DINÂMICAS DE COMUNIDADES EM REDES SOCIAIS DE GRANDE DIMENSÃO

por

Vítor Manuel Araújo Cerqueira

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

Orientado por

João Gama
Márcia Oliveira

Faculdade de Economia

Universidade do Porto

2014

Nota Biográfica

No dia 16 de fevereiro de 1990 nasceu Vítor Manuel Araújo Cerqueira, na vila de Ponte de Lima.

Em 2008, Vítor ingressou na Faculdade de Ciências da Universidade do Porto, instituição na qual se licenciou em Matemática Aplicada quatro anos mais tarde. Em setembro de 2012 iniciou o seu percurso no mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão, na Faculdade de Economia da Universidade do Porto.

Entre julho e outubro de 2014 foi bolseiro de investigação no LIAAD, Laboratório de Inteligência Artificial e Apoio à Decisão, na instituição INESC Porto.

Agradecimentos

Primeiro, gostaria de agradecer aos meus orientadores. Ao professor João Gama, pela disponibilidade, incentivos e pela oportunidade de trabalhar no LIAAD, onde aprendi imenso. À Márcia, pelo apoio e dedicação constantes, que foram fundamentais ao longo de toda esta caminhada. Aproveito também para agradecer a todos os meus colegas do laboratório, especialmente ao Rui Sarmento, pela ajuda a perceber parte das tarefas realizadas.

Agradeço à empresa detentora da base de dados analisada e aos seus colaboradores, especialmente ao Nuno, pela oportunidade de participar num projeto aliciante. Agradeço também ao Carlos e ao Daniel, que me acompanharam no projeto.

Finalmente, agradeço à minha família e amigos. Um obrigado especial à minha irmã Cristina, pela inspiração, a quem dedico o trabalho.

Resumo

Em redes sociais, grupos de indivíduos ou outras entidades mais ligados entre si do que a outros são tipicamente designados por comunidades. Analisar as dinâmicas de comunidades em redes sociais permite perceber como esses grupos se comportam ao longo do tempo. Esta estratégia de análise de redes possibilita o estudo de novas medidas, na forma como os negócios atuam sobre os seus clientes. Assim, esses negócios podem melhorar a experiência de consumo dos seus clientes. No caso particular de uma rede social de chamadas telefónicas, as operadora de telecomunicações ganham informações importantes acerca das dinâmicas da rede, através da descoberta dos padrões de evolução das suas comunidades. Uma forma de descobrir esses padrões é através da identificação e caraterização das comunidades e a monitorização das mesmas ao longo do tempo. As operadoras de telecomunicações têm em mãos redes de grande dimensão com milhões de ligações, difíceis de analisar. Esta dificuldade surge, não só em termos de dimensão, mas também pela natureza complexa dos dados. Nesse sentido, é proposta uma metodologia, para o estudo das dinâmicas de redes de grande dimensão. Esta metodologia foi aplicada empiricamente numa rede de chamadas telefónicas de grande dimensão, pertencente a uma grande empresa de telecomunicações. Através de um método de amostragem não enviesado, foi retirada uma amostra representativa da rede original, com milhões de ligações, ao longo de um horizonte temporal de seis meses. A modularidade obtida ao executar o algoritmo de deteção de comunidades dá indicação de uma estrutura de comunidade significativa na rede estudada. Além disso, através de uma abordagem multi-critério e de um método apropriado, as comunidades foram selecionadas e analisadas ao longo do

horizonte temporal. As dinâmicas dessas comunidades comprovam a existência de uma estabilidade na evolução da rede social estudada, o que poderá ajudar tomada de decisões, por parte das operadoras de telecomunicações.
Palavras-Chave: Redes Sociais de Grande Dimensão, Evolução, Análise de Comunidades

Abstract

In social networks, groups of people or entities that are connected with each other more than to others are usually known as communities. Through the analysis of the dynamics of communities in social networks, one can understand how those communities behave over time. This network analysis strategy opens the door to the study of new hypothesis in the way businesses act upon their customers. Thus, businesses may improve the quality of service provided. In the particular case of a network of phone calls, telecom operators gain an important insight about the network's dynamics, by discovering the evolution patterns of its communities. One way to learn the evolution patterns of the communities of the network, is to identify and characterize those communities, and monitor them over time. Telecom operators have in hands large scale networks with millions of links, difficult to analyse. This difficulty arises not only due to their increased size, but also because of the complex nature of the data. Thus, it is proposed a methodology, developed to analyse the dynamics of communities in large scale networks. This methodology was empirically applied to a large scale telecommunication network from a major mobile network operators. Through an unbiased sampling method it was taken a representative sample of a large scale network of phone calls, with millions of links, over a period of six months. The modularity obtained when executing the community detection algorithm suggests that there is a meaningful community structure in the network. Furthermore, based on a multicriteria approach and a fitting framework, the communities where selected and analysed over time. Analysing the dynamics of the communities, the results indicate the existence of a natural evolution of the network studied, which may help the

telecom operators in decision making.

Keywords: Large Scale Social Networks, Evolution, Community Analysis.

Índice

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Neste capítulo é discutido o problema abordado e a motivação para estudá-lo. São também apresentados os objetivos mais específicos, assim como a organização do documento.

1.1 Motivação

Muitos fenómenos do mundo real podem ser naturalmente representados através de grafos dinâmicos, capazes de capturar o estado da rede social subjacente em vários instantes de tempo. A estrutura da maioria das redes sociais encontradas no mundo real, tais como redes de estudantes de uma universidade, redes de clientes de uma empresa ou redes alimentares, exibe estrutura de comunidades (Newman and Girvan, 2004). Isto significa que é comum encontrar regiões na rede mais conectadas. Estes sub-grafos são normalmente designados por comunidades, uma vez que agregam entidades que se encontram mais ligadas entre si do que outras entidades na rede. O problema da deteção destas comunidades em redes sociais foi amplamente abordado na literatura (Fortunato, 2010), mas na grande maioria dos casos, os métodos desenvolvidos assumem que as redes são estáticas. Porém, na realidade, a maioria dos fenómenos tem uma natureza dinâmica, uma vez que sofrem constantes mudanças e se caraterizam por uma série de eventos temporais, tais como o crescimentos, degradação ou dispersão. Uma das formas de estudar a evolução

de redes sociais dinâmicas é através da análise da evolução das comunidades que a compõem. Numa perspetiva empresarial, tem interesse estudar a forma como estes grupos se comportam ao longo do tempo, com o propósito de, por exemplo, compreender alterações nas preferências (padrões de consumo dos clientes e deteção de eventuais indícios de abandono). Este tipo de informação pode ser utilizado para redefinir a estratégia de marketing, personalizar campanhas promocionais de acordo com as caraterísticas partilhadas pelos clientes pertencentes a uma dada comunidade, adotar medidas pro-ativas para evitar o churn, desenvolver sistemas de recomendação de produtos e serviços que melhorem a experiência de consumo do cliente e a sua relação com a empresa.

Este é um tema que desafia a utilização de metodologias que envolvem o conhecimento analítico geral nas áreas de análise de redes sociais, agrupamento de dados e manuseamento de dados em grande escala para a identificação de comunidades e padrões de evolução das mesmas.

1.2 Objetivos

O objetivo principal é analisar as dinâmicas de comunidades em redes sociais de grande dimensão e proceder à sua caraterização. Foram propostas várias metodologias para a monitorização de comunidades em redes ao longo do tempo, tais como Berger-Wolf and Saia (2006), Asur et al. (2009) e Oliveira et al. (2014), cuja abordagem tipicamente passa por detetar as comunidades em vários períodos de tempo. No entanto, estes métodos podem não estar bem adaptados para redes volumosas com milhares ou centenas de milhares de comunidades. À medida que entramos na era de *Big Data*, a nossa capacidade de acumular dados aumentou de forma considerável. Daí surge a necessidade de um melhoramento das metodologias tradicionais, de modo a proporcionar análises mais eficazes. Neste contexto, é apresentada uma metodologia, complementar às referidas, cuja funcionalidade é a monitorização de comunidades em redes de grande dimensão. A metodologia proposta abrange os seguintes tópicos: A amostragem não enviesada para redes de grande dimen-

são, a deteção de comunidades, uma seleção multi-critério das comunidades detetadas, a evolução das comunidades e a caraterização das comunidades. Esta metodologia surge no âmbito do projeto do caso de estudo. No caso de estudo, o conjunto de dados trata-se de uma rede de chamadas, de uma das principais operadoras de telecomunicações nacionais. Uma empresa desta envergadura exige soluções a questões de negócio, tais como a evolução das comunidades, que envolvem o manuseamento de uma rede de grande dimensão e de uma complexidade acrescida.

1.3 Organização

A dissertação divide-se em três partes principais. A base teórica na qual assenta o trabalho desenvolvido, a metodologia proposta e o caso de estudo, onde se aplicou essa metodologia. A primeira parte engloba o enquadramento teórico que suportou o estudo e que abrange os capítulos 2 e 3. No capítulo 2 é discutida a literatura referente à tarefa de deteção de comunidades em redes sociais, desde a definição do problema até à explicação de possíveis abordagens e algoritmos, tais como o Método de Louvain. No capítulo 3 é descrita a metodologia do MECnet, que foi originalmente desenvolvido para a monitorização da evolução de comunidades em redes.

No capítulo 4 é discutida a metodologia proposta para o estudo das dinâmicas de comunidades em redes sociais de grande dimensão.

Finalmente, no capítulo 5 é aplicada e interpretada a metodologia proposta, numa rede de chamadas de grande dimensão. As tarefas consistem, de forma geral, no método de amostragem, preparação dos dados, identificação e seleção das comunidades e respetiva monitorização ao longo do tempo. É ainda analisado o ciclo de vida das comunidades mais estáveis e traçados os seus perfis, em termos visuais e analíticos. As conclusões são discutidas no capítulo 6.

Capítulo 2

Deteção de Comunidades

Neste capítulo é abordado o tema da deteção das comunidades em redes sociais. Além da explicação do algoritmo adotado para o caso de estudo e de outros métodos de deteção de comunidades, é ainda feito o paralelismo com o agrupamento tradicional de dados e demonstrada a aplicabilidade da tarefa de deteção de comunidades através de exemplos mencionados na literatura.

2.1 Comunidades em Redes Sociais

Podemos pensar numa rede como sendo um conjunto de elementos, onde cada um deles está ligado com outros, cuja representação assume a forma de um grafo, onde os elementos ou vértices estão ligados por arestas. Uma rede social reflete, portanto, a estrutura social de indivíduos ou organizações e as suas relações, como indica Baruah and Angelov (2012). Assim, as comunidades são grupos de elementos da rede, que, em princípio, têm uma determinada propriedade em comum. Quando se fala em comunidade, é assumido intuitivamente um contexto social. Como é natural, as pessoas têm propensão para formar grupos, seja em família, amigos ou no trabalho e é neste contexto que se formam as comunidades. As comunidades também surgem noutros sistemas em rede, como por exemplo, nos ramos de biologia, engenharia e ciência dos computadores. Em redes de in-

teração de proteínas, há uma determinada probabilidade para a formação de comunidades de proteínas com a mesma função específica dentro de uma célula, como é referido por Pizzuti et al. (2012) e Chen and Yuan (2006).

Fortunato (2010) refere que estudar a estrutura das comunidades tem aplicações concretas importantes. Identificar grupos de pessoas com interesses ou padrões de comportamento semelhantes numa rede de relações de compra permite a criação de sistemas de recomendação eficientes, que melhoram substancialmente as oportunidades de negócio. Em redes de telecomunicações o operador pode usar este tipo de informação para o desenvolvimento de medidas e estratégias de Marketing direcionadas, que proporcionam um aumento na qualidade de serviço prestado à sua base de clientes.

A análise das comunidades identificadas numa rede social permite a classificação dos indivíduos de acordo com a sua posição estrutural na comunidade em que está inserido. Um indivíduo com uma posição preponderante dentro da sua comunidade ocupa nela um papel fundamental em termos de estabilidade e controlo. Por outro lado, indivíduos que partilham várias ligações com nós de outras comunidades têm uma função de mediação de relações entre diferentes comunidades. Em redes de telecomunicações, esta informação pode ser usada para enfrentar problemas como a prevenção de Churn ou a captação de clientes. Clauset et al. (2004) usam o seu algoritmo de deteção de comunidades em redes para analisar os padrões de compra numa grande companhia de retalho *online*, extraindo comunidades significativas e revelando grande padrões, que retratam os hábitos de compra dos consumidores. Blondel et al. (2008) identifica a principal língua das comunidades de uma operadora de telecomunicações belga, numa rede com 118 milhões de nós e mais de um bilião de ligações.

2.2 Definição do Problema

De forma geral, o problema da deteção de comunidades é um problema de agrupamento de dados. Jain (2010) afirma que a organização de um conjunto de objetos em grupos é

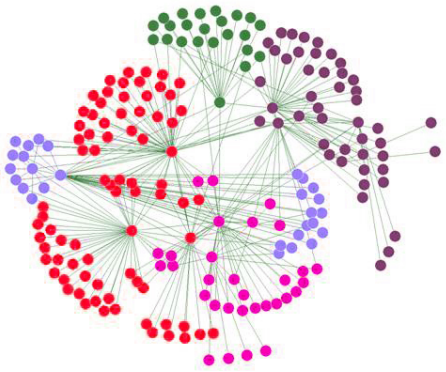


Figura 2.1: Exemplo da representação de uma rede social e discriminação das suas comunidades.

uma das formas mais básicas de entendimento e aprendizagem para o Homem. Assim, a resolução da tarefa passa por associar cada indivíduo, ou nó da rede, a um grupo, ou comunidade, de um modo relevante. Na análise de agrupamentos os objetos são agrupados de acordo com certas características intrínsecas ou conforme uma dada medida de semelhança ou dissemelhança. No caso particular da deteção de comunidades, os algoritmos tipicamente não incorporam informação sobre os atributos dos nós, baseando-se na estrutura das ligações existentes na rede. O intuito de agrupar os dados é estruturá-los e estudar as suas características gerais, sendo, por isso, uma análise de natureza exploratória.

O agrupamento de dados é uma análise de aprendizagem não supervisionada, ou seja, não envolve a utilização de classes pré-definidas e não existe informação *a priori* sobre o número de grupos ou a sua estrutura. A classificação e a análise discriminante são exemplos de outro tipo de aprendizagem, supervisionada. O facto de não haver informação *a priori*, faz com que o agrupamento de dados seja uma análise mais difícil e, ao mesmo tempo, mais aliciante do ponto de vista do investigador.

Como diz Jain (2010), um grupo pode ser definido como um conjunto de objetos compactos e isolados, que variam em termos de tamanho, forma e densidade. Existe uma subjetividade inerente ao conceito de grupo, cuja significância e interpretação dependem

do ponto de vista e requerem conhecimento do domínio. Em termos práticos, se tivermos um conjunto A = { Maria Sharapova; Roger Federer; Cristiano Ronaldo, Luisão; Marta da Silva; Rafael Nadal } e quisermos agrupar os objetos em diferentes classes, a solução é diferente se considerarmos uma abordagem em termos de género ou em termos de modalidade desportiva.

Há várias abordagens possíveis para realizar o agrupamento de dados. Um bom método de agrupamento de dados apresenta grupos com medida de semelhança elevada entre os seus objetos, enquanto essa medida de semelhança é baixa entre objetos pertencentes a diferentes grupos. Por outras palavras, uma boa estrutura de agrupamento apresenta uma elevada semelhança intra-classes e baixa semelhança inter-classes.

Jain (2010) indica ainda que esta análise é usada em diversos propósitos: A de descobrir a estrutura subjacente; ganhar sensibilidade dos dados; gerar hipóteses; detetar anomalias e identificar caraterísticas salientes; classificação natural, para perceber o grau de semelhança entre observações; compressão, como uma forma de organizar os dados e sumarizá-los.

2.3 Possíveis Abordagens

Os algoritmos de agrupamento podem ser, segundo Jain (2010), divididos em dois tipos: algoritmos particionais e algoritmos hierárquicos. Os primeiros criam várias partições em simultâneo e avaliam-nas segundo um determinado critério, sendo o mais comum, o critério da minimização do erro quadrado. Estes são algoritmos com uma estrutura não hierárquica, cujo argumento de entrada por norma requer a predefinição do número de grupos pretendido. O argumento de entrada pode ser uma matriz $n \times d$, com n objetos inseridos num espaço com d dimensões, ou uma matriz de semelhança $n \times n$, que pode proceder de uma matriz $n \times d$. Um dos algoritmos particionais mais usados na literatura é o K-médias, descrito por MacQueen (1967), no qual a função custo é a função do erro quadrado.

