INSTITUTO SUPERIOR DA CIÊNCIAS DO TRABALHO E DA EMPRESA
Departamento de Ciências e Tecnologias de Informação

Detecção de comunidades no sistema de correio electrónico universitário

DAVID MANUEL DE SOUSA RODRIGUES

Tese submetida como requisito parcial para a obtenção do grau de
MESTRE EM CIÊNCIAS DA COMPLEXIDADE

Orientador:
Professor Jorge Louçã, Professor Auxiliar,
Instituto Superior de Ciências do Trabalho e da Empresa

Março, 2009

Conteúdo

Lista de Tabelas

Lista de Figuras

Resumo

O estudo de sistemas estruturados em redes sociais conheceu inúmeros desenvolvimentos na aplicação da teoria de grafos às ciências sociais. Um dos aspectos recentes tem sido o da detecção de módulos, ou comunidades, em redes sociais. Diversos algoritmos e estratégias tem sido desenvolvidos para identificar a estrutura existente por detrás das interacções sociais.

Através de um estudo de caso, mostrámos a existência de comunidades de comunicação informal que utiliza a rede de correio electrónico do ISCTE, através da aplicação de algoritmos hierárquicos de detecção de comunidades. Analisámos a estrutura hierárquica da rede através de *k-cores* e verificámos que a as comunidades de comunicação informal formadas ultrapassam as fronteiras dos departamentos institucionais através do método de percolação de cliques. Às comunidades detectadas aplicámos uma medida de variação de informação para determinar a distância entre os diversos departamentos.

Construímos um modelo de simulação multi-agente, para mimar o sistema de comunicação informal através de correio electrónico, CIUCEU, que nos permitiu verificar a influência da vizinhança "social" dos agentes na criação e manutenção da estrutura da rede de professores do ISCTE. Analisámos ainda a utilização de simulações alimentadas por dados reais, concluindo sobre as implicações da utilização de dados reais sobre o desenho da simulação.

Palavras chave: detecção de comunidades, percolação de cliques, modelação multi-agente, complexidade, redes sociais, análise de *k-cores*, algoritmos hierárquicos, modularidade

Abstract

The study of structured systems in social networks has gone through several developments by the use of graph theory in social sciences. On aspect that has been given considerable attention in recent years is the module or community detection in social networks. Several algorithms and strategies have been developed to identify the structure behind social interaction.

Through a case study we show the existence of communities based on informal communication that use the email system at ISCTE. We applied a set of hierarchical algorithms to detect communities. Also, we analyzed the hierarchical structure through the *k-cores* method and verified the transitivity of the communities detected through clique percolation to put in evidence that informal communities are transversal to the institution departments. We also used a information variation measure to compare distances between different clusterings.

We built a multi-agent simulation to model the informal communication mechanism of the email system, CIUCEU. This is used to verify the dependence of the system on the notion of social neighborhood, in the teachers network of ISCTE. We also analyzed the usage of real data and concluded on its implications of the sampling and drawing os multi-agent simulations.

Keywords: community detection, clique percolation, multi-agent simulation, complexity, social networks, *k-core* analysis, hierarchical algorithms, modularity

Agradecimentos

A execução de uma tarefa como a produção de uma tese, apesar de ser por momentos um trabalho solitário, não seria possível sem o apoio e incentivo de muitas pessoas.

Em particular, queria agradecer ao meu orientador, o Professor Jorge Louçã, pelo empenho e motivação que me transmitiu através das inúmeras conversas que tivemos sobre este e outros assuntos relacionados com o estudo de sistemas complexos.

Ao Professor John Symons pelas conversas tivemos e comentários que fez ao longo do processo de desenvolvimento e escrita da tese.

Ao Professor Manuel Sequeira pela disponibilidade e apoio à esta tese, tendo aceite facilitar os dados e participar nas discussões que levaram à implementação de um sistema efectivo de anonimização de dados, o que permitiu realizar este estudo com a salvaguarda de que os utilizadores não poderiam ser identificados nem o conteúdo das suas mensagens conhecido.

Ao João Paulo Pires, pela paciência que teve para com os meus caprichos e pedidos, nunca se tendo negado a nada do que lhe pedi e estando sempre pronto para ajudar no que foi preciso e à equipa do DSI em geral pelo apoio nesta aventura.

À minha família que sempre me incentivou a continuar a cada telefonema que fazia mostrando as minhas dúvidas.

Queria agradecer muito especialmente à Mafalda, minha companheira, que me aturou as crises e desesperos e partilhou as minhas descobertas e alegrias com uma paciência que só um amor muito grande pode justificar.

Capítulo 1

Introdução

> *"Let me put it this way: Planet Earth has never benn as tiny as it is now. It shrunk - relatively speaking of course - due to the quickening pulse of both physical and verbal communication. This topic has come up before, but we had never framed it quite this way. We never talked about the fact that anyone on Earth, at my or anyone's will, can now learn in just a few minutes what I think or do, and what I want or what I would like to do. If I wanted to convince myself of the above fact: in couple of days I could be - Hocus pocus! - where I want to be."*
>
> Frigyes Karinthy *in Chain-Links (Karinthy, 1929)*

Este trabalho surge do desejo de compreensão do funcionamento dos sistemas estruturados que apresentam características de auto-organização e que surgem de forma informal em redes sociais do nosso dia a dia. Tais sistemas são abundantes no nosso mundo e a sua caracterização revela que nem sempre se organizaram da forma que os que desenvolveram o sistema idealizaram, assumindo características próprias. Esta 'apropriação' do sistema por parte dos utilizadores é característica de redes sociais, fazendo com que as representações topológicas da comunicação não coincidam com os quatro tipos de redes clássicos (grelha regular, aleatórias, *small-world* e *scale-free*)(Hamill e Gilbert, 2008). As redes sociais não são naturalmente aleatórias, uma vez que as pessoas tem tendência a ligar-se a pessoas suas conhecidas e com quem apresentam alguma afinidade. A estrutura das suas ligações não obedece também a uma grelha rígida pré-definida uma vez que as relações sociais mudam

com o passar do tempo e não obedecem a padrões repetitivos ao longo de grelha formada pelos indivíduos. No que diz respeito a redes do tipo *scale-free*(Barabasi e Albert, 1999), estas também não são adequadas para explicar as redes sociais onde as pessoas se inserem, uma vez que neste modelo há o estabelecimento de novas ligações em função do número de ligações já estabelecidas pela pessoa ligada. Embora o fenómeno da popularidade seja importante para o estabelecimento de novas ligações nas redes sociais, não é o único mecanismo pelo qual as pessoas se conectam, isto para além da impossibilidade de um indivíduo conhecer em detalhe o número de ligações que outro mantém. Para além disso nas redes sociais há uma tendência para que pessoas com muitas ligações estejam também elas ligadas a pessoas com bastantes ligações. Este fenómeno não se verifica nas redes *scale-free*. As redes do tipo *small-world* podem ser descritas por um modelo proposto por Strogatz e Watts (1998) em que se obtém a rede a partir de uma grelha regular na qual aleatoriamente algumas ligações são destruídas e substituídas por outras. A rede assim produzida apresenta elevada transitividade e distâncias curtas entre os seus membros. Encontra-se entre as redes aleatórias e as grelhas regulares, no entanto esta reorganização não encontra paralelo nas redes sociais. As pessoas não se desligam e reconectam de uma forma aleatória nem a rede social de partida é do tipo grelha regular. Newman *et al.* (2006, p. 292) reconhecem mesmo que o modelo *small-world* "não é geralmente um bom modelo de redes reais, incluindo redes sociais".

A questão da detecção de comunidades em ambientes sociais é actualmente um tópico de grande importância em áreas tão distintas como a Biologia(Jeong *et al.*, 2000), detectando unidades funcionais, ou na industria correlacionando preços de acções, ajudando prestadores de serviços a identificar grupos de interesse para o seus clientes, ou na World Wide Web classificando e agrupando websites (Brin e Page, 1998), (Kleinberg, 1999).

A noção de comunidade apresenta-se assim como uma noção difusa que pode ser definida como algo intermédio entre a rede global e a unidade fundamental, o nó. Partindo deste, normalmente nós com muitas ligações assumem um papel fundamental no fluxo de informação. Num nível mais acima surgem os chamados motivos locais, que não são mais que agregados de nós ligados de alguma forma particular. Estes motivos ao contrário dos nós com muitas ligações, podem não assumir um papel central no fluxo de informação, mas são normalmente importantes na execução de determinadas tarefas.

As comunidades são uma unidade que surge numa escala superior à dos motivos locais. São normalmente constituídas por conjuntos de nós que estão mais densamente conectados entre si do que com o resto da rede, formando grupos coesos, embora a sua definição possa variar de autor para autor. Por fim, chega-se ao nível da rede total, com suas propriedades agregadas, que embora úteis não permitem identificar partes.

Atendendo à incapacidade dos modelos de redes tradicionais para explicar a estrutura existente nas redes sociais e as comunidades em que se dividem, à procura de uma nova abordagem para este tipo de problema em redes informais e às redes oportunistas que se podem caracterizar à semelhança das redes do tipo iClouds (Heinemann *et al.*, 2003a), (Heinemann *et al.*, 2003b) ou P2P, temos como motivação encontrar mecanismos que possam ajudar a compreender este tipo de fenómenos sociais ao nível da sua estrutura.

Esta tese é motivada também pelo grande interesse na área de simulação social, nomeadamente na simulação de comunicação baseada em sistemas de agentes (Louçã *et al.*, 2007b) e sistemas complexos e de grandes dimensões. Estamos interessados na compreensão dos mecanismos que levam ao aparecimento de fenómenos de emergência a partir da dinâmica das relações existentes entre diferentes redes (Symons *et al.*, 2007) e compreender a sua dinâmica e intenções sem atender ao conteúdo semântico das suas actividades (Louçã *et al.*, 2007a).

Estes interesses levaram-nos a procurar conciliar neste trabalho os dois campos, por um lado a análise de redes de comunicação informal, por outro a simulação social e simulação multi-agente.

1.1 Objectivos

O objectivo do presente trabalho surge da dúvida relativa à estrutura da rede informal que se cria quando se está a lidar com redes sociais. Temos como objectivos comparar diversos algoritmos de detecção de comunidades em redes sociais, testando a sua adequação ao estudo deste tipo de redes e propondo um conjunto de conceitos e ferramentas que possam ser aplicadas no seu estudo, nomeadamente a utilização dos algoritmos de detecção de comunidades associados à simulação de agentes, para compreender a formação de estruturas informais de comunicação.

Utilizando o caso do correio electrónico do Instituto Superior das Ciências do Trabalho e da Empresa (ISCTE), pretendemos apresentar um modelo do comportamento da rede de comunicação informal.

Por items, os objectivos desta tese são:

- Identificar quais os mecanismos que se encontram actualmente ao dispor da comunidade científica para caracterizar redes sociais.
- Compreender a estrutura existente em redes de comunicação informal. Este ponto

será atingido através de um conjunto de objectivos operacionais respeitantes ao caso de estudo:

- Determinar quais as comunidades que compõem e utilizam regularmente o sistema de correio electrónico ISCTE.
- Identificar quais as características macroscópicas da rede de correio electrónico actual.
- Apresentar um modelo de funcionamento das redes de comunicação informal baseadas no correio electrónico e mostrar a vantagem de utilizar modelos assentes em simulação social para o caracterizar.

Através de um estudo de caso a análise dos diversos métodos mostrará qual ou quais aqueles métodos que serão mais adequados para o estudo de redes sociais que possuem inúmeros utilizadores e apresentam grande volatilidade no seu funcionamento e composição.

1.2 Contribuição científica

Este trabalho apresenta as seguintes contribuições:

- Faz um levantamento das técnicas mais usuais para a detecção de comunidades em redes sociais, abordando diferentes estratégias para o levantamento da estrutura dessas redes.
- Faz o levantamento do funcionamento do sistema de correio electrónico do corpo de estudo, identificando o problema da pouca ubiquidade presente no sistema.
- Propõe uma estratégia baseada na utilização de simulação social, juntamente com os algoritmos tradicionais de caracterização de redes e detecção de comunidades, para a análise de dados relativos a redes sociais.

1.3 Estrutura da tese

Este trabalho foi realizado de acordo com o plano de dissertação proposto no seguimento do ano curricular do Mestrado em Ciências da Complexidade e pretendeu aproveitar os recursos existentes no ISCTE, nomeadamente o seu serviço de correio electrónico. Esta disponibilidade permitiu a pesquisa na área de detecção de comunidades e de simulação social que são as áreas centrais dos nossos interesses académicos presentes.

Depois de uma preparação inicial do tema, através de pesquisa e recolha bibliográfica, foi iniciado um período de recolha de dados junto da Direcção de Serviços de Informática (DSI) do ISCTE. Esta fase englobou o desenvolvimento de software específico para a recolha e garantia de anonimato relativamente aos dados recolhidos, pois para o estudo em causa apenas estávamos interessados na estrutura da rede de comunicação e não em informações acerca de indivíduos ou nos conteúdos e trocas de informação entre membros do ISCTE. Paralelamente à recolha dos dados, foi possível proceder à implementação de alguns algoritmos para a detecção de comunidades.

Este documento está organizado por capítulos, sendo iniciado por este introdutório e seguindo-se mais quatro.

No capítulo 2 começaremos por fazer uma introdução de enquadramento ao estudo das redes sociais, seguindo-se um conjunto de noções básicas de ordem matemática necessárias para a compreensão da discussão no capítulo 4. Ainda no capítulo 2 fazemos o levantamento dos algoritmos mais utilizados na detecção de comunidades em redes e descrevemos os modelos tradicionais de redes. Neste capítulo, e consequentemente no resto do trabalho, optámos por manter a nomenclatura ligada à teoria de grafos em inglês. Tal deve-se a que muito do trabalho sobre a matéria está publicado em literatura anglo-saxónica e não encontrou ainda o seu caminho para a língua portuguesa. Optou-se assim por manter a versão original, assinalada com *itálico*. Nos termos em que já há versões portuguesas mas onde estas ainda não fazem parte do uso diário da comunidade científica faz-se normalmente referência às várias alternativas na primeira vez que o termo surge.

Seguidamente, no capítulo 3 colocamos as questões principais às quais este trabalho procura responder, apresentando a hipótese de que a detecção de comunidades através da análise topológica, sem conhecimento semântico da rede, permite caracterizar comunidades e hierarquias dentro dos processos de comunicação informal do ISCTE e fazemos uma breve discussão das suas implicações.

O capítulo 4 é dedicado a explicar o caso de estudo que foi seleccionado, começando por fazer a sua caracterização inicial e expondo os problemas que foi necessário resolver para assegurar a qualidade dos dados fornecidos pelos serviços do ISCTE. É apresentado o resultado da análise feita na detecção de comunidades do sistema de correio electrónico do ISCTE e de seguida o modelo de funcionamento do sistema de correio electrónico, mostrando como a utilização de simulação social pode ser aplicada à análise de redes sociais.

No capítulo 5 apresentamos a discussão dos resultados obtidos, fazemos a validação das hipóteses apresentadas no capítulo 3 e ainda levantamos algumas questões de investigação

que aqui não foram respondidas, apontando caminhos futuros para o estudo das redes de comunicação informal.

Capítulo 2

Estado da Arte

> *"The strange mind-game that clatters in me all the times goes like this: how can I think, with three, four, or at most five links of the chain, trivial, everyday things of life. How can I link one phenomenon to another? How can I join the relative and the ephemeral with steady, permanent things – how can I tie up the part with the whole?"*
>
> Frigyes Karinthy *in Chain-Links (Karinthy, 1929)*

2.1 Introdução ao estado da arte

O estudo de redes, como a Internet ou os sistemas sociais ou biológicos, tem vindo a ser aprofundado de forma intensa nos anos mais recentes. Desde o campo da física à ciência de computadores, investigadores têm encontrado sistemas que podem ser representados por grafos, dos quais se pode obter informação útil sobre o sistema em estudo.

O estado da arte que se segue começa por fazer uma breve introdução à história do estudo de redes na secção 2.2. Nela referiremos brevemente alguns momentos importantes para a história da disciplina, referindo os principais desenvolvimentos e contribuições na área. De seguida, na secção 2.3 abordamos algumas noções básicas para o estudo de redes. Esta secção está principalmente focada no tratamento matemático de grafos e apresenta as noções básicas necessárias para a discussão subsequente deste tema. Naturalmente, tratando-se de uma área vasta, não se condensou nesta secção toda a teoria de grafos. Para um estudo mais completo da teoria de grafos aconselha-se a leitura de Diestel (2005), uma referência fundamental nesta área. Na secção 2.4 expomos algumas medidas utilizadas para a caracterização de redes. Estas estão divididas em duas sub-secções: uma

dedicada às medidas de centralidade e outra às medidas de densidade. A secção 2.5 descreve os modelos clássicos utilizados normalmente para a caracterização das redes sociais. Seguidamente, na secção 2.6 discutimos as diversas técnicas e algoritmos disponíveis actualmente para a identificação de comunidades em redes sociais. Na secção 2.7 apresentamos alguns trabalhos realizados na detecção de comunidades em redes de email, introduzindo assim o que tem vindo a ser feito na área do nosso caso de estudo.

2.2 O estudo de redes sociais

O estudo de redes tem uma longa tradição na matemática. Um dos primeiros problemas matemáticos que foi preciso resolver recorrendo à "teoria dos grafos" é o conhecido problema das pontes de Konignsberg (Euler, 1741).

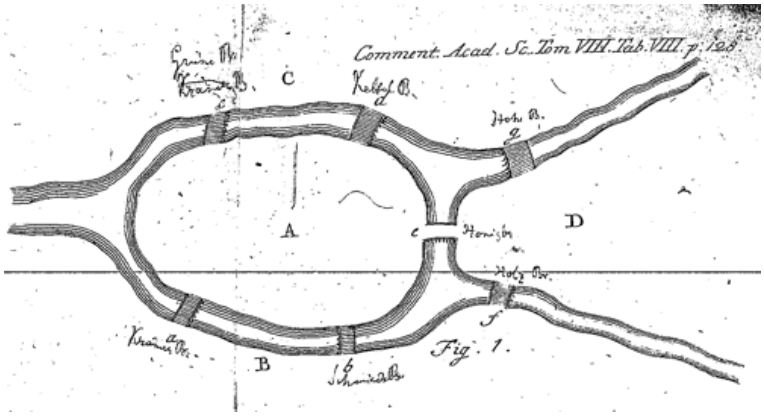

Figura 2.1: Diagrama das Pontes de Konigsberg no documento original (Euler, 1741)

A cidade de Konigsberg (hoje Kalinegrado, na Russia) está situada nas margens do rio Pregel. No meio do rio existem duas ilhas. Sete pontes unem estas quatro massas de terra. O problema que se punha era o de saber se seria possível percorrer todas as sete pontes num caminho que não repetisse pontes. A lenda conta então que a população da cidade passava inúmeras horas à procura da solução. Euler (1741) resolveu a questão de uma vez por todas, provando a inexistência de uma solução. A prova utiliza a noção de grafo que abstrai os pormenores físicos do problema e mapeia a ilha a um conjunto de nós e vértices. Desta forma, Euler prova que não é possível um tal caminho através das pontes, uma vez que qualquer caminho que percorra um grafo passando uma única vez em cada ligação apenas aceita no máximo 2 vértices com um número impar de ligações (por se tratar do vértice de início ou fim; todos os restantes terão de ter sempre um número par de ligações: uma entrada e uma saída). Ora o desenho da rede das 7 pontes (ligações)

e 4 massas de terra (vértices) mostra que os 4 vértices possuem todos um número impar de ligações pelo que tal caminho, que atravesse todas as ligações sem repetição, não é possível (Newman *et al.*, 2006).

Nos anos 50 o interesse na aplicação de técnicas quantitativas ao domínio da sociologia e antropologia levaram a que a linguagem matemática da Teoria dos Grafos fosse adoptada pelos cientistas sociais para ajudar na compreensão dos dados obtidos em estudos etnográficos (Newman *et al.*, 2006). Muita da terminologia utilizada em ciências sociais foi então adoptada directamente da "Teoria de Grafos".

2.2.1 A descoberta dos *random graphs*

O interesse científico no estudo de redes cresceu no final dos anos 40, início de 50. Durante esse período um dos maiores pensadores do campo foi Anatol Rapoport, que trabalhou principalmente na área da matemática e biologia numa altura em que estas duas disciplinas estavam muito desligadas (Newman *et al.*, 2006). Os seus trabalhos foram de particular importância para a compreensão das redes, tendo desenvolvido os métodos de análise das propriedades agregadas das redes, em vez de olhar para as propriedades particulares de cada um dos nós ou ligações. Foi também de sua autoria, em parceria com Ray Solomonoff, o primeiro artigo (Solomonoff e Rapoport, 1951) que aborda o estudo sistemático daquilo que hoje é conhecido por redes aleatórias. Neste artigo ao autores demonstram uma das propriedades mais importantes destas redes, que é a observação de que, quando o rácio entre ligações e nós aumenta, a dada altura a rede sofre uma transformação brusca do seu estado de agregação, passando de um conjunto de nós desligados, ou agrupados em pequenos grupos, para um estado em que surge um componente gigante conectado, embora possam ainda existir vários componentes pequenos dispersos (Watts, 2003, pp. 45-46).

2.2.2 Os pais da teoria de grafos

Seguidamente aos estudos de Solomonoff e Rapoport (1951), Erdós e Rényi (1960) também se interessaram pela "Teoria dos Grafos" e publicaram, no princípio de 1960, alguns artigos sobre o assunto, numa altura em que a comunidade científica se voltava com mais interesse para o estudo de redes. Erdós e Rényi (1960) publicaram um conjunto de artigos onde caracterizam as redes aleatórias. Um desses artigos (Erdós e Rényi, 1960) mostra a evolução da estrutura da rede à medida que o valor médio do número de ligações aumenta. Os autores mostram que muitas das propriedades dos grafos aleatórios emergem

subitamente e não gradualmente, à medida que mais ligações vão sendo adicionadas de forma aleatória ao grafo existente. Os autores provaram este surgimento abrupto de propriedades utilizando a seguinte definição: se a probabilidade de um grafo apresentar uma propriedade Q que tende para 1 quando o tamanho (número de nós) tende para infinito ($N \rightarrow \infty$), então quase todos os grafos de dimensão N têm a propriedade Q. Depois de estudarem o comportamento de diversas propriedades como uma função da probabilidade p de existência de uma ligação entre quaisquer dois vértices, mostraram que para muitas propriedades existe uma probabilidade crítica $p_c(N)$ tal que se esta probabilidade crescer mais devagar que $p(N)$ quando $N \rightarrow \infty$, então o grafo não apresenta a propriedade. Pelo contrário, o grafo apresenta a propriedade quando a probabilidade crítica cresce mais depressa que $p(N)$. Esta descoberta permitiu definir uma série de expoentes críticos para o limiar onde algumas propriedades dos grafos existem, nomeadamente a existência de ligações, ciclos ou cliques. A caracterização de redes assumiu a partir deste trabalho uma perspectiva estocástica em vez de puramente determinista, como até então.

2.2.3 Percolação

Uma área próxima do estudo das redes aleatórias é a que diz respeito à Teoria da Percolação, que foi objecto de estudo durante muito tempo por parte da física e da química (Watts, 2003, pp. 183-187). O modelo original de percolação foi introduzido por Broadbent e Hammersley (1957) e Hammersley (1957). No modelo de percolação das ligações procura-se estudar as propriedades do sistema de tal forma que as ligações numa grelha regular, ou genericamente numa rede, são ocupadas ou não, conforme uma probabilidade de ocupação p. A uma dada probabilidade crítica p_c dá-se uma transição de fase, rápida e aparentemente expontânea, passando o sistema, que até então se encontrava num estado desagregado, a estar num estado em que que se forma um componente gigante de permeação, na terminologia da área um *percolation cluster*. Para além da química e da física, a Teoria de Percolação é utilizada no estudo de redes para determinar por exemplo, os processos e difusão de epidemias (Newman *et al.*, 2001). Callaway *et al.* (2000) estudaram a robustez e a fragilidade de redes através de processos de percolação, a fim de detectar a sua resiliência em caso de ataques dirigidos ou aleatórios.

2.2.4 Questões importantes do estudo de grafos

No final dos anos 50, na mesma altura em que Erdós e Rényi começaram os seus trabalhos sobre grafos aleatórios, a comunidade de investigadores da área da sociologia começou a

mostrar-se interessada pelas aplicações da Teoria de Grafos. De Sola Pool e Kochen (1978) escreveram em 1958 um artigo que só viria a ser publicado em 1978, mas que circulou como pré-publicação pela comunidade académica (Newman *et al.*, 2006). Nele, os autores abordam pela primeira vez questões que o campo da sociologia abordaria nos anos que se seguiriam. O artigo (de Sola Pool e Kochen, 1978) seria publicado no número 1 da revista Social Networks e terá fornecido, entre outros, a inspiração para a experiência de Travers e Milgram (1969) sobre o efeito "small world". Entre as questões levantadas pelos autores na introdução do artigo, destacam-se:

- Quantos indivíduos conhece cada um dos membros de uma rede? Ou em termos da teoria de grafos, qual o grau (*degree*) de cada pessoa na rede?
- Qual é a distribuição desses conhecimentos ou grau? Qual o valor médio e quais os valores maiores e mais pequenos?
- Que tipos de indivíduos tem o maior número de contactos? São estes os indivíduos mais influentes na rede social?
- Como é que os contactos se organizam? Qual é efectivamente a estrutura da rede?
- Qual é a probabilidade de que dois indivíduos escolhidos ao acaso sejam conhecidos?
- Qual a probabilidade de terem um amigo comum?
- Qual é a probabilidade de o caminho mais curto que os une necessitar de dois intermediários? Ou mais que dois?

De Sola Pool e Kochen (1978) começam por discutir no artigo a dificuldade colocada aos cientistas sociais na determinação do número de contactos sociais que cada indivíduo tem. Apontam dois problemas, que têm a ver com a ambiguidade do que exactamente constitui um contacto social e com o facto de as pessoas não serem boas a estimar o número de contactos que possuem. Assim, os autores recorreram aos modelos de grafos de Solomonoff e Rapoport (1951) e baseiam os seus estudos em redes aleatórias. Assumindo que cada pessoa teria em média cerca de 1000 contactos previram que quaisquer duas pessoas na terra estariam conectadas através de um caminho geodésico com apenas mais 2 conhecidos (distância geodésica 3). Esta foi a primeira vez que o tema "small world" foi tratado de forma científica.

2.2.5 Os seis passos de separação de Milgram

Embora a ideia de utilização de redes começasse a ser frequente nas ciências sociais, só quando o experimentalista Stanley Milgram efectuou a famosa experiência "small world" é que este campo de estudo se tornou conhecido. O trabalho de de Sola Pool e Kochen (1978) inspirou Milgram a idealizar uma experiência que serviria para explorar as redes de tipo "small world". Contudo, os resultados das primeiras experiências foram publicadas sem um caracter científico. As experiênias foram posteriormente repetidas, associando-se Milgram a Jeffrey Travers, um matemático, por forma a obter um tratamento mais rigoroso da experiência. Em 1969 é então publicado o artigo (Travers e Milgram, 1969) que contém os dados da repetição da experiência, que pode ser descrita sucintamente da seguinte forma:

Um determinado número de indivíduos foi seleccionado aleatoriamente e foi definido um indivíduo alvo. Um embrulho de correio foi enviado a cada um dos indivíduos, contendo um livro de registos onde era pedido que os participantes introduzissem alguns dados. O objectivo era que os indivíduos enviassem este correio a alguém seu conhecido (do seu círculo próximo) e que eles achassem que pudesse conhecer o indivíduo alvo, ou então que pudesse por sua vez conhecer alguém adequado para tal tarefa. A definição de 'conhecido' era a de alguém cujo tratamento fosse feito na base do primeiro nome, uma vez que na cultura anglo-saxónica tal tratamento é normalmente restrito a pessoas bastante próximas. O correio seria assim enviado e o processo repetido até que este chegasse ao alvo ou então se perdesse. A cada passo era também pedido que, para além da inscrição do seu nome no livro de registos, fosse enviado um postal de correio aos autores da experiência, a fim de poderem acompanhar o percurso dos embrulhos, mesmo os que não chegaram ao destino. Foram enviados embrulhos a 296 indivíduos e destes 296 embrulhos, 64 chegaram ao destino, sendo que o número de intermediários se situou entre 1 e 11, com uma mediana de 5.2. 5 intermediários significa uma distância geodésica[1] de 6. Esta experiência ficou conhecida como *"os seis passos de separação de Milgram"*.

A validade da distância geodésica média obtida foi discutida pelos próprios no artigo, uma vez que o trabalho de de Sola Pool e Kochen (1978) apontava para valores mais baixos (3). No entanto, White (1970) determinou uma correlação existente nos dados que aumentou a separação para 8. Contudo há outros factores que podem forçar o valor a descer. Por exemplo, não há a garantia que o caminho percorrido pelos embrulhos seja efectivamente o caminho mais curto entre os indivíduos. É apenas o caminho que eles acham ser o mais curto, podendo haver outro que seja melhor, pelo que a separação real pode ser muito

[1] número de ligações mínimas para unir dois nós ou vértices, ver adiante ponto 2.3.8

menor do que o estudo sugere (Newman *et al.*, 2006).

