U.PORTO

FEP

ECONOMIA E GESTÃO

Redes Sociais e Classificação Conceptual: Abordagem Complementar para um Sistema de Recomendação de Coautorias

por

Tiago Filipe Pacheco Ferreira

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

Orientada por:

Prof. Dra. Maria Paula de Pinho de Brito Duarte Silva

2013

Nota Biográfica

Licenciado em Economia na Faculdade de Economia da Universidade do Porto em 2011. Durante a licenciatura realizou, em 2008 um estágio de Verão no Banco Santander Totta; entre 2010 e 2011 foi membro do departamento de Recursos Humanos da AIESEC - Association Internationale des Etudiants en Sciences Economiques et Commerciales.

Em 2011 ingressa no Mestrado em Análise de Dados e Sitemas de Apoio à Decisão da Faculdde de Economia da Universidade do Porto.

Foi colaborador no Financial Service Shared Center da adidas Group na Maia, entre 2011 e 2013, pcupando o cargo de Junior Account em 2011 e de Assistant Account entre 2012 e 2013.

Em Setembro de 2013 iniciou-se como colaborador da Porto Editora, ocupando o cargo de Adjunto da Direção Comercial.

Agradecimentos

À Professora Dra Maria Paula Brito , minha orientadora, pelas ideias, sugestões e paciência, incutindo-me ao longo do processo motivação para que este fosse levado até ao fim.

Aos meus pais, irmão e restante família que sempre me deram muita força para continuar, e que estiveram presentes em todos os momentos com boa disposição para me alegrar e dar motivação.

À Tânia Rocha o meu especial agradecimento, por estar sempre do meu lado para me dar boa disposição e motivação. Por todos os momentos em que o cansaço começava a vencer e mesmo assim me conseguia dar força para continuar.

Finalmente, aos meus amigos que sempre me apoiaram.

Resumo

Esta dissertação tem como o principal intuito a criação de uma aplicação de aconselhamento aliando dois métodos, a Análise de Redes e a Análise Conceptual Formal. Utilizando uma base de dados de coautorias em publicações foi possível criar uma ferramenta capaz de aconselhar a um qualquer autor da base de dados, todo um conjunto de autores que também já publicaram na mesma área.

Numa sociedade que diariamente é capaz de produzir dados em massa, é cada vez mais necessária a criação de ferramentas para a sua análise e interpretação. A utilização de redes sociais e o armazenamento de dados online pela sociedade disponibiliza todo um conjunto de informações úteis para a tomada de decisões.

Ao longo desta dissertação foram estudados os dois métodos mencionados, com o intuito de desenvolver um método que, aliando os resultados produzidos por ambos, pudesse usá-los em conjunto na tomada de decisões. Este sistema é a contribuição inovadora desta dissertação, dado conciliar dois métodos distintos para criar um Sistema de Recomendação.

PALAVRAS CHAVE

Análise Conceptual Formal, Coautorias, Redes Sociais, Sistema de Recomendação.

Abstract

This Master Dissertation's main objective is the creation of a Recommendation System using two methods, Network Analysis and Formal Concept Analysis. Using a publication co-authoring data base, we have developed a tool capable of advising each author on the data base, a group of other authors that had also published on the same areas.

In a society that is capable of producing daily a huge amount of data, it is each and every day more necessary to create new tools for its analysis and interpretation. The usage of social networks and the storage of online data by the Society, makes a large amount of data available, which is useful in the decision making.

Along this dissertation we have analyzed both methods mentioned, with the objective of combining them together in the decision making application. The proposed system is the added value of this dissertation, using two distinct methods to create a Recommendation System.

KEYWORDS

Co-Authorship, Formal Concept Analysis, Recommendation system, Social Networks.

Índice

Índice de Figuras

Índice de Tabelas

1 Introdução

Atualmente a informação tem muito valor, sendo feitos importantes investimentos na sua manipulação e análise desta. Com a informatização da sociedade, quase toda a gente possui uma grande quantidade de informação armazenada na gigantesca *World Wide Web*. A Análise de Redes Sociais constitui hoje um tema de investigação em grande desenvolvimento, com uma dinâmica notável, sendo constantemente apresentados trabalhos que visam esta temática em variadíssimas áreas do conhecimento. Este facto deve-se sobretudo à sua atualidade, ao ritmo a que se desenvolve, e às múltiplas questões que esta recente realidade pode levantar. A Análise Conceptual Formal é uma metodologia que tem por objetivo determinar uma hierarquia de conceitos ou ontologia formal a partir de um conjunto de objetos e atributos; apoia-se na teoria de reticulados e relações de ordem. Hoje em dia, a Análise Conceptual Formal tem aplicações em áreas tão diversas como o *data mining, text mining, machine learning, knowledge management, semantic web, software development* e Biologia. Esta área tem também crescido como área de investigação, no entanto, não ao mesmo ritmo da Análise de Redes Sociais.

A análise de redes tem por base a Teoria de Grafos, esta foi referida pela primeira vez em 1736 por Leonhard Euler no *paper "Seven Bridges of Königsberg"*, e muito desenvolvida por G. Birkhoff e outros nos anos 1930's. Ao longo do tempo, muitos conceitos e modelos estatísticos foram propostos para a sua análise. Os *softwares* utilizados foram-se multiplicando, tendo tornado possível a incorporação da dinâmica temporal nas redes.

Nesta dissertação foram utilizadas ambas as metodologias de análise com o intuito de se complementarem com vista a definir um Sistema de Recomendação. A base de dados principal utilizada neste estudo é composta por informações relativas a coautorias. A ideia principal é a criação e análise de método de aconselhamento dos autores de forma a, com maior facilidade identificarem investigadores trabalhando nas mesmas áreas, que poderão vir a ser eventuais coautores para as suas publicações.

Esta dissertação é composta por sete Capítulos e 34 Anexos. No Capítulo 2 é feito o Estado da arte relativo a ambas as análises abordadas. No Capítulo 3 são estudados conceitos sobre Grafos, Redes e Redes Sociais. O Capítulo 4 é dedicado à Análise

Conceptual Formal. Após a exposição da teoria subjacente a este estudo, no Capítulo 5 é explicada a Metodologia a ser posteriormente utilizada na análise dos dados de coautoria, com recurso a dados sobre o comércio de cereais na União Europeia. Por fim o Capítulo 6 é dedicado à análise dos dados de coautoria, e à proposta do Sistema de Recomendação. As conclusões finais são apresentadas no Capítulo 7.

2 Estado da Arte

Estudadas há várias décadas, a Análise de Redes e a Análise Conceptual Formal têm vindo a ganhar importância ao longo dos anos. Embora a análise conceptual formal seja estudada há mais tempo, a análise de redes tem ganho uma elevada importância no âmbito da investigação.

Com a generalização na sociedade de conceitos como os de redes, redes sociais e redes dinâmicas, investigadores em todo o mundo têm desenvolvido estudos ao redor desta temática, desvendando os seus diversos usos. Com a consciencialização do Homem como ser social e a forte ligação do mesmo ao mundo informático, o estudo das redes sociais tem tido um forte contributo para o estudo dos comportamentos do Homem na sociedade. Com a criação de aplicações como o Facebook, Hi5, Linked in, etc., as comunicações e as ligações informatizadas possibilitam a criação uma enorme quantidade de dados utilizados para o estudo comportamental da sociedade. Estes estudos têm por base o estudo das redes; inicialmente estas eram consideradas estanques no espaço temporal, no entanto atualmente já são feitas de forma dinâmica no espaço temporal. Assim, através da teoria das redes têm-se estudado as possibilidades comunicativas e de aprendizagem nas redes sociais na internet, maioritariamente nos adolescentes. A proliferação da utilização da teoria das redes tem originado a criação de variados programas como o *AllegroGraph, AutoMap, Centrifuge Visual Network Analytics, Commetrix, Detica NetReveal, Gephi, etc.*. Estes programas são um importante contributo para os estudos com redes.

Esta metodologia tem em muito contribuído para o mundo da ciência, Powell *et al.* (2005) desenvolveram e testaram quatro alternativas de aproximação para a estrutura e dinamização de colaborações entre organizações no ramo da biotecnologia. Neste estudo Powell *et al.* testaram estas aproximações assumindo quatro métodos: uma aproximação por vantagem acumulativa, uma aproximação por homogeneização, uma por seguimento da tendência atual e por fim uma por multiconetividade. No entanto, como é depreendido estes desenvolvimentos podem ser utilizados noutros ramos inclusive na sociologia estudando os comportamentos de aproximação interpessoais. M. E. J. Newman tem-se tornado um importante estudioso na área da análise de redes, estudando e desenvolvendo várias metodologias para esta análise. Em 2003, (Newman

(2003)) Newman elaborou um artigo sobre as estruturas e funções de redes complexas, abordando vários desenvolvimentos na área, como conceitos sobre correlações de redes, modelos de grafos aleatórios, modelos de crescimento de redes e aproximação preferencial, processos dinâmicos na rede, etc.. Em 2004 (Newman (2004)), incidiu os estudos em redes com pesos, aqui quebrou um tabu da complexidade de análise destas redes, comprovando que é possível a utilização de técnicas usuais para análise das redes sem pesos em redes com pesos, usando um simples mapeamento de uma rede com pesos para um multigrafo sem pesos. Também em 2004 (Newman e Girvan, 2004), publicou um estudo juntamente com M. Girvan para a deteção de comunidades em redes. Aqui Newman e Girvan desenvolveram algoritmos para a separação da rede em subgrupos (comunidades), em que estes possuíam uma nova medida para definir a força da estrutura das comunidades; esta era capaz de devolver uma medida objetiva para a escolha do número de comunidades em que a rede se devia dividir. A deteção de comunidades tem-se tornado muito relevante no estudo das redes, assim como Newman e Girvan, também Bo Yang, Jin Di, Jiming Liu e Dayou Liu dedicaram os seus esforços no desenvolvimento de um sistema de deteção de comunidades na análise de redes. Yang e a restante equipa decidiram explorar a estrutura da deteção de comunidades por uma perspetiva probabilística e criaram um algoritmo para tal denominado de PMC (Yang *et al.* (2012)). Como se pode verificar muitos são os estudos feitos nesta temático, e assim continuará pois esta metodologia tem dado muitos contributos nas mais diversas áreas.

A Análise Conceptual Formal tem como principal função a definição e hierarquização de grupos. Assim como a análise de redes, esta análise tem tido um importante contributo no estudo da sociedade. No entanto, esta teoria apenas se começou a difundir pelo universo da investigação há cerca de dez anos, antes disso, poucos investigadores recorriam a esta teoria. Em 2000 vários investigadores implementaram em várias aplicações de larga escala, com mais notoriedade na implementação no *Knowledge Exploration System* para o uso na engenharia civil em cooperação com o *Ministry for Civil Engineering of North-Rhine Westfalia*. No entanto, não conseguiram obter muito sucesso a nível mundial (Eschenfelder *et al.* (2000)).

Nos últimos dez anos, a Análise Conceptual Formal tem crescido internacionalmente na comunidade de investigadores, disponibilizando-se aplicações para as mais diversas

disciplinas, como aplicações linguísticas, psicologia, inteligência artificial, reaquisição de informação, etc.

Os desenvolvimentos dos estudos das Correspondências de *Galois* na Matemática, foram deveras importantes para esta metodologia (Artin (1998)). Sistemas baseados na teoria das Correspondências de *Galois* são atualmente muito utilizados em aplicações para inteligência artificial. Algoritmos como GALOIS (Carpineto e Romano (1993), TITANIC (Stumme *et al.* (2002)), GALÍCIA (Valtchev *et al.* (2002)) e Zoo M (Pernelle *et al.* (2002)) foram todos desenvolvidos com base no reticulado de conceitos de *Galois*. A principal funcionalidade destes algoritmos centra-se na procura de agrupamentos/comunidades, que são maximais para um conjunto de atributos, o que mediante a temática dos estudos pode ser utilizado na mais diversas aplicações.

Com a proliferação destas metodologias, novos desenvolvimentos têm sido feitos, aumentando assim o número de algoritmos disponíveis para serem utilizados numa análise conceptual - é o caso do FCbO analisado e utilizado nesta dissertação (Krajca (2010) e Outrata e Vychodil (2012)).

Num mundo e num tempo em que a informação e o seu tratamento são temáticas importantes, os sistemas de recomendação são muito procurados. Já não é possível apenas a utilização de programas para catalogar e armazenar dados, é também necessária a utilização destes dados para fornecer informações relevantes e potencialmente úteis.

Mathias Bank e Juergen Franke (Bank e Franke (2010)), fizeram um relevante contributo nesta área com a criação de um sistema de recomendação passível de ser utilizado em vários portais de *e-commerce*. A inovação destes investigadores foi a inclusão de redes sociais de confiança como fonte de dados alternativa, demonstrando como é possível utilizar estes dados no cálculo das medidas de satisfação e relevância dos produtos e providenciando informações relevantes de potenciais clientes.

Num estudo de Patrick du Boucher-Ryan e Derek Bridge (Boucher-Ryan e Bridge (2006)) foram apresentados dois novos algoritmos para pesquisa de vizinhos num sistema de recomendação colaborativo. Ambos algoritmos usam o reticulado de conceitos como *índex* para a matriz de hierarquização do recomendador. Com este sistema houve um elevado incremento na facilidade de pesquisa de vizinhos, garantindo que não se perde precisão ou cobertura.

3 Grafos, Redes e Redes Sociais

3.1 Teoria dos Grafos

A teoria dos Grafos utiliza estruturas matemáticas que modelizam relações entre objetos de um determinado grupo. A primeira referência a este tema foi publicada em 1736 por Leonhard Euler no *paper "Seven Bridges of Königsberg"*, que trata o mediático problema das sete pontes de *Königsberg*. O problema retrata a cidade de *Königsberg* construída no Rio *Pregel* incluindo duas grandes ilhas ligadas entre si e ao continente por sete pontes. O problema consistia em saber se seria possível seguir um caminho que atravessa cada uma das pontes exatamente uma vez e retornar ao ponto de partida. Para a resolução deste problema Leonard Euler construiu um grafo representando as ligações possíveis.

Um grafo, é definido como um conjunto de vértices e arestas (Newman (2003)). Numa outra nomenclatura os grafos podem ser denominados de redes, os vértices de nós e as arestas de ligações. Os vértices estão ligados pelas arestas determinando uma relação de algum tipo entre eles, quer seja dirigida ou não dirigida. Um grafo é aqui representado por G=(V,E) em que V é um conjunto de vértices e E um conjunto de arestas.

Segundo Easley e Kleinberg (2010), redes são:

"Any collection of objects in which some pairs of these objects are connected by links"

As relações nos grafos são não dirigidas quando não existe uma ordem na ligação dos vértices, ou seja, é indiferente dizer que o vértice x está ligado ao vértice y ou dizer que o vértice y está ligado ao vértice x. Pelo contrário, as relações dirigidas possuem uma ordem de ligação, pois aqui o vértice x pode ter uma ligação com o vértice y, mas o inverso não se verificar.

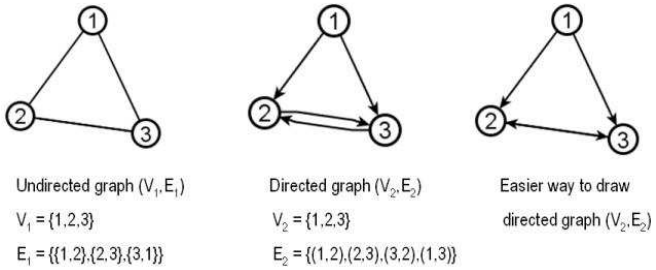

Figura 1 – Representações de grafos dirigidos e não dirigidos. O primeiro Grafo é não dirigido pois as ligações não possuem uma direção, o segundo e terceiro grafo são dirigidos pois as ligações têm uma direção (Hoppe (2007)).

Os Grafos tradicionais possuem no entanto algumas limitações, nomeadamente o facto de serem estacionárias no espaço temporal. No estudo de casos reais os objetos (vértices) possuem um grande dinamismo, é complicado obter toda a informação sobre estes e acima de tudo é irrealista admitir que a rede representada não sofre alterações ao longo do tempo (Hoppe (2009)).

3.2 Conceitos sobre redes

Para além dos conceitos básicos de redes, nós e ligações, existem também alguns conceitos importantes para a análise das redes.

De acordo com Diestel (2005), a ordem de um grafo G é a cardinalidade de V(G), ou seja, o número total de vértices n. Analogamente, o tamanho de G é a cardinalidade de E(G), isto é, o número total de arestas m. Para os grafos indiretos o número máximo de arestas é $m_{max} = \frac{n(n-1)}{2}$ e para os diretos é $m_{max} = n(n-1)$[(1)]. Um ***loop*** é uma aresta que liga um nó a si mesmo, enquanto que uma aresta simples faz a ligação entre dois vértices. Um **caminho** (*path*) de um grafo corresponde a uma sequência de vértices em que pares consecutivos de vértices estão ligados por arestas. Um **ciclo** (*cycle*) é um caminho fechado, em que o nó final coincide com o nó inicial.

Uma **componente conexa** é uma parte conexa maximal do grafo (subgrafo), onde para qualquer par de vértices existe pelo menos um caminho de um vértice para o outro. O subgrafo tem de ser uma parte autónoma do grafo, não ligada a uma parte maior (Easley e Kleinberg (2010)).

---

(1) Se forem permitidos *loops*, é necessário adicionar n às referidas fórmulas, obtendo $m\_max = n(n+1)/2$ para os grafos indiretos e de $m\_max = n(n+1)$ para os grafos diretos.

Numa matriz de adjacência A, em que a_{ij} corresponde às suas entradas, $a_{ij} = 1$ se *i* e *j* estiverem ligados por uma aresta (vizinhos diretos), e $a_{ij} = 0$ no caso contrário.

3.2.1 Ligações e nós

Existem três tipos de ligações relevantes: as pontes *(Bridges)*, os buracos estruturais *(Structural Holes)* e os *cliques*.

As **pontes** são ligações entre dois nós pertencentes a redes diferentes, originando assim uma união entre duas redes distintas criando uma rede maior constituída por dois subgrupos. Num contexto de redes sociais, as pontes constituem um grande interesse para os indivíduos acederem a novas informações e recursos, dado que elas facilitam a difusão de informação entre as comunidades (Kossinets e Watts (2006)). Dentro da temática das pontes, podemos também encontrar pontes locais *(Local Bridges)*, que ao invés de ligar dois grupos distintos, aproximam subgrupos criando uma ligação mais direta, ou seja mais rápida. Este tipo de pontes é ideal para facilitar a difusão de informações dentro da própria rede.

Buracos estruturais são buracos estáticos numa rede que impedem a comunicação entre componentes conexos distintas, mas que podem estrategicamente ser preenchidos adicionando uma ligação (Burt (1992)). Estes buracos existem maioritariamente em redes grandes e diversificadas. A utilização de pontes locais é uma possível solução para aumentar a facilidade de difusão da informação dentre da rede. As redes sociais são um bom exemplo de presença destes buracos. Por exemplo, uma pessoa possui, normalmente, vários grupos de "amigos" que apesar de pertencerem todos à mesma rede podem no entanto nem se conhecer, devido às diversas atividades e diferentes ambientes que a pessoa pode possuir.

Um ***clique*** é um grupo de vértices de um subgrafo em que todos os pares de vértices estão ligados entre si, ou seja, todos os vértices são vizinhos uns dos outros. Normalmente, um *clique* é denotado por K_n, onde n indica o número de vértices, ou seja, o tamanho do *clique*.

O tipo de nós mais importante para analisar são os ***hubs***. Estes caracterizam-se pela elevada quantidade de ligações que possuem. Naturalmente, os *hubs* são a principal fonte de difusão de informações dado serem os que conseguem partilhar informações com mais facilidade e para um maior número de nós. Aqui também se pode introduzir o

conceito de **Autoridades** ou *Authorities*, nós que recebem muitas ligações. Este conceito só se aplica em grafos com ligações dirigidas *(directed graphs).*

3.3 Medidas estatísticas para análise de grafos

As medidas estatísticas que irão ser estudadas contribuem para uma melhor análise de redes (Oliveira e Gama (2011)).

O conceito de **distância geodésica** é o conceito básico necessário para todas as análises estatísticas. A distância geodésica entre dois vértices é o número de arestas que os ligam pelo caminho mais curto, ou seja, pelo caminho constituído pelo menor número de nós. No entanto, nem em todos os grafos todos os pontos possuem ligações entre si. Nestes casos, a distância entre eles é considerada infinita. A distância geodésica média entre todas as combinações de vértices numa rede é usualmente denotada por *l* e dada por:

$$l = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij} \tag{3.1}$$

onde d_{ij}é a distância geodésica entre os vértices i e j. É de notar que é usado $\frac{1}{2}n(n+1)$ em vez de $\frac{1}{2}n(n-1)$, para incluir a distância de cada vértice a si próprio (que normalmente é nula). Como referido anteriormente, existe a possibilidade de a distância entre dois vértices ser infinita, o que iria causar que a média fosse também infinita. Para combater tal problema é possível utilizar a distância geodésica média harmonizada, alterada de modo a transformar as distâncias infinitas em distâncias nulas. É dada por:

$$l^{-1} = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} \frac{1}{d(i,j)} \tag{3.2}$$

A **excentricidade** de um vértice *v* pertencente a um grafo G, denotada por ϵ_v, é a distância de *v* ao vértice mais distante de *v*, utilizando o caminho mais curto. A excentricidade esta relacionada com conceitos como **raio** e **diâmetro** do grafo, em que o raio *R(G)* é dado pelo menor valor das excentricidades dos vértices de *V(G),* e o diâmetro *D(G)* é o maior destes valores.

$$\epsilon_v = \max_{i \in V(G) \setminus v} d(v, i) \tag{3.3}$$

A **densidade** de um grafo explica o nível geral de conectividade da rede e caracteriza-a como dispersa *(sparse)*, quando possui uma densidade baixa ou densa *(dense)*, quando a densidade é elevada. A densidade é a proporção de arestas (*m*) de um grafo *(G)* relativamente ao número máximo de arestas m_{max} (Equação 4).

$$\rho(G) = \frac{m}{m_{max}} = \frac{m}{\frac{1}{2}n(n-1)} = \frac{2m}{n(n-1)}, 0 \leq \rho \leq 1 \tag{3.4}$$

Quando lidando com grafos diretos, a equação é similar, definida pela proporção de arcos presentes no grafo *D* (Equação 5).

$$\rho(D) = \frac{m}{m_{max}} = \frac{m}{n(n-1)}, 0 \leq \rho \leq 1 \tag{3.5}$$

Em ambas as situações, a densidade varia de um mínimo de 0, quando o grafo não possui nenhuma aresta/arco, e um máximo de 1, quando o grafo é completo e possui arestas ligando todos os vértices, fazendo com que $m = m_{max} = \frac{n(n-1)}{2}$.

Relativamente às medidas orientadas para as arestas, a ***embeddedness*** de uma aresta é dada pelo número de vizinhos comuns partilhados pelos seus vértices. Uma outra medida é a **reciprocidade** *(reciprocity)*, que é uma quantidade específica para os grafos dirigidos que mede a tendência de pares de vértices para formar ligações simétricas entre eles. A reciprocidade é denotada por *r(D)*, e definida pela proporção de díades[(2)] num grafo (Equação 6).

$$r(D) = \frac{s(D)}{C_2^n} = \frac{mut(D) + null(D)}{C_2^n}, 0 < r < 1 \tag{3.6}$$

onde *mut(D)* corresponde às díades com ligações mútuas e *null(D)* às díades sem ligação.

[(2)] Segundo Wasserman e Faust (1994), as díades são subgrafos de grafos dirigidos, constituídos por dois nós e um possível arco entre eles. As díades simétricas podem nulas ou mútuas e o número de díades simétricas num grafo *D* é denominado de *s(D)*.

Como se pode observar na Figura 2 as díades podem ser nulas, assimétricas ou mútuas. Uma díade nula possui dois vértices sem qualquer ligação entre si. As díades assimétricas possuem uma aresta direcionada em apenas uma direção, enquanto que nas mútuas a aresta de ligação possui as duas direções (Wasserman e Faust (1994)). Assim, o número de ligações simétricas é a soma entre o número de ligações mútuas *(mut(D))* e o número de ligações nulas *(null(D)) –> s(D)=mut(D)+null(D).*

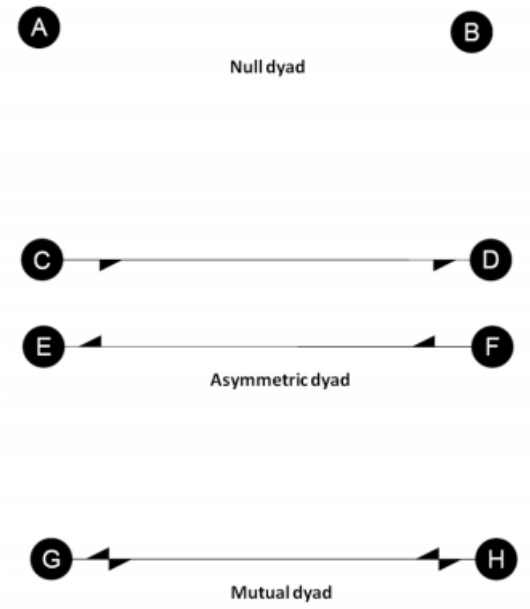

Figura 2 – Representação dos diversos tipos de arestas possíveis num grafo dirigido. [A,B] não possuem uma aresta, logo não têm qualquer ligação. [C,D] possuem uma aresta na direção C para D. [E,F] possuem uma aresta na direção F para E. [C,D] e [E,F] constituem uma díade assimétrica porque a direção é só num sentido. [G,H] é uma díade mútua a sua ligação é constituída por uma aresta com ambos os sentidos, significando que G possui uma relação com H e H com G. (Oliveira e Gama (2011))

Em suma, o valor da reciprocidade representa a probabilidade de dois vértices partilharem o mesmo tipo de ligação num grafo direto. No entanto, a definição mais popular de reciprocidade considera ligações mútuas em vez de simétricas[(3)]. Estas são calculadas pelo rácio entre o número de díades mútuas e o número de díades não nulas, demonstrado na Equação 7. Nos grafos indiretos, a reciprocidade é sempre máxima *(r(G)=1),* dado que todos os pares de vértices são simétricos (ou mútuos).

