Escola de Tecnologias e Arquitetura

Departamento de Ciências e Tecnologias de Informação

IDENTIFICAÇÃO E AVALIAÇÃO DE COMUNIDADES EM REDES DE COAUTORIAS

David Valente Fernandes

Dissertação submetida como requisito parcial para obtenção do grau de

Mestre em Informática

Orientadora:

Professora Doutora, Professora Auxiliar, Maria João Cortinhal, ISCTE-IUL

Coorientador:

Professor Doutor, Professor Auxiliar, Nuno Manuel Cruz David, ISCTE-IUL

Junho de 2017

Resumo

A investigação científica desenvolvida no seio universitário tem assumido um papel cada vez mais relevante na vida destas instituições, quer pela reputação e respeitabilidade que criam, quer pelo apoio financeiro nacional e internacional que é dado às unidades de investigação em função da sua produção científica. Assim, é crucial para o ISCTE-IUL desenvolver mecanismos que promovam e incentivem essa mesma produção.

O Ciência-IUL é o repositório digital das publicações científicas produzidas pelos autores do ISCTE-IUL. Esta dissertação propõe transformar a informação presente no mesmo numa rede de coautorias onde se possa aplicar algoritmos de identificação de comunidades. A identificação de comunidades de modo automático permite detetar padrões de partilha do conhecimento dentro do ISCTE-IUL que de outra forma, seriam impercetíveis.

Para analisar a rede de coautorias e as comunidades identificadas ao longo do tempo e tomar decisões sobre as mesmas, é desenvolvida uma base de dados onde a informação é persistida e o acesso à mesma é feita através de uma interface gráfica desenvolvida para o efeito. Assim, os responsáveis do ISCTE-IUL poderão a qualquer momento visualizar as redes de partilha de conhecimento e realizar análises temporais sobre a evolução das mesmas.

Palavras-chave: grafo, rede, coautoria, comunidade, MCL, ABCD

Abstract

Scientific research developed in universities has played an increasingly important role in the life of these institutions, both for the reputation and respectability they create as well as for the national and international financial support given to research units in terms of their scientific output. Thus, it is crucial for the ISCTE-IUL to develop mechanisms that promote and encourage scientific production.

The Science-IUL is the digital repository of the scientific publications produced by the authors of ISCTE-IUL. This dissertation proposes to transform the information present in Science-IUL into a network of co-authorships where community identification algorithms can be applied. Automatically identifying communities enables detection of patterns of knowledge sharing within ISCTE-IUL, which otherwise would be imperceptible.

In order to analyse the network of co-authorships and the communities identified over time, as well as to make decisions about such communities, a database is developed where information is persisted and whose access is achived through a graphical interface. In this way, ISCTE-IUL officials will be able to visualize the knowledge-sharing networks at any time and carry out temporal analysis of their evolution.

Keywords: graph, network, co-authorship, community, MCL, ABCD

Índice

Glossário

ISCTE-IUL - Instituição pública de ensino universitário criada em 1972, também conhecido como ISCTE-Instituto Universitário de Lisboa.

Ciência-IUL – Repositório digital da investigação e ciência produzida no ISCTE-IUL.

API – Acrónimo com origem no Inglês, representa o conjunto de informação sobre funcionalidades disponibilizadas por um software para terceiros saberem como os consumir.

REST – Em inglês, *Representational State Transfer*, é uma arquitetura desenvolvida para a partilha de recursos na *internet*.

JSON – Em inglês, *JavaScript Object Notation*, é um formato para partilha de informação de modo estruturado.

Web – O sistema que define como são acedidos e manipulados documentos através da *internet*.

URL – Em inglês, *Uniform Resource Locator*, é uma referência única de um recurso na *internet*, normalmente utilizado para o obter.

HTML – Em inglês, *HyperText Markup Language*, é uma linguagem que define um conjunto de marcadores que são utilizados para construir uma página Web.

Scimago – Portal público na Web que permite consultar informação associada a publicações científicas.

1. Introdução

1.1. Motivação e Enquadramento

O ISCTE-IUL dispõe atualmente de uma plataforma, o Ciência-IUL, que permite aos seus utilizadores introduzir a sua produção científica e pesquisar a informação acerca dos seus pares. Contudo, e apesar da quantidade e qualidade dos dados que se encontram registados, o tipo de pesquisa e informação que se pode extrair a partir da referida plataforma é, em alguns aspetos, bastante limitada. Sendo assim, a principal motivação deste trabalho é a premissa de que enriquecendo a informação disponibilizada, nomeadamente a partir da análise de redes de coautorias construídas implicitamente a partir da própria plataforma, poderá ser um meio de potenciar novas colaborações científicas, nomeadamente as de carácter multidisciplinar.

Nos últimos anos, a investigação científica desenvolvida no seio universitário tem assumido um papel cada vez mais relevante. Não nos podemos esquecer que quer a classificação em rankings internacionais, quer o apoio financeiro às unidades de investigação é atribuído em função da produção científica. Sendo assim, é crucial para o ISCTE-IUL desenvolver mecanismos que, por um lado, lhe permitam conhecer os padrões de produção e, por outro lado, promovam e incentivem essa mesma produção.

No final pretende-se disponibilizar uma interface que permita efetuar uma análise exploratória e analítica da rede de coautorias atual e das comunidades científicas identificadas na mesma. Pretende-se assim contribuir para a disponibilização de mais e melhor informação à comunidade ISCTE-IUL, tendo em vista poder potenciar as colaborações científicas entre os pares.

1.2. Objetivos

O principal objetivo principal desta dissertação é disponibilizar uma ferramenta que permita construir uma rede de coautorias a partir da informação presente no Ciência-IUL e, a partir dela, identificar e avaliar automaticamente comunidades de colaboração científica. A rede será modelada como um grafo e serão utilizadas as suas propriedades para identificar as comunidades do mesmo. A identificação de comunidades será feita recorrendo a algoritmos existentes na literatura que permitem a identificação de comunidades sem a necessidade de parametrização específica a cada tentativa de identificação. A identificação de comunidades tem de ser válida para a rede do presente, mas também para a evolução da rede no futuro, sendo que os algoritmos terão que ser

flexíveis quanto aos dados que tratam. Assim, também será possível realizar uma análise da evolução da rede e das comunidades ao longo do tempo.

