FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

FEUP

Identificação de comunidades e exploração visual em redes sociais: rede do comércio internacional europeu

Luis Matias Nunes de Pina Moura

Mestrado Integrado em Engenharia Informática e Computação

Orientador: Francisco José de Oliveira Restivo (Doutor)

Julho de 2012

Identificação de comunidades e exploração visual em redes sociais: rede do comércio internacional europeu

Luis Matias Nunes de Pina Moura

Mestrado Integrado em Engenharia Informática e Computação

Aprovado em provas públicas pelo Júri:

Presidente: Doutor Rui Filipe Maranhão de Abreu, Prof. Catedrático da FEUP

Vogal Externo: Doutor Luís Manuel Borges Gouveia, Prof. Associado da U. Lusófona

Orientador: Doutor Francisco José de Oliveira Restivo, Prof. Associado da FEUP

---

18 de Julho de 2012

Resumo

As redes sociais são constituídas por vértices (atores, ou nós), e por arestas (relações ou ligações), cada elemento tendo eventualmente atributos a si associados. A visualização de uma rede e o cálculo de algumas métricas apropriadas permite muitas vezes revelar factos sobre o seu funcionamento que doutro modo ficariam escondidos. No entanto, a complexidade das redes sociais modeladas a partir de redes reais com muitos nós e elevada densidade de ligações faz com que a sua visualização se torne confusa e difícil de interpretar.

Nesta dissertação, é ilustrada a utilização de um algoritmo de identificação de comunidades (*k-clique percolation*) em conjugação com uma plataforma de visualização de redes interativa *web-based* para a exploração de uma rede social de elevada densidade. Como caso de estudo, foi escolhida a rede do comércio entre países europeus, desde o final dos anos 80 até à atualidade.

Abstract

Social networks are made of nodes (social actors, institutions,…) and edges (representing relationships, dependencies, …). Each of those elements may have attributes of their own. A visual representation of such a network and the calculation of appropriate metrics often helps in uncovering facts and properties of the network that otherwise would remain hidden. However, making visual representations of real world social networks, which are often complex and dense, poses its own challenges.

In this dissertation, it is shown how community identification techniques (k-clique percolation) can be used along with an interactive dynamic network visualization application to explore a high density social network. The approached case study is the international trade network between European nations, since the late 80's until the current time.

Agradecimentos

Agradeço ao Prof. Francisco Restivo e à minha família e colegas todo o apoio prestado.

Luis Moura

Índice

Introdução **1**
1.1 Contexto/Enquadramento 2
1.2 Motivação e Objetivos 2
Elementos que motivaram o desenvolvimento desta dissertação: 2
Objetivos: 2
Metodologia: 3
1.3 Estrutura da Dissertação 3

Revisão Bibliográfica **5**
2.1 Introdução 5
2.2 Análise de redes complexas 5
2.2.1 Fenómeno *Small World* 7
2.2.2 Fenómenos de agregação e transitividade (*clustering*) 7
2.2.3 Distribuição dos graus dos nós *Scale Free* 7
2.3 Análise de redes sociais (SNA) 7
2.3.1 Representação de redes sociais 8
2.3.2 Dados medidos sobre redes sociais 9
2.3.2.1 Distribuição de graus de ligação 9
2.3.2.2 Cliques, coesividade e *clustering* 9
2.3.2.3 Análise espectral 10
2.3.2.4 Medidas locais 10
2.3.3 Análise de redes sociais na economia 11
2.4 Representação visual de informação 12
2.4.1 Eficiência 13
2.4.2 Elementos gráficos multifuncionais 14
2.4.3 Densidade de Dados 16
2.4.4 Considerações Estéticas 18
2.4.5 Visualização de redes complexas 18
2.5 Identificação de comunidades 19
2.5.1 Redes com comunidades sobrepostas, *k-clique percolation* 19
2.6 Ferramentas para análise de redes sociais 21

Identificação de Comunidades e Visualização de Rede Sociais 23
3.1 Introdução ao problema 23
3.2 Abordagem e Resultados 24
3.2.1 Recolha e tratamento de dados 24
3.2.2 Identificação de comunidades 25
3.2.3 Visualização 27
3.2.4 Layout 28
3.2.5 Filtragem de informação 28
3.2.6 Interação 29
3.3 Resumo e Conclusões sobre resultados obtidos 30

Implementação 33
4.1 Arquitetura da aplicação 33
4.1.1 Importação de dados 34
4.1.2 Geração de grafo de rede 34
4.1.3 Comunicação com a aplicação cliente 34
4.1.4 Interface e visualização da rede: *paper.js* 35
4.2 Utilização de ferramentas de análise de redes sociais 35
4.2.1 Biblioteca networkx de python 36
4.2.2 CFinder 36

Conclusões e Trabalho Futuro 37
5.1 Satisfação dos Objetivos 37
5.2 Trabalho Futuro 37

Referências 39

Lista de Figuras

Figura 1: Ars Electronica - Collaboration Structure of Cultural Projects (sup. esq.), ITP Student List Conversations (sup. dir.), Internet Map (inf.). 6
Figura 2 : grafo (esq.) e matriz de adjacência (dir.) de uma mesma rede 8
Figura 3: cliques nas famílias de Aterro, Costa Rica - (Freeman 2002) 10
Figura 4 : E. J. Marey - horário dos comboios em França, 1880 (grelha atenuada) 13
Figura 5 : Leonard P. Ayres, *The War with Germany* (Washingyon, D.C., 1919) , p. 102. 14
Figura 6 : Gráfico limpo (esq.) e com lixo visual (dir) 15
Figura 7: Axel, Kolman et Al. , Origin of a Signal Intensity Loss Artifact in Fat-Saturation MR Imaging 16
Figura 8 : Jacques Bertin, *Semiologie Graphique* (Paris, 2ª edição, 1973), p. 152. 17
Figura 9 Partição de rede gerada no CFinder, com *layout* ajustado manualmente 26
Figura 10 Comunidades no mapa da europa 27
Figura 11 antes (esq.) e depois (dir.) de filtragem da opacidade das arestas 29
Figura 12 *Screenshot* da aplicação 30
Figura 13 Estrutura da aplicação 33

Capítulo 1

Introdução

Este trabalho foi inspirado pelo interesse do autor no tema da análise de redes sociais, como conceito aplicável à observação de qualquer tipo de estrutura social, desde relações entre pessoas, empresas, países, e às suas potencialidades, nomeadamente em perceber o funcionamento dessas redes de relações, e de que forma a sua estrutura está organizada. A noção de que uma quantidade enorme de estruturas encontradas na natureza se organiza de forma semelhante, incluindo a estrutura dos nossos relacionamentos, quando superficialmente parecem ter uma organização arbitrária, e os mecanismos que levam a esse tipo de organização são questões que têm suscitado um grande interesse da parte da comunidade científica (e não só) nos últimos anos, especialmente com o aparecimento de *sites* de redes sociais na internet (e do estudo sobre a topologia da WWW, ela própria podendo ser modelada como rede social), que trouxeram esse conceito para o vocabulário do dia-a-dia em grande parte do mundo.

Neste documento, é feita uma breve revisão sobre a investigação científica feita até à data nos temas da análise de redes sociais e visualização de informação, e são apontadas algumas obras ou artigos de referência sobre o tema.

Esta dissertação foi feita acompanhando o desenvolvimento de uma aplicação para visualização de redes sociais, orientada para a análise da rede de comércio entre países europeus, que é usada como caso de estudo. Dadas as limitações de tempo e o formato específico de uma tese de mestrado, seria impossível abranger de forma completa o tema, pelo que se optou por focar em dois pontos de especial interesse na análise de redes sociais: a identificação de comunidades, e a visualização da rede, procurando-se pesquisar um método útil de ligar ambas as questões, aproveitando como plataforma/ferramenta a aplicação em desenvolvimento. Como será exposto na secção de revisão bibliográfica, tanto o tema de identificação de comunidades como o tema de visualização de redes sociais têm uma grande diversidade de abordagens e aplicações, e estão em grande desenvolvimento neste momento. Procura-se evidenciar, dentro dessa variedade de abordagens, uma que seja adequada ao caso de

estudo, e que abra caminho para a identificação de pontos de interesse para investigação e desenvolvimento da aplicação no futuro.

1.1 Contexto/Enquadramento

Esta dissertação é uma proposta autónoma e não faz parte de nenhum projeto existente previamente. É inspirada pelo desenvolvimento de uma aplicação web para visualização de redes sociais na economia. Os temas centrais do projeto (análise e visualização de redes sociais) têm sido explorados com grande interesse pela comunidade científica. Existem muitos temas científicos passíveis de estudo no contexto de um projeto de aplicação informática para análise de redes económicas, como, por exemplo as áreas de análise e visualização de redes, o tratamento de dados variantes no tempo, a forma como os dados são guardados, o estudo das tecnologias usadas ou o estudo da usabilidade de aplicações visuais interativas no contexto das redes sociais. O foco desta tese é a utilização de métodos de identificação de comunidades e a sua conjugação com técnicas de visualização interativa de redes sociais para a exploração da rede, usando como caso de estudo a rede de comércio europeu entre nações. As soluções tecnológicas usadas são também um ponto de interesse.

1.2 Motivação e Objetivos

Elementos que motivaram o desenvolvimento desta dissertação:

- Forma de dirigir o desenvolvimento de uma aplicação *web-based* para a análise da rede de comércio internacional;
- O autor tem um interesse pessoal pela análise de redes sociais, tema que já abordou em alguns trabalhos curriculares no MIEIC, e sobre o qual se tem informado também fora do contexto de trabalho académico;
- Na atualidade, a compreensão do funcionamento da economia como rede é um tema de destaque. As técnicas de análise de redes sociais adequam-se à análise de redes económicas e, sendo uma abordagem orientada pela topologia da rede e seus efeitos sobre os seus elementos, é bastante diferente das abordagens usadas tradicionalmente na área da economia, tendo por isso o potencial de fornecer uma perspetival diferente e potencialmente útil sobre as mesmas;
- Com a tecnologia web atual, é possível criar aplicações complexas cuja operação é baseada no *web browser*. A exploração desse tipo de tecnologia é um ponto de interesse.

Objetivos:

- Criar um sistema de visualização interativa que aproveite elementos de análise de redes sociais;
- Rever métodos de identificação de comunidades e identificar o mais adequado para este tipo de rede;
- Aproveitar a informação resultante do processo de identificação de comunidades para enriquecer a visualização da rede
- Criar um ponto de partida para o desenvolvimento prático de uma solução *web-based* para análise da rede económica internacional.

Metodologia:

Sendo um trabalho individual, a metodologia adotada para o seu desenvolvimento foi orientada a tarefas organizadas por prioridade. O trabalho foi desenvolvido seguindo uma fase inicial de investigação sobre os temas abordados, um período de adaptação às tecnologias usadas, uma fase de desenvolvimento da aplicação informática, e uma fase de análise de resultados e escrita da dissertação.

1.3 Estrutura da Dissertação

Para além da introdução, esta dissertação contém mais 4 capítulos. No capítulo 0 (Bibliografia), é descrito o estado da arte e são apresentados trabalhos relacionados. No capítulo 3 (Identificação de Comunidade e Visualização de Redes Sociais), é exposto em maior detalhe o problema e as soluções adotadas no caso de estudo, e alguns resultados obtidos. No capítulo 4 (Implementação), é apresentada com maior detalhe a implementação das soluções informáticas usadas. No capítulo 5 (Conclusões e Trabalho Futuro), são apresentadas as conclusões da dissertação, e ideias para um possível desenvolvimento futuro de trabalho sobre o mesmo tema.

Capítulo 2

Revisão Bibliográfica

2.1 Introdução

O tema desta dissertação enquadra-se nas áreas da análise de redes complexas, análise de redes sociais, identificação de comunidades em redes e visualização de informação. De seguida é feita, para cada uma dessas áreas, uma breve revisão bibliográfica, de forma a enquadrar a dissertação no contexto atual, e para que sejam evidenciados alguns trabalhos já feitos sobre o mesmo tema.

2.2 Análise de redes complexas

Uma grande variedade de sistemas de grande interesse e importância podem ser descritos como estruturas complexas sob a forma de rede (entre outros: células de seres vivos, a internet (WWW), rede de energia elétrica, relações sociais entre as pessoas). O desejo de compreender a topologia e dinâmica deste tipo de sistemas (em que a distância física é geralmente irrelevante, e cuja complexidade e ambiguidade das ligações entre elementos torna difícil a modelação e compreensão usando uma perspetiva reducionista) levou a comunidade de investigadores de física estatística, especialmente desde o início do século XXI, a interessar-se em desenvolver modelos estatísticos que lhes permitam analisar este tipo de problema, tradicionalmente abordado na área de teoria dos grafos.