2.2.6 As redes *scale-free* de Price

O trabalho de Price (1965) surge do campo da ciência da informação. Trata-se do primeiro a estudar redes de grande dimensão, nomeadamente redes de citações científicas onde cada vértice representa um artigo publicado e onde cada ligação (direccionada) representa uma citação no primeiro vértice ao artigo citado (segundo vértice). Foi um dos primeiros a sugerir que se olhasse para o mundo das citações como uma rede, tendo apresentado uma análise completa de uma rede de citações científicas (Newman *et al.*, 2006).

Como a rede de citações científicas é efectivamente uma rede dirigida (uma rede em que as ligações entre dois nós tem direcção de um nó para outro e portanto não é simétrica), Price estudou o *degree* nas suas versões *Outdegree* e *Indegree* e verificou que em ambos os casos a rede de citações obedece a uma distribuição de lei de potência com índices -3 e -2, respectivamente. Esta distribuição na forma de uma lei de potência é normal nas posteriormente chamadas redes do tipo "scale-free". Para além deste trabalho inicial de análise, Price propôs em 1976, num artigo intitulado *A general theory of bibliometric and other cumulative advantage processes* (Price, 1976), um modelo para a explicação da formação destas redes. O modelo diz que os artigos que têm mais citações recebem novas citações na proporção das citações que já possuem. Chamou a este processo "vantagem cumulativa" e o modelo que propôs efectivamente é gerador de distribuições de lei de potência. Este processo é hoje mais conhecido pela designação de "ligação preferencial", ocorrendo em muitos fenómenos sociais (Newman *et al.*, 2006, ver sec. 4.3).

Os trabalhos mais recentes levaram a que a Teoria de Grafos fosse cada vez mais aplicada a problemas reais. A nova ciência que cresceu em torno do estudo de redes aponta para 3 caminhos: foca-se nas propriedades de redes reais, estando preocupada tanto com questões empíricas como teóricas; assume que as redes não são estáticas preocupando-se cada vez mais com o seu aspecto dinâmico; procura perceber as redes, não apenas como objectos topológicos, mas num enquadramento operacional sobre o qual os sistemas dinâmicos distribuídos são construídos (Newman *et al.*, 2006).

De seguida fazemos uma revisão sucinta de um conjunto de ferramentas necessárias para proceder ao estudo de redes através da Teoria de Grafos. Este campo hoje em dia é vasto e muito do formalismo matemático necessário não está aqui tratado, uma vez que pretendemos no ponto seguinte abordar o minimamente necessário para compreender o trabalho de investigação desenvolvido. Para um conhecimento mais aprofundado do

formalismo destas matérias sugerimos a consulta do livro "Graph Theory" de Diestel (2005).

2.3 Noções básicas de redes

O estud de redes requer uma representação que seja tratável matematicamente. A Teoria de Grafos descreve formas de representar um gráfico sob a forma matricial, para que o seu tratamento matemático seja facilitado.

Com o passar do tempo, vários tipos de representação matricial foram sendo apresentados, de acordo com a necessidades e objectivos em estudo. A seguir descrevem-se algumas representações úteis e usuais em Teoria de Grafos, assim como alguns conceitos que são necessários para o posterior tratamento que faremos neste trabalho.

2.3.1 Grafo

A forma mais prática de representar uma rede é através de um grafo. Um grafo é uma representação de uma rede através de linhas e pontos de intersecção, ou formalmente um grupo de objectos (ou nós) e um mapeamento das relações existentes entre os diversos objectos (Diestel, 2005).

$$\text{Grafo} = \text{Grupo} < \text{Objectos}, \text{Relações} > \tag{2.1}$$

2.3.2 Grafo de linha

É importante introduzirmos a noção de grafo de linha de um grafo não direccionado, porque a sua noção vai ser necessária para pontos seguintes. O grafo de linha de um grafo G é um grafo onde as ligações entre vértices de G são consideradas elas próprias como vértices neste grafo e onde, por outro lado, cada vértice de G é considerado no grafo de linha como uma ligação. Na prática, corresponde a uma inversão de papeis dos nós e ligações do grafo original. No grafo de linha os nós passam a ser tratados como ligações e as ligações como nós.

2.3.3 Degree

O *degree*, grau ou valência de um nó ou vértice de um grafo é dado pelo número de ligações a outros nós que esse nó possui. O *degree* é o número de vizinhos de um determinado nó. O *degree* pode ser ainda dividido em *outdegree* e *indegree*, caso estejamos perante um grafo em que as relações entre nós não sejam equivalentes nos dois sentidos. Mais à frente, na secção sobre medidas de rede, voltaremos com mais detalhe a esta definição.

2.3.4 Matriz de adjacência

A matriz de adjacência é uma matriz quadrada $n \times n$ onde cada elemento $\{1, ..., n\}$ corresponde a um nó da rede. Esta matriz permite definir as ligações existentes entre os diversos nós de uma rede, da seguinte forma:

$$A : N \times N \rightarrow \{0, 1\} \tag{2.2}$$

$$A_{i,j} = \begin{cases} 1 & \text{se existir ligação entre } i \text{ e } j \text{ e } i \neq j \\ 0 & \text{caso contrário} \end{cases} \tag{2.3}$$

Tal como é apresentada na formulação anterior, a matriz de adjacência considera que todas as ligações da rede são equivalentes, não fazendo distinção entre elas. Para além disso não são permitidas auto-ligações (situações em que um nó estabelece uma ligação consigo próprio) (Diestel, 2005).

No caso de estarmos perante uma rede de ligações não direccionadas, então $A_{i,j} = A_{j,i}$ e a matriz de adjacência é simétrica. No caso de se tratar de um gráfico direccionado, então pode acontecer que $A_{i,j} \neq A_{j,i}$ e aí interpreta-se $A_{i,j}$ como a existência de uma ligação de i para j.

2.3.5 Matriz de similaridade ou de distâncias

A matriz de similaridade é uma matriz em tudo semelhante à matriz de adjacência, com a diferença de que as ligações entre os diversos nós não têm peso unitário mas antes apresentam valores reais que indicam o peso da ligação existente entre esses dois nós. Os grafos de similaridade são construídos de forma semelhante, mas considerando que a existência de uma ligação entre dois nós implica a existência de um valor mínimo de

similaridade entre os dois. Este valor pode ser 0 ou então pode ser definido um limite mínimo abaixo do qual se considera não existirem ligações entre os nós.

A construção da matriz e grafo recorre ao conceito de similaridade entre nós. Esta é definida de tal forma que, dados um conjunto de pontos $x_1, ..., x_n$ e uma qualquer noção de similaridade $s_{ij} \geq 0$ entre todos os pontos x_i e x_j, possa ser representado um grafo de similaridade $G = (V, E)$. Neste grafo cada vértice v_i corresponde a um ponto x_i e dois vértices estarão conectados se a sua similaridade s_{ij} for superior a um determinado valor limite e a sua ligação terá o peso dado por s_{ij} (von Luxburg, 2007).

Há diversas construções usuais para os grafos de similaridade perante um dado grupo de pontos $x_1, ..., x_n$ e suas similaridades (ou distâncias). Ao construir estes grafos é importante ter em consideração as relações locais entre os pontos a representar. Atendendo à definição de similaridade, é preciso notar que a distância entre dois pontos é tanto menor quanto maior for a similaridade, devendo encontrar-se uma forma de transformar uma na outra. A construção dos grafos de similaridade ou de distâncias é na prática igual (von Luxburg, 2007).

2.3.5.1 Grafos de ε-vizinhança

Na construção dos grafos de ε-vizinhança pretende-se conectar todos os pontos cujas distâncias sejam inferiores a ε. Como nestes estes grafos as distâncias apresentam sensivelmente a mesma escala (no máximo ε), são normalmente considerados grafos com ligações não ponderadas, pois a atribuição de peso às ligações não acrescenta mais informação acerca dos pontos do grafo (von Luxburg, 2007).

2.3.5.2 Grafos de k-vizinhança

Neste caso o objectivo é ligar cada um dos vértices do grafo de tal forma que a ligação de um vértice v_i a outro v_j só é possível se v_j for um vizinho de v_i e esteja nos k-vizinhos mais próximos (von Luxburg, 2007).

2.3.6 Matriz de incidência

O caso da matriz de adjacência permite representar matricialmente as relações entre elementos de um mesmo grupo (os nós ou vértices da rede, p.ex.). Porém, podemos estar interessados em estabelecer um tratamento matemático das relações entre elementos de

diferentes grupos. Para isso podemos utilizar a denominada matriz de incidência. No caso típico, os dois grupos que estamos interessados em relacionar são os nós e as ligações entre nós.

Considerando $M = \{\text{nós}\}$ e $N = \{\text{ligações}\}$, a matriz de incidência é:

$$C : M \times N \rightarrow \{0, 1\} \tag{2.4}$$

$$C_{i,j} = \begin{cases} 1 & \text{se existir uma relação entre o nó } i \text{ e a ligação } j \\ 0 & \text{caso contrário} \end{cases} \tag{2.5}$$

A matriz de incidência pode ser correlacionada com a matriz de adjacência através de:

$$A(L(G)) = C^T C - 2I \tag{2.6}$$

onde I é a matriz identidade e $A(L(G))$ é a matriz adjacente do grafo de linha L do grafo G.

2.3.7 Matriz de Laplace

Uma outra forma de representação matricial útil para a análise matemática de grafos é a chamada representação Laplaciana. A matriz de Laplace é definida para um grupo de nós M tal que:

$$L : M \times M \rightarrow \mathbb{N} \tag{2.7}$$

$$L_{i,j} = \begin{cases} \deg(i) & i = j \\ \begin{cases} -1 & \text{i e j relacionados} \\ 0 & \text{caso contrário} \end{cases} & i \neq j \end{cases} \tag{2.8}$$

Esta matriz coloca na diagonal principal o valor de *degree*. Caso se esteja a trabalhar com grafos direccionados ter-se-á que definir duas matrizes Laplacianas, uma para o *outdegree* e outra para *indegree*.

2.3.8 Caminho geodésico e distância geodésica

A noção de caminho geodésico é importante na medida em que vai ser muito utilizada na determinação de outras medidas referidas adiante neste documento.

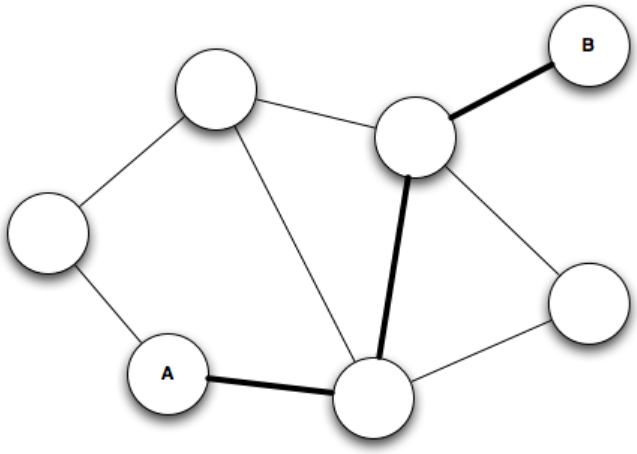

Figura 2.2: Caminho geodésico AB com distância geodésica 3.

O *caminho geodésico*, do inglês *Geodesic Path*, ou *Distância mais curta*, é o percurso definido pelo número mínimo de passos que é preciso dar para ir de um nó até outro. Na figura 2.2 o caminho geodésico entre os nós A e B, que está assinalado através de ligações a cheio, tem a distância geodésica 3. A distância geodésica representa o número mínimo de ligações existentes entre dois nós na rede. Para além da noção de caminho e de distância, sobre todas as distâncias é também comum definir a média, obtendo-se assim a chamada distância geodésica média, em inglês a *average path length*.

2.3.9 *k-core*

Um *k-core* é o maior subgrafo de uma rede onde os vértices possuem pelo menos k ligações. Uma forma de obter um determinado *k-core* é através da remoção iterativa de todos os nós de uma rede que possuem um *degree* inferior a k. Isto significa que um subgrafo *3-core* não possui nenhum nó com degree 1 ou 2. A decomposição de uma rede em *k-cores* permite descrever a rede numa forma hierárquica, sendo esta constituída por uma série de *k-cores* encapsulados, à semelhança de uma boneca russa (Dorogovtsev *et al.*, 2005).

2.3.10 Clique

Um *clique* num grafo é um subgrafo tal que todos os nós desse subgrafo estão ligados uns aos outros dentro do subgrafo. Em termos matemáticos um *clique* pode ser definido por um grupo de vértices V, tal que para todos os pares de vértices v_i, v_j, existe uma ligação entre v_i e v_j.

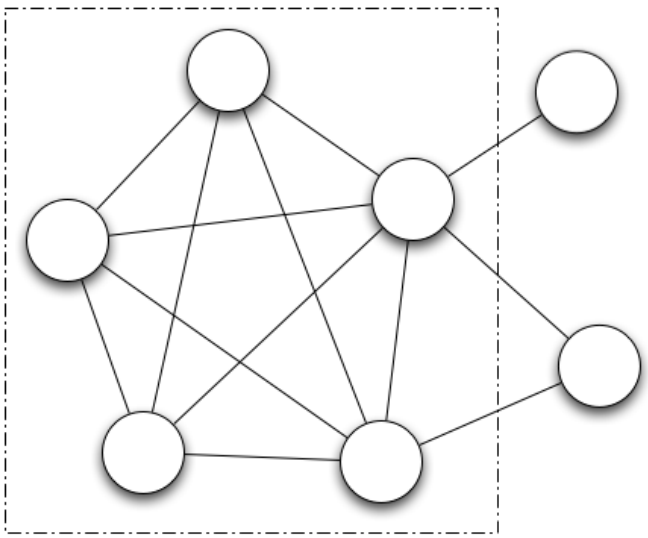

Figura 2.3: Clique com 5 vértices.

2.4 Caracterização de redes

A caracterização de determinada rede pode ser feita segundo duas perspectivas diferentes: focando o interesse na posição de determinados vértices e no papel que assumem dentro da estrutura global da rede, ou olhando para a rede de forma global e procurando encontrar medidas agregadas que sumariem alguma propriedade da rede em questão. Na primeira situação estamos perante medidas de centralidade (ver sec. 2.4.1), enquanto na segunda estamos perante mediadas de densidade (ver sec. 2.4.2).

2.4.1 Medidas de centralidade

Existem diversas medidas pelas quais se pode caracterizar uma rede. Uma das famílias de medidas mais importantes são as de centralidade. Estas medem de uma forma geral a importância relativa de cada um dos nós dentro da estrutura da rede.

2.4.1.1 *Degree* (grau)

Esta medida reflecte simplesmente o número de ligações existentes em cada nó da rede. Caso se trate de uma rede direccionada o *degree* é o número total de ligações ou o número de vizinhos imediatos. O *degree* pode ainda ser dividido em duas submedidas, o *Outdegree* e o *Indegree*, conforme se considerem as ligações "estabelecidas" do nó em causa para outros nós, ou se considere as ligações "recebidas"por esse mesmo nó.

A noção de *degree* pode ser traduzida facilmente em termos da matriz de adjacência (ver sec. 2.3.4)

$$\deg_i = \begin{cases} \sum_{j=1}^{n} A_{i,j} & \text{matriz não direccionada e } \mathit{OutDegree} \\ \sum_{j=1}^{n} A_{j,i} & \mathit{InDegree} \end{cases} \tag{2.9}$$

2.4.1.2 *Betweenness*

A *betweenness* é uma medida que dá a importância de um determinado nó em termos do controlo que exerce no fluxo de informação através da rede. Esta ideia é talvez melhor explicada se considerarmos que a *betweenness* representa a fracção de caminhos geodésicos [sec. 2.3.8] que atravessam um determinado nó. Em redes onde há fluxo de informação entre diversos nós, a *betweenness* reflecte o volume de informação que passa através de

cada vértice. É uma medida da influência que esse nó exerce na transmissão de informação na rede.

$$\text{Betweeness}_i = \frac{\sum GP(i)}{\sum GP_n} \tag{2.10}$$

Betweenness foi introduzida em simultâneo por Anthonisse (1971) e Freeman (1977) e dá conta da importância que um nó desempenha como intermediário. Enquanto alguns índices de centralidade requerem condições *à priori*, tais como a não direccionalidade do grafo ou o facto de ser um grafo conectado, Anthonisse (1971) define *betweenness* em grafos direccionados e Freeman (1977) aborda a aplicabilidade a grafos não dirigidos. Por outro lado, Brandes (Brandes, 2008) faz um levantamento de variantes do cálculo da *betweenness* quando generalizado a outros tipos de dados de rede e faz também o levantamento de algoritmos necessários para a computação da *betweenness* de forma eficiente.

2.4.1.3 *Closenness*

A medida de centralidade *closenness*, por outro lado, dá uma medida da proximidade de um determinado nó aos restantes nós. É definida em termos da distância média que o nó em questão está de todos os outros.

$$\text{Closenness}_i = \frac{\sum_{j=1\, j\neq i}^{n} GP(i,j)}{N-1} \tag{2.11}$$

Na equação anterior considera-se que $GP(i,j)$ é o caminho geodésico de i para j

2.4.1.4 *Eigenvector*

A determinação de centralidade de cada nó através dos vectores próprios é uma forma elaborada de centralidade que permite quantificar a importância de cada um nó em função não só do número de ligações que possui (*in* ou *out*), mas também da importância que aqueles que estão perto de si têm. Ou seja, estabelecer ligações a nós importantes fortalece a importância do nó que se liga.

$$x_i = \frac{1}{\lambda} \sum A_{i,j} x_j, \;\; Ax = \lambda x \tag{2.12}$$

2.4.1.5 *PageRank*

O algoritmo PageRank surgiu do trabalho de Brin e Page (1998) na universidade de Stanford, relativo à classificação e indexação de páginas web no motor de busca Google, sendo uma marca registada da empresa *Google, Inc.*

O algoritmo faz a análise das ligações existentes entre diversos nós (no caso páginas web) e atribui um peso numérico a cada nó conforme o número de ligações oriundas de outros nós. Isto faria do *PageRank* algo semelhante ao *inDegree*. No entanto, a originalidade dos autores passou por normalizar o número de links e não contabilizar todas as ligações recebidas com o mesmo peso (Brin e Page, 1998). O cálculo do *PageRank* PR é dado pela equação 2.13, onde d é um factor de moderação, usualmente 0.85 e $C(A)$ é o número de ligações que saem do nó A.

$$PR(A) = (1-d) + d(\frac{PR(T1)}{C(T1)} + ... + \frac{PR(Tn)}{C(Tn)}) \tag{2.13}$$

Na prática o algoritmo de *PageRank* é calculado iterativamente e corresponde ao vector próprio principal da matriz de adjacência normalizada (Brin e Page, 1998).

2.4.2 Medidas de densidade

2.4.2.1 Coeficiente de *clustering*

O coeficiente de *clustering* pode ser definido tendo em atenção duas situações distintas ao nível da escala utilizada para analisar a rede. Se a análise é feita sob uma perspectiva local, corresponde à chamada transitividade local, senão define-se como propriedade macroscópica e apresenta um valor agregado característico da rede em estudo.

Localmente o coeficiente de *clustering* quantifica o quão próximo um vértice de um grafo está perto de formar um clique com os seus vizinhos imediatos. A medida foi proposta por Strogatz e Watts (1998) e é uma das formas de verificar se uma rede pode ser classificada como *small-world* (ver adiante a secção sobre redes de tipo *small world* para descrição detalhada). A introdução da designação *Small-World Networks* por parte dos autores derivou do fenómeno *small-world*, popularmente conhecido como "seis graus de separação".

A medição do *clustering* global da rede C, por seu lado, depende dos valores de *clustering* de cada um dos seus nós C_v, e pode ser dada pela equação 2.14 que representa o rácio entre

o número de ligações existentes entre os vizinhos que estão a uma distância geodésica 1 e o número máximo de ligações possível $\frac{1}{2}k_v(k_v-1)$.

$$C =< C_v >_v = \Big\langle \frac{2E_v}{k_v(k_v-1)} \Big\rangle_v \tag{2.14}$$

Paralelamente a esta definição de *clustering*, uma variação do coeficiente de *clustering* é dada pelo rácio de "triplos" completamente conectados. Um "triplo" é um clique composto por exactamente 3 nós. Neste caso o coeficiente de *clustering* é dado por:

$$C_\Delta = \frac{3 \times \text{número de triângulos de um grafo}}{\text{número de triplos ligados a vértices}} \tag{2.15}$$

O cálculo do *clustering* dado pelas equações 2.14 e 2.15 não é equivalente, pois os limites físicos da amostra devem ser considerados (Ebel *et al.*, 2002).

2.5 Modelos de redes sociais

A compreensão das redes, nomeadamente as de cariz social, implica o estudo da sua estrutura do ponto de vista da análise da teoria de grafos, mas também a criação de modelos explicativos dessa mesma estrutura. Para além dos resultados estatísticos, é conveniente estudar como a dinâmica social gera estruturas iguais às observadas.

Os modelos teóricos da estrutura de redes podem ser divididos em quatro classes essenciais. A primeira engloba as redes do tipo "malha regular", onde todos os vértices fazem parte de uma matriz ou grelha e apresentam o mesmo *degree*, isto é, a estrutura da rede se repete.

Os modelos mais antigos são, naturalmente, os modelos de redes aleatórias, onde os trabalhos de Solomonoff e Rapoport (1951) foram pioneiros, seguindo-se os trabalhos de Erdós e Rényi (1960) onde um grafo aleatório $G_{n,p}$ consiste em n vértices e p denota a probabilidade de existir uma ligação entre dois pares de vértices. A vantagem das redes aleatórias é de permir tratamento analítico de algumas das suas propriedades. Um caso especial de redes aleatórias são as redes aleatórias geométricas, que são geradas colocando aleatoriamente N vértices num quadrado unitário e depois conectando os pares de vértices que se situem a uma distância inferior a um parâmetro de controlo (Dall e Christensen, 2002).

Mais tarde, surgiram os trabalhos sobre o fenómeno "*small world*". Os modelos geradores

para este tipo de rede são apresentados por Strogatz e Watts (1998), procurando reproduzir as propriedades de redes reais onde se verifica um excesso de *clustering*, em relação a um grafo aleatório e uma distância geodésica média baixa. Estas redes do tipo "*small world*" são normalmente geradas através da ligação entre vértices de uma rede regular com uma determinada probabilidade p. Este processo resulta então em redes que exibem propriedades do tipo "*small world*".

Mais recentemente surgiram os modelos que descrevem as redes do tipo "*scale-free*" introduzidos por Barabasi e Albert (1999). Os modelos para a construção destas redes são motivados pelas medições empíricas das distribuições de *degreee* observadas na Internet e na *World Wide Web*, que se verifica obedecerem a uma lei de potência. A construção do modelo para este tipo de redes é baseado na forma como se pensa que estas redes são estabelecidas no mundo real. A cada iteração do processo de construção há o crescimento do número total de nós da rede, por adição sequencial de um nó à rede e há o estabelecimento de uma ligação desse nó a um outro já existente segundo uma probabilidade variável, de acordo com o número de ligações que cada um dos outros nós já possui. O modelo é assim descrito em termos de *a*) crescimento e *b*) ligação preferencial (desJardins *et al.*, 2008).

Mais recentemente tem-se discutido as diferenças que as redes sociais apresentam para estes modelos de formação de redes. Newman e Park (2003) afirmam que as redes sociais são fundamentalmente diferentes de outras redes, focando-se em duas propriedades. Por um lado analisam a correlação de *degree*, observando que os *degrees* de nós adjacentes na rede estão positivamente correlacionados nas redes sociais e por outro lado, analisam a existência de transitividade elevada (*clustering*), *i.e.*, a propensão para que pares de nós estejam ligados entre si se possuírem um vizinho comum.

Hamill e Gilbert (2008) discutem a validade dos quatro modelos tradicionais de redes (grelhas regulares, aleatórias, *small-world* e *scale-free*) no estudo das redes sociais. Segundo estes autores, nenhum dos modelos é apropriado para aplicar a redes sociais porque estas tendem a conter poucas pessoas muito conectadas, como no caso das redes de tipo *scale-free*, mas não nos modelos de tipo *small-world*, apresentando por outro lado *clustering* elevado típico das redes *small-world* mas não das *scale-free*. As redes de tipo regular naturalmente não se aplicam, pois é evidente que tipicamente nas redes sociais os nós não apresentam todos as mesmas características. As redes aleatórias não podem naturalmente servir de modelos para as redes sociais uma vez que os contactos não são estabelecidos aleatoriamente em relação ao total da população mas antes são restritos por limitações de similaridade e geografia (Hamill e Gilbert, 2008).

O estudo de redes sociais pode ser abordado sob diferentes perspectivas, quer ao nível do posicionamento do observador, quer na globalidade da perspectiva adoptada para a análise da rede em estudo.

Peter Mardsen (Carrington *et al.*, 2005, cap. 2) aborda diferentes trabalhos realizados recentemente, nomeadamente abordando os estudos egocêntricos, os estudos mono-modais (com relações simples ou múltiplas), os estudos bimodais e os estudos de estrutura social cognitiva (CSS em inglês, de *cognitive social structure*). O autor acrescenta ainda referências para dois tipos de estudo que estão a ser recentemente efectuados: o estudo de redes por amostragem e o estudo por caminhadas aleatória (em inglês *random walk*).

Os limites do desenho de redes podem ser definidos de acordo com três princípios: baseados na posição dos actores, baseados em eventos e baseados nas relações sociais. Todos estes princípios podem ser utilizados em algoritmos de classificação dos actores das redes. Usualmente opta-se por utilizar as relações sociais, ou seja as ligações entre nós, como base para os algoritmos de detecção de comunidades, que apresentamos a seguir.

2.6 Detecção de comunidades

As técnicas de *clustering* são profícuas na análise exploratória de dados, com aplicações que vão da estatística à ciência de computadores, biologia ou psicologia. Em todas as ciências que têm que lidar com dados empíricos, uma das primeiras classificações que se tenta fazer dos dados é saber se podem ser agrupados através de alguma propriedade que se manifeste semelhante dentro de cada um dos grupos identificados. No entanto, todos os algoritmos de *clustering* encontram dados para os quais não conseguem encontrar uma boa partição dos pontos. Muitos foram desenvolvidos para lidarem efectivamente com situações que os métodos tradicionais não conseguiam resolver. Alguns métodos são robustos, sendo capazes de separar eficientemente grupos em conjuntos de dados muito diferentes, enquanto outros são mais específicos e necessitam de condições iniciais adequadas para que os seus resultados sejam satisfatórios (Shortreed, 2006).

As técnicas de *clustering* podem ser divididas em dois grupos principais, de acordo com a abordagem que fazem ao problema do particionamento. Podem ser globais, olhando para as redes como um todo e procurando a partir do conhecimento global da rede dividi-la em comunidades, ou módulos. Nesta categoria incluem-se por exemplo os métodos hierárquicos. As técnicas de particionamento podem, por outro lado, ser locais quando atendem aos padrões existentes em partes particulares da rede, como quando se estuda a existência de cliques completos. Este é o caso do método de percolação de cliques, onde

as comunidades são obtidas identificando os cliques completos adjacentes que constituem as comunidades.

2.6.1 Clustering hierárquico

Os algoritmos de particionamento hierárquicos procuram encontrar o particionamento de um grafo a partir dos grupos em que o grafo foi particionado anteriormente. Podem ser do tipo aglomerativo ("Bottom-Up") ou divisores ("Up-Down"). No primeiro, todos os elementos do grafo são separados em grupos contendo apenas 1 elemento. A partir daqui são aglomerados sucessivamente em grupos maiores até que uma determinada função de qualidade não consiga ser mais optimizada. Ambos os processos produzem um dendrograma das divisões efectuadas. A função de qualidade pode ser utilizada para definir o ponto de corte desse dendrograma, por forma a encontrar o valor óptimo da divisão das comunidades.

2.6.1.1 Algoritmo de Girvan e Newman

No artigo "Community structure in social and biological networks" (Girvan e Newman, 2001) é introduzido um novo algoritmo para a detecção de comunidades. Os autores referem que até 2001 a investigação principal se centrou nas propriedades de *small-world*, ou o reconhecimento de que a distância média entre vértices é pequena, nas análises das distribuições de potência onde a distribuição do *degree* dos vértices normalmente obedece a uma lei de potência ou exponencial, e na transitividade da rede (*clustering*) que basicamente diz que dois vértices que tenham uma ligação a um terceiro vértice comum, têm uma probabilidade maior de também estarem conectados por uma ligação. Neste artigo Girvan e Newman propõem que as redes sejam também analisadas olhando para uma outra propriedade: a estrutura de comunidades. Os autores relembram a condição típica das redes sociais, onde é facilmente observável a existência de comunidades sem que tal apareça claramente definido nos métodos estatísticos tradicionais. Os autores utilizam a expressão para o *clustering* do grafo dada pela equação 2.15 e definem o algoritmo de particionamento que ficou conhecido pelos seus nomes. Este algoritmo pretende ser capaz de detectar comunidades a partir das ligações que sejam menos centrais às comunidades, em oposição aos métodos de *clustering* hierárquico tradicionais, onde se procura construir uma medida para detectar quais as ligações mais centrais de uma comunidade. Assim, em vez de se construir comunidades através da adição de vértices a cada um dos grupos, os autores propõe a divisão do grafo original e o particionamento através da

remoção das ligações 'menos importantes'. Para isso, a medida adoptada para quantificar a importância de cada uma das ligações no seio do grafo é a *betweenness*. Esta, sendo definida pelo número de caminhos geodésicos que passam pelo vértice em questão, é uma medida da influência que esse vértice tem no fluxo de informação através da rede, particularmente nos casos em que o fluxo se processe preferencialmente através dos caminhos mais curtos. Os autores generalizam a *betweenness* dos vértices e aplicam-na às ligações, para determinar quais as ligações mais importantes para 'conectar' diferentes regiões da rede. Definem assim uma variação da medida a que chamam *edge betweenness*, que é o número de caminhos geodésicos entre vértices que passam através dessa ligação. Desta forma, se uma rede contiver várias comunidades que estejam ligadas por poucas ligações, estas ligações assumirão um papel importante na transmissão de informação entre comunidades, pelo que apresentarão valores elevados de *edge betweenness*. Removendo estas ligações separam-se as comunidades, revelando dessa forma a estrutura do grafo.