[(3)] Simétrico é um conceito mais vago, dado que engloba as ligações mútuas e as nulas.

$$r(D) = \frac{mut(D)}{mut(D)+asym(D)}, 0 < r < 1 \quad (3.7)$$

A ***betweenness of an edge*** (Equação 8) é definida como o número de caminhos geodésicos entre vértices que utilizam uma determinada aresta (Newman e Girvan (2004)). Esta equação mede a proporção dos caminhos mais curtos a passar por uma dada aresta; b_e é a *betweenness of an edge*, $\sigma_{uv}(e)$ expressa o número de caminhos mais curtos contendo a aresta *e*, σ_{uv} é o número de caminhos mais curtos entre os vértices *u* e *v:*

$$b_e = \sum_{u,v \in V(G) \setminus u,v} \frac{\sigma_{uv}(e)}{\sigma_{uv}} \quad (3.8)$$

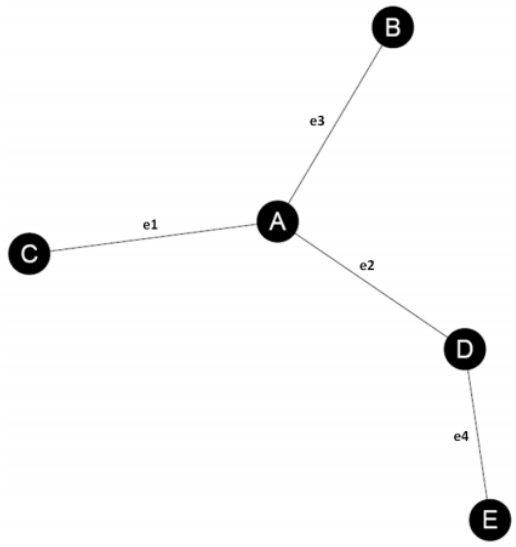

Figura 3 – Exemplo de um grafo constituído pelos vértices {A,B,C,D,E} e pelas arestas {e1,e2,e3,e4}. (Oliveira e Gama (2011))

Na Figura 3, como existe apenas uma distância geodésica entre cada par de vértices $\sigma_{uv} = 1, \forall_{u,v \in V(G) \setminus u,v}$, as medidas de *betweenness* serão as seguintes:

$$b_{e1} = |\{(C,A),(C,A,B),(C,A,D),(C,A,D,E)\}| = 4$$
$$b_{e2} = |\{(A,D),(B,A,D),(A,D,E),(C,A,D),(B,A,D,E),(C,A,D,E)\}| = 6$$
$$b_{e3} = |\{(B,A),(B,A,C),(B,A,D),(B,A,D,E)\}| = 4$$
$$b_{e4} = |\{(E,D),(E,D,A),(E,D,A,C),(E,D,A,B)\}| = 4$$

3.3.1 Medidas de centralidade

As medidas de centralidade vão permitir determinar a posição de um determinado vértice na estrutura de um grafo. Assim vai ser possível descobrir quais os vértices mais centrais, ou seja, com mais ligações. Estes vértices são importantes pois é através deles que a informação consegue fluir com mais rapidez. As medidas mais utilizadas são o *degree, betweenness, closeness e eigenvector centrality*. As primeiras três medidas foram propostas por Freeman (1979), para serem utilizadas apenas em redes binárias, ou seja sem pesos[4] associados às arestas. Mais tarde, Opsahl *et al.* (2010) propuseram essas mesmas medidas, mas para redes com pesos, ou seja redes que podiam conter valores nos vértices e/ou arestas. A quarta e última medida foi proposta por Bonachi (1987).

O **grau** (*degree*) de um vértice *v* mede o nível de envolvimento do mesmo na rede, é calculado utilizando a Equação 9, e determina o número de vizinhos do vértice. N_v corresponde à vizinhança do vértice *v*, em que no caso dos grafos indiretos é constituída por qualquer vértice que esteja ligado a *v*. No entanto, nos grafos diretos é necessário ter em atenção a direção da ligação, e aqui teremos dois tipos de *grau*, o *In-degree* e o *Out-degree*. No ***In-degree***, a vizinhança é constituída por todos os vértices que estejam ligados ao vértice *v* (direcionados para *v*), enquanto que para o ***Out-degree***, a vizinhança é definida por todos os vértices aos quais o vértice *v* se liga (direcionado para fora de *v*).

$$k_v = |N_v|, 0 \leq k_v \leq n \tag{3.9}$$

A níveis mais gerais, é possível determinar a conectividade do grafo (Equação 10), apenas calculando a média do *grau* de todos os vértices do grafo (Costa *et al*. (2008)).

$$\bar{k} = \frac{1}{n}\sum_{i=1}^{n} k_i \tag{3.10}$$

[4] Uma rede com pesos é uma rede que possui arestas com forças diferentes, podendo estas serem mais fortes ou mais fracas. Estas são matematicamente representadas por uma matriz de adjacência com entradas que não são apenas zero e um, mas que são iguais aos pesos das arestas. (*Newman* (2004))

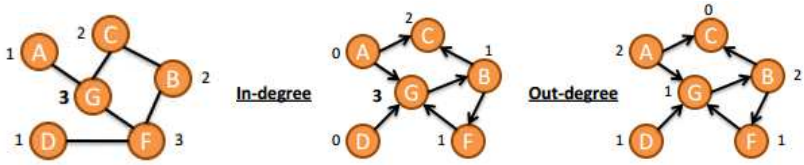

Figura 4 – Grafo não dirigido com o cálculo dos respetivos graus em cada vértice e dois grafos dirigidos com o cálculo do *In-degree* e do *Out-degree*, respetivamente. (Oliveira e Gama (2012))

No caso da Figura 4 o *grau* do vértice G é de $k_G = 3$ para o grafo não dirigido, nos grafos dirigidos o *In-degree do vértice G é de* $k_G = 3$ e o *Out-degree* de $k_G = 1$. $\bar{k} = 2$ é a média do grau dos vértices no grafo não dirigido.

Para os grafos com pesos o equivalente ao *grau* é a ***strength***, dada pelo somatório dos pesos das ligações adjacentes ao vértice *v* (Equação 11).

$$k_v^w = \sum_{u \in N_v} w_{vu} \tag{3.11}$$

onde w_{vu} são os pesos das ligações adjacentes ao vértice *v*.

A medida de ***betweenness de um vértice*** é dada pelo número de caminhos geodésicos entre dois outros vértices que passam pelo vértice (Equação 12). Vértices com elevada *betweenness* ocupam uma posição importante na rede, pois permitem que sejam usados como meio de ligação entre diferentes regiões da rede.

$$b_v = \sum_{s,t \in V(G) \setminus s,t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{3.12}$$

Tomando o exemplo da Figura 3, as medidas de *betweenness* dos vértices seriam:

$$b_A = |\{(C,A,B),(C,A,D),(B,A,D),(C,A,D,E),(B,A,D,E)\}| = 5$$
$$b_B = 0$$
$$b_C = 0$$
$$b_D = |\{(A,D,E),(C,A,D,E),(B,A,D,E)\}| = 3$$
$$b_E = 0$$

A terceira medida de *Freeman's* é a ***closeness*** (Equação 13), esta serve para medir a posição global do vértice na rede. Nas redes sociais, esta medida é maioritariamente utilizada para calcular o quão rápido um ator consegue chegar a qualquer outro na rede. Isto permite verificar a facilidade de difusão da informação pela rede.

$$Cl_v = \frac{n-1}{\sum_{u \in V(G) \setminus v} d(u,v)} \tag{3.13}$$

em que *d(u,v)* é a distância geodésica entre os vértices *u* e *v*.

A quarta medida é a ***eigenvector centrality***, proposta por Bonachi (1987). Esta medida é baseada na ideia de que o poder e estatuto de um vértice são definidos pelo poder e estatuto dos seus vizinhos. O *eigenvector* de um vértice é proporcional à soma das *eigenvector centralities* dos seus vizinhos diretos.

A *eigenvector centrality* é dada pela Equação 14, em que $x_i \backslash x_j$ corresponde à centralidade dos vértices *i* e *j*, a_{ij} é a entrada na matriz de adjacência A. Por fim λ representa o maior valor próprio de A. Esta medida será das mais adequadas para o cálculo grau do vértice, pois não tem apenas em conta a quantidade de vizinhos mas também a qualidade dos mesmos.

$$x_i = \frac{1}{\lambda} \sum_{j=1}^{n} a_{ij} x_j \tag{3.14}$$

Por fim a **Modularidade** *(Modularity)* permite a deteção e definição de comunidades numa rede. Uma comunidade é constituída por um grupo de nós muito próximos entre si, mas distantes dos restantes (McSweeney (2009)). Inicialmente cada nó *n* possui a sua comunidade, no entanto depois são analisados todos os nós vizinhos *v* calculando a existência de um ganho na modularidade ao retirar n da sua comunidade e incluir na comunidade *v* (Mendonça *et al.* (2009)). Dada uma partição (conjunto de nós) P e uma rede G modularidade é definida por:

$$Q(P,G) =< N, E >) = \sum_{C_i \in P} \frac{l(C_i)}{|E|} - (\frac{d(C_i)}{2.|E|})^2, \tag{3.15}$$

onde N é o conjunto total de nós da rede, E corresponde ao conjunto de ligações, C_i é uma classe da partição, $l(C_i)$ é o número de ligações que ligam os nós dentro de C_i e $d(C_i)$ é o grau de C_i, ou seja é o número de ligações que ligam os nós de C_i aos nós exteriores (McSweeney (2009)).

3.4 Análise de Redes sociais

A Análise de Redes Sociais (ou SNA) foi criada com o intuito de estudar o comportamento da sociedade. Análise de Redes Sociais é o mapeamento e estudo de relações e fluxos entre pessoas, grupos, organizações, entre outros. A SNA fornece análise tanto visual como matemática das relações humanas (Krebs (2000)). Este tipo de análise foca-se em pequenas redes com dois ou três tipos de ligações, mas apenas com um tipo de nós, estudada apenas num espaço temporal estanque e com informação quase perfeita (Carley (2003)). Esta metodologia de estudo de redes sociais foi um marco importante no estudo da sociedade, pois já permitiu o estudo de redes com mais de um tipo de ligações. Atualmente estudam novos tipos de redes, as redes dinâmicas, que introduziram a análise ao longo do tempo e com diversos tipos de nós, o que permitirá um estudo mais aprofundado do comportamento da sociedade.

3.4.1 Alguns conceitos específicos das redes sociais

Grande parte das teorias relativas a redes socias é baseada nos conceitos das teorias de redes e grafos. No entanto, os seguintes conceitos são característicos apenas das SNA.

Triadic Closure: Este conceito é apenas aplicado quando a análise da rede é feita ao longo do tempo, pois é baseado no princípio de que se duas pessoas possuem um amigo em comum então é muito provável que estes se tornem também amigos no futuro. Este é o princípio da transitividade (Rapoport (1953)) e é mais facilmente verificado quando o conjunto de indivíduos beneficia com a relação.

Structural Equivalence: É uma noção matemática que expressa as semelhanças entre atores numa rede social baseada nos vizinhos que partilham, ou seja, nas ligações idênticas que possuem (Lorrain e White (1971)). Deste modo, dois atores dizem-se estruturalmente equivalentes quando possuem exatamente os mesmos vizinhos, o que significa que podem trocar de lugares sem que se tenha de alterar a estrutura da rede.

Uma aproximação da *Structural Equivalence* pode ser usada para encontrar grupos, dado permitir identificar pares similares de atores.

Brokers: São considerados facilitadores (intermediários) nas interações entre membros de diferentes grupos, atuando como pontes *(Bridges)* em buracos estruturais *(Structural holes).*

Peers: Atores que são semelhantes no que respeita a idade, educação e classe social.

A análise das redes sociais é normalmente efetuada tendo como foco a sociedade como um todo *(Sociocentric approach)* ou os atores *(Egocentric approach)*. Como é facilmente depreendido, estas análises são diferenciadas pelo tipo de foco, enquanto que uma se foca maioritariamente na sociedade analisando o seu comportamento, a outra prefere um foco nos atores conseguindo uma análise especializada do seu comportamento. Estes tipos de análise foram desenvolvidos independentemente mas aproximadamente ao mesmo tempo (Chung *et al.* (2005)).

O foco na sociedade como um todo é a técnica mais utilizada, pois permite adquirir dados de toda a sociedade, ou seja de toda a rede. Este foco identifica tendências nos relacionamentos dos atores presentes no grupo em estudo, e explica os resultados dessas tendências.

Quando o foco é centrado nas relações de cada indivíduo com os seus vizinhos perdemos a noção do comportamento da rede como um todo, no entanto conseguimos informações mais pormenorizadas do comportamento do individuo em estudo. Este ator é denominado de *Ego*, os indivíduos com que ele está ligado são os *Alters* e as redes do *Ego* podem ser representadas com o *Ego* ou sem o *Ego.* Para esta última, é tido por base que não é necessária a presença do *Ego* pois todos os atores estão ligados a ele.

A escolha do foco a usar vai depender apenas do tipo de análise que se quer efetuar. É mais usual o estudo dos comportamentos de uma sociedade, no entanto o foco no *Ego* pode ser bastante útil quando se pretende estudar apenas um indivíduo da rede.

3.5 Relações e ligações em redes de larga escala

Vários investigadores em Física e Sociologia fazem estudos da estrutura das redes de larga escala. As redes de larga escala tipicamente têm características próprias de estrutura local, como um nível de *clustering*[5] elevado, e de estruturas globais, como a distância média entre os nós elevada. Isto significa que estas redes vão possuir vários grupos com uma distância curta entre os nós do grupo, no entanto com uma grande distância entre os nós de grupos diferentes. As características globais e locais ajudam a definir a topologia das redes *Small World,* ou pequeno mundo, são grandes redes com *clustering* local e uma distância relativamente curta entre nós. Watts (1999), mostrou que ao adicionar apenas algumas ligações remotas a uma rede de larga escala onde o grau de *clustering* local é elevado, é suficiente para criar uma rede de pequeno mundo.

Nestas redes é possível encontrar nós predominantes, apresentando um número anormalmente elevado de ligações, denominados de *hubs*. Pesquisas efetuadas nesta área verificaram que normalmente as redes possuem nós com pequenas ligações e apenas uma pequena quantidade de nós com uma elevada quantidade de ligações (*hubs*). Lotka (1926) e Price (1965), mostraram que nas redes a proporção de nós de grau k, normalmente varia segundo a função $1/k^{\alpha}$, onde *alpha* é o *power coefficient*[6].

Quando é adicionado um novo nó este vai ter de escolher qual o nó, ou nós, ao quais se vai ligar, os *hubs* vão ser considerados preferenciais, levando a uma dinâmica de "*rich-get-richer"* ou acumulativa. Este fenómeno é verificado no popular jogo "*Kevin Bacon game*", onde os novos atores de Hollywood tendem a começar a sua carreira como atores secundários acompanhando atores famosos. No processo de expansão de redes, aqueles que experienciam o sucesso cedo, recebem a parte da figura dominante ("*lion's share*") de subsequentes recompensas (Merton (1973)). Albert e Barabási (2002), formularam que um novo nó decidirá ligar-se a um nó já existente, dependendo do número de ligações que este nó já possua.

---

[5] *Clustering* é uma metodologia de análise de dados usado para identificar grupos que partilham características comuns. Envolve a pesquisa de elementos que permitem identificar grupos ou *clusters* demonstrando características comuns (QFINANCE, 2009). Os grupos podem ter um *clustering* elevado ou baixo consoante os indivíduos são muito ou pouco similares.

[6] Mostra a eficiência com que um nó se liga a outro.

Powell, Koput e Smith-Doerr (1996), em trabalhos realizados na indústria da biotecnologia, concluíram que os nós (organizações) mais antigos, com menos ligações tinham mais probabilidades de falhar. Certamente, nós mais antigos têm mais tempo que os mais recentes para estabelecer ligações, mas Powell *et al.* (1996) concluíram que as ligações que estas tinham e as atividades que praticavam eram essenciais. Deste modo, a decisão de um nó se ligar a outro não depende apenas do número de ligações que este possui, mas também da "qualidade" das ligações e das atividades que efetua.

A possibilidade de introduzir atividades no estudo de redes, veio pôr em evidência mais métodos de agrupamento. Numa rede verifica-se uma tendência dos nós para imitar o comportamento dominante da maioria, criando-se assim uma cascata de interações em que todos os nós da rede atuam da mesma forma (White (1981); DiMaggio e Powell (1983)). Neste contexto, as ações são ativadas por um sentimento de necessidade, por um desejo de acompanhar os outros atuando apropriadamente (March e Olsen (1989)).

Uma outra alternativa ao comportamento *rich-get-richer*, é a homogeneização (McPherson e Smith-Lovin (1987)), em que os nós da rede preferem interagir com nós que sejam parecidos com eles; é mais fácil interagirem dado terem o mesmo tipo de comportamentos e gostos.

Por último, os nós podem estabelecer ligações através de um desejo de diversidade e multiconetividade (Powell (1990)). Aqui os nós preferem procurar algo diferente, transformando a sua rede numa mais alargada e diversificada. Deste modo, os nós têm acesso a uma quantidade de informação e relações muito maior do que, por exemplo, na teoria da homogeneização em que os nós partilham os mesmos interesses e consequentemente os mesmos conhecimentos. Assim os nós têm a oportunidade de tornar a sua rede mais diversificada, conseguindo uma maior quantidade de informação.

4 Análise Conceptual Formal

A *Formal Conceptual Analysis* (FCA) ou Análise Conceptual Formal foi introduzida em 1982 por Rudolf Wille (Wille (1982)). Este tipo de análise de dados não era muito popular na época devido à grande incidência da investigação na Matemática. No entanto, foi a partir do século XXI que esta começou a ser mais utilizada em base de dados de larga escala, pois até então foi apenas sobretudo desenvolvida por Wille e a sua equipa. A primeira uilização desta metodologia em Análsie de Dados remonta no entanto a Barbut e Monjardet(1970), em França.

Esta metodologia de análise tem por base duas definições principais, a de **extensão** e a da **intensão**. As extensões são os conjuntos maximais de objetos que partilham propriedades comuns, as intensões consistem em conjuntos maximais de atributos partilhados por conjuntos de objetos. Um conceito é formado por uma intensão e respetiva extensão. Aqui objetos correspondem a "indivíduos" na base de dados (G), atributos (M), como o próprio nome indica, são as características dos indivíduos. É regularmente utilizado uma relação binária (I) para definir a presença (1) ou ausência (0) do atributo no objeto. Assim, matematicamente os dados são representados pelo tripleto (G,M,I), usualmente designado por "contexto", em que (g,m) ϵ I, significa que, o objeto g possui o atributo m.

Para um A ⊆ G, definimos:

$$A' \coloneqq \{m \in M | \forall g \in A : (g, m) \in I\} \tag{4.16}$$

Para um B ⊆ M, definimos:

$$B' \coloneqq \{g \in G | \forall m \in B : (g, m) \in I\} \tag{4.17}$$

Assim A' é um conjunto de atributos comuns a todos os objetos em A e B' é o conjunto de objetos possuindo todos os atributos em B. Deste modo, um conceito do contexto (G,M,I), é definido por um par (A,B) onde A ⊆ G, B ⊆ M, A'=B e B'=A. A extensão do conceito é A, enquanto a intensão é B. Os conceitos possuem uma estrutura de **reticulado, - o denominado reticulado de *Galois*,** que representa a hierarquia dos conceitos, apresentando-a num formato de grafo (Godin, Missaoui e Alaoui *(*1995)).

O conjunto total dos conceitos do contexto (G,M,I) é denotado por C(G,M,I). Uma relação ordinal (≤) é facilmente definida no conjunto de conceitos através da Relação (18).

$$(A_1, B_1) \leq (A_2, B_2) : \Leftrightarrow A_1 \subseteq A_2 (\Leftrightarrow B_1 \supseteq B_2) \qquad (4.18)$$

Para melhor perceber este modelo, é apresentado abaixo um exemplo retirado de Carpineto e Romano (1993), sobre os planetas do nosso sistema solar.

	size			distance from sun		moon	
	small	medium	large	near	far	yes	no
Mercury	x			x			x
Venus	x			x			x
Earth	x			x		x	
Mars	x			x		x	
Jupiter			x		x	x	
Saturn			x		x	x	
Uranus		x			x	x	
Neptune		x			x	x	
Pluto	x				x	x	

Tabela 1 – Contexto formal dos planetas, possui informações relativas ao tamanho do planeta, podendo ser pequeno *(small)*, médio *(medium)* ou grande *(large)*, relativas à sua distância do sol, admitindo as possibilidades perto *(near)* e longe *(far)*, e por fim relativas à presença ou não de lua podendo esta ser sim *(yes)* ou não *(no)*. Os planetas da amostra são Mercúrio *(Mercury)*, Vénus *(Venus)*, Terra *(Earth)*, Marte *(Mars)*, Júpiter *(Jupiter)*, Saturno *(Saturn)*, Úrano *(Uranus)*, Neptuno *(Neptune)* e Plutão *(Pluto)* (Carpineto e Romano (1993)).

Apenas pela visualização da tabela é possível verificar que alguns planetas possuem exatamente as mesmas características. Utilizando as notações acima mencionadas, temos por exemplo:

A={Mercury, Venus} A'={small, near, no}

B={large, far} B'={Jupiter, Saturn}

Assim, por exemplo, C={{*Mercury,Venus*},{*small,near,no*}} é um conceito.

O grafo que representa o reticulado de *Galois* organiza os conceitos em vários níveis de agrupamento. Á medida que se sobe no grafo, o número de atributos vai diminuindo e são incorporados mais planetas; quanto menos atributos no conceito maior o respetivo número de planetas. No primeiro nível estão representados todos os atributos (intensão), que compreensivelmente nenhum planeta possui, logo este nível está associado a um

conjunto vazio de objetos (extensão). No segundo nível já são contemplados menos atributos na intensão de cada conceito, logo já se observa a existência de objetos nas extensões. No primeiro conceito, a extensão é composta pelos objetos *Mercury e Venus*, enquanto que a intensão contempla os atributos *size small, distance near e moon no*. À medida que se vai subindo de nível no grafo, o cardinal da intensão vai diminuindo, no entanto o cardinal da extensão vai aumentando. Mais uma vez no primeiro conceito mas do terceiro nível a intensão é constituída pelos atributos *small size* e *distance near*, no entanto aqui a extensão já inclui *Mercury, Venus, Earth* e *Mars.* O último nível é oposto ao primeiro, enquanto no primeiro a intensão era constituída por todos os atributos e a extensão era um conjunto vazio ({},{ss,sm,sl,dn,df,my,mn}), aqui a extensão é composta por todos os objetos, pois a intensão é um conjunto vazio de atributos ({Me,V,E,Ma,P,J,S,U,N},{}).

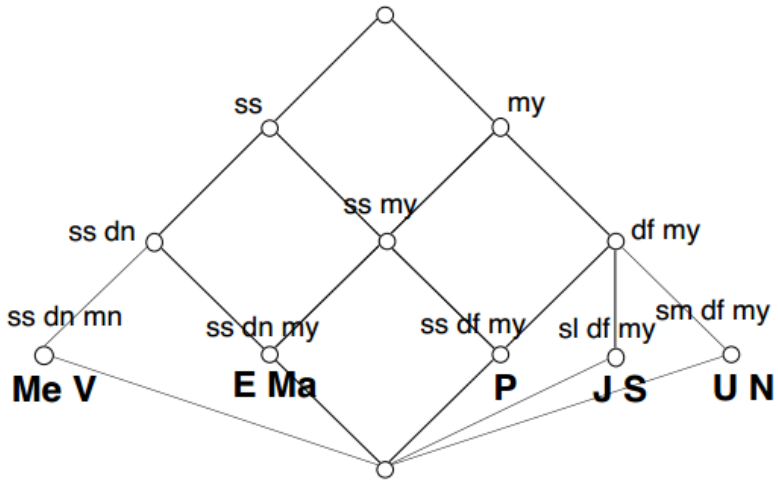

Figura 5 – Reticulado de conceitos do contexto dos planetas (Tabela 1) em que ss – size small; sm – size médium; sl – size large; dn – distance from the sun near; df – distance from the sun far; ms – moon yes; mn – moon no. Me – Mercury; V – Venus; E – Earth; Ma – Mars; P – Pluto; J – Jupiter; S – Saturn; U – Uranus; N – Neptune (Carpineto e Romano (1993)).

4.1 Algoritmo FCbO: Fast Close-by-One

O algoritmo FCbO é uma versão refinada do algoritmo CbO (*Close-by-One)* de *Kuznetsov* (Kuznetsov (1993) e Kuznetsov (1999)). O CbO é um algoritmo para o cálculo do conjunto de conceitos associado a um contexto (G,M, I), utilizando uma ordem de "baixo para cima". Com este algoritmo, um novo conceito é definido inicialmente calculando a interseção entre a intensão do atual conceito com alguns objetos situados no exterior da extensão do conceito. Esta será a nova intensão do conceito, a não ser que um conceito com a mesma intensão já tenha sido calculado. Depois a extensão do conceito atual é estendida de forma a incluir todos os outros objetos que contenham todos os atributos da nova intensão, assim como os seus atributos. O mecanismo de verificação da criação do conceito define uma ordem total no conjunto dos objetos. Depois é verificado se a nova intensão não está incluída num outro objeto exterior à extensão do atual conceito e possui uma ordem inferior à do objeto utilizado para calcular a intensão (Yevtushenko (2004)).

O FCbO (Krajca (2010) e Outrata e Vychodil (2012)) veio resolver um problema presente na maioria dos algoritmos de Análise Conceptual Formal, o cálculo múltiplo. O algoritmo FCbO consegue atingir uma melhor performance do que o CbO ao reduzir o número total de conceitos calculados várias vezes. A redução é feita introduzindo um teste de canonicidade adicional que reduz eficazmente a árvore do CbO durante o cálculo.

Neste algoritmo o teste canônico original usado no CbO é usado depois de o conceito formal ser determinado. Considerando um contexto formal (G,M,I), $^{\uparrow I}$: $2^X \rightarrow 2^Y$ e $^{\downarrow I}$: $2^Y \rightarrow 2X$, com B ⊆ M e j ∉ B:

$$B \cap M_j = D \cap M_j, \text{onde } D = (B \cup \{j\})^{\downarrow I \uparrow I} \; e \; M_j = \{m \in M | m < j\} \quad (4.21)$$

FCbO aplica um teste adicional realizado antes de D ser calculado, eliminando assim o cálculo de $^{\downarrow I \uparrow I}$. É de notar que (21) falha se $B \oslash j \neq \emptyset$, onde:

$$B \oslash j = (D \backslash B) \cap M_j = ((B \cup \{j\})^{\downarrow I \uparrow I} \backslash B) \cap M_j \quad (22)$$

O novo teste de canonicidade explora o facto de se (21) falha dado B e $j \notin B$, a monotonia de $^{\downarrow I \uparrow I}$ defende que o teste também falhará para cada $B' \supseteq B$ tal que $j \notin B'$. Assim, o novo teste de canonicidade é baseado nas seguintes declarações:

Sejam $B \subseteq M, j \notin B$, e $B \oslash j \neq \emptyset$. Então, para cada $B' \supseteq B$ tal que $j \notin B'$ e $B \oslash j \nsubseteq B'$, temos que $B^{'} \oslash j \neq \emptyset$.

O FCbO pode ser visto como uma extensão do CbO em que a informação é propagada sobre os conjuntos (22) que fazem parte do novo teste. De forma a aplicar o novo teste é necessário trocar a estratégia de procura do algoritmo de uma procura de "baixo para cima", usada no CbO, para uma combinação entre a procura de "baixo para cima" com uma de "cima para baixo". A grande alteração que melhorou o algoritmo FCbO foi que ao contrário do CbO, quando um teste tem êxito este não é processado recursivamente mas a informação sobre os conceitos é armazenada numa lista. Só depois de cada atributo ser processado é que são feitas as chamadas recursivas. Os testes de canonicidade são feitos com base na informação armazenada. Isto vai fazer com que a árvore com os conceitos do FCbO seja mais reduzida do que a do CbO, sendo assim mais sucinta e mais precisa.