O aumento do interesse nesta área de estudo está fortemente relacionado com um conjunto de fatores, por exemplo:

- A adoção de ferramentas informáticas na generalidade das áreas de conhecimento faz com que existam atualmente bases de dados extensas com informação sobre a topologia de várias redes complexas reais.

- A existência da WWW e o aumento no trabalho interdisciplinar entre a comunidade científica facilita o acesso a essas bases de dados.
- O poder de processamento dos computadores atuam permite o tratamento de dados para estudo de redes com uma enorme quantidade de nós e arestas, algo que há poucas décadas atrás não era possível.

Sendo que o conhecimento humano já é bastante avançado em termos de análise reducionista, existe cada vez mais interesse em perceber a topologia da interação entre os componentes dos sistemas, de forma a perceber o funcionamento destes como um todo. Esse interesse é também alimentado pelo aparecimento de estudos empíricos sobre um conjunto diverso de redes complexas que evidenciam princípios de organização estrutural que são transversais a várias redes de natureza muito diferente. Alguns exemplos na Figura 1, retirados

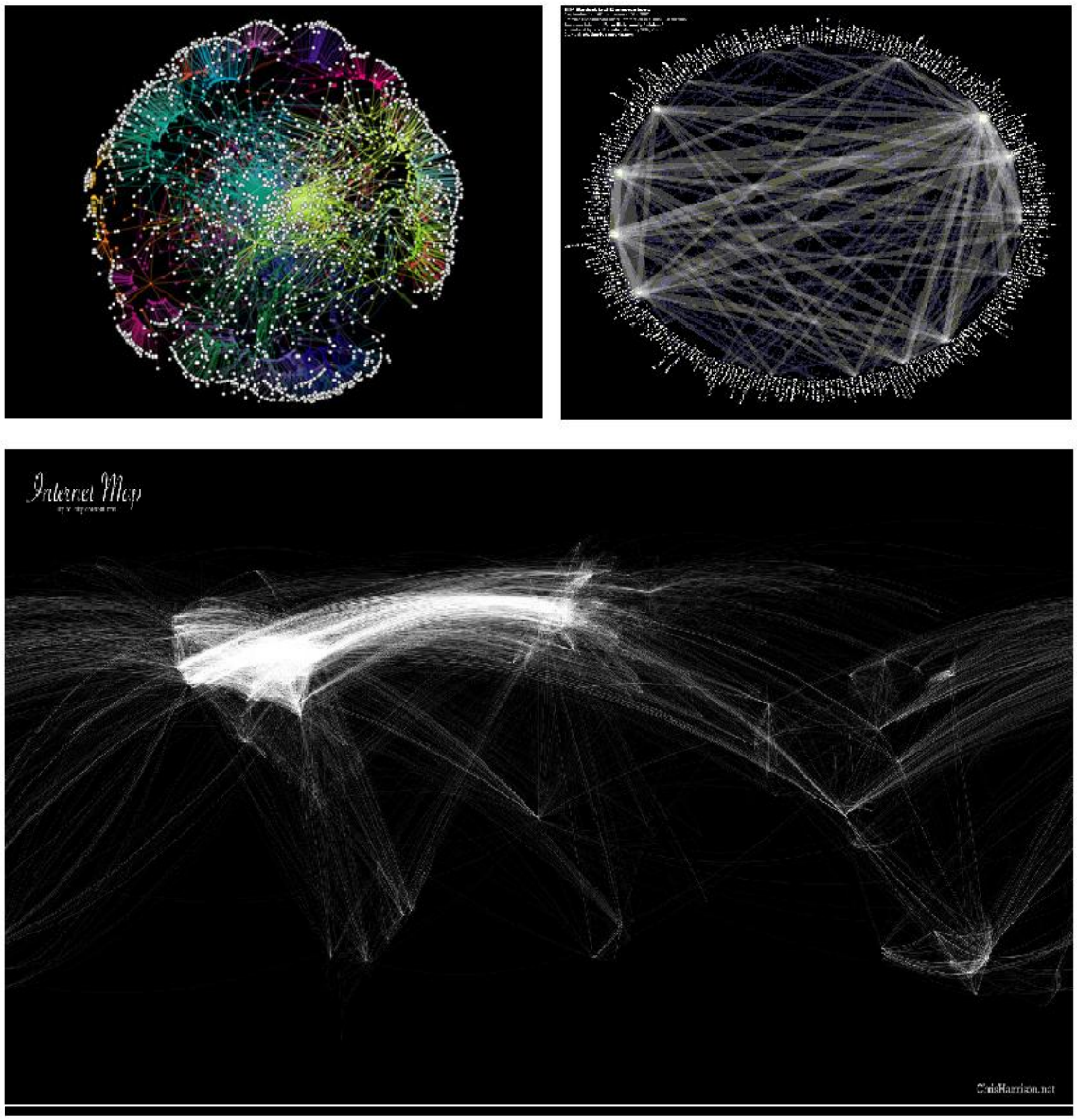

Figura 1: Ars Electronica - Collaboration Structure of Cultural Projects (sup. esq.), ITP Student List Conversations (sup. dir.), Internet Map (inf.).

de [1], ilustram a complexidade de redes pertencentes a domínios diferentes.

No seu artigo [2], Barabási apresenta um conjunto de modelos e ferramentas analíticas para análise da topografia e dinâmica de sistemas complexos que foram desenvolvidas no sentido de responder a esta questão. Estes refletem alguns conceitos centrais no pensamento atual sobre redes complexas, e são apresentados de seguida.

2.2.1 Fenómeno *Small World*

Nas redes complexas observadas empiricamente, é frequente observar-se que o tamanho médio do caminho entre dois nós cresce logaritmicamente com o tamanho do sistema [3]. Isto significa que a partir de qualquer nó da rede se consegue aceder a qualquer outro nó efetuando poucos saltos intermédios. As redes em que este fenómeno é observado são caracterizadas como redes *small world*.

2.2.2 Fenómenos de agregação e transitividade (*clustering*)

As redes complexas costumam exibir fenómenos de agregação dos nós, isto é, os vizinhos de um determinado nó têm grande probabilidade de estarem interligados, organizando-se como comunidades dentro da rede. A identificação dessas comunidades é apresentada na secção 2.5 Identificação de Comunidades.

2.2.3 Distribuição dos graus dos nós *Scale Free*.

No estudo de redes, o grau de um nó é o número de ligações que este possui. A distribuição de graus em redes complexas costuma diferir bastante da observada numa rede gerada aleatoriamente (distribuição de Poisson). Um tipo de distribuição encontrada em casos paradigmáticos do estudo de redes complexas é a chamada *Power Law Distribution*. Este tipo de distribuição é interessante pois é invariante face à escala. Dada uma relação $f(x) = ax^k$, escalar a constante x num factor constante causa apenas um escalamento proporcional da função em si. As redes que exibem esta propriedade são designadas como *Scale Free*.

Pode-se encontrar em [4] uma definição formal mais completa, assim como a ligação deste atributo à robustez e dinâmica da rede.

2.3 Análise de redes sociais (SNA)

Uma rede social é uma estrutura composta por entidades sociais (nós), ligadas por arestas que representam dependências entre elas (relações de amizade, intercâmbio financeiro,

parentesco, etc..). A sociedade humana pode ser vista como uma enorme rede social, em que cada nó é um indivíduo, organização ou nação, e as ligações são interações sociais.

A análise de redes sociais caracteriza-se pela sua perspetival de análise de redes baseada em conceitos relacionais, tentando perceber qual o impacto da estrutura de relações sobre características dos nós, isto é, em vez de se focar na análise e relação de atributos autónomos dos indivíduos (como a idade, sexo ou nacionalidade), tenta descobrir de que forma a estrutura das relações desses indivíduos (a sua posição na rede) influencia as suas características individuais, ou foca-se em propriedades estruturais do sistema no seu conjunto.

Esta perspetiva implica a utilização de métricas bastante distintas das usadas numa análise baseada em amostragem e relacionamento de dados estatísticos sobre os indivíduos [5].

Sendo que as redes sociais de grande dimensão são redes complexas, a procura de modelos para as caracterizar especificamente é um tema de estudo atual, podendo-se referir [6, 7], onde são analisadas diferenças das redes sociais face aos modelos tradicionalmente usados para caracterizar redes complexas (entre outros: *scale-free*, *small world*, aleatórias).

2.3.1 Representação de redes sociais

A natureza de uma rede, definida por nós e respetivas ligações, leva naturalmente a que estas sejam usualmente representadas visualmente sob a forma de grafos. As redes sociais são também representáveis sob a forma de matriz de adjacência, tendo nos eixos os nós da rede, e números ou símbolos nas células representando as ligações entre nós (Figura 2).

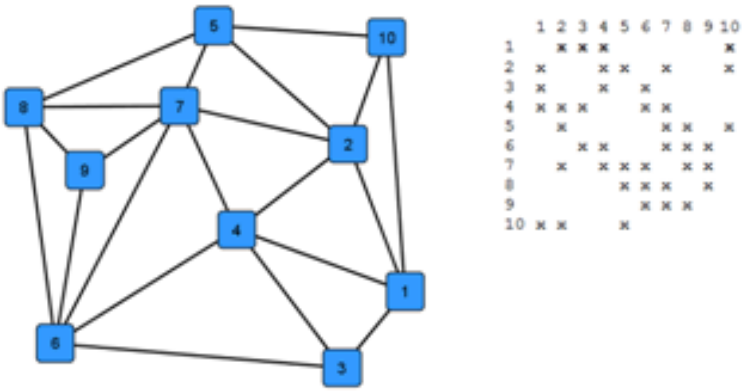

Figura 2: grafo (esq.) e matriz de adjacência (dir.) de uma mesma

A forma da representação de redes sociais foi evoluindo com os progressos tecnológicos e científicos, passando de desenhos manuais de redes de pequena dimensão para representações digitais interativas de redes de elevada dimensão e complexidade. A disposição dos nós também

evoluiu, de representações *ad hoc* para a utilização de algoritmos sofisticados de desenho de grafos, resultando em visualizações de grande riqueza de informação e impacto estético.

Esta evolução é descrita em [8], sendo que é mostrado que a visualização de redes sociais é um aspeto central da sua análise.

2.3.2 Dados medidos sobre redes sociais

Apesar de ter as suas origens na sociologia e antropologia do início do século XX, a perspetiva de análise de redes sociais tem vindo a ganhar nos últimos anos uma grande relevância em áreas como a medicina, a física ou a economia, devido à sua aplicação e relevância na área da análise de redes complexas.

Em análise de redes sociais (SNA), procura-se normalmente identificar dois tipos de padrões de ligações: grupos sociais (conjuntos de atores que estão fortemente interligados) e posições sociais (conjuntos de atores que estão ligados ao sistema de forma semelhante). Nesta secção são apresentadas algumas medidas tipicamente usadas como forma de evidenciar esses padrões, ilustrando a perspetiva relacional da SNA. A rede é usualmente abstraída num grafo. [5].

São de seguida apresentadas algumas métricas frequentemente usadas na análise de redes sociais.

2.3.2.1 Distribuição de graus de ligação

A distribuição probabilística dos graus dos nós é frequentemente usada para ajudar a caracterizar a estrutura de uma rede social, tal como é referido na secção 2.2.3 (distribuição das ligações em redes complexas).

2.3.2.2 Cliques, coesividade e *clustering*

Um aspeto importante das redes sociais é a medida e forma em que os seus nós estão interligados. Existe uma variedade de conceitos que ajudam a observar o nível de coesão da rede, e a forma como esta está organizada:

O c*lustering* ou coesividade é a propriedade de dois nós ligados a um mesmo terceiro nó terem uma probabilidade elevada de estarem interligados. O estudo desta medida permite indiciar a existência de comunidades dentro da rede (*clusters*).

Um motivo é um padrão de ligações que é identificado em número superior ao que ocorreria numa versão aleatória de uma mesma rede (com o mesmo número de ligações, nós e distribuição de graus).

Estruturas comunitárias são frequentemente definidas como sub-grafos coesivos, isto é, cujos nós estão fortemente interligados. Este nível de coesão pode ser quantificado de diferentes formas, mas comunidades podem genericamente ser descritas como agrupamentos de nós densamente ligados entre si, e fracamente ligados ao resto dos elementos da rede. A definição mais coesa de uma comunidade é uma *clique*, um sub-grafo cujos nós são completamente interligados, como é exemplificado na Figura 3, onde é representada a estrutura social das famílias instaladas no vale Turrialba na Costa Rica, onde que os nós são dispostos de forma a evidenciar as *cliques* encontradas nessa estrutura.