O algoritmo de particionamento de grafos proposto por Girvan e Newman (2001) é então:

1. Calcular o valor de *edge betweenness* para todos as ligações do grafo;
2. Remover a ligação com o valor mais alto de *edge betweenness*;
3. Recalcular a *edge betweenness* para todas a ligações afectadas pela remoção;
4. Repetir a partir do passo 2 até que não restem ligações.

Os autores fazem ainda notar a importância do passo 3, em que se recalcula o valor da *edge betweenness* das ligações afectadas. Pode acontecer que em comunidades ligadas por mais que uma ligação, estas ligações não tenham o mesmo valor de *edge betweenness* e após a remoção da ligação com valor mais alto, caso não fossem recalculados os valores, esta ligação poderia não ser removida. O cálculo da propriedade permite que a ligação assuma agora um papel mais importante na ligação entre comunidades, evitando ficar esquecida. O algoritmo produz um dendrograma, representando as divisões existentes ao longo do processo de divisão da rede em subgrafos. Desta forma, a escolha do ponto do dendrograma onde é definido o corte afecta o número e composição das comunidades identificadas.

2.6.1.2 Modularidade

Para encontrar o corte 'óptimo' do dendrograma e consequentemente encontrar as comunidades existentes numa rede, Newman e Girvan (2003) propuseram a medida de qualidade 'modularidade'. Esta medida é baseada na medida de *assortative mixing* (Newman, 2002), proposta anteriormente, onde 'modularidade' (Q) é definida por:

$$Q = \sum_i (e_{ij} - a_i^2) = \text{Tr e} - \|\text{e}^2\|, \tag{2.16}$$

e_{ij} é a fracção de todas as ligações na rede original, que ligam vértices da comunidade i à comunidade j na matriz simétrica $\mathbf{e}$ $(k \times k)$, onde k é o número total de comunidades. Tr $\mathbf{e}$ é o traço da matriz, que é a soma dos elementos da diagonal da matriz e dá a fracção de ligações que ligam vértices dentro da mesma comunidade Tr $\mathbf{e}$= $\sum e_{ii}$ e onde $a_i = \sum_j e_{ij}$ é a fracção de ligações que conectam nós dentro da comunidade i. O calculo da modularidade é então efectuado a cada divisão da rede dada pelo dendrograma e o valor óptimo será encontrado para a posição onde o valor de Q seja máximo.

Newman (2006) descreve uma abordagem ao particionamento de grafos por forma a detectar e caracterizar comunidades. Esta abordagem expande o trabalho realizado anteriormente, pois os métodos de detecção de comunidades em que a função de qualidade é conhecida como "modularidade" podem ser melhorados utilizando os vectores próprios de uma nova matriz característica do grafo, que o autor denomina de matriz de modularidade. Segundo o autor esta alteração da matriz de característica (que substituí a matriz de similaridade) leva a a um algoritmo espectral para a detecção de comunidades, mais rápido e que apresenta melhores resultados (Newman, 2006).

Atendendo à noção de que uma comunidade (ou módulo) é um subgrafo de uma rede onde a densidade de ligações é mais elevada que a densidade de ligações estabelecidas com outras comunidades (ou módulos), o método da modularidade óptima passa necessariamente pela noção de que existe uma divisão da comunidade que é óptima e que esta comunidade pode efectivamente ser dividida.

Para uma rede constituída por n vértices e supondo que existe uma divisão possível, $s_i = 1$ se o vértice i pertencer ao grupo 1 e $s_i = -1$ se pertencer ao grupo 2. A modularidade Q é dada por:

$$Q = \frac{1}{4m} \sum_{ij} \big(A_{ij} - \frac{k_i k_j}{2m}\big)(s_i s_j + 1) = \frac{1}{2m} \sum_{ij} \big(A_{ij} - \frac{k_i k_j}{2m}\big) s_i s_j \tag{2.17}$$

A equação 2.17 pode ser resumida a

$$Q = \frac{1}{4m} s^T B s \tag{2.18}$$

onde s é o vector coluna cujos elementos sõo s_i e onde a matriz B, real e simétrica é composta por

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m} \tag{2.19}$$

denominada matriz de modularidade.

Escrevendo a equação 2.18 como uma combinação linear dos vectores próprios normalizados de B, u_i de tal forma que $s = \sum_{i=1}^{n} a_i u_i$ com $a_i = u_i^T s$, obtém-se

$$Q = \frac{1}{4m} \sum_i a_i u_i^T B \sum_j a_j u_j = \frac{1}{4m} \sum_{i=1}^{n} (u_i^T s)^2 \beta_i \tag{2.20}$$

onde β_i é o valor próprio de B correspondente ao vector próprio u_i.

Assumindo que os valores próprios estão ordenados de forma decrescente, a maximização da modularidade é conseguida dividindo a rede e escolhendo s proporcional ao vector próprio u_1 (Newman, 2006). No entanto, tal não é possível uma vez que há a restrição de s_i ser 1 ou -1. No entanto, pode-se fazer esta separação utilizando o sinal dos elementos do vector próprio, atribuindo cada vértice a um grupo, conforme o sinal do correspondente componente de u_1 (Newman, 2006).

A divisão em mais do que duas comunidades não permite que possa ser utilizado o algoritmo anterior, o que levaria a que as ligações entre comunidades fossem apagadas, uma vez que tal modificaria o valor do *degree* dos vértices após a remoção das ligações entre comunidades. Assim, o autor propôs a adição de uma contribuição ΔQ à modularidade, após a divisão de um grupo g de tamanho n_g em dois:

$$\Delta Q = \frac{1}{2m}\Big[\frac{1}{2}\sum_{i,j\in g} B_{ij}(s_i s_j + 1) - \sum_{i,j\in g} B_{ij}\Big] \tag{2.21}$$

$$= \frac{1}{4m}\Big[\frac{1}{2}\sum_{i,j\in g} B_{ij} s_i s_j - \sum_{i,j\in g} B_{ij}\Big] \tag{2.22}$$

$$= \frac{1}{4m}\sum_{i,j\in g}\Big[B_{ij} - \delta\sum_{k\in g} B_{ik}\Big] s_i s_j \tag{2.23}$$

$$= \frac{1}{4m} s^T B^{(g)} s \tag{2.24}$$

onde $B^{(g)}$ é a matriz $n_g \times n_g$ com elementos indexados pelos índices i, j dos vértices dentro do grupo g e tendo o valor:

$$B_{ij}^{(g)} = B_{ij} - \delta_{ij}\sum_{k\in g} B_{ik} \tag{2.25}$$

O processo iterativo de divisão da rede é então executado enquanto a cada passo a divisão da rede apresente um ΔQ positivo, ou seja um incremento da modularidade. Caso o valor de ΔQ não seja positivo, com uma subsequente divisão da rede, então o processo deve ser parado, permitindo assim ao método ter um critério de paragem (Newman, 2006).

2.6.1.3 Algoritmo *fast community* de Clauset Newman e Moore

Um dos algoritmos utilizados na detecção de comunidades é proposto por Clauset *et al.* (2004). É um algoritmo hierárquico aglomerativo optimizado para a detecção de comunidades em redes de dimensões consideráveis. O algoritmo é baseado na modularidade (Newman e Girvan, 2003), uma medida da rede baseada na *assortative mixing* proposta por Newman (2002). A modularidade é uma propriedade da divisão de uma rede em comunidades que mede a qualidade de uma divisão, no sentido de haver mais ligações entre os elementos de uma comunidade do que as ligações existentes a conectar comunidades.

Partindo de uma matriz de adjacência A_{vw} os autores definem a modularidade como sendo:

$$Q = \frac{1}{2m}\sum_{vw}\Big[A_{vw} - \frac{k_v k_w}{2m}\Big]\delta(c_v, c_w) \tag{2.26}$$

na equação 2.26, a função $\delta(i, j)$ é igual a 1 se $i = j$ e 0 caso contrário e c_i é a comunidade i,

k_i é o *degree* do nó i e $m = \frac{1}{2}\sum_{vw} A_{vw}$ é número total de vértices do grafo. Assim definida, a modularidade representa o valor de desvio de uma rede em relação a uma rede aleatória. Valores de modularidade superior a 0.3 são bons indicadores da presença de uma estrutura de comunidades significativa na rede (Newman e Girvan, 2003).

O algoritmo proposto determina, a partir de uma rede constituída por tantas comunidades quantos nós da rede, os diversos incrementos de modularidade que qualquer atribuição de um elemento a uma determinada comunidade implica, seleccionando o máximo ΔQ para efectivamente agregar essas comunidades.

$$\Delta Q_{i,j} = \begin{cases} \frac{1}{2m} - \frac{k_i k_j}{(2m)^2} & \text{se i e j conectados} \\ 0 & \text{caso contrário} \end{cases} \tag{2.27}$$

$$a_i = \frac{k_i}{2m} \tag{2.28}$$

O algoritmo é definido por:

1. Calcular os valores de $\Delta Q_{i,j}$ e a_i através das equações 2.27 e 2.28 e popular uma matriz H com o valor máximo de cada linha da matriz ΔQ
2. Seleccionar os valores mais altos de $\Delta Q_{i,j}$ de H e juntar as comunidades correspondentes, actualizando a matriz ΔQ H e a_i e incrementar o valor de Q com $\Delta Q_{i,j}$
3. Repetir o ponto 2 até que apenas reste uma comunidade.

2.6.1.4 Resolução da modularidade

A utilização da 'modularidade' como função de qualidade para critério de corte do dendrograma dos processos de detecção de comunidades hierárquicos foi discutida por Fortunato e Barthelemy (2006), tendo os autores estudado o comportamento desta função para redes de diversas escalas. Os autores evidenciam no seu trabalho que a função 'modularidade' pode não ser capaz de identificar a existência de comunidades de tamanho inferior a um determinado valor limite, valor esse que depende do tamanho total da rede em causa e do grau de ligação existente entre comunidades. Os autores calcularam um limite para o tamanho das comunidades detectadas dado por:

$$l_s < 2l_R^{min} = \sqrt{2L} \tag{2.29}$$

onde para uma comunidade S, l_s é o número de ligações internas à comunidade S, l_R^{min} é o limite de resolução extremo e L é o número total de ligações existentes. Os autores propõem que para situações em que se verifique a desigualdade da equação 2.29, a divisão através da função de modularidade pode mascarar dentro da mesma comunidade várias sub-comunidades. Uma das estratégias que propõem é a aplicação dos métodos baseados em modularidade às sub-redes constituídas apenas pelos módulos onde a desigualdade 2.29 se verifique.

2.6.2 Clustering particional

2.6.2.1 K-means

O método *k-means* é um dos algoritmos não supervisionadas mais simples, capaz de resolver o problema de particionamento de um conjunto de dados, nomeadamente de redes. O objectivo deste algoritmo é o de particionar uma população N-dimensional em k comunidades. O procedimento é facilmente programável e requer poucos recursos computacionais e é apropriado para particionar redes de dimensões elevadas (Macqueen, 1967). O algoritmo faz a definição prévia do número de divisões k em que estamos interessados e prossegue de acordo com os seguintes passos:

1. Definição aleatória de k partições.
2. Cálculo dos centroides de cada uma destas partições.
3. Recolocação dos nós nas partições correspondentes aos centroides mais próximos.
4. Iterar os dois últimos pontos até que o sistema estabilize e não haja mais recolocações.

O método k-means apesar da sua rapidez, tem o inconveniente de não produzir resultados consistentes, uma vez que depende da distribuição aleatória inicial de pontos. No entanto há variantes deste método que foram desenvolvidas para tentar colmatar esta deficiência. Um problema é a necessidade de determinar à partida o valor de k. Uma estimativa inicial do número de partições pode ser dada por (Mardia *et al.*, 1980):

$$k \approx \sqrt{\frac{n}{2}} \tag{2.30}$$

Existem diversas variantes deste método, nomeadamente o *k-medoids*, em tudo semelhante ao algoritmo *k-means*, mas onde se atribui o centro das partições ao ponto mais próximo

em vez de ser calculado um centroide.

2.6.3 Decomposição espectral

As principais ferramentas para a utilização desta técnica são as chamadas matrizes espectrais. Há uma área de pesquisa completamente dedicada ao estudo destas matrizes, chamada *teoria espectral de grafos*. As matrizes utilizadas são as matrizes laplacianas, que, no entanto, assumem neste caso uma formulação diferente da apresentada anteriormente no ponto 2.3.7. Tal deve-se ao facto de na literatura se encontrarem muitas definições diferentes, pois diversos autores propõem versões próprias para a definir.

Von Luxburg (2007) define matrizes laplacianas a partir de grafos G não direccionados, com ligações entre vértices ponderados por uma martriz de pesos W, onde $w_{ij} = w_{ji} \geq 0$. Este autor assume que estas matrizes não têm que ser normalizadas, e propõe construir matrizes laplacianas não normalizadas e matrizes laplacianas normalizadas.

não-normalizadas: a matriz Laplaciana do grafo G é dada por:

$$L = D - W \tag{2.31}$$

onde L é a matriz Laplaciana, D a matriz com o valor do *degree* na diagonal e W a matriz de pesos das ligações.

normalizadas: neste caso há duas matrizes definidas na literatura. Ambas estão relacionadas uma com a outra e são definidas por:

$$L_{sym} := D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2} \tag{2.32}$$

$$L_{rw} := D^{-1} L = I - D^{-1} W \tag{2.33}$$

A primeira matriz L_{sym} é uma matriz simétrica enquanto a segunda L_{rw} está relacionada com a caminhada aleatória (von Luxburg, 2007).

2.6.3.1 Algoritmos para *spectral clustering*

Os algoritmos mais usuais em *Spectral Clustering* assumem que temos um conjunto de n pontos $x_1, x_2, ..., x_n$ que podem ser objectos arbitrários. São medidas as similaridades entre vértices $s_{ij} = s(x_i, x_j)$, através de uma qualquer função que seja simétrica e não negativa e assumem também que construímos a matriz $S = (s_{ij})_{i,j=1...n}$.

***Spectral clustering* não-normalizado**

Entrada: Matriz de similaridade $S \in \mathbf{R}^{n \times n}$ e o número de partições a construir k.

- Construir a matriz de similaridade de acordo com uma das estratégias definidas em 2.3.5, sendo W a matriz de pesos.
- Calcular a matriz Laplaciana não normalizada L
- Calcular os primeiros k vectores próprios $u_1, ..., u_k$ de L
- Construir a matriz $U \in \mathbf{R}(n \times k)$ contendo os vectores $u_1, ..., u_k$ nas colunas
- Para $i = 1, ..., n$, fazer $y_i \in \mathbf{R}^k$ ser o vector correspondente a cada linha i de U
- Particionar os pontos $(y_i)_{i=1,...,n}$ em $\mathbf{R}^k$ com o algoritmo k-means nas partições $C_1, ..., C_k$

Saída: Partições $A_1, ..., A_k$ com $A_i = \{j | y_i \in C_i\}$

Para o caso de particionamentos normalizados há dois algoritmos alternativos:

***Spectral Clustering* normalizado de acordo com Shi e Malik (2000)**

Entrada: Matriz de similaridade $S \in \mathbf{R}^{n \times n}$ e o número de partições a construir k.

- Construir a matriz de similaridade de acordo com uma das estratégias definidas em 2.3.5, sendo W a matriz de pesos.
- Calcular a matriz Laplaciana não normalizada L
- Calcular os primeiros k vectores próprios $u_1, ..., u_k$ da solução de $Lu = \lambda Du$
- Construir a matriz $U \in \mathbf{R}(n \times k)$ contendo os vectores $u_1, ..., u_k$ nas colunas
- Para $i = 1, ..., n$, fazer $y_i \in \mathbf{R}^k$ ser o vector correspondente a cada linha i de U
- Particionar os pontos $(y_i)_{i=1,...,n}$ em $\mathbf{R}^k$ com o algoritmo k-means nas partições $C_1, ..., C_k$

Saída: Partições $A_1, ..., A_k$ com $A_i = \{j | y_i \in C_i\}$

***Spectral Clustering* normalizado de acordo com Ng, Jordan e Weiss (2002)**

Entrada: Matriz de similaridade $S \in \mathbf{R}^{n \times n}$ e o número de partições a construir k.

- Construir a matriz de similaridade de acordo com uma das estratégias definidas em 2.3.5, sendo W a matriz de pesos.

- Calcular a matriz Laplaciana normalizada L_{sym}
- Calcular os primeiros k vectores próprios $u_1, ..., u_k$ de L_{sym}
- Construir a matriz $U \in \mathbf{R}(n \times k)$ contendo os vectores $u_1, ..., u_k$ nas colunas
- Construir a matriz $T \in \mathbf{R}(n \times k)$ a partir de U normalizando as linhas para normal 1, que é dado por $t_{ij} = u_{ij}/(\sum_k u_{ik}^2)^{1/2}$
- Para $i = 1, ..., n$, fazer $y_i \in \mathbf{R}^k$ ser o vector correspondente a cada linha i de T
- Particionar os pontos $(y_i)_{i=1,...,n}$ em $\mathbf{R}^k$ com o algoritmo k-means nas partições $C_1, ..., C_k$

Saída: Partições $A_1, ..., A_k$ com $A_i = \{j | y_i \in C_i\}$

Os 3 algoritmos descritos acima são semelhantes. As diferenças residem no facto de, em cada caso, o grafo Laplaciano ser diferente. Em todos os casos o importante é conseguir mudar a representação dos pontos abstractos x_i para $y_i \in \mathbf{R}^k$.

Sob o ponto de vista do Corte de Grafos (*Graph Cut*), dada uma matriz de similaridade e uma matriz de adjacência W, a forma mais simples de construir uma partição é resolver o problema do número mínimo de cortes necessários para separar o grafo. O problema é bastante fácil de resolver, mas o método nem sempre é fiável, levando a que por vezes o particionamento do grafo não seja o ideal (von Luxburg, 2007).

2.6.3.2 Caminhada aleatória

A caminhada aleatória num grafo é um processo estocástico de saltos consecutivos nos vértices do grafo. Von Luxburg (2007) mostra que o particionamento espectral pode ser interpretado como a tentativa de encontrar a partição do grafo tal que a caminhada aleatória permaneça o maior número de passos dentro dessa partição e raramente salte entre partições.

2.6.3.3 Distância de comutação (distância de resistência)

A distância de comutação c_{ij} entre dois vértices v_i e v_j é o tempo esperado médio que leva uma caminhada aleatória para ir de v_i a v_j e de regresso.

A vantagem da distância de comutação sobre a distância geodésica é que a primeira decresce se existirem diversos caminhos curtos (não necessariamente iguais ao geodésico)

entre os dois vértices. Em vez de olhar para um caminho (o geodésico) a distância de comutação é neste sentido apropriada para utilização em particionamento, uma vez que vértices unidos por caminhos curtos e que estejam numa zona de alta densidade de ligações estão naturalmente mais próximos que vértices unidos por um caminho curto, mas em diferentes zonas de alta densidade de ligações. Esta distância de comutação c_{ij} apresenta a vantagem de $\sqrt{c_{ij}}$ poder ser considerada uma função da distância euclidiana dos vértices do grafo (von Luxburg, 2007).

2.6.3.4 Aplicação prática

Von Luxburg (2007) faz uma discussão dos aspectos práticos da aplicação do particionamento espectral. Há diversas escolhas a fazer que a autora discute em pormenor e que é preciso ter em conta:

Construção do grafo de similaridade: Segundo a autora a construção do grafo necessário para proceder ao particionamento espectral não é trivial uma vez que se conhece pouco as implicações teóricas das diversas construções. O primeiro problema passa pela definição da função de similaridade entre pontos. Se o grafo a construir for um grafo de vizinhanças, é necessário que a medida escolhida para definir a similaridade tenha significado para o problema em causa. No caso dos pontos de estudo existirem num espaço euclidiano $\mathbf{R}^d$, uma função tipicamente utilizada para a similaridade é dada por:

$$s(x_i, x_j) = \exp\left(-||x_i - x_j||^2/(2\sigma^2)\right) \tag{2.34}$$

No entanto, esta equação introduz mais um parâmetro (σ) que será preciso definir posteriormente. A autora aconselha a que a função de similaridade escolhida dependa do domínio de onde os dados são retirados.

O tipo de grafo de similaridade: A escolha seguinte tem a ver com os aspectos do tipo de grafo pretendido, seja o de ε-vizinhança(2.3.5.1) ou k-vizinhança(2.3.5.2). Luxburg analisa as diferenças entre as diversas opções, salientando principalmente que quando se está a trabalhar com dados que se sabe à partida conterem diferentes escalas, o método da ε-vizinhança apresenta problemas, uma vez que a escolha do valor de ε fará com que alguns grupos se apresentem fortemente conectados enquanto outros ficarão mais 'soltos'.

2.6.4 Percolação de cliques

Um dos aspectos mais interessantes da teoria de grafos tem a ver com a existência de uma probabilidade crítica de estabelecimento de ligações, a partir da qual surge um componente gigante. Isto significa que abaixo de um determinado valor p_c de ligação, a rede é composta por sub-redes isoladas, enquanto que quando se atinge essa probabilidade p_c a rede é composta por um único componente. Este fenómeno é semelhante à transição de percolação estudado na matemática ou em mecânica estatística (Stauffer e Aharony, 1994), (Bunde e Havlin, 1995). Na realidade, a transição de percolação e a emergência do componente gigante são o mesmo fenómeno embora expresso em linguagens diferentes (Albert e Barabasi, 2001).

Considerando uma grelha de dimensão d, cujas arestas estejam presentes com uma probabilidade p e ausentes com uma probabilidade $1 - p$, a teoria de percolação estuda a emergência de caminhos que percolam através da grelha, isto é, que atravessam de um lado ao outro da grelha. Para valores pequenos de p apenas algumas arestas estão presentes, pelo que apenas pequenos grupos de nós estão ligados por arestas. Contudo, ao ser atingida a probabilidade crítica p_c , limiar de percolação, surge um *cluster* de nós de percolação ligados por arestas. Este *cluster* é também denominado *cluster* infinito.

As propriedade principais a reter no que diz respeito à percolação são:

A probabilidade de percolação P, que denota a probabilidade de que um determinado nó pertencer ao *cluster* infinito.

$$P = P_p(|C| = \infty) = 1 - \sum_{s<\infty} P_p(|C| = s) \tag{2.35}$$

O tamanho médio do *cluster*, $< s >$

$$< s >= E_p(|C|) = \sum_{s=1}^{\infty} sP_p(|C|) = s) \tag{2.36}$$

que dá o valor esperado do tamanho dos clusters formados. $< S >$ é infinito quando $P > 0$ ($P > 0$ se $p > p_c$). É útil nessas situações trabalhar com os tamanhos médio esperados dos clusters finitos, ignorando o cluster infinito. Por fim, outra propriedade é a distribuição de tamanhos dos clusters n_s, definida como

$$N_s = \frac{1}{s} P_p(|C| = s) \tag{2.37}$$

onde o valor de N_s não coincide com a probabilidade de um nó fazer parte de um *cluster* de tamanho s.

A grande diferença entre a teoria da percolação e a percolação de cliques é que a teoria de percolação é definida para grelhas, enquanto numa rede de outro tipo se pode definir uma distância não métrica através das ligações e qualquer nó pode estar ligado a um outro qualquer nó. No entanto e atendendo a que ambos se encontram no limite, quando a dimensão tende para infinito, muitos dos resultados da percolação podem ser adaptados para outras redes (Albert e Barabasi, 2001).

A detecção de comunidades normalmente recorre a algoritmos que implicam a subdivisão de rede através da quebra de ligações entre grupos que estejam densamente conexos. No entanto, esses algoritmos não prevêem a possibilidade de que as comunidades se sobreponham e que um determinado membro ou membros façam parte de mais do que uma comunidade. A utilização de percolação de cliques permite definir um método capaz de identificar comunidades que se sobrepõe em redes de grandes dimensões (Derenyi *et al.*, 2005).

A técnica da percolação por cliques define algumas noções importantes:

A noção de k-clique, que não mais é que um clique (2.3.10) composto por nós com *degree* k dentro desse sub-grafo.

Adjacência de k-cliques: diz-se que dois k-cliques são adjacentes se partilharem $k-1$ vértices, *i.e.* se diferirem apenas de um nó da rede.

Cadeia de k-cliques: diz-se de um sub-grafo que seja a união de uma sequência de k-cliques adjacentes.

Conectividade k-clique (do inglês *k-clique connectedness*): dois k-cliques dizem-se conectados se ambos fazem parte da mesma cadeia k-clique.

Cluster de percolação k-clique: é o sub-grafo máximo de k-cliques conectados *i.e.* trata-se da união de todos os k-cliques que estão conectados a um determinado k-clique (Derenyi *et al.*, 2005).

Os requisitos principais para as técnicas de detecção de comunidades, ou módulos, é que sejam locais, que sejam baseados na densidade de ligações e que sejam tolerantes a erros (a remoção ou inserção de um link deve apenas alterar módulos próximos). Por outro lado, grupos densos em grafos reais muitas vezes sobrepõe-se a outros grupos. É fácil imaginar que, por exemplo, uma pessoa inserida num grupo social possa efectivamente pertencer a diferentes grupos (família, colegas, amizades). A proibição de sobreposição durante a

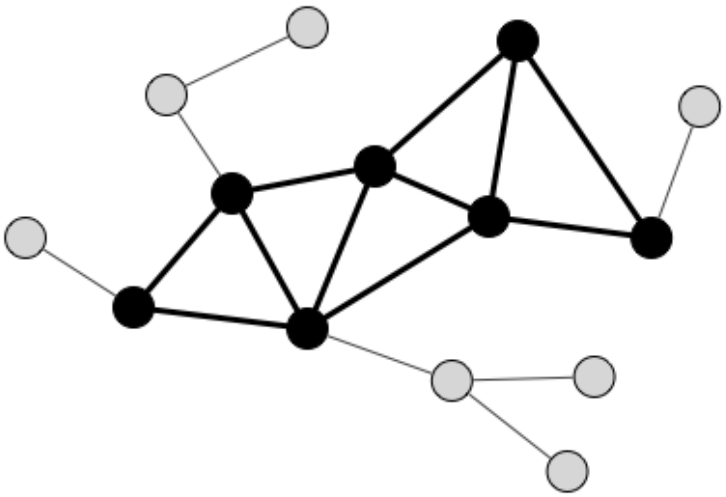

Figura 2.4: Grafo evidênciando uma cadeia de 3-cliques

identificação das comunidades aumentará naturalmente a probabilidade de ocorrência de falsos negativos (Palla *et al.*, 2007).

A definição de percolação por cliques é baseada em k-cliques e baseia-se na procura dos *clusters* de percolação k-cliques, com a vantagem de permitir a sobreposição de comunidades, algo que noutros métodos não é permitido (Palla *et al.*, 2005). Para além da versão normal do algoritmo de detecção de comunidades por percolação de cliques, que foi proposta para comunidades com ligações não direccionadas, foi proposta em 2007 uma versão aplicável a grafos dirigidos (Palla *et al.*, 2007). Ebel *et al.* em 2002 utilizaram esta abordagem para estudar a topologia da rede de correio electrónico.

2.7 Detecção de comunidades em redes de email

Uma das primeiras questões que se coloca quando se pretende proceder à detecção de comunidades na rede de correio electrónico tem a ver com a necessidade de filtrar os dados obtidos. Um dos problemas que se coloca é a quantidade de correio electrónico indesejado que é recebido nas caixas de correio, vulgo *spam*. Têm sido propostos vários métodos para a detecção e eliminação deste tipo de mensagens, mas isto tem levado a surgimento de um problema novo que é o aparecimento de falsos positivos. Sistemas baseados na análise de conteúdos e teste *bayesinanos* tem sido propostos, assim como a manutenção de listas negras e listas brancas (Garriss *et al.*, 2006). Para além destes sistemas mais tradicionais de detecção de correio electrónico indesejado tem-se assistido recentemente a uma abordagem em que se aproveita o conhecimento que se tem da estrutura das

redes sociais dos indivíduos por forma a filtrar o correio electrónico que chega à caixa de correio do utilizador. A técnica proposta por Kim (2007) utiliza uma decomposição espectral da matriz laplaciana construída a partir dos cabeçalhos das mensagens de correio electrónico recebidas na caixa de correio do utilizador. A partir da decomposição da rede de contactos gerada pelas mensagens em sub-redes, os autores classificam cada uma das sub-redes quanto aos seus coeficientes de *clustering* sendo que sub-redes com valores altos de *clustering* podem ser classificadas como não sendo de correio indesejado(Kim, 2007). Neste caso a utilização de técnicas de detecção de comunidades foi capaz de particionar correctamente o correio electrónico no corpo de estudo em desejado e indesejado sem recorrer à análise de conteúdo semântico das mensagens.