Para a informação ser relevante para o ISCTE-IUL é também um objetivo a conceptualização de uma base de dados que permita persistir o grafo da rede de coautorias e os vários grafos de coautorias com comunidades identificadas. De modo a se conseguir aceder à base de dados, uma interface *web* tem de ser desenvolvida de modo a permitir efetuar uma análise exploratória e analítica, quer na rede de coautorias, quer nas comunidades identificadas pelos algoritmos. Assim, os responsáveis do ISCTE-IUL poderão visualizar as comunidades existentes e com isso podem promover a formação de novas colaborações científicas, por exemplo.

Esta dissertação procura essencialmente responder às seguintes questões:

- Como recolher, limpar e transformar a informação não orientada a grafos disponibilizada pelo Ciência-IUL de modo a construir uma rede de coautorias?
- Como modelar um grafo, com os seus vértices e arestas, de modo a este refletir as propriedades de uma rede de coautorias?
- Qual é o modelo de base de dados necessário para persistir, processar e consultar a rede construída e comunidades identificadas?
- Quais são os algoritmos que permitem identificar comunidades na rede de coautorias sem a necessidade de efetuar parametrizações a cada tentativa de identificação?
- Qual é a interface gráfica que permitirá a um membro da comunidade do ISCTE-IUL explorar e recolher a informação disponibilizada?

1.3. Contribuições

As contribuições desta dissertação são:

- A construção de uma rede de coautorias a partir dos autores e publicações do Ciência-IUL;
- A identificação de comunidades na rede de coautorias utilizando algoritmos recolhidos da literatura da área: ABCD e MCL;
- O desenvolvimento de uma aplicação, onde se inclui uma base de dados e uma interface gráfica, para construir a rede de coautorias, identificar as comunidades e analisar os resultados.

1.4. Estrutura do Documento

A dissertação encontra-se dividida em cinco capítulos. No capítulo 1 é feita a apresentação do trabalho e o seu enquadramento. Nos capítulos 2, 3 e 4 são apresentados o estado da arte, onde se descrevem as características gerais de uma rede de coautorias, os algoritmos genéricos de identificação de comunidades utilizados e o trabalho desenvolvido para a identificação de comunidades científicas, respetivamente. No capítulo 5 é apresentada a arquitetura da solução, descrevendo os seus diferentes módulos: recolha dos dados, construção da rede de coautorias, identificação de comunidades, persistência na base de dados e interface gráfica para aceder aos resultados produzidos. Também são discutidos vários aspetos sobre as decisões técnicas feitas no desenvolvimento da mesma. No capítulo 6 são apresentados os resultados obtidos, a sua avaliação e o que podemos concluir sobre os mesmos. No capítulo 7 são apresentadas as conclusões finais da dissertação, as suas limitações e o trabalho futuro.

2. Redes e Grafos

É aceite na generalidade que o artigo de Euler em 1736 sobre o problema das sete pontes de Königsberg é a primeira aplicação da teoria de grafos e marca o nascimento deste ramo da matemática (Ruohonen, 2013). É um ramo extenso, quer em bibliografia quer em conteúdo, pelo que neste capítulo irão ser introduzidos apenas os conceitos que se aplicam às redes de coautorias.

Um grafo G é um objeto composto por um conjunto V não vazio de vértices e um conjunto E, que pode ser vazio, de arestas. Em notação matemática, G = (V, E) em que $V=\{v_1,v_2,...,v_v\}$ e $E=\{(v_i,v_j): v_i,v_j \in V\}$. Na Figura 1 ilustra-se um pequeno exemplo de um grafo com 5 vértices e 5 arestas. Neste trabalho serão apenas considerados grafos finitos, isto é, em que os conjuntos V e E são também finitos. O grau de um vértice é medido como o total de arestas que se ligam ao mesmo; um grafo diz-se regular se todos os vértices tiverem o mesmo grau.

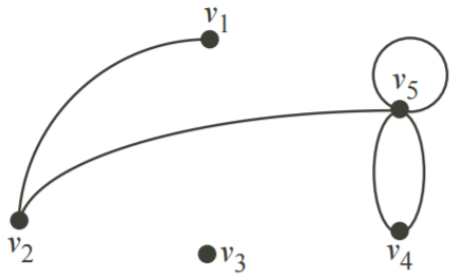

Figura 1 - Um grafo G com V = {v1, v2, v3, v4, v5} e E = {(v1, v2), (v2,v5),(v5,v5),(v5,v4),(v4,v5)} (Ruohonen, 2013)

Um grafo que não contenha múltiplas arestas entre qualquer par de vértices nem lacetes, ou seja, uma aresta de um vértice para si próprio (ver Figura 1, vértice v_5), diz-se um grafo simples. Um grafo simples que contem todas as arestas possíveis entre os seus vértices é um grafo completo e tem a notação de K_n. Adicionalmente, um grafo G1 = (V1, E1) designa-se por subgrafo de um grafo G2 = (V2, E2) se V1 $\subseteq$ V2 e $E1=\{(v_i,v_j) \in E2: v_i, v_j \in V1\}$.

Os grafos podem ser direcionados ou não direcionados. Num grafo direcionado as arestas são substituídas por arcos que indicam que a ligação entre os vértices tem uma direção específica. Neste trabalho serão apenas considerados grafos não direcionados.

O caminho de um vértice v para um vértice u num grafo G = (V, E) é uma sequência de arestas que se podem definir como um conjunto de valores: $\{v, v_1\}$, $\{v_1, v_2\}$, …, $\{v_{k-1}, v_k\}$, $\{v_k, u\}$. Se existir tal caminho, dizemos que os vértices u e v estão ligados. (Schaeffer, 2007).

A densidade de um grafo G = (V,E) mede a proporção de arestas que estão no conjunto E face ao número máximo possível de arestas entre os vértices do conjunto V. Um grafo não direcionado pode ter no máximo $|V| * (|V| - 1) / 2$ arestas e, portanto, a sua densidade é dada por $2 * |E| / (|V| * (|V| - 1))$. (Ruohonen, 2013)

Quer os vértices, quer as arestas, podem ter pesos associados, que traduzem características inerentes ao grafo em estudo. Quando tal acontece, os grafos tomam a designação de redes.

2.1. Redes de Colaboração Científica

Em termos genéricos, designa-se por rede de coautoria uma rede que traduza as colaborações científicas entre um conjunto de autores. Assim, o estudo destas redes de coautoria permite analisar os padrões de partilha de conhecimento dentro duma comunidade académica. O tipo de colaborações científicas, por exemplo, artigos científicos, livros ou projetos, bem como os autores a serem considerados, varia com o objeto de estudo. Normalmente, e de acordo com a nomenclatura da teoria de grafos, considera-se um autor um vértice e a colaboração entre dois autores uma aresta. Uma aresta pode representar mais que uma colaboração, pelo que o peso associado à aresta é a propriedade que quantifica a colaboração entre autores em termos de atratividade.