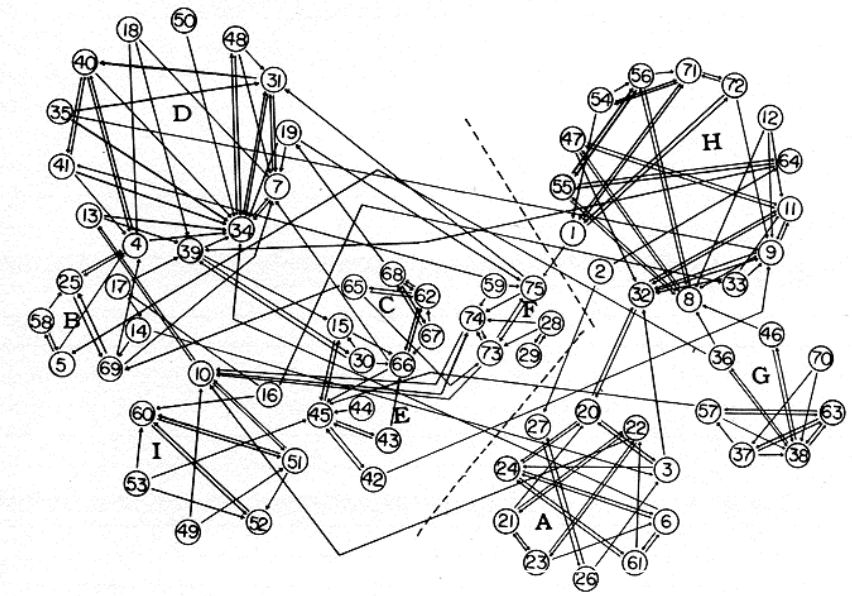

Figura 3: cliques nas famílias de Aterro, Costa Rica - (Freeman 2002)

2.3.2.3 Análise espectral

O espectro de um grafo é o conjunto de valores próprios (*eigenvalues*) da sua matriz de adjacência. Os *eigenvalues* e *eigenvectors* são usados como forma de caracterizar tanto modelos como redes reais, e para descobrir a presença de subgrupos coesivos e outras características locais das redes, analisando padrões de variância da matriz [9].

2.3.2.4 Medidas locais

Para além de medidas sobre a rede no seu todo, existe um grande número de medidas de centralidade aplicadas a cada nó, que fornecem informação que permite comparar nós e ver de que forma um determinado nó está inserido na estrutura da rede [10]:

- Grau: número de ligações de um determinado nó. No caso de uma rede direcionada, refere-se as medidas de *out-degree* e *in-degree* para o número de arestas que sai e entra no nó.
- Proximidade: facilidade de um dado nó chegar a outros nós.

- *Betweenness*: importância de um nó em termos de facilitar a interligação de outros nós.
- Prestígio e medidas relacionadas com *eigenvectors*: a forma como a vizinhança de um nó afeta o seu prestígio na rede. Existem vários métodos analíticos baseados em *eigenvectors* que evidenciam formas de analisar o prestígio de um nó. Estas medidas de centralidade têm em conta, para além da centralidade do nó (em termos de grau), a centralidade dos seus vizinhos, atribuindo um valor mais elevado conforme a centralidade da sua vizinhança. Isto faz com que estas medidas de centralidade refletem a estrutura da rede na avaliação do prestígio do nó. [9].

Uma descrição bastante completa de métricas para estudo da estrutura e dinâmica destas redes pode ser consultada em [11]. Para além da sua natureza descritiva, este tipo de medidas permite tirar conclusões sobre a robustez da rede face à remoção de nós, ou a disseminação de fenómenos de contágio através da rede, dando-lhe portanto grande relevância prática.

2.3.3 Análise de redes sociais na economia

Muito do interesse sobre redes sociais está ligado ao facto de a sua estrutura ser determinante para a forma como as sociedades e economias funcionam. A economia e, consequentemente, o comércio são uma parte fundamental da organização social humana. No entanto, a teoria económica tradicional não explorou muito esta área até aos anos 90. Com o recente desenvolvimento da área da análise de redes complexas, e a ubiquidade desse tipo de rede, gerou-se um interesse em redor desta questão. Com a observação de que a posição de uma entidade numa rede económica tem um impacto significativo sobre os proveitos que a entidade dela pode tirar, depreendeu-se que essas entidades tentarão moldar as suas relações dentro da rede, o que levou ao desenvolvimento de teorias de modelação e desenvolvimento estratégico de redes económicas. [12] e [13].

Com o já amplamente reconhecido fenómeno da globalização [14] , o comércio é uma forma de interação entre países cujo impacto vai muito para além do da troca de bens, podendo por exemplo ser um canal de propagação de crises, já que perturbações económicas originadas num país central ao comércio se podem repercutir por toda a rede. É também interessante ver o impacto das políticas de liberalização sobre a estrutura da rede.

A rede (não direcionada) do comércio internacional entre nações enquadra-se no contexto da análise de redes complexas, visto que esta exibe propriedades características destas, nomeadamente, uma distribuição de graus livre de escala, é uma rede *small-world*, tem um elevado grau de *clustering* e uma correlação (neste caso competitiva) dos graus entre vértices. [15]

A rede direcionada enquadra-se no modelo de evolução *fitness model*, em que a forma como as ligações entre nós evolui está relacionada com uma variável de capacidade de encaixe

(neste caso o produto interno bruto), sendo que os nós com mais elevada capacidade de encaixe atraem mais ligações, à custa de nós com menos poder de encaixe. [16]

Existe já alguma investigação dedicada à análise da rede de comércio internacional (apesar de não se focarem no aspeto da visualização das redes), facto de grande importância, por dois motivos. Em primeiro lugar, a sua existência mostra o interesse e validade deste tipo de análise aplicado à rede de comércio entre nações. Em segundo lugar, esses trabalhos contêm informação relevante sobre a metodologia e estratégias utilizadas no processamento analítico dos dados. Alguns exemplos:

- No artigo [17] é observada uma reciprocidade elevada nas relações bilaterais de importação e exportação, o que permite uma análise não direcionada das relações sem que se percam características topológicas importantes. Esta simplificação reduz significativamente a complexidade dos cálculos sobre a rede, em comparação com uma análise em que as ligações são direcionadas. Pode-se ver outro exemplo no artigo[18] em que são usados apenas os dados de importação pois estes foram considerados mais fiáveis em relação aos de exportação.
- A análise da rede considerando tanto as ligações binárias como as ligações pesadas fornece informação que se complementa
- A distinção entre a rede da economia real e financeira é relevante pois estas têm propriedades estruturais distintas, sendo a rede da economia real bastante mais ligada que a financeira.

Este tipo de considerações permite procurar informação mais útil e orientar a análise dos dados de uma forma mais eficiente do que usando uma estratégia puramente exploratória.

Fica aqui a referência para estes e outros artigos de interesse encontrados sobre o tema da rede de comércio internacional:

[3, 10, 12, 15, 17-32].

2.4 Representação visual de informação

Uma das maiores vantagens da representação visual de dados (gráficos) é a elevada quantidade de informação que pode ser rapidamente interpretada caso seja bem apresentada. A informação importante de milhões de medidas pode ser obtida em poucos instantes. Algumas características de uma boa representação[33]:

- Permitir a perceção imediata de padrões estruturais e de propriedades emergentes não antecipadas.
- Frequentemente, consegue-se identificar problemas na qualidade dos dados refletidos como irregularidades na visualização.

- Permitir a análise dos dados tanto em grande como em pequena escala.

Visualizações de qualidade (com informação transmitida de forma rigorosa e intuitiva) são portanto uma ferramenta poderosa para raciocinar sobre informação quantitativa.

Edward Tufte expõe um conjunto de princípios para a criação de visualizações de qualidade (Tufte 1990):

2.4.1 Eficiência

Pode-se referir como considerações ligadas à eficiência a remoção de redundâncias e de lixo visual.

- Remoção de redundâncias

Gráficos de dados devem chamar a atenção para o sentido e substância dos dados, e nada mais. Deve-se portanto tentar maximizar no gráfico a proporção da tinta usada para transmitir informação sobre os dados sem redundâncias. Uma forma de implementar este princípio é através do uso de um processo de edição com o objetivo de eliminar toda a informação duplicada.

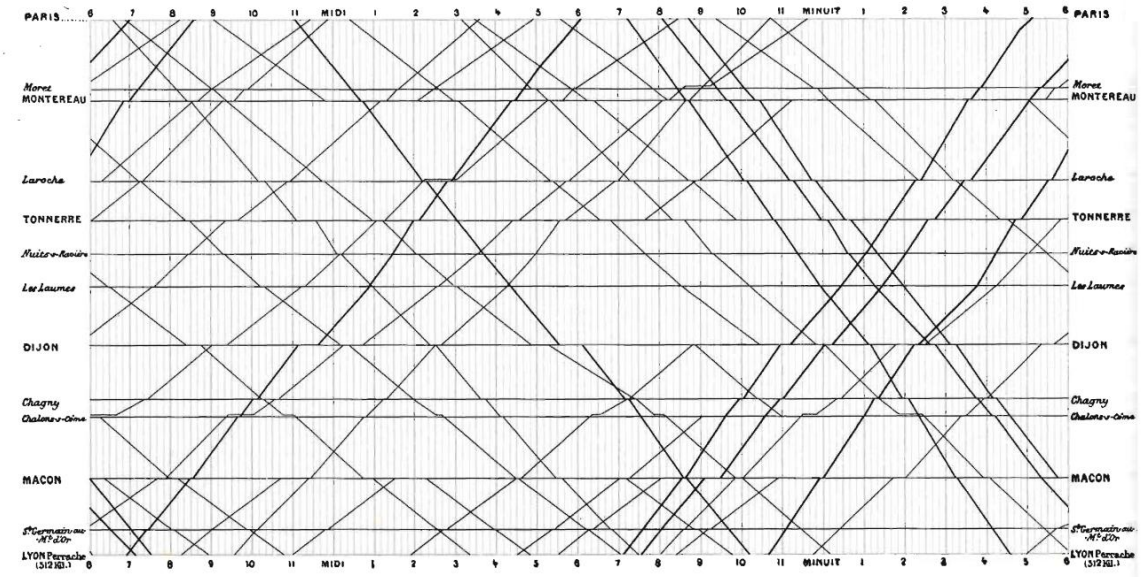

Figura 4 : E. J. Marey - horário dos comboios em França, 1880 (grelha atenuada)

Existem no entanto situações em que a redundância pode ser útil, nomeadamente para facilitar a análise de informação cíclica ou facilitar a comparação entre várias partes da imagem. Um exemplo é um horário de comboio, em que a redundância de informação faz com que seja possível identificar todos os horários possíveis (Figura 4).

- Remoção de lixo visual

É muito frequente encontrar, em todo o tipo de publicações, gráficos contendo elementos de decoração que não transmitem nenhuma informação, com o intuito de fazer o gráfico parecer mais científico e preciso, ou para torná-lo mais interessante artisticamente. Independentemente desse intuito, esse tipo de decoração que não transmite informação (e frequentemente dificulta ou distorce a leitura desta) é lixo visual, e deve ser removido.

"Graphical decoration, which prospers in technical publications as well as in commercial and media graphics, comes cheaper than the hard work required to produce intriguing numbers and secure evidence" [34]

Alguns exemplos desse tipo de decoração obstrutiva são a utilização de padrões de linhas Moiré (que provocam interferência na visão da imagem, e ilusões óticas), grelhas e ornamentações artísticas. Figura 4, pode-se ver uma boa utilização de uma grelha, em que esta foi atenuada de forma a não interferir com a leitura da informação. Para além de servir de lógica de edição, estes princípios de eficiência devem ser integrados no processo de conceção dos gráficos.

2.4.2 Elementos gráficos multifuncionais

Elementos visuais que preenchem convencionalmente uma função de design podem ser usados como forma de transmitir informação sobre os dados, ou até mostrar vários tipos de informação em simultâneo.