Para além do natural interesse na detecção e filtragem do correio electrónico a análise pode ser dirigida apenas ao correio electrónico legitimo a fim de identificar as propriedades das redes subjacentes às trocas de mensagens válidas.

Tyler *et al.* (2003) estudaram a rede de correio electrónico dos laboratórios de investigação da empresa Hewllet-Packard, e faziram uma análise dos ficheiros de *logs* de correio electrónico para tentar detectar comunidades e a estrutura informal das relações e interesses existentes. Tentaram assim perceber a dinâmica da informação dentro da empresa.

Os autores mostram que estas redes informais coexistem com a estrutura formal da organização e servem-na a diversos níveis, tais como a resolução de objectivos conflituosos ou problemas de projectos internos. Para além disso as redes informais funcionam como um meio de aprendizagem e transmissão de conhecimento dentro da empresa (Tyler *et al.*, 2003).

Devido ao valor que estas comunidades de prática possuem para as organizações, é desejável um método que seja rápido, prático e preciso. Os autores apresentam um método automatizado para identificar estas redes dentro da organização. Utilizaram a rede de correio electrónico do laboratório para construir uma rede das trocas de mensagens e particionaram esta rede para identificar comunidades.

O sistema de correio electrónico incluiu cerca de 1 milhão de mensagens e cobriu o período de cerca de dois meses (Tyler *et al.*, 2003).

Devido ao elevado número de mensagens os autores definiram um número limiar de mensagens que é necessário existir entre dois utilizadores para que estes fossem considerados conectados. Os grafos construídos desta forma revelaram possuir uma distribuição que obedece a uma lei de potência.

No estudo realizado por Ebel *et al.* (2002) os autores analisaram a rede de correio electrónico da universidade de Kiel, na Alemanha, a partir dos dados de registo de mensagens enviadas durante o período de 112 dias. Os nós da rede correspondem aos endereços de correio electrónico dos alunos e as ligações correspondem à existência de um envio de mensagens entre eles. No caso a rede resultante consistiu em 59812 nós das quais 5165 correspondem a contas de alunos, sendo que apresenta um *degree* médio de $< k >= 2.88$ que contém vários componentes separados com menos de 150 nós e um componente gigante com 56969 nós onde o *degree* médio é $< k_{giant} = 2.96 >$.

Os autores verificaram que a distribuição de *degree* obedece a uma lei de potência e que apresenta uma comportamento exponencial na cauda da distribuição (para valores de *degree*>100)

$$n(k) \propto k^{-1.81}$$

A medição dos coeficientes de *clustering* da rede mostrou discrepâncias entre os valores calculados através da equação 2.14 e da equação 2.15. O processo de medição faz com que os vizinhos dos nós exteriores à rede da universidade sejam pouco conhecidos e os autores obtiveram portanto valores mais baixos para C_Δ do que C. Apesar de tudo, os autores verificaram também que, apesar das limitações de medição devidas às ligações ao exterior da Universidade introduzirem um desvio, o *clustering* da rede era 1 a 2 ordens de grandeza superior ao de redes aleatórias com igual distribuição de *degree*.

2.8 Conclusão do estado da arte

O estudo de redes sociais e a aplicação dos diversos métodos e algoritmos expostos neste capítulo dependem naturalmente do corpo de estudo a analisar. As noções básicas do ponto 2.3 são as ferramentas sobre as quais outras noções são construídas, adicionando graus de complexidade a este estudo. Atendendo ao corpo de estudo, que será descrito pormenorizadamente no capítulo 4, algumas das noções apresentadas não serão aplicadas neste trabalho embora tenham sido referidas neste levantamento do estado da arte. Tal é o caso da medida *PageRank* do ponto 2.4.1.5 uma vez que esta é utilizada para redes direccionadas e no nosso caso de estudo se optou por tratar a rede como não-direccionada. Também os temas tratados nos pontos 2.6.2 e 2.6.3 que abordam o *clustering* particional e a decomposição espectral respectivamente, não foram aplicados ao caso de estudo uma vez que são abordagens genéricas aos problemas de classificação de dados, nomeadamente

segmentação de imagem e classificação de dados, optando-se por estudar e aplicar aqueles que foram desenvolvidos particularmente para os problemas de redes. Assim, entre os algoritmos de carácter global, foram estudados os dois algoritmos hierárquicos baseados na modularidade, sendo um aglomerativo e o outro divisivo. Fez-se também a caracterização através da determinação dos *k-cores* da rede. Nos algoritmos de carácter local, utilizou-se a percolação de cliques para verificar a tranversalidade dos grupos de comunicação informal aos grupos institucionais do corpo de estudo.

O estado da arte exposto ao longo destas páginas serviu de enquadramento para a colocação de questões sobre os sistemas de comunicação informal em sistemas de correio electrónico. No próximo capítulo abordamos dessas questões, que nos levaram à formulação de uma hipótese sobre a estrutura deste tipo de redes.

Capítulo 3

Hipótese

Para definir a nossa hipótese de investigação, começamos por colocar uma série de questões que foram sendo levantadas durante os estudos prévios relativamente ao trabalho desenvolvido. Estas perguntas levaram à identificação de uma questão fundamental, que procurou ser respondida através da formulação de uma hipótese.

Durante a fase de preparação de trabalho, de entre muitas outras questões, acabámos por nos cingir àquelas que nos pareceram mais pertinentes:

- Como funciona o sistema de correio electrónico do ISCTE? É transparente, *i.e.* não revela estruturas sociais existentes na universidade, ou por outro lado permite ter uma impressão digital do funcionamento da universidade?
- A comunidade de alunos agrega-se em torno das turmas ou dos cursos existentes, ou o seu comportamento é transversal e não é centrado apenas num circulo próximo de amizades?
- Há diferenças estruturais significativas entre as redes de alunos, professores e funcionários?
- Qual é a volatilidade da estrutura de redes? Notam-se diferenças entre alunos de primeiro ano e alunos de anos subsequentes?
- Nota-se a integração dos alunos recentemente chegados à instituição, sejam eles caloiros, erasmus, ou bolseiros de algum tipo?
- Como evoluem as comunidades ao longo do tempo? Há uniões? Cisões?
- Podemos detectar essas comunidades apenas a partir dos dados de registo das comunicações entre os intervenientes?

- Os algoritmos de detecção de comunidades apresentam resultados substancialmente diferentes? Em termos computacionais quais os que apresentam utilização mais prática?

Estas questões iniciais levaram a que, na elaboração do trabalho, se concluísse que poderiam ser resumidas a uma pergunta fundamental, à qual uma resposta seria também solução para estas questões prévias:

Pergunta fundamental: Qual a estrutura da rede de comunicação informal composta pelos utilizadores do serviço de correio electrónico do ISCTE?

A resposta a esta questão revelar-se-á esclarecedora para as questões iniciais.

Assim e no seguimento desta questão, formulamos uma hipótese a respeito do sistema em estudo.

- **Hipótese**: A detecção de comunidades, através da análise topológica sem observação da componente semântica, é possível, encerrando na estrutura topológica do grafo que a representa, informação suficiente para caracterizar comunidades e hierarquias dentro da rede de comunicação informal do ISCTE.

A prova desta hipótese efectuou-se através de um caso de estudo. No capítulo 4 fazemos a caracterização do corpo de estudos. Aplicamos os algoritmos de detecção de comunidades e desenvolvemos um modelo de agentes para concluir sobre a validade desta hipótese.

Capítulo 4

O correio electrónico do ISCTE

Com o objectivo de testar a nossa hipótese de investigação, desenvolvemos um caso de estudo que aplica os mecanismos de detecção de comunidades. Começamos por fazer uma descrição do corpo de estudo, apresentando a sua caracterização inicial. Seguem-se os resultados da aplicação dos métodos de detecção de comunidades e de caracterização da estrutura da rede estudada. Seguidamente abordamos o problema da modelação do sistema, propondo um modelo baseado em agentes para análise. Por fim, fazemos um resumo comparativo dos diversos resultados, relacionando os diferentes métodos e abordagens.

4.1 O caso de estudo

O caso de estudo escolhido foi o sistema de comunicação através de correio electrónico do ISCTE. Tal sistema apresenta algumas característica interessantes, das quais destacamos:

- Acessibilidade aos dados em bruto, mediante garantia de anonimização.
- Volume de dados: a dimensão da instituição permite aceder a um conjunto elevado de dados, nomeadamente fazer o cruzamento dos dados do sistema de correio electrónico com os dados existentes no sistema de gestão académica Fenix[1].
- Várias sub-redes disponíveis para estudo: Professores, Alunos, Funcionários.

[1] O sistema Fenix é um sistema de registo de central de todos os intervenientes na vida académica, permitindo aos serviços fazer a gestão de todos os processos respeitantes aos seus participantes e servindo de plataforma para a disseminação de conteúdos e informação para as actividade da universidade.

- Comunidades bem conhecidas *à priori*, como sejam os departamentos ou cursos a que os professores e os alunos estão associados.

O corpo de estudo utilizado para a produção deste trabalho foi constituído pelos dados recolhidos nos servidores de correio electrónico do ISCTE ao longo do 10 semanas, entre 23 de Abril de 2008 e 23 de Junho de 2008. A figura 4.1 apresenta o número de mensagens processadas por hora ao longo do período em estudo.

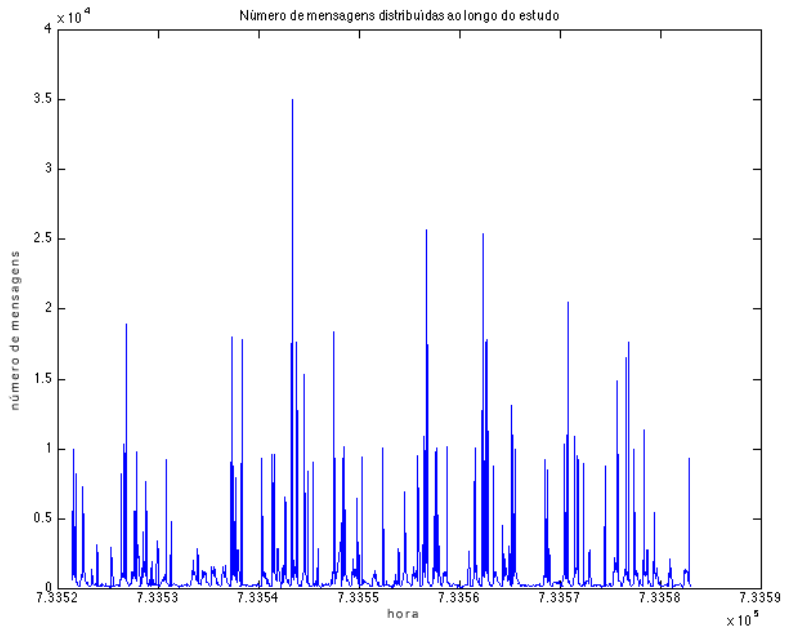

Figura 4.1: Histograma do número de mensagens processadas pelo serviço

A figura 4.1 revela a existência de períodos horários onde o volume de mensagens processadas pelo sistema de correio electrónico é elevado, tendo ultrapassado em 4 momentos as 20000 mensagens processadas por hora. Também é identificável o ciclo semanal de envio de mensagens com um menor número de mensagens processadas durante os fins de semana e feriados.

Os ficheiros de registo originais (*logs*) não incluem nenhum conteúdo das mensagens trocadas entre os membros do ISCTE, indicando apenas o emissor e o(s) destinatário(s) das mensagens trocadas. Para além de não acedermos a nenhum campo com conteúdo semântico e a fim de garantir a privacidade dos utentes do sistema, foi criado um sistema de anonimização automática dos ficheiros de registos, ficando os originais junto do servidor de correio electrónico. A Direcção de Serviços Informáticos (DSI) permitiu-nos aceder apenas aos dados anonimizados (ver diagrama da figura 4.2). Desta forma, os dados provenientes dos *logs* do correio electrónico e as correspondentes tabelas de características

provenientes do sistema Fenix foram utilizadas sem que houvesse nelas qualquer dado que permitisse identificar os utilizadores do serviço de correio electrónico do ISCTE.

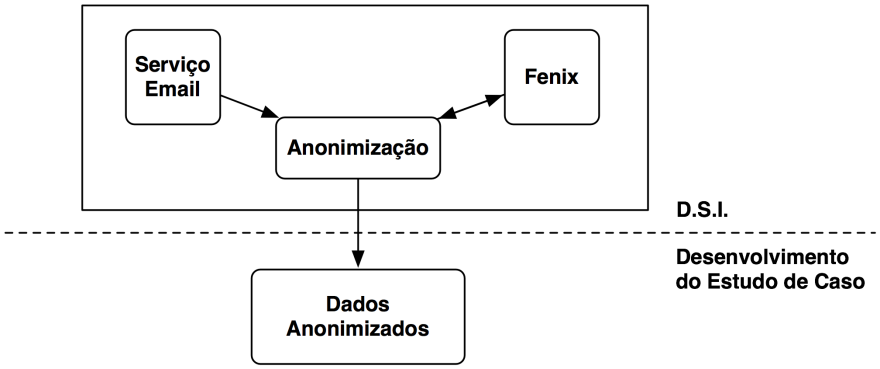

Figura 4.2: Diagrama de anonimização dos dados dos logs de email do ISCTE

4.1.1 Preparação dos dados

O serviço de correio electrónico do ISCTE apresenta um volume elevado de tráfego, sendo que o grosso das mensagens enviadas são-no para fora do sistema, ou de fora para alguém dentro do sistema. As mensagens trocadas internamente são minoritárias, mas são contudo as que são possíveis de controlar, de forma a que se possa fazer um mapeamento entre os recipientes e as suas características (curso, departamento, sexo, etc...). Esta necessidade de filtrar os endereços de correio electrónico externos levou a que os resultados possam apresentar algum desvio em relação à realidade, uma vez que muitos membros do ISCTE podem optar por utilizar nos seus contactos um correio electrónico externo, embora lhes tenha sido atribuído um endereço interno a partir do momento em que ingressam no ISCTE.

Para além desta questão, relativa à utilização de emails externos à rede do ISCTE, o sistema apresenta também o problema do correio electrónico indesejado (vulgo *spam*). Embora o sistema possua meios próprios de controlo de *spam*, haverá naturalmente alguma percentagem de correio indesejado que chega aos recipientes. A filtragem destes emails é também necessária.

Por outro lado, o sistema de correio electrónico do ISCTE possui listas de distribuição de mensagens. Estas, se forem consideradas nós das redes de recipientes, serão *hubs* importantes da rede. No entanto, não podem ser mapeadas a pessoas e servem principalmente para distribuição de informação institucional e portanto não serão reflexo de uma estrutura auto-organizada latente, mas antes de uma matriz imposta pela organização do

ISCTE, não incorporando informação de cariz pessoal e portanto não sendo útil para a identificação de comunidades (Tyler *et al.*, 2003).

Na preparação dos dados foram executados 3 passos de filtragem:

- Anonimização dos dados dos ficheiros de registo do servidor de correio electrónico do ISCTE e das correspondências com os dados do sistema Fenix. Esta operação foi realizada na D.S.I., antes da libertação dos dados para o desenvolvimento do caso de estudo.
- Remoção das mensagens enviadas a partir de endereços de correio electrónico exteriores ao ISCTE e das mensagens de correio indesejado.
- Remoção dos endereços das listas de distribuição interna do ISCTE.

4.1.2 Caracterização dos dados do caso de estudo

O conjunto de dados utilizado foi obtido a partir dos registos dos servidores de correio electrónico do ISCTE ao longo de 2 meses. Durante estes 62 dias, o serviço de email recebeu pedidos para envio de 1670313 mensagens, englobando 4235349 endereços de correio electrónico diferentes. O número de mensagens enviadas pelo sistema foi de 242544, sendo as restantes descartadas por não serem autorizadas (*spam*, origem não autorizada, etc...).

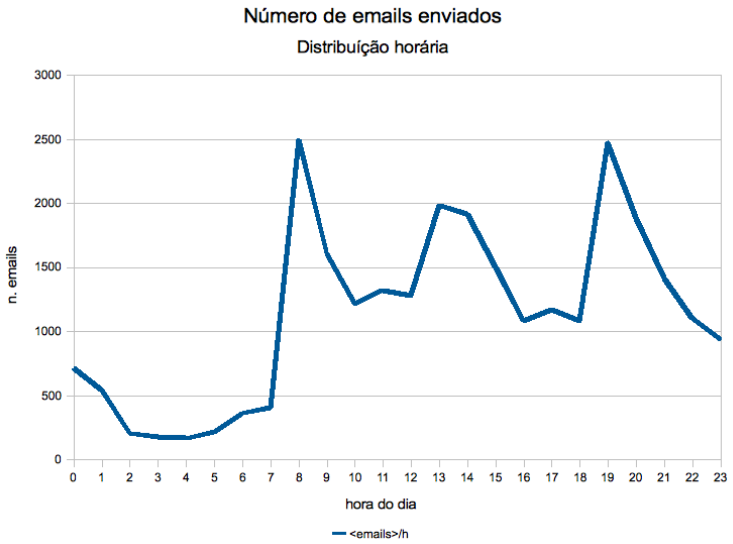

Figura 4.3: Distribuição horária do número de emails processados pelos sistema

O número médio de mensagens de correio electrónico processadas pelo sistema por hora, no período em estudo, foi de 1137. A figura 4.3 mostra a distribuição horária do número de

mensagens enviadas, onde se verifica claramente a existência de uma variação horária relativa ao volume de mensagens processado. São identificáveis visualmente sete zonas:

- Período nocturno entre as 0h e as 8h, onde o volume de mensagens processado é inferior à média diária.
- Período entre as 8h e as 10h da manhã, que corresponde ao início das actividades no ISCTE.
- Período entre as 10h e as 13h.
- Período de almoço entre 13h e 15h, onde o número de mensagens volta a aumentar.
- Período entre as 15h e as 19h, que representa o período mais calmo em termos de envio e recepção de mensagens da zona diurna.
- Período entre as 19h e 21h, onde novamente se verifica grande actividade de envio de mensagens, à semelhança do que se verifica no período das 8h-10h durante a manhã.
- Período entre 21h e 24h, onde os níveis de tráfego caem para valores semelhantes aos dos períodos intermédios verificados durante o dia.

A distribuição do número médio horário de emails enviados semanalmente é dada pela figura 4.4:

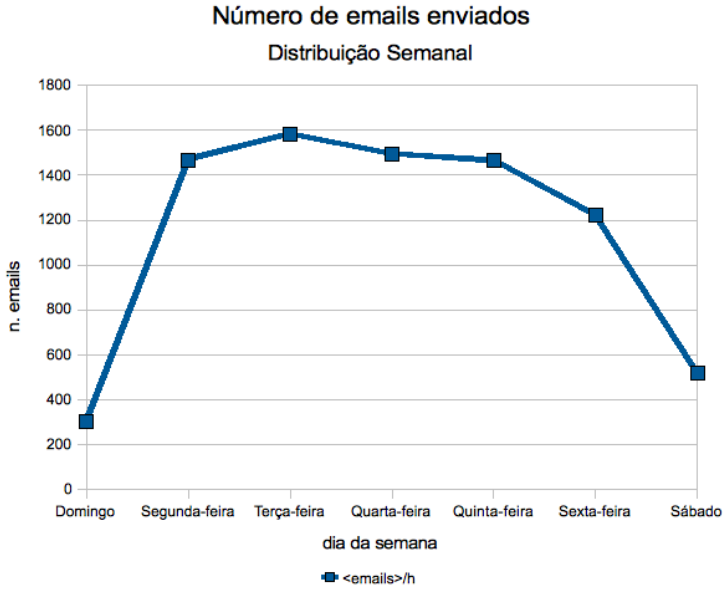

Figura 4.4: Distribuição semanal do número de emails processados pelos sitema

Notam-se facilmente dois regimes de funcionamento, o compreendido pelos 5 dias úteis

onde o volume de mensagens processadas pelo sistema é elevado e o regime de fim-de-semana onde o sistema processa muito menos mensagens. Nota-se que comparando sexta-feira com os restantes dias da semana, o envio de mensagens de correio electrónico é menor.

Da análise dos dados verificámos que foram trocadas mensagens entre 51702 endereços diferentes de correio electrónico do domínio "iscte.pt". Este número de endereços inclui duplicados, na medida em que há casos em que o mesmo utilizador utiliza diversos *alias* para a mesma conta de email. No serviço de email estudado todos os alunos possuem um endereço de correio electrónico com o número mecanográfico precedido de uma letra indicativa do nível a que pertencem, mas possuem também um *alias* com o seu nome para facilitar a memorização e distribuição dos endereços de correio electrónico aos seus contactos. Este número inclui também os endereços das listas de distribuição, serviços académicos e posições oficiais, que naturalmente não correspondem a pessoas em concreto dentro do ISCTE e que quando cruzados com os dados do sistema Fenix revelam números inferiores.

Deste corpo de 51702 endereços de email, o sistema Fenix indicou que 11833 endereços de correio electrónico pertencem a alunos, 240 pertencem a alunos estrangeiros em Erasmus no ISCTE, 1164 são respeitantes a docentes da instituição e 426 dizem respeito a funcionários do ISCTE.

A classificação nas categorias anteriores fez com que ficassem por classificar 38279 endereços de correio electrónico. Estes podem-se distribuir em duas categorias: ou o sistema Fenix não tem registos deles em base de dados, ou então são casos de pessoas que não se enquadram nas 4 classes de utilizadores do ISCTE. Verificou-se que destes 38279 endereços, 38188 pertencem a endereços de correio electrónico que não estavam no sistema Fenix, podendo corresponder a listas de distribuição, endereços de correio electrónico de serviços ou endereços de correio electrónico de projectos e que portanto têm correspondência a uma pessoa concreta no sistema Fenix. Os restantes 91 endereços de correio electrónico pertencem a pessoas que não se enquadram na classificação anterior mas que estão inseridos na base de dados do sistema Fenix. Poderão tratar-se de más inserções ou casos de desistências após pré-inscrição, mas será algo que eventualmente será analisado no futuro pela Direcção de Serviços de Informática do ISCTE.

Em termos de distribuição de idades os 5 grupos (Alunos, Alunos de Erasmos, Empregados, Professores, e Não classificados) apresentam os valores médios da tabela 4.1.

Ao longo do período analisado foram contabilizados 242544 emails entre utilizadores do ISCTE correspondentes a 11764 nós (que enviaram emails para dentro do ISCTE), que

Tabela 4.1: Distribuição de idades por grupo de utilizador

Categoria	Idade	Observações
Alunos	27,6	(11698 emails)
Aluno Erasmus	25,0	(218 emails)
Funionários	42,0	(426 emails)
Docentes	48,7	(1153 emails)
Outros	22,7	(83 emails)

estabelecem o máximo de 49933 ligações.

A rede tem uma densidade de 0,07% (percentagem de ligações existentes sobre o número de ligações possíveis).

O *average degree* é de $8,49$, o que quer dizer que cada nó (utilizador) está ligado em média a outros $8,49$ nós, quer como emissor quer como receptor.

Os cinco grupos anteriores podem ser reduzidos fundamentalmente a três: Professores, Funcionários e Alunos, sendo que os Alunos de Erasmus apresentam uma percentagem pequena, assim como os não classificados.

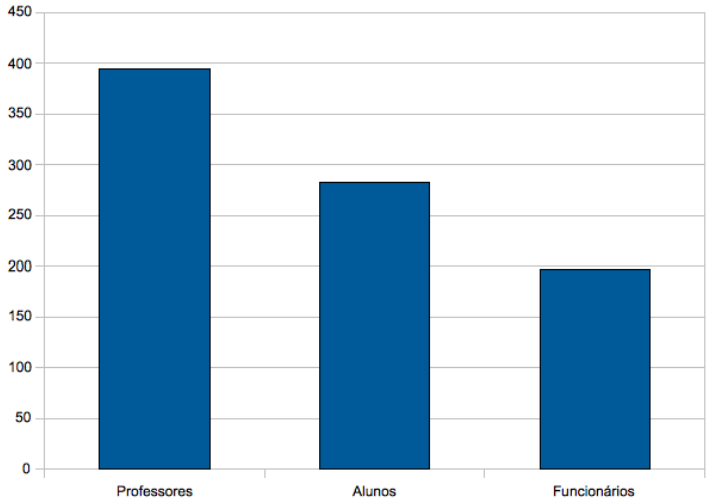

Figura 4.5: Número de nós em cada uma das 3 sub-redes

A análise de cada uma destas 3 sub-redes revelou que a adopção do sistema de correio electrónico por parte de cada um destes grupos não é uniforme. Em termos percentuais verificou-se que a adopção do sistema de correio electrónico por parte dos alunos para a troca de mensagens entre si é relativamente baixa (2.4%) quando comparada com a verificada para professores e funcionários da instituição (34.3% e 46.2% respectivamente).

As três redes apresentam as seguintes características:

Tabela 4.2: Número de nós em cada sub-rede do ISCTE

Sub-rede	n.º de membros	n.º total	percentagem
Professores	395	1153	34.3%
Alunos	279	11698	2.4%
Funcionários	197	426	46.2%

4.1.2.1 Rede de professores

A rede não direccionada formada pelas mensagens de correio electrónico trocadas pelos professores do ISCTE é composta por 2097 ligações e 395 nós (sendo os nós os professores e as ligações a existência de uma ou mais mensagens enviadas entre eles) agregados num único componente (não há grupos desligados).

Esta rede apresenta um *degree* médio de 10.6 e a distribuição de probabilidade do *degree* aproxima-se de uma distribuição exponencial, como se pode ver na figura 4.6. A densidade desta rede é de 2.69%.

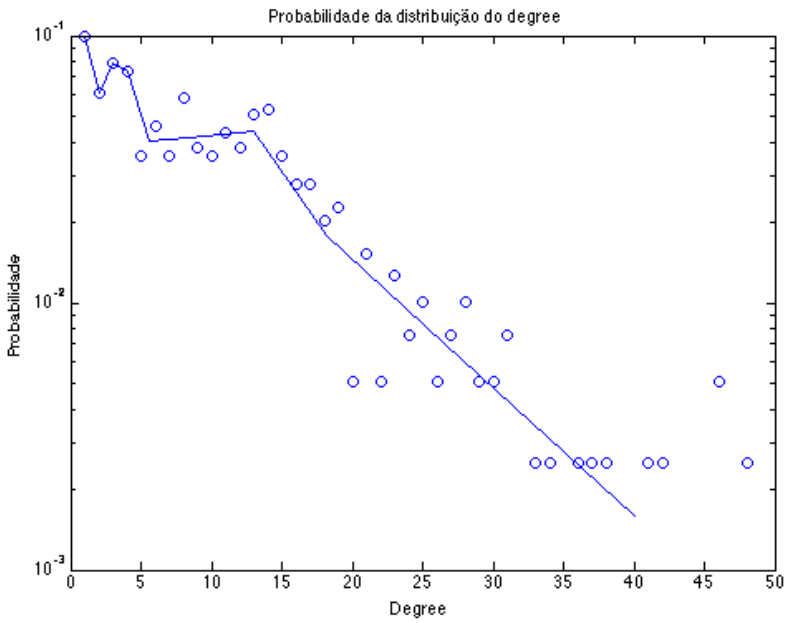

Figura 4.6: Distribuição do *degree* na rede de professores

O diâmetro desta rede é de 8, o que significa que no máximo são precisas 8 ligações para unir os dois professores mais distantes da rede de professores.

Definido um evento como o envio de uma mensagem de correio electrónico para um ou mais destinatários, pode-se verificar que a distribuição do número de eventos em função do números de destinatários obedece a uma lei de potência com expoente -2.7, como

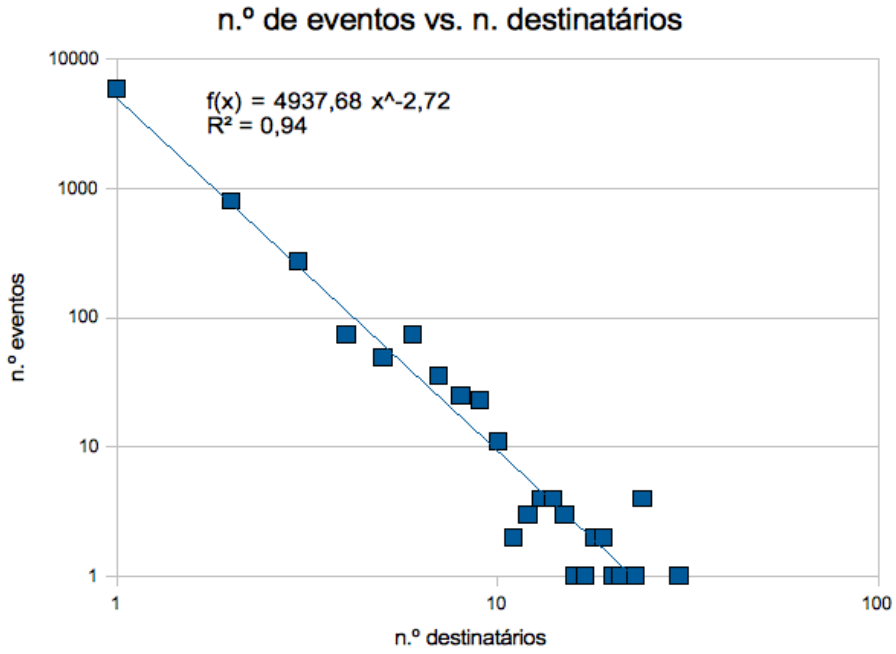

Figura 4.7: Distribuição do n.º de eventos em função do n.º de destinatários para a rede de professores

verificado na figura 4.7.

