\frac{2}{3}$ de semelhança entre entornos muito dificilmente resultaria em suspeitos de *churn* rotacional, dado que existe tendência para continuarem a contactar com a mesma comunidade. O resultado deste filtro será uma lista composta pelos suspeitos de *churn rotacional*.

4. **Construção do entorno após o dia X do mês 4 para cada suspeito de *churn* rotacional**

   Antes de avançarmos para este passo, foi necessário excluir todos os potenciais *churners* que tinham apenas um elemento no seu entorno até à data da sua saída. Acontece que, tendo apenas um elemento no seu entorno, teríamos imensos suspeitos para cada um desses potenciais *churners* e provavelmente elevadas probabilidades baseadas num entorno com pouca informação e que pouca sustentação daria à probabilidade. Desse modo, evitamos a redundância do excesso de suspeitos nestes casos, considerando a partir deste ponto apenas os potenciais *churners* que tenham mais de um elemento no seu entorno. Assim, passamos de uma amostra de 100 potenciais *churners* para apenas 64. De seguida, procedemos à construção do entorno dos suspeitos de *churn rotacional*, para cada potencial *churner*, de um modo idêntico ao já apresentado anteriormente mas com a pesquisa na informação anonimizada de comunicações a ser feita após o dia X do mês 4. O resultado desta etapa será posteriormente utilizado para depois de ser comparado com o entorno do potencial *churner*, no período complementar, e ser atribuída uma probabilidade de se tratar do mesmo indivíduo.

5. ***Matching* entre os potenciais *churners* e os seus suspeitos**

   Neste passo é finalmente calculada a probabilidade de um suspeito de *churn* rotacional ser o mesmo indivíduo que o respetivo potencial *churner*. São utilizados os entornos construídos no Passo 2 e os entornos construídos no passo

anterior, e será feita a comparação dos mesmos. A probabilidade é dada pela intersecção entre estes dois conjuntos dividida pela sua união, conforme indicado no Capítulo 3. Uma probabilidade de 80% indica-nos que o suspeito tem 80% de probabilidade de ser o mesmo indivíduo que o potencial *churner*.

6. **Determinação da data de entrada dos suspeitos de *churn* rotacional**

   Nesta última etapa da aplicação da metodologia é determinada a data da primeira comunicação de cada um dos suspeitos de *churn* rotacional. Assim, serão apenas considerados como suspeitos de *churn* rotacional os elementos do passo anterior cuja data da primeira comunicação seja após o dia X do mês 4. Com isto, alcançamos a ideia de que o potencial *churner* abandona a operadora mas que eventualmente pode regressar com uma identificação diferente, contatando tendencialmente dentro da mesma comunidade.

4.2 Análise de Resultados

4.2.1 Introdução

Neste capítulo iremos apresentar e analisar os resultados obtidos bem como algumas simulações. Começamos com uma análise global dos resultados obtidos. Depois, avançamos para a descrição do caso de dois potenciais *churners*, culminando no cálculo da probabilidade entre estes e os seus suspeitos. De seguida vamos analisar a sensibilidade da metodologia face ao número de elementos do entorno do potencial *churner*. Por fim, iremos testar a metodologia proposta em elementos que não sejam potenciais *churners*, procurando averiguar a capacidade do modelo em os encontrar como suspeitos de *churn* rotacional, através da análise do valor de semelhança atribuído.

4.2.2 Análise Global

A nível global, dos 64 potenciais *churners* considerados, 43 têm zero suspeitos. Para os restantes 21 potenciais *churners* encontrámos no total 57 suspeitos. No entanto, no total apenas foram obtidas probabilidades para 34 elementos. Alguns suspeitos comunicam com pelo menos $\frac{2}{3}$ dos elementos com que o potencial *churner* comunica, em períodos complementares. Porém, estes não fazem parte do seu entorno *core*. Logo, dado que a intersecção entre o suspeito e o potencial *churner* será igual a zero, a probabilidade para estes casos não será calculada.

No total das 34 probabilidades calculadas, apenas 18 suspeitos estão na base de partida. Isto significa que, de todos os suspeitos encontrados, apenas 18

deles pertencem a um dos tarifários escolhidos para este estudo. É conveniente referir que apenas temos informação sobre as datas da primeira e última comunicação para os clientes que pertençam à base de partida. Para além disso, dos 18 clientes, apenas 10 têm entrada na rede posterior ao dia X do mês 4. Assim, dos 34 suspeitos para os quais temos probabilidade de semelhança com um dos potenciais *churners*, apenas 10 são considerados suspeitos de *churn* rotacional, conforme se pode confirmar na Tabela 4.2.

Na Tabela 4.2 podemos ver a lista dos potenciais *churners* e dos seus suspeitos e as respetivas probabilidades, representadas pela coluna "P", sendo estes os resultados finais obtidos a partir da aplicação da metodologia aos dados.

	CHURNER	SUSPEITO	P
1	34779748	34761956	0.429
2	34938263	33601107	0.400
3	39649574	39651032	0.143
4	40123168	40706446	0.500
5	40123168	41234766	0.667
6	40123168	41451394	0.400
7	40234761	41278045	0.125
8	40234761	41127704	0.125
9	40595600	41494948	0.571
10	41024994	41121228	0.250

Tabela 4.2: Resultados finais obtidos

Em relação a estes 10 suspeitos de *churn* rotacional, as datas da primeira comunicação destes na rede ocorrem entre após o momento X do mês 4 e no final do mês 5. As probabilidades variam entre $12,5\%$ e $66,7\%$, mostrando que foram encontradas algumas semelhanças entre potenciais *churners* e outros clientes com atividade num período complementar. Podemos também verificar que são encontrados suspeitos para sete potenciais *churners* distintos.

Dos elementos inicialmente considerados como suspeitos e com primeira comunicação antes do dia X do mês 4, 6 entram no primeiro dia de dados disponível, um no dia seguinte e o outro no final do mês 3. Neste caso as probabilidades também variam entre $12,5\%$ e $66,7\%$.