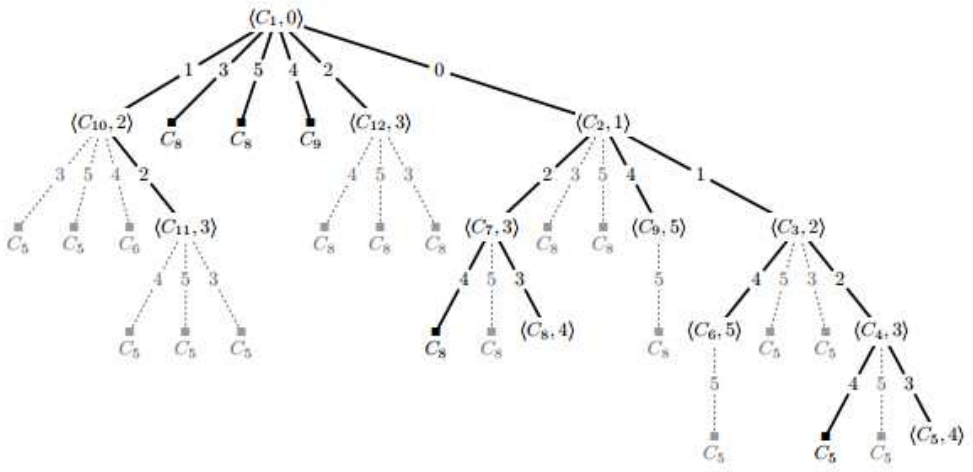

Figura 6 – Exemplo de uma árvore CbO reduzida para uma árvore FCbO. As linhas contínuas representam a árvore FCbO e as linhas tracejadas representam os conceitos da árvore CbO que não pertencem à árvore FCbO (Krajca et al. (2010)).

5 Metodologias de Análise de Redes e de Análise Conceptual Formal

Este capítulo servirá para explicar as metodologias posteriormente utilizadas na análise dos dados de coautorias, sendo estes os principais dados analisados nesta dissertação. Para o efeito, são utilizadas duas tabelas de dados para a explicação, uma retirada do *website* das Nações Unidas http://comtrade.un.org/db/dqBasicQuery.aspx, constituída pelas relações comerciais entre 28 Estados Membros da União Europeia, e outra criada com algumas características demográficas dos mesmos países, esta retirada da revista *"2012 World Population Data Sheet" – Population Reference Bureau*. Para evitar uma excessiva quantidade de informação para a primeira tabela são apenas consideradas as exportações de Cereais no ano de 2012. Devido ao seu tamanho é recomendada a visualização da base de dados originais através do *link* presente no Anexo 1. No entanto, desta tabela é apenas necessário retirar as ligações comerciais e os seus valores de modo a criar uma rede para posterior análise no *package Gephi*. Presente no Anexo 2, a rede de ligações comerciais é constituída por uma primeira coluna com o país exportador, uma segunda coluna com o país que recebe o bem e uma terceira com o valor comercial anual das transações. A segunda tabela de dados, corresponde a características demográficas dos 28 Estados Membros da União Europeia. Esta tabela possui um código binário que reflete se o país possui uma determinada característica (1) ou não (0) (Anexo 3). Como é compreensível, a tabela não possuía os dados em código binário nem estava preparada para tal. Na tabela 3 encontram-se os dados antes do pré-processamento, isto é, antes da categorização das variáveis originais (ver Anexo 3). Por exemplo, a variável *"Births per 1000 Population"* foi categorizada em "Nascimentos por cada mil habitantes inferiores a 10" *("Births per 1000 Population <10")* ou "Nascimentos por cada mil habitantes superiores a 10" *("Births per 1000 Population >10").*

	Population mid-2012 (millions)	Births per 1,000 Population	Deaths per 1,000 Population	Rate of Natural Increase %	Net Migration Rate per 1,000	2050 Population as a Multiple of 2012	Infant Mortality Rate
Denmark	5,6	11	9	0,1	4	1,1	3,5
Estonia	1,3	11	11	0	0	0,9	3,3
Finland	5,4	11	9	0,2	3	1,1	2,4
Ireland	4,7	16	6	1	-7	1,4	3,5
Latvia	2	9	14	-0,5	-4	0,8	5,7
Lithuania	3,2	11	13	-0,2	-12	0,9	4,3
Sweden	9,5	12	10	0,2	5	1,1	2,1
United Kingdom	63,2	13	9	0,4	4	1,3	4,3
Austria	8,5	9	9	0	4	1,1	3,6
Belgium	11,1	12	10	0,2	7	1,2	3,5
France	63,6	13	9	0,4	1	1,1	3,5
Germany	81,8	8	10	-0,2	3	0,9	3,4
Luxembourg	0,5	11	7	0,3	16	1,3	3
Netherlands	16,7	11	8	0,2	2	1	3,8
Bulgaria	7,2	10	15	-0,5	-1	0,8	8,5
Hungary	9,9	9	13	-0,4	2	0,9	4,9
Poland	38,2	10	10	0,1	0	0,9	4,8
Romania	21,4	9	13	-0,4	0	0,9	9,9
Slovakia	5,4	11	9	0,2	0	1	5,3
Croatia	4,3	9	12	-0,2	-1	0,9	4,4
Greece	10,8	10	10	0,1	4	1	3,8
Italy	60,9	9	10	-0,1	4	1	3,4
Malta	0,4	10	7	0,2	6	0,9	5,5
Portugal	10,6	9	10	-0,1	1	1	2,5
Slovenia	2,1	11	9	0,1	1	1	3
Spain	46,2	10	8	0,2	-2	1	3,2

Tabela 2 – Tabela original retirada do relatório *"2012 World Population Data Sheet"*

5.1 Metodologia das Redes

A criação das redes é efetuada com recurso a um *software* específico para criação e análise de redes denominado *Gephi*. Esta plataforma gratuita desenvolvida pela *Gephi Consortium*, vai permitir criar e analisar as redes das duas bases de dados apresentadas. Com este programa a visualização e compreensão das redes será facilitada; o programa fornece também uma ampla quantidade de estatísticas úteis para a análise.

O *package Gephi* possui a opção de fazer o *upload* e configuração de uma rede através do formato csv, no entanto esta não é muito eficaz, assim a criação de um ficheiro de extensão .gdf é aconselhada. Este tipo de ficheiro necessita de dois tipos de entradas facilmente codificadas num programa de texto como o *Notepad* ++, uma

inicial constituindo a informação dos nós e uma segunda e final com a informação das ligações (Anexo 4).

Para codificar os nós são necessárias no mínimo duas entradas, uma com o código único do nó e outra com o nome atribuído a esse código. Para esta análise foi criada mais uma entrada com o valor das exportações em milhões. Cada linha representa um nó e as suas características são separadas por uma vírgula, estas podem ser classificadas como texto usando o código VARCHAR ou como número usando o código INT (números inteiros). Esta codificação da informação do nó é colocada no cabeçalho em frente ao nome atribuído a essa informação (Figura 7).

```
nodedef>name VARCHAR,label VARCHAR,valor exportado INT
Austria,Austria,440
Belgium,Belgium,742
Bulgaria,Bulgaria,858
Croatia,Croatia,78
Czech Rep.,Czech Rep.,709
Estonia,Estonia,77
```

Figura 7 – Parte da codificação dos nós da rede constituído por 3 entradas, código do país, nome do país e valor das exportações em milhões de Euros.

A codificação das relações é similar, devendo ser colocada imediatamente abaixo do final da informação sobre os nós. Aqui usam-se apenas os códigos dos nós e não o seu nome de forma a evitar erros, pois vários nós, dependendo do tipo de informação, podem ter o mesmo nome. Nos exemplos aqui analisados vão ser utilizadas 3 entradas, uma com o nó de partida, uma com o nó de chegada e outra com um valor atribuído à relação (ver Figura 8).

```
Romania,Romania,0
Cyprus,Cyprus,0
edgedef>node1 VARCHAR,node2 VARCHAR,Weight INT
Austria,Belgium,1300649
Austria,Bulgaria,1647292
Austria,Croatia,1630369
Austria,Czech Rep.,11688936
Austria,Denmark,369739
```

Figura 8 – Parte da codificação das relações, constituída pelo país exportador, pelo país recetor do bem e pelo valor do bem em Euros. A mesma codificação é utilizada para o tipo de informação aplicando VARCHAR às entradas de texto e INT às entradas numéricas.

No *package Gephi*, o primeiro passo consiste em definir o tipo de grafo, devendo escolher-se "dirigido" ou "não dirigido". Considerando a transferência dos cereais, o mais indicado para esta situação é escolher um grafo dirigido, pois existe um sentido na transferência.

A rede inicial é criada de forma aleatória, posicionando os nós sem uma lógica aparente (ver Figura 9). Para facilitar a análise é possível utilizar um conjunto de algoritmos que alteram o posicionamento dos nós mediante uma determinada lógica. O algoritmo utilizado é denominado de *Force Atlas*, este algoritmo tem a capacidade de aproximar os nós mediante a força das suas ligações, criando assim grupos dentro da própria rede (Figura 10).

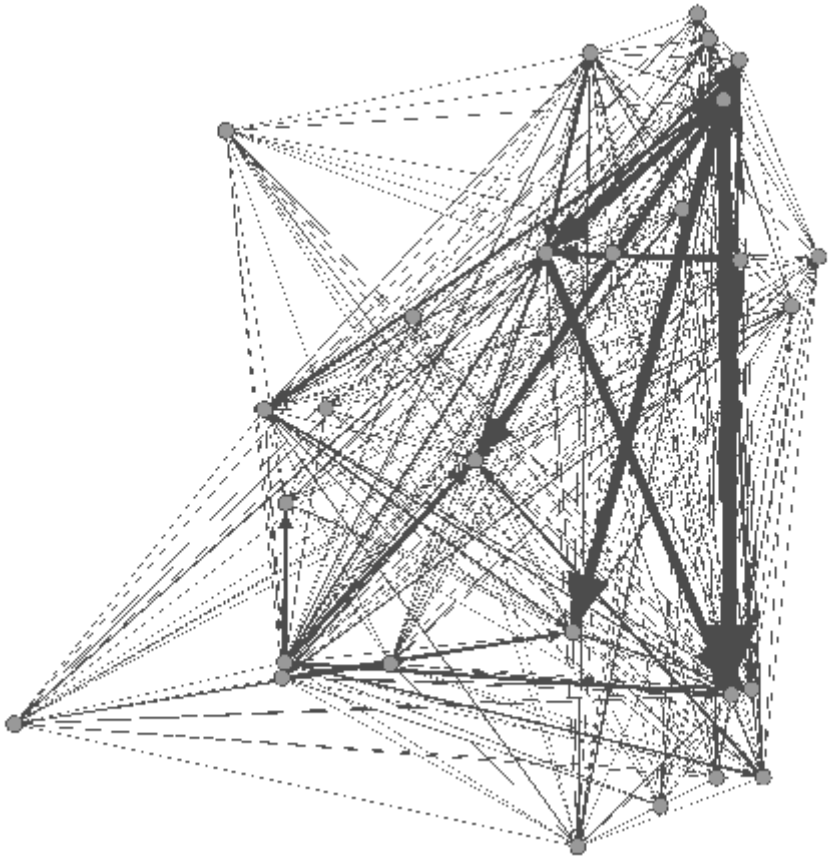

Figura 9 – Rede das exportações entre 28 Estados Membros da União Europeia. Rede inicial sem qualquer alteração do posicionamento dos nós. Esta pode ser diferente sempre que aberta no *package Gephi.*

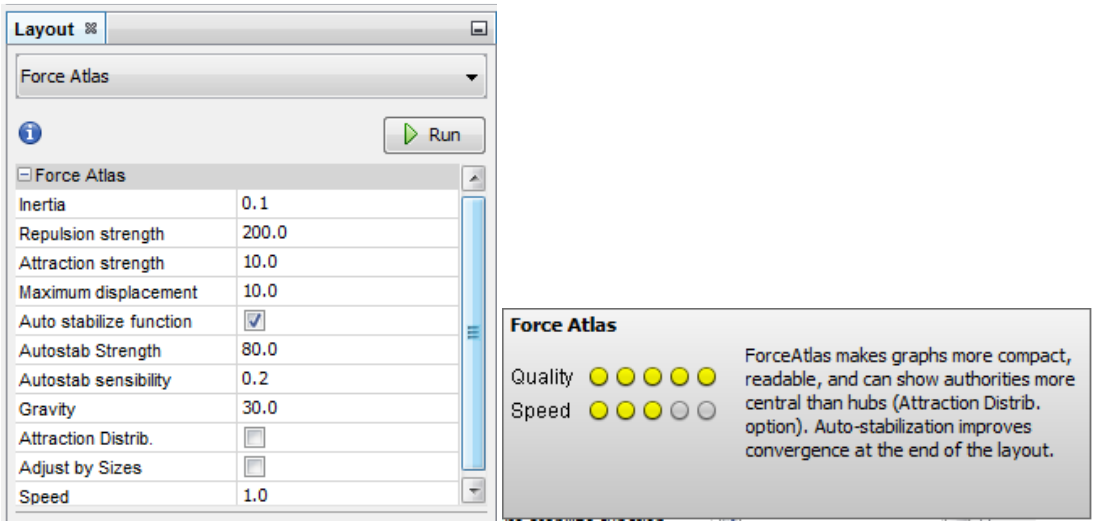

Figura 10 – Parâmetros do algoritmo e a sua explicação.

Como demonstra a Figura 11, a utilização do algoritmo não produziu bons resultados. Como os nós possuem uma relação muito forte entre si, o algoritmo juntou demasiadamente todos os nós. Felizmente é possível alterar os parâmetros do algoritmo, neste caso basta diminuir a força de atração (*Attraction strength)*, aumentar a força de repulsão (*Repulsion strength*) ou alterar ambas. Para ser possível ter uma boa visualização da rede a força de repulsão foi alterada para 9999999999 e a força de atração para 5, criando assim a rede representada na Figura 12.

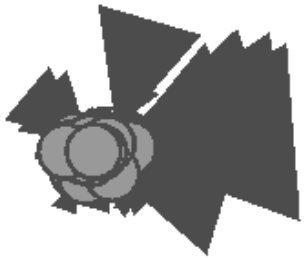

Figura 11 – Rede utilizando os parâmetros de origem do algoritmo *Force Atlas*.

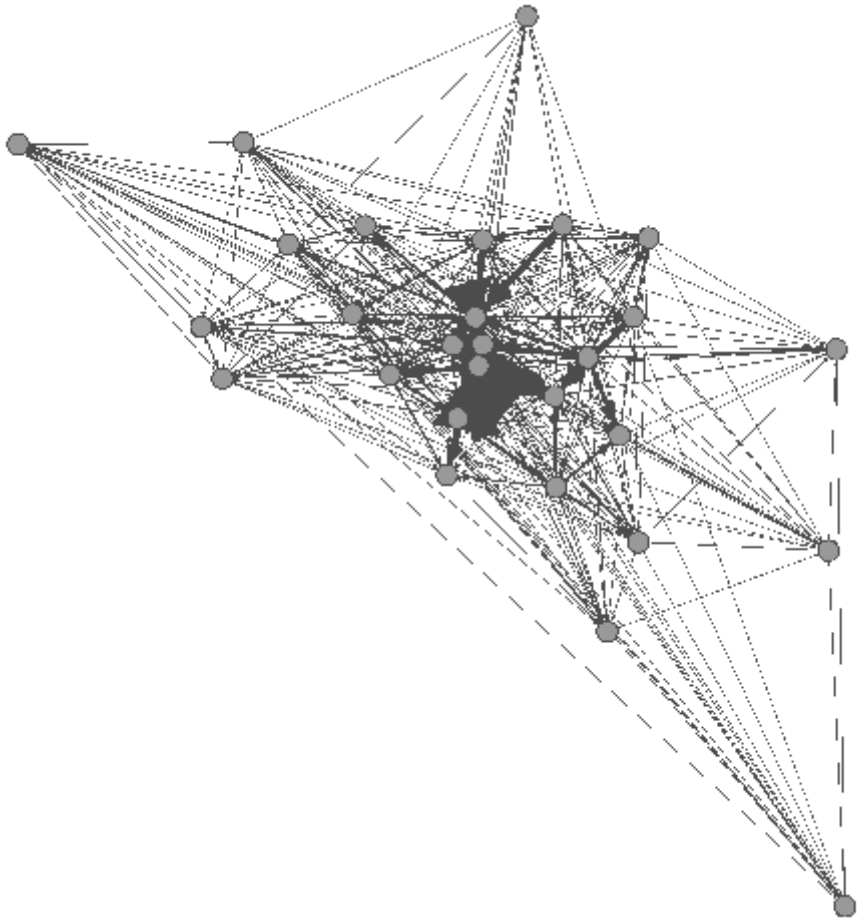

Figura 12 – Rede criada com as alterações da força de repulsão para 9999999999 e a força de atração para 5.[7]

A análise propriamente dita inicia-se agora, começando por fazer alterações na visualização da rede através da introdução de parâmetros de hierarquização quer dos nós quer das relações. Na Figura 12 já é possível verificar que algumas ligações têm linhas mais espessas que outras, isto tem a ver com o valor das exportações, quanto maior o valor das exportações mais espessa a linha da ligação.

O *package Gephi* possui uma opção de identificação dos nós. Uma vez os nós identificados, é possível iniciar as análises e explicar os respetivos resultados. Começando por uma análise dos nós, foi feita na Figura 13 uma coloração, variando de verde, nos nós com grau mais elevado, para vermelho nos nós com grau baixo. Como era expectável pela utilização do algoritmo *Force Atlas*, os nós com mais ligações, ou

[7] As falhas que se observam nas ligações são apenas problemas de resolução e não linhas descontínuas.

seja com um grau maior, encontram-se localizados no centro da rede devido à *Attraction Distribution Option* referida na Figura 10.

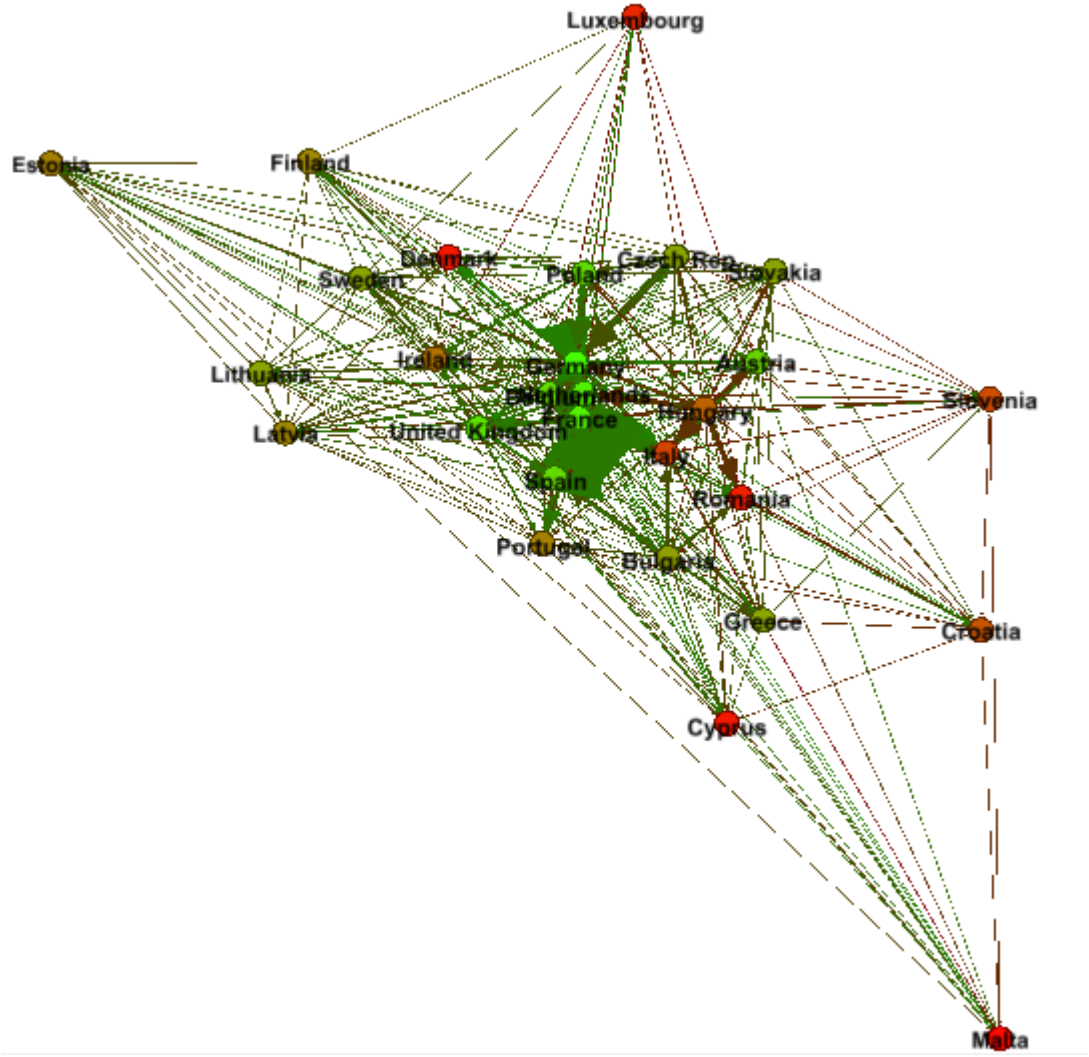

Figura 13 – Rede de países após coloração dos nós mediante o seu grau: verde para os nós com grau mais elevado e graduando para vermelho com a diminuição do grau dos nós.

Na Figura 14 é relativamente percetível que os países com mais ligações são a Alemanha, Holanda e França. No entanto podem surgir algumas dúvidas com o Reino Unido, a Bélgica, e Espanha e talvez com a Áustria também. Para facilitar esta análise é possível aumentar o tamanho dos nós, neste caso em função do grau. Assim, analisando a Figura 15 já é mais fácil validar a teoria inicial, que os países com um grau maior são a Alemanha, Holanda e França. Quanto maior a amplitude entre o tamanho mínimo e o máximo dos nós definida nas opções desta análise, mais fácil é identificação de os nós.

Isto porque o tamanho dos nós com grau baixo será muito inferior ao dos com grau elevado.

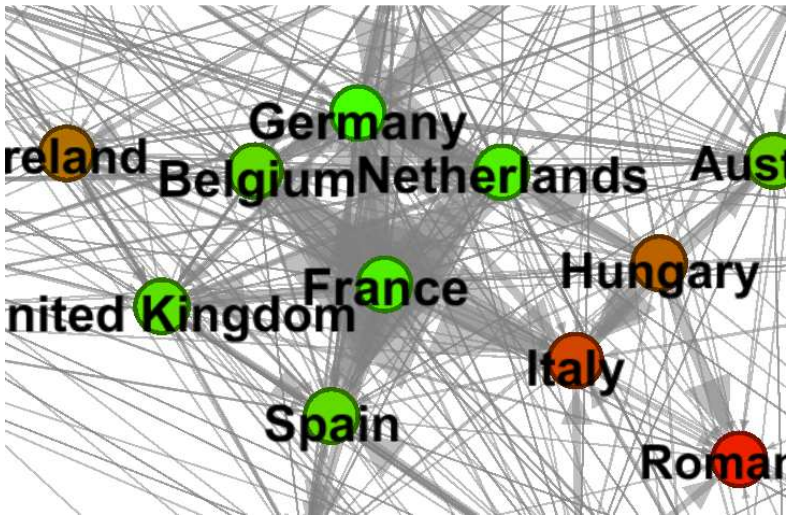

Figura 14 – Imagem da rede com foco nos países com grau mais elevado.

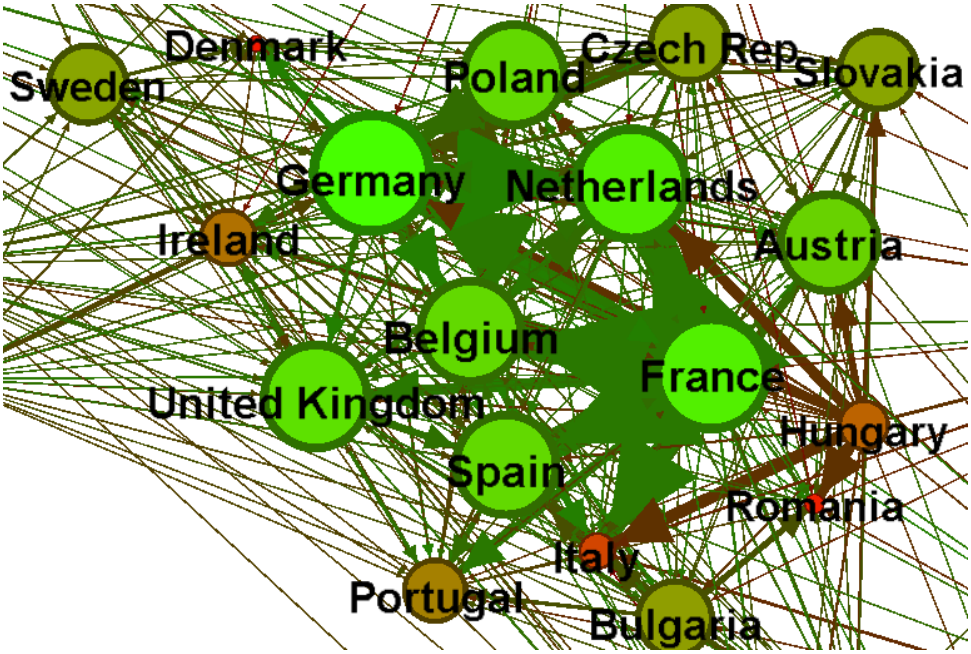

Figura 15 – Imagem da rede com foco nos países com grau mais elevado, após a alteração do tamanho dos nós consoante o grau. Foi definido um tamanho mínimo de 2 e um máximo de 25, de modo a serem mais visíveis as diferenças entre graus.

Como a rede em análise é uma rede dirigida vai ser possível fazer as análises referidas acima para o *In-Degree* e o *Out-degree*, identificando assim quais os países que importam de mais países e os que exportam para mais países, respetivamente. Fazendo a hierarquização dos nós pelo *In-degree* com a coloração dos nós e a alteração

do tamanho, usando os mesmo métodos acima mencionados, é possível verificar (ver Figura 16) que o país que importa de mais países é a Itália seguida da Alemanha. É também visível que a Hungria é de longe o país que recebe cereais de menos países, podendo assim concluir-se que as importações de cereais deste país, se existentes, são maioritariamente feitas de países de fora da União Europeia.

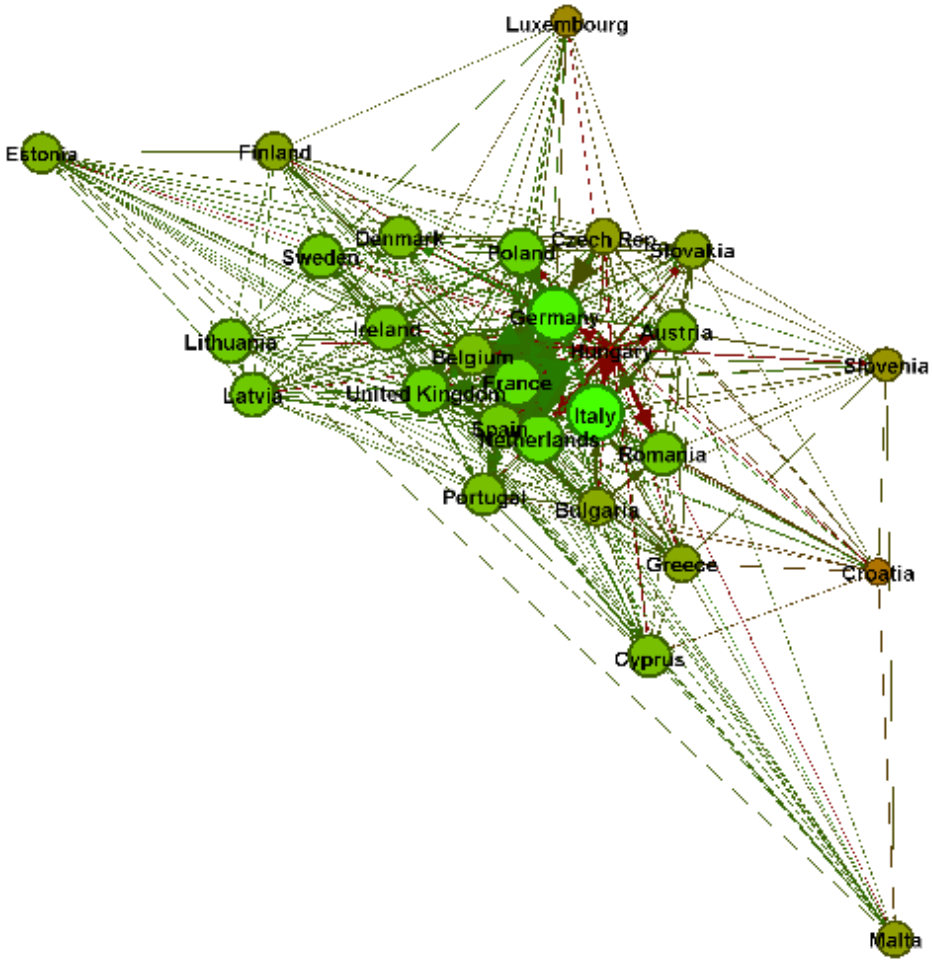

Figura 16 – Rede hierarquizada consoante o *In-Degree*, ou seja, pelo número de ligações de importação que cada país possui.