4.1.2.2 Rede de alunos

A rede de alunos apresenta muito menos utilizadores do que seria de esperar, atendendo a que se trata de uma população superior à dos professores. No entanto, a sua adesão ao serviço de correio electrónico é muito inferior, pelo que o volume de dados obtidos para a formação desta rede é muito menor. A justificação desta característica poderá passar pelo facto de os alunos já possuírem, à entrada na universidade, endereços de correio electrónico alternativos e não adoptarem por um endereço de correio electrónico novo.

Durante o período de análise apenas houve o estabelecimento de 1027 ligações entre 283 nós da rede de alunos.

Verificou-se a existência de 3 componentes nesta rede: um grupo principal englobando 279 nós e dois pares de utilizadores que apenas trocaram correio electrónico entre si e que estão desconectados dos outros dois componentes da rede.

A densidade da rede de alunos é de 2.57% e apresenta um *average degree* de 7.3.

A distribuição de *degree* desta rede é também do tipo exponêncial, como se pode verificar

na figura 4.8.

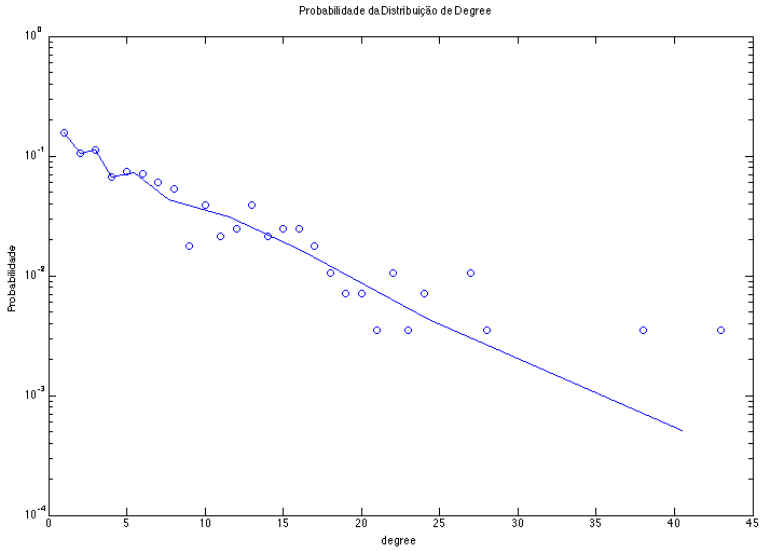

Figura 4.8: Distribuição do *degree* na rede de alunos

O diâmetro da rede de alunos é 10, sendo que se está a considerar aqui o caso do componente principal da rede e não os dois componentes isolados.

4.1.2.3 Rede de funcionários

A rede de funcionários do ISCTE foi das 3 redes a que apresentou maior percentagem de adesão ao sistema de correio electrónico (tabela 4.2), com 46.2% dos utilizadores registados no sistema Fenix a pertencerem à rede formada pela troca de mensagens entre funcionários da instituição.

A rede de funcionários do ISCTE é composta por 197 nós e 964 ligações agrupadas num único componente. A rede de funcionários do ISCTE apresenta um *average degree* de 9.8 e uma densidade de ligações de 4.99%, sendo das 3 redes a mais densa.

A distribuição de *degree* desta rede, ao contrário das outras duas, não parece apresentar uma distribuição exponêncial, como se pode ver na figura 4.9

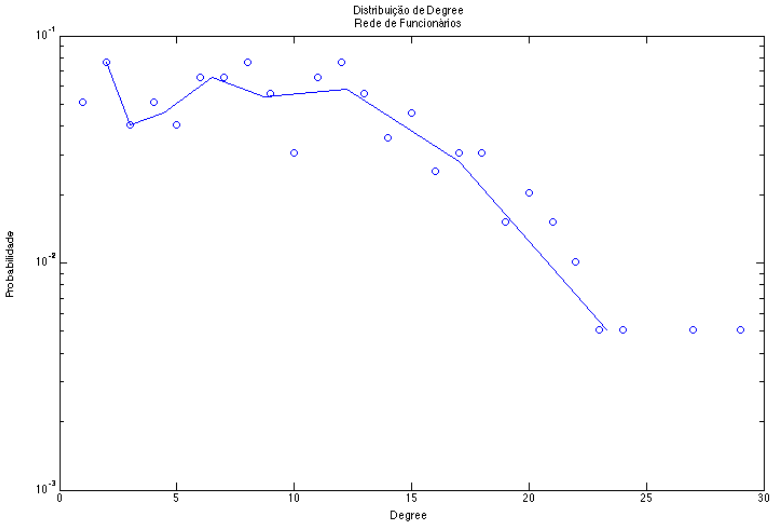

Figura 4.9: Distribuição do *degree* na rede de funcionários

4.2 Detecção de comunidades

Atendendo aos resultados da caracterização expostos na secção anterior, optámos por apenas efectuar o estudo dos procedimentos de detecção de comunidades à rede de professores, uma vez que para os funcionários do ISCTE não há um mapeamento nos dados do sistema Fenix que permita tirar conclusões sobre os grupos existentes e para o caso da rede de alunos verificou-se que estes apresentam uma adesão baixa ao sistema de correio electrónico disponibilizado pela instituição. No caso dos alunos os resultados sofreriam de uma amostragem muito grande, não sendo relevantes para inferir conclusões sobre a dinâmica de formação de redes informais de comunicação.

A rede de professores foi estudada sob dois algoritmos hierárquicos: o algoritmo Girvan-Newman e o algoritmo Clauset-Newman-Moore. Ambos são de caracter global sendo que o primeiro é do tipo divisivo e o segundo aglomerativo. Ainda foram estudadas duas estratégias de caracter local: a análise de *k-cores* para identificar a estrutura hierárquica da rede de professores e a percolação de cliques para identificar comunidades transversais aos diversos departamentos.

4.2.1 Algoritmo Grivan-Newman

O algoritmo de Girvan-Newman (ver secção 2.6.1.1) utiliza a noção de modularidade como medida de qualidade para determinar em que ponto do dendrograma se deve fazer o corte a fim de obter a divisão 'óptima' da rede de professores. O valor de modularidade máximo foi obtido para um valor de $Q = 0,588$, correspondendo a uma divisão de professores em 14 comunidades, como apresentado na tabela 4.3. O valor de modularidade $Q > 0,3$ é indicativo da existência de estrutura na rede de professores (Clauset *et al.*, 2004).

Tabela 4.3: Comunidades identificadas pelo algoritmo de Girvan-Newman

Grupo	1	Grupo	2	Grupo	3
DS	3	DMQ	54	DPSO	19
Outros	1	DCG	28	Outros	9
		Outros	12	DS	8
		DF	11	DE	1
		DC	5		
		DS	2		
		DPSO	1		
Total	4		113		37
Grupo	4	Grupo	5	Grupo	6
DCTI	83	DS	17	DE	43
DMQ	5	Outros	10	Outros	4
DS	1	SAD	5	SAD	2
Outros	2	DMQ	1	DC	1
Total	91		33		50
Grupo	7	Grupo	8	Grupo	9
DS	2	DA	24	SAAU	6
Outros	1	DH	17		
		SAAU	2		
		Outros	1		
Total	3		44		6
Grupo	10	Grupo	11	Grupo	12
DS	2	DF	1	ACEA	2
Outros	2				
Total	4		1		2
Grupo	13	Grupo	14		
SAAU	6	DS	1		
Total	6		1		

4.2.2 Algoritmo Clauset-Newman-Moore

O algoritmo de Clauset-Newman-Moore (ver secção 2.6.1.3) utiliza, da mesma forma que o de Girvan-Newman, a noção de modularidade como medida de qualidade para determinar em que ponto se obtém a divisão 'óptima' da rede de professores. O valor de modularidade máximo foi obtido para um valor de $Q = 0,585$, correspondendo a uma divisão de professores em 7 comunidades, como apresentado na tabela 4.4.

Tabela 4.4: Comunidades identificadas pelo algoritmo de Clauset-Newman-Moore

Grupo	1	Grupo	2	Grupo	3
DCTI	80	DPSO	19	DCG	25
DMQ	5	DCTI	3	SAD	5
DCG	2	DS	5	DF	7
Outros	2	DE	1	DMQ	3
		Outros	9	DC	5
				Outros	10
Total	89	Total	37	Total	55
Grupo	4	Grupo	5	Grupo	6
DMQ	52	DE	42	DH	17
DS	30	DC	1	DA	24
SAD	2	Outros	4	SAAU	2
DE	1			DF	5
DPSO	1			DS	1
ACEA	2			DCG	1
Outros	14			Outros	3
Total	102	Total	47	Total	53
Grupo	7				
SAAU	12				
Total	12				

4.2.3 k-cores

A análise da rede através dos *k-cores* (ver secção 2.3.9) permite caracterizar a rede para além da distribuição de *degree* e analisá-la em busca de hierarquias. Pela análise efectuada verifica-se a existência de nós de diversos departamentos em todos os valores de k-core mais baixos. No entanto, quando se analisam os valores de k mais elevados (8 e 9), verifica-se que os núcleos centrais da rede de professores são formados em torno de alguns departamentos específicos (DMQ, DCG, DA, DCTI e DE) colocando em evidência o facto de apresentarem números de ligações entre os seus membros muito elevados. Isto é particularmente evidente para o caso de $k = 9$, onde se verifica que o clique completo

é quase exclusivamente constituído por membros dos departamentos DCTI e DE (tabela 4.5). Uma das observações efectuadas no processo de decomposição em *k-cores* é a de que este não divide a rede em componentes separados. Pelo contrário, a cada nível extra de k que é removido, o núcleo central permanece único, não se dividindo nunca, o que evidência uma estrutura hierárquica forte. A tabela 4.5 representa as diferentes camadas que se vão retirando ao núcleo central de nós, sendo que este núcleo nunca se divide pela remoção das camadas exteriores.

Tabela 4.5: Distribuição de membros por *k-core*

	1	2	3	4	5	6	7	8	9
DCTI	9	3	7	9	5	2	1	6	43
DMQ	4	3		3	3	8	9	34	
DE	3	4	1	1		1	1	5	32
DS	9	4	8	3	3	8	7	5	1
DCG	2	1	3			1	7	19	
DPSO		1	1	1	2	3	9	7	1
DA			4	3		1	2	15	
DH	2	1	2	1	2	3	6		
SAAU	5	8		1					
DF	1		6	2				5	
SAD	1	2	1	2		2			
DC	3		2		2				
ACEA	1	1							1
Outros		1		3					

A distribuição de professores por diversas camadas dos *k-cores* não é homogénea: por exemplo os departamentos DE e DCTI apresentam uma percentagem alta de professores nos núcleo $k = 9$ enquanto outros departamentos, como DS, apresentam professores distribuídos ao longo de todas as camadas. Outros ainda, como o SAAU tem professores que se encontram apenas nos níveis mais baixos e portanto apresentam-se mais periféricos em termos hierárquicos.

4.2.4 Percolação de cliques

A maior desvantagem da utilização dos algoritmos globais tem a ver com a impossibilidade de contabilizar a sobreposição de comunidades. A utilização do método de percolação de cliques (ver secção 2.6.4) permite verificar a sobreposição de comunidades. A análise através deste método revelou que, utilizando os valores de $k =< 3, ..., 7 >$, para $k = 6, 7$ os dois módulos encontrados não possuem nenhuma sobreposição, estando isolados um do outro. Para os restantes valores de k o método detectou diferentes números de comunidades, sendo que em alguns casos comunidades aparecem isoladas, como no caso de $k = 3, 4$ e 5 em que apresentam sobreposição parcial de comunidades. Este fenómeno pode verificar-se nas figuras seguintes, que acompanham os resultados da identificação de comunidades.

4.2.4.1 Comunidades detectadas para $k = 3$

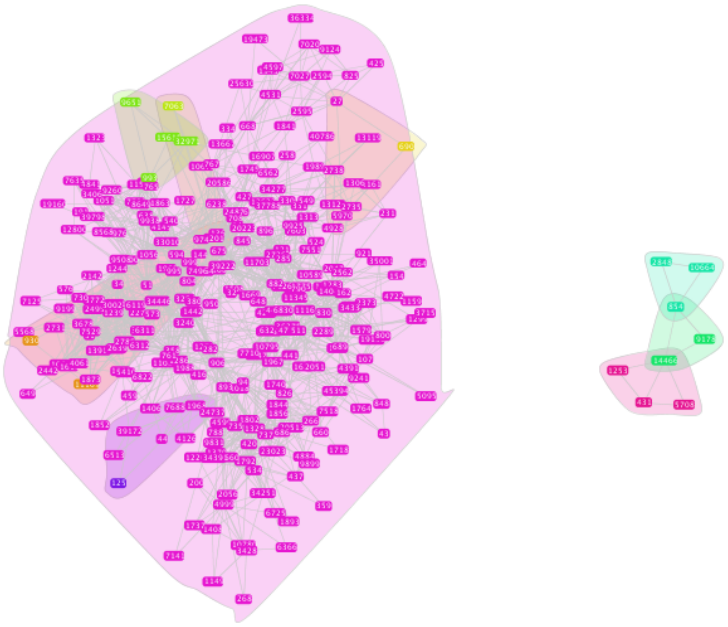

Figura 4.10: Distribuição por departamento: k=3

Neste caso verifica-se a existência de um grupo de elevadas dimensões que engloba membros de quase todos os departamentos (grupo 3), verificando-se a existência de vários grupos de pequenas dimensões constituídos normalmente por membros de 1 ou 2 departamentos.

Tabela 4.6: Percolação de Cliques $k = 3$

Grupo	Departamento	número elementos	Total
1	SAAU	3	3
2	DS	2	
	DMQ	1	3
3	DMQ	58	
	DCTI	50	
	DE	43	
	DCG	27	
	DS	24	
	DPSO	24	
	DA	17	
	DH	14	
	SAD	6	
	DF	5	
	DC	3	
	Outros	3	274
4	SAAU	4	4
5	DS	4	4
6	SAAU	3	3
7	DS	2	
	DMQ	1	3
8	DS	3	3
9	DF	2	
	DH	1	3
			300

4.2.4.2 Comunidades detectadas para $k = 4$

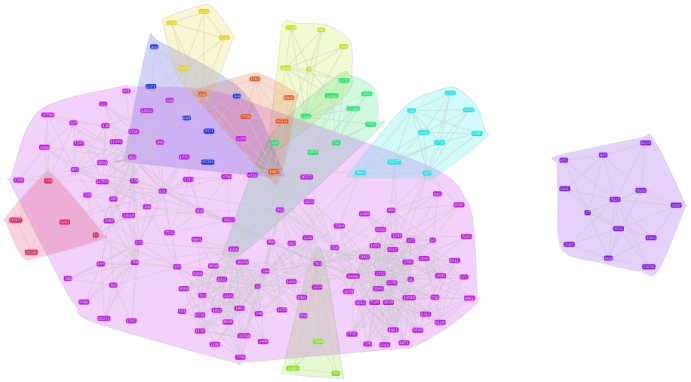

Figura 4.11: Distribuição por departamento: k=4

Tabela 4.7: Distribuição por departamento: k=4

Grupo	Departamento	número de elementos	Total
1	DH	12	
	DA	1	13
2	SAD	4	4
3	DMQ	5	5
4	DMQ	42	
	DE	39	
	DPSO	22	
	DS	7	
	DA	4	
	DF	3	
	DCTI	3	
	DCG	2	
	SAD	1	123
5	DCG	5	
	DCTI	3	
	DS	1	
	DMQ	1	10
6	DCG	5	
	DS	1	6
7	DCTI	9	9
8	DCTI	10	
	DPSO	1	11
9	DA	9	9
10	DCG	5	5
			195

A utilização do valor $k = 4$ coloca em evidência que as comunidades de comunicação

informal que se formam não se cingem ao interior dos departamentos do ISCTE. Embora apareçam alguns grupos constituídos unicamente por membros de um único departamento (grupos 2,3,7,9 e 10), verifica-se que os restantes grupos apresentam membros de diversos departamentos. Inversamente, é também evidente que os diversos departamentos têm os seus membros distribuídos por diversos grupos.

4.2.4.3 Comunidades detectadas para $k = 5$

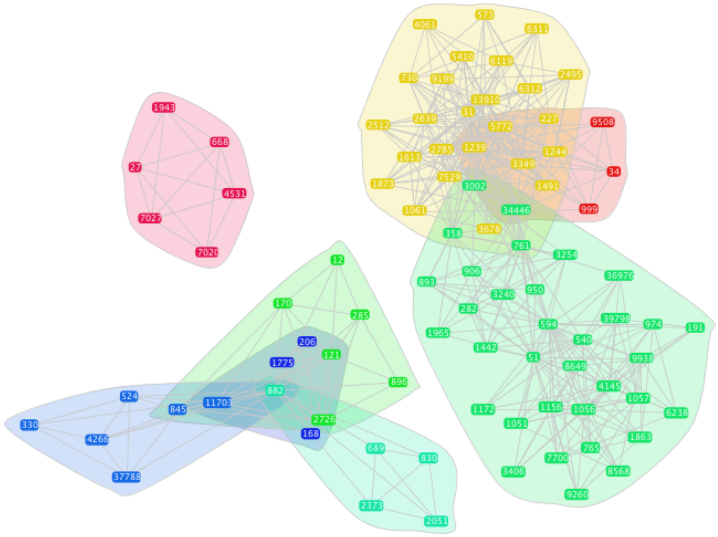

Figura 4.12: Distribuição por departamento: k=5

O método de detecção de comunidades por percolação de cliques mostra que há grupos constituídos unicamente por elementos de um único departamento (2,4,6,7 e 8), assim como grupos transversais a vários departamentos (1,3 e 5), embora normalmente haja um departamento dominante. Tal é facilmente identificável no caso do grupo 1, onde o departamento DPSO é maioritário, assim como no Grupo 5 que é constituído por professores do departamento DE à excepção de um elemento. Verifica-se igualmente que alguns departamentos aparecem em vários grupos, o que significa que as ligações informais extravasam claramente a organização institucional.

Tabela 4.8: Distribuição por departamento: k=5

Grupo	Departamento	número de elementos	Total
1	DPSO	21	
	DS	4	
	DCTI	3	
	DE	3	
	SAD	1	
	DMQ	1	
	DA	1	34
2	DMQ	7	7
3	DCG	3	
	DCTI	1	
	DMQ	1	5
4	DH	6	6
5	DE	28	
	SAD	1	29
6	DE	9	9
7	DMQ	9	9
8	DMQ	6	6
			105

4.2.4.4 Comunidades detectadas para $k = 6$

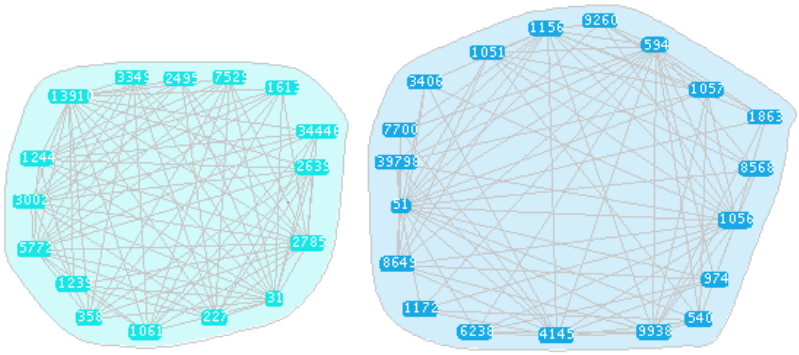

Figura 4.13: Distribuição por departamento: k=6

Tabela 4.9: Distribuição por departamento: k=6

Grupo	Departamento	número de elementos	Total
1	DPSO	17	
	DMQ	1	
	DCTI	1 19	
2	DE	16	16
			35

Neste caso verifica-se que o Grupo 1 é constituído maioritariamente por elementos do departamento DPSO, mas que mesmo assim a comunidade atravessa as fronteiras do departamento e inclui também membros dos departamentos DMQ e DCTI.

4.2.4.5 Comunidades detectadas para $k = 7$

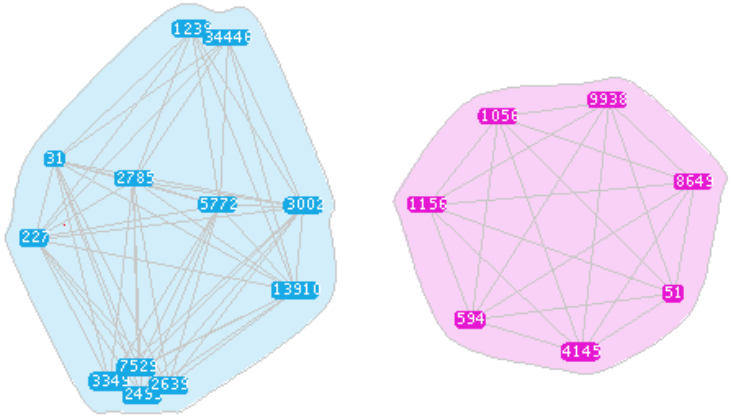

Figura 4.14: Distribuição por departamento: k=7

Tabela 4.10: Distribuição por departamento: k=7

Grupo	Departamento	número de elementos	Total
1	DPSO	7	7
2	DE	12	12
			19

Neste caso, verifica-se que apenas dois departamentos possuem um elevado número de ligações internas correspondentes aos dois grupos detectados.

4.2.4.6 Conclusão sobre a análise por *k-cores*

As tabelas 4.6, 4.7, 4.8, 4.9 e 4.10 colocam em evidência que a aplicação deste método depende da escolha do valor de k, uma vez que este influencia de forma decisiva o resultado do método de percolação de cliques. Para valores altos de k ($k = 6, 7$), o número de comunidades encontradas é de apenas duas, sendo que apenas uma fracção muito baixa dos professores do ISCTE está presente nestas comunidades altamente conectadas. Por outro lado, com $k = 3$ surge uma comunidade gigante (grupo 3 da tabela 4.6) que engloba a quase totalidade dos nós da rede (total = 395). Os valores intermédios são naturalmente aqueles que são mais interessantes do ponto de vista prático, uma vez que permitem verificar a formação de redes informais para além dos departamentos em causa.

4.3 Comparação dos resultados

Para a comparação dos resultados obtidos, convém primeiro definir o particionamento de base em relação ao qual se efectuam as comparações. Este particionamento é o definido pela distribuição da rede de professores pelos respectivos departamentos.

Tabela 4.11: Distribuição dos professores por departamento do ISCTE

Departamento	n.º professores	percentagem
DCTI	83	21,01%
DMQ	60	15,19%
DE	44	11,14%
DS	36	9,11%
DCG	28	7,09%
DA	24	6,08%
DPSO	20	5,06%
DH	17	4,30%
SAAU	14	3,54%
DF	12	3,04%
SAD	7	1,77%
DC	6	1,52%
ACEA	2	0,51%
não ident.	42	10,63%
Total	395	100,00%

A distribuição do número de professores professores apresenta um decaimento exponencial como pode ser verificado pela figura 4.15. Os dados da tabela 4.11 são a base das comunidades institucionais definidas pela hierarquia do ISCTE, tal como foram obtidas do sistema Fenix. A seguir faremos a comparação dos diversos métodos tendo em atenção esta base.

A comparação dos diversos algoritmos recorre à construção de matrizes de associação, também conhecidas como matrizes de confusão ou tabelas de contigência, que servem de base aos cálculos efectuados na determinação da variação de informação de diferentes algoritmos (Meilă, 2007).

Tendo dois particionamentos $C =< C_1, ..., C_k >$ e $C' =< C'_1, ..., C'_{k'} >$, a matriz de associação é uma matriz $k \times k'$ tal que o elemento kk' representa o número de pontos na intersecção das partições C_k de C e $C'_{k'}$ de C'

$$n_{kk'} = |C_k \cap C'_{k'}| \tag{4.1}$$

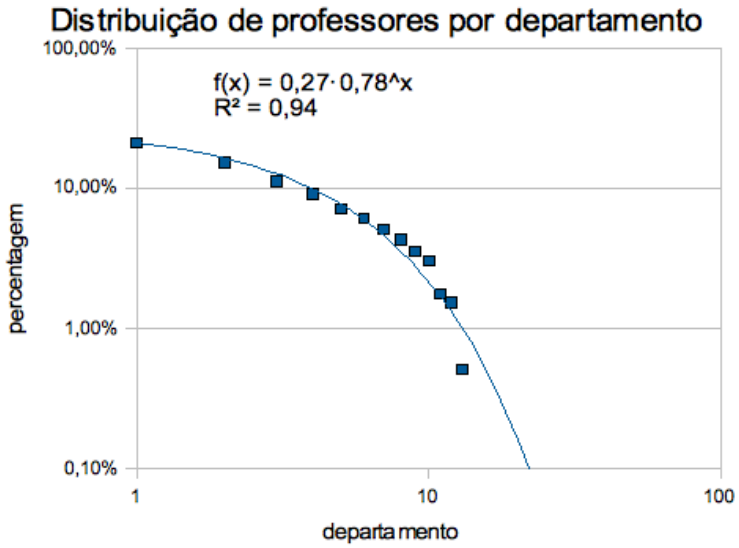

Figura 4.15: Distribuição do número de professores por departamento.

A definição da variação de informação para comparação de dois particionamentos é calculada estabelecendo primeiramente o valor da informação existente em cada um dos particionamentos e o valor da informação que um particionamento tem sobre o outro particionamento.

Considerando um dado particionamento, a probabilidade de que nó k pertença a uma determinada partição C_k é dada pela equação 4.2, onde n_k é o número de elementos na partição C_k e n o número total de elementos:

$$P(k) = \frac{n_k}{n} \tag{4.2}$$

A incerteza associada a esta medida será dada pela entropia da variável ($P(k)$), dada por 4.3

$$H(C) = -\sum_{k=1}^{K} P(k)\log(P(k)) \tag{4.3}$$

sendo $H(C)$ a entropia associada ao particionamento C. Este valor é sempre não-negativo e assume o valor zero apenas quando não há incerteza, nomeadamente quando o número de partições num determinado particionamento é 1.

A definição da informação mútua ($I(C, C')$) entre dois particionamentos, ou seja a informação que um particionamento tem sobre o outro, é dada pela probabilidade $P(k, k')$

que representa a probabilidade de um ponto pertencente à partição C_k se encontrar na partição $C'_{k'}$. Esta probabilidade é dada pela equação 4.3.

$$P(k,k') = \frac{|C_k \cap C'_{k'}|}{n} \tag{4.4}$$

Assim, define-se a informação mútua ($I(C,C')$) como sendo igual à informação mútua associada a duas variáveis aleatórias:

$$I(C,C') = \sum_{k=1}^{K}\sum_{k'=1}^{K'} P(k,k') \log \frac{P(k,k')}{P(k)P'(k')} \tag{4.5}$$

Meilă (2007) propõe que o critério utilizado para comparação de dois particionamentos seja então a quantidade *variação de informação*, VI, tal que:

$$VI(C,C') = H(C) + H(C') - 2I(C,C') \tag{4.6}$$

A medida VI é efectivamente uma métrica, uma vez que é sempre não-negativa, é simétrica e apresenta desigualdade triangular.

A variação de informação pode ainda ser normalizada (V e V_{k^*}) caso se pretendam obter distâncias entre 0 e 1, assumindo então a forma da equação 4.7 no caso da normalização ser feita quando o conjunto de dados é o mesmo, ou a forma da equação 4.8 quando o número de partições é o mesmo (K^*):

$$V(C,C') = \frac{1}{\log n} VI(C,C') \tag{4.7}$$

$$V_{k^*}(C,C') = \frac{1}{2\log K^*} VI(C,C') \tag{4.8}$$

4.3.1 Algoritmos hierárquicos

4.3.1.1 Girvan-Newman

O algoritmo detectou as comunidades apresentadas na tabela 4.3. A matriz de associação entre este algoritmo e os departamentos existentes no ISCTE da tabela 4.11 é dada pela matriz da tabela 4.12:

Tabela 4.12: Matriz de associação entre o algoritmo de Grivan-Newman e os departamentos do ISCTE

0	0	1	3	0	0	0	0	0	0	0	0	0	0
0	0	13	2	1	54	0	11	0	0	28	5	0	0
0	0	9	8	19	0	0	0	1	0	0	0	0	0
0	0	1	1	0	5	0	0	0	83	0	0	0	0
5	0	10	17	0	1	0	0	0	0	0	0	0	0
2	0	4	0	0	0	0	0	43	0	0	1	0	0
0	0	1	2	0	0	0	0	0	0	0	0	0	0
0	2	1	0	0	0	17	0	0	0	0	0	24	0
0	6	0	0	0	0	0	0	0	0	0	0	0	0
0	0	2	2	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	2
0	6	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0	0

Utilizando a matriz de associação para calcular o valor da distância de informação obtemos então os resultados da tabela 4.13:

Tabela 4.13: Variação de informação entre o algoritmo de Grivan-Newman e os departamentos do ISCTE

H(ISCTE)	2.346
H(Girvan-Newman)	1.945
I	1.474
VI	1.342
V	0.224

Por outro lado, é possível comparar este valor com uma situação de um hipotético algoritmo de particionamento aleatório. Nesse caso a probabilidade de um determinado elemento pertencer a uma determinada partição seria de $1/n_k$. Mantendo as dimensões do particionamento do algoritmo de Girvan-Newman e a distribuição dos professores pelos departamentos teríamos para esta situação um valor de variação de informação normalizada $V = 0.833$.