A partir de 1990 as potencialidades destas redes começaram a ser estudadas por vários autores, principalmente ao nível da análise estatística, e permitiram identificar, por exemplo, qual a frequência de artigos coautorados. Mas foi apenas a partir de 2000, com o aparecimento de bases de dados *online* de colaboração científica como o Ciência-IUL, que as construções de redes de coautoria se tornaram possíveis (Newman, 2004), criando como consequência representações visuais enriquecidas como a da Figura 2.

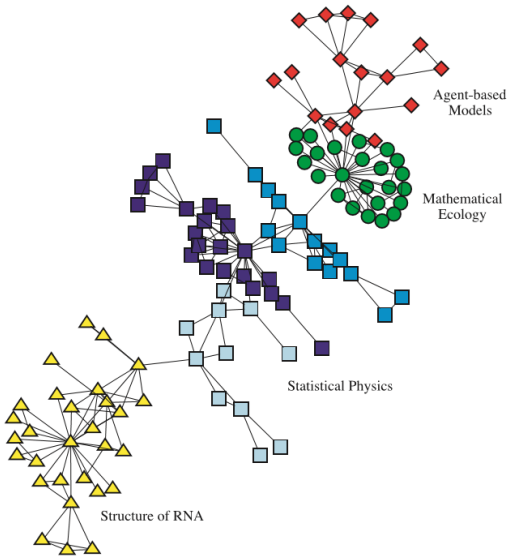

Figura 2 – Rede de coautorias de uma instituição privada com várias comunidades distintas (Newman, Coauthorship networks and patterns of scientific collaboration, 2004)

Na análise de redes de coautoria podem ser utilizadas para além da representação gráfica um conjunto de medidas quantitativas. Na Figura 2 ilustra-se uma representação gráfica de uma rede de coautorias em que foram utilizados diferentes formatos geométricos nos vértices para distinguir os agrupamentos de coautorias em quatro diferentes áreas científicas. Como se pode observar, apenas uma pequena percentagem de autores tem o número mais elevado de colaborações. Designando por grau de um vértice o número de arestas que contêm esse vértice, este facto traduz que a distribuição de grau dos vértices será provavelmente *fat-tailed*, ou seja, apenas uma pequena percentagem de autores tem o maior número de colaborações.

A distância entre dois vértices é o número mínimo de arestas que separam os dois e é uma das medidas quantitativas utilizada no estudo destas redes. Por exemplo, dois autores estarão a uma distância de 1 se tiverem coautorado um artigo, a uma distância de 2 se não tiverem coautorado um artigo, mas partilharem os dois um autor com quem tenham coautorado um artigo. Observando a Figura 2 verifica-se que dentro de cada área científica as distâncias entre os autores tendem a ser mais reduzidas se comparadas com a distância a autores de outras áreas científicas.

A centralidade de um autor é outra das medidas utilizada na análise de uma rede de coautorias. Existem diferentes medidas de centralidade, mas todas elas expressam a influência que um autor tem na rede em que está inserido. Por exemplo, a centralidade de intermediação de um autor A é calculada com base no número de caminhos mais curtos

entre autores que passam por A. Identificam-se assim autores que servem para unir partes distintas da rede que, de outra forma, não comunicariam.

3. Identificação de Comunidades

O termo "comunidade" apareceu pela primeira vez em 1887 no livro "Gemeinschaft und Gesellchaft" de Ferdinand Tönnies (Deepjyoti Choudhury, 2013). Ferdinand Tönnies elaborou que cada indivíduo com os seus valores e crenças se liga a outros que partilham esses mesmos valores e crenças, criando assim uma comunidade. A comunidade pode assim ser entendida como uma construção humana que interage e tem objetivos comuns, sejam eles políticos, económicos ou simplesmente lúdicos. Neste sentido, o ISCTE-IUL pode ser considerado como sendo uma comunidade, a comunidade ISCTE-IUL.

A partir do repositório do Ciência-IUL é possível recolher os autores pertencentes à comunidade ISCTE-IUL que tenham, pelo menos, uma colaboração científica com outro autor da comunidade. Por colaboração científica entende-se uma coautoria em publicações como artigos, revistas, capítulos de livros, entre outros. A representação matemática deste conjunto de autores bem como das relações entre si, que traduzem as suas colaborações cientificas, será aqui designada por rede de coautorias. A identificação de comunidades é feita a partir da análise da estrutura desta rede de coautorias. As comunidades serão assim subconjuntos de autores que estejam mais relacionados com autores da mesma comunidade do que com autores de outras comunidades. Todas estas comunidades pertencem à comunidade geral ISCTE-IUL.

A deteção de comunidades em redes tem sido um problema vastamente estudado e que, em teoria de grafos, é designado por *graph clustering problem* (Ruifang Liua, 2014), ou seja, um problema de construção de agrupamentos num grafo, em que cada agrupamento é constituído por um subconjunto de vértices, e respetivas arestas. A determinação da pertença dos vértices aos agrupamentos é, em geral, feita com base na otimização duma determinada função que traduza o conceito de comunidade como sendo um subconjunto de vértices mais ou melhores conectados com vértices do mesmo agrupamento do que com vértices de outros agrupamentos (Figura 3).

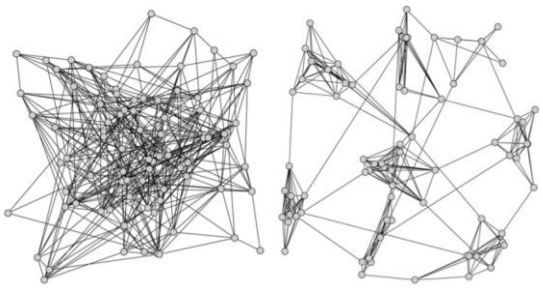

Figura 3 - Um grafo (esquerda) e seus agrupamentos (direita) (Schaeffer, 2007)

A abordagem mais direta para obter agrupamentos num grafo é obter uma divisão dos vértices em grupos que minimizem o número de arestas entre esses mesmos grupos, uma abordagem conhecida como *minimum cut* (Newman, 2006). Contudo, esta abordagem é bastante limitativa pois baseia-se na estrutura do grafo sem ter em consideração as propriedades do mesmo.