A próxima análise a ser efetuada será a hierarquização dos nós consoante o número de ligações para o exterior (*Out-Degree)*, ou seja, neste caso o número de países para os quais exporta. Da Figura 17 é de notar um resultado interessante, a Hungria que era o país que tinha menos Estados Membros a exportarem cereais para si, é agora um dos países que exporta para mais países. Através deste resultado é possível deduzir que a Hungria é um grande produtor de cereais, sendo autossuficiente e ainda possuindo uma grande capacidade de exportação. De acordo com a Comissão Europeia (União

Europeia, (1995-2012)) cerca de dois terços do território da Hungria é consagrado à agricultura, e destes, 60% dos solos aráveis são dedicados à produção de cereais (União Europeia, 1995-2012).

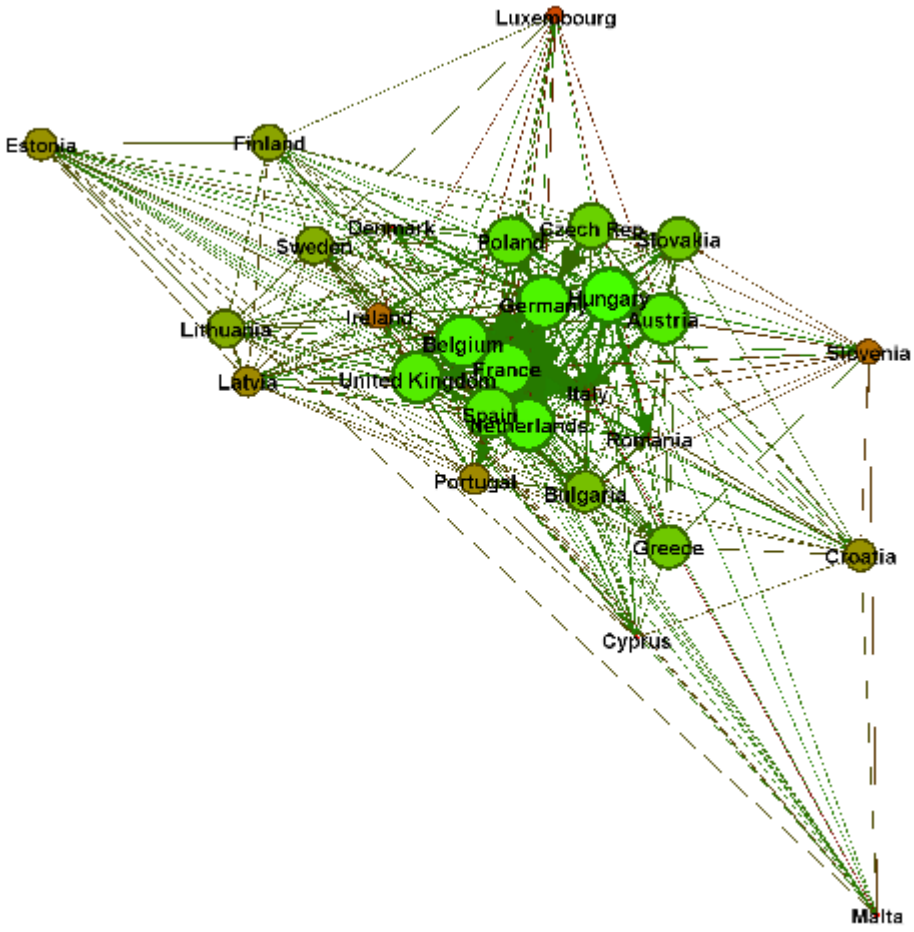

Figura 17 – Rede hierarquizada consoante o *Out-Degree*, ou seja, pelo número de ligações de exportação que cada país possui.

Com a próxima análise vai ser possível verificar se a Hungria é de facto um grande exportador de cereais. O package *Gephi* também permite hierarquizar os nós da rede através da variável incutida aos nós, que neste caso é o valor exportado. Com isto os nós correspondentes a países com maior valor de exportação vão possuir um tamanho maior e cor verde. Como é possível verificar na Figura 18, o país que mais se destaca é a França sendo claramente o país que possui o maior valor de exportações. A Hungria aparentemente não conseguiu um resultado muito bom, no entanto, observando-se com atenção, esta possui um valor de exportações aproximado ao da Alemanha. Como esta hierarquização opera com base na comparação entre os nós, se um dos nós possuir um

valor abruptamente elevado todos os outros irão ter resultados maus. É o que se verifica aqui: por a França ter um valor exportado muito elevado todos os outros têm aparentemente maus resultados. No entanto é possível verificar que apenas a Hungria e a Alemanha se aproximaram mais da França, sendo que para os outros países, esta continua bastante distante.

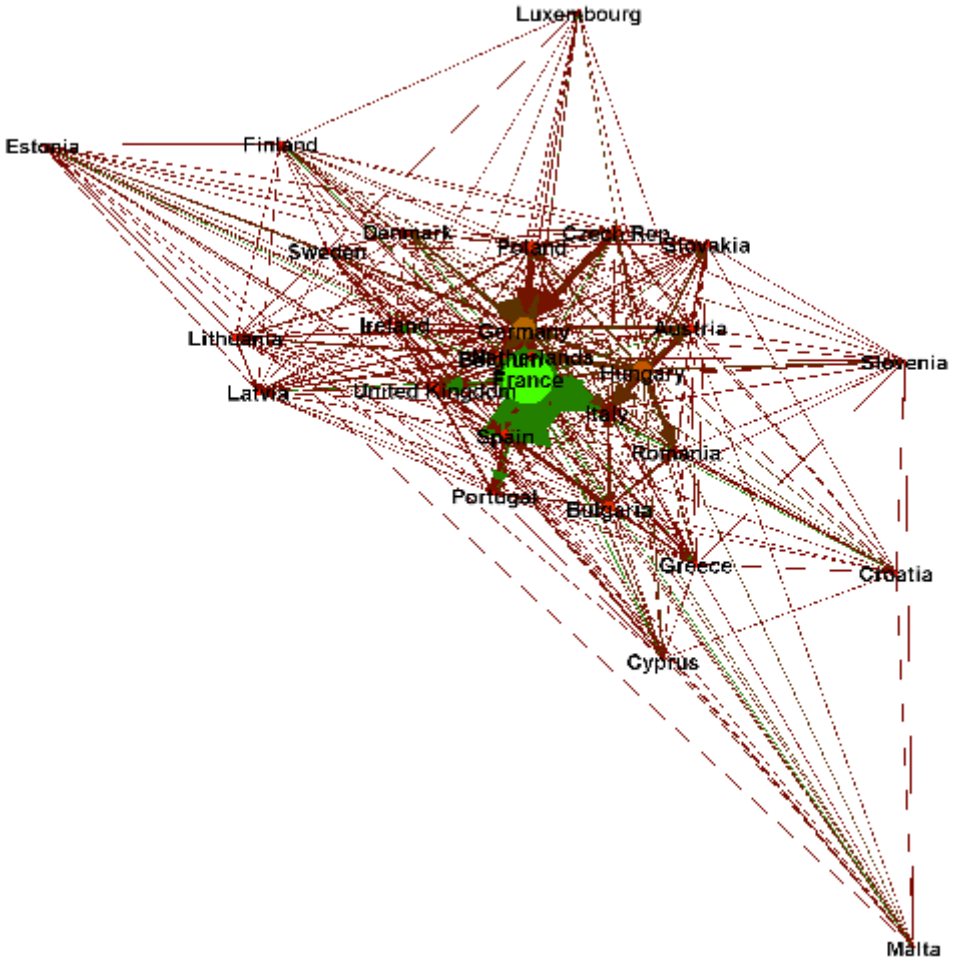

Figura 18 – Rede hierarquizada pelo valor exportado, colorindo a verde os países com maior valor exportado e a vermelho os com menor.

É de notar que a coloração e a alteração do tamanho não necessitam de ser feitos apenas para a mesma variável, sendo possível fazer combinações diferentes para estudos específicos.

Para finalizar a análise gráfica da rede, falta apenas analisar o comportamento das arestas ou ligações. Como já é possível verificar na Figuras acima, as ligações já têm tamanhos diferentes consoante o seu valor. Esta opção é de origem, a única alteração

possível de ser feita é a coloração, que até agora era sempre a mesma. Em suma, esta análise não é muito utilizada pois não acrescenta muita informação à rede. Na Figura 19 é possível verificar que mais uma vez devido aos elevados valores das exportações francesas todos os outros países ficaram com má avaliação.

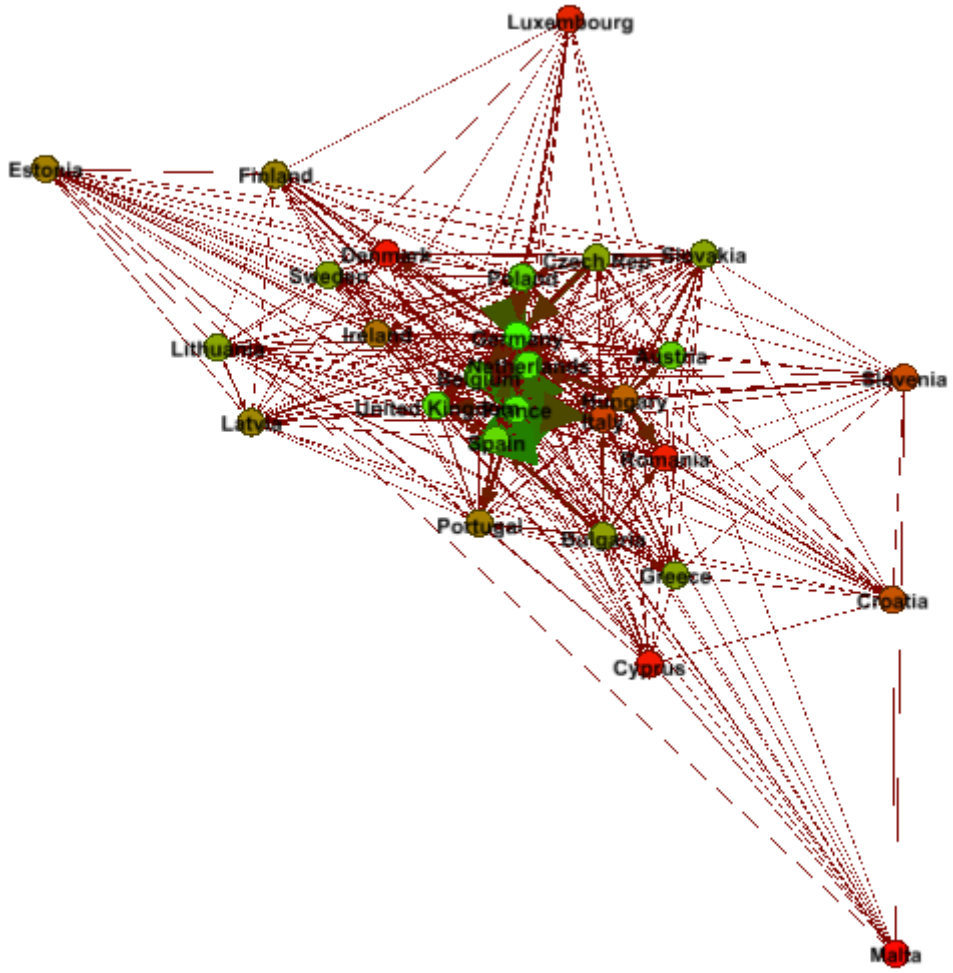

Figura 19 – Rede com os nós coloridos mediante o grau e as ligações coloridas mediante o valor exportado.

5.1.1 Análises estatísticas da rede

O *package Gephi* é também capaz de fazer análises estatísticas de modo a ser possível quantificar as análises das redes descritas. As análises acima podem ser muito úteis, no entanto como estas são alterações visuais, por vezes não é muito simples fazer uma análise cuidada das redes, sendo sempre necessário elaborar algumas análises numéricas.

As primeiras análises a serem efetuadas serão análises generalizadas à rede em si. A primeira análise efetuada calcula o grau médio da rede tendo em conta o Grau, o *In-Degreee* e o *Out-Degree*. Para este cálculo basta somar o número de ligações dividindo pelo número de nós, o resultado fornecido pelo *package Gephi*, apresentado no Anexo 5, indica um grau médio de 16,25. Este resultado é o mesmo para o *In-Degree, Out-Degree* e como o cálculo do grau incorpora os dois a média deste é também 16,25. As contagens são apresentadas como demonstrado na Figura 20, em que o eixo do x representa o grau e o eixo do y reflete a frequência de nós com esse mesmo grau. Neste exemplo é possível referir que existem 2 nós com grau 17, 1 nó com grau 23, 5 com grau 36, etc.

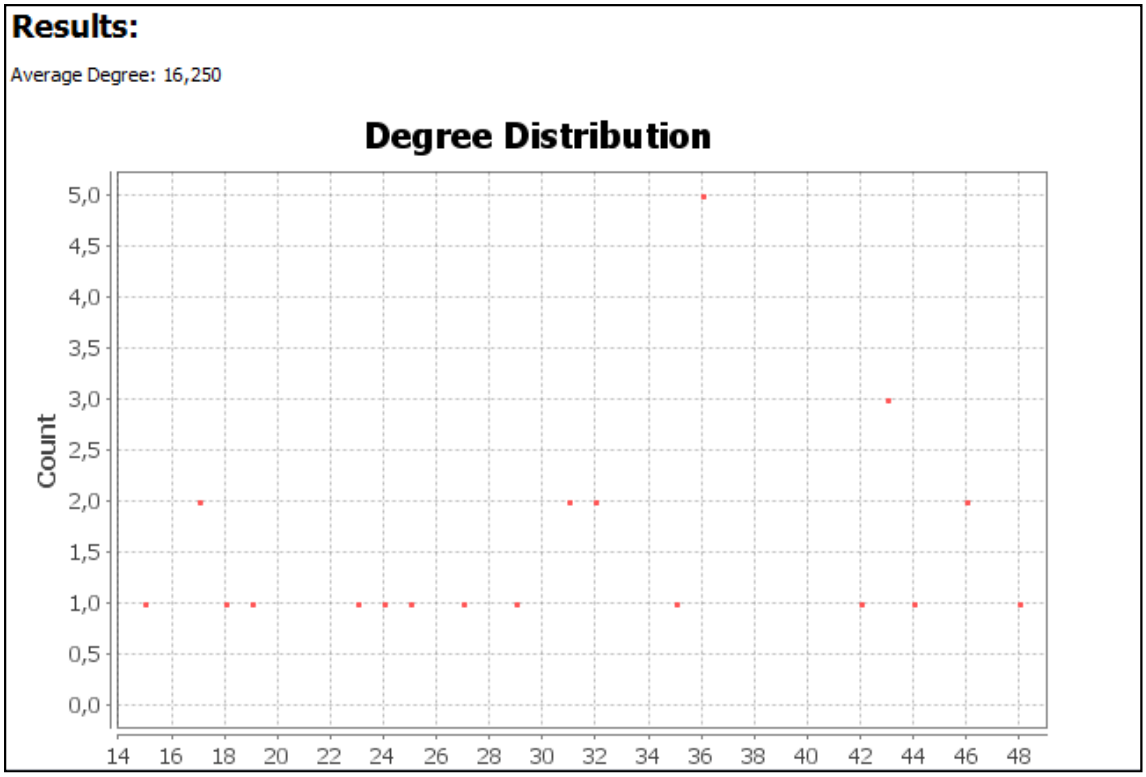

Figura 20 – Gráfico ilustrativo da distribuição do grau. O eixo do x corresponde ao grau e o eixo do y a índices de frequência dos nós com esse grau.

Recorrendo ao Anexo 6 é de maior relevância mencionar que os países com um grau mais elevado são a Alemanha (48), França (46), Holanda (46) e Reino Unido (44). Este facto pode não ter um significado relevante por a rede ser dirigida, tendo assim de ser complementado com os valores do *In-Degree* e do *Out-Degree*; os países com um *In-Degree* maior, ou seja com mais ligações importadoras, são a Itália (23), a Alemanha (22), a França (20) e a Holanda (20); os países com o *Out-Degree* maior, ou seja os que

exportam para mais países, são a Hungria (27), Alemanha (26), França (26), Holanda (26) e Bélgica (26).

Através do *Data Laboratory* no *Gephi*, onde estão armazenados os resultados das estatísticas, é possível estudar os resultados e saber quais os países que mais contribuem para as estatísticas. Há 3 países que se destacam nos resultados do *Weighted Out-Degree* (ver Figura 21), estes são a França, a Alemanha e a Hungria. Para o *Weighted In-Degree* os países com valores mais elevados, e portanto os que importam mais e de maior número de países, são a Holanda, Alemanha e Itália.

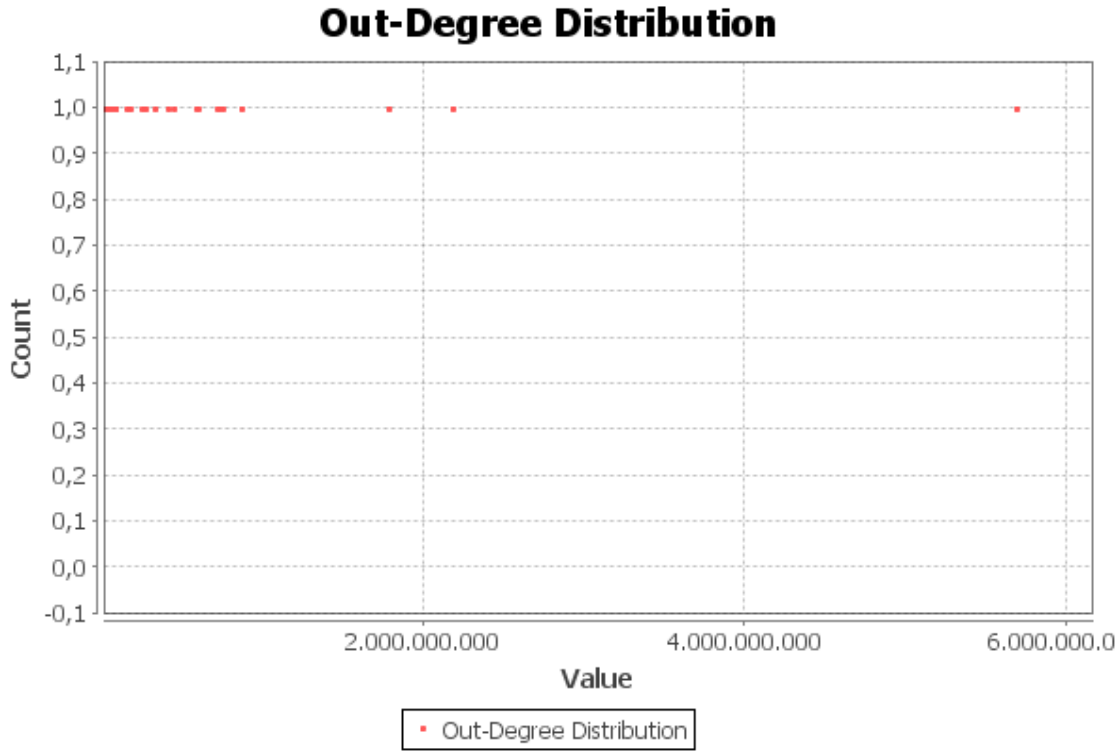

Figura 21 – Gráfico representativo dos resultados do *Weighted Out-Degree*.

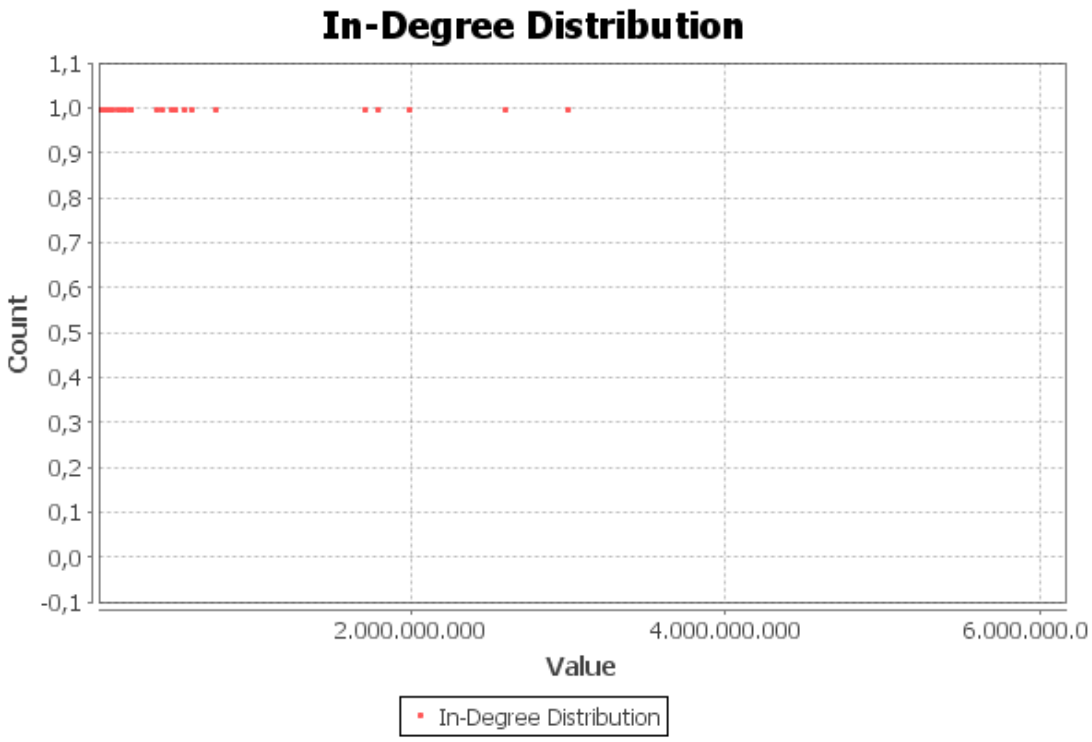

Figura 22 - Gráfico representativo dos resultados do *Weighted In-Degree*.

No cálculo da distância da rede o *package Gephi* efetua também análises para a *Betweenness Centrality, Closeness Centrality e a Eccentricy.* Com estas análises são-nos fornecidos a distância média de um caminho, igual a 1,24, e o número de caminhos mais curtos existentes na rede, 600. Estes resultados estão presentes no Anexo 7, assim é possível concluir que em média é necessário passar por 1,24 nós para chegar de um determinado nó até outro.

Analisando a Densidade de uma rede é possível identificar o quão perto a rede está de ser uma rede completa, ou seja com a totalidade dos nós ligados entre si. A densidade da rede é de 0,602 o que indica que mais de metade das ligações já estão efetuadas, no entanto ainda faltam muitas para que a rede esteja completa perfazendo assim uma densidade de 1. Neste caso específico todos os países exportariam para todos os países, fazendo com que estivessem todos ligados.

A próxima medida denominada de HITS vai analisar dois valores distintos para cada nó, o primeiro valor (denominado *Authority*) mede quão valiosa é a informação armazenada no nó. O segundo valor (denominado de *Hub*) mede a qualidade das ligações de cada nó. Observando as Figuras 23 e 24 conclui-se que a informação

armazenada nos nós não é muito valiosa e que a qualidade das ligações é fraca, uma vez que os valores estão todos muito próximos de zero.

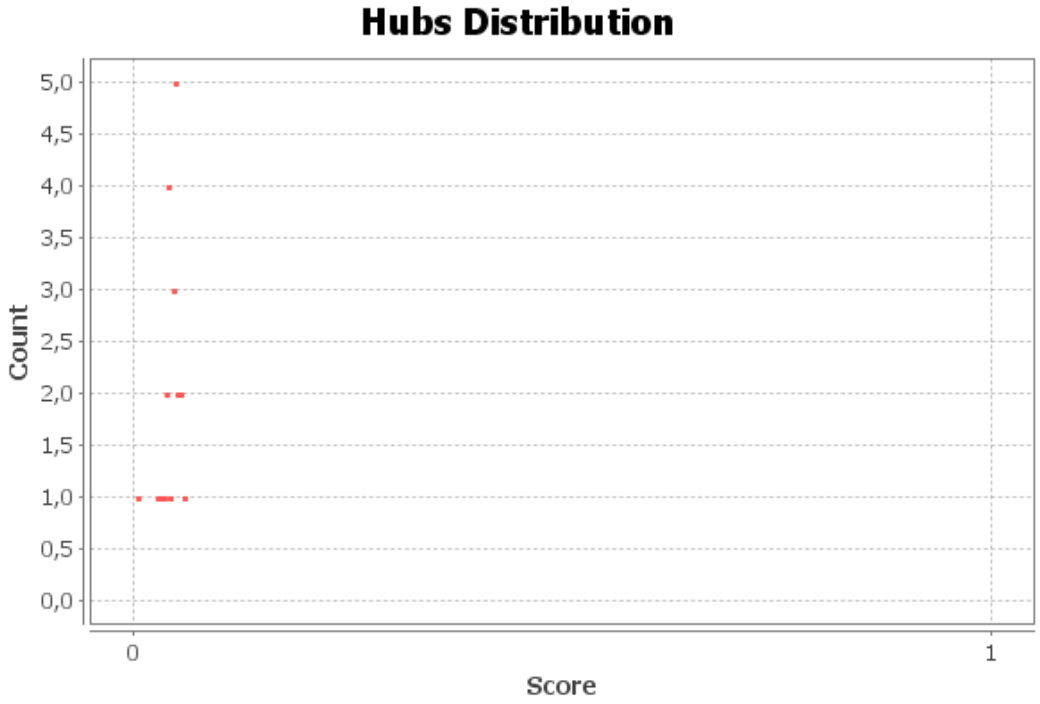

Figura 23 – Distribuição de *Hubs*, permite medir a qualidade das ligações de cada nó. Quanto mais próximo de 1 melhor é a qualidade.

Figura 24 – Distribuição de *Authority*, permite medir a qualidade da informação armazenada em cada nó. Quanto mais próximo de 1 melhor é a qualidade e mais valiosa é a informação.

A deteção de comunidades é bastante importante na análise de redes sociais, o *package Gephi* possui ferramentas de deteção de comunidades. Como indicado na Figura 25, existe a possibilidade de procurar comunidades pequenas ou comunidades grandes. Como a rede é relativamente pequena não tem lógica procurar comunidades grandes, logo a Resolução escolhida foi de 0,7. Não se considerando pesos nas arestas, é necessário desseleccionar a opção *Use weights*.