4.3.1.2 Clauset-Newman-Moore

O algoritmo detectou as comunidades apresentadas na tabela 4.4. A matriz de associação entre este algoritmo e os departamentos existentes no ISCTE da tabela 4.11 é dada pela matriz da tabela 4.14:

Tabela 4.14: Matriz de associação entre o algoritmo de Clauset-Newman-Moore e os departamentos do ISCTE

0	0	2	0	0	5	0	0	0	80	2	0	0	0
0	0	9	5	19	0	0	0	1	3	0	0	0	0
5	0	10	0	0	3	0	7	0	0	25	5	0	0
2	0	14	30	1	52	0	0	1	0	0	0	0	2
0	0	4	0	0	0	0	0	42	0	0	1	0	0
0	2	3	1	0	0	17	5	0	0	1	0	24	0
0	12	0	0	0	0	0	0	0	0	0	0	0	0

Utilizando a matriz de associação para calcular o valor da distância de informação obtemos então os resultados da tabela 4.15:

Tabela 4.15: Variação de informação entre o algoritmo de Clauset-Newman-Moore e os departamentos do ISCTE

H(ISCTE)	2.346
H(Clauset-Newman-Moore)	1.811
I	1.372
VI	1.413
V	0.236

Também é possível comparar este valor com uma situação de um hipotético algoritmo de particionamento aleatório. Mantendo as dimensões do particionamento do algoritmo de Clauset-Newman-Moore e a distribuição dos professores pelos departamentos teríamos para esta situação um valor de variação de informação normalizada $V = 0.718$.

4.3.1.3 Girvan-Newman vs. Clauset-Newman-Moore

Da mesma forma que se pode calcular a variação de informação entre um particionamento e o particionamento base do ISCTE, também se pode fazer o mesmo cálculo comparando algoritmos directamente. Assim, comparando os dois algoritmos anteriores obtém-se a matriz de associação 4.16.

Tabela 4.16: Matriz de associação entre os algoritmos de Girvan-Newman e Clauset-Newman-Moore

0	0	0	4	0	0	0
2	0	48	57	0	7	0
0	33	0	4	0	0	0
86	4	0	0	0	0	0
0	0	7	26	0	0	0
0	0	0	3	47	0	0
0	0	0	3	0	0	0
0	0	0	0	0	44	0
0	0	0	0	0	0	6
0	0	0	4	0	0	0
0	0	0	0	0	1	0
0	0	0	2	0	0	0
0	0	0	0	0	0	6
0	0	0	0	0	1	0

Utilizando a matriz de associação para calcular o valor da distância de informação obtemos então os resultados da tabela 4.17:

Tabela 4.17: Variação de informação entre os algoritmos de Girvan-Newman e Clauset-Newman-Moore

H(Girvan-Neman)	1.945
H(Clauset-Newman-Moore)	1.811
I	1.390
VI	0.976
V	0.163

4.3.2 *k*-core

A análise de *k-cores* da rede de professores do ISCTE revela a seguinte composição de cada um dos *k-cores*:

Verificou-se que para todos os *k-cores* a rede nunca se dividiu em vários componentes, mostrando que a cada nível hierárquico há sempre um núcleo coeso que não é dividido

Tabela 4.18: N.º de elementos de cada *k-core*

k-core	n.º elementos	Total do k-core
1	40	395
2	29	355
3	35	326
4	29	291
5	17	262
6	29	245
7	42	216
8	96	174
9	78	78

pelo aumento do valor de k na análise de *k-cores*.

4.3.3 Percolação de cliques

A percolação de cliques mostrou que o componente gigante da rede inicial sofre divisões em diversos grupos isolados para todos os valores de k. Para além disso, para valores de $k = 3, 4, 5$ os componentes isolados são constituídos por subgrupos que se sobrepõem. Ao representar para cada um destes sub-componentes, obtêm-se as relações existentes entre comunidades através da sobreposição de determinados elementos.

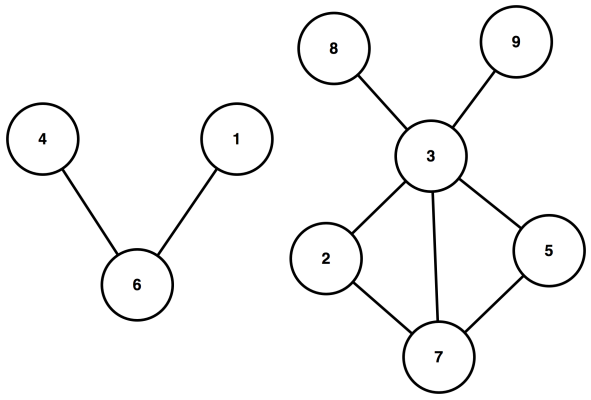

Figura 4.16: Ligações entre Comunidades: $k = 3$

Nestes diagramas (figuras 4.16, 4.17 e 4.18) não foram representadas as comunidades que não estabelecem nenhuma ligação com outras comunidades e portanto não apresentam sobreposição, tal como acontece para $k = 6, 7$. Assim, na figura 4.17 não foi representado o grupo 1 e na figura 4.18 não foi representado o grupo 4.

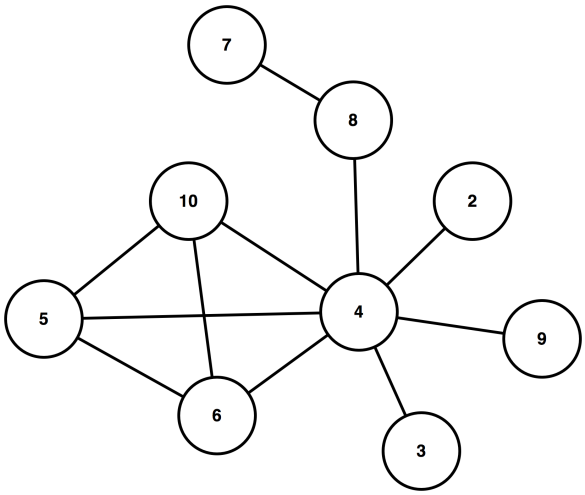

Figura 4.17: Ligações entre Comunidades: $k = 4$

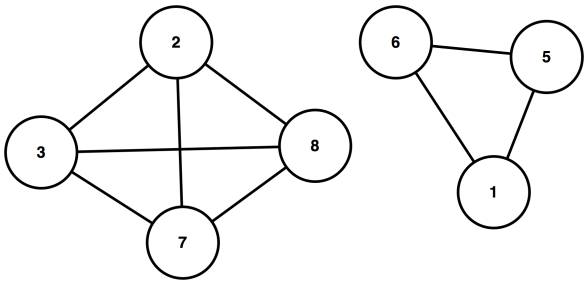

Figura 4.18: Ligações entre Comunidades: $k = 5$

4.3.4 Resumo

As diversas abordagens para a identificação de comunidades produzem resultados diferentes. Enquanto as abordagens globais não eliminam elementos das comunidades identificadas, as abordagens locais procuram, pelo contrário, encontrar núcleos fortemente conectados, mesmo que tal implique afastar desses núcleos os elementos periféricos. Resumindo as principais características encontradas na tabela 4.19:

As principais conclusões que se tiram dos diversos métodos, relativamente aos diferentes resultados que apresentam, são naturalmente ligadas à forma e objectivos de cada tipo de algoritmo. Enquanto os algoritmos de detecção globais baseados na densidade de ligações entre comunidades conseguem fazer a classificação de todos os nós presentes na rede, os algoritmos locais não o fazem[2]. Os algoritmos globais conseguem dividir a rede em grupos

[2]O caso do *k-core* com $k = 1$ é incluído na tabela 4.19 mas na prática nunca é aplicado, uma vez que $k = 1$ corresponde à eliminação da rede inicial de todos os elementos que tenham *degree*=0. Como não há elementos isolados na rede de professores $k = 1$ corresponde à rede total.

Tabela 4.19: Características dos resultados obtidos

Método	Âmbito	Componentes	Grupos	N.º classificados
Girvan-Newman	Global	14	14	395
Clauset-Newman-Moore	Global	7	7	395
k-core	Local			
$k = 1$		1	-	395
$k = 2$		1	-	355
$k = 3$		1	-	326
$k = 4$		1	-	291
$k = 5$		1	-	262
$k = 6$		1	-	245
$k = 7$		1	-	216
$k = 8$		1	-	174
$k = 9$		1	-	78
Percolação de cliques	Local			
$k = 3$		2	9	300
$k = 4$		2	10	195
$k = 5$		3	8	105
$k = 6$		2	2	35
$k = 7$		2	2	19

de forma a que se possa rapidamente definir grupos para análise detalhada, sendo que não permitem a sobreposição de grupos. Um elemento não pode pertencer a mais que um grupo em simultâneo. Por outro lado, a análise através da percolação de cliques permite garantir que todos os elementos dentro dessa comunidade estão fortemente conectados e podem participar em outras comunidades, permitindo a sobreposição de comunidades. O método de $k-cores$ não permite identificar comunidades, a menos que o componente em análise seja subdividido pelo incremento de k, uma vez que o seu objectivo é sobretudo fazer uma análise vertical ao longo dos diversos *k-cores*. Embora a comparação com os dados reais da análise de $k-cores$ mostre que para valores de k elevados, estes núcleos são constituídos maioritariamente por elementos de alguns departamentos, isto não mostra por si só a estrutura de interligação desses grupos, mostrando antes a colocação hierárquica dos elementos desses grupos no total da rede, em termos de *degree*.

A análise das comunidades do ISCTE, revelou diversas propriedades interessantes. No entanto, a análise não mostra a dependência das comunidades formadas com algumas propriedades. Na próxima secção, desenvolvemos um modelo de agentes para estudar o influência da noção de 'vizinhança social' na formação de comunidades de comunicação informal.

4.4 Modelação do sistema de correio electrónico

A modelação da rede social emergente da troca de mensagens entre professores do ISCTE pretendeu criar uma reprodução *in silico* dos eventos verificados ao longo do período em estudo, por forma a verificar e testar algumas ideias quanto à formação destas redes informais de comunicação.

O desenvolvimento de simulações baseadas em agentes tem vindo a despertar interesse em anos recentes. Um dos aspectos onde tem surgido interesse é o da modelação de dados reais, incluindo a interacção dos agentes com o mundo exterior (Janssen e Ostrom, 2006), tendo surgido desenvolvimentos na produção de linguagens de programação que permitam a exploração de ambientes externos à simulação por parte dos agentes (Dastani, 2008).

Ao invés de fazer uma simulação baseada em agentes, autónoma da realidade, procurámos averiguar a possibilidade de incluir dados reais na simulação a fim de treinar o modelo, de forma a que este possa de alguma forma criar uma ontologia do processo e depois evoluir de acordo com a ontologia desenvolvida. A noção de ontologia é aqui definida em termos globais pelo sistema de agentes e não concretizada no conhecimento que cada agente tem do sistema. Estamos interessados no comportamento do sistema como um todo e não nos comportamentos de cada agente do modelo.

Os agentes construídos neste modelo não são efectivamente capazes de aprendizagem, no entanto a fase de treino permite alimentar o modelo com dados que tornarão os resultados da simulação mais próximos da realidade. Um dos problemas desta abordagem é o de distinguir se estamos efectivamente a treinar o sistema para melhorar resultados ou se apenas nos limitamos a reproduzir a realidade. É preciso encontrar um equilíbrio entre a liberdade completa do sistema, que produziria uma situação aleatória e um número de restrições onde a simulação apenas decalca o que existe nos ficheiros de treino. Algures entre estes dois extremos é onde a simulação se torna interessante e onde se poderá detectar a emergência de comportamentos novos, ou a validação de outros que não tenham sido incluídos na fase de treino mas que tenham efectivamente sido observados nos dados reais.

O volume de dados proveniente da rede informal criada com a troca de mensagens de correio electrónico entre professores do ISCTE foi dividido em dois grupos. O primeiro serve de treino para a simulação e o segundo é utilizado para validação e comparação de resultados da simulação. O modelo de comunicação informal entre utilizadores de correio electrónico universitário (CIUCEU) criado é descrito seguidamente em detalhe.

4.4.1 O modelo CIUCEU

O modelo de agentes foi desenvolvido utilizando a plataforma de simulação multi-agente MASON (Luke *et al.*, 2009). Foram definidos 3 tipos de agentes, um representando a entidade social (o Professor) e 2 tipos de agentes de controlo, treino e monitorização, que permitem que a cada passo da simulação se possa fazer o treino do sistema, gerar eventos, recolher dados ou exportar resultados para posterior tratamento.

Atendendo a que a recolha de dados efectuada não é uma figura estática de uma realidade, mas antes um processo de aquisição contínuo a partir do qual procuramos abstrair uma realidade, convém ter em atenção a dinâmica dos próprios dados no processo de modelação. Ao analisarmos os dados reais, verificamos que a rede de professores teve 7376 eventos ao longo dos 62 dias. Se olharmos para estes dados como um objecto global podemos ser tentados a dizer, por exemplo, que o *average degree* é de 10. Partindo do princípio que há um limite para o valor do *degree* de cada utilizador, será de esperar que o valor da média convirja para um patamar no qual se manterá ao longo do tempo. No entanto, este patamar só poderá ser observado se o período de observações se estender por um período considerável de tempo. Como esse não é o caso do nosso modelo não podemos dizer que cada agente terá 10 contactos em média. Se observarmos a dinâmica da média verificamos que nos dados reais a taxa de crescimento do *average degree* é de $0,11\%$ a cada evento. Assim, para que o modelo incorpore dados reais, tem que ter em conta a dinâmica destes dados, aplicando uma taxa de crescimento semelhante (no caso metade do valor anterior, uma vez que estamos a considerar ligações não direccionadas no modelo).

O modelo corre em duas fases distintas: regime de treino e regime livre, como exemplificado na figura 4.19. Inicialmente, os agentes não possuem quaisquer contactos na sua agenda e a probabilidade de contactar alguém é zero. Com o decorrer do treino, cada agente vai efectuando os contactos que lhe são indicados pela realidade e desta forma vão preenchendo a sua agenda de contactos com pessoas que poderão vir a contactar. A probabilidade de contacto é definida pelo número de contactos prévios. Na segunda fase, a de regime livre, um evento é despoletado a cada passo da simulação e o agente para quem esse evento é despoletado envia uma mensagem de correio electrónico a alguém da sua agenda de contactos. Nesta situação, como os agentes não conhecem mais ninguém fora da sua agenda, o *average degree* não aumenta. Isto não seria problemático se se tivesse verificado que na realidade o valor do número de contactos das pessoas já tinha atingido um patamar. No entanto, tal não é verdade e portanto é necessário incluir um mecanismo de aumento do número de contactos, de forma a aproximar o modelo dos dados reais.

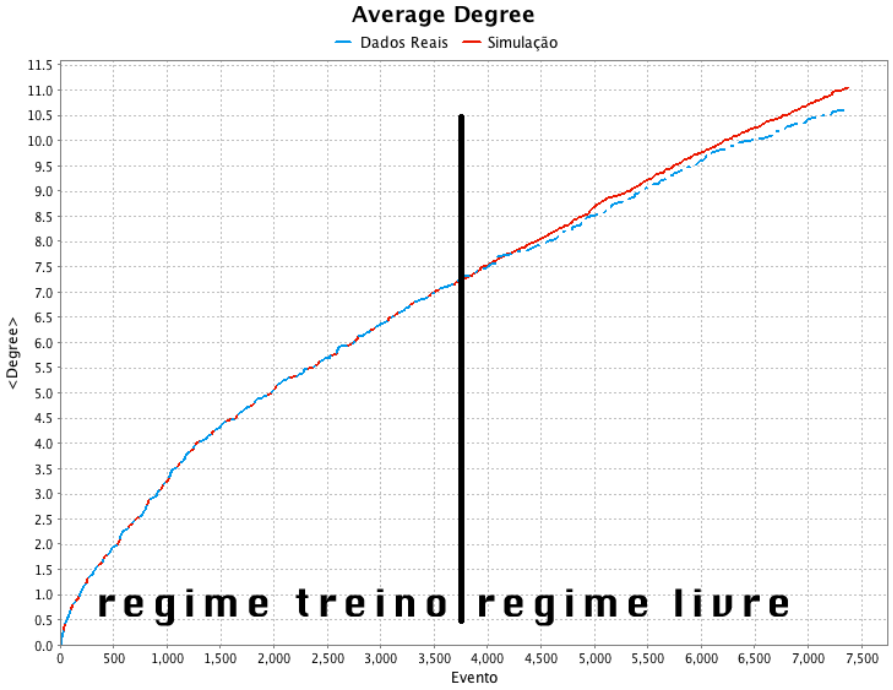

Figura 4.19: Exemplo da evolução do *degree* da rede de professores para 50% de treino.

Esse mecanismo passa pela definição de uma probabilidade p_r que define uma taxa de crescimento residual do sistema e que permite que durante o regime livre a simulação se adeqúe à dinâmica do modelo. Para além desta probabilidade é ainda necessário que a simulação gere novos contactos de forma a apresentar o efeito de *assortative mixing* e a noção de conhecimento local da realidade por parte do agente.

Seguidamente o CIUCEU é descrito em mais detalhe, apresentando o propósito, as variáveis de estado e escalas, a visão do processo e a calendarização de eventos, os conceitos de design, a inicialização, os dados de entrada ou treino e os submodelos incluídos no CIUCEU.

4.4.1.1 Propósito

O propósito da simulação é o de averiguar a influência da noção de vizinhança "social" no estabelecimento de redes de comunicação informais baseadas no sistema de correio electrónico. O modelo procura também averiguar até que ponto é possível capturar informação sobre a estrutura de comunidades utilizando dados de treino reais, quer para aplicação em problemas de amostragem, quer para aplicação em casos preditivos.

4.4.1.2 Variáveis de estado e escalas

Tabela 4.20: Parâmetros do modelo, descrição e valores utilizados.

Parâmetro	Descrição	Valor
Vizinhanca	Distância utilizada para a transitividade local	2
ProfDepart	Ficheiro com o mapeamento entre o ID número dos agentes e o existente nos departamento dos dados reais	<ficheiro txt>
Ficheiro	Dados de Treino com os eventos colocados um por cada linha do ficheiro na forma *remetente dest1 dest2 ... destN*	<ficheiro txt>
Treino	Coloca o modelo em modo de treino ou corrida Livre	Activado
PercMutation	Probabilidade de Eventos Aleatórios	4.0×10^{-4}
PercTreino	Fracção dos dados reais introduzida utilizada no Ficheiro de Treino	$[0, 1]$
NumDepartamentos	Número de departamentos para modelos aleatórios	15
NumProfessores	Indicação do número de Agentes a criar em modelos aleatórios	395

Agentes - O sistema é composto por agentes que representam os professores integrados no sistema de correio electrónico. A cada professor é atribuído um departamento. Esta atribuição pode ser feita de forma aleatória ou, caso estejam a ser utilizados dados reais para treino, de acordo com a distribuição de professores existentes no conjunto de dados de treino.

Tempo - O tempo da simulação não é um tempo continuo, mas antes um tempo discreto. Cada momento do tempo corresponde a um evento de envio de uma mensagem de correio electrónico para um ou mais destinatários. Não é necessário ter uma descrição probabilística da ocorrência de eventos ao longo do período em análise. Para determinar o tempo total da simulação (número de eventos) são utilizados os dados de treino que permitem calcular o tempo da corrida completa. Isto significa que, para um conjunto de dados reais que tenham ocorrido durante o período de 2 meses, estes são transformados numa sequência de eventos e o número de eventos ocorridos é então considerado para definir o tempo total (número de *steps*) da simulação.

Gerador de eventos - Este sistema controla o processo de treino dos agentes sendo responsável por colocar a simulação em modo de treino ou de corrida livre. Em cada *time step* o gerador de eventos obriga o agente a reproduzir o evento dos dados de treino,

se em regime de treino, ou é responsável por gerar oportunidades de eventos aleatórios para algum agente da simulação, se em regime livre. Em cada *time step* só acontece um evento, pelo que a decisão sobre o seu acontecimento não pode pertencer ao agente mas antes tem que ser supervisionada pelo gerador de eventos.

4.4.1.3 Visão do processo e calendarização de eventos

O modelo proposto assume que a manutenção das redes de relações sociais tem um custo, o que implica uma limitação do número de relações que são estabelecidas pelos agentes do modelo. Esta limitação provêm da noção de que, com o passar do tempo, a falta de reforço das relações sociais fará deteriorar a sua qualidade, fazendo eventualmente com que relações se desvaneçam. No entanto, para efeitos práticos deste caso de estudo, em que se pretende que o modelo seja comparado com os dados reais de 2 meses, nos dados reais a taxa de diminuição das relações é suficientemente baixa para assumir que a sua influência na estrutura final seja diminuta. Por outro lado, o modelo inclui a noção de que as ligações se efectuam preferencialmente entre pessoas que apresentam homofilia. Esta característica é evidenciada através da geração de novas ligações, de acordo com um submodelo de *assortative mixing* (ver ponto 4.4.1.7).

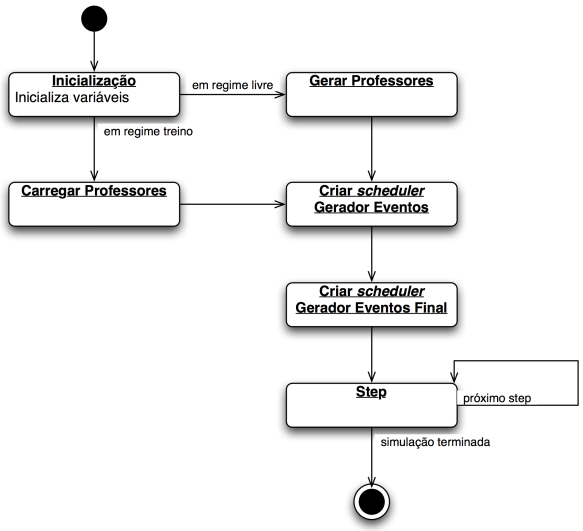

Figura 4.20: Diagrama de estados geral da simulação.

O funcionamento do modelo, exemplificado na figura 4.20, passa por uma fase de inicalização de variáveis, verificando em seguida se o utilizador pretende utilizar ou não dados de treino nessa corrida. Em caso positivo, os dados respeitantes aos professores são carregados e o processo passa à criação da calendarização da simulação (*scheduler*). Caso

contrário, os professores serão criados aleatoriamente de acordo com os valores introduzidos pelo utilizador. Depois de criadas as instâncias do modelo o controlo é passado ao sistema de calendarização que iterativamente faz a simulação avançar até ao seu ponto final ou até ser parada pelo utilizador.

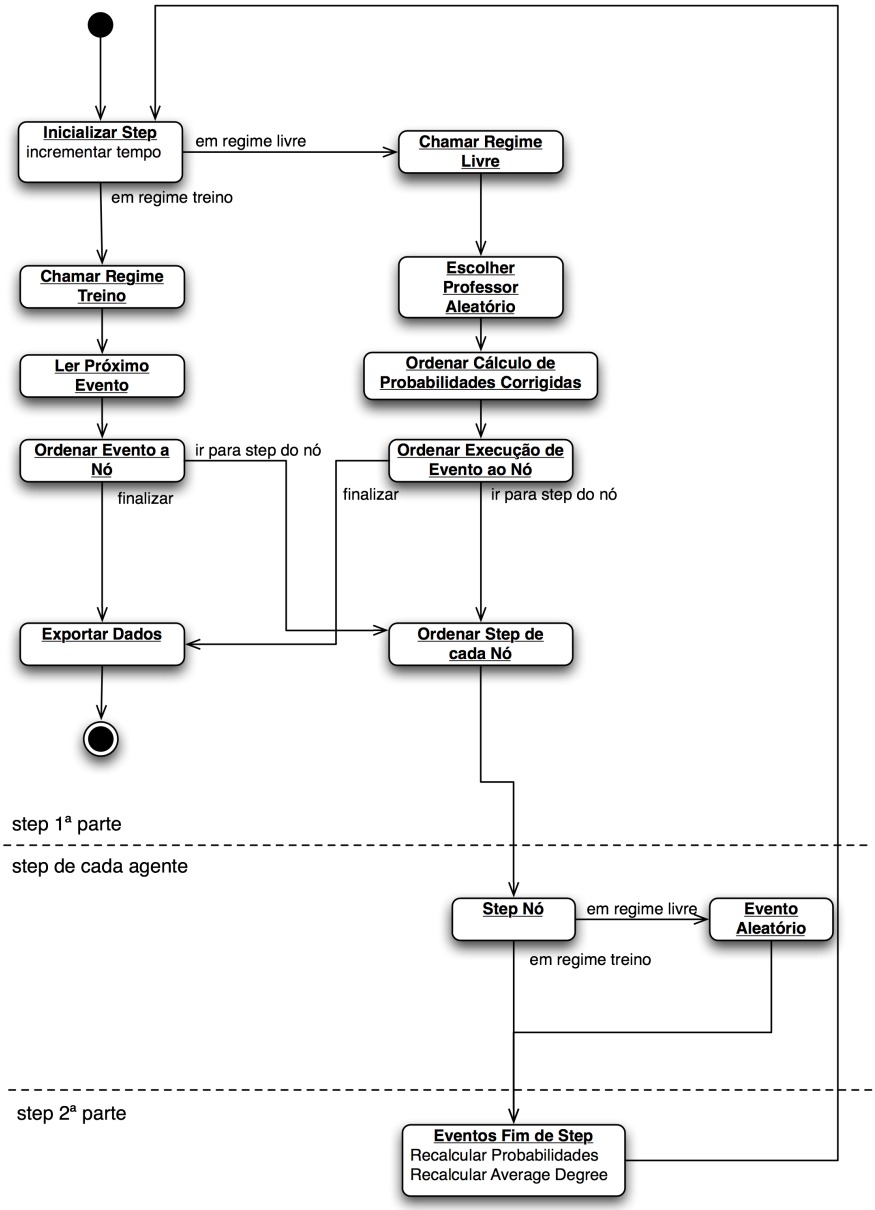

Figura 4.21: Diagrama de estados da calendarização de eventos (*scheduler*).

O processo de calendarização está definido em 3 níveis, como se pode observar na figura 4.21[3]. A calendarização de eventos decorre discretamente, sendo o tempo uma unidade

[3]O diagrama 4.21 não respeita de forma rigorosa a sintaxe UML, pois para representar as diferentes etapas de cada step do *schelduler* o diagrama de estados foi dividido em 3 zonas.

abstracta contabilizada através do acontecimento de eventos. Assumindo que não há eventos simultâneos, *i.e.* que o envio de mensagens é efectuado de forma sequencial através do servidor de correio electrónico, cada ciclo de tempo de simulação corresponde a uma oportunidade de envio de mensagem entre utilizadores. Em cada *step* o utilizador escolhido pode portanto efectuar o envio de uma mensagem para um ou vários recipientes. A sequência de cada *step* da simulação começa então por verificar se a simulação se encontra em regime de treino ou em regime livre. Caso esteja em regime de treino, o evento seguinte dos dados de treino é lido e a sua execução é cumprida pelo professor em causa. Caso esteja em regime livre, é escolhido um professor aleatoriamente para efectuar o envio de uma mensagem, de acordo com o modelo matemático descrito na parte submodelos. Após este início do *step*, cada um dos professores verifica se está em treino ou não. Se não estiver, procederá à geração de um evento aleatório, de acordo com o submodelo "Eventos Aleatórios" descrito mais à frente. Depois de todos os professores terem procedido ao seu *step* interno, o controlo do *step* volta ao ciclo anterior para ser finalizado com o cálculo de estatísticas e de probabilidades que dependem de todos os nós terem já efectuado os seus *steps* internos. O processo recomeça novamente do início para mais um *step*, até que a simulação termine ou seja interrompida pelo utilizador.

4.4.1.4 Conceitos de design

Evento - Atendendo à estrutura dos dados, a definição do que é um evento corresponde ao envio de uma mensagem de correio electrónico. Esta mensagem tem apenas um remetente e um ou mais destinatários.