Na realidade, como Jain (2010) menciona, pouco é conhecido sobre a estrutura das comunidades numa rede. É invulgar saber em quantas comunidades se divide uma rede ou outros dados relativos ao conjunto dos membros. Neste contexto, os métodos particionais para o agrupamento de dados tornam-se pouco praticáveis, visto que é necessário fazer suposições aceitáveis no que toca ao tamanho e número das comunidades.

Por outro lado, a rede pode ter uma estrutura hierárquica, ou seja, apresentar vários níveis de comunidades entre os seus elementos, com comunidades mais pequenas inseridas noutras maiores. Para estes casos, pode ser adequado o recurso a algoritmos de agrupamento por hierarquias, que revelam a estrutura hierárquica de uma rede, visível através de dendrogramas. Inicialmente é necessário definir uma medida de semelhança entre os nós da rede. Essa medida é calculada para cada par de nós, mesmo os que não estão conectados, o que origina uma matriz de semelhança X, $n \times n$.

Os métodos de agrupamento hierárquico podem ser divididos em duas classes: Algoritmos aglomerativos, nos quais em cada passo as comunidades são fundidas se o seu coeficiente de semelhança for suficientemente alto; e os algoritmos divisivos, onde as comunidades são separadas iterativamente, retirando ligações de nós com baixo nível de semelhança.

Os primeiros, algoritmos aglomerativos, seguem um processo *bottom-up*, começando com os nós como comunidades individuais e acabando com a rede como uma só comunidade. Os algoritmos divisivos são técnicas *top-down* e funcionam de forma oposta. Posteriormente, são adotadas condições de paragem, segundo um determinado critério, como por exemplo a otimização de uma função de qualidade. É também fundamental definir uma medida para quantificar a distância entre as diferentes comunidades. O principal inconveniente desta abordagem, e que a torna inviável no tratamento de grandes quantidades de dados, é a sua elevada complexidade computacional.

Um dos algoritmos aglomerativos mais populares é o das ligações médias, explicado por Hastie et al. (2001).

No fundo, Jain (2010) afirma que a escolha do algoritmo de deteção de grupos depende

da função objetivo, dos geradores de modelos probabilísticos e de heurísticas.

2.3.1 Dados Dinâmicos

Dados dinâmicos, ao contrário de dados estáticos, podem alterar-se ao longo do tempo, como páginas web ou blogs. Daí vem que à medida que o tempo passa e os dados se modificam, as técnicas de agrupamento devem ser atualizadas de forma adequada. Para Jain (2010), estas caraterísticas e volumes de dados elevados obrigam a novas exigências aos algoritmos tradicionais de deteção de grupos, para processar e sumarizar esses dados em tempo útil. Além disso, é necessária a habilidade de adaptação às mudanças na distribuição dos dados, a habilidade de deteção de comunidades emergentes e distingui-las de *outliers*, e a habilidade de detetar a fusão de comunidades e descartar as já caducadas. Devido à necessidade de processar estes dados num curto espaço de tempo, foram propostos vários algoritmos para resolver estes pontos. Tipicamente, esse algoritmos são extensões de outros algoritmos mais simples, como o K-médias, como os métodos propostos por Guha et al. (2003) ou Aggarwal et al. (2003), que descrevem técnicas para agrupar grandes quantidades de *streams* de dados, de forma eficiente.

2.3.2 Dados Relacionais

Em dados relacionais ou dados de redes, a ideia principal é detetar as partições de um grande grafo, tal como uma rede de telecomunicações, e dividi-lo em sub-grafos coesos com base na sua estrutura de ligação. Um modelo geral probabilístico foi primeiramente proposto por Taskar et al. (2002), onde diferentes entidades relacionadas são distribuições modeladas que se condicionam umas às outras.

Tal como a rede do caso de estudo, a maior parte das redes reais são redes dinâmicas, onde as arestas das rede surgem e desaparecem à medida que o tempo passa. Daí vem a necessidade de modelar esse comportamento evolutivo das redes, tanto na filiação das comunidades, como noutras caraterísticas preponderantes. Quando se tratam de redes vo-

lumosas, esta escala de dados exige novos métodos que recolham a informação da sua estrutura de forma eficiente, já que as técnicas mais tradicionais estão inaptas para o tratamento deste tipo de dados massivos, pelo seu custo computacional.
A abordagem natural passa por decompor a rede noutras mais pequenas, as designadas comunidades, que são compostas por nós bastante interligados.
Já existem vários algoritmos na literatura, que detetam razoavelmente boas partições num curto espaço de tempo. Esta procura por algoritmos rápidos sofreu um grande aumento nos últimos anos, devido à crescente disponibilidade de conjuntos de dados de redes volumosas e o impacto que elas têm no dia-a-dia. A função de modularidade, descrita por Newman and Girvan (2004), é um critério muito popular para descobrir a estrutura de comunidades em redes, que considera a estrutura das ligações.

2.4 Métodos Tradicionais de Deteção de Comunidades

Nesta secção serão abordados alguns métodos que permitem identificar comunidades em redes. Estes métodos efetuam análise de agrupamentos em dados relacionais. A intermediação e a modularidade são duas métricas usadas para a descoberta de comunidades em redes.

2.4.1 Métodos Baseados na Intermediação

A intermediação é uma medida que favorece ligações inter-comunidades em detrimento de ligações intra-comunidades. A lógica da utilização desta medida passa por separar as várias comunidades existentes numa rede, removendo as ligações mais fracas entre as elas, isolando-as. Em termos práticos, uma medida de intermediação é baseada no caminho geodésico: são encontrados os caminhos mais curtos entre todos os pares de nós e conta-se quantos correm por cada ligação, ou por cada nó, dependendo do tipo de intermediação que estamos a calcular.
De forma geral, a intermediação de um nó v pode ser calculado da seguinte forma:

$$I(v) = \sum_{v \neq w \neq z} \frac{\pi_{wz}(v)}{\pi_{wz}}, \tag{2.1}$$

Onde π_{wz} é o número total de caminhos mais curtos do nó w para o nó z e $\pi_{wz}(v)$ é o número de menores caminhos que passam por v.

Intuitivamente, a intermediação é uma medida que traduz a suscetibilidade de uma pessoa servir de via direta entre outras duas pessoas.

O algoritmo divisivo proposto por Girvan and Newman (2002) identifica os limites das comunidades numa rede através da intermediação. Apesar de ter sido aplicado de forma bem sucedida em várias redes, como redes de emails, redes sociais e outras, este é um algoritmo que não escala bem para redes realmente volumosas. É uma técnica que envolve um custo computacional que restringe a sua utilidade a redes com no máximo alguns milhares de nós.

2.4.2 Métodos Baseados na Modularidade

A função objetivo de modularidade, descrita por Newman and Girvan (2004), é um critério bastante utilizado para detetar a estrutura das comunidades em redes. De forma simples, é uma métrica que quantifica a qualidade das partições, medindo o quão bem estão estruturadas as comunidades encontradas em comparação com um grafo construído de forma aleatória. A modularidade é definida por Newman and Girvan (2004) como

$$Q = \frac{1}{2m} \sum \left[A_{vw} - \frac{k_v \times k_w}{2m} \right] \delta(c_v, c_w)$$

Onde m é o número total de ligações da rede, A_{vw} representa a matriz de adjacências da rede, k_v significa o grau do nó v. A função δ pode ter os valores 1 ou 0, caso o par de ligações (c_v, c_w) pertença à mesma comunidade ou não. Se a fração dos vértices que pertencem à mesma comunidade numa rede for igual ao que é esperado num grafo aleatório, o valor de Q é nulo.

Objetivamente, a modularidade mede o desvio entre a hipótese das ligações das comunidades terem sido geradas pela sua estrutura de comunidades e a hipótese de terem sido geradas aleatoriamente. A modularidade apresenta um valor entre -1 e 1 e, na prática, uma estrutura de comunidades significativa deverá apresentar valores acima de 0.3. Infelizmente, a otimização da modularidade exata é um problema computacionalmente difícil, por isso são consideradas apenas aproximações ao lidar com redes de grande dimensão.
Um dos algoritmos mais rápidos da literatura baseado na otimização da modularidade em redes de grandes dimensões foi proposto por Clauset et al. (2004). A técnica funde de forma recorrente as comunidades de forma a maximizar a modularidade, tratando-se, assim, de um algoritmo "ganancioso". Para além disso, é uma aplicação que tende a produzir comunidades com frações de nós elevadas, mesmo em redes sintéticas, que não têm uma estrutura hierárquica significativa. Este facto condiciona o algoritmo e limita a sua capacidade em redes com mais de um milhão de nós. Outro exemplo é o caso de Guimera and Amaral (2005), que através de *simulated annealing* procuraram maximizar a modularidade. Outro algoritmo importante que é baseado na otimização da função de modularidade, é o Método de Louvain.

2.5 O Método de Louvain

O algoritmo de Louvain, introduzido por Blondel et al. (2008) e adotado para o caso de estudo nesta dissertação, é uma técnica usada para extrair a estrutura hierárquica das comunidades em redes de grandes dimensões de forma completa, possibilitando o acesso a diferentes resoluções da deteção de comunidades. Como afirmam os autores, esta é uma ferramenta heurística, baseada na otimização da função da modularidade.
Para além de proporcionar uma excelente qualidade nas comunidades detetadas, ultrapassa qualquer outro algoritmo na literatura em termos de tempo de computação. Simulações em grandes redes modulares *ad-hoc* sugerem que a sua complexidade é linear em dados típicos e dispersos. Além disso, é um algoritmo altamente intuitivo e de fácil im-

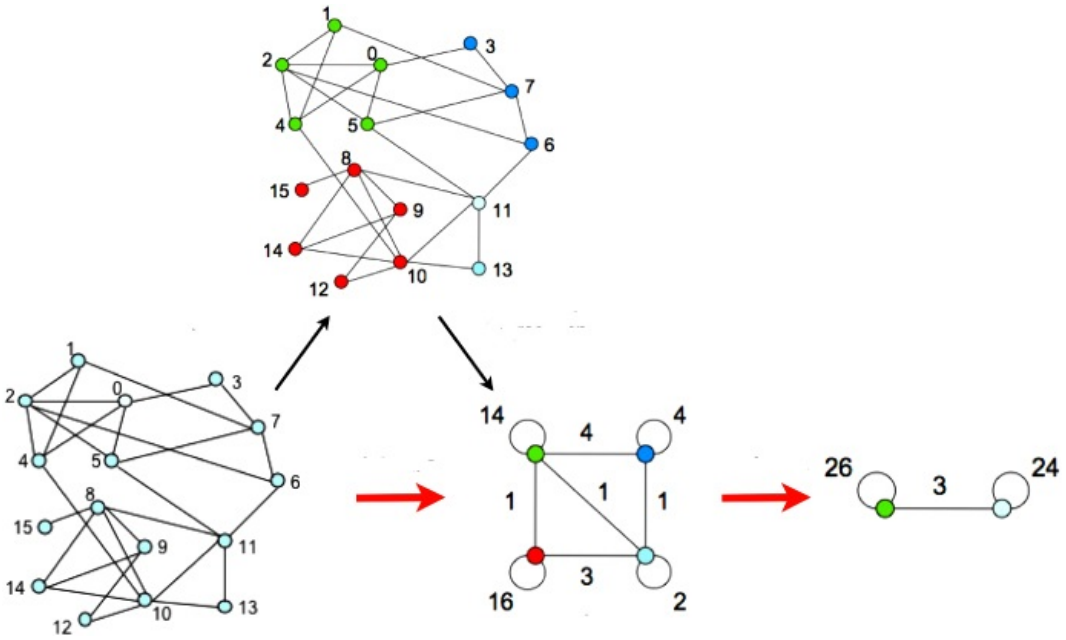


Figura 2.2: Visualização dos passos do algoritmo Método de Louvain. Cada passagem assinalada por uma seta de cor vermelha representa uma combinação das duas fases descritas. As passagens são iteradas até não haver incremento possível na modularidade.

plementação.

O Método de Louvain é um algoritmo não determinístico e que não assegura um máximo global da modularidade. Porém, vários testes confirmam que tem uma excelente precisão que muitas vezes gera decomposições cuja modularidade é muito próxima de 1.

O algoritmo é dividido em duas fases, que são repetidas iterativamente até que um máximo de modularidade seja obtido. Primeiro, o algoritmo procura as comunidades através de uma otimização local da modularidade. Na fase seguinte, o algoritmo volta a construir a rede em análise, na qual os nós são as comunidades detetadas na fase anterior. Estas duas fases são repetidas iterativamente até se maximizar a modularidade, resultando na partição desejada da rede em comunidades. O tamanho da hierarquia que é construída é dada pelo número de combinações das duas fases.

Capítulo 3

Evolução das Comunidades Detetadas

Neste capítulo é apresentada a metodologia MECnet. Na sua essência, o MECnet possibilita a monitorização da evolução de comunidades em redes. Além disso, são também mencionados outros métodos, que também permitem a modelação do comportamento evolutivo das redes.

3.1 Introdução Teórica

As comunidades de uma rede social são dinâmicas e encontram-se em constante mutação. Todos os dias ocorrem eventos que alteram de alguma forma a vida de um indivíduo e que podem eventualmente afetar a sua relação com os indivíduos que pertencem à comunidade em que ele se insere. É neste contexto que se torna importante estudar as comunidades da rede ao longo do tempo. Como Wu et al. (2009) afirma, ao estudar a evolução das comunidades numa rede, analistas têm uma visão mais profunda acerca dos padrões de comunicação das comunidades e dos seus modelos de evolução, o que faz com que a gestão dessas comunidades seja mais fácil em aplicações reais.

Em dados que evoluem ao longo do tempo, a estrutura dos grupos também evolui ao longo do tempo, daí vem que as técnicas de deteção de comunidades têm de ser adaptadas ou, alternativamente, complementadas com metodologias de monitorização da evolução de

comunidades.

O plano para estudar a evolução das comunidades ao longo do tempo consiste, primeiramente, em dividir a rede em vários períodos temporais. Em cada período temporal considerado são detetadas as comunidades. Depois, comparando o conjunto das comunidades, é possível detetar e avaliar transições que ocorrem nas comunidades.

Neste contexto, o processo de mapeamento consiste na descoberta da correspondência das comunidades, obtidas no período temporal x_t e no período temporal x_{t+1}. Assim, é monitorizada a evolução da estrutura das comunidades ao longo do tempo e são encontradas mudanças e possíveis correlações.

Berger-Wolf and Saia (2006), Asur et al. (2009) e Brodka et al. (2013) propuseram metodologias para a análise dinâmica de rede sociais, tendo em conta mudanças nos elementos das comunidades e outros traços caraterísticos.

Para realizar esta tarefa é utilizada a metodologia MECnet, desenvolvida por Oliveira et al. (2014), que trata a monitorização da transição de comunidades. O modelo de representação de grupos eficaz, a facilidade de implementação e o sistema de monitorização torna esta metodologia perfeitamente enquadrada ao problema proposto. Além disso, a contribuição dos autores para esta dissertação, torna a utilização do método ainda mais apelativo.

3.2 MECnet - Monitorizando a Evolução de Comunidades em Redes

O MEC (Oliveira and Gama, 2010) trata-se de uma metodologia que estuda a evolução de grupos ao longo do tempo através da deteção e categorização das transições sofridas por esses grupos, em diferentes intervalos de tempo. O MECnet é a aplicação deste método a dados relacionais.

No MECnet (Oliveira et al., 2014) , as comunidades são representadas por enumeração e

o processo de mapeamento é baseado em probabilidades condicionadas. Para a classificação das transições que ocorrem nos grupos e categorização de padrões e conceitos que evoluem ao longo do tempo é considerada uma taxonomia.

3.2.1 Representação das Comunidades

Como explicado por Oliveira et al. (2014), as comunidades são representadas em enumeração, ou seja, cada uma das comunidades é definida pelo conjunto de elementos que lhe é atribuído pelo algoritmo de deteção de comunidades.

Seja n_i, $(i = 1, ..., N)$, o i-ésimo nó da rede $G(N, E)$, onde N representa o número total de nós da rede e onde E representa o conjunto de ligações. Uma possível representação de uma comunidade detetada no instante de tempo t pode ser definido da seguinte forma: $C_m(t) = (n_1, ..., n_i, ..., n_S)$, onde m representa o índice da comunidade $(m = 1, ..., C$, com C a denotar o número de comunidades detetadas$)$ e S representa o número de nós pertencentes à comunidade C_m.