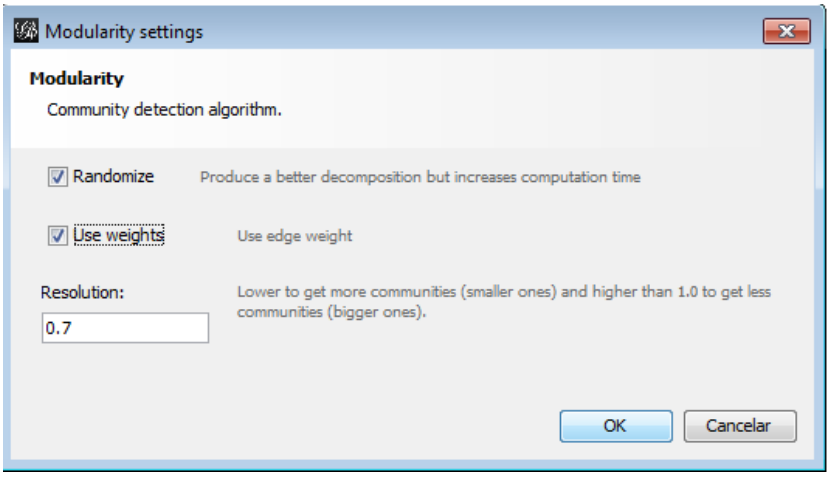

Figura 25 – Opções assumidas para a deteção de comunidades, optando por uma resolução de 0.7 para a deteção de comunidades relativamente pequenas.

Como apresentado na Tabela 4 é possível identificar oito comunidades,

Label	Mod. Class	Label	Mod. Class	Label	Mod. Class	Label	Mod. Class
Estonia	0	Croatia	1	U. Kingdom	2	Austria	5
Finland	0	Czech Rep	1	Belgium	3	Germany	5
Ireland	0	Greece	1	Netherlands	3	Slovakia	6
Latvia	0	Slovenia	1	Luxembourg	4	Italy	6
Lithuania	0	Romania	1	Poland	4	France	7
Denmark	0	Hungary	2	Portugal	4	Spain	7
Bulgaria	1	Malta	2	Sweden	4	Cyprus	7

Tabela 3 – Tabela com a composição das 8 comunidades em que *Label* corresponde à identificação do país e *Mod. Class* é a classe de modularidade, cada número corresponde a uma comunidade.

5.1.2 Análise geral dos nós

Para a análise dos nós existem apenas duas estatísticas definidas no *package Gephi*, o coeficiente de agrupamento médio (*Avg. Clustering Coefficient*) e o *Eigenvector centrality*.

O **coeficiente de agrupamento**, indica os efeitos das redes *Small-World*. Assim vai ser possível verificar quão integrados estão os nós nas suas vizinhanças. A média deste coeficiente fornece uma indicação generalizada dos agrupamentos da rede. A média fornecida após correr o algoritmo é de 0,67 indicando assim uma boa integração dos nós nas suas vizinhanças.

O ***Eigenvector centrality*** fornece informações sobre a importância dos nós na rede baseado nas ligações de cada nó. De acordo com a Figura 26, a maioria dos nós tem um *Eigenvector centrality* superior a 0,5 demonstrando que estes possuem todos importância na rede. São de destacar pela positiva a Alemanha, Holanda e Itália com coeficientes próximos de 1 e pela negativa a Hungria com um coeficiente próximo de 0.

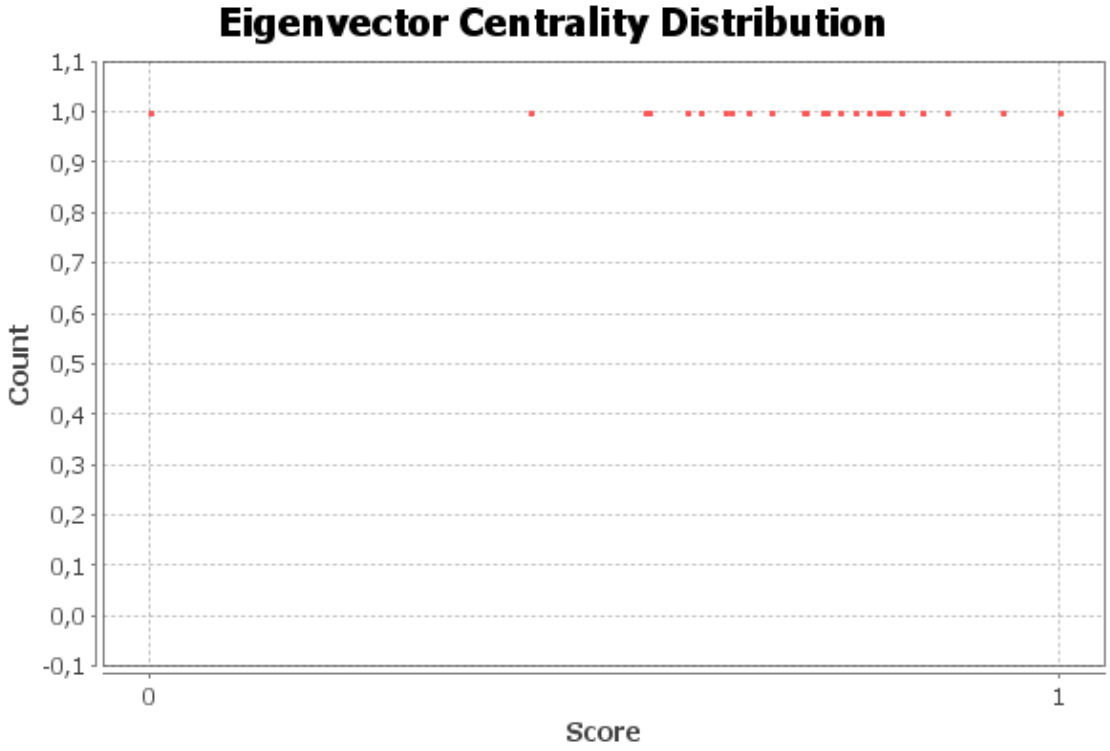

Figura 26 – Gráfico com os valores da Distribuição do *Eigenvector Centrality* de cada nó. Cada ponto corresponde a um nó marcando assim o valor do *Eigenvector*.

À medida que se efetuam as análises acima mencionadas, os resultados generalizados para a rede são apresentados com as estatísticas (Figura 27), no entanto os resultados relativos a cada nó são armazenados na tabela de dados com a informação dos nós. Estes resultados estão disponíveis no Anexo 8.

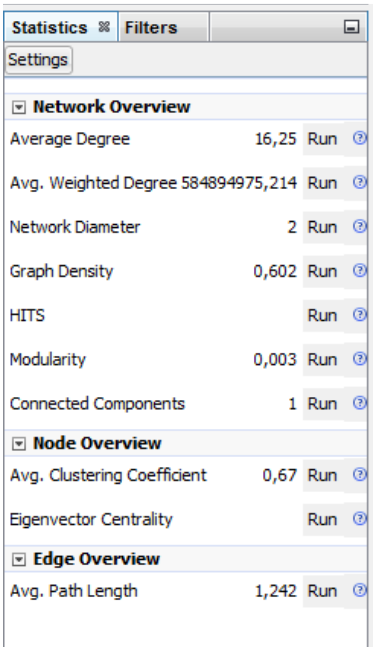

Figura 27 – Conjunto de estatísticas disponíveis no *Gephi* e seus resultados para a rede em análise.

5.2 Metodologia da Análise Conceptual

Neste capítulo irá ser abordada uma análise diferente com o intuito de complementar a análise de redes. A análise conceptual vai permitir agrupar os países mediante as suas intensões, ou seja atributos em comum. Para esta análise é necessária a utilização de dois programas distintos, porque não foi encontrado um *software* gratuito que conseguisse determinar os conceitos e calcular as intensões e respetivas extensões. Deste modo foi usado para a determinação dos conceitos e o cálculo das intensões o FCbO, um programa que faz o cálculo formal dos conceitos através do algoritmo FCbO, e foi criado um código em linguagem R capaz de determinar as extensões mediante os resultados obtidos do FCbO.

Para que os dados pudessem ser utilizados no software FCbO estes tiveram de ser convertidos para o tipo de ficheiro .dat, onde é necessário converter os “1” para o número da coluna que este representa. Analisando a Tabela 5, temos dois países com “1” nas características que cada um possuí e 0 nas que não possuem. Para transformar em formato .dat é substituído o 1 pelo número da coluna ficando apenas os números das colunas com as características que o país possui (Tabela 6).

Países	Pop mid 2012 (M) < 20	Pop mid 2012 (M) 20 < 50	Pop mid 2012 (M) > 50	Births per 1,000 Population < 10	Births per 1,000 Population < 10	Deaths per 1,000 Population < 10	Deaths per 1,000 Population < 10	Rate of Natural Increase % Positive	Rate of Natural Increase % Negative	Net Migration Rate per 1,000 positive	Net Migration Rate per 1,000 negative	2050 Population as a Multiple of 2012 <1	2050 Population as a Multiple of 2012 > 1	Infant Mortality Rate < 5	Infant Mortality Rate >5
Denmark	1	0	0	0	1	1	0	0	1	0	1	0	1	1	0
Estonia	1	0	0	0	1	0	1	1	0	1	0	1	0	1	0

Tabela 4 – Tabela com código binário representando as características de dois dos Países em análise.

Países	Pop mid 2012 (M) < 20	Pop mid 2012 (M) 20 < 50	Pop mid 2012 (M) > 50	Births per 1,000 Population < 10	Births per 1,000 Population < 10	Deaths per 1,000 Population < 10	Deaths per 1,000 Population < 10	Rate of Natural Increase % Positive	Rate of Natural Increase % Negative	Net Migration Rate per 1,000 positive	Net Migration Rate per 1,000 negative	2050 Population as a Multiple of 2012 <1	2050 Population as a Multiple of 2012 > 1	Infant Mortality Rate < 5	Infant Mortality Rate >5
Denmark	0				4	5			8		10		12	13	
Estonia	0				4		6	7		9		11		13	

Tabela 5 – Tabela com a alteração para o formato .dat

Apesar da apresentação na Tabela 6 o programa apenas necessita dos valores, ignorando tanto os nomes dos países como os cabeçalhos, ficando com a apresentação representada no Anexo 9. Pode ser necessário para a utilização do FCbO a criação de um executável do *MsDos* com o diretório apenas para abrir o programa. Este programa tem uma utilização bastante simples, sendo neste caso apenas necessário usar o código representado na Figura 28. Aqui é possível verificar que a primeira parte corresponde ao diretório da localização do programa, "fcbo" é o nome do ficheiro do programa e de seguida tem o respetivo código, em que "países.dat" é o nome do ficheiro onde estão os dados e "resultadospaises.dat" o nome para o ficheiro com os resultados que irá ser criado. Neste caso, a primeira coluna da tabela de dados (primeiro atributo da Tabela 6) é representada por 0 (zero).

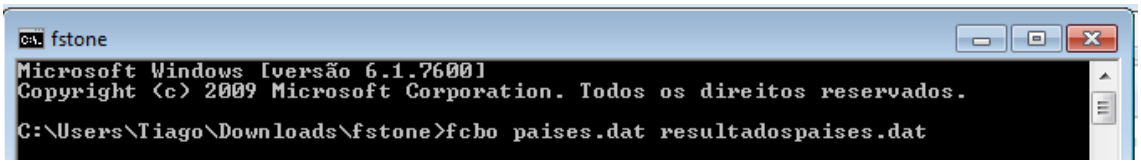

Figura 28 – Representação do código para utilização do *software* FCbO no ficheiro de dados países.dat.

Representado por uma amostra na Figura 29 e na totalidade no Anexo 10, o ficheiro de dados resultantes do FCbO contém as intensões dos conceitos do conjunto dos 28 Estados Membros da EU. Cada linha representa uma intensão e cada número representa um atributo, sendo que as intensões são, neste caso, um conjunto de atributos presentes num determinado conjunto de países. Vamos então ter, por exemplo, um conjunto de

países que possuem uma taxa de mortalidade infantil acima de 5 e uma população em 2050 inferior à de 2012, intensão = {14, 11}.

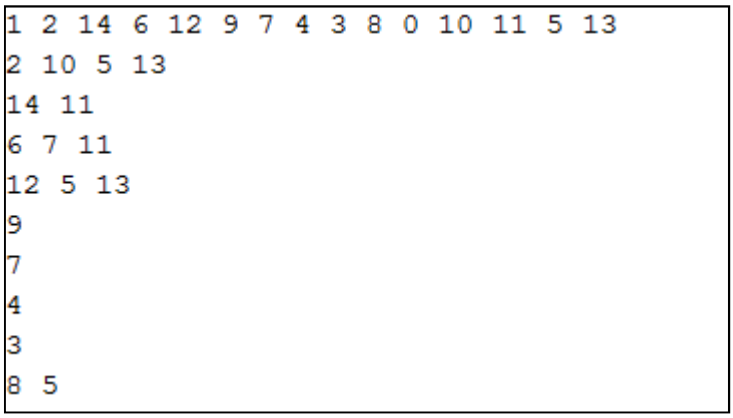

Figura 29 – Amostra dos resultados obtidos através da utilização do FCbO, constituídos pelas intensões dos conceitos formados a partir dos dados demográficos dos países.

Como os resultados do FCbO não incluem as extensões foi necessário criar um outro método para arranjar as extensões. Para tal foi usado o *software* R criando um código que vai usar dois tipos de dados: A tabela do Anexo 9 contendo os atributos dos países, esta necessitou de uma pequena alteração, como nem todos os países possuíam o mesmo número de características havia linhas com diferente número de colunas e portanto, para igualar o número de colunas foi usado o número “999”; A segunda entrada de dados é classificada como lista, exatamente por não ter o mesmo número de colunas, e vai ser constituída pelas intensões (tabela do Anexo 10).

Páises	C1	C2	C3	C4	C5	C6	C7
Denmark	0	4	5	8	10	12	13
Estonia	0	4	6	7	9	11	13
Finland	0	4	5	8	10	12	13
Ireland	0	4	5	8	9	12	13
Latvia	0	3	6	7	9	11	14
Lithuania	0	4	6	7	9	11	13
Sweden	0	4	5	8	10	12	13
United Kingdom	2	4	5	8	10	12	13
Austria	0	3	5	7	10	12	13
Belgium	4	5	8	10	12	13	999
France	2	4	5	8	10	12	13
Germany	2	3	5	7	10	11	13
Luxembourg	0	4	5	8	10	12	13
Netherlands	4	5	8	10	11	13	999
Bulgaria	0	3	6	7	9	11	14
Hungary	0	3	6	7	10	11	13
Poland	3	5	8	9	11	13	999
Romania	3	6	7	9	11	14	999
Slovakia	0	4	5	8	9	11	14
Croatia	0	3	6	7	9	11	13
Greece	3	5	8	10	11	13	999
Italy	2	3	5	7	10	11	13
Malta	0	3	5	8	10	11	14
Portugal	3	5	7	10	11	13	999
Slovenia	0	4	5	8	10	11	13
Spain	3	5	8	9	11	13	999

Tabela 6 – Tabela com os atributos e usando o número 999 para que todos os países tenham o mesmo número de colunas

Presente no Anexo 11, a função criada para este caso funciona de uma maneira simples: usando como variável a lista de intensões o algoritmo vai percorrer todas as entradas da lista uma a uma gravando no final de cada análise o resultado na lista denominada *res.final*. O algoritmo ao abordar uma entrada (intensão) vai primeiro verificar a o seu cardinal através de uma função *if* (if (length(dados[[i]])==1), exemplo para testar se tem um valor apenas), caso o cardinal da intensão do conceito coincida com o valor representado na função é feito o cálculo da extensão, caso contrário passa para outra função *if* igualando o cardinal a um outro valor. O Código 1 representa um exemplo para a função de cálculo das extensões. Este exemplo é usado para intensões com apenas dois valores; para calcular as extensões o algoritmo vai verificar coluna a

coluna nos dados da Tabela 7 se algum país possui as características presentes na intensão. Isto é feito individualmente, procurando um país que possua o primeiro valor da intensão (dados[[i]][1]) numa das 7 colunas da Tabela 7 (Código 2), utilizando o "|" representando "ou". Isto é, verifica se o valor da intensão está na primeira coluna (C1) ou na segunda coluna (C2), ou na terceira e assim sucessivamente. Após verificar isto para o primeiro valor o algoritmo faz o mesmo para o segundo valor da intensão; se a intensão possuir mais valores faz o mesmo para todos. Como é pretendido que o país possua todos os valores das intensões, as pesquisas dos vários valores da intensão são separados por um "&". Como este código está a fazer uma pesquisa dentro do *data frame* criado com os valores da Tabela 7 (teste[…]) é preciso identificar no final qual a coluna contendo a informação que é pretendido que o R devolva, sendo esta o "id", onde estão colocados os nomes dos países.

```
resultado<-
teste[(teste$C1==dados[[i]][1]|teste$C2==dados[[i]][1]|teste$C3==dados[[i]][1]|teste$C4==dados[[i]][1]|teste$C5==dados[[i]][1]|teste$C6==dados[[i]][1]|teste$C7==dados[[i]][1])&(teste$C1==dados[[i]][2]|teste$C2==dados[[i]][2]|teste$C3==dados[[i]][2]|teste$C4==dados[[i]][2]|teste$C5==dados[[i]][2]|teste$C6==dados[[i]][2]|teste$C7==dados[[i]][2]),"id"]
```

Código 1 – Código R para calcular as extensões de intensões com dois valores.

```
(teste$C1==dados[[i]][1]|teste$C2==dados[[i]][1]|teste$C3==dados[[i]][1]|teste$C4==dados[[i]][1]|teste$C5==dados[[i]][1]|teste$C6==dados[[i]][1]|teste$C7==dados[[i]][1])
```

Código 2 – Código R para verificar se algum país possui o primeiro atributo presente na intensão.

Na Figura 30 está representada uma amostra dos resultados obtidos a partir do programa em R, os resultados completos encontram-se no Anexo 12. Na Figura 30 representam-se as extensões das 3 primeiras intensões, é no entanto de notar que não foi tida em conta a primeira intensão (verificar na Figura 29) pois esta representava todos os atributos, e nenhum país possuía todos os atributos. Relembrando a Figura 29, a segunda intensão possuía os atributos 2, 10, 5 e 13 o que significa que os países Reino Unido, França, Alemanha e Itália possuem estes 4 atributos. Já a Letónia, Bulgária, Roménia, Eslováquia e Malta têm em comum os atributos 14 e 11, e por fim a Estónia, Letónia, Lituânia, Bulgária, Hungria, Roménia e Croácia têm em comum os atributos 6, 7 e 11.

```
#resultados do R finais

[1] United Kingdom France         Germany        Italy
[1] Latvia   Bulgaria Romania  Slovakia Malta
[1] Estonia   Latvia    Lithuania Bulgaria  Hungary   Romania   Croatia
```

```
2 10 5 13
14 11
6 7 11
```

Figura 30 – Imagem superior com uma amostra dos resultados retirados do programa em R, estes resultados são as extensões calculadas a partir das intensões e da tabela de dados inicial. Estas extensões representam os diversos países que possuem os vários atributos mencionados nas intensões. Imagem inferior com as intensões referentes às extensões apresentadas (ver Tabela 5 ou 6 para identificação das intensões).

Com estes resultados é possível a criação de grupos, pois as extensões integram vários países com as mesmas características. Assim, esta técnica também é considerada uma técnica de agrupamento, pois vai sempre agrupando os países em pequenos grupos (as extensões dos conceitos).

6 Análise dos dados de coautoria

A análise que se segue irá ser efetuada aos dados de coautoria, sendo esta a principal análise desta dissertação. A explicação dos métodos não será tão detalhada, mas a explicação dos resultados será mais pormenorizada.

Este conjunto de dados foi gentilmente fornecido pelo Professor Dr. Fernando Silva e pela Drª Sylwia Bugla parceiros no LIAAD (*Laboratory of Artificial Intelligence and Decision Support*). O conjunto de dados é constituído por duas tabelas, uma representando as relações de coautoria (Anexo 13) e a outra representando as áreas das publicações de cada autor (Anexo 14). Ambas as tabelas possuem dados relativos apenas à área da Economia, podendo no entanto pertencer a várias ISI[8] ou áreas de publicação. No total, são analisados 233 autores e são tidas em conta 263 ISI sendo que a ISI 1 contempla as publicações que não possuem categoria e as restantes 262 representam uma categoria específica. De forma a evitar a duplicação de autores, estes estão representados por um código e não pelo respetivo nome.

Na tabela das relações de coautoria, as duas primeiras colunas representam dois autores que trabalharam em conjunto. A terceira coluna contém o número de publicações efetuadas pelos dois autores mencionados nas duas primeiras colunas. Cada linha desta tabela contém uma relação entre dois autores, sendo que estes, individualmente, podem estar em mais que uma linha, dado poderem trabalhar com outros autores. No entanto, a mesma ligação não se pode encontrar em mais do que uma linha. Esta tabela vai ser a base para a criação da rede de coautorias no *package Gephi*.

A segunda tabela vai servir para caracterizar e agrupar os autores. Para tal, esta é uma tabela de dupla entrada com código binário. Esta tabela possui em cada linha um autor e em cada coluna uma ISI. A publicação de um determinado autor numa ISI será codificada com 1, em suma, os valores 1 irão indicar as áreas das publicações efetuadas por cada autor. Um autor pode conter várias ISI, e logicamente uma ISI pode corresponder a mais do que um autor. As ISI estão representadas em código sendo que para saber qual a área de cada ISI pode consultar-se o Anexo 15.

[8] ISI é um código que corresponde à categoria do tema do jornal onde a publicação foi publicada. É possível verificar a lista em http://ip-science.thomsonreuters.com/mjl/scope/scope_scie/#AA

6.1 Análise da Rede

É aqui encontrada a única diferença entre os dados relativos aos países e os presentes dados, dado que na análise anterior as ligação eram direcionadas e agora não são, pois nas coautorias os autores trabalham em conjunto. Esta rede de coautoria é constituída por 212 nós e 213 ligações, isto é, existem 212 autores e 213 parcerias. Aqui tanto os nós como as ligações possuem valores associados, sendo que para os nós foi contabilizado o número de publicações de cada autor (ver Anexo 16) e nas ligações é contabilizado o número de publicações que os dois autores fizeram em conjunto (ver Anexo 17). É no entanto de notar que para as ligações não foi usado o número de publicações mas antes o peso relativo, dividindo o número de publicações de cada ligação pelo número total de publicações da rede. Isto foi efetuado apenas para que não existam variações muito elevadas que aumentem excessivamente o tamanho das arestas dificultando a sua visualização e análise da rede.

Para esta análise foi utilizado novamente o algoritmo *Force Atlas*, no entanto apenas foi alterada a força de atração de 10.0 para 1.0 aumentando assim a distância entre os nós.

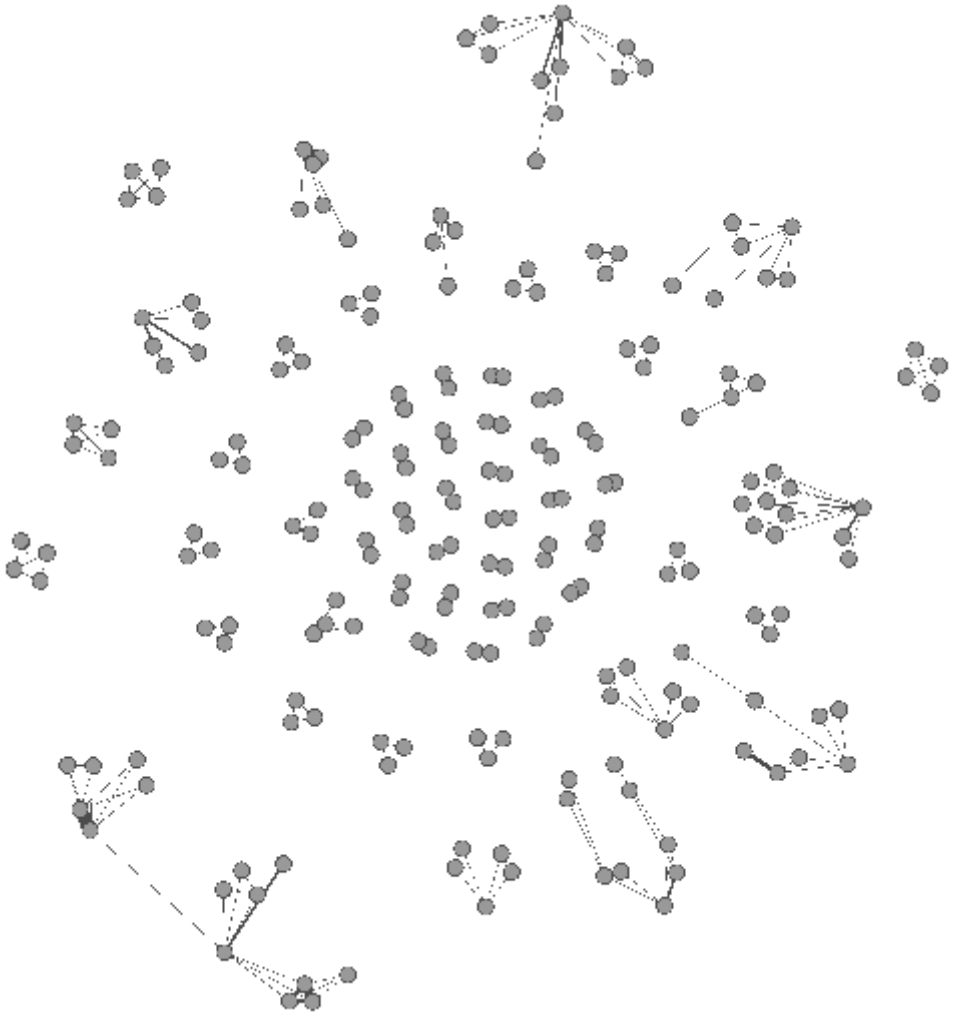

Figura 31 – Rede de coautorias configurada com o algoritmo *Force Atlas*, com a alteração da força de atração de 10 para 1.

Esta é uma rede constituída por muitos pequenos grupos, ou seja possui um agrupamento local, grupos pequenos e unidos mas espaçados entre si. Um comprovativo deste facto é a existência de várias cliques na rede, como as da Figura 32. É já também visível a existência de pontes locais, mostrando que os autores estão a querer diversificar a sua rede e os seus conhecimentos para outros grupos, talvez com publicações em áreas diferentes (ver Figura 33).

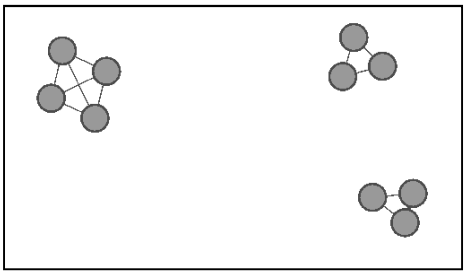

Figura 32 – Exemplos de *cliques* existentes na rede. *Clique*s são grupos da rede em que todos os nós estão ligados entre si.