Um outro conceito utilizado em algoritmos para identificar agrupamentos é o de caminho. Note-se que não existe nenhuma definição do que constitui um agrupamento num grafo que seja universalmente aceite (Schaeffer, 2007). No entanto, existe um entendimento que cada agrupamento tem de estar intuitivamente ligado: devem existir vários caminhos a ligar um par de vértices (u, v) que esteja dentro do mesmo agrupamento. Estes caminhos são constituídos por arestas internas (ao agrupamento), ou seja, arestas que ligam vértices dentro do agrupamento, em contraposição às arestas externas que ligam um vértice do agrupamento com um vértice que não pertence ao agrupamento. De realçar que u e v não têm necessariamente de estar diretamente ligados por uma aresta para se considerar que estão dentro do mesmo agrupamento.

Um dos algoritmos propostos para identificar agrupamentos, e que assenta no conceito de caminho é o *k-clique* (Schaeffer, 2007). Neste algoritmo, considera-se que dois vértices (u, v) estão dentro do mesmo agrupamento se o valor do caminho mais curto entre u e v é menor ou igual a *k*, em que o valor de um caminho representa o número de arestas ou, em caso de grafos com pesos nas arestas, a soma dos pesos das arestas que pertencem ao caminho. De modo a definir o valor de *k* deve-se ter em conta o diâmetro do grafo a ser estudado, isto é, a máxima distância possível entre os vértices. Se *k* for um valor muito próximo ao do diâmetro serão criados agrupamentos muito grandes, praticamente do tamanho do grafo. Por outro lado, se *k* for um valor muito pequeno existirá a tendência para que um agrupamento natural acabe por ser dividido em outros agrupamentos mais pequenos. Uma vez que requer a parametrização de *k*, algo que se pretende evitar, este algoritmo não foi utilizado.

Estes são apenas dois exemplos de algoritmos para deteção de comunidades. Neste estudo, não só porque se pretendia evitar o recurso a algoritmos que exigissem parametrizações manuais, mas também porque a rede de coautorias é uma rede com pesos nas arestas e, por isso, essa informação deve contribuir para a definição das comunidades, optámos por utilizar algoritmos especificamente desenhados para redes com pesos nas arestas: o *Attractiveness-based community detection* (ABCD) e o *Markov Cluster*

Algorithm (MCL). Note-se que estes algoritmos consideram que cada vértice pode ser incluído numa única comunidade, ou seja, cada autor não pode pertencer a várias comunidades.

Antes de passarmos à descrição destes algoritmos convém esclarecer como vamos quantificar a relação de proximidade entre os vértices da rede de coautorias, nomeadamente através da atribuição de um peso às mesmas. Assim, o peso de cada aresta, P_e, será calculado através da equação de Newman (Newman, 2004)

$$P_e = \sum_{i=0}^{k} \frac{1}{(ni-1)} \text{ (Equação 1)}$$

em que *k* representa o número de publicações coautoradas por um mesmo par de autores e *ni* o número total de autores dessas mesmas publicações.

A equação de Newman traduz que cada artigo coautorado adiciona à relação entre dois autores o fator 1/(n–1). A lógica subjacente é que se uma coautoria envolve *n* autores, então cada autor divide o seu tempo com os restantes n-1 autores do artigo e, por isso, a força da colaboração varia inversamente a n-1. Por exemplo, uma coautoria constituída por uma publicação em que dez autores participaram não deve ser "tão forte" quanto uma em que apenas dois colaboraram. Na Figura 4 apresenta-se a representação gráfica dum grafo em que o peso das arestas foi calculado de acordo com a equação de Newman. O peso das arestas é traduzido a partir da largura das arestas: quanto maior for a largura mais forte é a relação de coautoria.

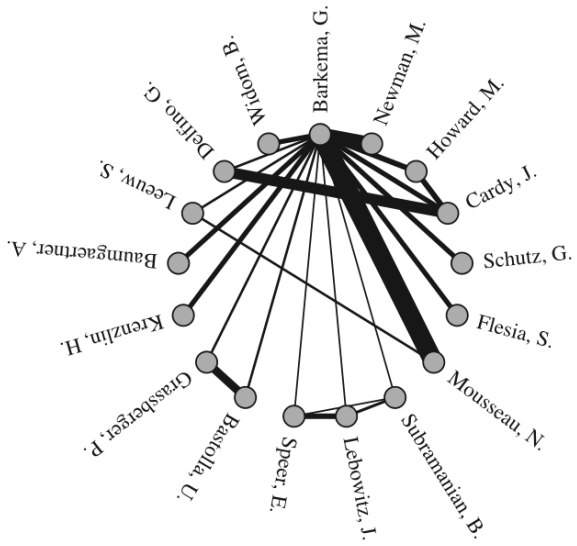

Figura 4 – Grafo de coautorias organizado segundo a equação de Newman (Newman, 2004)

Ao longo deste documento, o termo peso de uma aresta será muitas vezes referenciado por força de atratividade de uma coautoria.

3.1. Algoritmo - Attractiveness-based community detection

O *attractiveness-based community detection* - ABCD (Ruifang Liua, 2014) é um algoritmo para deteção de comunidades em grafos com pesos, ou seja, grafos em que os vértices e/ou as arestas têm um peso (valor numérico) associado. Assenta em dois conceitos: densidade de um agrupamento e atratividade entre agrupamentos A densidade de um agrupamento i, Wi, é dado por

$$W_i = \frac{\sum_{a=1}^{Q_i} W_a}{Q_i} \quad \text{(Equação 2)}$$

em que Q_i representa o número de vértices do agrupamento i e Wa o peso de cada vértice que pertence ao agrupamento. Note-se que a Equação 2 traduz que a densidade de um agrupamento i, W_i, não é mais que o peso médio dos vértices que a ele pertencem.

A atratividade entre dois agrupamentos i e j, S_{ij}, é definida por

$$S_{ij} = \frac{\sum_{e=1}^{q} Pe}{Q_i \times Q_j} \quad \text{(Equação 3)}$$

em que q representa o número de arestas que ligam vértices do agrupamento i com vértices do agrupamento j, e P_e o peso associado a essas arestas.

É de notar que $Q_i \times Q_j$ representa o número máximo de arestas que podem existir entre os vértices dos agrupamentos i e j, e não os que existem. Desta forma, a atratividade S_{ij} não traduz exatamente o peso médio das arestas entre os agrupamentos i e j. Mais ainda, o seu valor tende a decrescer à medida que o número de vértices por agrupamento aumenta.