Ao representar as comunidade desta forma não há azo a perda de informação, uma vez que sabemos a todo o instante quais os elementos que compõem as comunidades, e permite a monitorização de cada elemento ao longo do tempo. Assim, são assegurados resultados das transições mais confiáveis e precisos (Oliveira and Gama, 2010).

Os argumentos de entrada do MECnet são os conjuntos de comunidades gerados a cada período temporal considerado, juntamente com os limiares pré-definidos pelo utilizador: o limiar de sobrevivência e o limiar de separação. O limiar de sobrevivência assume um valor mínimo de 0.5, o que na prática significa que uma comunidade deve manter metade dos seus elementos para que estes se considerem pertencentes ao mesmo grupo, em períodos temporais diferentes. O limiar de separação ajuda a definir os eventos de transição detetados.

3.2.2 Mapeamento das Comunidades

O MECnet é uma metodologia baseada em eventos, funcionando em duas fases. A primeira consiste na deteção de comunidades em cada período temporal analisado, assegurada no caso de estudo de forma independente com o Método de Louvain. Na fase posterior, efetua-se o mapeamento das comunidades. De acordo com a metodologia proposta por Oliveira et al. (2014), é avaliado cada par de possíveis conexões entre comunidades detetadas em intervalos de tempo consecutivos, com base na proporção de elementos mútuos a ambas as comunidades. A proporção referida é calculada através do cálculo de probabilidades condicionadas, da seguinte forma:

$$weight(C_m(t_{i)}, C_u(t_{i+\Delta t})) = P(X \in C_u(t_{i+\Delta t}) \mid X \in C_m(t_i)) = \\ = \frac{\sum P(x \in C_m(t_i) \cap C_u(_{i+\Delta t}))}{\sum P(x \in C_m(t_i))}$$

Onde X é o conjunto de elementos atribuídos à comunidade $C_m(t_i)$ e $P(X \in C_u(t_{i+\Delta t}) \mid X \in C_m(t_i))$ representa a probabilidade de X pertencer à comunidade C_u detetada no período de tempo $t_{i+\Delta t}$, sabendo que X pertence a C_m, detetada no período de tempo anterior t_i.

3.2.3 Taxonomia das Transições

Com o conhecimento da proporção de elementos que pertence a cada par de comunidades geradas em períodos de tempo consecutivos, é possibilitada a construção do ciclo de vida das comunidades, através de um grafo evolutivo. Um grafo evolutivo é um conjunto sequencial de grafos bipartidos. As ligações deste grafo evolutivo são dadas pelas probabilidades condicionadas calculadas na fase de mapeamento. Assim, o ciclo de vida de uma determinada comunidade representa a sua evolução no horizonte temporal, tal como é exemplificado na figura 3.1.

Ao longo do ciclo de vida das comunidades são representadas as transições ocorridas, de acordo com uma taxonomia.

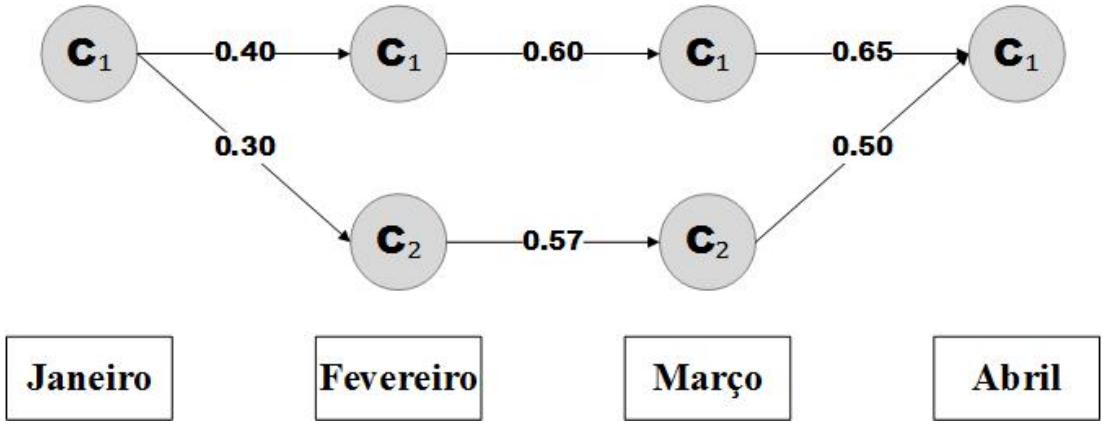


Figura 3.1: Exemplo da representação do ciclo de vida das comunidades. A comunidade C_1 de janeiro separa-se em duas comunidades no mês de fevereiro, a comunidade C_1 e comunidade C_2. De fevereiro para março, ambas as comunidades (C_1 e C_2) sobrevivem. De março para abril, a comunidade C_1 funde-se com a comunidade C_2, formando a comunidade C_1.

A taxonomia considerada para a categorização dos eventos de transição das comunidades em dois intervalos de tempo consecutivos está definida por Oliveira and Gama (2010) da seguinte forma:

- Nascimento - Uma nova comunidade aparece;
- Morte - Uma comunidade desaparece;
- Separação - Uma comunidade separa-se em duas ou mais comunidades;
- Fusão - Duas ou mais comunidades fundem-se numa comunidade só;
- Sobrevivência - Uma comunidade mantém a sua estrutura, não acontecendo nenhuma das transições anteriores.

Estas são transições externas. Além destas, para as comunidades que sobrevivem é possível detetar transições internas, em termos de tamanho e compacidade. Estas definições

suportam a identificação, ao longo do tempo, dos eventos que afetam as interações sociais. A figura 3.2 dá exemplos de cada um dos eventos, formalizados, de acordo com Oliveira and Gama (2010), na tabela 3.1.

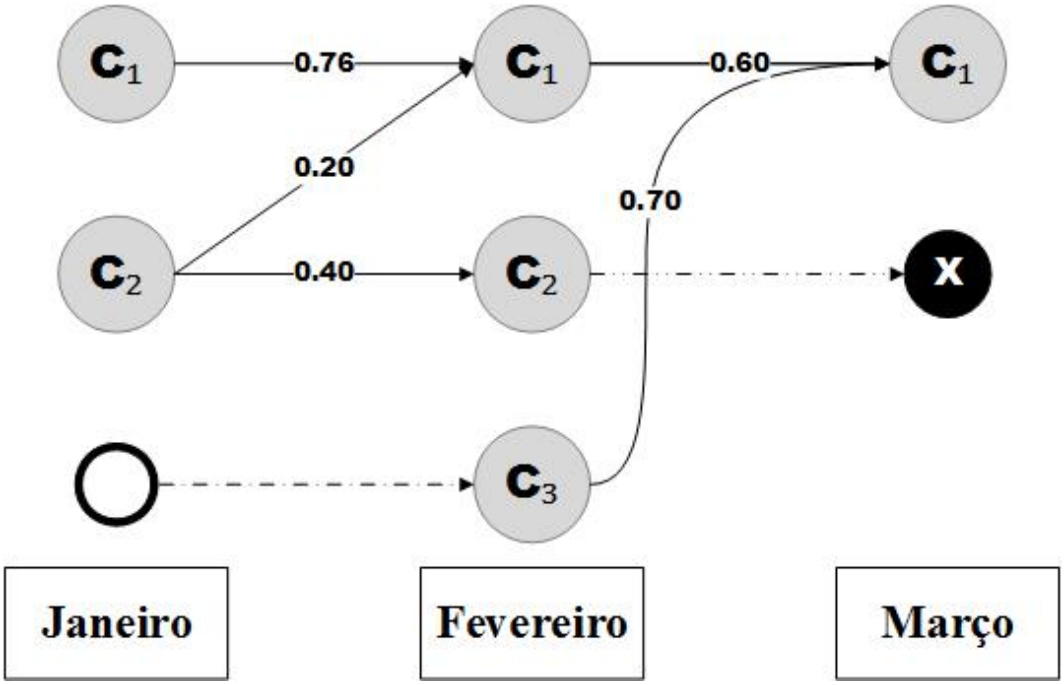


Figura 3.2: Exemplo da representação dos eventos de transição detetados. A comunidade C_1 de janeiro sobrevive na comunidade C_1 de fevereiro. Também entre janeiro e fevereiro, C_2 (janeiro) separa-se em C_1 e C_2 (fevereiro). Entretanto nasce em fevereiro, a comunidade C_3. De fevereiro para março, as comunidade C_1 e C_3 do mês de fevereiro fundem-se na comunidade C_1, do mês de março. A comunidade C_2 de fevereiro desaparece em março.

Taxonomia da Transição	**Notação**	**Definição Formal**
Nascimento de Comunidade	$\emptyset \rightarrow C_u\left(t_{i+\Delta t}\right)$	$0 < weight\ (\ C_m\left(t_i\right), C_u\left(t_{i+\Delta t}\right)\ ) < \tau \forall m$
Morte de Comunidade	$C_m\left(t_i\right) \rightarrow \emptyset$	$weight\ (\ C_m\left(t_i\right), C_u\left(t_{i+\Delta t}\right)\ ) < \lambda \forall u$
Separação de Comunidade	$C_m\left(t_i\right) \subseteq \{\ C_1\left(t_{i+\Delta t}\right), ..., C_r\left(t_{i+\Delta t}\right)\ \}$	$(\exists u \exists v\ :\ weight\ (\ C_m\left(t_i\right), C_u\left(t_{i+\Delta t}\right)\ ) \geqslant \lambda \bigwedge weight\ (\ C_m\left(t_i\right), C_v\left(t_{i+\Delta t}\right)\ ) \geqslant \lambda) \bigwedge \sum_{u=1}^{r} weight\ (\ C_m\left(t_i\right), C_u\left(t_{i+\Delta t}\right) \geqslant \tau$
Fusão de Comunidades	$\{\ C_1\left(t_i\right), ..., C_p\left(t_i\right)\ \} \subseteq C_u\left(t_{i+\Delta t}\right)$	$(weight\ (\ C_m\left(t_i\right), C_u\left(t_{i+\Delta t}\right)) \geqslant \tau \bigwedge \exists C_p \in \varepsilon_i\ \{C_m\} : (weight\ (\ C_m\left(t_i\right), C_u\left(t_{i+\Delta t}\right)) \geqslant \tau$
Sobrevivência de Comunidade	$C_m\left(t_i\right) \rightarrow C_u\left(t_{i+\Delta t}\right)$	$(weight\ (\ C_m\left(t_i\right), C_u\left(t_{i+\Delta t}\right)) \geqslant \tau \bigwedge \nexists C_p \in \varepsilon_i\ \{C_m\} : (weight\ (\ C_m\left(t_i\right), C_u\left(t_{i+\Delta t}\right)) \geqslant \tau$

Tabela 3.1: Definição formal das transições externas que ocorrem nas comunidades em dois intervalos de tempo consecutivos.

Capítulo 4

Metodologia Proposta

Neste capítulo é proposta uma metodologia para estudar as dinâmicas de comunidades em redes de grande dimensão.

Ultimamente tem havido um crescimento considerável no volume de informação disponível, o que fez com que algumas metodologias se tenham tornado desatualizadas para o estudo de dados tipicamente volumosos. Neste contexto, é importante desenvolver metodologias eficientes para analisar esse tipo de dados, de forma a melhorar as decisões nos negócios. No caso particular de redes sociais, as relações de cada indivíduo fornecem informações importantes que devem ser bem analisadas.

De forma geral, o estudo passa por identificar e caraterizar as comunidades em redes de grande dimensão e fazer a monitorização das mesmas ao longo do tempo. A figura 4.1 ilustra a estrutura da metodologia proposta. Na primeira fase é aplicado um método de amostragem não enviesado para redes de grande dimensão. De seguida, é executado o algoritmo de deteção de comunidades a cada período temporal considerado. Depois, é proposta uma seleção multi-critério das comunidades detetadas, para, na fase seguinte, estudar a evolução das comunidades selecionadas no horizonte temporal analisado. Finalmente, é descrito o método para classificar e caraterizar as comunidades selecionadas.

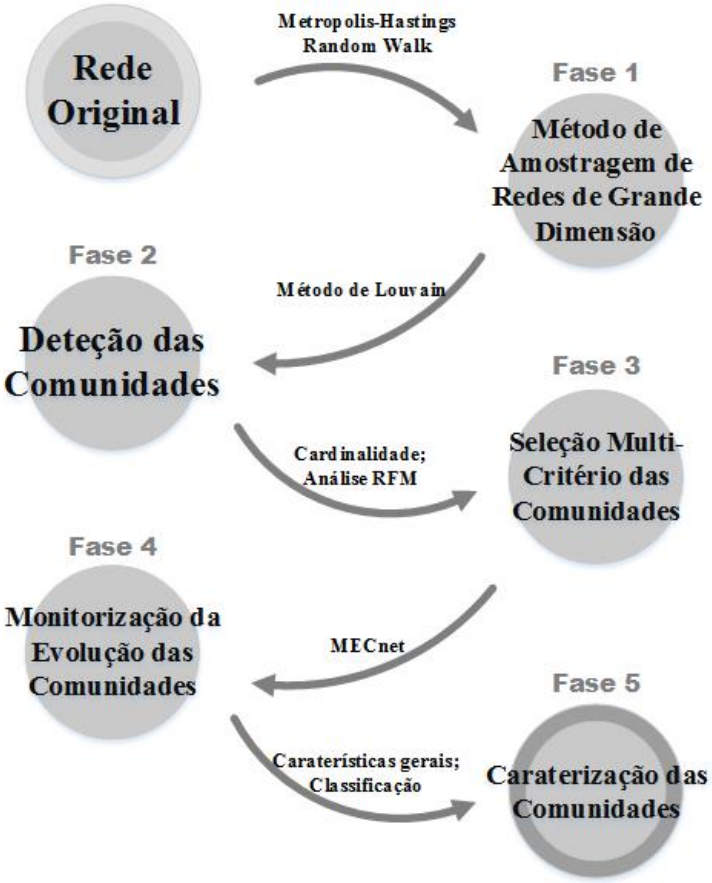


Figura 4.1: Estrutura da metodologia proposta para estudar as dinâmicas de comunidades em redes de grande dimensão.

4.1 Amostragem

Para analisar redes de grande dimensão de forma eficiente é importante efetuar uma amostragem da rede. Seria de esperar que esta problemática se devesse a questões de tempo computacional, mas na verdade o fator principal é em termos de capacidade de armazenamento de dados. O Método de Louvain, o algoritmo de deteção de comunidades é bastante rápido na sua execução, como já foi referido, e não apresenta dificuldades em lidar com a totalidade da rede, num curto espaço de tempo.

Um dos objetivos da amostragem é que a rede obtida através do processo represente um

retrato fiel da rede original. Ou seja, pretende-se obter uma amostra representativa da rede que preserve as suas caraterísticas estruturais. Uma amostragem simples seria escolher um determinado número de elementos da rede, de forma aleatória, e construir a rede de ligações desses elementos com a sua vizinhança. No entanto, este processo proporcionaria resultados enviesados, devido à falta de profundidade na rede, visto que são considerados apenas vizinhos diretos dos elementos amostrados. A inclusão de um segundo grau de ligações na rede, isto é, as ligações entre os vizinhos dos vizinhos dos elementos amostrados, é exequível. No entanto, a amostra continuaria com um enviesamento considerável para elementos com maior grau, ou seja, maior número de ligações.

Se ignorássemos a intenção de manter a estrutura da rede original, uma boa condução deste processo seria escolher a parte dos elementos com maior número de ligações e construir a respetiva rede com o segundo grau de vizinhos. Desta forma, seria obtida uma parte substancial da rede, e certamente a porção mais ativa e influente. Para além disso, o ruído seria reduzido, uma vez que excluiria uma grande parte das ligações que não têm uma frequência considerável.

4.1.1 Amostragem Não Enviesada em Redes de Grande Dimensão

Uma proposta de amostragem que proporciona resultados não enviesados, é o algoritmo *Metropolis-Hastings Random Walk*. Este método já foi aplicado de forma bem sucedida por Gjoka et al. (2010), obtendo uma amostra representativa dos utilizadores do *Facebook*. Algoritmos de procura em largura (*Breadth First Search*), e caminhos aleatórios (*Random Walk*) são outras abordagens populares de amostragem em redes de grande dimensão. No entanto, Wang et al. (2010) mostraram que ambos conduzem a resultados enviesados.