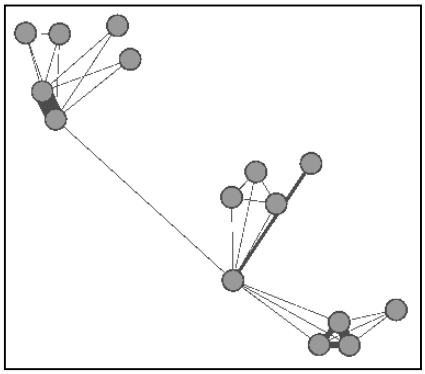

Figura 33 – Exemplo de ponte unindo dois pequenos grupos da rede, criando assim uma ligação e consequente aproximação entre os dois grupos.

A hierarquização dos nós é inicialmente efetuada através do grau e futuramente através do número de publicações dos autores. A hierarquização foi novamente representada através da coloração e alteração do tamanho dos nós. Na Figura 34 está representada essa hierarquização, como era de esperar pelo formato da rede o grau dos nós é muito baixo, à exceção de alguns grupos que estão melhor formados.

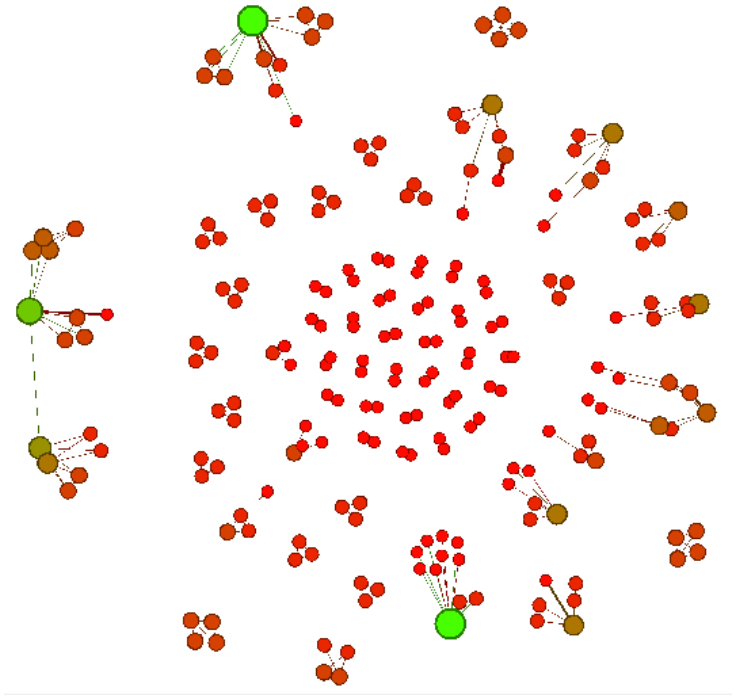

Figura 34 – Rede hierarquizada pelo grau através da coloração dos nós, sendo verdes os nós com grau mais elevado e vermelhos com grau menor. A hierarquização é feita também através do tamanho dos nós num intervalo de tamanho [10,25] em que o tamanho aumenta em proporção do aumento do grau.

É de facto intrigante verificar que nos grupos representados na Figura 35 existe apenas um nó com um grau elevado e os restantes possuem um grau muito inferior. Isto significa que os *Hubs* destes grupos publicam com muitos autores no entanto esses autores não publicam uns com os outros, ou publicam com poucos. Este caso está bem presente no segundo grupo da Figura 35, em que o autor 2744 publica com todos os outros autores do grupo, no entanto esses não publicam uns com os outros, à exceção dos autores 16664 e 8086. Este facto já por si indica que é necessário haver mais cooperação entre os autores; pode no entanto significar que os autores 9111, 2744 e 1154 são autores mais conceituados e têm muitos outros a querer trabalhar com eles.

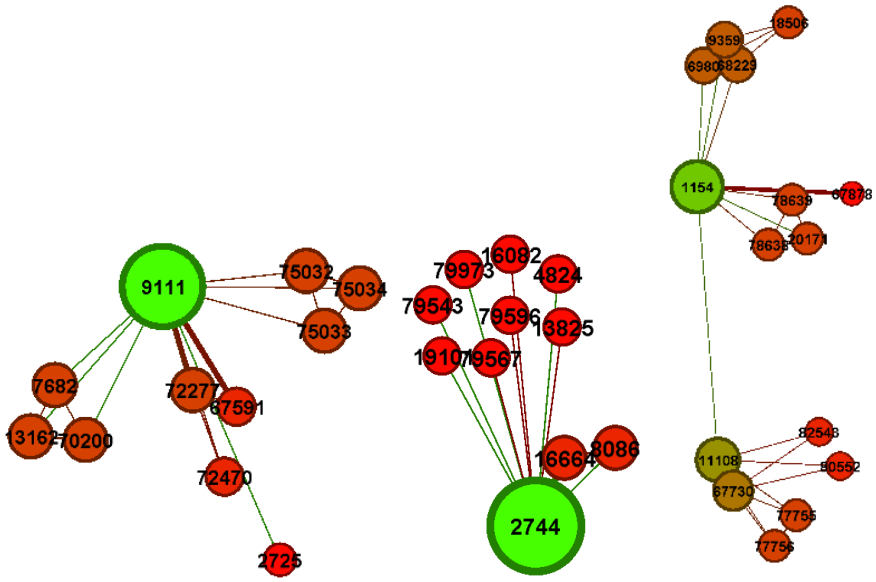

Figura 35 – Os três grupos com os nós com maior grau, sendo que o nó 9111 e o 2744 têm um grau de 10 e o 1154 de 9.

Na próxima análise é aplicado o mesmo método de hierarquização mas para o número de publicações. É no entanto de notar que o número de publicações não é propriamente dependente do grau dos nós, ou seja, o número de publicações não parece depender da quantidade de parcerias. Como representado na Figura 36, apenas o autor 2744 apresenta um grau e um número de publicações elevado, os restantes não têm uma prestação muito boa nesta análise. Nesta análise os autores com melhores resultados são o já referido 2744, o 18645 e o 11108.

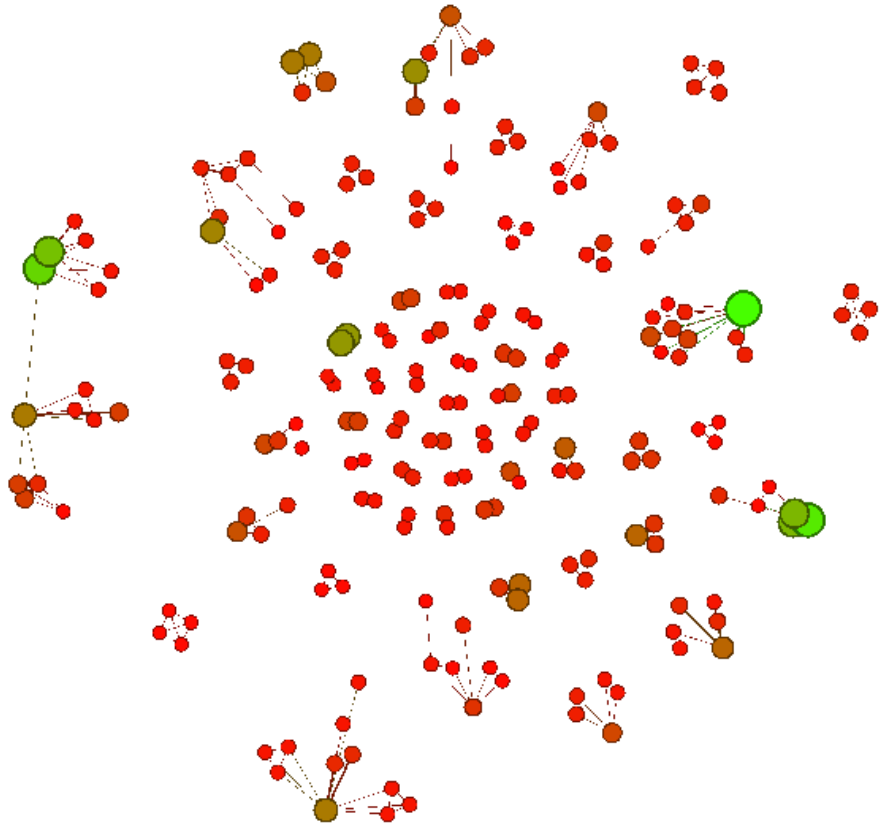

Figura 36 - Rede hierarquizada pelo número de publicações através da coloração dos nós, sendo verdes os nós com mais publicações e vermelhos os com menos publicações. A hierarquização é feita também através do tamanho dos nós num intervalo de tamanho [10,25] em que o tamanho aumenta em proporção do aumento do número de publicações.

Foi também feita uma comparação do comportamento dos nós através do grau e através do número de publicações (Figura 37). Para esta análise apenas o nó 2744 possui bons resultados tanto no grau como nas publicações, sendo que os restantes não possuem um comportamento coincidente entre o grau de o número de publicações. Complementando estes resultados foi calculado o coeficiente de correlação entre o grau e o número de publicações, obtendo r = 0,522934. O que significa que o grau e o número de publicações possuem uma correlação não muito elevada mas positiva.

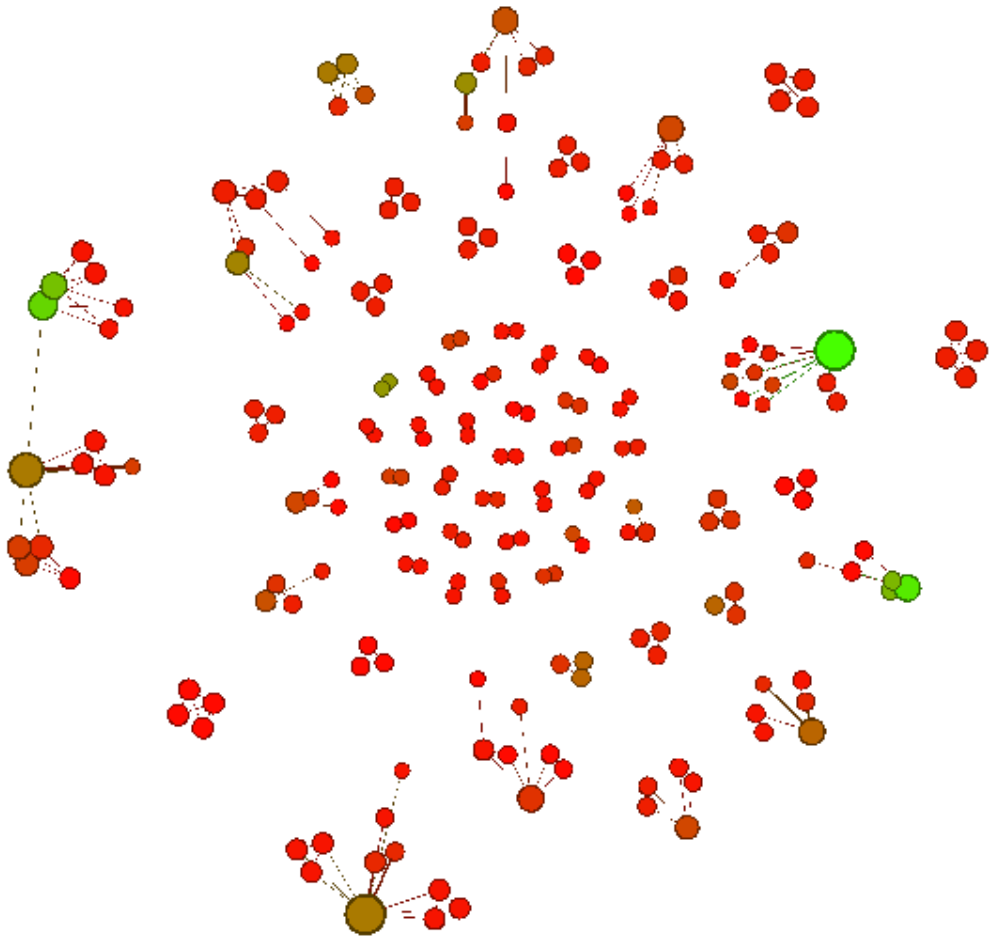

Figura 37 - Rede hierarquizada pelo número de publicações através da coloração dos nós, sendo verdes os nós com mais publicações e vermelhos os com menos publicações. A hierarquização é feita também através do tamanho dos nós num intervalo de tamanho [10,25], no entanto esta é relativa ao grau.

A coloração dos nós consoante o peso das publicações na rede é, como já referido acima, uma análise pouco relevante dado que as ligações estão por definição hierarquizadas pelo valor associado através do tamanho. Para além disto, já foi efetuado através do número de publicações de cada nó. Esta análise permite descobrir quais os pares com mais obras publicadas em conjunto (Figura 38). Este conjunto é sem dúvida o constituído pelos autores 11108 e 67730, que possuem um peso de 2,985 seguidos pelos grupos 14209 e 66596, 18645 e 14209 e o grupo 18645 e 66596. Este resultado é deveras interessante pois verifica-se que estas três díades são constituídas por apenas 3

autores que vão trabalhando uns com os outros. Este facto mostra a existência de uma grande cumplicidade entre os autores.

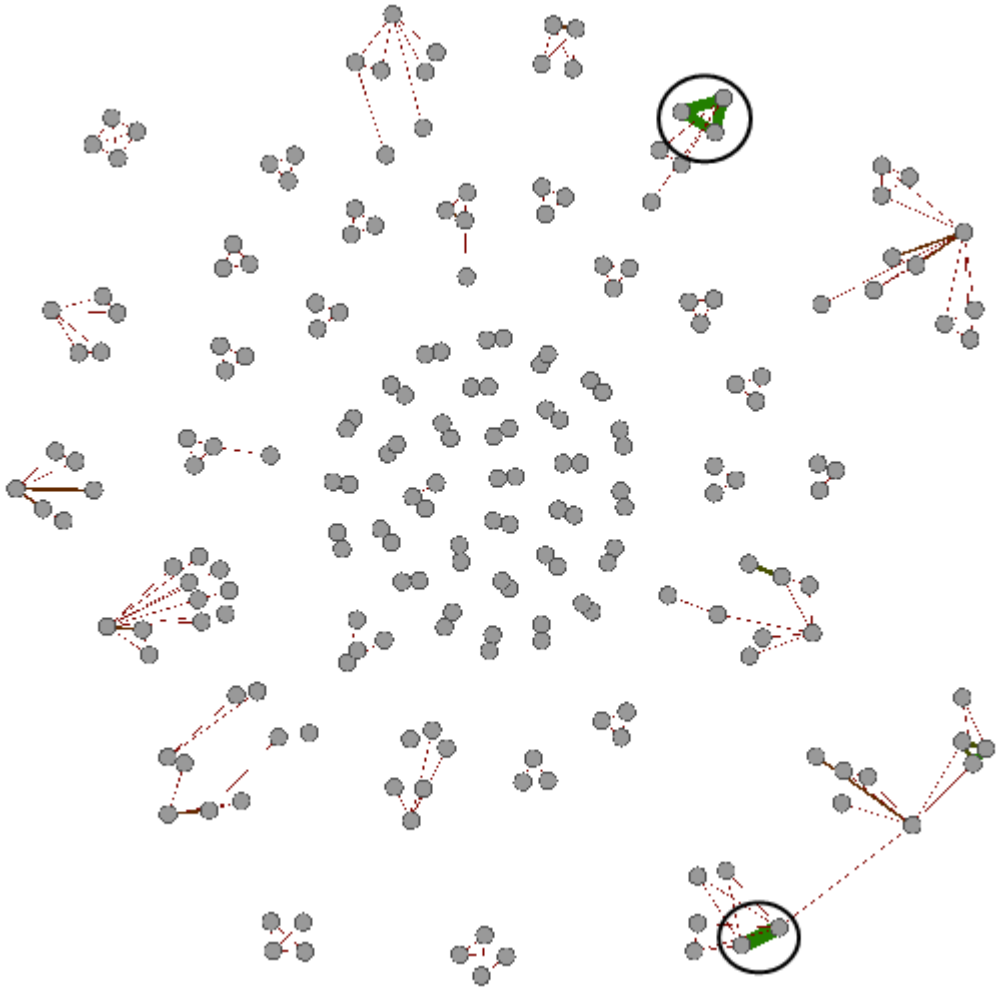

Figura 38 – Rede com as ligações hierarquizadas pelo peso do número de publicações na rede através da coloração, sendo verde para as ligações com um peso elevado e vermelho com peso baixo, e por definição através do tamanho sendo as linhas mais espessas quanto maior o peso. As ligações com maior peso estão realçadas através de uma linha circular.

6.1.1 Análises estatísticas da rede

Como já referido o *package Gephi* permite o cálculo de algumas análises estatísticas sobre a rede, dando assim uma imagem mais pormenorizada. Os resultados aqui apresentados vão centrar-se apenas nas médias e nos nós mais relevantes, sendo que os resultados para cada nó são apresentados no Anexo 18.

No cálculo do grau médio agora é apenas analisado o grau dos nós, pois como as ligações são não dirigidas não existe nem o *In-degree* nem o *Out-degree*. O grau médio

desta rede é de 2,009 significando que em média um autor está ligado a dois outros autores. Observa-se na Figura 39 que a maioria dos nós possuem um grau baixo, de 1 ou 2 nós. Apenas dois nós têm um grau de 10, sendo estes o 2744 e o 9111, que fazem aumentar a média. Mais uma vez, é de referir que na amostra analisada não existe muita cooperação entre os autores.

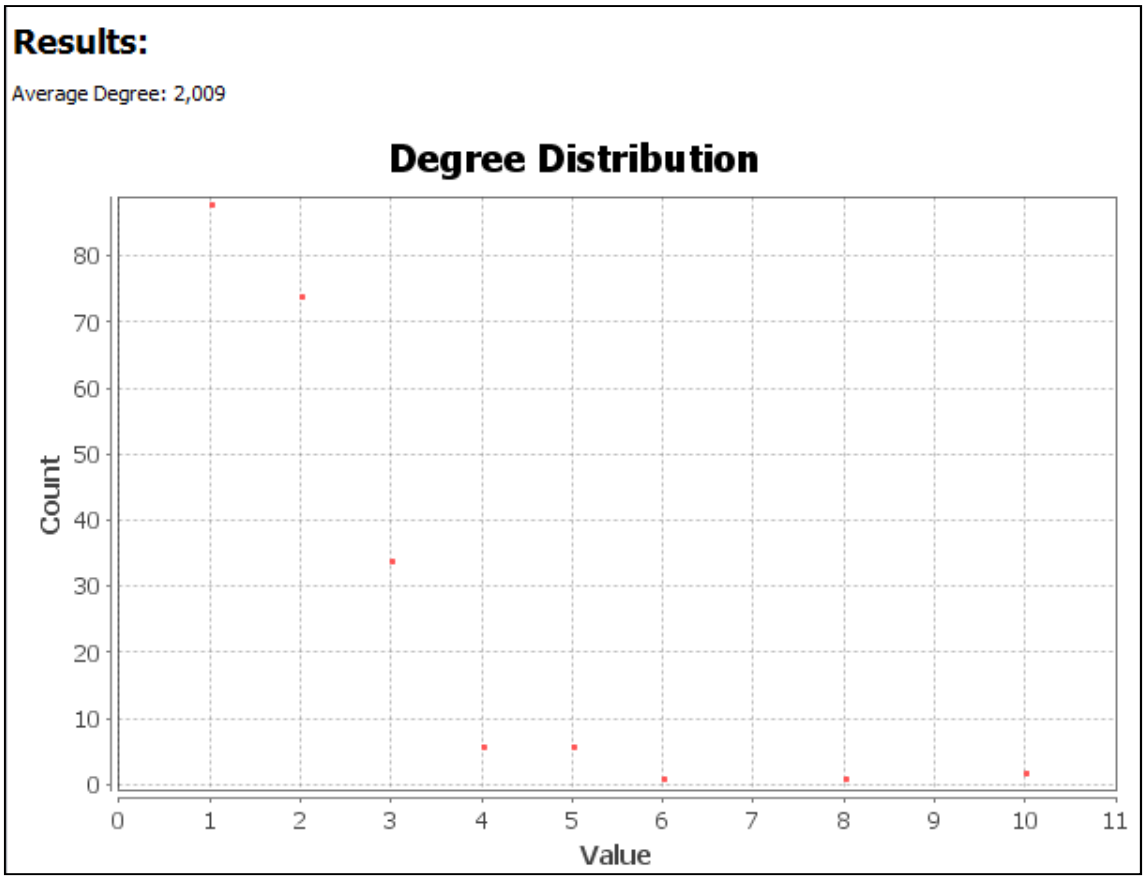

Figura 39 – Gráfico da distribuição do grau dos nós. Eixo X corresponde ao valor do grau e o eixo Y corresponde ao índice de frequência de nós com esse grau.

Enquanto a média do grau apenas soma o número de ligações de um nó, na *Average Wheighted Degree* ou média ponderada do grau, a média é calculada somando os pesos das ligações dos nós. Para esta rede de autores a média ponderada do grau é de 0,943. Apesar de esta média ser mais fidedigna é, no entanto mais difícil de interpretar. Mesmo assim, é facilmente visível pelo gráfico da Figura 40 que esta média é muito baixa.

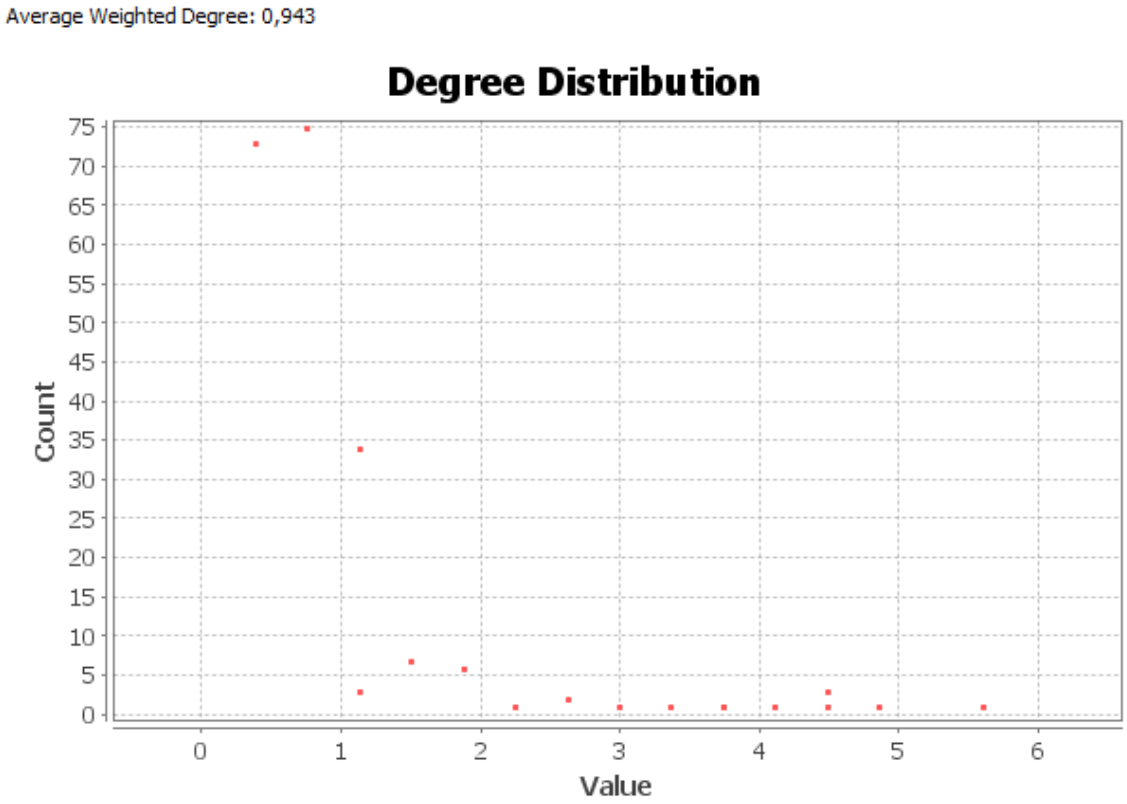

Figura 40 – Gráfico da distribuição do grau ponderado dos nós. Eixo X corresponde ao valor do grau ponderado e o eixo Y corresponde à contagem de nós.

A análise do diâmetro da rede fornece-nos várias estatísticas relacionadas com as distâncias: o diâmetro, o raio, o comprimento do caminho médio e o número de caminhos mais curtos. Esta rede possui um diâmetro de 4, significando que a distância máxima entre qualquer par de nós é de apenas 4 nós. Isto pode ser causado pela existência de muitos pequenos grupos que não estão unidos, criando assim muitas falhas de ligação, falhas estas que como já referido configuram uma distância nula. O raio da rede representa a menor distância entre um par de nós, que logicamente é de 1. Na ligação de qualquer par de nós, a distância média é de 1,714, ou seja, em média a distância das ligações entre quaisquer dois autores da rede é de 1,714 nós. Mais uma vez devido à existência de muitos pequenos grupos sem ligações existem uma grande quantidade de caminhos mais curtos, 936.

Na Figura 41 está representado o gráfico com a *Betweenness Centrality Distribution,* obtendo-se um valor interessante: a maioria dos nós não aparece no meio da ligação de um par de nós através de caminhos mais curtos. O que é compreensível devido aos

resultados acima obtidos, a distância média entre dois nós é de apenas 1,714, ou seja em média existe apenas 1 nó no meio de uma ligação entre dois pares. Dado este resultado ser para caminhos normais, nos caminhos mais curtos a distância seria ainda mais curta, provocando assim este elevado número de nós que não intercedem ligações mais curtas entre nós. É no entanto possível verificar que um dos nós possui uma *Betweenness Centrality* de 67, este nó é o autor 1154. O autor 11108 (47,5) e o 2744 (44) também possuem bons resultados; é também de notar que estes três autores possuem um grau elevado podendo este ser um indicador justifica os bons resultados na *Betweenness Centrality.*

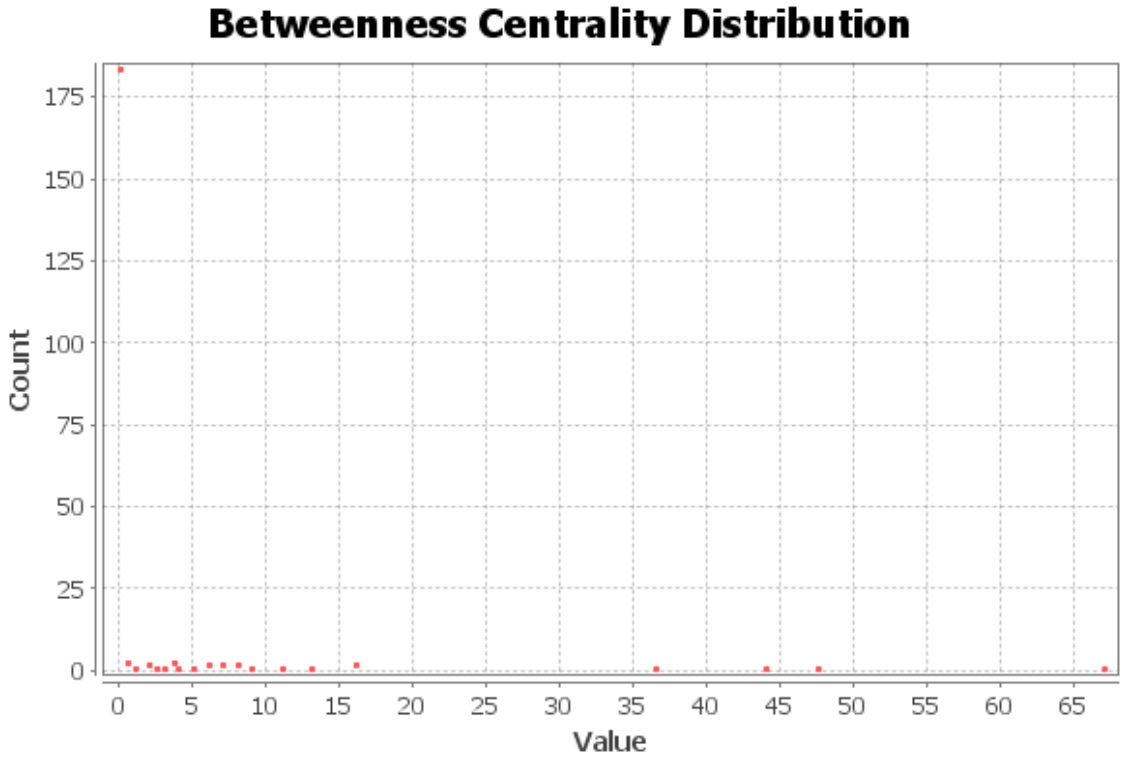

Figura 41 – Distribuição da *Betweenness Centrality* da rede. No Eixo X estão representados os possíveis valores da *Betweenness Centrality* e no Eixo Y está representada a contagem.