Este algoritmo foi aplicado com sucesso a conjuntos de dados da plataforma de *blogging* Sina Weibo e da rede social Renren, ambas de origem chinesa. É um algoritmo típico de aglomeração sucessiva, ou seja, que vai agrupando vértices em comunidades até ser atingida uma condição de paragem. No início do algoritmo cada vértice constitui um agrupamento, ou seja, existem tantos agrupamentos quantos os vértices. Depois, em cada iteração, são escolhidos os dois agrupamentos i e j com maior atratividade (com maior valor de S_{ij}) e para os quais se verifiquem as seguintes condições:

1. $q \geq Q_i$ e $q \geq Q_j$, para todo o $j \neq i$
2. $S_{ij} \geq W_i + W_j$

Note-se que sempre que ocorre um empate, o par de agrupamentos a agrupar é escolhido aleatoriamente. O algoritmo termina quando não existir nenhum par de agrupamentos que verifique as duas condições anteriormente enumeradas.

Quer a Equação 2, quer a Equação 3, requerem pesos associados às arestas e aos vértices. Neste trabalho considerou-se que o peso das arestas é dado pela equação de Newman (Equação 1). O peso de cada vértice (autor) é a média dos pesos das arestas que o contêm. Desta forma, estabelece-se uma relação direta entre os autores e a atratividade das suas coautorias.

A partir das condições 1) e 2) anteriormente referidas, é possível concluir que o algoritmo ABCD procura promover comunidades densas de vértices. A condição 1) requer que o número de arestas entre os agrupamentos a unir seja, pelo menos, igual ao número de vértices do agrupamento de maior dimensão – e que tenham uma grande atratividade entre eles. A condição 2) impõe que a união de dois agrupamentos i e j só pode ocorrer se a atratividade entre os mesmos, calculada a partir de um peso ponderado das arestas que os unem, tenha um valor, pelo menos, igual à soma do peso médio dos vértices pertences ao agrupamento i com o peso médio dos vértices pertencentes ao agrupamento j.

No caso em estudo e tendo em conta que o peso dos vértices representa o peso médio das arestas que o contêm, verificou-se que a condição 2 se tornava muito restritiva, impedindo a formação de comunidades. Para ultrapassar esta limitação, foram introduzidas duas alterações ao algoritmo. A primeira consistiu em alterar a expressão de cálculo da atratividade entre dois agrupamentos i e j para:

$$S_{ij} = \sum_{e=1}^{q} P_e \quad \text{(Equação 4)}$$

A segunda consistiu em alterar a condição 2) anteriormente referida pela condição:

$$S_{ij} \geq \frac{W_i + W_j}{2}$$

Desta forma, a atratividade entre agrupamentos passa a ser quantificada recorrendo apenas ao peso das arestas entre os agrupamentos. Para além disso, a alteração da condição 2 irá permitir que a união de agrupamentos se possa efetuar desde que o peso das arestas entre os mesmos seja, pelo menos, igual ao peso médio dos vértices desses agrupamentos.

3.2. Algoritmo – Markov Cluster Algorithm

O algoritmo de agrupamentos de Markov, Markov Cluster Algorithm (MCL), baseia-se no princípio que um agrupamento terá muitas ligações entre si e poucas ligações com outros agrupamentos (Dongen, 2000). Isto significa que dados dois vértices u e v pertencentes a um mesmo agrupamento, a probabilidade de que o caminho de u a v contenha vértices externos ao agrupamento deve ser baixa. Em consequência, num passeio aleatório de um qualquer vértice para um vértice do mesmo agrupamento existe maior probabilidade de percorrer arestas dentro do agrupamento do que arestas externas ao agrupamento. O MCL assenta assim no conceito de passeios aleatórios num grafo.

Os passeios aleatórios num grafo podem ser descritos por meio de cadeias de Markov em que a sequência de variáveis na cadeia é representada por uma sequência de matrizes de transição de probabilidade, $M(k)$, $k \geq 0$.

Consideremos um grafo com n vértices e seja $M(0)$ a matriz de transição de probabilidades de ordem 0. Note-se que esta matriz é uma matriz quadrada de ordem n, ou seja, com nxn elementos. Nesta matriz inicial de transição de probabilidades, $M(0)$, cada entrada (i,j) irá conter o valor da probabilidade de, num passeio aleatório, sair do vértice j e chegar ao vértice i passando por 0 vértices intermédios. Consequentemente, cada coluna j da matriz $M(0)$ irá conter as probabilidades associadas a passeios aleatórios que se iniciem no vértice j e que não passem por qualquer vértice intermédio.

Esta matriz inicial é construída a partir da matriz de adjacência: matriz que em cada posição (i,j) contem o valor 1 se a aresta (i,j) pertencer ao grafo e 0 caso contrário; caso às arestas estejam associados pesos o valor 1 é substituído pelo respetivo peso.

Dado que a matriz de transição de probabilidades tem que ser estocástica por coluna, ou seja, a soma dos valores de qualquer coluna tem que perfazer o valor 1, a matriz de adjacência tem que ser normalizada por coluna: somam-se todos os elementos da coluna e depois divide-se cada um deles pelo valor da soma. Note-se que a exigência de que matriz de transição de probabilidades tenha que ser estocástica por coluna advém do facto de que cada coluna j contem as probabilidades de, num passeio aleatório, sair do vértice j, e logo a sua soma tem que ser 1.

Uma vez determinada a matriz inicial, $M(0)$, o algoritmo prossegue de forma iterativa até que o critério de paragem se verifique. O critério de paragem assenta na convergência, ou

seja, até que não existam diferenças significativas entre duas consecutivas matrizes de transição.

Em cada iteração $k \geq 1$, uma nova matriz M(k) é calculada de forma determinística usando para o efeito dois operadores: *Expansion* e *Inflation*.

O operador *Expansion* eleva a matriz à potência p, em que p é um parâmetro dado. O efeito deste operador é alterar as probabilidades de transição de forma a que as mesmas reflitam a introdução de (p-1) vértices intermédios no passeio aleatório de qualquer vértice j para qualquer vértice i. Por exemplo, se à matriz M(0) for aplicada este operador com p=2, significaria que cada entrada (i,j) na nova matriz iria representar a probabilidade de num passeio aleatório de j para i passar por um vértice intermédio. Note-se que este vértice intermédio pode ser o próprio i ou o próprio j.