Metropolis-Hastings Random Walk

O algoritmo funciona como explicado a seguir e ilustrado na figura 4.2. Primeiramente, é considerado um nó aleatório v da rede e estabelecido um critério de paragem. Enquanto

esse critério não for encontrado, o algoritmo procura e seleciona um nó w, aleatoriamente, dos vizinhos de v. De seguida, é gerado um α da distribuição uniforme $U(0,1)$. Se $\alpha \leq \frac{k_v}{k_w}$, onde k_v e k_w representam o grau de v e o grau de w, respetivamente, então v é aceite e adicionado à amostra. Depois, w torna-se o nó de referência e repete-se o processo. Caso a condição não se verifique, é escolhido outro vizinho de v.

É sempre aceite a mudança para nós com grau menor, mas rejeitadas algumas mudanças para nós com grau superior. Este fator elimina o enviesamento para nós com maior número de ligações. O método deverá ser aplicado à rede completa, isto é, considerando todos os períodos de tempo analisados, de forma a manter a consistência dos nós amostrados. Caso contrário, se aplicarmos o algoritmo nos diferentes períodos de tempo, poderemos obter uma amostra muito diferente da inicial.

4.2 Deteção das Comunidades

Depois da amostragem da rede, as comunidades são detetadas através da execução do Método de Louvain, introduzido por Blondel et al. (2008) e descrito no segundo capítulo desta tese. O algoritmo de deteção lê como dados de entrada cada ligação existente na rede, juntamente com o respetivo peso. Para além de mostrar a modularidade obtida para a partição gerada, é ainda associado a cada um dos indivíduos, o índice da comunidade a que pertence.

4.3 Seleção das Comunidades

Em redes de grande dimensão, o algoritmo de deteção de comunidades tipicamente produz um grande número de comunidades, o que faz com que seja necessário delinear critérios de seleção de comunidades. Os critérios adotados na metodologia proposta são: a cardinalidade, isto é, o número de elementos incluídos em cada comunidade; a análise RFM (Recência, Frequência e Valor Monetário).

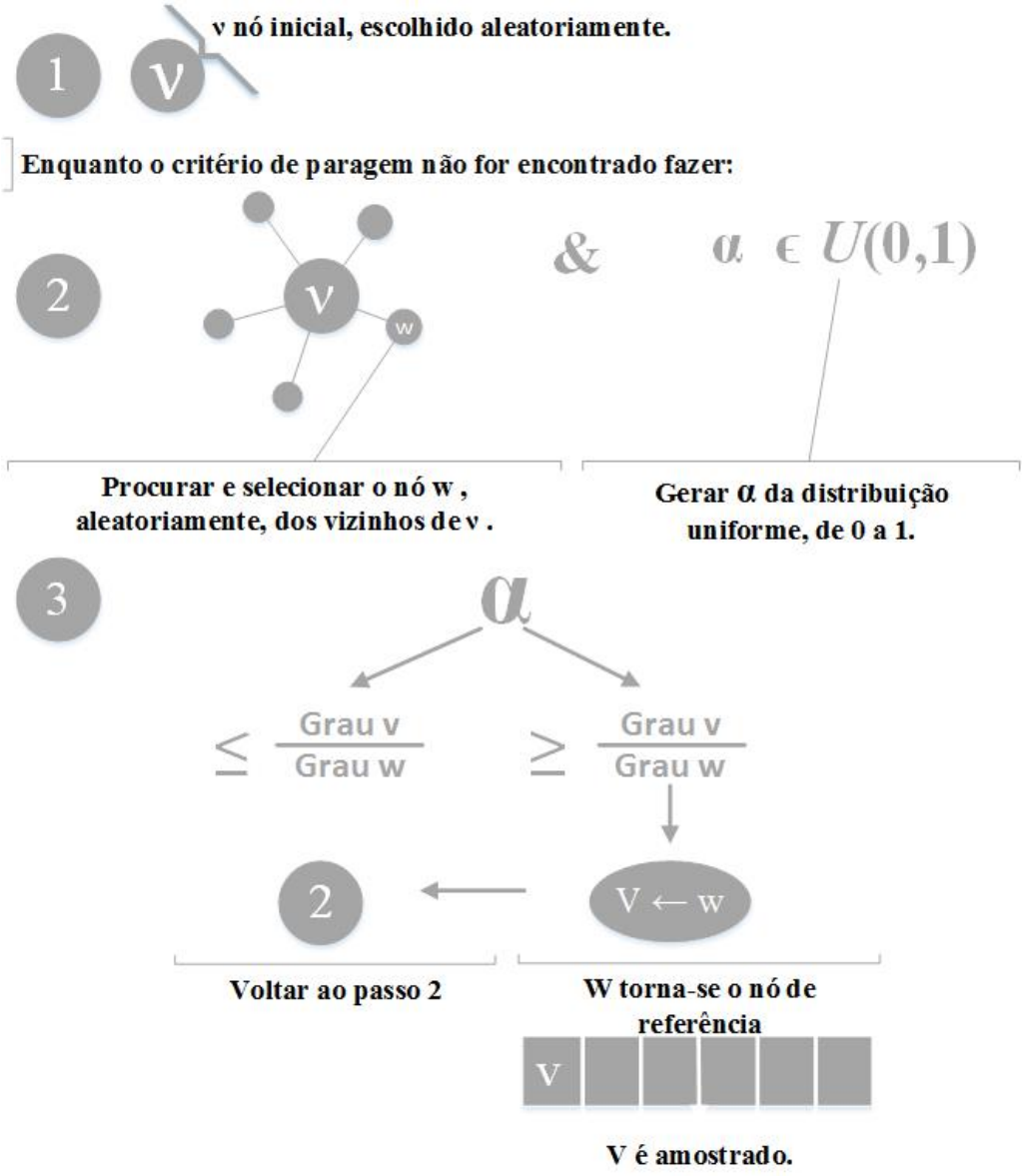


Figura 4.2: Implementação do algoritmo de amostragem para redes de grande dimensão *Metropolis-Hastings Random Walk*.

4.3.1 Cardinalidade das Comunidades

Tipicamente, em redes sociais de grande dimensão, a distribuição da cardinalidade das suas comunidades segue uma Lei de Potência. De acordo com Clauset et al. (2009), uma Lei de Potência é uma relação entre duas variáveis, que ocorre quando uma varia como potência da outra. Nas comunidades detetadas, este fenómeno está associado à distribuição assimétrica positiva acentuada do número de elementos de cada comunidade, onde normalmente se observam muitas comunidades com um número de elementos reduzido e um número pequeno de comunidades com dimensão significativa. Para verificar a ocorrência deste fenómeno nos dados, é executado um teste estatístico, com o recurso à *package* "$poweRlaw$" do software R (Team, 2008).

Para estimar qual o melhor ponto de corte na distribuição, é analisado o decréscimo do número de elementos abrangido por cada percentagem de comunidades considerada, com estas ordenadas por ordem decrescente de cardinalidade. Assim, o corte é executado onde for verificada uma quebra natural da distribuição. Embora seja um procedimento empírico, este revela-se eficaz na seleção das comunidades que representam a maior proporção da rede.

4.3.2 Análise RFM

Um modelo de RFM (recência, frequência e monetária) é uma estratégia de marketing tipicamente usada para determinar de forma quantitativa quais os melhores clientes de uma empresa, de acordo com três vertentes: Recência, referindo-se ao tempo que decorreu desde a última vez que o cliente utilizou o serviço; Frequência, que é o número de vezes que o cliente utilizou o serviço num dado intervalo de tempo; Valor monetário, que corresponde ao montante total gasto pelo cliente no serviço. Segundo Birant (2011), este tipo de análise é tipicamente usado para segmentar uma população de potenciais consumidores em grupos mais ou menos propensos a reagir positivamente às ofertas de mercado. Além de poder ser usado como uma técnica de segmentação, esta análise permite encon-

trar quais os consumidores ou grupo de consumidores mais atrativos.

A implementação do modelo é executada através da ordenação decrescente dos clientes nas três componentes mencionadas. Depois, esses clientes são, em cada componente, separados em cinco quantis. Aos primeiros 20 por cento é atribuída a pontuação máxima de 5, os seguintes 20 por cento têm pontuação 4, e por aí em diante. Finalmente, os clientes são classificados, concatenando as suas pontuações nas três componentes: recência, frequência e monetária. Os melhores clientes estão no quantil 5 de cada vertente e são aqueles que utilizaram o serviço mais recentemente, mais frequentemente e despenderam mais dinheiro.

Existem várias versões do modelo RFM. Numa versão pesada, como Miglautsch (2000) explica, cada uma das componentes é multiplicada por um valor de peso, de acordo com a importância relativa de cada componente.

Em redes de grande dimensão, para além de servir como uma caraterização prévia das comunidades, este modelo apoia a decisão sobre quais as comunidades que podem ser descartadas do estudo, e quais aquelas que são realmente interessantes para o estudo, do ponto de vista do negócio.

4.4 Evolução das Comunidades

Após a seleção das comunidades detetadas, procede-se à monitorização da evolução das comunidades através da aplicação da metodologia MECnet. Primeiramente, e com o auxílio do método baseado em probabilidades condicionadas do MECnet, são estabelecidas as correspondências das comunidades obtidas em cada par de instantes temporais, ao longo do horizonte temporal considerado. É ainda aplicada uma metodologia para classificar as comunidades no tipo de evolução que têm ao longo do seu ciclo de vida: crescimento, estagnação ou declínio.

4.5 Caraterização das Comunidades

Nesta fase é explicada a metodologia que servirá para classificar e caraterizar as comunidades detetadas.

Depois de estudar a evolução das comunidades ao longo do horizonte temporal analisado, torna-se importante procurar inferir hipóteses que expliquem os fenómenos evolutivos detetados. Nesse sentido, as comunidades são caraterizadas tendo por base certas variáveis presentes na base de dados e que não entram na fase de agrupamento. Além destas, são também analisados alguns atributos relacionais das comunidades, como a densidade ou centralidades.

4.5.1 Classificação das Comunidades no Tipo de Evolução

Para examinar as transições com o intuito de estudar as dinâmicas de crescimento e degradação das comunidades, é importante categorizar essas comunidades quanto ao tipo de evolução que estas experienciam. Neste enquadramento, será utilizada a taxonomia ilustrada na figura 4.3. Estas definições são usadas para classificar o tipo de evolução das comunidades no seu ciclo de vida, e não a cada transição entre pares de instantes de tempo sequenciais.

4.5.2 Análise de Rede Sociais

É importante caraterizar as comunidades para compreender o que as distingue e a lógica subjacente ao respetivo agrupamento. O cálculo de métricas oriundas da Análise de Redes Sociais às comunidades é útil para, não só identificar possíveis causas que expliquem certas transições detetadas, mas também para promover a coesão social e o crescimento das comunidades. Assim, as métricas apresentadas suportam o estudo da estrutura e dinâmicas de uma rede social, ao nível das respetivas comunidades.

Para esta pesquisa foi usado o software R (Team, 2008), com a *package igraph*.

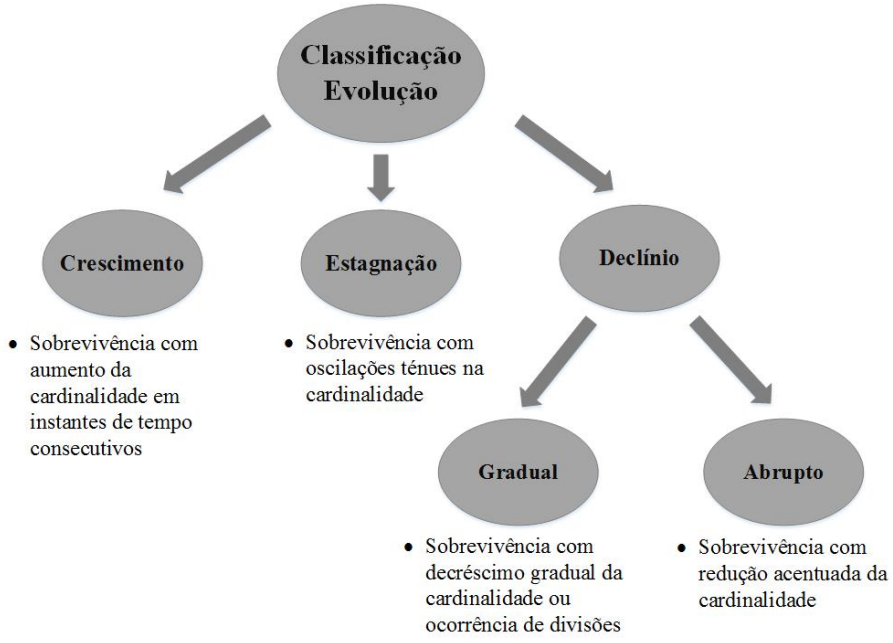


Figura 4.3: Taxonomia para classificação das comunidades no seu tipo de evolução. Consideram-se os seguintes eventos: crescimento, estagnação ou declínio.

Utilizadores chave

Os indivíduos centrais de uma comunidade são provavelmente mais influentes dentro da comunidade que outros mais periféricos. Atores centrais têm acesso mais fácil a informação no contexto da comunidade e comunicam ideias e opiniões aos outros, de forma mais eficiente.

As medidas relacionais calculadas ao nível dos atores são a centralidade do grau, intermediação, proximidade e vetor próprio. É ainda examinada a transitividade local de cada membro da comunidade.

O grau de um utilizador é uma propriedade da sua estrutura, sendo dado pelo número das suas ligações diretas. Esta é uma medida local da conectividade de um indivíduo de uma comunidade, que traduz a sua popularidade.

No contexto do estudo de comunidades e respetiva evolução, analisar o grau dos seus

membros não é suficiente para descrever a centralidade de um utilizador. A habilidade de um indivíduo atuar como intermediário entre diferentes grupos e pessoas e a probabilidade de uma mensagem originada em qualquer ponto da rede chegar a todos os nós são medidas importantes para a classificação dos elementos na sua influência na comunidade em que se inserem.

A intermediação já foi abordada aquando da deteção de comunidades e é definida pela probabilidade de um utilizador ser o caminho mais curto entre outras duas pessoas na comunidade. Segundo Freeman (1979), a proximidade é uma medida de alcance, que mede a "rapidez"com que a informação chega a toda a comunidade, a partir de um nó inicial.

A centralidade do vetor próprio, explicada por Freeman (1979) é uma medida proporcional à soma das centralidades do vetor próprio de todos os utilizadores na sua vizinhança, ou seja, é uma medida que traduz o quanto um utilizador está conectado a outros indivíduos bem conectados. Esta é uma forma semelhante à qual a Google utiliza para classificar as páginas web, com ligações a páginas bem conectadas a terem maior preponderância.

Caraterísticas Estruturais das Comunidades

Para caraterizar as comunidades ao nível da sua estrutura são usadas medidas de análise de redes sociais mais generalizadas à globalidade da rede, sendo essas métricas a densidade, transitividade global e centralizações do grau, intermediação e proximidade.

Densidade

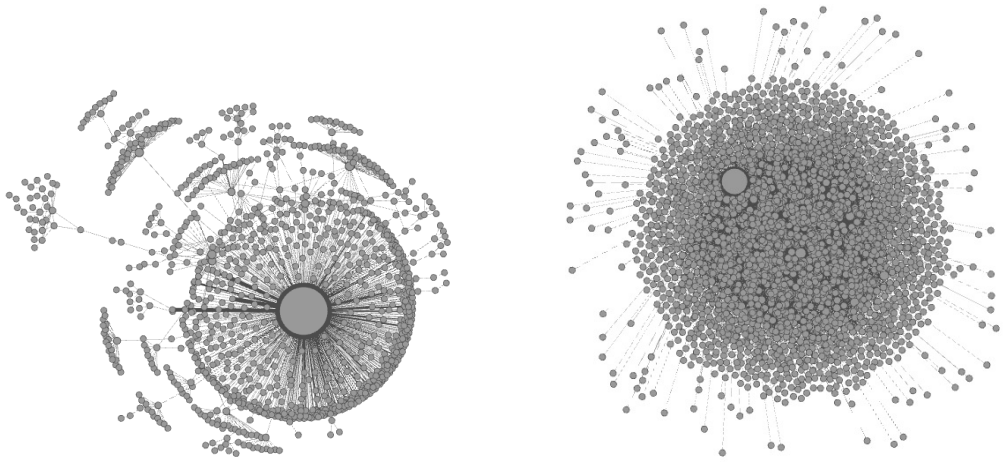
A densidade de uma comunidade é o rácio entre o número de ligações existentes entre os indivíduos da comunidade, e o número total de ligações possíveis entre os indivíduos dessa mesma comunidade. Esta pode ser uma medida interessante para comparar diferentes comunidades, e serve para inferir o quão bem ligada está a comunidade. Uma comunidade perfeitamente conectada tem densidade igual à unidade e chama-se *clique*.