A *Closeness Centrality* fornece a distância média a partir de um determinado nó para todos os outros da rede. Aqui temos a maioria dos nós como uma *Closeness Centrality* de 1, sendo que para os restantes a distância média varia entre o 1 e 3. Mais uma vez se verifica o quão curtos os caminhos são, caso a rede estivesse mais ligada, estas distâncias seriam maiores pois existiriam mais caminhos possíveis para ligar os nós.

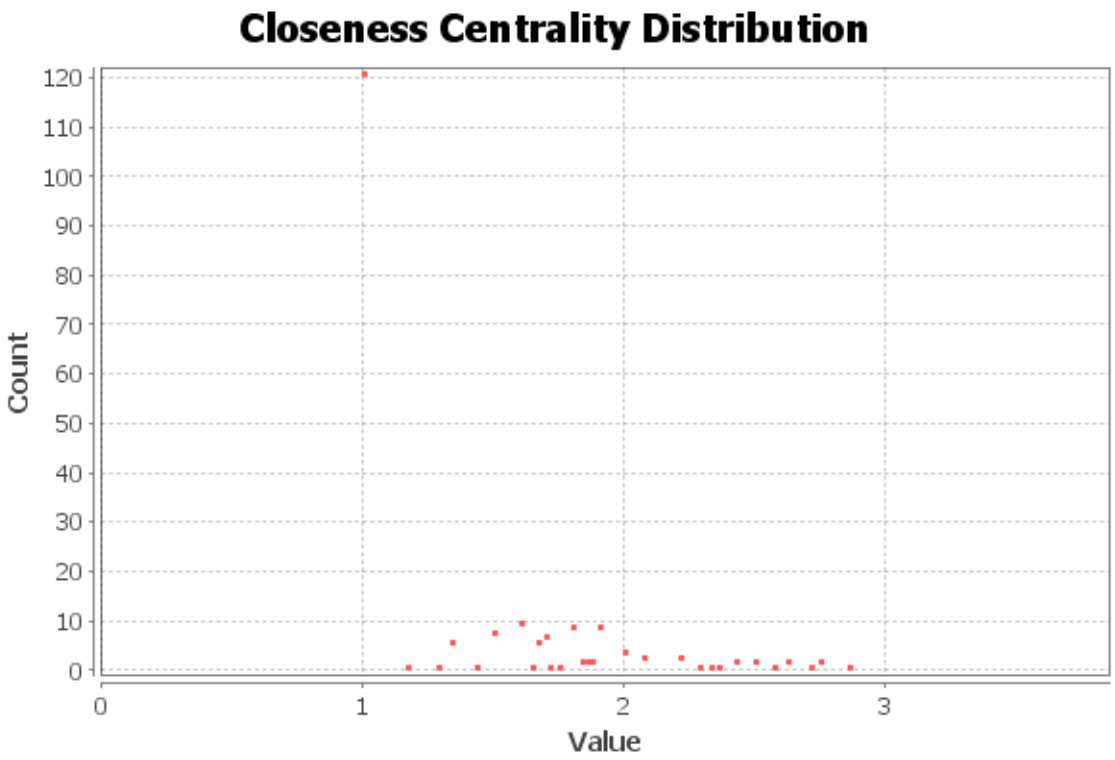

Figura 42 – Gráfico da *Closeness Centrality Distribution* com o Eixo X representando os valores desta medida e no Eixo Y a contagem destes valores.

A *Eccentricity Distribution* vai calcular a distância de um determinado nó ao correspondente nó mais longe na rede. Mais uma vez se observa que a maioria dos nós tem o seu nó mais distante a uma distância de apenas 1 nó. Isto mostra mais uma vez a falta de ligações entre os autores na rede. A maior distância é de apenas 4, muito curta para uma rede com este tamanho.

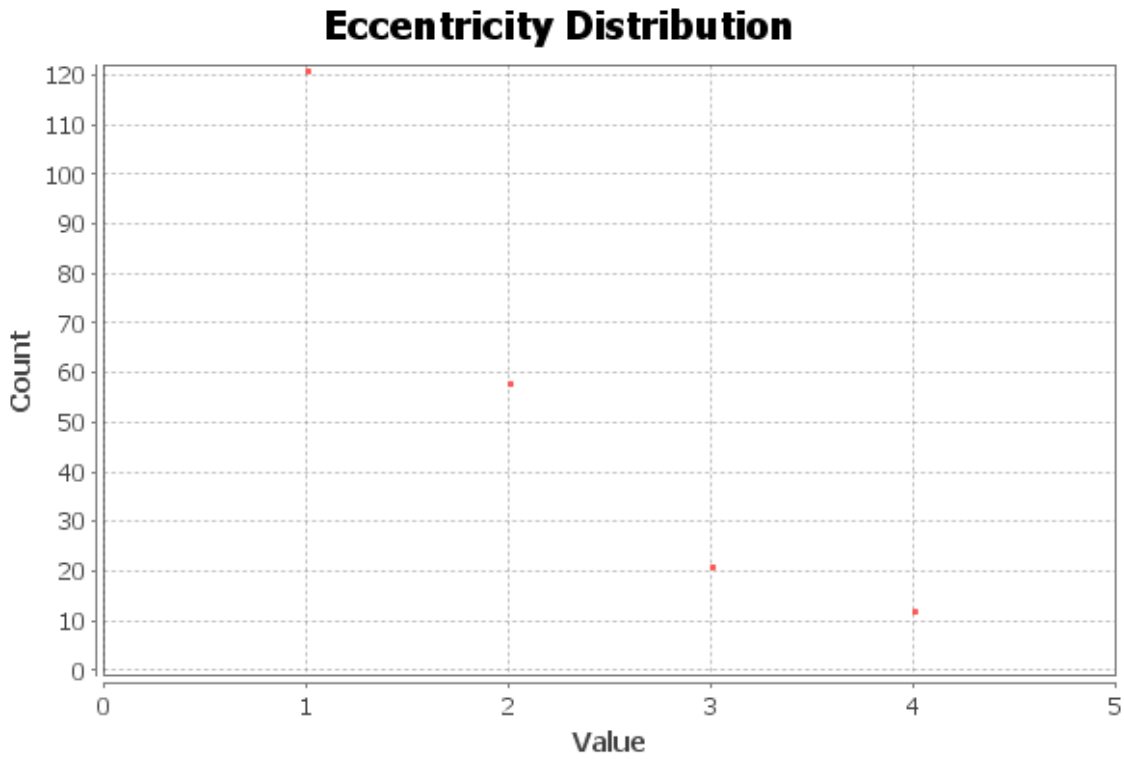

Figura 43 – Gráfico da *Eccentricity Centrality Distribution* com o Eixo X representando os valores desta medida e no Eixo Y a contagem destes valores.

Como é de esperar tanto pelos resultados obtidos como pelas conclusões já retiradas, a densidade do grafo é muito baixa, de 0,010. Como quanto mais completa a rede, mais próxima será de 1, esta rede está muito incompleta. Mesmo sendo os dados todos de publicações na área da Economia existem muito poucas ligações entre os autores. Podendo talvez existir uma especialização dos autores na área geral de publicação e mais especificamente na ISI.

Na Figura 44 estão representadas as distribuições de *Hubs e Authority*. Infelizmente os resultados destas estão muito aquém do desejado, dado que tanto as ligações como os nós possuem informação de fraca qualidade e pouco valiosa.

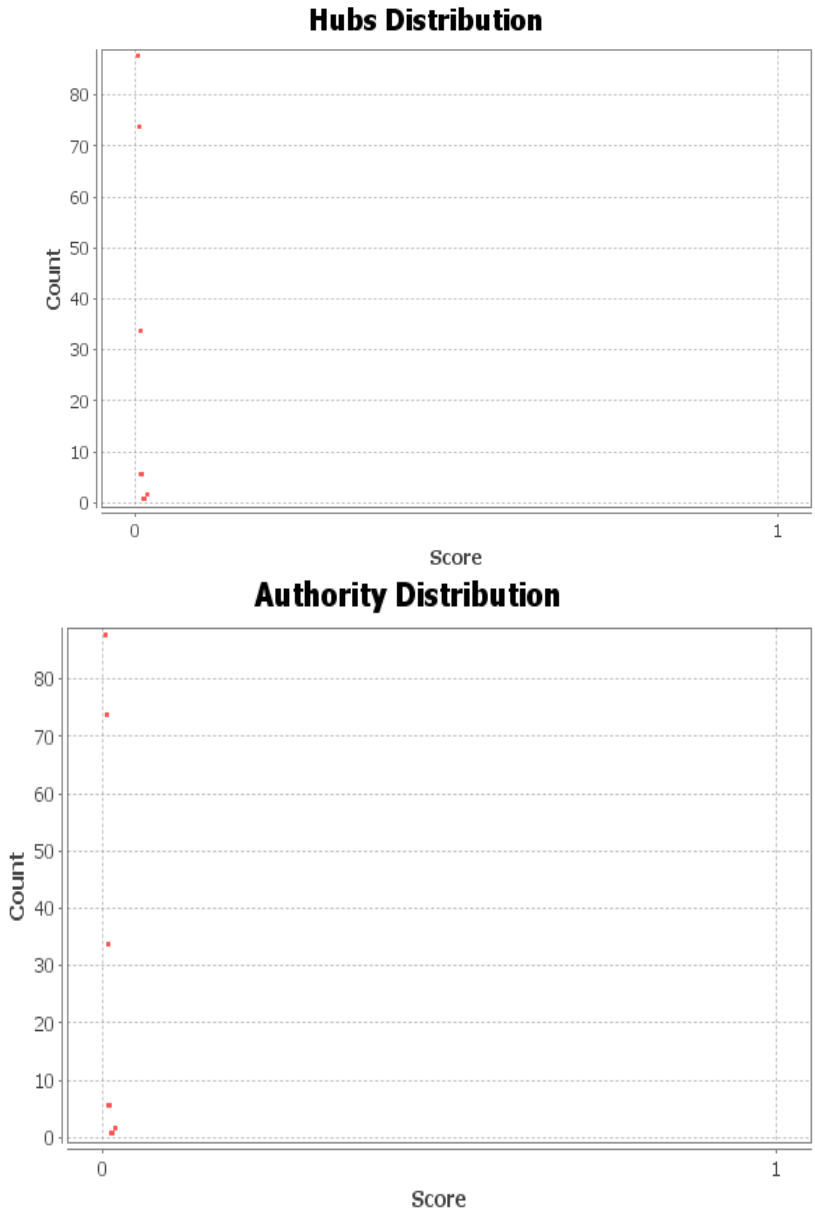

Figura 44 – Gráficos do *Hubs e Authority Distributions* com o Eixo X representando os valores destas medidas e no Eixo Y a contagem destes valores.

Para a deteção de comunidades, foi novamente ignorado o peso dos nós. Para esta rede e tendo em conta os resultados já obtidos, é expectável que existam bastantes comunidades; a resolução da *Modularity* é de 1. Com estas características o *package Gephi* conseguiu detetar 61 comunidades (ver Anexo 19).

Numa abordagem mais direcionada para o estudo dos nós, é verificado que em média o *clustering* dos nós nem é muito baixo, ficando nos 0,498. Ao analisar a rede é

imaginável que tal possa acontecer pois embora a rede seja um pouco dispersa, existem muitos grupos de nós bastante unidos. Tal é verificado na Figura 45.

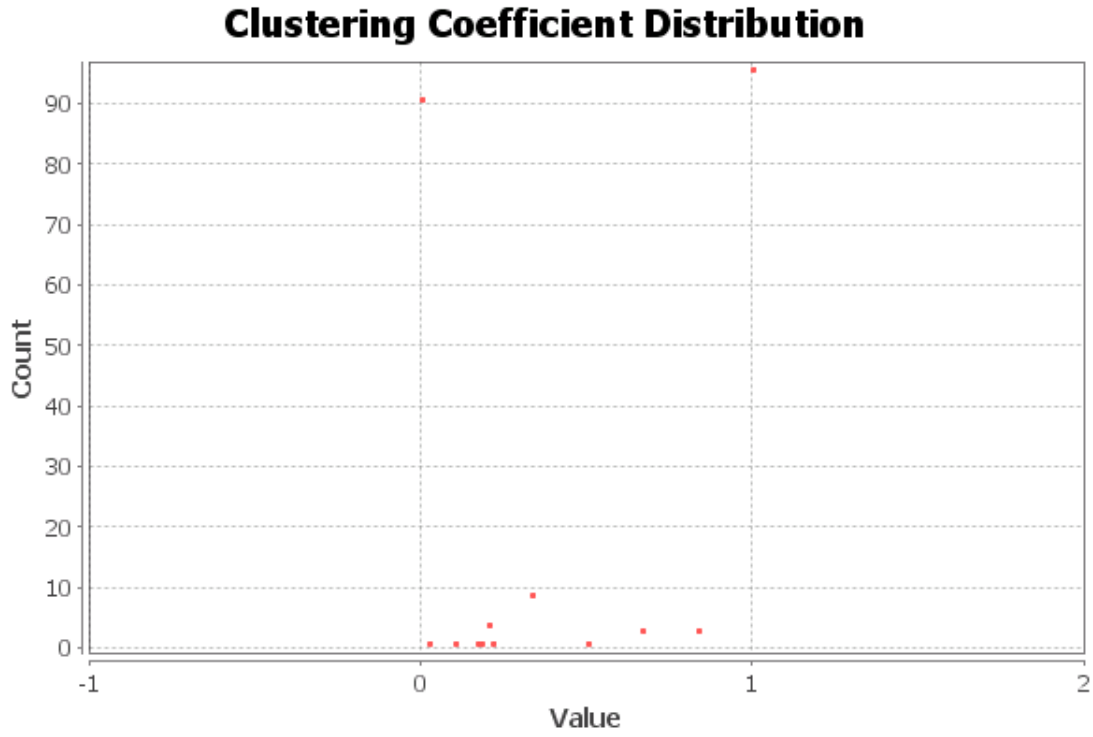

Figura 45 - Gráfico da *Clustering Coefficient Distribution* com o Eixo X representando os valores desta medida e no Eixo Y a contagem destes valores. Os pontos vermelhos representam a quantidade de nós que possuem um determinado valor de *Clustering Coefficient.*

O cálculo do *Eigenvector Centrality* é mais uma medida de análise dos nós da rede, possibilitando analisar se um nó é ou não importante na rede baseado nas suas ligações. É de prever que não existam muitos nós com muita importância, pois estes possuem poucas ligações e maioritariamente fracas. Como é visível na Figura 46, a maioria dos nós possuem um *Eigenvector Centrality* inferior a 0,5, o que significa que a maioria dos nós não possui uma grande importância na rede. Existem no entanto alguns nós que se destacam com ótimos resultados como o 9111 (1), o 1154 (0,997) e o 11108 (0,709), valores estes observáveis no Anexo 20. Já era de esperar serem estes os autores mais relevantes na rede dado que têm conseguido bons resultados em quase todas as análises efetuadas à rede.

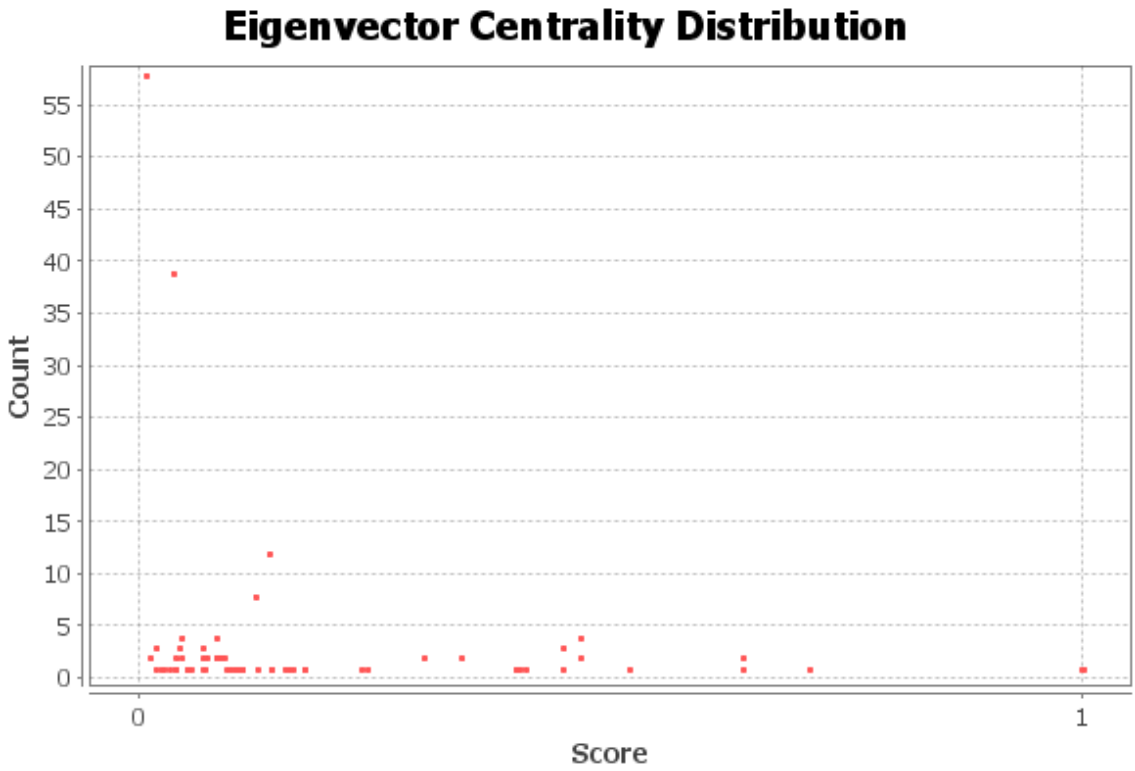

Figura 46 – Gráfico da *Eigenvector Centrality Distribution* com o Eixo do X representando os valores desta medida e no Eixo do Y a contagem destes valores.

Em suma, das análises efetuadas à rede é verificável que esta rede é muito dispersa, existindo poucos autores com relevância na rede. É necessário apostar na partilha de conhecimentos e experiências, como estas publicações têm como área "Economia" deveria ser possível aos autores trabalharam mais em conjunto. No entanto, é também compreensível que os autores se especializem em determinados ISI de modo a conseguirem estudos mais aprofundados e mais exaustivos, e assim criando grupos pequenos como os do caso em estudo. Contudo, como já referenciado nesta rede, existem os autores 9111, 1154, 11108 e o 2744 que possuem muito bons resultados dentro da rede. Estes devem ser os autores mais conceituados possuindo mais publicações e mais ligações. É no entanto de notar que estes autores pertencem a grupos separados, podendo uma das razões ser o facto de eles publicarem em ISI diferentes, ou então por serem os mais conceituados competirem entre si. Alegando ser esta uma justificação plausível, deveriam então existir muitos mais nós como estes, sendo que seria possível isto acontecer para cada ISI ou grupo de ISI similares. No entanto isto apenas aconteceria se houvesse um número similar de autores a publicar em todas as ISI

ou grupos de ISI, algo que claramente não acontece, sendo determinadas ISI mais importantes que outras.

6.2 Análise Conceptual

A análise conceptual vai permitir a caracterização dos autores criando grupos de autores mediante as ISI das suas publicações. No Anexo 21 estão presentes os dados já preparados para serem tratados no FCbO, com isto vamos obter as intensões dos conceitos, através das quais futuramente irão ser determinadas as respetivas extensões. Como esta metodologia já foi previamente explicada é apenas relevante a apresentação dos resultados tanto das intensões como das extensões. Assim as intensões estão apresentadas no Anexo 22 e os resultados das extensões no Anexo 23. É importante também mencionar que o código R utilizado para estes dados teve de sofrer algumas alterações, porque o número de colunas do *input data* possui mais colunas que o do exemplo dos países (Anexo 24). Através dos algoritmos utilizados foi possível encontrar 113 grupos de atores, correspondentes aos conceitos identificados.

Estes resultados não permitem muitas análises estatísticas, no entanto é possível determinar a quantidade de conceitos a que cada autor pertence e a respetiva percentagem. Isto tem relevância pois permite definir os atores que estão mais entrosados na comunidade e possuem publicações num maior número de ISI's. Mais uma vez os autores 2744 e 11108 estão muito bem posicionados sendo estes os que pertencem a mais grupos (Tabela 8). O autor 1154 presente em quinto lugar, está presente em 23 grupos, ou seja, em cerca de 20% dos grupos. Para testar a teoria de autores com publicações em mais ISI estarem mais propensos a pertencer a mais grupos foi calculado na Tabela 9 o número de ISI em que os autores publicaram. Os 5 autores que pertencem a mais grupos estão também entre os autores que publicam em mais ISI: como o algoritmo do FCbO pode criar um grupo de autores para todas as possíveis combinações de atributos (ISI), logo quanto mais atributos o autor tiver maior a probabilidade de pertencer a um maior número de grupos.

Posição	Autores	Número de conceitos	Percentagem de grupos
1º	2744	41	0,362831858
2º	11108	32	0,283185841
3º	67730	31	0,274336283
4º	67878	23	0,203539823
5º	1154	23	0,203539823
6º	18645	22	0,194690265
7º	25544	18	0,159292035
8º	4824	18	0,159292035

Tabela 7 –Autores, número de conceitos a que o autor pertence e respetiva percentagem.

Autor	Número de ISI
2744	10
18645	9
3794	9
557	9
67288	8
13825	7
11108	7
69565	7
67730	7
66596	6
14209	6
1154	6
67878	6

Tabela 8 –Autores e número de ISI em que cada um já publicou, os autores destacados a amarelo são os autores mais mediáticos nas análises até agora efetuadas.

Esta análise é maioritariamente concebida com o propósito de juntar os nós similares em grupos, para com os dados da rede de coautorias calcular as percentagens de autores que trabalham em conjunto e possuem as mesmas características, para os que trabalham em conjunto apesar de não partilharem as mesmas características e por fim a percentagem de autores que possuem as mesmas características mas nunca publicaram em conjunto. Esta é a principal estatística a ser analisada, pois o intuito do trabalho é o de identificar autores que poderiam trabalhar em conjunto mas ainda não o fizeram.

Para efeitos comparativos, é também feita a análise conceptual para a matriz de adjacência da tabela das relações entre os autores. Analogamente às análises conceptuais já efetuadas, são definidos conceitos com os autores com características similares. Estes resultados serão superficialmente comparados com o processo de deteção de comunidades do *package Gephi*. A matriz de adjacência utilizada encontra-se no Anexo 25, devido a especificidades da tabela .dat criada com a matriz de adjacência (ver Anexo 26) é necessária a devida adaptação do código em linguagem R que vai detetar as extensões através das intensões fornecidas pelo *software* FCbO (intensões e código no Anexo 27, e extensões no Anexo 28). Analisando a deteção de comunidades do *package Gephi*, depreende-se que o FCbO é mais generalista ou abrangente pois detetou 134 conceitos, enquanto que o *package Gephi* detetou 61 comunidades (ver Anexo 19). No entanto, caso nos resultados do FCbO existam muitos conceitos com apenas 1 autor, eles não serão muito uteis. É importante referir que as comunidades criadas no *package Gephi* são em pouco similares às criadas pelo FCbO.

Para as próximas análises foi feito um refinamento dos grupos de autores nos resultados da Análise Conceptual inicial. Isto, de modo a que estas obtenham resultados viáveis e mais próximos da realidade. Como é compreensível existem ISI em que há mais autores a publicarem, por serem mais populares ou mais genéricas, e existem ISI em há muito poucos autores a publicar por serem o mais específicas. Isto originou a que houvesse grupos que fossem constituídos por muito poucos autores e grupos compostos por uma elevada quantidade de autores. Para refinar estes dados foram eliminados da lista de grupos os grupos com apenas 1 autor, e os dois maiores grupos que possuíam 104 e 105 autores, restando 99 dos 113 conceitos iniciais.

6.3 Análise Conjunta: Sistema de Recomendação de Parcerias

No presente capítulo irão ser apresentadas algumas análises conjuntas dos dados da rede de coautorias e dos conceitos obtidos pela Análise Conceptual Formal. No final do capítulo será apresentado o Sistema de Recomendação que visa ser uma ferramenta de ajuda capaz de recomendar coautores a um determinado autor, tendo em conta os autores com quem já fez parcerias e as áreas em que publica.

Inicialmente é analisado se os autores que publicaram em conjunto estão juntos em pelo menos um dos conceitos obtidos. Para tal foi criado um código usando a linguagem de programação do Excel, Visual Basic Application. O algoritmo na Figura 47 funciona com base nos dados do Anexo 29, em que na primeira folha se estão as parcerias nas publicações, sendo que cada linha corresponde a uma publicação. Na segunda folha encontra-se a lista de grupos criados pelo algoritmo FCbO (extensões dos conceitos obtidos), em que cada linha corresponde a um grupo e os elementos são autores com características similares. Deste modo o algoritmo começa por escolher o primeiro autor da primeira linha de publicações e de seguida procura-o nos grupos da folha dois, quando encontra o autor, vai procurar nesse mesmo grupo o segundo autor da primeira publicação. Se o encontrar nesse grupo vai escrever “OK” na célula em frente aos autores da publicação caso contrário continua a pesquisa nos restantes grupos da mesma forma. Isto é repetido para todos os autores e consequentemente todas as publicações.

```
Sub teste()
Dim x, y, z

x = Worksheets(1).Cells(3, 6)
y = Worksheets(1).Cells(5, 6)
z = Worksheets(1).Cells(7, 6)

For i = 1 To x
For j = 1 To y
For k = 1 To z

If Worksheets(2).Cells(j, k) = Worksheets(1).Cells(i, 1) Then
For w = 1 To z
If Worksheets(2).Cells(j, w) = Worksheets(1).Cells(i, 2) Then
Worksheets(1).Cells(i, 3) = "OK"
End If
Next
End If
Next
Next
Next

End Sub
```

Figura 47 – Código utilizado para descobrir se numa determinada parceria ambos os autores pertencem a um dos grupos da lista.

Com este refinamento os resultados obtidos são bastante satisfatórios sendo que a maioria das parcerias para uma publicação são criadas por autores com características idênticas. Em 86% das publicações (183) os autores pertencem ao mesmo grupo partilhando assim as mesmas características. No final, apenas 14 % das publicações foram feitas por autores de diferentes grupos, permitindo concluir que por norma os autores escolhem um coautor com base nas suas antigas publicações e áreas de trabalho.

De seguida interessa saber se os autores com as mesmas características trabalham entre si, e se sim em que percentagem. Assim, é possível verificar se é feita uma boa gestão da base de dados de autores e se autores com áreas de trabalho semelhantes cooperam entre si.