Por fim, temos os clientes para os quais temos probabilidade calculada mas que não pertencem a um dos tarifários pré-pagos selecionados. Em relação a estes clientes, as probabilidades variam entre 11% e 100%, sendo que não temos mais informações sobre os mesmos. Assim, não são considerados clientes suspeitos de *churn* rotacional, dado não dispormos de informação sobre a data da sua primeira comunicação. Por esse motivo, não podemos garantir que esta tenha ocorrido após o dia X do mês 4. No Anexo A são apresentadas as distribuições e as tabelas de frequências das probabilidades dos clientes que não são suspeitos de *churn* rotacional (data de primeira comunicação antes do dia X do mês 4) bem como dos suspeitos que não utilizam um dos tarifários

disponíveis (Figura A.1, Figura A.2, Tabela A.1 e Tabela A.2). Na Figura 4.1 podemos ver graficamente a distribuição das probabilidades obtidas para todos os suspeitos de *churn* rotacional.

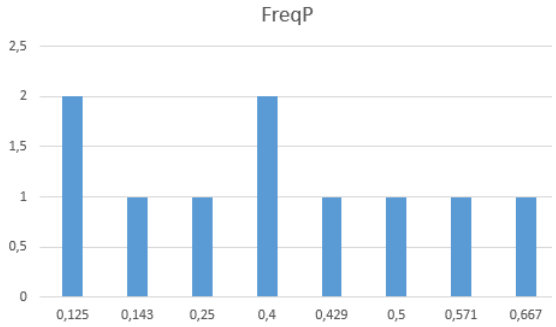

Figura 4.1: Distribuição das probabilidades obtidas

4.2.3 Análise individual de dois potenciais *churners*

O cliente 25091705 tem a sua primeira comunicação no primeiro dia de dados disponível e a última no dia X do mês 4, sendo assim considerado um potencial *churner* e membro de um dos tarifários pré-pagos usados para análise. Na Tabela 4.3 temos o entorno deste cliente até ao dia X do mês 4, isto é, no período A.

	MSISDN_REF	MSISDN_ENTORNO	TOTAL_DURATION	DURATION_REF	PCT_DURATION	Rank_25091705	CUM_PCT
1	25091705	24613760	12001	15415	0.779	1	0.779
2	25091705	6693358	1728	15415	0.112	2	0.891

Tabela 4.3: Entorno do potencial *churner* 25091705 até ao dia X do mês 4

Podemos então verificar que o potencial *churner* tem um entorno *core* composto por dois elementos e que representam $89,1\%$ do seu total de comunicações. As comunicações com o cliente 24613760 representam $77,9\%$ desse valor, o equivalente a cerca de três horas e vinte minutos de comunicações.

De seguida, analisa-se o entorno destes dois elementos (24613760 e 6693358) após o dia X do mês 4 até ao fim dos dados disponíveis. Após ser feita a união destes dois entornos, selecionamos os elementos distintos do entorno e fazemos a contagem, de modo a descobrirmos com quantos elementos dos clientes 24613760 e 6693358 comunica cada elemento do entorno desta união. O total de elementos que contactam pelo menos uma vez com estes dois clientes é 66, sendo que apenas quatro deles têm comunicações com os dois clientes. Deste modo, apenas quatro clientes respeitam a condição de comunicarem com pelo menos $\frac{2}{3}$, o que significa que para o cliente 25091705 teremos quatro

suspeitos de *churn* rotacional para os quais iremos construir o entorno após o dia X do mês 4. As principais tabelas que resultam deste processo estão disponíveis no Anexo A (Tabela A.3 até Tabela A.7)

Na Tabela 4.4 temos o entorno do suspeito 351920123744. Na Tabela 4.5 temos o entorno do suspeito 9740505. Os restantes estão disponíveis no Anexo A, nas tabelas A.8 e A.9.

	MSISDN_REF	MSISDN_ENTORNO	TOTAL_DURATION	DURATION_REF	PCT_DURATION	CUM_PCT
1	351920123744	24613760	319	398	0.802	0.802
2	351920123744	6693358	79	398	0.198	1.000

Tabela 4.4: Entorno do suspeito de *churn* rotacional, 351920123744

	MSISDN_REF	MSISDN_ENTORNO	TOTAL_DURATION	DURATION_REF	PCT_DURATION	CUM_PCT
1	9740505	24613760	4295	7208	0.596	0.596
2	9740505	22158532	1027	7208	0.142	0.738
3	9740505	6693358	775	7208	0.108	0.846

Tabela 4.5: Entorno do suspeito de churn rotacional, 9740505

Assim, na Tabela 4.4 verificamos que este cliente tem exatamente os mesmos elementos do entorno que o potencial *churner* em análise. Os dois elementos com que o nosso potencial *churner* comunicava $89,1\%$ do seu total de comunicações representam 100% das comunicações do suspeito de *churn* rotacional 351920123744. Isto mostra-nos que após o abandono do potencial *churner*, o suspeito apenas comunica com estes dois elementos. No entanto, o total de comunicações é bem menor quando comparado com o total de comunicações do entorno do potencial *churner*.

Na Tabela 4.5, para além dos dois elementos do entorno do potencial *churner* 25091705, temos mais um cliente que integra o entorno do suspeito.

De seguida vamos poder verificar quais as probabilidades atribuídas pela metodologia proposta a este potencial *churner* e respetivos suspeitos. Podemos verificar os resultados na Tabela 4.6. As probabilidades estão representadas na coluna "P".

	CHURNER	SUSPEITO	INTU_S	UNIONS_U	P
1	25091705	9740505	2	3	0.667
2	25091705	22158532	2	6	0.333
3	25091705	351916495685	1	5	0.200
4	25091705	351920123744	2	2	1.000

Tabela 4.6: Probabilidades atribuídas ao potencial *churner* 25091705 e aos seus suspeitos

Já referimos anteriormente que a probabilidade é o valor da intersecção entre o entorno do potencial *churner* e o entorno do suspeito, a dividir pela união

entre ambos. Assim, no caso do suspeito 9740505 temos dois elementos em comum com o potencial *churner*, o que nos dá uma intersecção igual a dois. Se unirmos os dois conjuntos obtemos um valor de três. Assim, a probabilidade de o suspeito 9740505 e o potencial *churner* 25091705 serem o mesmo indivíduo é de $66,7\%$. No caso do suspeito 351920123744, comparamos dois conjuntos iguais. Deste modo, tanto o valor da intersecção como da união será igual a dois e portanto teremos uma probabilidade de serem o mesmo indivíduo de 100%. Para os restantes suspeitos as probabilidades encontradas são menores.