Centralização

Freeman (1979) e Wasserman and Faust (1994) referem-se à centralização como sendo um método geral para calcular a centralidade ao nível do grafo, com base nas medidas de centralidade dos indivíduos. Através desta medida pode-se perceber a uniformidade da distribuição da centralidade dos indivíduos de uma comunidade. A centralização C de um grafo G é definida por

$$C(G) = \sum_{v} \left(\max_{w} c_w - c_v \right), \tag{4.1}$$

Onde c_v e c_v indica a centralidade do indivíduos w e v, respetivamente.
Como se pode verificar graficamente na figura 4.4 se a centralização for alta, então existe um nó interagindo com muitos outros. Pelo contrário, se a centralização for baixa, as interações são distribuídas de forma mais equilibrada.



(a) Comunidade centralizada (b) Comunidade não centralizada

Figura 4.4: Diferença na topologia das redes, para níveis altos e baixos de centralização do grau.

Transitividade Global

Como é referido por Wasserman and Faust (1994), a transitividade de uma rede mede a probabilidade dos indivíduos ligados a um determinado indivíduo estarem também conectados entre si. Formalmente trata-se do quociente entre o número de triplos fechados (*closed triples*) e o número de triplos de indivíduos ligados. Um triplo consiste em três nós que se encontram conectados por duas (triplo aberto) ou três (triplo fechado) ligações.

Capítulo 5

Caso de Estudo

Neste capítulo é aplicada a metodologia proposta, numa rede social de grande dimensão. Por questões de privacidade dos dados, alguns valores são omitidos do estudo. As principais ferramentas de programação utilizadas foram o *Python*, para a implementação do algoritmo de amostragem e criação de dicionários e o *R* para testes estatísticos, cálculo de medidas de análise de redes sociais e execução de uma parte da metodologia. Toda a gestão da base de dados foi executada explorando as potencialidades do *Microsoft SQL Server 2014 Enterprise Edition*.

5.1 Descrição dos Dados

Os dados do caso de estudo consistem numa rede de chamadas de uma grande operadora de telecomunicações nacional.

No conjunto de dados é detalhada informação acerca de cada chamada que envolve a operadora e é também fornecida informação sobre as caraterísticas gerais dos clientes da empresa, tais como o seu plano de atividades. O conjunto de comunicações abrangem um total de mais de mil milhões de registos, referentes a um horizonte temporal total de seis meses. As interações tanto podem ser em termos de chamadas de voz como mensagens de imagens ou texto. Os dados de uma empresa de telecomunicações fornecem informações

fundamentais acerca das atividades de um indivíduo e das suas relações. Informações do tipo quem liga a quem e com que frequência podem ser usadas para criar a rede social de um indivíduo.

5.2 Amostragem da Rede

Inicialmente, foi aplicado o algoritmo de amostragem *Metropolis-Hastings Random Walk*. O algoritmo foi implementado na linguagem $Python$, sendo o critério de paragem o número de iterações. As iterações foram efetuadas na totalidade da rede, isto é, com os dados agregados no horizonte temporal de seis meses. Ao fim de cerca de 21.000 iterações, foram extraídos 6.600 utilizadores. A rede foi construída com dois níveis de profundidade partindo desses 6.600 indivíduos, ou seja, uma rede contendo todas as ligações envolvendo os 6.600 nós mais os seus vizinhos diretos e ainda os vizinhos destes. Ao todo esta é uma rede amostrada bastante substancial, com milhões de registos de comunicações.

5.3 Preparação dos dados

O conjunto de dados é dividido por mês, formando seis subconjuntos que são analisados separadamente. Um ponto fulcral em relação à deteção das comunidades é encontrar aquelas que são de natureza robusta, ou seja, as comunidades que são persistentes ao longo dos vários intervalos de tempo estudados. Por isso, antes de proceder à deteção das comunidades da rede, é importante eliminar o ruído existente na base de dados, que pode afetar a eficácia do algoritmo Método de Louvain. Este pré-processamento é repartido em três fases: Tipo de chamadas de serviço, duração das chamadas e peso mínimo das ligações.

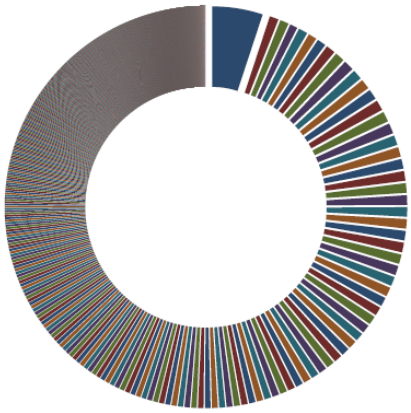


Figura 5.1: Distribuição da duração das chamadas ao longo de um período de um mês, em segundos. O fragmento destacado representa as chamadas com três segundos de duração, o fragmento seguinte, as comunicações com quatro segundos de duração, e assim sucessivamente.

5.3.1 Tipo de Chamadas de Serviço

No contexto do problema da descoberta da comunidade em que cada indivíduo se insere, não faz sentido considerar as ligações estabelecidas entre um cliente e os números de serviço da operadora, pois estes não têm qualquer influência na rede social em que cada indivíduo se insere. O alcance de um número de serviço abrange toda a rede, o que se reflete numa maior centralidade na rede, centralidade esta que se destaca de todas as outras. Este fenómeno afeta fortemente a tarefa de deteção de comunidades. Com base neste raciocínio, todas as chamadas de serviço foram eliminadas dos seis conjuntos de chamadas.

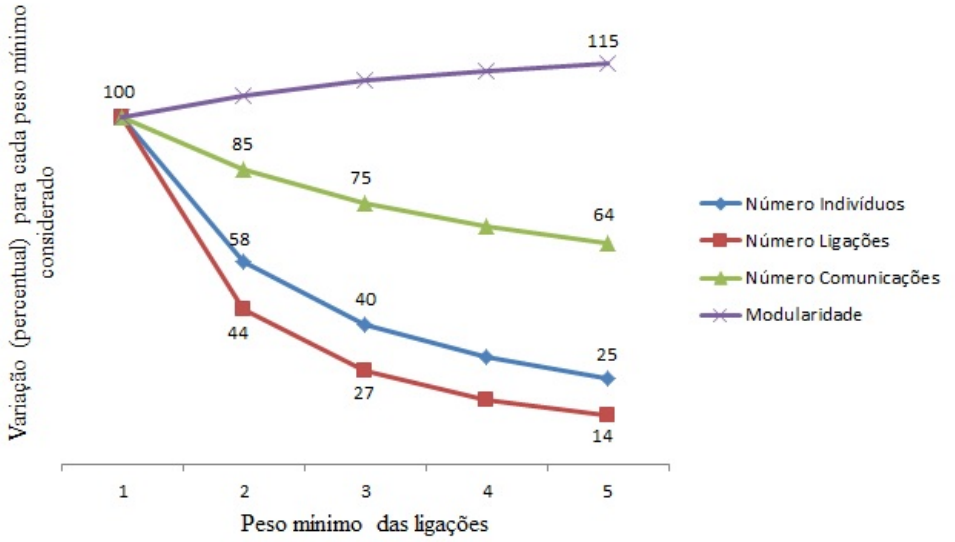


Figura 5.2: Variação percentual do número de indivíduos, número de ligações distintas e volume de comunicações, para diferentes pesos mínimos de ligação considerados.

5.3.2 Duração das Chamadas

A figura 5.1 representa a distribuição da duração das chamadas ao longo de um mês, em segundos. O fragmento destacado representa as chamadas de duração de três segundos, sendo que o seguinte está associado às chamadas de quatro segundos de duração, e assim sucessivamente. Analisando a distribuição, e tendo em conta o significado de uma chamada de voz, considera-se que as chamadas até três segundos são irrelevantes para o contexto do problema. Assim, são contabilizadas apenas as chamadas a partir de quatro segundos.

5.3.3 Peso Mínimo das Ligações

Todos os dias as pessoas interagem com outras sem que haja uma intenção continuada no ato, como por exemplo, quando queremos reservar um lugar num restaurante ou num hotel, ou até mesmo chamadas por engano. Esse tipo de comunicações também podem

ser consideradas pouco pertinentes para a tarefa de deteção de comunidades.

A figura 5.2 mostra, em termos percentuais, a variação no número de ligações distintas, a variação no número de diferentes indivíduos, a variação no número de comunicações e o incremento da modularidade, para diferentes pesos mínimos de ligação considerados. Quando se refere o peso de uma ligação, significa a contagem de comunicações entre cada par de nós da rede. As comunicações são contabilizadas tanto em termos de chamadas de voz, como mensagens de texto ou de imagem. Analisando a figura 5.2, a ideia de que de facto existem bastantes ligações irrelevantes faz sentido. Observa-se um decréscimo contínuo nos valores das variáveis, com exceção da modularidade, que aumenta gradualmente, à medida que o peso mínimo das ligações aumenta. O resultado era esperado, já que ao restringir o peso mínimo das ligações tem-se normalmente menos interações a considerar. No entanto, pode verificar-se um corte abrupto na primeira passagem (do peso mínimo 1 para peso mínimo 2), isto é, quando se tem em conta apenas as ligações com pelo menos duas comunicações recíprocas, no período temporal de um mês. Embora não seja tão acentuado, o mesmo verifica-se na segunda passagem (do peso mínimo 2 para peso mínimo 3), sendo ainda assim, uma descida preponderante.

Os dados apresentados corroboram a ideia de que de facto existem ligações esporádicas e de pouca relevância, por isso são apenas consideradas ligações com pelo menos três comunicações recíprocas entre cada par de indivíduos da rede, isto é, com peso mínimo de ligação igual a 3.

5.4 Deteção das Comunidades

Depois de levar a cabo todo o processamento preliminar dos dados, o algoritmo Método de Louvain é executado para cada um dos seis instantes de tempo considerados no estudo. O tamanho de cada rede gerada é ilustrado na tabela 5.1. Tal como foi mencionado anteriormente, neste passo obtém-se a modularidade para a partição gerada e é ainda associado a cada um dos indivíduos, o índice da comunidade a que pertence.

Tabela 5.1: Dimensão da rede de comunicações, relativamente ao número de indivíduos, ligações e total de comunicações, por mês.

Mês	Indivíduos (em Milhões)	Ligações (em Milhões)	Comunicações (em Milhões)
Julho	1,4	2,2	21,2
Agosto	1,4	2,0	18,2
Setembro	1,3	2,0	18,8
Outubro	1,3	2,0	19,3
Novembro	1,4	2,0	18,0
Dezembro	1,9	2,9	23,2

5.4.1 As Comunidades Detetadas

Os resultados do algoritmo, reportados na tabela 5.2, superaram as expetativas em todos os seis ensaios. A modularidade a rondar valores na ordem dos 0,90 demonstra que realmente existe uma estrutura de comunidades bastante significativa na rede estudada. Embora o número de comunidades descobertas pelo algoritmo seja elevado, a distribuição da dimensão das comunidades segue uma Lei da Potência. Este é um facto que será explorado no processo de seleção das comunidades que serão estudadas ao longo dos seis meses.

5.5 Seleção das Comunidades Detetadas

Como foi mencionado anteriormente, as comunidades que nos interessam estudar são aquelas que persistem ao longo do tempo. Esta ideia, aliada ao facto do algoritmo ter detetado um grande número de comunidades, reforça a importância da tarefa de seleção de comunidades.

Tabela 5.2: Número de comunidades detetadas em cada mês estudado e respetiva modularidade, utilizando o Método de Louvain.

	Mês	**Comunidades Detetadas**	**Modularidade**	
	Julho	7.495	0,90	
	Agosto	5.033	0,91	
	Setembro	5.988	0,91	
	Outubro	5.668	0,90	
	Novembro	5.678	0,90	
	Dezembro	5.610	0,89	

5.5.1 Cardinalidade das Comunidades

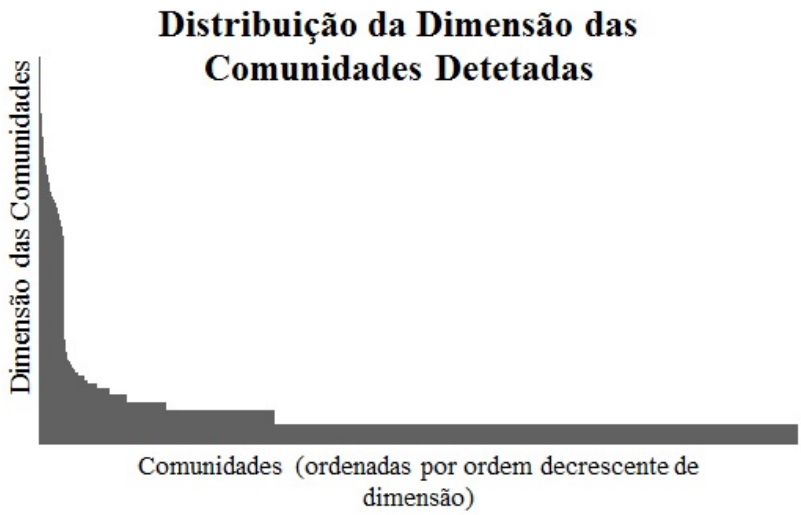


Figura 5.3: A distribuição da dimensão das comunidades detetadas segue uma Lei de Potência.

Em 2.500 testes estatísticos, 2.412 comprovam que, de facto, a distribuição da dimensão das comunidades, esboçada na figura 5.3, segue uma Lei de Potência com um valor

de parâmetro de escala α de aproximadamente 1.96.

Na figura 5.4 está reportado o decréscimo da percentagem de utilizadores abrangida, consoante a percentagem de comunidades considerada, com estas ordenadas por ordem decrescente de cardinalidade. Nestes moldes, é percetível que a quebra natural ocorre quando são consideradas apenas menos de 3% das comunidades com maior dimensão, sendo esse o corte que se aplicou. Concluindo, são apenas considerados os primeiros 3% de comunidades com maior cardinalidade e que abrangem aproximadamente $98,1\%$ da totalidade dos indivíduos da rede. Em princípio, podemos pensar nos indivíduos da rede que acabam descartados deste estudo como sendo indivíduos pouco influentes, visto que pertencem a comunidades pequenas.

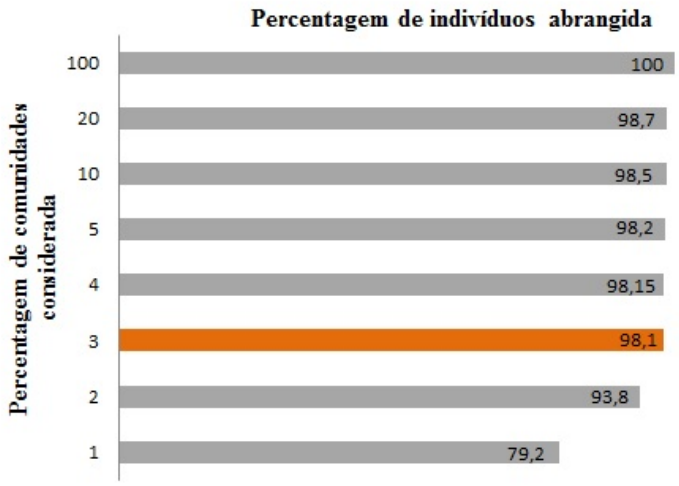


Figura 5.4: Decréscimo da percentagem de utilizadores abrangida, consoante a percentagem de comunidades considerada, com estas ordenadas por ordem decrescente de cardinalidade.

5.5.2 Análise RFM

Através de uma análise de recência, frequência e monetária às comunidades da rede, podemos inferir quais são as comunidades mais importantes nas três vertentes referidas. Estas vertentes são definidas de acordo com o contexto do problema da seguinte forma:

Recência

O tempo médio, em horas, decorrido entre cada comunicação efetuada, é uma boa medida para quantificar o nível de atividade da comunidade.

Frequência

O número médio de comunicações efetuadas pelos utilizadores de cada comunidade.

Monetária

A componente monetária indica o valor total de prémio, retribuído à empresa pela comunidade.