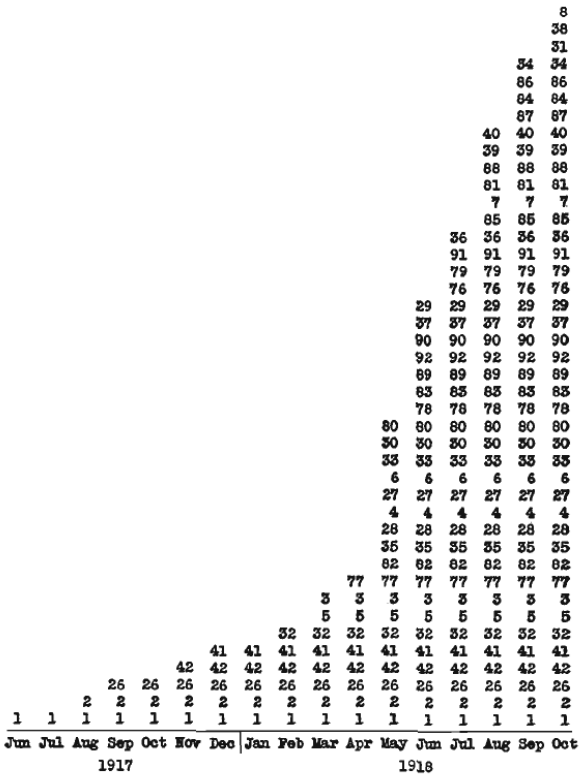

Figura 5 : Leonard P. Ayres, *The War with Germany* (Washingyon, D.C., 1919) , p. 102.

Esse tipo de elementos, se trabalhados com cuidado e subtileza, permitem transmitir informação complexa e multivariada. De seguida, são ilustrados alguns princípios sob a forma de exemplos.

No gráfico apresentado na Figura 5,que representa a quantidade de divisões militares americanas presentes em França na Primeira Guerra Mundial no período entre Junho de 1917 e Outubro de 1918, as barras são desenhadas usando a designação de cada divisão. Isto faz com que cada barra transmita três tipos de informação em simultâneo [34]:

- O número de divisões presentes no território em cada mês
- A designação das divisões para cada mês
- A duração da presença de cada divisão militar

No caso daFigura 6., a legenda da escala é transformada de forma a transmitir a amplitude de valores, assim como os valores de cada ponto. Foram também eliminadas as linhas de escala horizontal. Compare-se com a versão convencional (template do Excell) do gráfico. Ao invés de auxiliar à compreensão dos dados, as linhas e escala completa tornam-se distrações visuais que acrescem pouco à informação transmitida pelo gráfico. No exemplo "limpo", a quantidade de informação transmitida é superior, sendo usada menos tinta para a transmitir, o que facilita a leitura, dando um aspeto mais limpo ao gráfico:

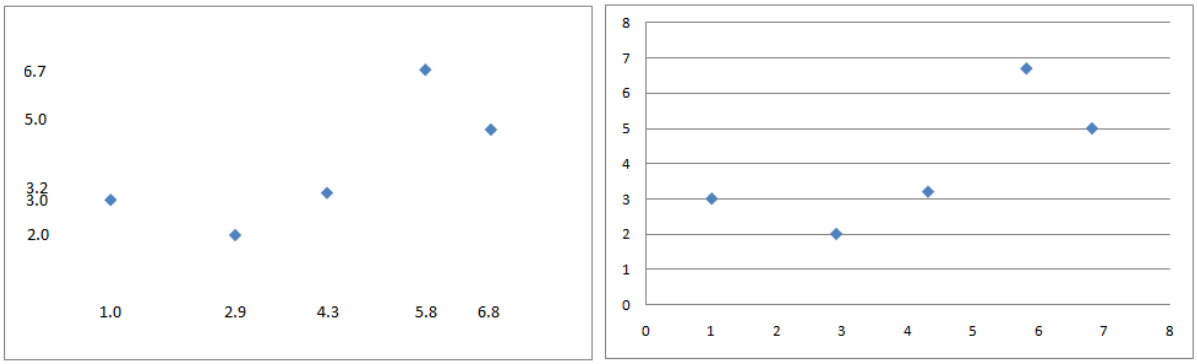

Figura 6 : Gráfico limpo (esq.) e com lixo visual (dir)

A grelha de um gráfico pode também ser usada como forma de transmitir informação, em vez de servir apenas de referência visual, como é ilustrado na Figura, onde é representado o campo magnético induzido por uma esfera magnetizada uniformemente.

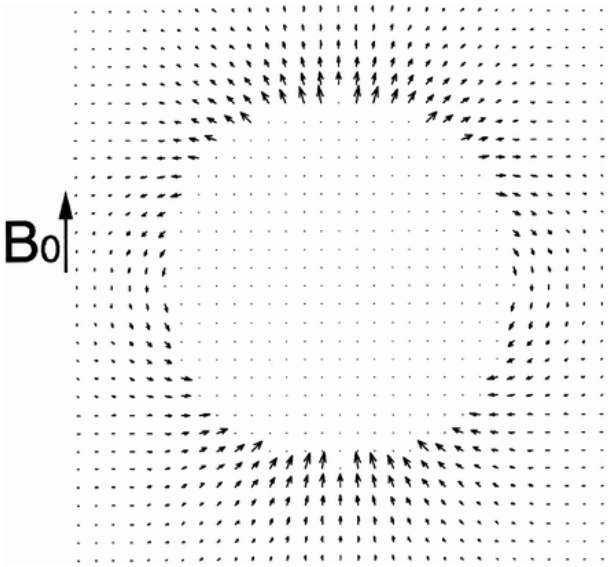

Figura 7: Axel, Kolman et Al. , Origin of a Signal Intensity Loss Artifact in Fat-Saturation MR Imaging

A complexidade de elementos multifuncionais pode transformar visualizações de informação em *puzzles* visuais. Uma forma de identificar essas situações é quando a leitura do gráfico tem de ser feita verbalmente em vez de visualmente, isto é, obriga a usar mnemónicas verbais para a sua interpretação. A utilização de cor gera frequentemente esse tipo de puzzles visuais. A mente humana não ordena naturalmente as cores, exceto talvez o vermelho para refletir níveis mais elevados. Para além disso, as cores tem frequentemente um significado cultural, que varia conforme o observador[35]. Ao observar a cor verde num gráfico, um observador chinês poderia interpreta-la como um valor de teor negativo na escala, visto que na China a cor verde simboliza a morte, enquanto um ocidental associaria essa cor a um valor positivo.

O perigo da utilização de elementos multifuncionais é que estes tendem a gerar puzzles gráficos, que tornam a interpretação do gráfico mais difícil.

2.4.3 Densidade de Dados

Os nossos olhos conseguem fazer um elevado número de distinções numa pequena área (usando uma grelha com linhas muito leves, conseguimos distinguir cerca de 100 pontos num centímetro quadrado). Se considerarmos a informação numérica de um gráfico como uma matriz de observações por variáveis, pode-se calcular a densidade de informação de um gráfico como número de entradas da matriz a dividir pela área do gráfico.

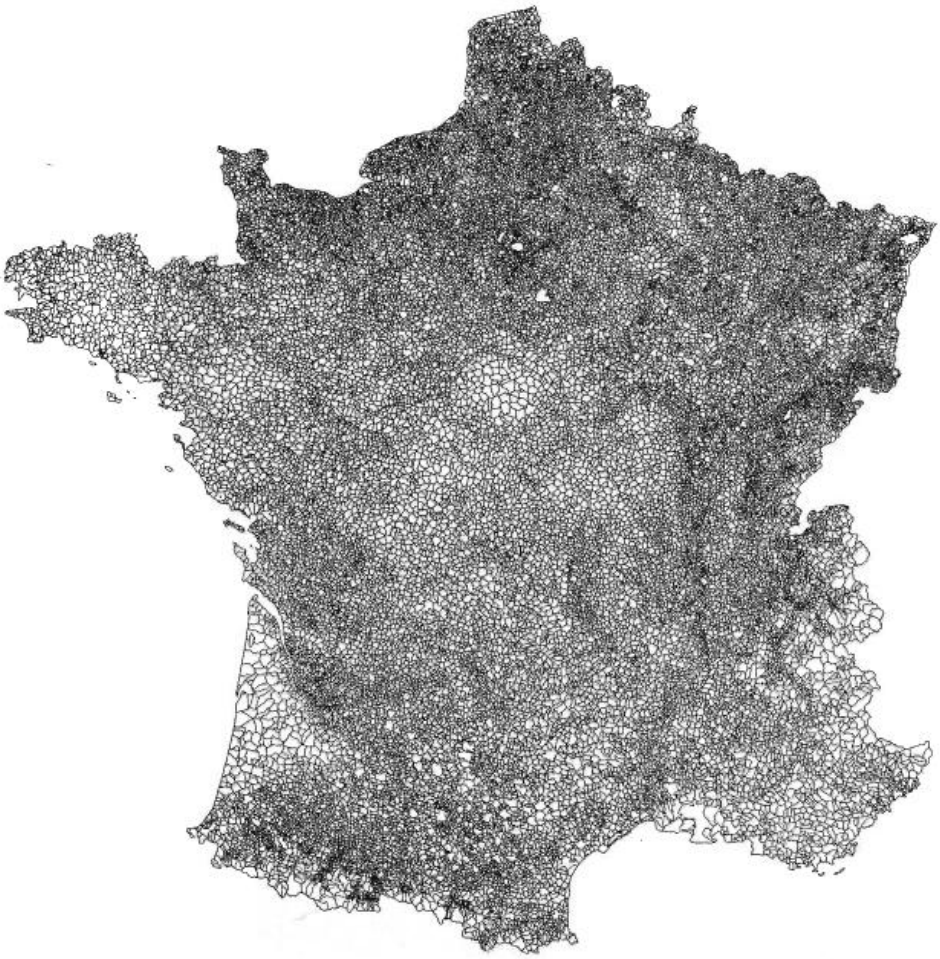

Figura 8 : Jacques Bertin, *Semiologie Graphique* (Paris, 2ª edição, 1973), p. 152.

Como exemplo pode-se observar a Figura 8, extraída de [34], em que se vê um mapa de França apresentando 30 000 freguesias. Seriam precisos cerca de 240 000 números para representar esta informação (30 000 latitudes, 30 000 longitudes e cerca de 6 dígitos para caracterizar a forma de cada freguesia) resultando numa densidade de informação de 1400 números por centímetro quadrado na imagem original.

Gráficos de informação devem ser baseados preferencialmente em matrizes de grande dimensão, já que o custo marginal de interpretação de informação visual é bastante reduzido. Para representar conjunto pequeno de dados, é frequentemente mais eficiente uma representação em tabela ou uma explicação textual da informação. Gráficos podem permitir a interpretação de conjuntos grandes de dados e complexos que de outra forma não poderiam ser transmitidos. Se a representação se torna demasiado densa, existe uma variedade de métodos para a clarificar, nomeadamente a utilização de dados transformados em medidas agregadas (médias de valores contíguos, ou representação de elementos estruturais, referidos na secção sobre análise de redes complexas). O princípio é portanto tentar maximizar a densidade de informação, dentro do razoável. O impacto negativo de lixo visual é ainda mais elevado em gráficos com alto teor de informação.

Outra forma de aumentar a densidade de informação é a redução do tamanho do gráfico, desde que este se mantenha legível. Gráficos bem elaborados podem ser reduzidos sem perda significativa de informação, permitindo inclusive a comparação de várias instâncias do mesmo gráfico, por exemplo para uma análise da evolução de um conjunto de dados ao longo do tempo, ou sob condições experimentais diferentes, de forma a permitir uma interpretação eficiente de mudanças nas relações entre variáveis.

2.4.4 Considerações Estéticas

Alguns princípios para elaborar gráficos esteticamente apelativos [34]:

- Ter cuidado na escolha do formato do *design* dos elementos gráficos.
- Conjugar a utilização de palavras, números e desenhos.
- Refletir um equilíbrio visual na imagem, usar uma escala e proporção relevantes.
- Utilizar uma complexidade de detalhe acessível.
- Tentar ter uma qualidade narrativa na exposição dos dados.
- Criar a representação de forma tecnicamente profissional e cuidada, e evita decoração sem conteúdo, incluindo lixo visual.

2.4.5 Visualização de redes complexas

Para além dos princípios genéricos de *design* já apresentados, existem técnicas de representação especialmente relevantes para a representação visual de redes complexas.
Visualizações interativas que permitam a navegação da imagem (*zoom*, rotação e translação) e a filtragem dos dados representados (por exemplo, na representação de uma rede, filtrar o que é apresentado em função de atributos dos nós) pode simplificar a leitura de gráficos complexos. Uma outra medida de interatividade útil é a animação ao longo do tempo, que permite evidenciar a evolução da rede, ou partes desta. No entanto essas técnicas são frequentemente insuficientes para lidar com a representação de redes complexas, nomeadamente redes com grande densidade de ligações. Apesar de serem de grande beleza, é frequente as representações de redes complexas serem ilegíveis, apenas transmitindo a noção da sua complexidade.
Os atributos inerentes aos nós não transmitem à partida informação estrutural/topológica que é essencial para uma análise que permita uma representação útil em termos de análise topológica da rede, sendo por isso importante o cálculo de medidas estruturais como é referido nas secções "Análise de Redes Complexas" e "Análise de Redes Sociais" para atribuir aos nós atributos que reflitam a topologia da rede, algoritmos de *clustering* ou estratégias de *data mining* (extracção de padrões a partir de conjuntos grandes de informação), de forma a evidenciar a estrutura da rede nas visualizações.