Quando determinamos as datas da primeira comunicação de cada um dos suspeitos relativos ao potencial *churner* 25091705 verificamos que nenhum será classificado de suspeito de *churn* rotacional. Isto acontece pois as datas de primeira comunicação são antes do dia X do mês 4.

Os restantes suspeitos do potencial *churner* 25091705 apresentam probabilidades de 20% e 100%. No entanto, não pertencem aos tarifários que escolhemos. Por esse motivo não temos mais informações sobre eles, não podendo deste modo garantir que a sua entrada ocorreu após o dia X do mês 4, condição exigida para que sejam considerados suspeitos de *churn* rotacional.

De seguida, vamos analisar o cliente 40595600. Este, tem a primeira comunicação no primeiro dia de dados disponíveis e a última comunicação no dia X do mês 4. Deste modo, é considerado um potencial *churner* e membro de um dos tarifários pré-pagos usados para análise. Na Tabela 4.7 apresentamos o entorno deste cliente até ao momento X do mês 4.

	MSISDN_REF	MSISDN_ENTORNO	TOTAL_DURATION	DURATION_REF	PCT_DURATION	CUM_PCT
1	40595600	35116533	3360	30308	0.111	0.111
2	40595600	351914090122	3131	30308	0.103	0.214
3	40595600	35116177	2633	30308	0.087	0.301
4	40595600	351911878467	2585	30308	0.085	0.386
5	40595600	351965839752	2370	30308	0.078	0.465
6	40595600	351916962358	1596	30308	0.053	0.517

Tabela 4.7: Entorno do potencial *churner* 40595600 até ao dia X do mês 4

Podemos verificar que a percentagem do total de comunicações com o potencial *churner* é semelhante para todos os clientes deste entorno, variando entre $5,3\%$ e $11,1\%$. Constatamos também que, no total, estes seis elementos representam $51,7\%$ do entorno do potencial *churner*, isto é, o seu entorno *core*.

Posteriormente analisamos o entorno destes seis elementos após o momento X do mês 4 até ao fim dos dados disponíveis. O passo seguinte é a sua união, que nos permite selecionar os elementos distintos do entorno e contar para quantos elementos do entorno do potencial *churner* estes contatam após o dia X do mês 4. Assim, construímos uma lista com todos os clientes que comunicam pelo menos uma vez com um dos seis elementos do entorno do potencial *churner*.

Para que a condição de comunicar com pelo menos $\frac{2}{3}$ do entorno do cliente 40595600 se verifique, um cliente da lista criada anteriormente tem de ter comunicado com pelo menos 4 elementos do entorno do potencial *churner*. Daqui, obtemos apenas um suspeito, que comunica com os seis elementos. Os restantes comunicam, no máximo, com três elementos do entorno do potencial *churner*.

Na Tabela 4.8 apresentamos o entorno após o dia X do mês 4 do único cliente considerado suspeito, com a identificação 41494948.

	MSISDN_REF	MSISDN_ENTORNO	TOTAL_DURATION	DURATION_REF	PCT_DURATION	CUM_PCT
1	41494948	351965027874	5446	25440	0.214	0.214
2	41494948	35116177	3457	25440	0.136	0.350
3	41494948	351914090122	3186	25440	0.125	0.475
4	41494948	351965839752	1854	25440	0.073	0.548
5	41494948	35116533	1496	25440	0.059	0.607

Tabela 4.8: Entorno do potencial suspeito 41494948 após o dia X do mês 4

Verificamos que, dois elementos do entorno *core* do potencial *churner* não são entorno *core* do suspeito, embora ele contate com eles no período B, isto é, após o dia X do mês 4, com percentagem do total de comunicações inferior a 5%. Ainda assim, o suspeito apresenta no seu entorno quatro clientes em comum com o entorno do potencial *churner*. Para além disso, o suspeito contata com um cliente com que o potencial *churner* não tem comunicações.

Na Tabela 4.9 apresentamos a probabilidade de semelhança atribuída ao potencial *churner* 40595600 e ao seu suspeito 41494948. A probabilidade está representada na coluna "P".

	CHURNER	SUSPEITO	INTU_S	UNIONS_U	P
1	40595600	41494948	4	7	0.571

Tabela 4.9: Probabilidade atribuída ao potencial *churner* 40595600 e ao seu suspeito

Deste modo, podemos considerar que o suspeito tem uma probabilidade de $57,1\%$ de tratar-se do mesmo cliente que o potencial *churner* 40595600. De seguida, determinamos que a data da primeira comunicação do único suspeito encontrado aconteceu poucos dias após o dia X do mês 4 e que a sua última comunicação ocorreu no último dia de dados disponíveis. Assim, podemos considerar este cliente como um suspeito de *churn* rotacional, com a probabilidade de $57,1\%$.

4.2.4 Sensibilidade da metodologia face ao número de elementos do entorno do potencial *churner*

De modo a analisarmos a sensibilidade da metodologia face a diferentes números de elementos do entorno do potencial *churner*, dividimo-los em grupos e reunimos os resultados para cada grupo. Assim, e com o objetivo de criar grupos equilibrados, foi feita a divisão do número possível de elementos do entorno dos potenciais *churners* em três grupos, que são apresentados de seguida:

- Grupo 1
  (a) Número de elementos do entorno do potencial *churner*: 2;
  (b) Número de potenciais *churners* que pertencem a este grupo: 24;
  (c) Percentagem dos potenciais *churners* presentes neste grupo: $37,5\%$;
- Grupo 2
  (a) Número de elementos do entorno do potencial *churner*: 3, 4 ou 5;
  (b) Número de potenciais *churners* que pertencem a este grupo: 25;
  (c) Percentagem dos potenciais *churners* presentes neste grupo: $39,05\%$
- Grupo 3
  (a) Número de elementos do entorno do potencial *churner*: Mais do que 6;
  (b) Número de potenciais *churners* que pertencem a este grupo: 15;
  (c) Percentagem dos potenciais *churners* presentes neste grupo: $23,45\%$;