Treino - O CIUCEU permite incluir dados reais por forma a treinar os agentes e assim melhorar o seu comportamento futuro. Em vez de partir de uma distribuição de probabilidades homogénea, em que cada agente pode estabelecer um evento com qualquer outro agente, a distribuição de probabilidade é efectivamente calculada a partir dos dados reais. Para isso estes foram colocados sob a forma de eventos sequenciais onde cada linha do ficheiro de treino corresponde ao envio de uma mensagem de correio electrónico. Depois, durante a execução da simulação o investigador pode escolher qual a percentagem dos dados que serão carregados pelo modelo e a simulação correrá a partir desse ponto em regime livre.

Emergência - A noção de emergência no CIUCEU encontra-se no surgimento de redes com estrutura, isto é, na formação de comunidades cuja densidade de ligações internas seja superior à densidade das ligações entre comunidades e também superior àquela que se possa verificar numa rede aleatória. A emergência de estrutura como propriedade global

não está definida à priori, estando apenas definidas regras locais para o estabelecimento de ligações entre os diversos nós, que asseguram *assortative mixing* e transitividade local. O fenómeno global do surgimento de estrutura na rede é então considerado o fenómeno emergente do modelo.

Vizinhança - Para que o modelo possa, em regime livre, comportar-se de forma a gerar redes que apresentem as características de *assortative mixing* e transitividade elevada, sem recorrer a um modelo de ligação preferencial das redes do tipo *scale-free*, e querendo que os agentes tenham um conhecimento apenas local do que é a 'sua realidade' é preciso definir uma distância de conhecimento (também chamada distância social), calculada através da distância geodésica. Quando esta distância é 1 então as redes locais correspondem às redes *Ego* onde o nó central tem conhecimento do seus *Alters* com quem tem ligações imediatas. Quando o valor da distância social é 2 os nós centrais tem conhecimento também de todos os *Alters* que se encontrem a uma distância geodésica 2 e assim sucessivamente.

Reciprocidade - Embora o envio de mensagens de correio electrónico seja um acto direccional, o modelo assume a existência de reciprocidade nas relações sociais. A ideia é que a ligação de A para B tenha a mesma importância que de B para A. Tal deve-se à assunção de que a emergência de relações sociais surge da interacção dos agentes e não necessariamente da direcção do envio da mensagem. No futuro o modelo poderá ser adaptado para produzir redes direccionadas onde a reciprocidade não se verifique.

Conhecimento - Os agentes têm uma visão local do sistema através da distribuição de probabilidade de estabelecimento de contactos. Esta não é global porque, para todos aqueles com quem não foi estabelecido nenhum contacto, uma probabilidade de 0 significa na prática a impossibilidade de interacção. O conhecimento local do agente é expandido para além desta distribuição de probabilidade derivada dos contactos anteriores, através da noção de vizinhança que lhe permite ter uma percepção das ligações dos seus *Alters*. Esta noção permite ao agente calcular uma probabilidade corrigida, que utilizará para efectuar novas ligações com pessoas da sua vizinhança.

4.4.1.5 Inicialização

O modelo é inicializado criando um espaço 2D onde os agentes são colocados aleatoriamente. Os agentes são instanciados a partir dos dados do ficheiro de treino em mesmo número e atribuídos aos seus departamentos de forma a que haja correspondência com o ficheiro de eventos que será carregado durante a fase inicial de treino. A colocação num espaço 2D é arbitrária, uma vez que não há um mapeamento do espaço físico e durante

a simulação apenas serão geradas ligações num espaço topológico. No entanto, esta representação permitirá futuramente acoplar ao modelo algoritmos de visualização a 2D da representação da rede gerada.

No final da inicialização o controlo da simulação é passado ao *scheduler*.

4.4.1.6 Dados de entrada ou treino

O modelo é iniciado com um ficheiro de dados contendo o número de professores e uma tabela de correspondência com os respectivos departamentos, por forma a colocar inicialmente os agentes do modelo de acordo com os departamentos reais do sistema. O outro conjunto de dados que é fornecido ao modelo é uma sequência de eventos de envio de mensagem de correio electrónico, tal como foram obtidos dos servidores de email do ISCTE. Este ficheiro regista uma mensagem por linha, com o emissor e respectivos destinatários. O ficheiro de treino será carregado na altura em que o modelo estiver em modo de treino. Um exemplo da estrutura do ficheiro de treino pode ser observado na tabela 4.21 onde estão representadas as primeiras 7 linhas do ficheiro.

Tabela 4.21: Exemplo do ficheiro de dados de treino do CIUCEU

0	1	2	
3	4		
7	8		
7	9		
10	11		
12	13	14	15
16	17		

4.4.1.7 Submodelos

Gerador de números aleatórios - a geração de eventos aleatórios ocorre na fase de regime livre da simulação e tem como objectivo atribuir eventos aos agentes, permitindo que estes enviem mensagens a outros agentes de acordo com a sua percepção do mundo local. Utilizámos o gerador de números aleatórios "*Mersenne twister*" (Matsumoto e Nishimura, 1998), que cria números pseudo-aleatórios uniformes.

Assortative mixing - A ideia de que agentes estabelecem ligações com agentes semelhantes é conseguida através da comparação dos valores de *degree* dos agentes. Para um conjunto de agentes $X =< x_1, x_2, ..., x_n >$ com a distribuição de *degree* $< k_1, k_2, ..., k_n >$

e distribuição de probabilidade histórica $< p_1, p_2, ..., p_n >$, a probabilidade corrigida *pc* entre os agentes i e j com $i \notin X, j \in X$ será dada por:

$$pc_{i,j} = \frac{\frac{p_j}{1+|k_i - k_j|}}{\sum_j \frac{p_j}{1+|k_i - k_j|}} \tag{4.9}$$

Desta forma, a probabilidade corrigida *pc* entra em linha de conta com a diferença de *degree* existente entre os dois agentes, fazendo com que tenha uma probabilidade maior quando o seu *degree* é próximo e menor quando os seus *degrees* forem muito diferentes.

Transitividade - A obtenção da transitividade elevada é conseguida através da restrição de aplicação da expressão para o *assortative mixing* a um conjunto de agentes que se situem numa vizinhança relativamente pequena do agente remetente.

A expressão da probabilidade corrigida 4.9 apenas corrige a probabilidade de contacto de acordo com a diferença de *degree* entre os nós onde já existiram contactos prévios ($p_j \neq 0$). Para além disso é necessário que dentro da vizinhança estabelecida, as probabilidades sejam calculadas de forma que as ligações estabelecidas sejam de alguma forma influenciadas pelas probabilidades dos contactos intermédios existentes.

Considerando por exemplo uma rede linear de 4 elementos e vizinhança 3, com os nós numerados sequencialmente $A =< n1 - n2 - n3 - n4 >$, a probabilidade de $n1$ se ligar a $n3$ ou $n4$ é 0, porque não foram historicamente feitos quaisquer contactos. Dentro desta vizinhança $v = 3$, para assegurar que o nó $n1$ pode também ligar-se a $n3$ ou $n4$, as suas probabilidades de ligação terão que ser calculadas de acordo com a distância a que se encontram do nó $n1$ segundo a seguinte expressão, onde a é um parâmetro ajustável para medir o peso da distância na determinação da probabilidade:

$$p_{1,2} = \frac{1}{1^a} p_{1,2} \ , \ p_{1,3} = \frac{1}{2^a} p_{2,3} \ , \ p_{1,4} = \frac{1}{3^a} p_{3,4}$$

Esta incorporação da distância permite que a probabilidade de ligação seja transitiva através da rede. A probabilidade pode ser generalizada para dois nós j e l, cuja distância entre eles seja 1 e para o nó i para o qual queremos calcular a probabilidade, que se encontra à distância geodésica d de l e $d-1$ de j:

$$p^*_{i,l} = \frac{1}{d^a} p_{j,l} \tag{4.10}$$

O valor $p^*_{i,l}$ da equação 4.10 representa o ajuste da probabilidade a uma determinada

distância sob um determinado caminho geodésico. No entanto, pode acontecer que i e l estejam conectados por mais do que um caminho geodésico dentro da vizinhança definida. Assim, é preciso calcular $p^*_{i,l}$ para todas as possibilidades e proceder a uma normalização, por forma a garantir que $\sum p^*_{i,l} = 1$, obtendo-se então:

$$p_{i,l} = \frac{\sum_{cam.geod} p^*_{i,l}}{\sum_k \sum_{cam.geod} p^*_{i,l}} \tag{4.11}$$

Eventos aleatórios - Apesar do modelo incorporar o conceito de transitividade e de *Assortative Mixing*, tal não é suficiente para garantir que no final estejamos perante um componente único na rede formada. Tal deve-se a ambos os conceitos só afectarem a formação de ligações entre nós da rede que já estejam de alguma forma conectados a outros nós (*degree*>0), mesmo que não estejam conectados ao nó que faz a nova ligação. Isto implica que o modelo não é capaz de criar novas ligações a nós isolados. Por que isso aconteça é preciso incluir uma probabilidade pequena, de a cada evento o destinatário ser alguém da universidade. Isto permite imitar a existência de páginas brancas de contactos da universidade e desta forma a possibilidade de um evento potenciar novos contactos.

4.4.2 Experimentação e resultados

O modelo foi primeiro afinado no valor da probabilidade dos eventos aleatórios com treino de 50%, vizinhança 1 e $a = 4$, de forma a assegurar que o valor do *degree* médio não divergia mais do que 10% do valor observado nos dados reais. Isto fez com que obtivéssemos o valor de 4.0×10^{-4} para a probabilidade dos eventos aleatórios.

O modelo foi testado para valores entre 10% e 90% de dados de treino e para diversos valores de vizinhança entre 1 e 5, por forma a avaliar o impacto do tamanho da vizinhança e também o impacto da quantidade de dados de treino nos resultados.

Os dados reais mostraram que o *degree* médio cresce de acordo com uma lei de potências, como se pode verificar pela figura 4.22 com um expoente característico de $0,593$.

Estando interessados no comportamento do modelo com a vizinhança e com a percentagem de treino utilizada como entrada, calculámos o degree médio, o n.º de ligações, a densidade da rede, a distância geodésica média e o coeficiente de *clustering*. Para todas as redes resultantes foi aplicado o algoritmo de Clauset-Newman-Moore, a fim de obter uma indicação do valor de modularidade Q para tentar perceber qual o efeito do modelo na estrutura da rede.

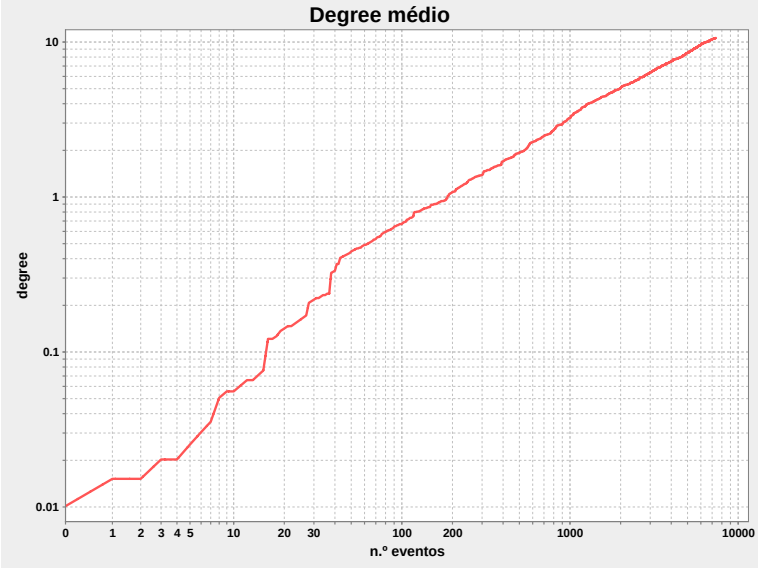

Figura 4.22: Evolução do *degree* médio real (log-log)

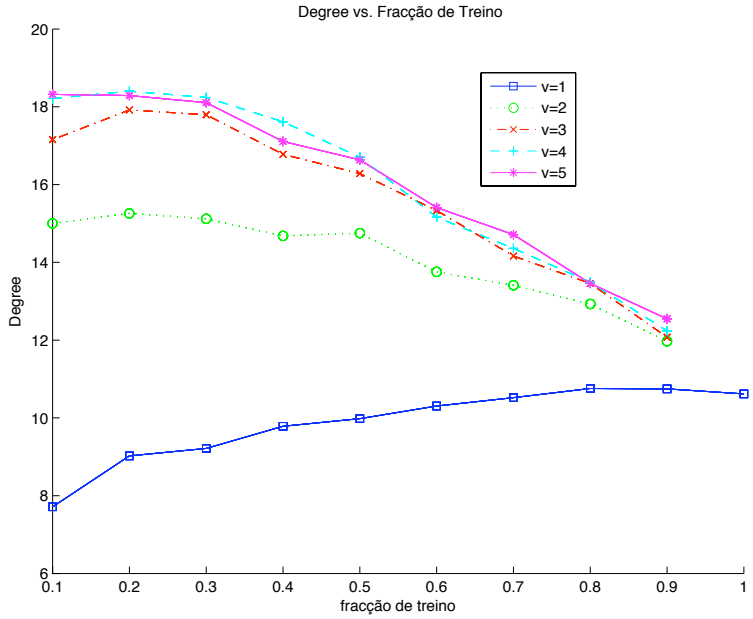

Figura 4.23: CIUCEU *degree* médio final *vs.* fracção de treino

O cálculo do *degree* médio final mostra que quando se combinam vizinhanças diferentes de 1 e percentagens de treino baixas os valores finais obtidos são bastante diferentes. Nota-se, como seria de esperar, a convergência dos valores de *degree* médio para o valor real com o aumento da percentagem de treino. Também se verifica que o aumento da vizinhança para valores superiores a 3 não parece provocar alterações significativas nos resultados do modelo.

O número de ligações formadas está naturalmente correlacionado com o o *degree* médio e o

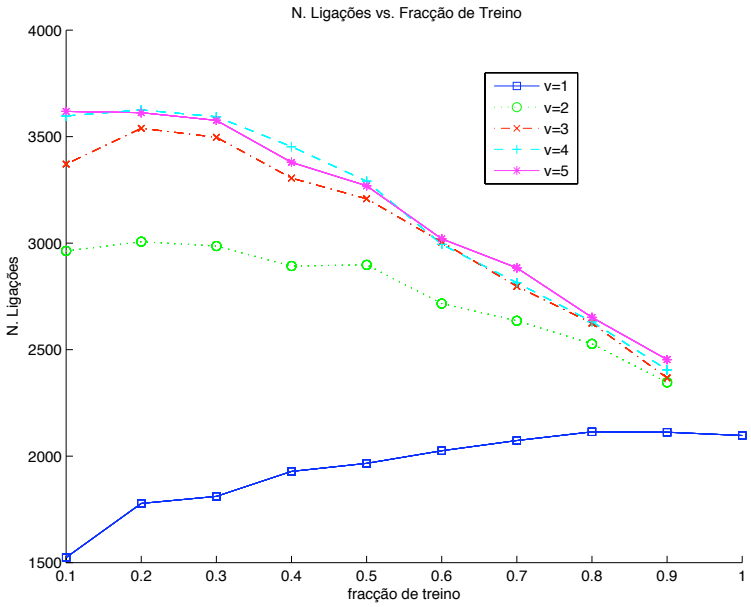

Figura 4.24: n.º de ligações final *vs.* fracção de treino

comportamento observado é em tudo semelhante ao da figura 4.23. Novamente, um valor da vizinhança superior a 3 não influência significativamente os resultados finais.

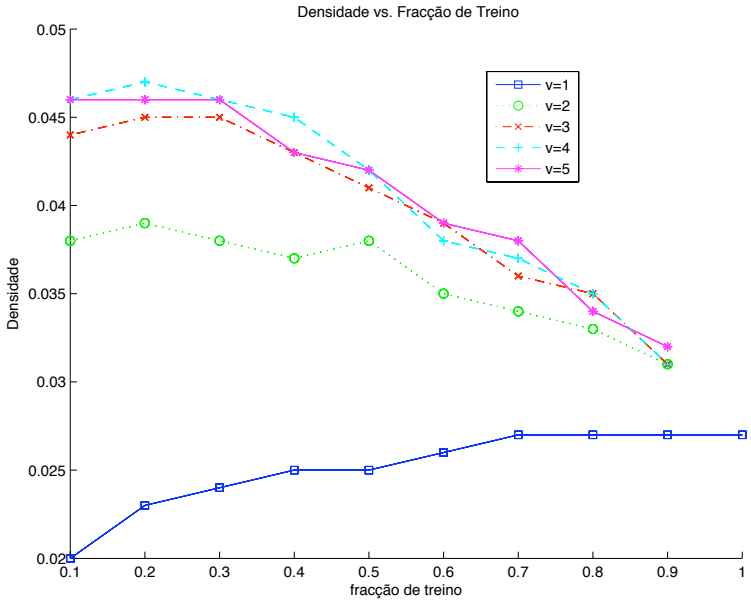

Figura 4.25: densidade final *vs.* fracção de treino

A conclusão respeitante à figura 4.24 é a mesma em relação aos resultados da densidade de ligações da rede. A rede original tem uma densidade de 2,7%. Observando a figura 4.25 observa-se que a utilização de vizinhança superior a 1 leva a aumentos de densidade que atingem 4,7% para os valores mais baixos de fracção de treino.

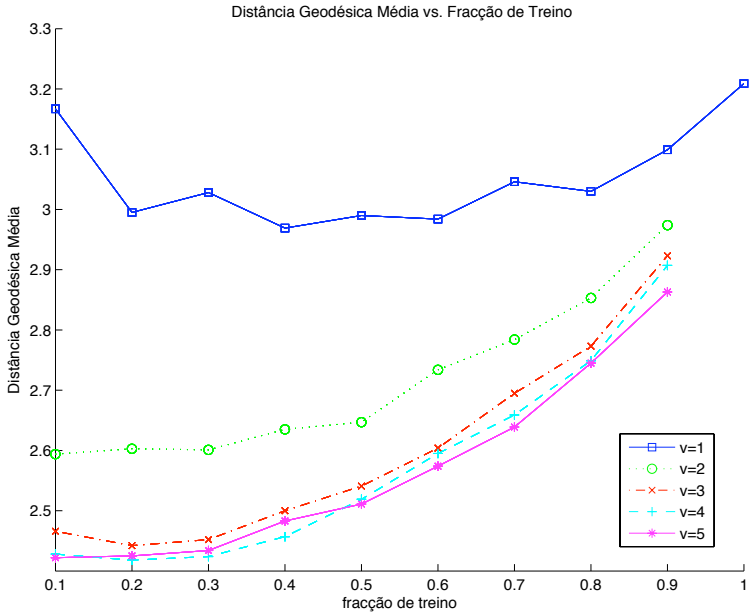

Figura 4.26: Distância geodésica média final *vs.* fracção de treino

Analisando a distância geodésica média verifica-se que o efeito da aleatoriedade existente no modelo faz descer o valor da distância geodésica média. No entanto, o efeito só é acentuado quando se utilizam vizinhanças superiores a 1, parecendo atingir um patamar mínimo para uma distância geodésica média de 2,43. O decréscimo da distância geodésica média não é tão pronunciado no caso da vizinhança 1, não mostrando os resultados uma dependência sensível à fracção de treino. No entanto, para valores de vizinhança superior essa dependência é observada.

O coeficiente de *clustering*, da figura 4.27, apresenta um caso curioso. Independentemente da vizinhança utilizada, a fracção de treino apresenta-se correlacionada com o resultado do coeficiente de *clustering*, aumentando à medida que a fracção de dados de treino utilizada também aumenta. No entanto, analisando a dependência dos valores de clustering com a vizinhança verifica-se que para $v = 2$ os valores de clustering são mais altos, voltando a descer para $v = 3, 4$ e 5 e são mais próximos dos valores de $v = 1$. Este fenómeno indica que um pequeno valor de vizinhança é importante para que o modelo apresente características de transitividade, mas o efeito duma vizinhança estendida produz o efeito contrário.

O objectivo de estudar a modularidade das redes produzidas pelo modelo tem por detrás a ideia de verificar se pequenas percentagens de dados de treino são ainda capazes de manter a estrutura existente da rede. Naturalmente, quanto menor for a percentagem de treino menor será a modularidade. No entanto, verificamos que mesmo para 10% de

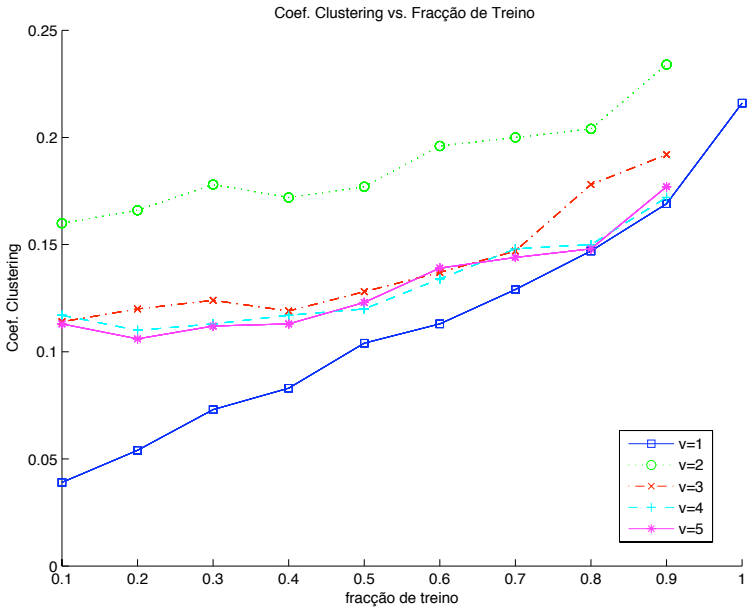

Figura 4.27: Coeficiente de *clustering* final *vs.* fracção de treino

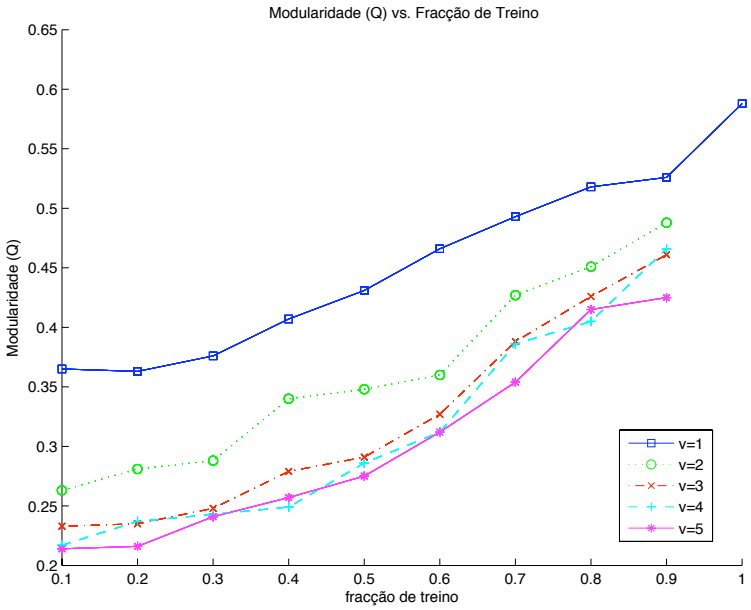

Figura 4.28: Modularidade final *vs.* fracção de treino

treino, no caso em que a vizinhança é 1, o modelo consegue manter uma modularidade superior a 0,3. Lembremo-nos que este é o valor mínimo considerado por Clauset, Newman e Moore para que haja estrutura numa rede. Verifica-se também que com o aumento da vizinhança o valor da modularidade final diminui, fazendo aumentar a percentagem de treino necessária para que a rede apresente ainda estrutura. A vizinhança parece ter um efeito negativo na estrutura da rede, o que faz sentido se pensarmos numa vizinhança

infinita, onde todos conheçam todos. Nesse caso estaríamos numa situação de uma rede aleatória sem estrutura.

Capítulo 5

Conclusão e Perspectivas

O trabalho desenvolvido procurou descobrir e analisar a estrutura do sistema de comunicação informal, apoiado no sistema de correio electrónico do ISCTE. Mostrámos no capítulo 4 várias evidências que suportam a nossa hipótese de trabalho, nomeadamente que a estrutura latente existente no sistema de correio electrónico do ISCTE representa informação suficiente para caracterizar comunidades e hierarquias dentro da instituição.

Na análise efectuada com algoritmos hierárquicos, nos pontos 4.2.1 e 4.2.2, mostrámos claramente que a utilização de algoritmos de detecção de comunidades baseados em propriedades globais é capaz de identificar comunidades na rede de professores do ISCTE, sem ser necessário o conhecimento do seu conteúdo semântico. A existência de estrutura é provada através do valor de modularidade obtido para a sua aplicação.

Mostrámos, no ponto 4.2.3, que a estrutura das rede de comunicação informal entre professores do ISCTE apresenta hierarquias em determinados departamentos. Os *k-cores* mais elevados e portanto contendo as pessoas com mais conexões, são dominados por 5 departamentos: DCTI, DMQ, DE, DCG e DA. Os seus membros aparecem predominantemente nos *k-cores* 8 e 9, indicando possuírem um papel mais preponderante na estrutura hierárquica da rede de comunicação informal existente.

Através dos resultados dos ponto 4.2.1 e 4.2.2 verificámos que as comunidades de comunicação informal apresentam diferenças para a estrutura de departamentos do ISCTE, indicando que as redes de comunicação informal ultrapassam as fronteiras dos departamentos. Este facto foi confirmado pela utilização do método de percolação de cliques no ponto 4.2.4, onde se verificou a transversalidade das comuniades formadas e inclusive a sobreposição entre elas, permitindo que alguns professores possam ser considerados como

pertencentes a mais do que uma comunidade.

Criámos um modelo multi-agente, no ponto 4.4, capaz de gerar redes com estrutura apenas a partir do conhecimento local que cada agente tem de um histórico de contactos. Analisámos a sua dependência com o tamanho da vizinhança "social" dos agentes e mostrámos que a rede de comunicação informal é gerada utilizando vizinhanças "sociais" de dimensão reduzida. Mostrámos também que a utilização de dados reais para treinar o modelo multi-agente permite uma aproximação dos resultados da simulação à realidade, sendo que a escolha da fracção de dados treino deve ter em linha de conta o tamanho da vizinhança "social" escolhida.

Pela análise efectuada na caracterização do sistema de correio electrónico do ISCTE do ponto 4.1.2, mostrámos os diferentes níveis de adopção do sistema pelas três categorias principais de intervenientes da vida académica. Verificámos, na tabela 4.2, a pequena adesão por parte dos alunos da universidade ao sistema de correio electrónico. Esta informação poderá ser analisada futuramente para criar políticas de incentivo à utilização pelos alunos do sistema disponibilizado pela universidade.

Verificámos no ponto 4.1.2 que, ao contrário do que existe na literatura, a distribuição de *degree* de cada uma das redes do membros do ISCTE não obedece a leis de potência, excepto em determinadas restrições, mas antes apresenta um decaimento exponencial.

Mostrámos no ponto 4.1.2 que o envio de correio electrónico tem uma distribuição temporal característica, quer diária, quer semanal. Este resultado é intuitivo e expectável. No entanto, quantificámos essa distribuição e os resultados poderão ser futuramente utilizados para aplicação prática de medidas de optimização da rede do ISCTE.

Este trabalho contribui para o avanço dos estudos de redes sociais em diversos aspectos, nomeadamente:

Contribui para o campo da simulação multi-agente de redes, através da análise do impacto que a utilização de dados reais tem no desenho, implementação e resultados de simulações multi-agente de redes com dados reais. Neste caso concreto mostrámos que a fracção de dados de amostragem a utilizar em simulação multi-agente terá que ter em linha de conta o objectivo da simulação (no caso detecção de estrutura) e o comportamento do modelo no espaço de variáveis.

Aplica uma medida de variação de informação, eminentemente estatística, a redes sociais e aos resultados de algoritmos de detecção de comunidades, permitindo determinar distâncias entre diferentes particionamentos. Desta forma, é possível ter uma medida comparável entre diversos algoritmos de detecção de comunidades sem estar dependente

dos mecanismos intrínsecos de cada classe de algoritmo. A utilização da variação de informação é independente do método escolhido, ao passo que a modularidade é intrínseca aos algoritmos utilizados.

Abre alguns caminhos para possíveis investigações, nomeadamente na investigação do fluxo de informação dentro de organizações de tipo universitário através das redes informais de comunicação. O trabalho pode ser útil na investigação e avaliação dos processos de aprendizagem. As técnicas utilizadas podem servir para compreender como o conhecimento é transmitido dentro das instituições de ensino. A partir deste trabalho, pode-se expandir o estudo para a criação de modelos que mimem o comportamento de alunos nos seus processos de aprendizagem.

Uma área para a qual esta investigação pode ser útil e onde o conhecimento é ainda reduzido é a área de redes sociais com atributos ou *tagged networks*, onde se associa conhecimento semântico, ou atributos, aos diversos nós e ligações. O estudo desse tipo de redes está ainda na fase embrionária e novos algoritmos e abordagens estão a ser desenvolvidos. O estudo aqui apresentado permite, pelas características dos dados existentes, ser expandido para essa categoria de análise.

Uma das ideias expostas neste trabalho tem a ver com a utilização de medidas de informação para classificação dos particionamentos obtidos. Esta é uma medida estatística baseada na teoria da informação. Um futuro desenvolvimento poderá passar pelo estudo de medidas para aplicação à detecção de comunidades e que possam igualmente medir os resultados de simulação multi-agente.