A descrição de um grande número de abordagens para visualização, algoritmos e técnicas de representação espacial (incluindo referências para pesquisa aprofundada) pode ser consultado

em [36], [37] e [38], sendo a última referência de destacar pela sua clareza e indicação de metodologia acompanhando todo o processo de criação da representação, expondo as várias fases de criação da representação: recolha de dados, tratamento e filtragem, procura de informação, escolha de representação, refinamento da representação e inclusão de interatividade.

2.5 Identificação de comunidades

A identificação de comunidades (referidas frequentemente como *clusters*, módulos ou subgrupos coesivos) é muito útil como forma de compreender e classificar a estrutura da rede e observar a sua evolução. O conceito mais frequentemente encontrado para comunidade é um grupo de nós mais densamente ligados entre si do que em relação ao resto da rede. Uma definição mais concreta e completa é especificada dependendo do caso de estudo, assim como a abordagem escolhida para a sua identificação [39]. A categorização dos nós (ou arestas) da rede como membros de comunidades fornece frequentemente um novo olhar sobre o seu papel na rede, ou características/semelhanças entre estes que de outra forma não seriam detetadas, algo que é importante para a análise exploratória de dados empíricos.

O interesse neste tema é ilustrado pela grande diversidade de métodos e algoritmos de identificação de comunidades que foram desenvolvidos, especialmente na última década. Uma recolha exaustiva de diferentes métodos pode ser consultada em [40], sendo que a maior parte devolve uma partição da rede em comunidades sem sobreposição.

Pode-se referir como exemplo a estratégia de definir uma função de qualidade que sirva de heurística para a partição da rede. A medida é otimizada de forma a determinar a divisão da rede ótima. Uma das funções mais utilizadas é a modularidade [41]. A rede é comparada a um modelo de rede estruturalmente semelhante, mas cuja distribuição de ligações é aleatória. A densidade de ligações nas comunidades encontrada no grafo é comparada às da rede aleatória, identificado assim zonas mais coesas, que indicam a presença de comunidades. É utilizada uma diversidade de algoritmos de pesquisa para efetuar essa otimização. [42]

2.5.1 Redes com comunidades sobrepostas, *k-clique percolation*

A maior parte dos métodos de identificação de comunidade mais estabelecidos não devolvem comunidades com sobreposição. No entanto, nas redes sociais observadas empiricamente, especialmente em casos em que existe um elevada densidade de arestas, uma partição representativa da rede tem de admitir que um mesmo nó possa pertencer a várias comunidades. Este tema tem por isso suscitado bastante interesse nos últimos anos.

Um método que se destaca ao rever a literatura sobre deteção de comunidades com sobreposição é o *k-clique percolation method*, apresentado em [43]:

Uma comunidade é neste caso definida como sendo composta por vários sub-grafos completos (cliques) que partilham entre si muitos dos seus nós. Cada comunidade k-clique é constituída por todas as cliques de tamanho *k* (*k-cliques*, sendo *k* o número de nós na clique) tais que estas possam se ligadas por uma série de k-cliques adjacentes. Uma k-clique é adjacente a outra se ambas partilham *k-1* nós. Este conceito de comunidade satisfaz os objetivos de representar grupos de nós fortemente coesos, deixando flexibilidade para que as comunidades tenham sobreposições. Existe a limitação de que há frequentemente nós que não são incluídos em nenhuma das comunidades identificadas. Dependendo do caso de estudo, isso pode levar à necessidade de desenvolver uma estratégia para os incluir nas comunidades já existentes.

O valor ideal para a variável *k* tem de ser descoberto com alguma experimentação. Ao aumentar o valor de *k*, o tamanho das comunidades diminui, e comunidades previamente identificadas podem ficar fragmentadas. No entanto, as comunidades encontradas são mais coesas. Tenta-se aumentar o valor de *k* até ao ponto em que o número de comunidades identificadas reduz drasticamente. Desta forma, procura-se obter uma partição da rede o mais estruturada possível.

Uma descrição pormenorizada da implementação do método, considerações sobre a sua eficiência computacional, escalabilidade e avaliação de resultados pode ser consultada em [40, 44].

Este método foi estendido para poder ser aplicado a redes com ligações pesadas [45] e direcionadas [46].

Para o caso de redes pesadas, é definida a intensidade (*I*) de cada k-clique como sendo a média geométrica do peso das arestas que dela fazem parte. Um valor mínimo para *I* é usado para que só as *k-cliques* com maior intensidade sejam aceites. Esta estratégia permite que existam arestas de peso baixo que possam ser consideradas como constituindo uma *k-clique*, desde que esta tenha um valor de intensidade relevante. Outra estratégia utilizada é impor um limite (*w*) ao peso das arestas. Mais uma vez, é preciso alguma experimentação para encontrar o valor ótimo de limite de peso de forma a perder o mínimo detalhe da rede possível, e conseguir no entanto ter partições da rede úteis.

Para lidar com redes direcionadas, é necessário decidir de que forma a direção das arestas é considerada na constituição das *k-cliques*. Neste caso, define-se como requisito a existência de uma ordenação de nós dentro da *k-clique* baseada no número de arestas saindo do nó que o ligam a outros elementos dentro da clique (*restricted out degree*). Para uma rede direcionada em que não existem ligações recíprocas (ie: duplas) entre nós, esse requisito é cumprido usando as seguintes condições:

- Qualquer aresta incluída na *k-clique* vai de um nó de maior ordem (com maior *restricted out degree*) para um de menor ordem;

- A k-clique não contém ciclos direcionados (caminho fechado);
- O *restricted out degree* de cada nó da *k-clique* é diferente.

Para uma rede direcionada em que existem ligações recíprocas, estas regras não se podem aplicar. Procura-se, no caso em que existam nós ligados por duas arestas, uma em cada sentido, remover a aresta que vai em direção ao nó de maior ordem, até que se verifiquem as condições referidas anteriormente. É referido no mesmo artigo [46] que se podem usar outros critérios para definir uma *k-clique* direcionada.

O método de *k-clique percolation* tem suscitado bastante interesse, e encontram-se estudos recentes debruçando-se sobre as suas variantes e performance, por exemplo [47].

2.6 Ferramentas para análise de redes sociais

Existe um grande número de ferramentas informáticas muito completas orientadas para o estudo de redes complexas, podem-se referir aplicações como o Pajek, o Ucinet ou o Gephi. Para o estudo de redes sociais online, vários dos principais serviços disponibilizam Application Programming Interfaces (API do Twitter ou do Facebook) para permitir a recolha e análise de dados sobre as redes. Existem também bibliotecas de código especializadas para uma variedade de linguagens de programação. Sendo que a aplicação informática desenvolvida em paralelo com esta dissertação é implementada em python, refere-se aqui o NetworkX [48], uma das mais usadas nessa linguagem.

O NetworkX contém funcionalidades para criação, manipulação e estudo da estrutura, dinâmica e características de redes complexas, oferecendo funcionalidades que permitem calcular a maior parte das medidas referidas na secção 2.3.2 (Dados medidos sobre Redes Sociais), assim como a implementação de alguns algoritmos para *layout* de grafos.

Capítulo 3

Identificação de Comunidades e Visualização de Rede Sociais

3.1 Introdução ao problema

Devido à complexidade característica de redes sociais densas e com muitos nós, e à natureza empírica da sua observação, o seu estudo costuma ter uma natureza exploratória. Muitas características inesperadas e interessantes da rede podem ser descobertas através da utilização de métodos de análise de redes sociais. Como é referido na secção 2.4 da Revisão Bibliográfica (Representação visual de informação), a representação visual dessa informação é muito útil para tentar perceber a composição e funcionamento da rede, mas não existem soluções pré-determinadas para todos os casos, sendo que a criação de visualizações de qualidade é um trabalho criativo e experimental por si só.

Sendo um problema multidisciplinar e complexo, não se pretende nesta dissertação abordar todos os seus aspetos com profundidade, nem efetuar um estudo de teor económico. A dissertação foca-se no aspeto de identificação de comunidades e procura-se uma solução para um sistema de navegação que permita explorar dados variantes no tempo, e que se debruce sobre a utilização de identificação de comunidades como componente importante para auxílio ao estudo da rede.

A escolha da rede de comércio entre nações europeias deve-se ao interesse e potencial existentes em estudar redes económicas com métodos de análise de redes sociais, como é evidenciado pelos numerosos estudos sobre a rede de comércio mundial referidos na revisão bibliográfica. O foco na rede de comércio europeu deve-se também em parte ao facto de os dados sobre transações (importações e exportações) entre países europeus estarem disponíveis livremente na web [49], e de forma estruturada o suficiente para serem utilizáveis sem um grande trabalho de tratamento de dados. Embora esta rede corresponda apenas a uma parte da

rede de comércio mundial, tem uma dimensão não trivial, e é uma rede com alguma complexidade devido à elevada densidade de ligações entre países. Os métodos usados para a analisar são portanto interessantes para a análise da rede de comércio mundial. A utilização desta porção da rede mais pequena torna também mais simples o processo de desenvolvimento da solução, e a comunidade observada, sendo composta de países geograficamente próximos, não é arbitrária.

Apesar de existirem muitas ferramentas informáticas, algumas muito completas, para análise e visualização de redes complexas, a escolha de tecnologias web atuais para o desenvolvimento de uma aplicação desse tipo (e integração de várias dessas tecnologias, nomeadamente Django/python, JSON, Javascript, e bibliotecas de processamento gráfico baseadas no *canvas* do HTML5 como o Paper.js ou o Processing.js) é também um ponto interessante, e uma componente importante do trabalho desenvolvido, sendo que esta dissertação é feita no Mestrado Integrado em Engenharia Informática e Computadores.

3.2 Abordagem e Resultados

O problema da identificação e visualização de comunidades numa rede social é logicamente repartido em vários passos. É necessário efetuar a recolha e tratamento adequado dos dados que são usados para modelar a rede, aplicar um método de identificação de comunidades adequado à rede observada, e desenvolver um sistema de visualização que tome proveito da informação obtida de forma a facilitar a exploração visual (e interativa) da rede, e dos atributos dos seus nós. Pretende-se nesta dissertação ilustrar uma abordagem a esse problema através do desenvolvimento de uma aplicação web interativa para análise exploratória de uma rede social. Apesar de não se pretender efetuar uma análise económica da rede em questão, a escolha da rede de comércio entre países europeus mostra uma aplicação concreta deste tipo de abordagem num domínio de grande interesse na atualidade, como é referido na secção 2.3.3(Análise de redes sociais na economia).

A abordagem ao problema é descrita mais em concreto nas secções seguintes, assim como alguns resultados obtidos.

3.2.1 Recolha e tratamento de dados

A informação necessária para a caracterização da rede foi recolhida da base de dados do Eurostat [49] e inserida na base de dados do servidor.

Estes dados abrangem o período entre 1988-2011. Para cada ano, são registados os nós e arestas que constituem a rede. Esta organização de dados deve-se a este ser o formato ideal para gerar os grafos de rede utilizando a ferramenta escolhida (networkX[50]).

Para cada aresta, para além do peso (valor agregado do total de exportações ou importações entre os países), foi calculado um valor de peso relativo expressando a percentagem do peso da aresta em relação ao volume total do nó (soma do peso de todas as arestas). Esta medida é utilizada de forma a atribuir às arestas um significado de importância relativa comparável de cada ligação, e é utilizada no processo de identificação de comunidades, e para alguns cálculos auxiliares ao nível da geração de representações visuais.

3.2.2 Identificação de comunidades

Foi utilizada a versão de linha de comandos do CFinder executada num script para identificação das comunidades usando **k-clique percolation** sobre cada fatia temporal da rede. O resultado foi inserido na base de dados do servidor de forma a que a informação sobre comunidades possa ser integrada na visualização da rede. A versão gráfica do CFinder também foi utilizada para verificação dos resultados e para a afinação das variáveis k e w usadas no algoritmo de CPM.

A escolha deste método de identificação de comunidades justifica-se pela natureza da rede de comércio europeu (rede muito densa com nós direcionados e com peso) e pela natureza das comunidades que se procura identificar, em que tem de se admitir sobreposição. Para outro tipo de dados, poderia ser mais adequado utilizar outro método de identificação de comunidades.