O operador *Inflation*, por sua vez, é aplicado às colunas das matrizes M: eleva à potência r todos as entradas (i,j) da matriz M e depois normaliza, por coluna, os valores obtidos para que a matriz resultante seja estocástica por coluna. Se r>1, o efeito deste operador será o de alterar as probabilidades associadas aos passeios aleatórios que se iniciam num determinado vértice j (representados na coluna j da matriz) de modo a favorecer os passeios mais prováveis (ao elevar à potência r>1, aumenta a diferença entre as probabilidades mais altas e mais baixas). Desta forma, reforça a "força de atração" aos vértices vizinhos "mais fortes", diminuindo-a aos vértices vizinhos "menos fortes".

O autor que primeiro sugeriu a aplicação deste algoritmo aos grafos em 2000, Stijn van Dongen, disponibiliza uma ferramenta com licença *GNU General Public License*, que o próprio desenvolveu (Dongen, MCL - a cluster algorithm for graphs, 2017). A mesma foi utilizada para aplicar aos grafos de coautorias desta dissertação, tendo-se utilizado o operador de *Expansion* com p=2 e o operador de *Inflation* com r=2.

3.3. Denominação e Caracterização das Comunidades

Tendo em conta o objeto de estudo deste trabalho, torna-se importante caracterizar as comunidades, ou seja, disponibilizar alguma informação que permita inferir, se possível, algumas dinâmicas subjacentes à sua formação. Torna-se necessário então definir uma metodologia para o efeito. Para caracterizar as comunidades, é disponibilizada a seguinte informação:

- O nome da comunidade;

- O nome, a escola, departamento e/ou centro de investigação de cada elemento da comunidade;
- O número total autores, de coautorias e de publicações científicas na comunidade;
- O número de autores por escola, por departamento e por centro de investigação;
- As categorias científicas das publicações.

Uma outra questão adicional que se coloca é como denominar cada uma das comunidades detetadas de forma automática. Uma das formas possíveis para atribuir um nome a uma comunidade científica, não necessariamente de coautorias, é através de metainformação associada às publicações associadas aos elementos que as compõem. Por exemplo, num estudo sobre deteção de comunidades numa rede de conferências, ou seja, uma rede em que os vértices representam conferências e as arestas o número de autores que publicaram em ambas as conferências, Cervera (Cervera, 2010) utilizou o nome das *k* conferências mais participadas pelos membros de uma determinada comunidade - sendo *k* um valor configurável - para denominar as comunidades. O autor defende que uma conferência assistida por vários autores corresponde em si mesmo à partilha de uma comunidade, i.e., um tópico de interesse.

No caso da rede de coautorias em estudo, os autores presentes no Ciência-IUL têm todos, pelo menos, um departamento e/ou um centro de investigação associado. O nome atribuído a uma comunidade é uma consequência da análise e processamento dessa informação: a designação da comunidade resulta da concatenação, através do carácter &, dos três departamentos e/ou centros de investigação que reúnem um maior número de autores na comunidade. Devido ao método genérico de atribuição de nome, embora pouco frequente, é possível que comunidades diferentes tenham o mesmo nome, facto que se aceita uma vez que a lista de autores será necessariamente diferente, sendo ainda assim facilmente diferenciáveis.

Em relação às categorias científicas, convém realçar que o Ciência-IUL possui informação acerca das categorias científicas atribuídas às publicações pelo *Shape of Science* da *Scimago Journal* (SCImago, 2016). Esta informação é disponibilizada pela administração do Ciência-IUL uma vez que a utilização da mesma se encontra licenciada pela *Scimago*. Esta caracterização adicional, a partir das categorias científicas das publicações, permite identificar as áreas científicas dominantes em cada uma das comunidades identificadas na rede de coautorias. A sua análise poderá, por exemplo, ser

uma forma de dar a conhecer e, se possível, dinamizar colaborações científicas de âmbito multidisciplinar.

4. Ferramentas para Análise de Redes Sociais

4.1. Socializador-IUL

Em 2016 foi submetida por Diogo Pinheiro, no âmbito de uma dissertação de mestrado, uma proposta de como identificar comunidades no ISCTE-IUL (Pinheiro, 2016). O seu trabalho permitiu concluir que apenas o Ciência-IUL apresenta uma fonte de dados fiável para a identificação de comunidades dentro do ISCTE-IUL. A informação da mesma encontra-se disponível através de uma API REST (Ciência-IUL, 2016), sendo o objeto Autor o conceito mais importante para a construção da rede de coautorias (Figura 5). O objeto Autor possui as publicações do mesmo, cada uma com um identificador único e é com base neste conceito que poderemos criar o vértice da coautoria, relacionando autores através das publicações que partilham entre si.

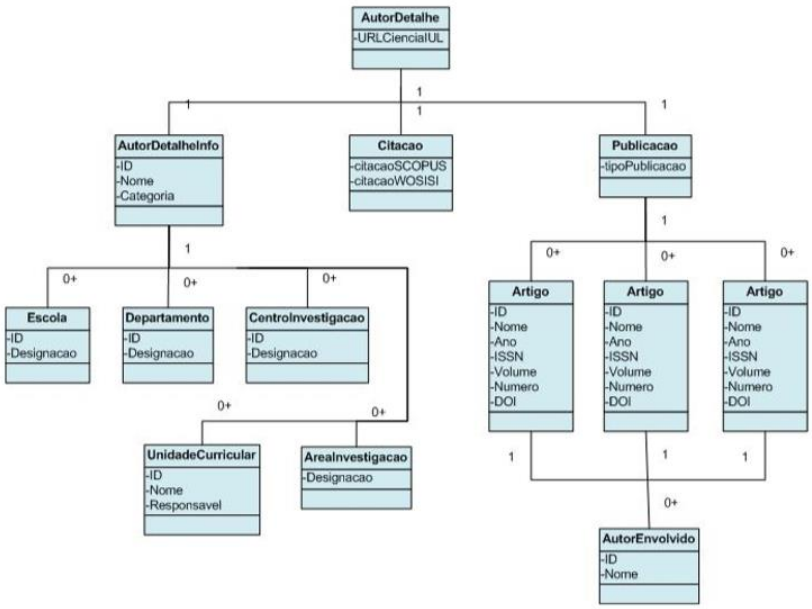

Figura 5 - Mapeamento UML do objeto Autor extraído do Ciência-IUL (Pinheiro, 2016)