Implementação do Modelo

A figura 5.5 mostra como funciona o modelo. O modelo é implementado baseado numa interpretação pesada, ou seja, é dada uma importância relativa a cada uma das variáveis: recência, frequência e monetária. Como podemos inferir intuitivamente, a monetária é a componente mais relevante para qualquer negócio, pois é aquela que permite o alcance do seu principal objetivo: a geração de lucro. Comparando as vertentes de recência e frequência, é atribuída maior importância à primeira, tal como foi optado anteriormente por Miglautsch (2000) . No contexto do problema, o tempo médio que passa entre cada chamada dá-nos uma melhor perspetiva em relação à atividade de uma comunidade, do que a frequência média de comunicações. Os valores de peso adotados são de 25, 15 e 60 pontos percentuais, para a recência, frequência e monetária, respetivamente. O método de seleção das comunidades a reter, é desenvolvido através do somatório da pontuação

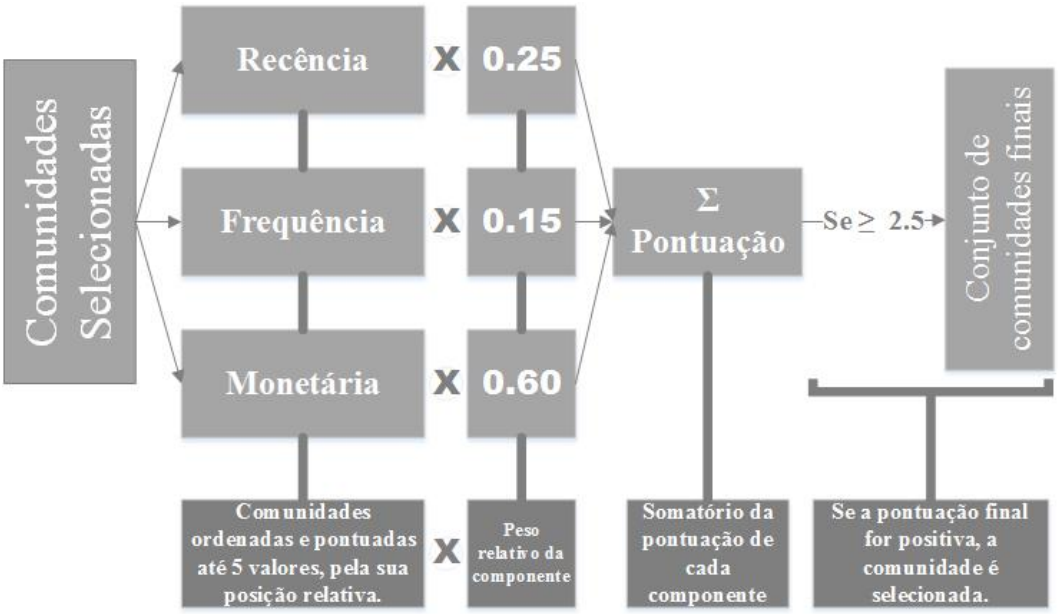


Figura 5.5: Explicação da implementação do modelo RFM. A análise é feita às comunidades selecionadas na secção 5.5.1. A componente monetária é a componente mais preponderante em qualquer negócio tradicional, enquanto que a recência dá uma melhor perspetiva em relação à frequência, no que toca à atividade de uma comunidade.

obtida por cada comunidade em cada uma das três componentes da análise. O somatório final varia de 0 a 5 valores. Se este for positivo, isto é, igual ou superior a 2,5, então a comunidade é considerada para o estudo. Caso contrário, a comunidade é descartada do estudo.

A tabela 5.3 mostra o número de comunidades em cada mês, após a aplicação destes critérios.

5.6 Modelação e Aplicação do MECnet

Inicialmente foram realizadas várias experiências, com o objetivo de afinar os parâmetros do MECnet, em concordância com a natureza dos dados. Com o limiar de sobrevivência e

Tabela 5.3: Número de comunidades selecionadas após aplicação dos critérios.

Mês	Comunidades
Julho	208
Agosto	210
Setembro	192
Outubro	170
Novembro	170
Dezembro	174

de cisão pré-definidos (0.5 e 0.1, respetivamente), as transições detetadas foram bastante escassas. Posto isto, optou-se por reduzir o rigor dos parâmetros. O limiar de cisão foi fixado em 0.0675. No que diz respeito ao limiar de sobrevivência, considera-se que uma comunidade se mantém viva a partir de um limiar de 0.25. Com estes parâmetros foram detetadas todas as transições entre todas as comunidades selecionadas e procuradas essencialmente aquelas comunidades que subsistem ao longo dos seis meses. Além disso, foi utilizado o método de classificação das comunidades no tipo de evolução, com o intuito referido.

5.6.1 Ciclos de vida

As figuras 5.7, 5.8, 5.9 e 5.10 exibem os ciclos de vida das comunidades que persistem ao longo do horizonte temporal estudado ou que têm correspondências e evoluções possivelmente interessantes e devem ser examinadas com atenção. Por outro lado, as transições detetadas entre comunidades cujo ciclo de vida é curto e errático foram excluídas desta fase de estudo. Por exemplo, ciclos de vida caraterizados por nascimentos, imediatamente seguidos de mortes tratam-se de comunidades com um tempo de vida útil reduzido, o que motiva a sua exclusão do estudo.

A figura 5.6 serve para perceber melhor as ilustrações dos ciclos de vida. Os números

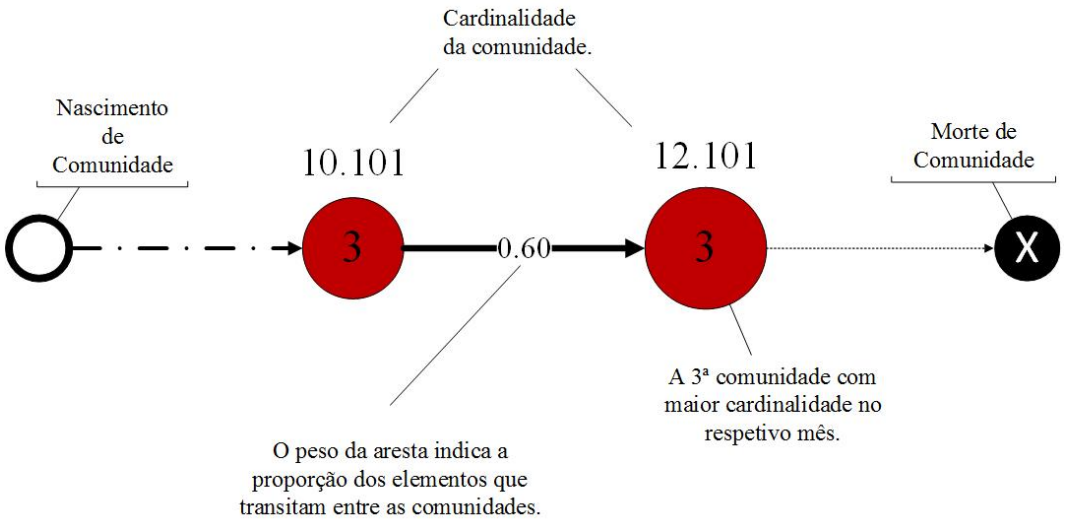


Figura 5.6: Legenda informativa acerca da representação gráfica do quadro dos ciclos de vida das comunidades selecionadas.

contidos no interior de cada comunidade representada por um círculo, simbolizam a sua ordem relativa em termos dimensionais, em cada mês. Por exemplo, em julho, a comunidade '2' é a segunda comunidade com maior cardinalidade nesse mês (Figura 5.7). Os pesos das arestas indicam a proporção de indivíduos que migrou de um mês para o outro. O número inteiro que consta acima de cada comunidade indica a cardinalidade da mesma. Para além disso, os círculos têm um tamanho proporcional à cardinalidade das comunidades. Círculos vazios simbolizam o nascimento de uma comunidade no mês seguinte, enquanto que círculos negros marcados com um 'X' simbolizam o desaparecimento de uma comunidade. As comunidades são denominadas por ordem alfabética ou, alternativamente, pela cor que escolhemos para as caraterizar.

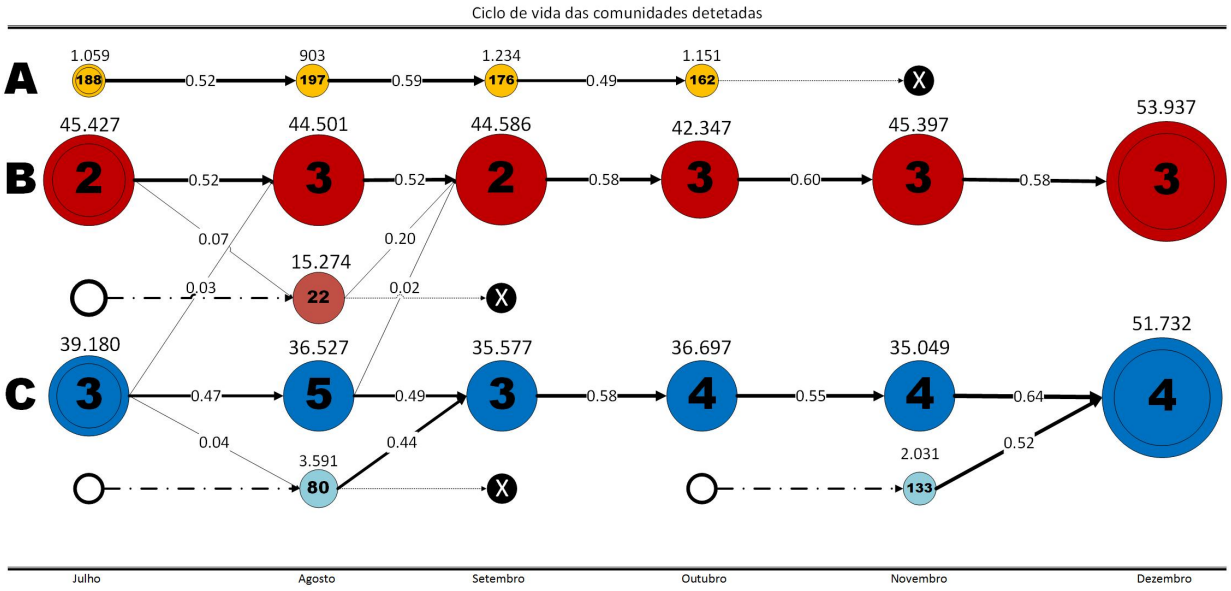


Figura 5.7: Ciclo de vida das comunidades 'A - Amarela','B - Vermelha' e 'C - Azul'. Os ciclos de vida têm um horizonte temporal máximo de seis meses, de julho a dezembro.

Comunidade A - Amarela

A comunidade Amarela representada na figura 5.7 é um sub-grafo que, apesar de ser bastante pequeno comparativamente com os outros, apresenta valores nos pesos das arestas razoáveis, sendo o mais alto de 0.59 (agosto para setembro). A comunidade mantém-se estagnada até outubro e perece no mês seguinte.

Comunidade B - Vermelha

A comunidade Vermelha (fig. 5.7) é, entre todas as comunidades detetadas pelo algoritmo, a comunidade que apresenta maior estabilidade, quer em termos dos valores nos pesos das arestas quer em termos de evolução da cardinalidade. Apesar de sofrer ligeiras oscilações na cardinalidade nos primeiros meses, nos últimos três instantes de tempo verifica-se um crescimento sucessivo, que se torna substancial de novembro para dezembro. Tendo por base este comportamento, é considerada uma comunidade em crescimento. É importante analisar também a migração de

um subconjunto de indivíduos oriundos da comunidade "2", para uma nova comunidade (denominada "22"), durante o mês de agosto. Esta importância de analisar este tipo de transições surge pela necessidade de descobrir se estas ocorrem devido a alterações sociais nas comunidades, isto é, alterações na interação entre os indivíduos que integram as comunidades, ou se são artifícios do algoritmo de deteção de comunidades.

Comunidade C - Azul

A seguir à Vermelha, a comunidade Azul (fig. 5.7) é a comunidade com maior estabilidade. A comunidade Azul está em crescimento, revelando valores nos pesos das arestas satisfatórios e uma cardinalidade elevada. Embora a cardinalidade oscile tenuemente, na última transição a comunidade cresce substancialmente, verificando-se também uma fusão com uma pequena comunidade (comunidade "133"). Analogamente ao que acontece na comunidade Vermelha, em agosto nasce a comunidade "80", parcialmente constituída por elementos que migraram da comunidade Azul.

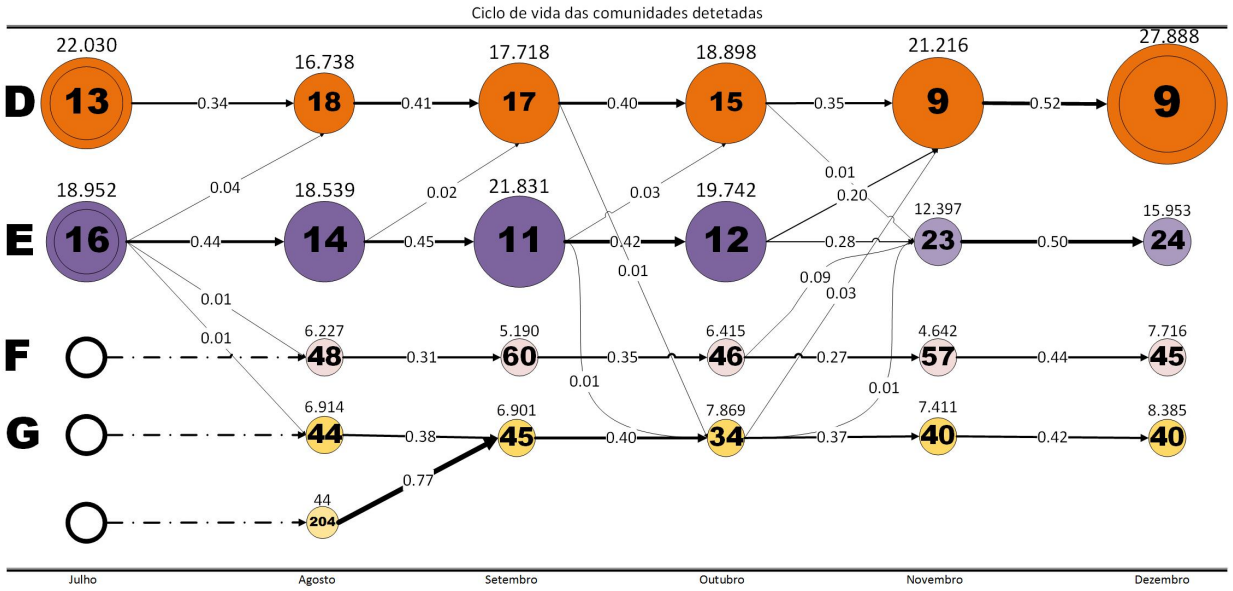


Figura 5.8: Ciclo de vida das comunidades 'D - Laranja','E - Roxa', 'F - Rosa' e 'G - Dourada'. Os ciclos de vida têm um horizonte temporal máximo de seis meses, de julho a dezembro.

Comunidade D - Laranja

A comunidade Laranja (fig. 5.8) cresce gradualmente a partir de agosto, apresentando uma cardinalidade considerável e valores nos pesos das arestas aceitáveis nos meses seguintes.

Comunidade E - Roxa

A comunidade Roxa (fig. 5.8) apresenta um declínio abrupto de outubro para novembro, que é justificado pela migração de 20% dos seus elementos para a comunidade "9"da comunidade Laranja. Será importante analisar com maior profundidade este evento temporal, visto que se trata de um acontecimento reincidente (migração dos elementos da comunidade roxa para a comunidade Laranja).

Comunidade F - Rosa

A comunidade Rosa (fig. 5.8) é uma comunidade que surge apenas em agosto e cujo ciclo de vida apresenta estabilidade. Esta apresenta uma cardinalidade mais baixa em relação às cardinalidades das comunidades que têm vindo a ser analisadas, mas os valores nos pesos das arestas demonstram alguma coesão entre os seus elementos.

Comunidade G - Dourada

Analogamente à Rosa (fig. 5.8), a comunidade Dourada nasce em agosto e sofre oscilações ligeiras na cardinalidade, revelando estagnação.