Consultando os estudos referentes ao comércio mundial referidos na revisão bibliográfica, estes indiciam que a rede sofre poucas alterações de ano para ano, pelo que se achou adequado utilizar apenas uma das fatias temporais como referencia para testar os valores de k e w. No entanto, se o objeto desta tese fosse a análise da rede em si, seria interessante efetuar uma pesquisa individual em cada fatia temporal para descobrir os valores ótimos de k e w para cada período observado, o que teria o potencial de fornecer resultados mais rigorosos, sendo esse um ponto de estudo no futuro.

Foram utilizados os valores das exportações de país para país como peso das arestas.

Apesar de os dados recolhidos abrangerem o período [1988-2010], a informação só está completa para todas as nações da rede no período [1998-2010]. Nesse intervalo, a densidade da rede varia, como seria de esperar, entre 99,8% e 100%, justificando a necessidade de se impor um filtro às arestas de menor valor. Visto existir uma grande disparidade no valor total absoluto de exportações de cada nação, foi efetuada uma relativização no peso das arestas. A cada aresta foi atribuído como peso a percentagem do seu valor absoluto relativamente ao valor total das exportações do seu nó de origem. Pretende-se assim encontrar as comunidades de países usando como critério de ligação a importância relativa da relação de exportação de bens entre cada país. Esta estratégia é empregue como exemplo para testar o algoritmo de identificação de comunidades, sendo que a interpretação dos resultados obtidos está condicionada por uma limitação: visto que só são consideradas as relações de comércio intraeuropeias, um determinado país pode na realidade ter uma percentagem do seu volume de exportações muito

significativa com destino a países extraeuropeus, algo que não é considerado, e que poderia afetar os resultados obtidos em termos das comunidades identificadas.

Pode-se dar como exemplo a partição da rede obtida para o ano 2002, usando os parâmetros k = 4 e w = 12.0, representada na Figura 9, gerada usando a ferramenta CFinder. Cada comunidade encontrada é representada como uma área colorida. Na legenda de cada nó, encontra-se o código indicativo do país (seguindo os indicativos nacionais usados na internet), e a proporção de arestas que apontam para o nó e saem dele, nas comunidades encontradas. Para estes parâmetros, oito países não foram classificados em nenhuma comunidade.

Apesar de não se poderem tirar conclusões sem um estudo mais aprofundado, observa-se uma relação entre a proximidade geográfica dos países e o seu agrupamento em comunidades, como se pode observar na Figura 10, em que cada país foi colorido conforme a comunidade a que pertence. Os países pertencentes a múltiplas comunidades estão marcados com círculos identificando as múltiplas comunidades a que pertencem. Os países que não foram alocados a nenhuma comunidade usando os parâmetros *k* e *w* utilizados estão pintados a cinzento. Consegue-se identificar claramente os pontos de sobreposição, sendo que a Alemanha (DE) é o único nó que está incluído em todas as comunidades. A Itália (IT) e o Reino Unido (UK) pertencem ambos a duas comunidades em simultâneo.

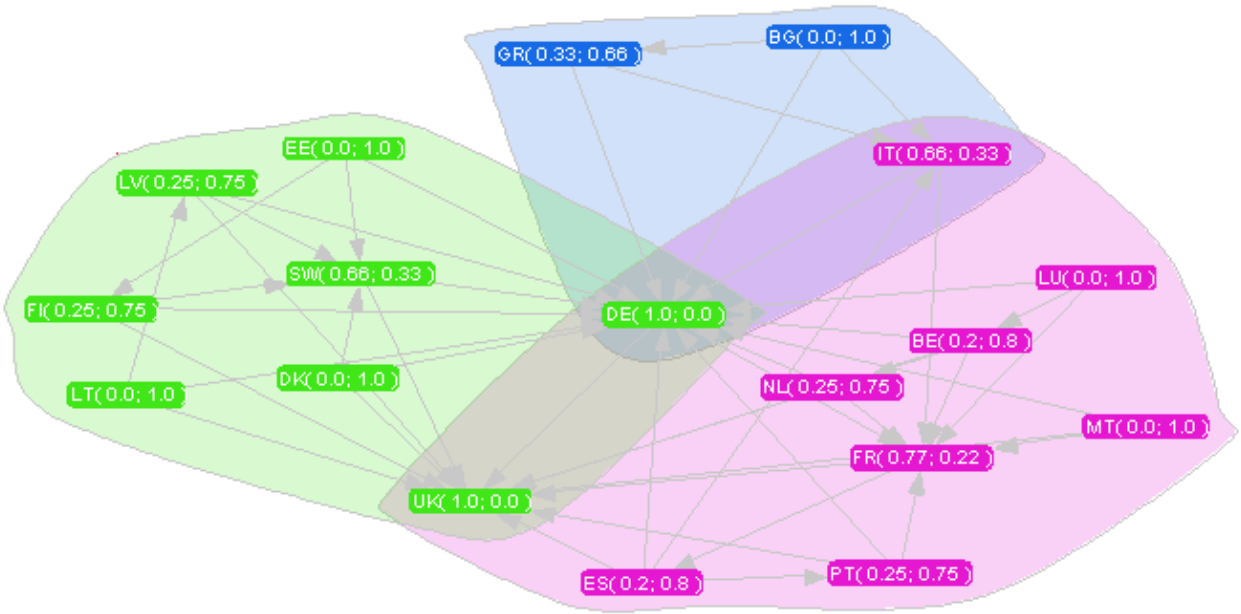

Figura 9 Partição de rede gerada no CFinder, com *layout* ajustado manualmente

Este tipo de informação dá de imediato uma nova perspetiva sobre a estrutura da rede, tendo o potencial para ser uma grande mais-valia para a sua exploração.

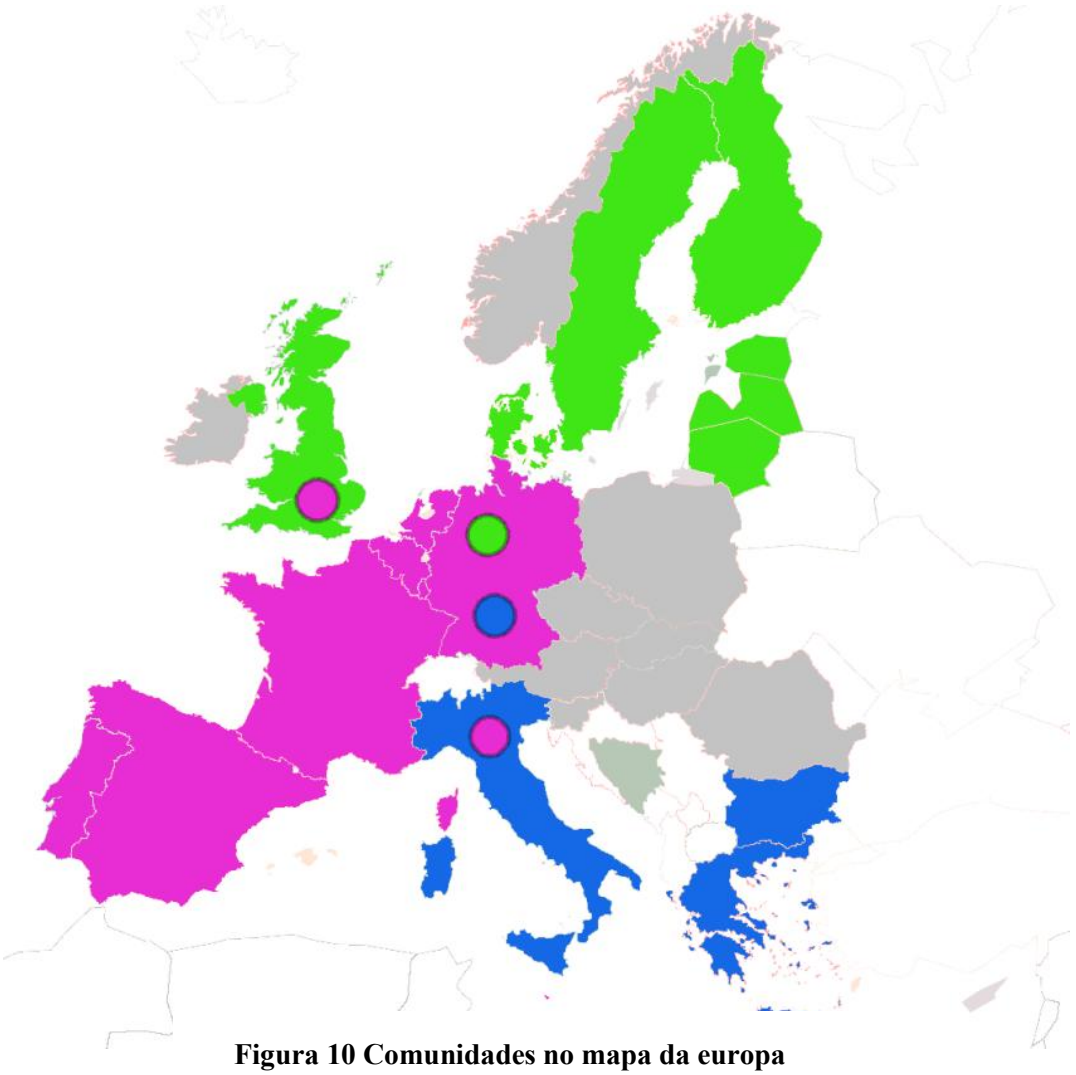

Figura 10 Comunidades no mapa da europa

3.2.3 Visualização

Como é exposto no capítulo 2.4 (Representação Visual de Informação) existe uma componente significativa de criatividade associada à criação de soluções para abordar um problema de visualização de informação, independentemente dos moldes em que este se enquadre. A solução proposta nesta dissertação pretende aproveitar alguns dos modelos típicos de visualização de redes encontrados (representação sob a forma de grafo, com utilização de alguns *layouts* bastante conhecidos), enquadrando-os num sistema interativo que facilite a exploração dos dados sobre a rede de comércio europeu. A utilização da interatividade é especialmente importante para facilitar essa exploração, devido à dimensão temporal dos dados, que cobrem um período de vários anos. A interface desenvolvida ao nível do utilizador integra portanto elementos de *layout* de grafos, filtragem de informação e interatividade, que são expostos nas secções seguintes:

3.2.4 Layout

Sendo que um dos problemas da visualização de uma rede é a disposição dos nós, foram calculadas coordenadas para os nós seguindo três *layouts* frequentemente usados:

- *Spring layout*: foi usado o algoritmo [51] para gerar *layout* em mola. Um problema encontrado é que, para cada fatia temporal, o *layout* obtido é bastante diferente devido à forma como o algoritmo funciona, o que torna impossível comparar vários layouts. Optou-se por utilizar um dos possíveis *layouts* calculados durante a navegação. Apesar de este poder não ser o ideal para todas as fatias temporais, a comparação de períodos é facilitada;
- Circular: evidencia melhor a densidade de ligações de cada nó, com a desvantagem que a posição dos nós deixa de ter significado, o que diminui a quantidade de informação representada;
- Disposição geográfica dos nós: fazendo corresponder os nós a posições geográficas sobre um mapa, o utilizador tem uma referência imediata do país a que cada nó corresponde. A relação da geografia com a densidade de ligações e comunidades encontradas pode ser facilmente analisada.

3.2.5 Filtragem de informação

Um dos métodos que facilita a exploração visual de informação é a filtragem da informação representada, permitindo ao utilizador focar-se em determinados pontos da rede. Pode-se referir:

Opacidade das arestas: Como forma de evidenciar as arestas de maior peso, foi atribuído um valor de opacidade às arestas proporcional ao seu peso. Desta forma, as arestas de maior peso ficam mais evidenciadas, enquanto que as de menor peso ficam apagadas na imagem. Sendo que a rede em estudo tem uma densidade de ligações muito elevada, este processo torna bastante mais clara a leitura da imagem, e a identificação da estrutura da rede. Isso é ilustrado na Figura 11, na qual duas representações idênticas da rede de comércio europeu (geradas pela aplicação desenvolvida, criando um grafo com um layout Fruchterman-Reingold [51]. Cada nó representa um país, e cada ligação a sobreposição do volume de exportações entre cada país) são apresentadas, uma com a opacidade das arestas ajustada ao seu peso, e outra com opacidade uniforme.