Para tal foi criado o algoritmo apresentado na Figura 48, que tem um funcionamento similar ao representado na Figura 47. O algoritmo começa por procurar o primeiro autor do primeiro grupo na primeira coluna da lista das parcerias, quando o encontra verifica se algum dos restantes autores do grupo são o autor com quem está a fazer parceria nessa publicação. Ou seja, se é o autor que se encontra na segunda coluna da mesma linha. Caso encontre parcerias vai acrescentar 1 na coluna do nº de parcerias e vai escrever o código dos autores na coluna de parcerias por grupo (Anexo 30)

```
Sub grupos()
Dim x, y, w, z
x = Worksheets(1).Cells(3, 7)
For i = 1 To x
y = Worksheets(3).Cells(i + 1, 1) - 1
w = Worksheets(3).Cells(i + 1, 1)
For j = 1 To y
z = Worksheets(1).Cells(3, 6)
For l = 1 To z
If Worksheets(2).Cells(i, j) = Worksheets(1).Cells(l, 1) Then
For r = j + 1 To w
If Worksheets(2).Cells(i, r) = Worksheets(1).Cells(l, 2) Then
Worksheets(3).Cells(i + 1, 2) = Worksheets(3).Cells(i + 1, 2) + 1
Worksheets(3).Cells(i + 1, 7) = Worksheets(3).Cells(i + 1, 7) & Worksheets(2).Cells(i, j) & " " & Worksheets(2).Cells(i, r) & ";"
End If
Next
End If
Next
For m = 1 To z
If Worksheets(2).Cells(i, j) = Worksheets(1).Cells(m, 2) Then
For s = j + 1 To w
If Worksheets(2).Cells(i, s) = Worksheets(1).Cells(m, 1) Then
Worksheets(3).Cells(i + 1, 2) = Worksheets(3).Cells(i + 1, 2) + 1
Worksheets(3).Cells(i + 1, 7) = Worksheets(3).Cells(i + 1, 7) & Worksheets(2).Cells(i, j) & " " & Worksheets(2).Cells(i, s) & ";"
End If
Next
End If
Next
Next
Next


End Sub
```

Figura 48 - Código utilizado para descobrir se numa determinada parceria ambos os autores pertencem a um dos grupos da lista

Com os resultados do algoritmo é possível calcular a média da percentagem de parcerias por grupo, sendo que esta, devido ao elevado número de autores em alguns grupos, é bastante baixa (35%). Logicamente a percentagem média de parcerias que não foram efetuadas num grupo é de 65%. Isto permite concluir que ainda existem muitas parcerias que podem ser criadas.

De forma a saber quais as parcerias que podem ser criadas é utilizado o algoritmo da Figura 49. Este é um algoritmo bastante simples; usando a mesma ideia dos algoritmos

anteriores começa por selecionar o primeiro autor da primeira linha e imprime todas as parcerias que este pode fazer, e assim sucessivamente por todos os autores.

```
Sub parcposs()

Dim x, y, w
x = Worksheets(1).Cells(3, 7)
For i = 1 To x
y = Worksheets(3).Cells(i + 1, 1) - 1
w = Worksheets(3).Cells(i + 1, 1)

For j = 1 To y
For l = j + 1 To w

Worksheets(3).Cells(i + 1, 8) = Worksheets(3).Cells(i + 1, 8) & Worksheets(2).Cells(i, j) & " " & Worksheets(2).Cells(i, l) & ";"

Next
Next
Next

End Sub
```

Figura 49 - Código utilizado para identificar todas as parcerias possíveis num determinado grupo.

O único problema deste método é que a análise das parcerias existentes num grupo e as que são passíveis de existir tem de ser feita manualmente, o que exige algum tempo.

Estas análises vieram confirmar o já expectável ao analisar a rede de coautorias: que ainda existem muitas parcerias que podem ser criadas. Para uma análise futura seria interessante realizar uma análise mais individualizada, criando uma lista por autor de parcerias já criadas e de parcerias passíveis de serem criadas. Para facilitar esta análise foi criado um ficheiro em Excel capaz de restituir os coautores de um determinado autor à escolha, e um conjunto de autores com quem poderia fazer parcerias, pois pertencem aos mesmos grupos do autor escolhido (Anexo 31).

A utilização do ficheiro é muito simples, na Figura 50 está representada a única parte que deve ser trabalhada pelo utilizador. Esta é composta por 3 campos e dois botões, um campo para os autores criado com uma *drop down* com a lista dos autores, um campo em que vão aparecer as parcerias já efetuadas e um outro campo onde irão aparecer as parcerias ainda não efetuadas mas que seriam plausíveis. Os dois botões estão associados a duas macros, o botão “Apagar info” apaga a informação relativa às parcerias (para a macro principal funcionar os campos precisam de estar limpos). O botão “Correr Macro” é o que ativa a macro que vai encontrar os autores para as parcerias.

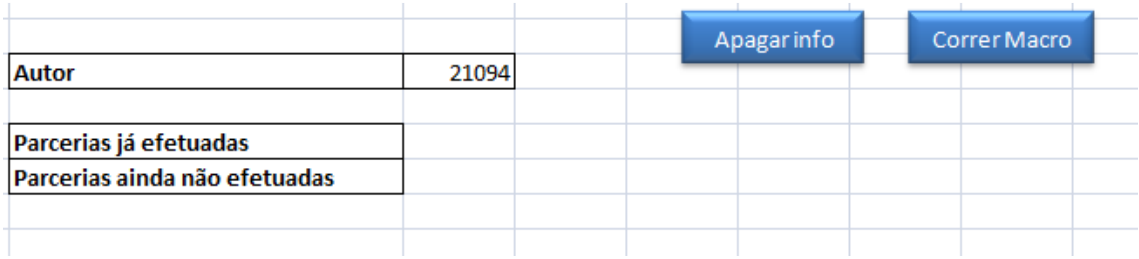

Figura 50 – Menu de funcionamento do Sistema de Recomendação

A macro para apagar a informação, é uma macro muito simples que apenas seleciona toda a informação à direita e para baixo da célula "C5".

```
Sub delete()
'
' delete Macro

    Range("C5").Select
    Range(Selection, Selection.End(xlToRight)).Select
    Range(Selection, Selection.End(xlToRight)).Select
    Range(Selection, Selection.End(xlDown)).Select
    Range(Selection, Selection.End(xlDown)).Select
    Selection.ClearContents
End Sub
```

Figura 51 – Algoritmo da macro para apagar a informação das parceiras

A macro para escolher os autores é feita com base em dois algoritmos, um para detetar as parcerias existentes e outro para detetar as parcerias passíveis de serem criadas. De forma a não ter muitos algoritmos incorporados no mesmo algoritmo foi criado, por motivos de segurança, um formulário que questiona se as células foram apagadas antes começar a correr a macro. Assim é possível escolher "Sim" para correr a macro ou "Não" para primeiro apagar a informação.

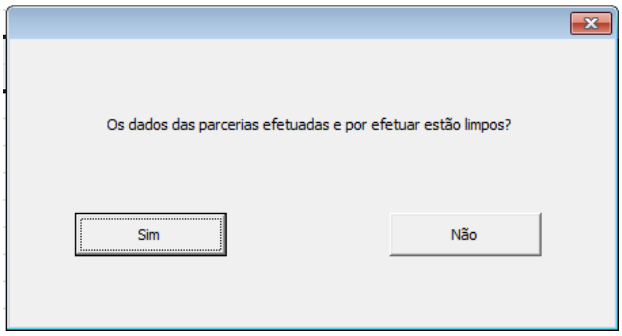

Figura 52 – Formulário inicial para correr a macro para encontrar os autores para parcerias, se pressionar o botão Sim a macro é corrida diretamente, se pressionar Não uma macro apaga então a informação.

Na Figura 53 está representada a primeira parte da macro principal, a variável X representa o número de linhas da folha “Parcerias” que contém todas as parcerias efetuadas, em que cada linha corresponde a uma parceria para uma publicação. De seguida temos duas funções *FOR* que fazem as variáveis “i” e “s” adotarem valores entre 1 e x, sendo que neste caso x é 213. Para cada função *FOR* existe uma função *IF* que vai procurar o autor que estamos a pesquisar inicialmente na primeira coluna, devolvendo assim o autor da segunda coluna como um coautor. E de seguida procura o autor na segunda coluna retribuindo o autor da primeira coluna presente na mesma linha como sendo um coautor. Como todos os dados dos coautores são armazenados numa única célula é preciso no final separar esses autores colocando um por cada célula.

```
Dim x, y, z, w

x = Worksheets(4).Cells(4, 7)

For i = 1 To x

If Worksheets(1).Cells(3, 3) = Worksheets(2).Cells(i, 1) Then
Worksheets(1).Cells(5, 3) = Worksheets(1).Cells(5, 3) & " " & Worksheets(2).Cells(i, 2)
End If
Next
For s = 1 To x
If Worksheets(1).Cells(3, 3) = Worksheets(2).Cells(s, 2) Then
Worksheets(1).Cells(5, 3) = Worksheets(1).Cells(5, 3) & " " & Worksheets(2).Cells(s, 1)
End If
Next
If Worksheets(1).Cells(5, 3) = "" Then
Else
Range("C5").Select
Selection.TextToColumns Destination:=Range("C5"), DataType:=xlDelimited, _
 TextQualifier:=xlDoubleQuote, ConsecutiveDelimiter:=True, Tab:=True, _
 Semicolon:=False, Comma:=False, Space:=True, Other:=False, FieldInfo _
 :=Array(Array(1, 1), Array(2, 1), Array(3, 1), Array(4, 1)), TrailingMinusNumbers:= _
True
End If
```

Figura 53 – Algoritmo para identificar os autores que já fizeram publicações com um determinado autor.

O próximo algoritmo é já mais complicado, tendo sido preciso criar algumas condições para que os valores fossem corretos e não contivessem informações desnecessárias e/ou erradas. O algoritmo começa por fazer uma pesquisa ao longo das linhas, ou seja dos conceitos, para cada linha (a escolha é feita da primeira para a última linha) o algoritmo começa então a pesquisar nessa linha todas as células à procura do autor em análise. Quando encontra o autor é iniciada uma nova pesquisa, na mesma linha, mas começando novamente do início e cada membro do conceito é comparado com os autores que já possuem ligações com o autor principal. Se o autor do conceito já estiver presente na lista de coautores então nesse momento é colocado um 1 na célula

"A6" e se não se encontrar já na lista não é feito nada. Depois de o autor do conceito selecionado ser comparado com todos os coautores é verificado a célula "A6" para aferir se o valor desta é igual a 1 ou não, se for igual a 1 significa que o autor já pertence à lista de coautores e caso contrário é escrito o código do autor na célula "C6". É importante salientar que caso o autor do conceito não apareça na lista de coautores é verificado também se este autor é o nosso autor principal, pois como estão a ser analisados os conceitos onde o autor aparece, logicamente o autor principal também vai aparecer na análise.

```
y = Worksheets(4).Cells(6, 7) + 3
z = Worksheets(4).Cells(5, 7)
For j = 1 To z
w = Worksheets(4).Cells(j, 9)
For l = 1 To w
If Worksheets(1).Cells(3, 3) = Worksheets(3).Cells(j, l) Then
For r = 1 To w
For h = 3 To y
If Worksheets(1).Cells(5, h) = Worksheets(3).Cells(j, r) Then
Worksheets(1).Cells(6, 1) = 1
Else
If Worksheets(3).Cells(j, r) = Worksheets(1).Cells(3, 3) Then
Worksheets(1).Cells(6, 1) = 1
End If
End If
Next
If Worksheets(1).Cells(6, 1) > 0 Then
Else
Worksheets(1).Cells(6, 3) = Worksheets(1).Cells(6, 3) & " " & Worksheets(3).Cells(j, r)
End If
Worksheets(1).Cells(6, 1) = 0
Next
End If
Next
Next
If Worksheets(1).Cells(6, 3) = "" Then
Else
Range("C6").Select
Selection.TextToColumns Destination:=Range("C6"), DataType:=xlDelimited, _
 TextQualifier:=xlDoubleQuote, ConsecutiveDelimiter:=True, Tab:=True, _
 Semicolon:=False, Comma:=False, Space:=True, Other:=False, FieldInfo _
 :=Array(Array(1, 1), Array(2, 1), Array(3, 1), Array(4, 1)), TrailingMinusNumbers:= _
True
End If
Unload Me
```

Figura 54 – Código para o preenchimento dos possíveis autores com que o autor em análise pode trabalhar dado estarem nos mesmos grupos.

Os resultados são apresentados como demonstrado na Figura 55, retribuindo separadamente os autores que já fizeram publicações com o autor principal e os autores que pertencem ao mesmo conceito mas não fizeram até então nenhuma parceria. Como estes dados são individualizados, não serão sujeitos a nenhuma análise posterior. Este resultado pode ser utilizado por algum autor que queira saber autores que estejam

relacionados consigo para publicarem juntos. A título de exemplo foi utilizado o autor 3923, este já publicou em parceria com o autor 3794 e com o 557. No entanto, podia publicar com os autores 1374, 1823, 81774, 4843, 67288, 15456, 7357, 9934, etc., porque estes autores pertencem a conceitos que o autor 3923 também pertence.

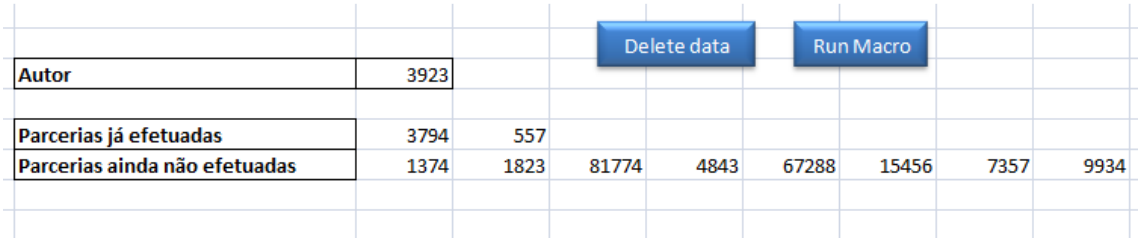

Figura 55 – Representação dos resultados obtidos depois de analisado o autor 3923, indicando os autores com quem publicou e os autores com quem ainda não publicou mas pertencem aos mesmos conceitos.

Devido aos bons resultados do processo de deteção de comunidades do *package Gephi* foi feita as análises com estas três macros de forma a comparar os resultados e verificar com qual destes se obtém melhores resultados.

Iniciando as análises pelo Anexo 32, é verificado o número de publicações que existem em que os seus autores pertencem ao mesmo grupo. Ora após esta análise é verificado que 99,53 % das publicações são feitas com autores do mesmo grupo.

Na análise do Anexo 33, é verificado que em média 87% das possíveis parcerias dentro de um conceito já se encontram efetuadas.

Por fim através do Anexo 34 é possível verificar que analisando alguns autores aleatoriamente verifica-se que o número de coautores aconselhados diminui drasticamente. Com isto, apesar de as comunidades serem bem feitas pelo *software Gephi*, estas não conseguem cumprir o seu objetivo no sistema de recomendação.

7 Conclusão

A presente dissertação aborda dois temas que estão atualmente em voga, no entanto em poucos trabalhos se encontram a operar em conjunto. A análise de redes tem tido um crescimento exponencial nos últimos anos e muitos desenvolvimentos se têm feito nesta área. Já para a análise conceptual, apesar de se investir mais nesta área criando novos e melhorados algoritmos, poucos investimentos se têm feito na criação de programas informáticos para a sua análise. Ao considerar em conjunto estas duas metodologias de análise é criada uma forte sinergia.

Conseguindo com sucesso aliar dois métodos distintos, e as respetivas mais-valias, a criação de um Sistema de Recomendação é o valor acrescentado desta dissertação. Devido à sua automatização, facilmente pode ser adaptado a dados de diferentes tipos e origens. Um possível trabalho futuro, será a utilização de um outro método de agrupamento capaz de criar grupos de indivíduos consoante as suas características, baseado em medidas de dissemelhança. Assim, seria possível analisar vários métodos e tentar apurar qual deles conseguia obter um melhor resultado a nível do sistema de recomendação.

Ao longo da dissertação foram efetuadas várias análises, inicialmente à rede de coautorias estudada, foco principal deste trabalho. Posteriormente foi feita uma análise aos conceitos determinados com base nas áreas de publicação dos autores envolvidos. A análise da rede serviu para perceber a sua composição e limitações. Claramente, a rede de coautorias “necessita” de aumentar as ligações entre os nós. Logicamente, numa rede de densidade elevada o sistema proposto não fará tanto sentido, pois os nós já possuem muitas ligações e estão bem entrosados na rede.

Posteriormente é então feita a análise conceptual, criando uma importante base para o sistema de recomendação. De facto, é através dos resultados da análise conceptual que se identificam grupos maximais de autores com as mesmas características, isto é, publicando nas mesmas áreas. Assim, para cada autor considerado, o sistema pode verificar quais os outros autores que trabalham nas mesmas áreas. Com o investimento em programas capazes de efetuar uma análise conceptual completa, esta análise seria mais simples. No nosso caso foi necessário o desenvolvimento de um algoritmo em linguagem R para a obtenção das extensões dos conceitos identificados.

Embora os investimentos na área da análise de redes sejam já muito elevados, é compreendido que esta metodologia se aplica intensamente em muitos aspetos do nosso dia a dia, pelo que se pode prever que este tipo de análise seja cada vez mais utilizado em investigação nos próximos anos.

8 Bibliografia

- Albert, R. and Barabási, A.L. (2002). “Statistical mechanics of complex networks.” Reviews of Modern Physics T4, 1: 47-97.
- Artin, E. (1998). “Galois Theory”, Dover Publications, ISBN 0486623424.
- Bank, M. e Franke, J. (2010). “Social Networks as Data Source for recommendation Systems”. In E-Commerce and Web Technologies (pp. 49-60). Springer Berlin Heidelberg.
- Barbut M., Monjardet B. (1970). “Ordre et Classification”, Algèbre et Combinatoire, Tomes I et II, Hachette, Paris.
- Birkhoff G. (1940). “Lattice theory”, American Mathematical Society Colloquium Publications, Vol.XXV, 1st edition, 1940 (3rd edition, 1967).
- Bonacich, P. (1987). “Power and centrality: A family of measures”. The American Journal of Sociology, 92(5):1170-1182.
- Boucher-Ryan, P. du, & Bridge, D. (2006). “Collaborative recommending using formal concept analysis”. Knowledge-Based Systems, 19(5), 309-315.
- Burt, R. S. (1992). “Structural Holes: The Social Structure of Competition”. Networks and organizations: Structure, form, and action. Harvard University Press, Massachusetts, USA 57-91.
- Carley, Kathleen M (2003), “Dynamic Network Analysis” in the Summary of the NRC workshop on Social Network Modeling and Analysis, Ron Breiger and Kathleen M. Carley (Eds.) (pp. 133-145). National Research Council.
- Carpineto, C. e Romano, G. (1993). “GALOIS: An order-theoretic approach to conceptual clustering”. In ICML (Vol. 90, pp. 33-40). Fondazione Ugo Bordoni Rome (Italy).
- Chung, K. K. S., Hossain, L., and Davis, J. (2005). “Exploring sociocentric and egocentric approaches for social network analysis”. In Proceedings of the International Conference on Knowledge Management in Asia Pacic, Wellington,New Zealand, November 27-29, pages 17.
- Costa, L., Jr., O. N. O., Travieso, G., Rodrigues, F. A., Boas, P. R. V., Antiqueira, L., Viana, M. P., e da Rocha, L. E. C. (2008). “Analyzing and modeling real-world

phenomena with complex networks: A survey of applications". Advances in Physics, 60(3), 329-412.

- Diestel, R. (2005). "Graph Theory". Graduate texts in mathematics, vol. 173. Springer-Verlag.
- DiMaggio, P. J. e Powell W. W. (1983). "The Iron Cage Revisited: Institutional isomorphism and collectivity rationality in organizational fields." American Sociological Review 48:147-60.
- Easley, D. e Kleinberg, J. (2010). "Networks, Crowds and Markets: Reasoning about a Highly Connected World". Cambridge of University Press, New York, USA
- Eschenfelder, D., Kollewe W., Skorsky, M., & Wille, R. (2000). Ein Erkundungssystem zum Baurecht: Methoden der Entwicklung eines TOSCANA-Systems. In G. Stumme, & R. Wille (Eds.), Begriffliche Wissensverarbeitung. Methoden und Anwendungen. Berlin: Springer, 254-272.
- Freeman, L.C. (1979). "Centrality in Social networks: Conceptual clarification". Social Networks, 1(3):215-239.
- Godin, R., Missaoui, R., Hassan, A. (1995). "Incremental concept formation algorithms based on Galois (concept) lattices" Appeared in Computational Intelligence (1995), 11(2), 246-267 Département d'Informatique, Université du Québec à Montréal.
- Hoppe, B. (2007). "Introduction to Network Mathematics". Boston University, http://webmathematics.net/
- Hoppe, B. (2009) in "Web Science" - http://webwhompers.com/course-overview/25.html
- Kossinets, G. and Watts, D. J. (2006). "Empirical analysis of an evolving social network". Science, 311(57):88-90.
- Krajca, P., Outrata, J., Vychodil, V. (2010). "Advances in algorithms based on CbO". Department of Computer Science, Palacky University, Czech Republic. In CLA (pp. 325-337).
- Krebs, V. (2000) in "Social Network Analysis , A Brief Introduction" - http://www.orgnet.com/sna.html

- Kuznetsov, S.O. (1993). “A fast algorithm for computing all intersections of objects in a finite semi-lattice”, Automat. Document. Math. Linguist. 27 (5) (1993) 11-21
- Kuznetsov, S.O. (1999), “Learning of simple conceptual graphs from positive and negative examples”, PKDD (1999) 384–391.
- Lorrain, F. and White, H. C. (1971). “Structural equivalence of individuals in social networks”. Journal of Mathematical Sociology, 1(1):49-80.
- Lotka, A. J. (1926) in “The Frequency Distribution of Scientific Productivity”. Journal of the Washington Academy of Science 16:317-323.
- Lucas, C. (2012), “Conceptual Clustering and Galois Concept Lattice”. PDMA – Doctoral Program in Applied Mathematics. Slides Presentation. University of Porto.
- March, J. G. and Olsen, J. P. (1989). “Rediscovering Institutions”. New York: The Free Press (pp 278-281).
- McPherson, J. M., and Smith-Lovin, L. (1987). “Homophily in voluntary organizations: Status distance and the composition of face-to-face groups.” American Sociological Review 52: 370-79.
- McSweeney, P. J. (2009). “Gephi Network Statistics”. Google Summer of Code 2009 Project Proposal.
- Mendonça, G., Machado, M., Dahis, R., Vasconcelos, R. (2009). “Detecção de Estruturas de Comunidades em Redes Complexas”, Disciplina de Inteligência Computacional.
- Merton, R. K. (1973). “The sociology of science: Theoretical and empirical investigations. The Normative Structure of Science” Pp. 267-78 in his The Sociology of Science: University of Chicago Press.
- Newman, M. E. J. (2003). “The Struture and Function of Complex Networks”. SIAM Review, 45(23):167-228.
- Newman, M. E. J. (2004). “Analysis of weighted networks”. Department of Physics and Center for the Study of Complex Systems University of Michigan, and Santa Fe Institute, Physical Review E 70.5 (2004): 056131.
- Newman, M. E. J. and Girvan, M. (2004). “Finding and evaluating community structure in networks”. Physical Review E, 69(2): 026113.

- Oliveira, M. and Gama, J. (2011). “An Overview of Graph Theory, Network Theory and Social Network Analysis” in Faculdade de Economia do Porto.
- Oliveira, M. and Gama, J. (2012). “Social Network Analysis”. Slides de Extracção de Conhecimentos de Dados II, FEP.
- Opsahl, T., Agneessens, F., & Skvoretz, (2010). “Node Centrality in weighted networks: Generalizing degree and shortest paths”. Social Networks, 32(3):245-251.
- Outrata, J., Vychodil, V. (2012). “Fast algorithm for computing fixpoints of Galois connections induced by object-attribute relational data”. Dept. Computer Science, Palacky University, Czech Republic
- Pernelle, N., Rousset, M. C., Soldano, H., & Ventos, V. (2002). “Zoom: a nested Galois lattices-based system for conceptual clustering”. Journal of Experimental & Theoretical Artificial Intelligence, 14(2-3), 157-187.
- Powell, W. W. (1990). “Neither Market Nor Hierarchy: Network Forms of Organization.” Research in Organizational Behavior 12:295-336. JAY Press Inc..
- Powell, W. W., Koput, K. and Smith-Doerr, L. (1996). “Interorganizational Collaboration and the Locus of Innovation: Networks of Learning in Biotechnology”. Administrative Science Quarterly 41(1): 116-45.
- Powell, W. W., White, D. R., Koput, K. W., & Owen-Smith, J. (2005). “Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences”. American journal of sociology, 110(4), 1132-1205.
- Price, D. J. d. S. (1965) in “Statistical studies of networks of scientific papers”. In Statistical Association Methods for Mechanized Documentation: Symposium Proceedings (Vol. 269, p. 187). US Government Printing Office., Science 149:510-515
- Priss, U. (2006) “Formal Concept Analysis in Information Science” in Cronin, Blaise (Ed.), Annual Review of Information Science and Technology, ARIST, 40(1), 521-543.
- QFINANCE – “The Ultimate Resource”, © 2009 Bloomsbury Information Ltd http://www.qfinance.com/dictionary/cluster-analysis.

- Rapoport, A. (1953). “Spread of information through a population with socio-structural bias: Assumption of transitivity”. Bulletin of Mathematical Biophysics, 15(4):523-533.
- Stumme, G. Taouil, R., Bastile, Y., Pasquier, N., Lakhal, L. (2002). “Computing iceberg concept lattices with TITANIC”, Data & Knowledge Engineering 42: 189-222.
- Stumme, G., Wille, R., Wille, U. (1998). “Conceptual Knowledge Discovery in Databases Using Formal Concept Analysis Methods” Technische Universität Darmstadt, Fachbereich Mathematik, D-64289 Darmstadt, Germany, IBM Research Division, Zurich Research Laboratory. (pp. 450-458). Springer Berlin Heidelberg.
- União Europeia, (1995-2012) ©. Comissão Europeia – “Situação agrícola e perspectivas nos países da Europa Central e Oriental; Hungria” in http://ec.europa.eu/agriculture/publi/peco/hungary/summary/sum_pt.htm
- Valtchev, P., Missaoui, R., Godin, R., & Meridji, M. (2002). “Generating frequent itemsets incrementally: two novel approaches based on Galois lattice theory”. Journal of Experimental & Theoretical Artificial Intelligence, 14(2-3), 115-142.
- Wasserman, S. and Faust, K. (1994). “Social Network Analysis: Methods and Applications”. Cambridge University Press, Cambridge
- Watts, D. (1999) in “Small Worlds”. Princeton: Princeton University Press (Vol. 8).
- White, H. (1981). “Where do markets come from?” American Journal of Sociology Vol. 87, No. 3 (Nov., 1981), pp. 517-547.
- Wille, R. (1982). “Restructuring lattice theory: an approach based on hierarchies of concepts”. In I. Rival (Ed.), Ordered sets. Reidel, Dordrecht-Boston,445-470.
- Yang, B., Di, J., Liu, J., & Liu, D. (2012). “Hierarchical community detection with applications to real-world network analysis”. Data & Knowledge Engineering.
- Yevtushenko, S. (2004). “Computing and visualizing concept lattices” (Doctoral dissertation, TU Darmstadt).