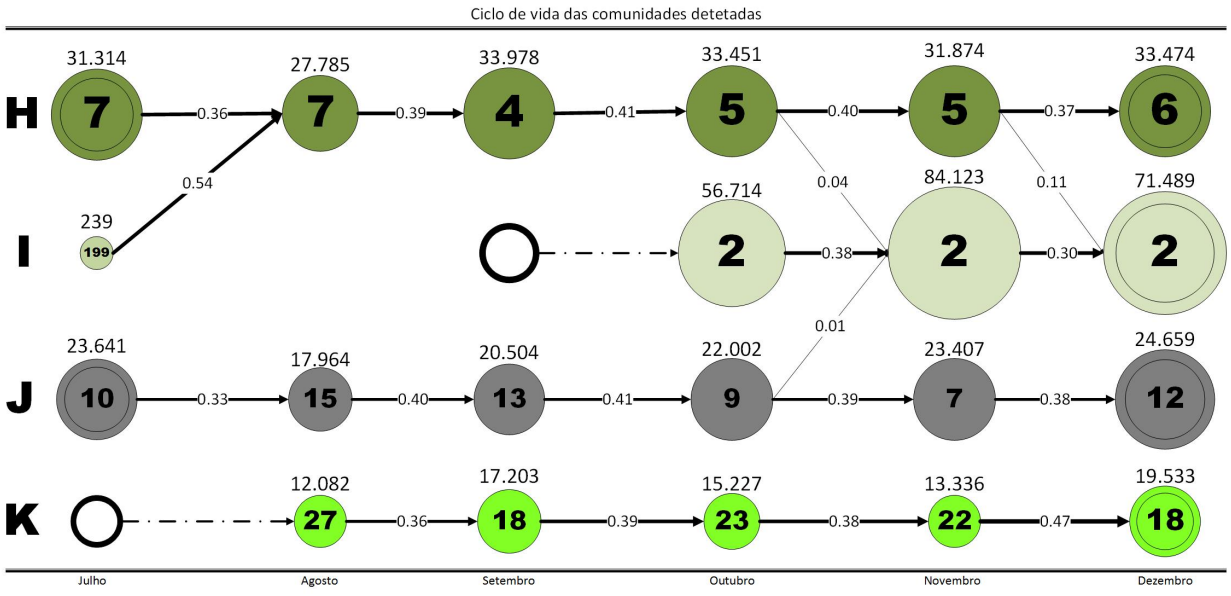


Figura 5.9: Ciclo de vida das comunidades 'H - Verde','I - Verde-Clara', 'J - Cinzenta' e 'K - Verde-Fluorescente'. Os ciclos de vida têm um horizonte temporal máximo de seis meses, de julho a dezembro.

Comunidade H - Verde

A cardinalidade da comunidade Verde (fig. 5.9) oscila ligeiramente entre julho e dezembro, revelando estagnação. Esta é uma comunidade sólida, com valores nos pesos das arestas razoáveis. De notar a fusão com uma comunidade pequena (comunidade "199") em agosto e ainda a migração de uma proporção dos seus elementos, nos dois últimos meses, para a comunidade Verde-Clara.

Comunidade I - Verde-Clara

A comunidade Verde-Clara (fig. 5.9) ganhou forma apenas em outubro. Apresenta uma cardinalidade elevada, tratando-se da segunda comunidade com maior dimensão nos seus três meses de vida. Apesar disso, é difícil inferir num tão curto ciclo de vida, qual o tipo de evolução que a comunidade apresenta, apesar de, com base nas variações da cardinalidade e nos pesos das arestas, existirem indícios de estagnação.

Comunidade J - Cinzenta

A comunidade Cinzenta (fig. 5.9) é uma comunidade estável, que está em crescimento estrito desde agosto, e que revela uma coesão razoável entre os seus elementos.

Comunidade K - Verde-Fluorescente

A comunidade Verde-Fluorescente (fig. 5.9), que surge apenas em agosto, aparenta estar estagnada, tendo por base as pequenas oscilações de cardinalidade detetadas. Relativamente ao peso das arestas, apresenta valores razoáveis e mostram aderência entre os elementos.

Comunidade L - Castanha

A comunidade Castanha (fig. 5.10) indicia um certo crescimento nas duas primeiras transições. Porém, esse crescimento quebra e observa-se um declínio sucessivo e acentuado de cardinalidade nos meses seguintes. Relativamente ao peso das arestas,

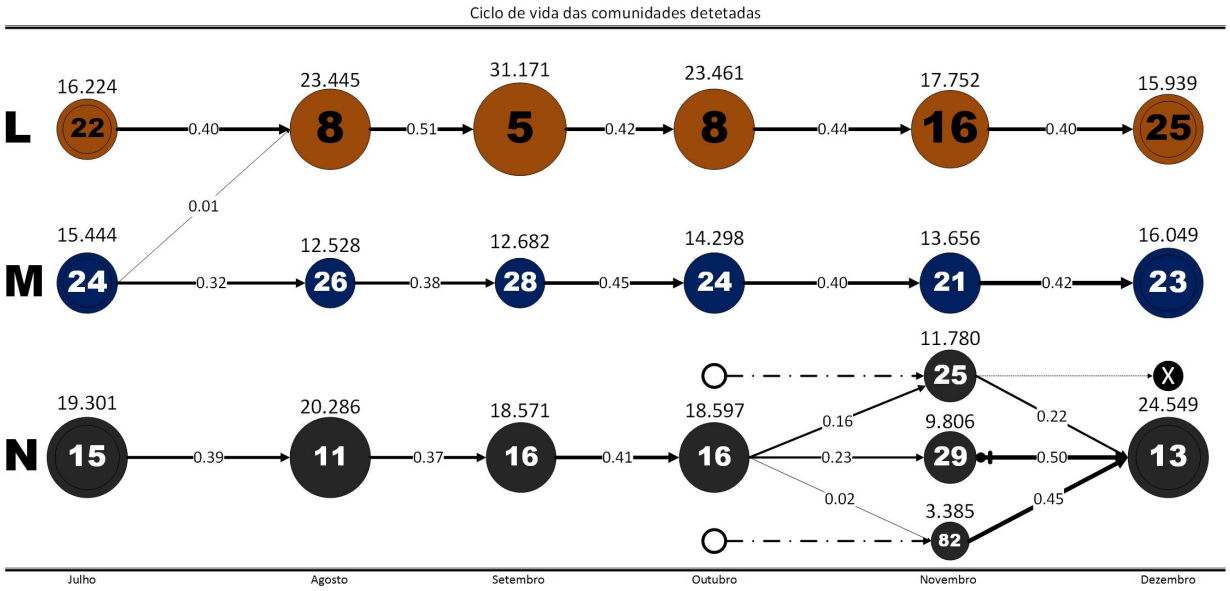


Figura 5.10: Ciclo de vida das comunidades 'L - Castanha','M - Azul-Escura' e 'N - Preta'. Os ciclos de vida têm um horizonte temporal máximo de seis meses, de julho a dezembro.

estes são razoáveis, o que nos diz que existe uma coerência dos elementos mais estáveis da comunidade, ao longo do horizonte temporal.

Comunidade M - Azul-Escura

A comunidade Azul-Escura (fig. 5.10) é semelhante às comunidades classificadas como "estagnadas". É uma comunidade cuja cardinalidade é razoável e oscila ligeiramente.

Comunidade N - Preta

A comunidade Preta (fig. 5.10), apesar de crescer no último mês, revela alguma estagnação ao longo do seu ciclo de vida. De referir as transições que ocorreram nos últimos dois meses, nos quais a comunidade se separou e voltou a unir-se (outubro '16' para novembro '25' e '29' e depois para dezembro '13'). Em novembro também ocorreu uma fusão com uma pequena comunidade.

Como podemos deduzir pelos quadros dos ciclos de vida, as taxas de sobrevivência são relativamente baixas, mas ainda assim consideráveis. Este facto poderá estar relacionado com a volatilidade inerente ao padrão de comunicações de um indivíduo comum. Daí advém a dificuldade em encontrar estabilidade na rede social dos utilizadores. Além disso, com a crescente utilização de computadores e dispositivos digitais, as pessoas tendem a recorrer a outro tipo de plataformas de comunicação, como é o caso das redes sociais online, em detrimento do tradicional telefone/telemóvel para comunicar. Contudo, existem comunidades que subsistem nesses moldes ao longo do horizonte temporal estudado.

Tal como foi mencionado na análise aos quadros dos ciclos de vida, é importante perceber a validade das transições referidas e verificar se algumas destas são ou não um artifício do algoritmo de deteção de comunidades. O Método de Louvain não é um algoritmo determinístico e os seus resultados podem variar. Uma dessas transições é a que ocorre entre a comunidade Laranja(**D**) e a comunidade Roxa(**E**), de outubro para novembro (fig. 5.8). A questão que se coloca é, porque é que 20% dos elementos da comunidade Roxa migram para a comunidade Laranja? Embora seja um acontecimento reincidente, como se pode ver na figura 5.8, esta transição é significativa, ao contrário das anteriores. Verificou-se que, ao longo do horizonte temporal, existem várias ligações entre ambas as comunidades, pelo que podemos concluir que estas estão próximas uma da outra, na rede estudada. Analisando os indivíduos centrais da comunidade Roxa verificou-se que alguns deles são intermediários entre as duas comunidades, Roxa e Laranja. A migração de 20% apresenta estas proporções porque tem por base a migração desses indivíduos centrais, que arrastam quase toda a sua vizinhança. Porém, embora estes sejam acontecimentos interessantes, são inconclusivos, sobretudo por causa da natureza subjetiva dos dados e as caraterísticas não determinísticas do algoritmo de deteção de comunidades.

5.7 Caraterização das Comunidades

A descrição das comunidades proporciona um conhecimento mais aprofundado acerca das mesmas. Depois de descrevermos as dinâmicas das comunidades ao longo do horizonte temporal, é importante examinar as respetivas caraterísticas. Em termos práticos, este tipo de análise permite inferir hipóteses, acerca do tipo de evolução detetado anteriormente e compreender melhor as eventuais causas que justificam algumas das transições observadas.

A descrição realizada é dividida em três fases: Descrição visual, descrição operacional e descrição relacional.

5.7.1 Descrição Visual

'Como Photoshop para grafos' é como os utilizadores do software Gephi (Bastian et al., 2009) o descrevem. Bastian et al. (2009) desenvolveram uma plataforma de visualização e exploração interativa para todo os tipo de redes. Esta é uma ferramenta que permite interagir com a representação, manipular a estrutura, forma e cor, com o intuito de revelar padrões e propriedades escondidas nas redes. É com o apoio desta aplicação que são representadas graficamente as comunidades em estudo.

A figura 5.11 representa graficamente a comunidade Rosa (**F**) em outubro. Os nós que constituem a comunidade têm tamanho proporcional ao seu número de ligações. Com o intuito de proporcionar uma visualização mais agradável foram omitidas ligações com peso menor do que 5. Do esboço pode-se visualizar uma homogeneidade do grau dos indivíduos.

5.7.2 Descrição Operacional

A caraterização operacional baseia-se em atributos do negócio, como é o caso da cardinalidade (número de clientes), da distribuição de planos de atividade, do volume de interações e do volume total da componente monetária associado à comunidade. Segui-

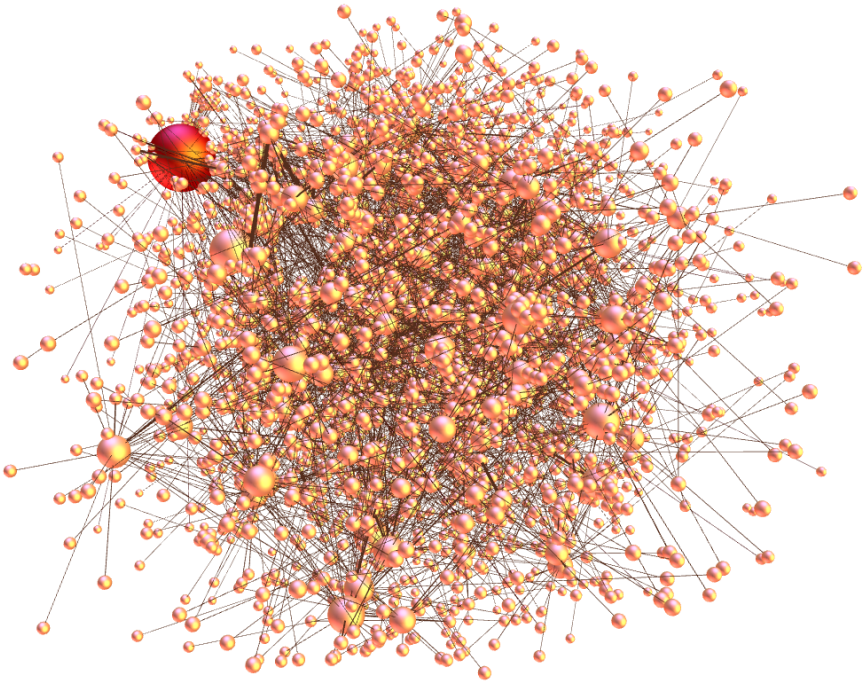


Figura 5.11: Representação gráfica da comunidade Rosa (**F**), no mês de outubro. A dimensão dos nós é proporcional ao seu número de ligações (isto é, o grau).

damente, irão ser apresentados gráficos que representam a evolução das variáveis mencionadas para a comunidade Laranja (**D**, figura 5.8). A escolha desta comunidade como exemplo ilustrativo foi motivada por ter uma cardinalidade considerável e ter caraterísticas representativas das comunidades analisadas nesta fase do estudo.

A figura 5.12 mostra a variação da cardinalidade da comunidade Laranja. Como se verificou no quadro do seu ciclo de vida, a cardinalidade decresce de julho para agosto, mas cresce gradualmente até dezembro, sendo que a sua variação total (isto é, a diferença entre a cardinalidade no último mês e no primeiro) é positiva. De referir que, em todos os meses, pouco mais de metade dos elementos da comunidade pertencem à operadora que forneceu os dados, o que mostra diversidade nas operadoras dos elementos da comunidade.

Da análise da figura 5.13, verifica-se a existência de um plano de atividade dominante

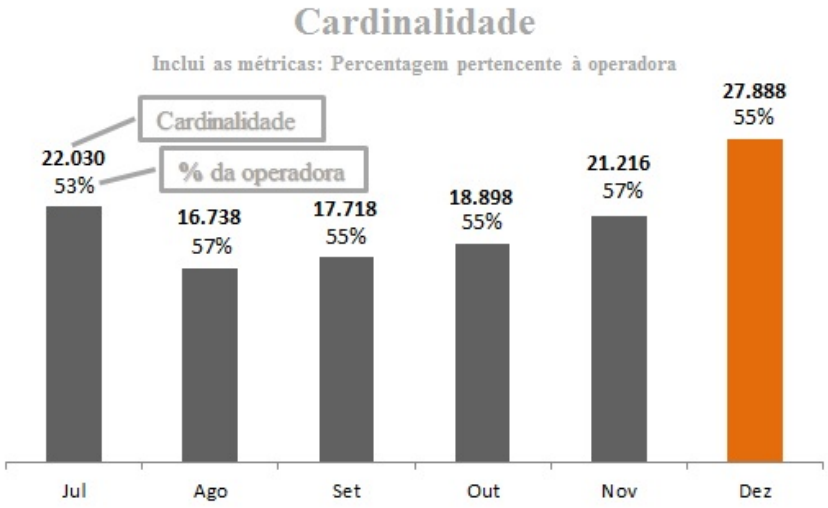


Figura 5.12: Variação da cardinalidade da comunidade Laranja desde julho a dezembro. As percentagens indicam a proporção de clientes que pertence à operadora que forneceu os dados, de julho a dezembro.

(plano A). Ou seja, a maioria dos indivíduos que constituem a comunidade Laranja adotou o planos de atividades A.

O volume de comunicações, ilustrado na figura 5.14, apresenta uma variação semelhante à observada para a cardinalidade, o que faz sentido. Havendo mais utilizadores dentro de uma comunidade, maior a probabilidade de existirem mais comunicações entre eles. Analogamente, o volume de prémio (figura 5.15) também varia consoante a cardinalidade. À medida que esta cresce, o prémio aumenta. As variáveis de cardinalidade, volume de comunicações e prémio atingem o seu pico em dezembro.