Nesse trabalho foi criada uma rede simples com os dados recolhidos, em que as arestas ligam os autores com, pelo menos, uma colaboração entre eles. No entanto, a rede apresenta uma limitação indicada pelo autor: "*Sendo assim, o número de ligações pode diferir do número real de colaborações, bastando para tal que entre o mesmo par de autores exista mais que uma colaboração científica do mesmo tipo*." (Pinheiro, 2016) Portanto, um ponto a explorar é a colocação de pesos nas arestas que ligam os autores de modo a refletir o número de colaborações e não apenas se existe, ou não, uma aresta. É também de realçar que os autores externos ao ISCTE-IUL não são uma mais valia para a

identificação das comunidades da instituição: "*os autores externos não possuem informações relevantes, tal como afiliação de escola, departamento, etc., e como tal não trazem grande vantagem ao serem incluídos na rede*." (Pinheiro, 2016)

Foi utilizado o software *Gephi* para detetar as comunidades da rede simples de autores, através do método de *Louvain*, desenvolvido por Vincent Blondel (Pinheiro, 2016). A utilização de um software dedicado facilitou a construção gráfica e a análise das redes de coautorias, mas limita a navegação, consulta e validação das redes por parte da comunidade ISCTE-IUL. Por outro lado, o método de *Louvain* utiliza o conceito de modularidade do grafo que, por sua vez, requer a parametrização de um limite de resolução. Foram utilizados vários valores para o mesmo, implicando que as comunidades tenham sido identificadas por um método de tentativa-erro desse parâmetro, necessariamente com intervenção humana, em vez de um modo automático usando um algoritmo que não necessite de parametrizações adicionais.

O processamento e tratamento da informação foi feita sem persistência em base de dados, o que inviabiliza a avaliação temporal da evolução da rede e das comunidades identificadas. Esta limitação também não permite adicionar novos algoritmos de identificação de comunidades uma vez que os dados foram preparados para serem processados pelo algoritmo utilizado e não persistidos num grafo genérico e pronto a ser utilizado por outros algoritmos.

4.2. Interface Gráfica - MIT Media Lab

Existem vários tipos de interfaces computacionais criados para visualizar grafos. Considerando a abertura e disponibilidade da internet, este trabalho focou-se nas interfaces *web* criadas para representar grafos e suas propriedades, recorrendo às tecnologias HTML, CSS e JavaScript. Existe uma distinção importante a fazer entre as bibliotecas em código aberto, que podemos utilizar para criar os grafos, e os projetos em código fechado, em que não podemos aceder e utilizar o código, mas podemos visualizar e analisar a interface sem restrições, servindo de inspiração conceptual.

Um dos projetos feitos com as tecnologias mais avançadas e interessantes foi o *Clinton Circle* (Lab, 2016), que permite aceder às relações entre as personalidades do Partido Democrático dos EUA tendo em conta os emails trocados pelos mesmos. Foi realizado pelo *MIT Media Lab*, um grupo de especialistas tecnológicos da *Massachusetts Institute of Technology* focado neste tipo de interfaces. O código da solução não está disponível,

mas podemos estabelecer um paralelismo com a rede de coautorias em que também temos pessoas, neste caso autores, como vértices, sendo que as arestas não correspondem a emails trocados, mas a publicações partilhadas.

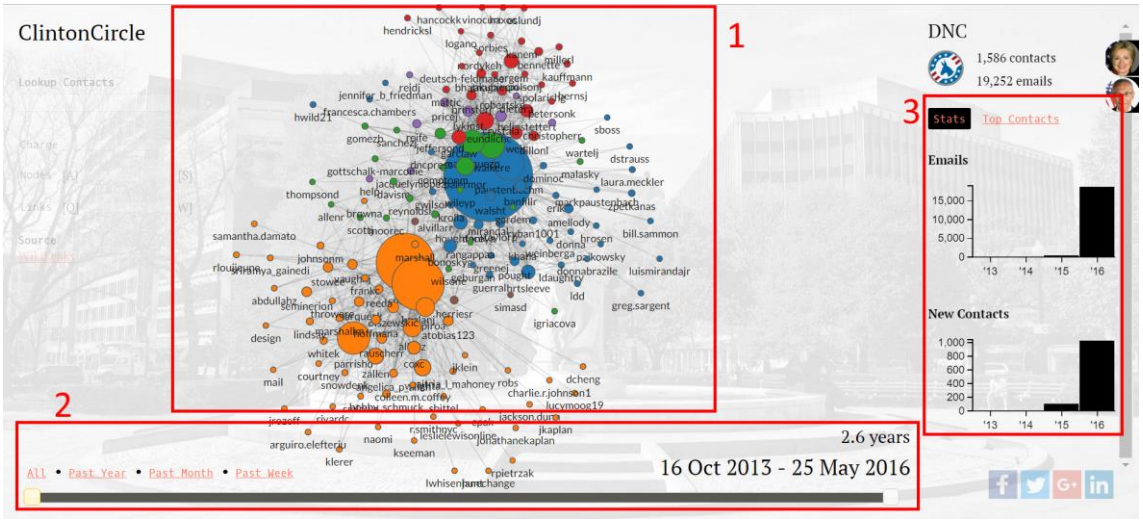

Figura 6 - Interface do grafo e suas comunidades do projeto *Clinton Circle* (Lab, 2016)

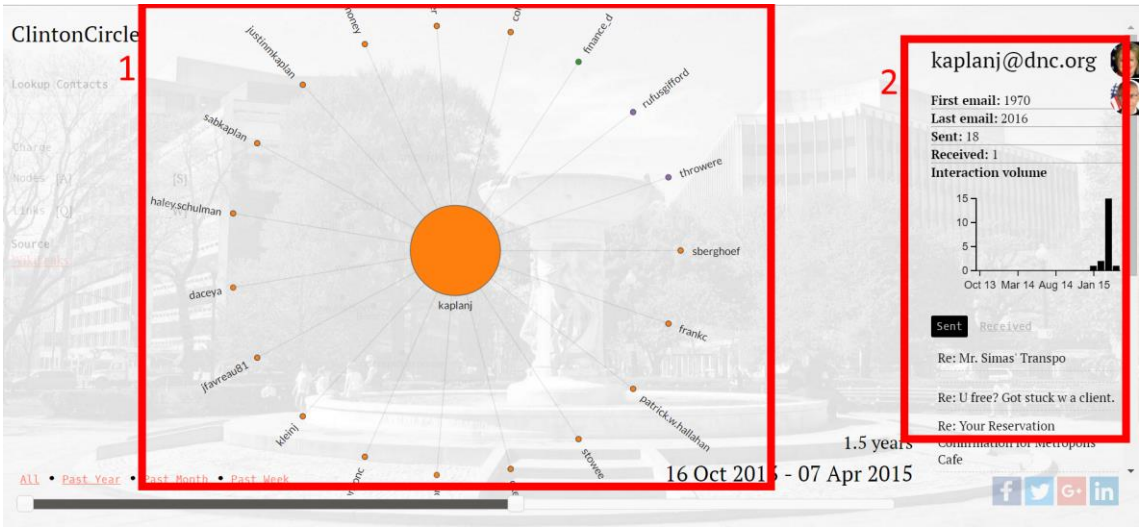

Figura 7 - Interface do projeto *Clinton Circle* quando selecionado um vértice em particular (Lab, 2016)