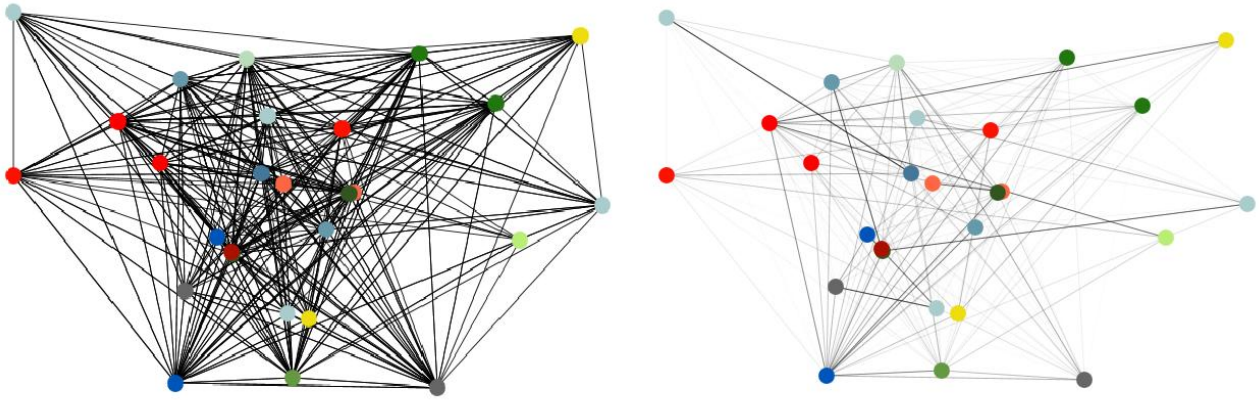

Figura 11 antes (esq.) e depois (dir.) de filtragem da opacidade das

- Highlight da seleção: Ao clicar-se num nó, as suas ligações são evidenciadas a azul brilhante. Isto faz com que o utilizador se possa focar num determinado nó e se aperceba facilmente dos seus vizinhos mais significativos, e da forma das suas ligações.
- Coloração dos nós: A cor dos nós pode também ser usada para evidenciar um atributo, como, por exemplo, a sua pertença a determinada comunidade identificada.

3.2.6 Interação

Como forma de trazer interatividade à exploração da rede, foi criado um painel suplementar à direita da representação do grafo de rede, que é ativado quando se clica num nó com o rato (Figura 12). Este painel auxiliar contém um gráfico de barras representando, no eixo y, o valor total do peso das ligações (sendo que a barra é dividida em secções verticais proporcionais à contribuição de cada vizinho do nó para o total). Para a rede em estudo, o valor representa o total de exportações para países europeus, mas pode ser usada qualquer outra variável. No eixo x, está o tempo. Cada barra corresponde a uma fatia temporal da rede. A escala nas abscissas é variável, podendo aparecer mais ou menos fatias no gráfico. Para além de permitir observar a evolução da variável do eixo y, e a partição de cada barra, é possível interagir com este painel. Ao clicar numa barra, a representação em grafo representada no painel do lado esquerdo é atualizada. Isto permite uma navegação através do tempo focada na observação de um nó, em paralelo com a rede no seu conjunto, e evidenciando o nó, sendo também possível, por exemplo, fazer um *highlight* das suas ligações e vizinhos.

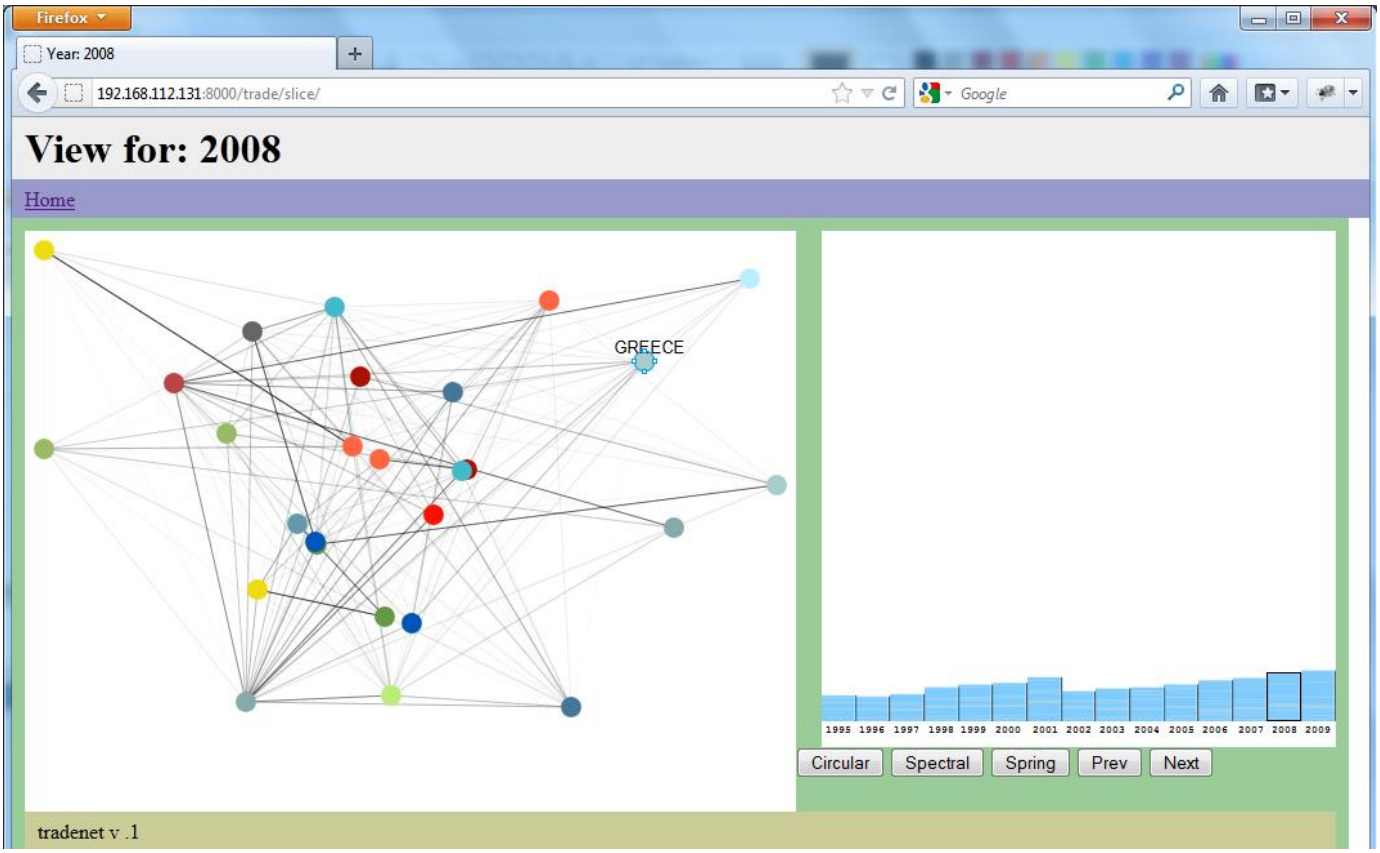

Figura 12 *Screenshot* da aplicação

3.3 Resumo e Conclusões sobre resultados obtidos

Foi criada uma aplicação que tira partido de ferramentas e bibliotecas de código para análise de redes sociais. Esta permite uma navegação interativa da rede, tendo o potencial para uma partilha comunitária de conhecimento, visto que é baseada numa plataforma web. Embora seja apenas um protótipo com muitas limitações, especialmente ao nível da usabilidade, e desenvolvido individualmente num tempo restrito, é um primeiro passo sólido para o desenvolvimento de uma aplicação com potencial. Seria, claro, necessário atingir um nível mais avançado no desenvolvimento da aplicação (e trabalho em conjunto com pessoal especializado no tema do caso de estudo) para obter uma versão que possa fornecer resultados que não sirvam apenas de exemplo metodológico, e entrar numa fase de teste de resultados e usabilidade.

A utilização do método de percolação de cliques permitiu, com algumas afinações dos parâmetros k e w, identificar comunidades. Como forma de validação, foram comparados os outputs de alguns anos contíguos, em que as comunidades se mantêm bastante estáveis de ano

para ano, e coerentes com a distribuição geográfica dos países, o que é uma primeira medida de que este método tem potencial para ser usado com esta rede de comércio.

Seria no entanto necessário um estudo de resultados mais aprofundado para se tirar conclusões mais definitivas sobre a qualidade do método.

Capítulo 4

Implementação

Este capítulo é dedicado à apresentação de detalhes de nível mais baixo relacionados com o enquadramento e implementação das soluções preconizadas no capítulo anterior, nomeadamente a estrutura da aplicação, as estruturas de dados usadas e a utilização de ferramentas de análise de redes sociais.

4.1 Arquitetura da aplicação

A aplicação está dividida em vários módulos independentes, correspondendo às várias fazes de processamento de dados necessárias para o seu funcionamento. Na Figura 13 é mostrada uma representação simplificada, (mas correspondente aos principais módulos implementados) da arquitetura da aplicação.

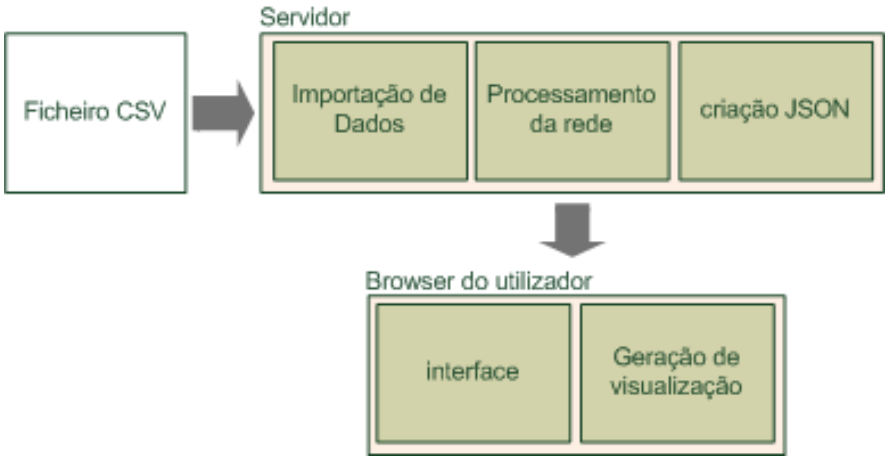

Figura 13 Estrutura da aplicação

O (s) ficheiro (s) CSV contém a descrição da rede, e são importados diretamente para a base de dados do servidor

Como forma de explorar a utilização de tecnologias de desenvolvimento web modernas, foi utilizada a *web framework* Django [52] (implementada na linguagem python [53] para construir a aplicação. No servidor, é feita a importação dos dados, a sua transformação num grafo de rede (e cálculo de medidas úteis para a sua análise), e é criada a estrutura de dados a ser enviada para o browser do utilizador, no formato JSON. No browser do utilizador, é apresentada uma página HTML que serve de suporte para a *interface* da aplicação, e onde são geradas as visualizações da rede, usando JavaScript e paper.js, sem a utilização de nenhum *plugin* externo.

4.1.1 Importação de dados

Os dados utilizados foram extraídos da base de dados do EuroStat [49] para ficheiros no formato CSV (*Comma Separated Value*) com a informação sobre o fluxo de exportações e importações agregadas entre países europeus, para o período de 1988 a 2010. Estes ficheiros foram importados para uma base de dados relacional no servidor, sendo filtrados para incluírem apenas as interações entre países europeus. Foram removidos os carateres não alfa numéricos, e normalizado o formato de *encoding* para UTF-8, e identificados registos duplicados ou irregulares, para tornar os dados importados íntegros.

4.1.2 Geração de grafo de rede

Foi usada a biblioteca networkX [50]para Python para gerar a estrutura de dados que suporta o grafo da rede. É gerado um grafo para cada fatia temporal da rede, e são nele carregados todos os atributos necessários à análise e geração de visualização, incluindo as coordenadas dos nós para *layout*, e qualquer atributo adicional que se queira associar à rede, nós ou arestas, para utilização na visualização ou com ferramentas de processamento de dados.

4.1.3 Comunicação com a aplicação cliente

Os objetos nx.Graph são convertidos em *arrays* associativos no formato JSON. Este formato foi escolhido pois um objeto json é diretamente convertível em estrutura de dados Javascript. Visto que o módulo da aplicação *client-side* é em JavaScript, isto torna o JSON o formato ideal a usar na transmissão dos dados, evitando um passo de conversão de dados que de outra forma seria necessário.

4.1.4 Interface e visualização da rede: *paper.js*

A aplicação é utilizada a partir de uma página web. As representações gráficas estão em campos *canvas* de HTML5, usando *paper.js* [54].