Uma área de investigação que pode beneficiar da análise aqui efectuada é a da análise de redes sociais a múltiplos níveis, percebendo como as diferentes redes sociais às quais um indivíduo pertence interagem entre si. O estudo do indivíduo enquanto central a um sistema composto por diversos níveis, sejam redes de trabalho, redes de amizades ou redes geográficas, pode beneficiar do trabalho de detecção de comunidades para a compreensão da noção de território. Este território engloba todas as interacções existentes entre essas redes através da participação do indivíduo nas suas redes. O estudo das relações entre múltiplos níveis é então uma área para a qual este trabalho pode servir de ponto de partida.

Outra área que não foi explorada neste trabalho mas que pode beneficiar das técnicas aqui expostas é a detecção de correio electrónico indesejado, podendo estas técnicas ajudar a desenvolver sistemas que beneficiem do conhecimento da estrutura da rede, por forma a evitar falsos positivos.

Bibliografia

Albert, Reka e Barabasi, Albert-Laszlo. Statistical mechanics of complex networks. *cond-mat/0106096* (2001). Reviews of Modern Physics 74, 47 (2002).
URL http://arxiv.org/abs/cond-mat/0106096

Anthonisse, J. M. The rush in a directed graph. Relatório técnico, Stichting Mathematicsh Centrum, Amsterdam (1971).

Barabasi, Albert-Laszlo e Albert, Reka. Emergence of scaling in random networks. *Science*, 286:509 (1999).
URL http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/9910332

Brandes, Ulrik. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30:136–145 (2008). doi:10.1016/j.socnet.2007.11.001.
URL http://www.sciencedirect.com/science/article/B6VD1-4RFJ4K1-1/1/d40b8e563e7f8b505bc3b5eadafd1a94

Brin, Sergey e Page, Lawrence. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30:107–117 (1998). doi:10.1.1.42.3243.
URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.3243

Broadbent, S. R. e Hammersley, J. M. Percolation processes. i. crystals and mazes. In *Proceedings of the Cambridge Philosophical Society*, volume 53, páginas 629–641 (1957).
URL http://adsabs.harvard.edu/abs/1957PCPS...53..629B

Bunde, Armin e Havlin, Shlomo. *Fractals in Science*. Springer (1995). ISBN 3540562214, 298 páginas.

Callaway, D. S, *et al.* Network robustness and fragility: Percolation on random graphs. *cond-mat/0007300* (2000). Phys. Rev. Lett. 85, 5468-5471 (2000).
URL http://arxiv.org/abs/cond-mat/0007300

Carrington, Peter J., Scott, John, e Wasserman, Stanley. *Models and Methods in*

Social Network Analysis. Cambridge University Press (2005). ISBN 0521600979, 344 páginas.

Clauset, Aaron, Newman, M. E. J, e Moore, Cristopher. Finding community structure in very large networks. *cond-mat/0408187* (2004). doi:doi:10.1103/PhysRevE.70.066111. Phys. Rev. E 70, 066111 (2004).
URL http://arxiv.org/abs/cond-mat/0408187

Dall, Jesper e Christensen, Michael. Random geometric graphs. *cond-mat/0203026* (2002). Phys. Rev. E 66, 016121 (2002).
URL http://arxiv.org/abs/cond-mat/0203026

Dastani, Mehdi. 2apl: a practical agent programming language. *Autonomous Agents and Multi-Agent Systems*, 16:214–248 (2008). doi:10.1007/s10458-008-9036-y.
URL http://dx.doi.org/10.1007/s10458-008-9036-y

Derenyi, Imre, Palla, Gergely, e Vicsek, Tamas. Clique percolation in random networks. *Physical Review Letters*, 94:160202 (2005).
URL doi:10.1103/PhysRevLett.94.160202

desJardins, Marie, Gaston, Matthew E., e Radev, Dragomir. Introduction to the special issue on ai and networks. *AI Magazine*, 29(3):11–15 (2008).

Diestel, Reinhard. *Graph Theory*. Springer, 3rd edição (2005). ISBN 3540261826, 415 páginas.

Dorogovtsev, S. N, Goltsev, A. V, e Mendes, J. F. F. k-core organization of complex networks. *cond-mat/0509102* (2005). Phys. Rev. Lett. 96, 040601 (2006).
URL http://arxiv.org/abs/cond-mat/0509102

Ebel, Holger, Mielsch, Lutz-Ingo, e Bornholdt, Stefan. Scale-free topology of e-mail networks. *Physical Review E*, 66:035103 (2002). doi:10.1103/PhysRevE.66.035103. Copyright (C) 2008 The American Physical Society; Please report any problems to prola@aps.org.
URL http://link.aps.org/abstract/PRE/v66/e035103

Erdós, P. e Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5:17–61 (1960).

Euler, Leonhard. Solvtio problematis ad geometriam sitvs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 8(53):128–140 (1741).

Fortunato, Santo e Barthelemy, Marc. Resolution limit in community detection.

physics/0607100 (2006). doi:doi:10.1073/pnas.0605965104. Proc. Natl. Acad. Sci. USA 104 (1), 36-41 (2007).
URL http://arxiv.org/abs/physics/0607100

Freeman, Linton C. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41 (1977).
URL http://links.jstor.org/sici?sici=0038-0431%28197703%2940%3A1%3C35%3AASOMOC%3E2.0.CO%3B2-H

Garriss, Scott, *et al.* Abstract re: Reliable email. In *Proceedings of the 3rd Symposium on Networked Systems Design and Implementation*, páginas 297–310. San Jose, CA (2006). doi:10.1.1.61.7781.
URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.7781

Girvan, Michelle e Newman, M. E. J. Community structure in social and biological networks. *cond-mat/0112110* (2001). Proc. Natl. Acad. Sci. USA 99, 7821-7826 (2002).
URL http://arxiv.org/abs/cond-mat/0112110

Hamill, Lynne e Gilbert, Nigel. A simple but more realistic agent-based model of a social network. In *Conf. European Social Simulation Assoc. (ESSA'08)* (2008).

Hammersley, J. M. Percolation processes. ii. the connective constant. In *Proceedings of the Cambridge Philosophical Society*, volume 53, páginas 642–645 (1957).
URL http://adsabs.harvard.edu/abs/1957PCPS...53..642H

Heinemann, Andreas, *et al.* Ad Hoc Collaboration and Information Services Using Information Clouds. In Torsten Braun, Nada Golmie, e Jochen Schiller, editores, *Proceedings of the 3rd Workshop on Applications and Services in Wireless Networks, (ASWN 2003)*, páginas 233–242. Institute of Computer Science and Applied Mathematics, University of Bern, Bern, Switzerland (2003a).
URL http://iclouds.tk.informatik.tu-darmstadt.de/iClouds/pdf/aswn2003.pdf

Heinemann, Andreas, *et al.* iClouds – Peer-to-Peer Information Sharing in Mobile Environments. In Harald Kosch, László Böszörményi, e Hermann Hellwagner, editores, *Euro-Par 2003. Parallel Processing, 9th International Euro-Par Conference*, volume 2790 de *Lecture Notes in Computer Science*, páginas 1038–1045. Springer, Klagenfurt, Austria (2003b).
URL http://iclouds.tk.informatik.tu-darmstadt.de/iClouds/pdf/europar2003.pdf

Janssen, Marco A. e Ostrom, Elinor. Empirically based, agent-base models. In *Ecology and Society*, volume 11(2). Resilience Alliance (2006).

Jeong, H, *et al.* The large-scale organization of metabolic networks. *Nature*, 407:651–4 (2000). ISSN 0028-0836. doi:11034217. PMID: 11034217.
URL `http://www.ncbi.nlm.nih.gov/pubmed/11034217`

Karinthy, Frigyes. *Everything is Different*. Budapest (1929).

Kim, Ungsik. Analysis of personal email networks using spectral decomposition. *International Journal of Computer Science and Network Security*, 7(4):185–188 (2007).

Kleinberg, Jon M. Authoritative sources in a hyperlinked environment. *J. ACM*, 46:604–632 (1999). doi:10.1145/324133.324140.
URL `http://portal.acm.org/citation.cfm?id=324140`

Louçã, Jorge, *et al.* Emergence in social networks: Modeling the intentional properties of multi-agent systems. In Fréderic Amblard, editor, *Proceedings of the 4th Conference of the European Social Simulation Association (ESSA'07)*, páginas 639–650. Toulouse, France (2007a).

Louçã, Jorge, *et al.* Pattern-oriented analysis of communication flow: The case study of cicada barbara lusitanica. In Ivan Zelinka, Zuzana Oplatková, e Alessandra Orsoni, editores, *21st EUROPEAN Conference on Modelling and Simulation ECMS 2007*, páginas 229–234. Prague, Czech Republic (2007b).

Luke, Sean, *et al.* Mason multiagent simulation toolkit (2009).
URL `http://cs.gmu.edu/~eclab/projects/mason/`

von Luxburg, U. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416 (2007).
URL `http://springerlink.metapress.com/content/jq1g17785n783661/fulltext.pdf`

Macqueen, JB. Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, páginas 297, 281 (1967).

Mardia, Kanti V., Kent, J. T., e Bibby, J. M. *Multivariate Analysis*. Academic Press (1980). ISBN 0124712525.

Matsumoto, Makoto e Nishimura, Takuji. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput.*

Simul., 8(1):3–30 (1998). doi:10.1145/272991.272995.
URL http://portal.acm.org/citation.cfm?doid=272991.272995

Meilă, Marina. Comparing clusterings—an information based distance. *J. Multivar. Anal.*, 98(5):873–895 (2007).
URL http://portal.acm.org/citation.cfm?id=1233220

Newman, M. E. J. Mixing patterns in networks. *cond-mat/0209450* (2002). Phys. Rev. E 67, 026126 (2003).
URL http://arxiv.org/abs/cond-mat/0209450

Newman, M. E. J. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582 (2006). doi:10.1073/pnas.0601602103.
URL http://www.pnas.org/cgi/content/abstract/103/23/8577

Newman, M. E. J e Girvan, M. Finding and evaluating community structure in networks. *cond-mat/0308217* (2003). Phys. Rev. E 69, 026113 (2004).
URL http://arxiv.org/abs/cond-mat/0308217

Newman, M. E. J, Jensen, I., e Ziff, R. M. Percolation and epidemics in a two-dimensional small world. *cond-mat/0108542* (2001). Phys. Rev. E 65, 021904 (2002).
URL http://arxiv.org/abs/cond-mat/0108542

Newman, M. E. J e Park, Juyong. Why social networks are different from other types of networks. *cond-mat/0305612* (2003). Phys. Rev. E 68, 036122 (2003).
URL http://arxiv.org/abs/cond-mat/0305612

Newman, Mark, Barabasi, Albert-Laszlo, e Watts, Duncan J. *The Structure and Dynamics of Networks:*. Princeton University Press, 1 edição (2006). ISBN 0691113572, 624 páginas.

Palla, Gergely, *et al.* Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814–8 (2005). ISSN 1476-4687. doi: nature03607. PMID: 15944704.
URL http://www.ncbi.nlm.nih.gov/pubmed/15944704

Palla, Gergely, *et al.* Directed network modules. *New Journal of Physics*, (9):186 (2007).

Price, Derek De Solla. Networks of Scientific Papers. *Science*, 149(3683):510–515 (1965). doi:10.1126/science.149.3683.510.
URL http://www.sciencemag.org

Price, Derek De Solla. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27:292–306 (1976). doi:10.1002/asi.4630270505.
URL http://dx.doi.org/10.1002/asi.4630270505

Shortreed, Susan. *Learning in Spectral Clustering*. Tese de Doutoramento, University of Washington Graduate School (2006).

de Sola Pool, Ithiel e Kochen, Manfred. Contacts and influence. *Social Networks*, 1(1):5–51 (1978). doi:http://dx.doi.org/10.1016/0378-8733(78)90011-4.
URL http://dx.doi.org/10.1016/0378-8733(78)90011-4

Solomonoff, Ray e Rapoport, Anatol. Connectivity of random nets. *Bulletin of Mathematical Biology*, 13:107–117 (1951). doi:10.1007/BF02478357.

Stauffer, Dietrich e Aharony, Ammon. *Introduction To Percolation Theory*. CRC, 1 edição (1994). ISBN 0748402535, 192 páginas.

Strogatz, Steven e Watts, Duncan. Collective dynamics of small-world networks. *Nature*, 293:420–442 (1998).
URL http://tam.cornell.edu/tam/cms/manage/upload/SS_nature_smallworld.pdf

Symons, John, *et al.* Detecting emergence in the interplay of networks. Arlington, Virginia (2007).

Travers, Jeffrey e Milgram, Stanley. An experimental study of the small world problem. *Sociometry*, 32(4):425–443 (1969).
URL http://www.jstor.org/stable/2786545

Tyler, Joshua R., Wilkinson, Dennis M., e Huberman, Bernardo A. *Email as spectroscopy: automated discovery of community structure within organizations*, páginas 81–96. Kluwer, B.V. (2003). ISBN 1-4020-1611-5.
URL http://portal.acm.org/citation.cfm?id=966268

Watts, Duncan J. *Six Degrees: The Science of a Connected Age*. W. W. Norton & Company, 1st edição (2003). ISBN 0393041425, 368 páginas.

White, Harrison C. Search parameters for the small world problem. *Social Forces*, 49(2):259–264 (1970).

Apêndice A

Figuras dos diversos *k-cores*

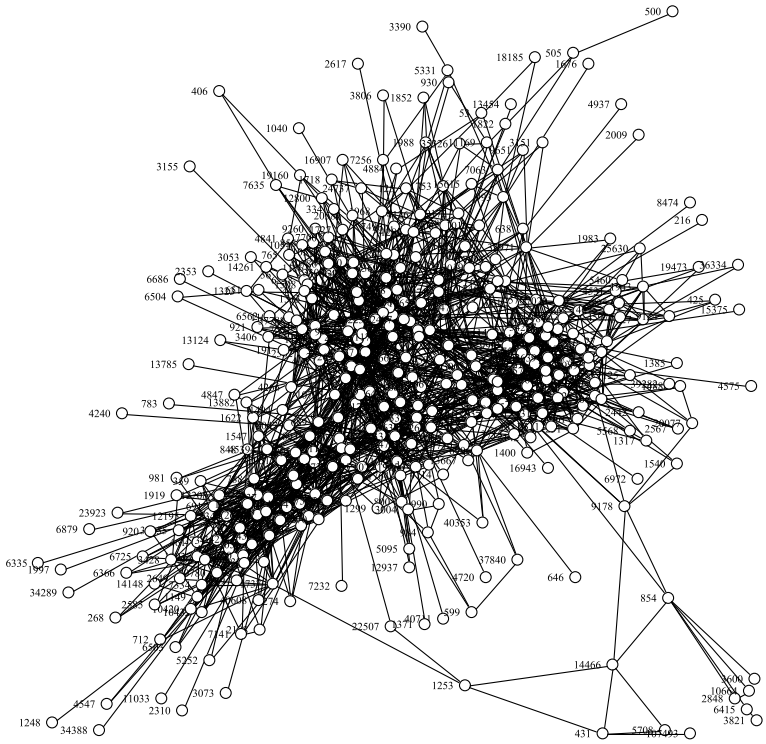

Figura A.1: k-core k=1

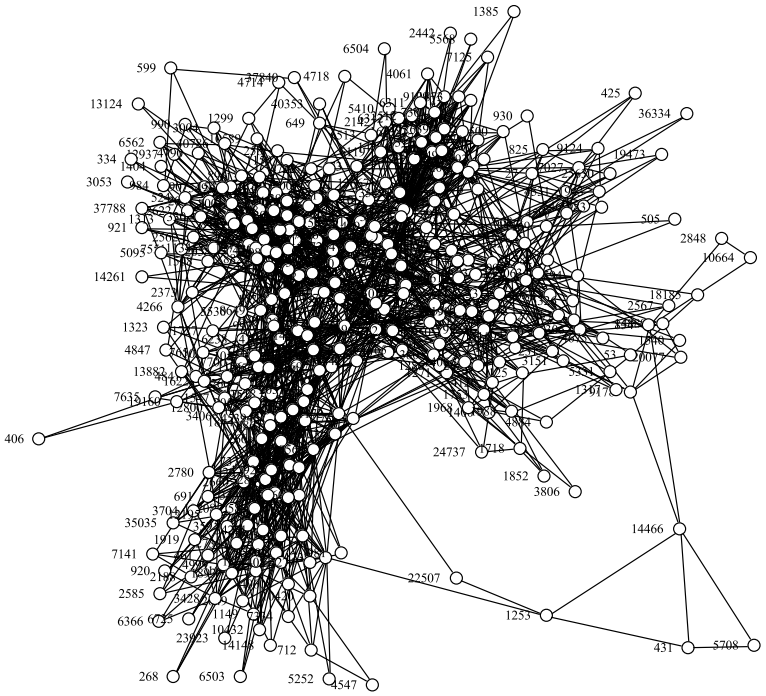

Figura A.2: k-core k=2

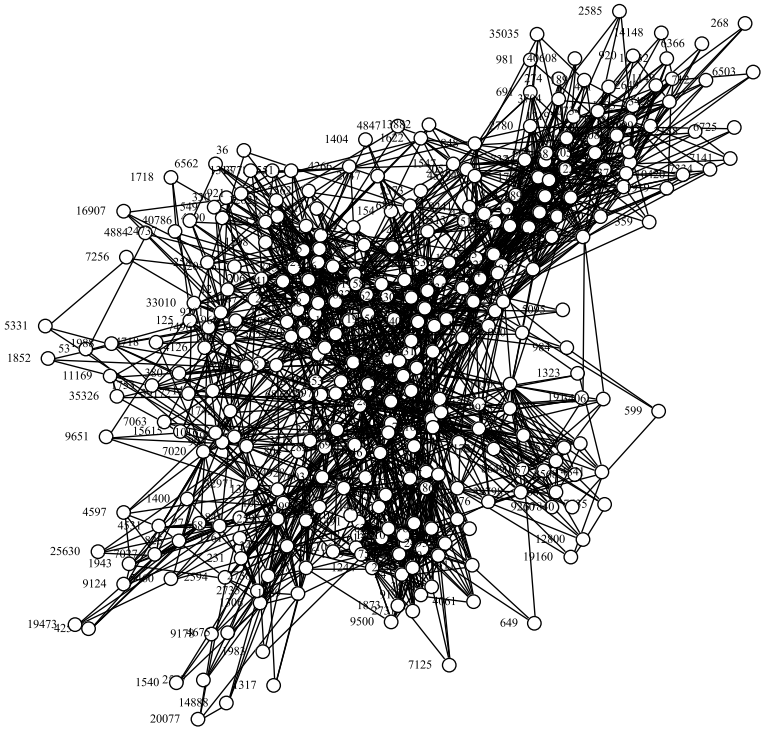

Figura A.3: k-core k=3

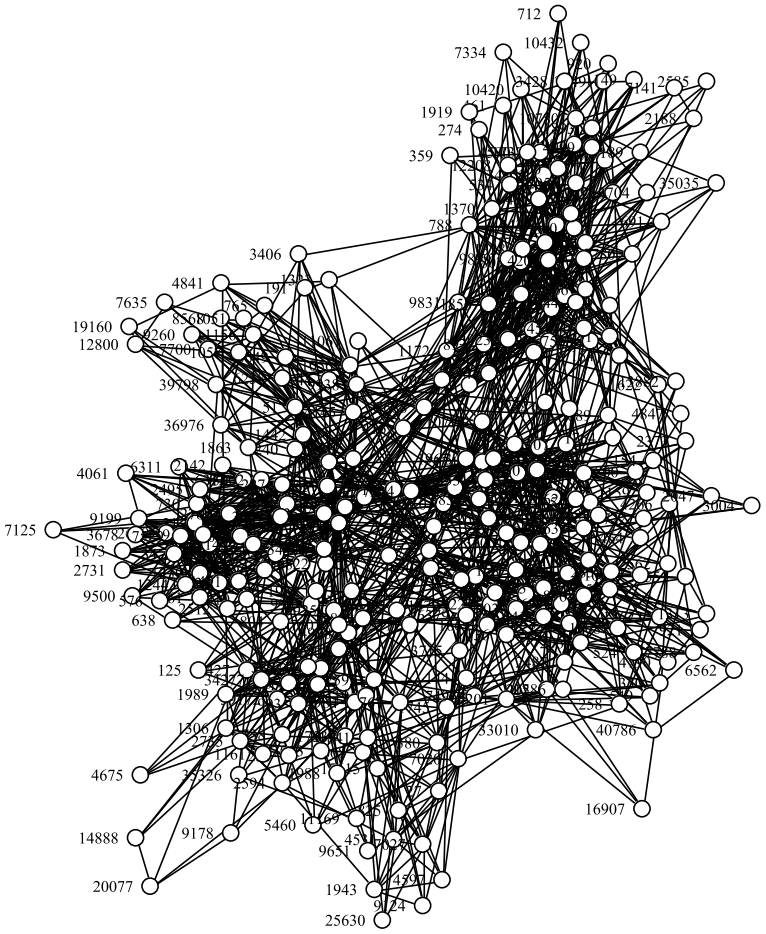

Figura A.4: k-core k=4

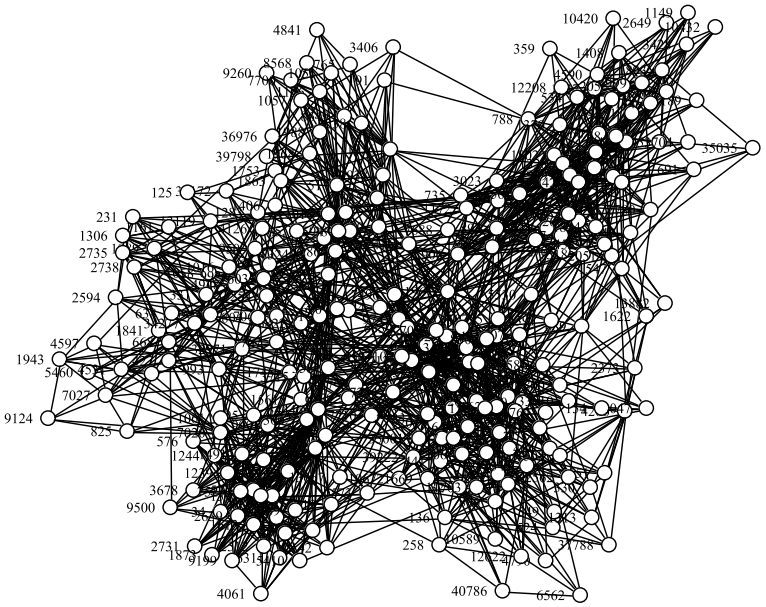

Figura A.5: k-core k=5

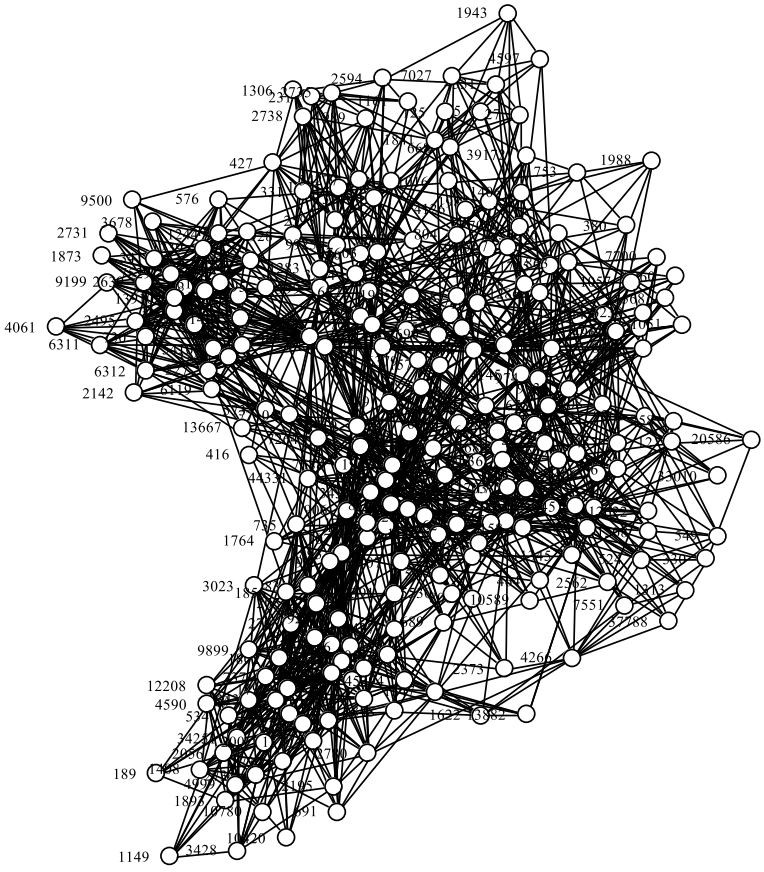

Figura A.6: k-core k=6

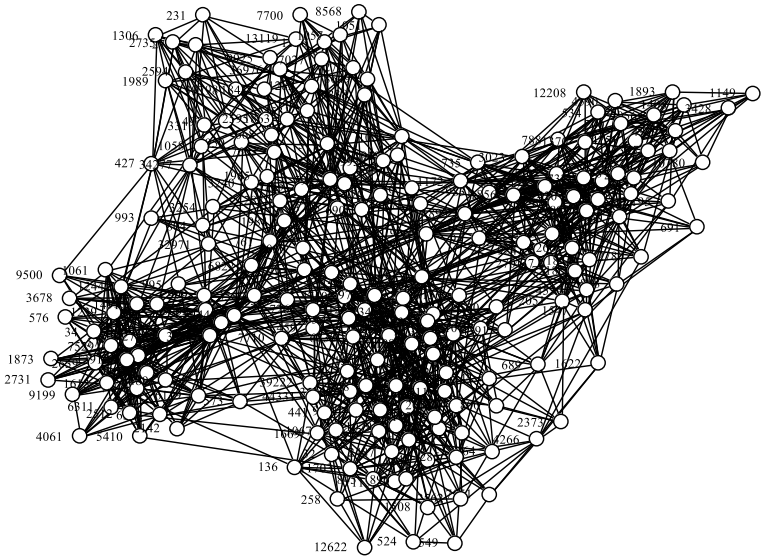

Figura A.7: k-core k=7

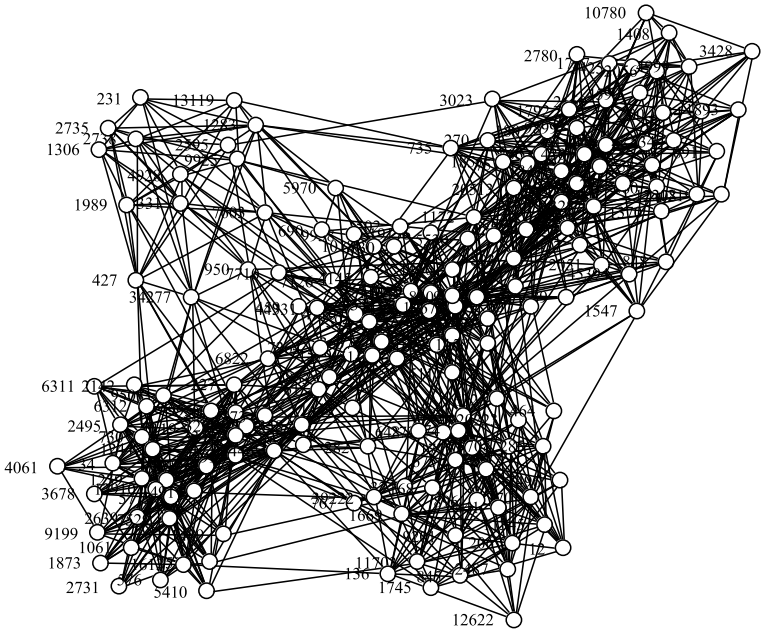

Figura A.8: k-core k=8

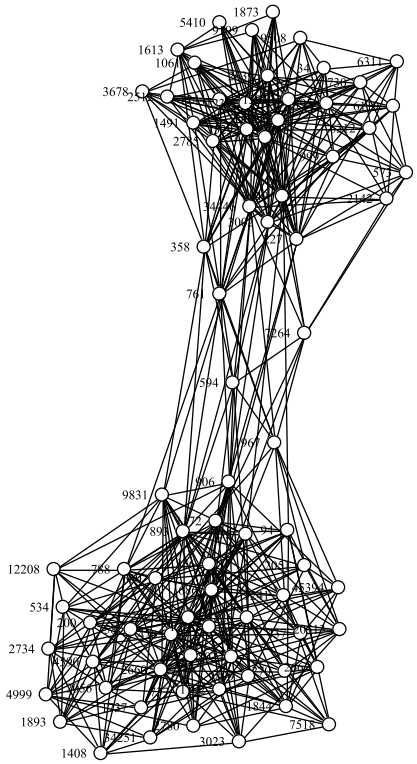

Figura A.9: k-core k=9

David Manuel de Sousa Rodrigues

(Licenciado em *Engenharia Química* pelo *Instituto Superior Técnico* da *Universidade Técnica de Lisboa*)

Curriculum Vitae

Fevereiro 2009