Assim, os potenciais *churners* presentes no Grupo 1 têm apenas dois elementos no seu entorno até ao dia X do mês 4 e representam $37,5\%$ do total de potenciais *churners*. Dos 24 clientes, para 18 deles foram encontrados zero suspeitos. No entanto, para os restantes seis elementos foram encontrados 30 suspeitos. Destes, apenas foram calculadas probabilidades para 11. De seguida, determinamos a data da primeira comunicação de cada um destes 11 suspeitos para os quais foi calculada probabilidade de semelhança com o respetivo potencial *churner*. Aqui, verificamos que apenas cinco serão considerados suspeitos de *churn* rotacional, dado que têm datas de primeira comunicação após o dia X do mês 4. Três clientes entram antes do dia X do mês 4, com probabilidades entre $33,3\%$ e $66,7\%$, não sendo portanto considerados suspeitos de *churn* rotacional. Os restantes três não utilizam um dos tarifários selecionados para este estudo, tendo probabilidades entre 20% e 100%. Na Figura 4.2 podemos ver a distribuição das probabilidades no grupo 1.

No Grupo 2, temos 25 potenciais *churners* com números de elementos do entorno situados entre três e cinco, e representam $39,05\%$ do total. Dos 25

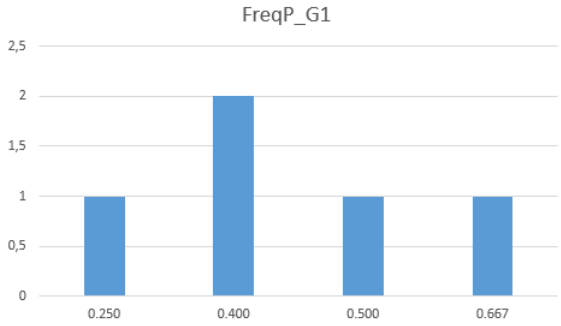

Figura 4.2: Distribuição das probabilidades do Grupo 1

clientes, 13 deles tinham zero suspeitos. Para os restantes 12 elementos foram encontrados no total 24 suspeitos. Destes, foram calculadas probabilidades para 20. De seguida, determinamos a data da primeira comunicação de cada um destes 20 suspeitos para os quais foi calculada probabilidade de semelhança com o respetivo potencial *churner*. Verificamos que apenas quatro são considerados suspeitos de *churn* rotacional, dado que apenas estes têm datas de primeira comunicação após o dia X do mês 4. Três clientes entraram antes desse dia, com probabilidades entre 20% e 40%. Os restantes 13 clientes não utilizam um dos tarifários selecionados para este estudo, tendo probabilidades entre $11,10\%$ e $66,7\%$. Na Figura 4.3 podemos verificar a distribuição das probabilidades do grupo 2.

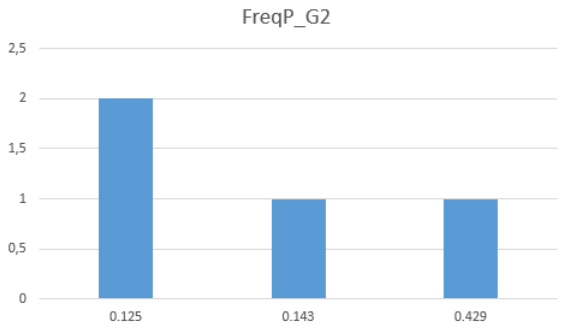

Figura 4.3: Distribuição das probabilidades do Grupo 2

No Grupo 3, composto por 15 potenciais *churners* com números de elementos do entorno iguais ou superiores a seis, temos $23,45\%$ do total. Dos 15 clientes, 12 deles não têm qualquer suspeito. Para os restantes três elementos foram encontrados três suspeitos, sendo que foram calculadas probabilidades para todos. De seguida, determinamos a data da primeira comunicação de cada um destes três suspeitos para os quais foi calculada probabilidade de semelhança com o respetivo potencial *churner*. Verificamos que apenas um é considerado

suspeito de *churn* rotacional, com probabilidade de $57,1\%$, dado que tem a sua primeira comunicação após o dia X do mês 4. Os restantes dois suspeitos entram antes desse dia, com probabilidades entre $12,5\%$ e $28,6\%$.

4.2.5 Simulação com elementos que não são potenciais *churners*

Esta simulação serve como um teste à metodologia proposta. Assumindo um indivíduo que tenha a sua primeira comunicação no primeiro dia de dados disponível e a última no último dia de dados disponível será de esperar que este mantenha um padrão de comunicações constante ao longo do tempo. Isto é, que durante estes cerca de seis meses, tenha a maioria das suas comunicações dentro de uma comunidade e que isto seja visível ao longo do tempo. Se isto for verdade, então é esperado que o próprio indivíduo seja encontrado pela metodologia e lhe seja atribuída uma probabilidade elevada.

Fizemos a simulação com quatro indivíduos. O primeiro teste foi com o ID 4, que tem as datas idênticas às do referido exemplo. Este cliente foi corretamente encontrado como suspeito no entanto teve uma probabilidade de apenas 25%. Consideramos que o baixo valor de semelhança atribuído não se deve ao mau funcionamento da metodologia mas sim à variação significativa do entorno de um período para outro e também ao filtro aplicado para determinar o entorno *core*. Neste caso, se mantivéssemos os elementos que acrescentam menos do que 5% ao total de comunicações a intersecção já seria igual a quatro. De referir que o número de elementos do entorno é igual em ambos os períodos, embora não seja composto pelos mesmos elementos. Podemos verificar no Anexo A, na Tabela A.12, o entorno *core* deste cliente até ao dia X do mês 4 e na Tabela A.13 a partir dessa data.