5.7.3 Descrição Relacional

As comunidades são sub-grafos com determinados atributos. Os atributos de um grafo ajudam a perceber a forma como os seus elementos se relacionam entre si. Para as comunidades em estudo, os atributos analisados nesta fase são: a densidade, a transitividade

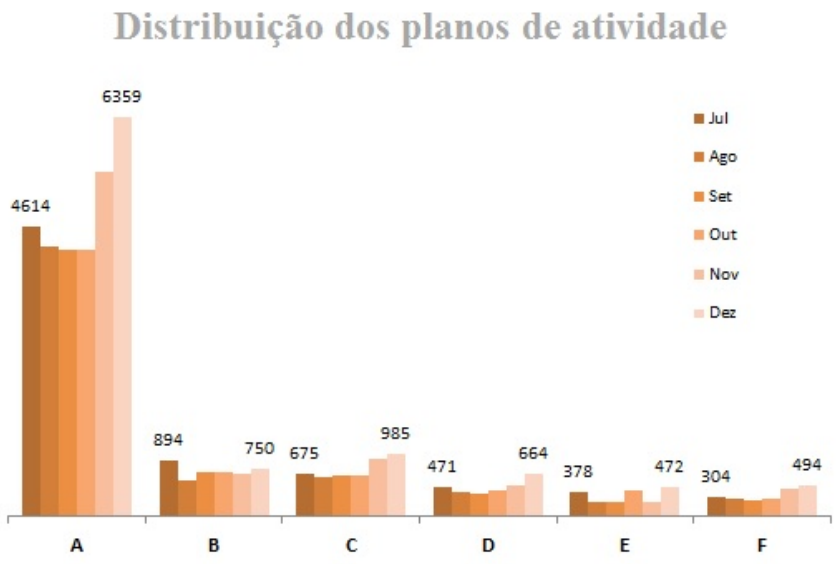


Figura 5.13: Distribuição dos seis principais planos de atividade dos indivíduos da comunidade Laranja, de julho a dezembro.

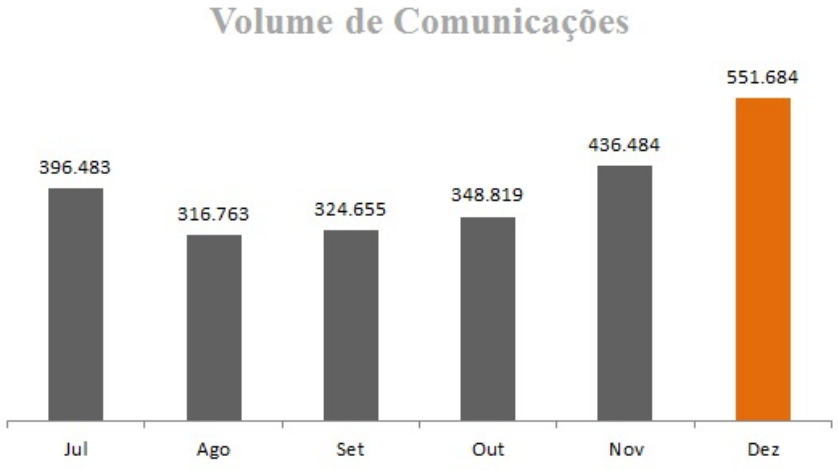


Figura 5.14: Volume de comunicações estabelecidas entre os elementos da comunidade Laranja, de julho a dezembro.

global e as centralizações do grau, da intermediação e da proximidade.

Enquanto que a densidade permite medir a coesão geral das comunidades, a centralização

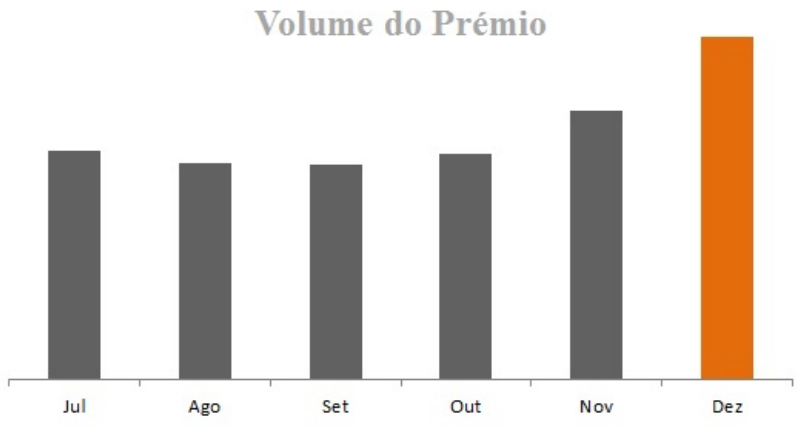
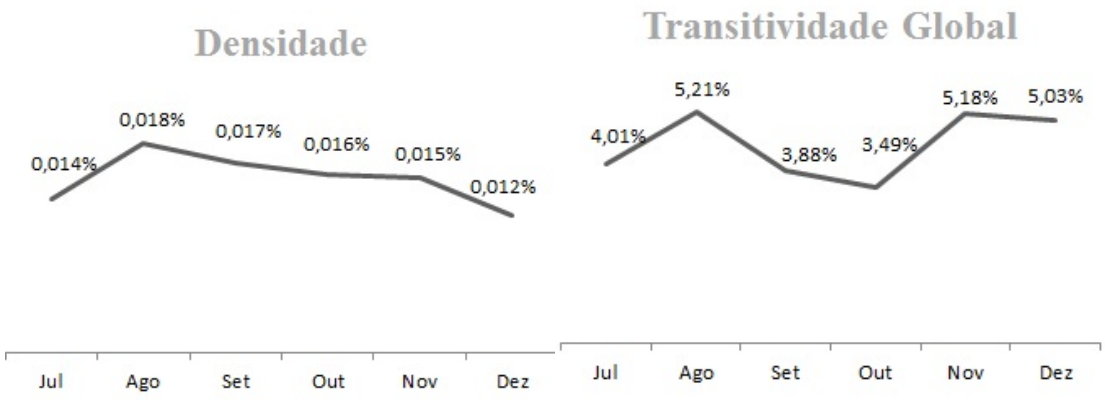


Figura 5.15: Volume do prémio total associado à comunidade Laranja, de julho a dezembro.

indica em que medida a coesão de uma comunidade é assegurada por um ou por vários elementos. Tal como foi referido anteriormente, a densidade é dependente do número de elementos da rede. Assim, à medida que a cardinalidade de uma comunidade cresce, a dispersão entre os seus elementos aumenta, diminuindo a densidade da comunidade. Isto acontece porque normalmente o número de ligações existentes não varia na mesma ordem que o número de ligações possíveis. A transitividade global dá uma indicação geral da topologia da rede. Para este tipo de redes sociais de grande dimensão, não se consegue extrair informação útil destas medidas, como se pode constatar através dos valores obtidos por estas métricas (figura 5.16). Porém, estas métricas podem revelar-se úteis na comparação de diferentes comunidades.

Relativamente às centralizações da comunidade Laranja, estas são baixas, logo a rede não está organizada à volta dos elementos centrais da comunidade.



(a) Densidade (b) Transitividade Global

Figura 5.16: Variação da densidade e da transitividade global da comunidade Laranja, entre julho e dezembro.

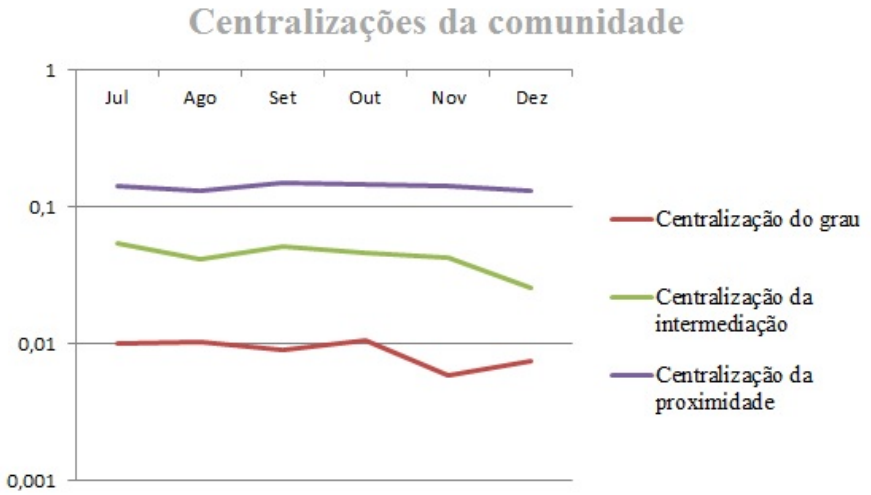


Figura 5.17: Variação das centralizações de grau, intermediação e proximidade da comunidade Laranja, de julho a dezembro. A centralização mais alta é a centralização da proximidade, sendo ainda assim reduzida.

Capítulo 6

Conclusões

Neste capítulo são apresentadas as conclusões obtidas, assim como as limitações do trabalho e recomendações para trabalho futuro.

6.1 Resultados

A análise da evolução de redes sociais através das dinâmicas das comunidades que as compõem pode ter aplicações práticas importantes. Do ponto de vista de negócios, estudar as caraterísticas comuns a grupos de clientes ajuda na tomada de decisões na área do marketing, com o objetivo de potencializar esses grupos. No caso particular das operadoras de telecomunicações, estas contêm dados acerca das atividades dos seus clientes e das suas relações. Essa informação permite-nos construir a rede social de cada um dos clientes e estudar o grupo social em que cada um se insere.

As comunidades de uma rede social nascem, evoluem e morrem. Para perceber como isso acontece é importante gerar e estudar hipóteses, que expliquem esses fenómenos evolutivos sofridos pelas comunidades, e que possam ser usadas pelo marketing. Neste contexto, foi proposta uma metodologia para o estudo das dinâmicas de comunidades em redes de grande dimensão. Com esta metodologia, é executada uma rápida identificação e caraterização das comunidades, e ainda a identificação dos padrões de evolução dessas mesmas

comunidades ao longo do tempo.

Ao analisar uma rede de chamadas telefónicas de grande dimensão, pertencente a uma grande operadora de telecomunicações nacional, comprovou-se a aplicabilidade da metodologia proposta.

6.2 Discussão e Limitações do Trabalho

A complexidade, dimensão e natureza dos dados do caso de estudo constituíram um obstáculo na elaboração deste trabalho. A dimensão dos dados limitou o trabalho, por falta de capacidade de armazenamento de dados. Nesse sentido, foi implementado o algoritmo de amostragem não enviesado *Metropolis-Hastings Random Walk*, que permitiu obter uma amostra representativa da rede original. Este é um método para amostragem em redes que ultrapassa técnicas mais tradicionais, como a pesquisa em largura ou os caminhos aleatórios.

A modularidade das comunidades obtidas com a execução do algoritmo de deteção de comunidades Método de Louvain rondou os 0.90, o que denota a existência de uma estrutura de comunidades significativa na rede, o que também foi influenciado por uma rigorosa preparação dos dados.

Através da cardinalidade e de uma análise RFM, foram extraídas as comunidades que, em princípio, têm um impacto mais significativo na rede. Embora seja difícil encontrar estabilidade nas dinâmicas evolutivas na rede social de um indivíduo comum, foi verificada uma coerência entre as comunidades selecionadas, ao longo do tempo. Com um espaçamento de um mês entre a execução de cada mapeamento das comunidades através da metodologia MECnet, existe uma certa harmonia entre as comunidades persistentes, principalmente entre aquelas com maior cardinalidade. Em termos práticos, esta ideia de articulação poderá aumentar a confiança quando a operadora decidir atuar sobre a sua rede, de que de facto existe nela uma evolução natural.

Depois do exame às caraterísticas gerais da comunidade, procuraram-se as causas que

levaram às transições detetadas, entre diferentes comunidades. Esta procura partiu sobretudo da monitorização do comportamento dos atores centrais das comunidades. Quer dizer, quando uma parte substancial de indivíduos migra para outra comunidade, analisar o quão determinante foram os utilizadores chave nesse processo. Embora realmente se tenha encontrado uma correlação entre os fatores, isto é, quando estão envolvidos indivíduos centrais, as transições são tipicamente mais fortes, não se pode tirar conclusões indiscutíveis. É arriscado justificar as transições com alterações sociais nas comunidades, já que o algoritmo de deteção de comunidade é um método não determinístico. As comunidades detetadas podem variar ligeiramente, o que faz com que esse tipo de transições possa ser algo artificial.

6.3 Trabalho Futuro

O método de amostragem utilizado proporcionou um retrato fiel daquilo que é a rede original. No entanto, com a crescente exploração de técnicas de manipulação de dados de grandes dimensões, seria interessante estudar a aplicabilidade desses modelos na rede do caso de estudo, de forma a melhorar a eficiência das tarefas abordadas, aludindo a totalidade da rede. Um exemplo de um método apropriado seria o *Hadoop Distributed File System*, proposto por Shvachko et al. (2010).

Procurar extrair a relação entre o tipo de evolução das comunidades e as variáveis de caraterização, elevaria a qualidade da caraterização das dinâmicas das comunidades. Quer dizer, por meio de um classificador, criar um modelo que justifique, em termos das variáveis utilizadas na caraterização das comunidades, o fenómeno evolutivo de crescimento, estagnação e declínio.

Bibliografia

Aggarwal, C. C., Han, J., Wang, J., and Yu, P. S. (2003). A framework for clustering evolving data streams. In *Proceedings of the 29th International Conference on Very Large Data Bases - Volume 29*, VLDB '03, pages 81–92. VLDB Endowment.

Asur, S., Parthasarathy, S., and Ucar, D. (2009). An event-based framework for characterizing the evolutionary behavior of interaction graphs. *ACM Transactions on Knowledge Discovery from Data*, 3(4):16:1–16:36.

Baruah, R. D. and Angelov, P. (2012). Evolving social network analysis: A case study on mobile phone data. In *Evolving and Adaptive Intelligent Systems (EAIS), 2012 IEEE Conference on*, pages 114–120. IEEE.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In Adar, E., Hurst, M., Finin, T., Glance, N. S., Nicolov, N., and Tseng, B. L., editors, *ICWSM*. The AAAI Press.

Berger-Wolf, T. Y. and Saia, J. (2006). A framework for analysis of dynamic social networks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 523–528, New York, NY, USA. ACM.

Birant, D. (2011). Knowledge-oriented applications in data mining. In (Ed.), P. K. F., editor, *Data Mining Using RFM Analysis*, pages 91–108. INTECH Open Access Publisher.

Blondel, V., Guillaume, J., Lambiotte, R., and Mech, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech*, 10:1–12.

Brodka, P., Saganowski, S., and Kazienko, P. (2013). Ged: the method for group evolution discovery in social networks. *Social Network Analysis and Mining*, 3(1):1–14.

Chen, J. and Yuan, B. (2006). Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics*, 22(18):2283–2290.

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

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174.

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

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

Gjoka, M., Kurant, M., Butts, C. T., and Markopoulou, A. (2010). Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. In *Proceedings of IEEE INFOCOM '10*, INFOCOM'10, pages 2498–2506, San Diego, California, USA. IEEE Press.

Guha, S., Meyerson, A., Mishra, N., Motwani, R., and O'Callaghan, L. (2003). Clustering data streams: Theory and practice. *IEEE Trans. on Knowl. and Data Eng.*, 15(3):515–528.

Guimera, R. and Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recogn. Lett.*, 31(8):651–666.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Cam, L. M. L. and Neyman, J., editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.

Miglautsch, J. R. (2000). Thoughts on RFM scoring. *The Journal of Database Marketing*, 8(1):67–72.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113.

Oliveira, M. D. B. and Gama, J. a. (2010). Mec - monitoring clusters' transitions. In Agotnes, T., editor, *STAIRS*, volume 222 of *Frontiers in Artificial Intelligence and Applications*, pages 212–224. IOS Press.

Oliveira, M. D. B., Guerreiro, A., and Gama, J. (2014). Dynamic communities in evolving customer networks: an analysis using landmark and sliding windows. *Social Netw. Analys. Mining*, 4(1).

Pizzuti, C., Rombo, S. E., and Marchiori, E. (2012). Complex detection in protein-protein interaction networks: A compact overview for researchers and practitioners. In *Proceedings of the 10th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, EvoBIO'12, pages 211–223, Berlin, Heidelberg. Springer-Verlag.

Shvachko, K., Kuang, H., Radia, S., and Chansler, R. (2010). The hadoop distributed file system. In *Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Sys-*

tems and Technologies (MSST), MSST '10, pages 1–10, Washington, DC, USA. IEEE Computer Society.

Taskar, B., Abbeel, P., and Koller, D. (2002). Discriminative probabilistic models for relational data. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, pages 485–492, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Team, R. D. C. (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN: 3-900051-07-0.

Wang, T., Chen, Y., Zhang, Z., Sun, P., Deng, B., and Li, X. (2010). Unbiased sampling in directed social graph. *SIGCOMM Comput. Commun. Rev.*, 40(4):401–402.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Wu, B., Ye, Q., Yang, S., and Wang, B. (2009). Group crm: a new telecom crm framework from social network perspective. In Wang, J., Zhou, S., and Zhang, D., editors, *CIKM-CNIKM*, pages 3–10. ACM.