O *paper.js* é uma *framework* de *scripting open source* para a geração de gráficos vetorizados, que corre no campo *canvas*. Foi escolhida especialmente por três das suas características. Em primeiro lugar, o paper.js foi escolhido pois está implementado em javaScript, pelo que é plenamente integrável na página HTML e pode facilmente interagir com qualquer elemento ou função definidos na página. A segunda característica de interesse é a utilização de objetos vetorizados no *paper.js*: visto que os dados com que a aplicação tem de lidar são um conjunto de nós e vértices representados numa tela, a possibilidade de os representar em objetos vetorizados faz com que se possa atualizar facilmente propriedades como a posição, a cor (ou qualquer atributo que lhes seja arbitrado) sem ser necessário criar objetos novos para cada atualização da tela, o que constitui um fator essencial para a performance da aplicação ao nível da geração de visualizações. A terceira característica útil do *paper.js* é que, visto este ser uma *framework* orientada para *designers*, tem uma estrutura de cena semelhante à de um editor de imagens como o *Photoshop*, o que inclui conceitos de camadas na imagem, permite aplicar propriedades a grupos de objetos em simultâneo ou alterar propriedades gráficas dos objetos de forma trivial, ou efetuar transformações matriciais, e as estruturas de dados nele usado refletem esse tipo de conceitos, o que facilita imenso o desenvolvimento da aplicação, tanto ao nível da geração da imagem, como ao nível da interatividade (um objeto vetorizado é reconhecido como objeto na tela, e não como um conjunto de bits coloridos, o que torna muito simples a deteção de cliques do rato sobre objetos, e outros tipos de interação, usando as funções de deteção de colisões incluídas na *framework*).

Convém referir que, numa primeira fase, foi usada outra biblioteca, o *processing.js*, também muito popular e capaz de gerar imagens de grande complexidade e beleza, mas a ausência de vectorização de objetos levou à mudança para o *paper.js*.

A utilização destas tecnologias é importante referir pois, toma partido de novas capacidades do HTML5, sem as quais não era possível criar este tipo de aplicação numa página web sem a utilização de *plugins* externos como o *adobe flash* ou o *java*.

4.2 Utilização de ferramentas de análise de redes sociais

Seguem-se algumas notas de interesse sobre a integração de bibliotecas/ferramentas de análise de redes na aplicação web desenvolvida, e a justificação da sua utilização.

4.2.1 Biblioteca networkx de python

Para definir o grafo, calcular *layouts* e medidas de SNA, foi usada a biblioteca networkx (*nx*) para a linguagem python [48]. Utilizando a informação da base de dados, foram gerados objetos nx.DiGraph (*Directed Graph*), sobre os quais foram feitos os cálculos de *layout*, e foram gerados os objetos JSON para transmissão para o browser-cliente. Os objetos nx.DiGraph foram guardados em ficheiros de texto, usando pickling, um processo de serialização do python. Desta forma, é possível carregar os objetos diretamente na memória do servidor, caso se queira efetuar cálculos de medidas de SNA ou algoritmos de layout/clustering, entre outros. O networkx é bastante usado na comunidade de análise de redes, e, sendo implementado em python, é uma escolha lógica.

4.2.2 CFinder

Para a identificação de comunidades, foi usada a versão Shell do CFinder, implementado pelos autores do método de kCPM, executada num script para todas as fatias temporais da rede, sendo os resultados guardados em memória como atributos dos objetos newtworkx. Graph referidos no subcapítulo anterior. Numa primeira fase, foram testadas algumas fatias temporais de forma isolada, usando a interface gráfica do CFinder, de forma a afinar os parâmetros do algoritmo (*k* e *w*).

Capítulo 5

Conclusões e Trabalho Futuro

5.1 Satisfação dos Objetivos

Foi desenvolvido um modelo de exploração visual e interativa para análise de uma rede social, baseada em tecnologia *web* atual, que é um ponto de partida para o desenvolvimento de uma análise sobre a rede de comércio entre nações. A aplicação está construída de forma a ser fácil acrescentar dados discriminado sobre transações de vários tipos de produto, permitindo uma eventual análise sectorial da rede, sendo assim feita a prova de conceito pretendida.

Foi testado o método de *k-clique percolation* para identificação de comunidades sobre a rede de comércio europeu, tendo-se obtido uma partição da rede promissora (na medida em que se consegue observar uma correspondência com a proximidade geográfica dos países), tendo em conta que esta foi gerada a título de exemplo, não fazendo sentido uma análise mais aprofundada que teria de depender de conhecimento sobre economia de forma a avaliar a sua qualidade.

5.2 Trabalho Futuro

Identificação de comunidades:

- Utilizando o método CPM, comparar os resultados obtidos nas várias fatias temporais da rede, e quantificar a variação das comunidades;
- Comparar os resultados obtidos entre as variantes do algoritmo com arestas pesadas e não pesadas, e entre rede direcionada e não-direcionada para a rede de comércio.

Desenvolvimento da aplicação:

- Transformar o protótipo numa aplicação pronta pra o utilizador, nomeadamente com a inclusão de funcionalidades de filtragem e *zoom* utilizáveis na interface;
- Aperfeiçoamentos dos *layouts*, nomeadamente, pesquisa de layouts construídos a partir da partição de comunidades da rede, de forma a que os nós de uma mesma comunidade estejam agrupados visualmente.

Este desenvolvimento implica um trabalho em conjunto com pessoas da área de economia, de forma a produzir resultados úteis para a aquisição de conhecimento sobre a rede económica em análise.

Vertente web colaborativa:

- Desenvolvimento de um módulo de importação de dados para utilizadores do sistema
- Sistema comunitário que facilite a partilha de redes e resultados, de forma a permitir trabalho colaborativo à distância.

Testes:

Devido às limitações de tempo inerentes a este projeto, não foi possível desenvolver uma fase de testes sobre os resultados, pelo que esse seria um ponto essencial para trabalho futuro.

Referências

[1] VisualComplexiy. (2011). Available: http://visualcomplexity.com/

[2] A. R. Barabasi, "Statistical mechanics of complex networks," *Reviews of Modern Physics,* vol. 74, p. 47, 2002.

[3] D. J. Watts and S. H. Strogatz., "Collective dynamics of 'small-world' networks," *Nature* pp. 440-442 1998.

[4] X. F. Wang and G. Chen, "Synchronization in scale-free dynamical networks: robustness and fragility," *IEEE Transactions on Circuits and Systems I-regular Papers,* vol. 49, pp. 54-62, 2002.

[5] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*: Cambridge University Press, 1994.

[6] M. E. J. Newman and J. Park, "Why social networks are different from other types of networks," *Physical Review E,* vol. 68, p. 036122, 2003.

[7] L. Hamill and N. Gilbert, "A simple but more realistic agent-based model of a social network," ed: Surrey Scholarship Online, 2008.

[8] L. C. Freeman, "Visualizing Social Networks," *Carnegie Mellon: Journal of Social Structure,* vol. 1, 2000 2000.

[9] W. Richards and A. Seary, "Eigen Analysis of Networks," *Carnegie Mellon: Journal of Social Structure,* vol. 1, 2000 2000.

[10] M. O. Jackson, *Social and Economic Networks*: Princeton University Press, 2008.

[11] S. Boccaletti, *et al.*, "Complex networks: Structure and dynamics," *Physics Reports,* vol. 424, pp. 175-308, 2006.

[12] M. O. Jackson, "An Overview of Social Networks and Economic Applications," in *Handbook of Social Economics*. vol. Volume 1, A. B. Jess Benhabib and O. J. Matthew, Eds., ed: North-Holland, 2011, pp. 511-585.

[13] S. Goyal, *An Introduction to the Economics of Networks*: Princeton University Press, 2009.

[14] S. Amīn, *Capitalism in the age of globalization: the management of contemporary society*: Zed Books, 1997.

[15] M. Serrano, *et al.*, "Topology of the world trade web," *Physical Review E,* vol. 68, p. 015101, 2003.

[16] G. Caldarelli, *Scale-free networks: complex webs in nature and technology*: Oxford University Press, 2007.

[17] S. Schiavo, *et al.*, "International trade and financial integration: a weighted network analysis," *Quantitative Finance,* vol. 10, pp. 389-399, 2010.

[18] S. Kim and E.-H. Shin, "A Longitudinal Analysis of Globalization and Regionalization in International Trade: A Social Network Approach," *Social Forces,* vol. 81, pp. 445-468, 2002.

[19] M. Zhou, "Intensification of geo-cultural homophily in global trade: Evidence from the gravity model," *Social Science Research,* vol. 40, pp. 193-209, 2011.

[20] E. E. Leamer, "The Heckster-Ohlin model in theory and practice," *Princeton Studies in International Finance,* vol. 77, February 1995 1995.

[21] G. Fagiolo, *et al.*, "The evolution of the world trade web: a weighted-network analysis," *Journal of Evolutionary Economics,* vol. 20, pp. 479-514, 2010.

[22] K. Bhattacharya, *et al.*, "The International Trade Network," in *Econophysics of Markets and Business Networks*, A. Chatterjee and B. K. Chakrabarti, Eds., ed: Springer Milan, 2007, pp. 139-147.

[23] D. Garlaschelli and M. I. Loffredo, "Structure and evolution of the world trade network," *Physica A,* vol. 355, pp. 138-144, 2005.

[24] H. Wang, *et al.*, "Symmetry in world trade network," *Journal of Systems Science and Complexity*, vol. 22, pp. 280-290, 2009.
[25] H. van Ees and R. Bachmann, "Transition Economies and Trust Building: A Network Perspective on EU Enlargement," *Cambridge Journal of Economics*, vol. 30, pp. 923-939, 2006.
[26] M. Serrano, *et al.*, "Patterns of dominant flows in the world trade web," *Journal of Economic Interaction and Coordination*, vol. 2, pp. 111-124, 2007.
[27] A. Mayer, "Online social networks in economics," *Decision Support Systems*, vol. 47, pp. 169-184, 2009.
[28] E. Janeba, "International trade and consumption network externalities," *European Economic Review*, vol. 51, pp. 781-803, 2007.
[29] T. Furusawa and H. Konishi, "Free trade networks," *Journal of International Economics*, vol. 72, pp. 310-335, 2007.
[30] G. J. Felbermayr and F. Toubal, "Cultural proximity and trade," *European Economic Review*, vol. 54, pp. 279-293, 2010.
[31] G. Fagiolo, *et al.*, "World-trade web: Topological properties, dynamics, and evolution," *Physical Review E*, vol. 79, p. 036115, 2009.
[32] A. Aviat and N. Coeurdacier, "The geography of trade in goods and asset holdings," *Journal of International Economics*, vol. 71, pp. 22-51, 2007.
[33] J. Steele and N. Iliinsky, *Beautiful Visualization: Looking at Data Through the Eyes of Experts*: O'Reilly Media, 2010.
[34] E. Tufte, *The visual display of quantitative information*: Graphics Press, 2001.
[35] C. Ware, *Information visualization: perception for design*: Morgan Kaufman, 2004.
[36] C. Chen, *Information visualization: beyond the horizon*: Springer, 2006.
[37] J. Zhang, *Visualization for information retrieval*: Springer, 2008.
[38] B. Fry, *Visualizing Data: Exploring and Explaining Data with the Processing Environment*: O'Reilly, 2007.
[39] J.-P. O. Mason A. Porter, Peter J. Mucha, "Communities in Networks," *Notices of the American Mathematical Society*, vol. 56, pp. 1082-1097, 2009.
[40] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, pp. 75-174, 2010.
[41] R. Guimerà, *et al.*, "The real communication network behind the formal chart: Community structure in organizations," *Journal of Economic Behavior & Organization*, vol. 61, pp. 653-667, 2006.
[42] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E*, vol. 74, p. 036104, 2006.
[43] G. Palla, *et al.*, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, pp. 814-818, 2005.
[44] I. Derényi, *et al.*, "Clique Percolation in Random Networks," *Physical Review Letters*, vol. 94, p. 160202, 2005.
[45] I. Farkas, *et al.*, "Weighted network modules," *New Journal of Physics*, vol. 9, p. 180, 2007.
[46] G. Palla, *et al.*, "Directed network modules," *New Journal of Physics*, vol. 9, p. 186, 2007.
[47] F. Reid, *et al.*, "Percolation Computation in Complex Networks," *Arxiv preprint arXiv:1205.0038*, 2012.
[48] networkx. (2012). *networkX - Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks*. Available: http://networkx.lanl.gov/
[49] E. Comission. *Eurostat*. Available: http://epp.eurostat.ec.europa.eu
[50] A. Hagberg, *et al.*, *Exploring network structure, dynamics, and function using networkx*, 2008.
[51] T. M. J. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Software: Practice and Experience*, vol. 21, pp. 1129-1164, 1991.
[52] Django. (2012). *Django Project*. Available: https://www.djangoproject.com/

[53] python. (2012). *Python Programming Language*. Available: http://python.org/
[54] paper.js. (2012). *Paper.js - Scriptographer ported to JavaScript and the browser, using HTML5 Canvas*. Available: http://paperjs.org/