No caso seguinte, relativo ao ID 6503426, também aqui as datas são iguais ao exemplo e o indivíduo é encontrado como único suspeito de *churn rotacional*. Mais uma vez temos uma probabilidade de apenas 25%. Aqui existe bastante influência da definição de entorno *core*. Se fixássemos o limite da percentagem do total de duração das chamadas nos $3,5\%$ já seriam mais dois elementos em comum com o entorno do cliente no primeiro período. Note-se que, neste caso, ao apenas selecionar os elementos que acrescentem mais do que 5% ao total de comunicações estamos a excluir 26 elementos do entorno deste indivíduo. Mais uma vez, a baixa probabilidade justifica-se pela variação significativa do entorno entre os dois períodos.

Podemos verificar no Anexo A, na Tabela A.14, o entorno *core* deste cliente até ao dia X do mês 4 e na Tabela A.15 a partir dessa data.

O ID 1006 tem também como data da primeira comunicação o primeiro dia de dados disponível e última comunicação no último dia de dados disponível. Este indivíduo é encontrado com sucesso como suspeito e é lhe atribuída uma probabilidade de ser o mesmo indivíduo de $83,3\%$, aproximadamente. Note-se que este indivíduo tem cinco elementos em comum, numa união de seis elementos, nos períodos complementares. Podemos verificar no Anexo A, na Tabela A.16, o entorno *core* deste cliente até ao dia X do mês 4 e na Tabela A.17 a partir dessa data.

No último caso testado, o ID 286, com datas iguais às dos restantes indivíduos, não tivemos sucesso em encontrá-lo como suspeito. Isto deveu-se ao facto de, no segundo período não cumprir a comunicação com $\frac{2}{3}$ dos elementos do entorno no primeiro período e como tal não é considerado como suspeito de *churn rotacional*. Podemos verificar no Anexo A, na Tabela A.18 o entorno core deste cliente até ao dia X do mês 4, sendo que neste caso não existe a partir desta data.

4.3 Sumário

Neste capítulo começamos por descrever o pré-processamento e filtros realizados aos dados, avançando posteriormente para a aplicação das várias etapas da metodologia em dados reais e anonimizados. Analisámos também os resultados, começando por o fazer a nível global e individual. De seguida foi apresentada uma análise da sensibilidade da metodologia à variação do número de elementos do entorno dos potenciais *churners*. Posteriormente a metodologia foi aplicada a elementos que não eram considerados *churners* de modo a verificar se os resultados estavam de acordo com o esperado. No próximo capítulo serão apresentadas as conclusões, limitações e sugestões de trabalhos futuros para este estudo.

Capítulo 5

Conclusões e Trabalhos Futuros

5.1 Conclusões

Consideramos que o objetivo proposto de identificar indivíduos semelhantes em períodos complementares foi atingido. A utilização do entorno dos potenciais *churners* e dos seus suspeitos para cálculo da semelhança entre estes mostrou-nos resultados interessantes e resultam de simples pesquisas aos dados disponíveis, sendo portanto uma metodologia prática.

Pelos resultados obtidos, acreditamos que a metodologia identificou com sucesso indivíduos semelhantes em períodos complementares e portanto conseguiu identificar os clientes com maior probabilidade de estarem a cometer *churn rotacional*. No entanto, as probabilidades encontradas não foram suficientemente elevadas para concluir com maior precisão sobre a existência de *churners* rotacionais. Para além disso, e partindo das simulações que realizámos com uma pequena amostra de *não-churners*, verificámos que em alguns casos o entorno dos indivíduos varia bastante ao longo do tempo. Do ponto de vista empresarial, a aplicação desta metodologia permitirá a identificação dos clientes com maior probabilidade de estarem a cometer *churn* rotacional e portanto a possibilidade de medir o fenómeno, corrigir indicadores (tanto de *churn* como de captação de clientes), rever comissionamentos de ativação de cartões, entre outros. Permite ainda o estudo destes dados de forma a, futuramente, poder antecipar casos de *churn* rotacional. Como limitações consideramos o facto de termos restringido a procura apenas aos registos de chamadas entre os clientes e apenas aos clientes utilizadores de um tarifário pré-pago. É conveniente recordar que para a adesão a este tipo de tarifários pouca informação é exigida. Aliando este facto ao de apenas conhecermos os registos de quem ligou para quem, a informação que dispomos é limitada.

5.2 Trabalhos Futuros

Dado que este trabalho incidiu sobre uma área ainda pouco estudada existem bastantes melhorias que podem ser realizadas. Podemos apostar numa melhoria contínua da precisão do modelo, com cruzamento com outro tipo de dados e testes de avaliação do modelo a cada iteração.

O aumento da dimensão temporal da análise poderá também fortalecer o modelo. Por exemplo, podemos identificar potenciais *churners* através do estudo de outros dias, para além do escolhido neste trabalho. Podemos também considerar a extensão dos dados disponíveis, sendo que, com cerca de um ano os resultados seriam mais robustos.

Seria também relevante estudar o impacto mensal do fenómeno do *churn* rotacional na operadora de telecomunicações ao longo de um ano, mostrando a importância que tem a deteção deste fenómeno.

Outra melhoria seria o enriquecimento da análise com atributos de negócio de forma a perceber melhor qual o impacto deste fenómeno na atividade empresarial. Por exemplo, quantificar o peso do aproveitamento deste tipo de clientes oportunistas, que agem por conveniência. Ou também, estudar em que lojas ou ações de terreno este fenómeno se verifica mais, através da ativação de novos cartões.

Poderá também tornar-se interessante alterar a definição de entorno *core*, aferindo a sensibilidade do modelo. Utilizámos como filtro que os clientes teriam de representar pelo menos 5% do total de comunicações do entorno do seu par. Poderia por exemplo ser testado como definição de entorno *core* e consequente filtro, que este fosse composto por 85% do total de comunicações.

Podemos também considerar que os dados existentes pouco antes de o potencial *churner* abandonar podem constituir ruído. Isto é, pouco antes de potencialmente abandonar a operadora, as comunicações do cliente podem ser poucas e não relevantes, constituindo ruído. Simular os resultados sem estes dados permitiria conhecer como o modelo reage a esta alteração e qual o impacto nos resultados obtidos.

Por fim, a utilização de ego-redes pode também acrescentar valor à resolução do problema de *churn* rotacional. Podemos por exemplo estudar apenas a ligação do ego (potencial *churner*) com os seus alters (entorno) e comparar a estrutura com a ego-rede do suspeito. Ou então, acrescentar mais um nível à ego-rede. Isto significa que iríamos estudar também a ligação dos alters (elementos do entorno) entre si. A utilização de ego-redes permitiria assim que mais informação fosse obtida uma vez que consideraria não só o número de ligações, mas também o padrão das ligações estabelecidas entre o ego e os respetivos alters.

Bibliografia

Argeseanu, S. C., Kelly, L., and Prabhakaran, D. (2013). Egocentric social network analysis of cardiovascular disease in south asian: Preliminary evidence from urban india. *Population Association of America, Annual Meeting*.

Babu, M. S., Devi, D. A., and Anuradha, M. (2015). A study on impact of big data analytics to indian e-commerce applications. *International Journal of Advanced Engineering, Management and Science (IJAEMS)*, 1(2):76–82.

Bonchi, F., Castillo, C., Gionis, A., and Jaimes, A. (2011). Social network analysis and mining for business applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):22–37.

Borgatti, S. P., Everett, M. G., and Freeman, L. C. (2002). Ucinet for windows: Software for social network analysis.

Burt, R. S. (2009). *Structural holes: The social structure of competition*. Harvard university press.

Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G., and Barabási, A.-L. (2008). Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015.

Chung, K. K., Hossain, L., and Davis, J. (2005). Exploring sociocentric and egocentric approaches for social network analysis. In *Proceedings of the 2nd international conference on knowledge management in Asia Pacific*, pages 27–29.

Coleman, J., Katz, E., and Menzel, H. (1957). The diffusion of an innovation among physicians. *Sociometry*, pages 253–270.

Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A. A., and Joshi, A. (2008). Social ties and their relevance to churn in mobile telecom networks. In *Proceedings of the 11th international*

conference on Extending database technology: Advances in database technology, pages 668–677. ACM.

Eagle, N., Pentland, A. S., and Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278.

Fisher, D. (2005). Using egocentric networks to understand communication. *Internet Computing, IEEE*, 9(5):20–28.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239.

Friedman, S. R. and Aral, S. (2001). Social networks, risk-potential networks, health, and disease. *Journal of Urban Health*, 78(3):411–418.

Guare, J. (1990). *Six degrees of separation: A play*. Random House LLC.

Hanneman, R. and Riddle, M. (2005). *Introduction to Social Network Methods*. University of California.

Jaccard, P. (1901). *Distribution de la Flore Alpine: dans le Bassin des dranses et dans quelques régions voisines*. Rouge.

Karnstedt, M., Hennessy, T., Chan, J., Basuchowdhuri, P., Hayes, C., and Strufe, T. (2010). Churn in social networks. In *Handbook of Social Network Technologies and Applications*, pages 185–220. Springer.

Kawale, J., Pal, A., and Srivastava, J. (2009). Churn prediction in mmorpgs: A social influence based approach. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 423–428. IEEE.

Kazienko, P., Ruta, D., and Bródka, P. (2009). The impact of customer churn on social value dynamics. *International Journal of Virtual Communities and Social Networking*, 1(3):62–74.

Kempe, D., Kleinberg, J., and Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM.

Kiss, C. and Bichler, M. (2008). Identification of influencersâmeasuring influence in customer networks. *Decision Support Systems*, 46(1):233–253.

Mattison, R. (2006). *The telco churn management handbook*. Lulu.com.

Milgram, S. (1967). The small world problem. *Psychology today*, 2(1):60–67.

Newman, M. E. (2000). Models of the small world. *Journal of Statistical Physics*, 101(3-4):819–841.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

Oliveira, M. and Gama, J. (2012). An overview of social network analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(2):99–115.

Richter, Y., Yom-Tov, E., and Slonim, N. (2010). Predicting customer churn in mobile networks through analysis of social groups. In *SDM*, volume 2010, pages 732–741.

Scott, J. (1988). Social network analysis. *Sociology*, 22(1):109–127.

Zafarani, R., Abbasi, M. A., and Liu, H. (2014). *Social Media Mining: An Introduction*. Cambridge University Press.

Anexo A

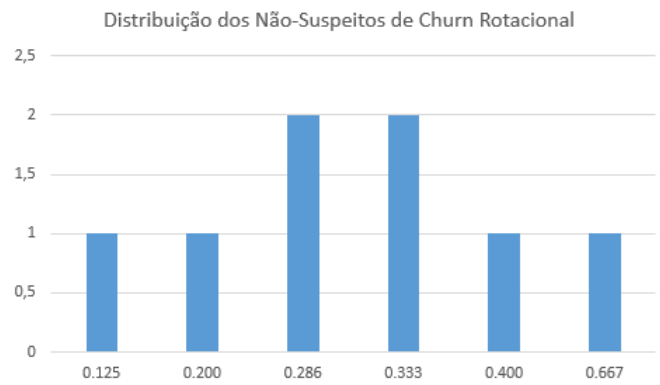

Figura A.1: Distribuição dos clientes que não são suspeitos de *Churn* Rotacional

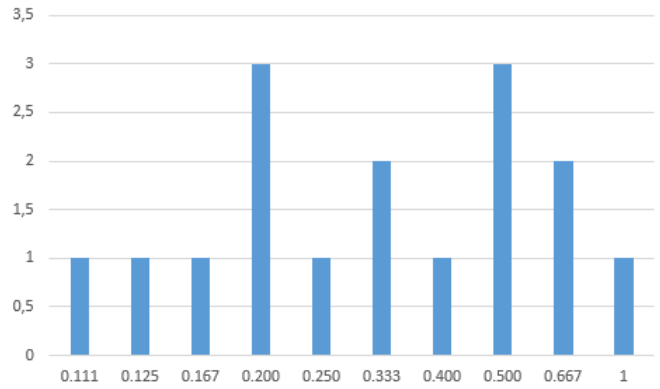

Figura A.2: Distribuição dos clientes que não utilizam um dos tarifários em estudo

	CHURNER	SUSPEITO	INTU_S	UNIONS_U	P	GRUPO	CONTRACT_ID	MIN_TIME	MAX_TIME
1	39037679	431802	1	3	0.333	G1	431802	30DEC2013	16JUN2014
2	38978582	3719180	1	5	0.200	G2	3719180	30DEC2013	16JUN2014
3	25091705	9740505	2	3	0.667	G1	9740505	01JAN2014	16JUN2014
4	41399381	21926011	1	8	0.125	G3	21926011	30DEC2013	16JUN2014
5	25091705	22158532	2	6	0.333	G1	22158532	30DEC2013	13JUN2014
6	32387429	26290121	2	7	0.286	G2	26290121	30DEC2013	16JUN2014
7	151340	34656698	2	5	0.400	G2	34656698	30DEC2013	13JUN2014
8	39725216	41063252	2	7	0.286	G3	41063252	21MAR2014	16JUN2014

Tabela A.1: Lista de todos os clientes que não são suspeitos de *Churn* Rotacional

	CHURNER	SUSPEITO	INTU_S	UNIONS_U	P	GRUPO	CONTRACT_ID	MIN_TIME	MAX_TIME
1	32387429	351963778804	1	9	0.111	G2	.	.	.
2	38306081	351910594831	1	8	0.125	G2	.	.	.
3	38978582	41878600	1	6	0.167	G2	.	.	.
4	40659417	351933458319	1	5	0.200	G2	.	.	.
5	25091705	351916495685	1	5	0.200	G1	.	.	.
6	40719322	41705178	1	5	0.200	G1	.	.	.
7	151340	34631272684	1	4	0.250	G2	.	.	.
8	151340	351925445745	2	6	0.333	G2	.	.	.
9	38170790	41494343	2	6	0.333	G2	.	.	.
10	7403128	41578664	2	5	0.400	G2	.	.	.
11	151340	351253048760	2	4	0.500	G2	.	.	.
12	38978582	351915338472	2	4	0.500	G2	.	.	.
13	7403128	351967877803	2	4	0.500	G2	.	.	.
14	7403128	351924101036	2	3	0.667	G2	.	.	.
15	4347466	41498679	2	3	0.667	G2	.	.	.
16	25091705	351920123744	2	2	1.000	G1	.	.	.

Tabela A.2: Lista de todos os clientes que não utilizam um dos tarifários em estudo

	MSISDN_REF	MSISDN_ENTORNO	TOTAL_DURATION	DURATION_REF	PCT_DURATION
1	24613760	6693358	17369	28872	0.602
2	24613760	9740505	4295	28872	0.149
3	24613760	351915565737	1869	28872	0.065
4	24613760	351910830806	731	28872	0.025
5	24613760	22158532	648	28872	0.022
6	24613760	14801831	574	28872	0.020
7	24613760	351913325241	472	28872	0.016
8	24613760	351966168812	351	28872	0.012
9	24613760	351920123744	319	28872	0.011
10	24613760	42057004	303	28872	0.010
11	24613760	351919937250	272	28872	0.009
12	24613760	351916495685	228	28872	0.008
13	24613760	351925963464	226	28872	0.008
14	24613760	41550037	193	28872	0.007
15	24613760	351923030269	179	28872	0.006

Tabela A.3: Entorno do indivíduo 24613760, elemento do entorno do potencial *churner* 25091705

	MSISDN_REF	MSISDN_ENTORNO	TOTAL_DURATION	DURATION_REF	PCT_DURATION
1	6693358	24613760	17369	71396	0.243
2	6693358	351914179187	12483	71396	0.175
3	6693358	351910311403	9634	71396	0.135
4	6693358	39447024	6581	71396	0.092
5	6693358	7354298	3976	71396	0.056
6	6693358	351915679587	3063	71396	0.043
7	6693358	33698971810	2519	71396	0.035
8	6693358	351912668995	2304	71396	0.032
9	6693358	351916495685	1890	71396	0.026
10	6693358	351919665191	1210	71396	0.017
11	6693358	351918743617	1209	71396	0.017
12	6693358	22158532	1173	71396	0.016
13	6693358	351910411602	936	71396	0.013
14	6693358	351914340760	890	71396	0.012
15	6693358	9740505	775	71396	0.011
16	6693358	33669235493	680	71396	0.010

Tabela A.4: Entorno do indivíduo 6693358, elemento do entorno do potencial *churner* 25091705

	MSISDN_REF	MSISDN_ENTORNO	TOTAL_DURATION	DURATION_REF	PCT_DURATION
1	6693358	3511734	88	71396	0.001
2	6693358	5652739	648	71396	0.009
3	6693358	7354298	3976	71396	0.056
4	6693358	9740505	775	71396	0.011
5	6693358	12235346	61	71396	0.001
6	6693358	22158532	1173	71396	0.016
7	6693358	24613760	17369	71396	0.243
8	6693358	39447024	6581	71396	0.092
9	6693358	40455051	359	71396	0.005
10	6693358	41512675	116	71396	0.002
11	6693358	33669235493	680	71396	0.010
12	6693358	33698971810	2519	71396	0.035
13	6693358	351308801418	357	71396	0.005
14	6693358	351910311403	9634	71396	0.135
15	6693358	351910411602	936	71396	0.013
16	6693358	351911078780	224	71396	0.003
17	6693358	351912453297	455	71396	0.006

Tabela A.5: União das Tabelas A.3 e A.4

	MSISDN_ENTORNO	NR_ENT
1	3511734	1
2	5652739	1
3	6693358	1
4	7354298	1
5	9740505	2
6	12235346	1
7	14801831	1
8	22158532	2
9	24613760	1
10	39447024	1
11	40455051	1
12	41512675	1
13	41550037	1
14	42057004	1

Tabela A.6: Número de elementos do entorno do potencial *churner* 25091705 para o qual contactam os suspeitos iniciais

	MSISDN_ENTORNO	NR_ENT
1	9740505	2
2	22158532	2
3	351916495685	2
4	351920123744	2

Tabela A.7: Suspeitos do potencial *churner* 25091705

	MSISDN_REF	MSISDN_ENTORNO	TOTAL_DURATION	DURATION_REF	PCT_DURATION	CUM_PCT
1	22158532	41512675	1526	6282	0.243	0.243
2	22158532	6693358	1173	6282	0.187	0.430
3	22158532	9740505	1027	6282	0.163	0.593
4	22158532	24613760	648	6282	0.103	0.696
5	22158532	351913968244	596	6282	0.095	0.791
6	22158532	351917224428	510	6282	0.081	0.872

Tabela A.8: Entorno do suspeito 22158532 depois do dia X do mês 4

	MSISDN_REF	MSISDN_ENTORNO	TOTAL_DURATION	DURATION_REF	PCT_DURATION	CUM_PCT
1	351916495685	6380338	5149	12362	0.417	0.417
2	351916495685	9307312	2112	12362	0.171	0.587
3	351916495685	6693358	1890	12362	0.153	0.740
4	351916495685	6953397	943	12362	0.076	0.817

Tabela A.9: Entorno do suspeito 351916495685 depois do dia X do mês 4

	CHURNER	SUSPEITO	INTU_S	UNIONS_U	P	GRUPO	CONTRACT_ID	MIN_TIME	MAX_TIME
1	34938263	33601107	2	5	0.400	G1	33601107	11APR2014	12APR2014
2	40123168	40706446	2	4	0.500	G1	40706446	13APR2014	22APR2014
3	41024994	41121228	1	4	0.250	G1	41121228	10APR2014	02JUN2014
4	40123168	41234766	2	3	0.667	G1	41234766	08APR2014	12APR2014
5	40123168	41451394	2	5	0.400	G1	41451394	29MAY2014	09JUN2014

Tabela A.10: Lista de todos os suspeitos de *churn* rotacional do Grupo 1

	CHURNER	SUSPEITO	INTU_S	UNIONS_U	P	GRUPO	CONTRACT_ID	MIN_TIME	MAX_TIME
1	34938263	33601107	2	5	0.400	G1	33601107	11APR2014	12APR2014
2	40123168	40706446	2	4	0.500	G1	40706446	13APR2014	22APR2014
3	41024994	41121228	1	4	0.250	G1	41121228	10APR2014	02JUN2014
4	40123168	41234766	2	3	0.667	G1	41234766	08APR2014	12APR2014
5	40123168	41451394	2	5	0.400	G1	41451394	29MAY2014	09JUN2014

Tabela A.11: Lista de todos os suspeitos de *churn* rotacional do Grupo 2

	MSISDN_REF	MSISDN_ENTORNO	TOTAL_DURATION	DURATION_REF	PCT_DURATION	CUM_PCT
1	4	351939315766	6991	27724	0.252	0.252
2	4	351918041102	3177	27724	0.115	0.367
3	4	351967292281	1880	27724	0.068	0.435
4	4	74561	1787	27724	0.064	0.499
5	4	351915294192	1498	27724	0.054	0.553

Tabela A.12: Entorno até ao dia X do mês 4 do ID 4

	MSISDN_REF	MSISDN_ENTORNO	TOTAL_DURATION	DURATION_REF	PCT_DURATION	CUM_PCT
1	4	351939315766	3925	23647	0.166	0.166
2	4	351918041102	3736	23647	0.158	0.324
3	4	351917498106	2388	23647	0.101	0.425
4	4	41157567	1733	23647	0.073	0.498
5	4	351919345623	1533	23647	0.065	0.563

Tabela A.13: Entorno após o dia X do mês 4 do ID 4

	MSISDN_REF	MSISDN_ENTORNO	TOTAL_DURATION	DURATION_REF	PCT_DURATION	CUM_PCT
1	6503426	38326634	24683	81400	0.303	0.303
2	6503426	351924102215	8813	81400	0.108	0.411
3	6503426	351967923622	8030	81400	0.099	0.510
4	6503426	351963097313	7171	81400	0.088	0.598
5	6503426	351963986930	4842	81400	0.059	0.658
6	6503426	351919333400	4085	81400	0.050	0.708

Tabela A.14: Entorno até ao dia X do mês 4 do ID 6503426

	MSISDN_REF	MSISDN_ENTORNO	TOTAL_DURATION	DURATION_REF	PCT_DURATION	CUM_PCT
1	6503426	38326634	14810	50858	0.291	0.291
2	6503426	351967923622	13347	50858	0.262	0.554
3	6503426	351212156007	3967	50858	0.078	0.632
4	6503426	8663864	2844	50858	0.056	0.688

Tabela A.15: Entorno após o dia X do mês 4 do ID 6503426

	MSISDN_REF	MSISDN_ENTORNO	TOTAL_DURATION	DURATION_REF	PCT_DURATION	CUM_PCT
1	1006	23682	9426	33996	0.277	0.277
2	1006	3803568	6227	33996	0.183	0.460
3	1006	23651	5813	33996	0.171	0.631
4	1006	14166474	3843	33996	0.113	0.744
5	1006	26027748	2165	33996	0.064	0.808
6	1006	351918557123	1717	33996	0.051	0.859

Tabela A.16: Entorno até ao dia X do mês 4 do ID 1006

	MSISDN_REF	MSISDN_ENTORNO	TOTAL_DURATION	DURATION_REF	PCT_DURATION	CUM_PCT
1	1006	23682	10045	34228	0.293	0.293
2	1006	26027748	7731	34228	0.226	0.519
3	1006	23651	5428	34228	0.159	0.678
4	1006	351918557123	2750	34228	0.080	0.758
5	1006	3803568	1967	34228	0.057	0.816

Tabela A.17: Entorno após o dia X do mês 4 do ID 1006

	MSISDN_REF	MSISDN_ENTORNO	TOTAL_DURATION	DURATION_REF	PCT_DURATION	CUM_PCT
1	286	1668407	1211	5921	0.205	0.205
2	286	351913551100	482	5921	0.081	0.286
3	286	351913886561	409	5921	0.069	0.355
4	286	351932683004	320	5921	0.054	0.409
5	286	20809147	307	5921	0.052	0.461

Tabela A.18: Entorno até ao dia X do mês 4 do ID 286