Na visão geral do grafo, Figura 6, existem três componentes essenciais para interpretar o grafo. Na secção 1 temos a sua representação gráfica, com os vértices a representar elementos do partido Democrático e as arestas indicando a troca de emails. De realçar que a cor de um determinado vértice marca a comunidade a que o mesmo pertence. Neste exemplo existem vértices com diferentes tamanhos pois o tamanho é proporcional ao total de emails enviados pelo vértice em questão. Na secção 2 é possível ajustar a rede ao intervalo temporal onde se pretende pesquisar. Na secção 3 pode-se ver uma apresentação de estatísticas gerais da rede; no caso das redes de coautoria serão representados o número

total de autores, de coautorias, de publicações e o número de autores por escola, departamento e centro de investigação.

A Figura 7, que decorre da seleção de um vértice em particular, poderá ser adaptada para se analisar um autor em detalhe. Na secção 1, o autor em análise ficará no centro da interface, com os autores com quem fez coautorias à sua volta. Os seus coautores têm a cor da comunidade a que pertencem pelo que conseguimos identificar as comunidades específicas a que cada um pertence. Na seção 2 poderá existir uma seção de estatísticas gerais do autor, como o número de publicações, a distribuição das publicações por datas, etc.

Um dos projetos desenvolvidos em tecnologias *web* que também é importante salientar é o *Shape of Science* da *Scimago Journal* (SCImago, 2016). O mesmo representa uma rede de colaborações científicas entre áreas tendo em conta os artigos presente na base de dados da mesma. A diferença é que neste caso os vértices correspondem a áreas científicas e não autores. Contudo, a abordagem foi semelhante ao *Clinton Circle* (Lab, 2016), o que demonstra que a abordagem de uma interface *web* funciona também em projetos de coautorias científicas.

5. Solução Proposta

De modo a permitir a construção de uma rede de coautorias e a identificação de comunidades nessa rede foi concebida uma arquitetura baseada em quatros módulos: construção do grafo simples, identificação de comunidades, base de dados e *website* (Figura 8). A natureza modular da solução permite a substituição futura de qualquer um dos módulos sem prejuízo de continuar a utilizar os restantes módulos. Por exemplo, poderá ser adicionado um novo algoritmo de identificação de comunidades que utiliza o grafo persistido na base de dados sem qualquer alteração aos módulos responsáveis pelos mesmos.

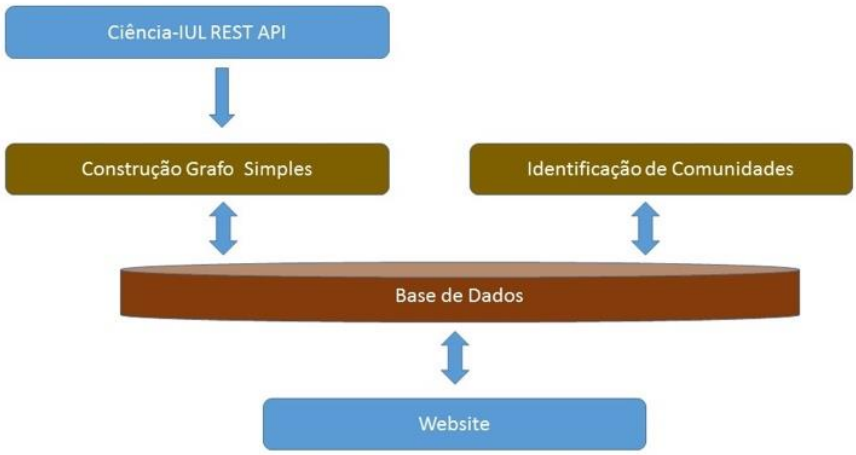

Figura 8 - Arquitetura da solução

No fluxo normal da aplicação, em primeiro lugar os dados são recolhidos através de uma REST API disponibilizada pelo Ciência-IUL para serem transformados num grafo com vértices de autores e arestas de coautorias entre estes. Depois de construído o grafo - no fundo uma rede simples sem comunidades identificadas - o mesmo é persistido na base de dados, podendo depois ser acedido pela interface gráfica, o *Website*. Após a construção do grafo simples, podem ser aplicados os algoritmos de identificação de comunidades, sendo os resultados dos mesmos também armazenados na mesma base de dados e acedidos pelo mesmo *Website*.

A componente de Identificação de Comunidades é onde os algoritmos de identificação de comunidades são executados. Tendo em consideração que é um componente à parte, que terá a sua própria calendarização e lógica, facilmente se pode adicionar ou retirar algoritmos à solução. Foram implementados dois algoritmos, o ABCD e o MCL. No entanto, a solução não fica necessariamente limitada pelos mesmos; podem ser

adicionados e executados novos algoritmos, o que permitirá comparar os resultados dos novos algoritmos com os já implementados nesta solução.

A interface gráfica da solução (*Website* na Figura 8) permite visualizar o grafo produzido, assim como as comunidades identificadas num navegador *web* comercial. As funcionalidades essenciais do mesmo são a visualização do grafo com a rede de coautorias, a visualização de um grafo com várias comunidades identificadas (todos os vértices da mesma comunidade são da mesma cor), a capacidade de consultar os detalhes do autor acedendo ao vértice que o representa e um conjunto de métricas descritivas gerais (números de autores, número de publicações, número de coautorias, etc).

5.1. Recolha de Dados

A informação guardada no Ciência-IUL encontra-se disponível através de uma API pública implementada com *Web Services* REST (Ciência-IUL, 2016). Estes serviços utilizam como protocolo de comunicação o HTTP e são endereçados através de um URL único. Não guardam o estado do lado do servidor e os recursos que devolvem estão relacionados com outros recursos que apenas são obtidos após novas chamadas a outros serviços (Leonard Richardson, 2007). A API REST do Ciência-IUL devolve os recursos no formato JSON, um formato de fácil leitura para humanos e de fácil processamento para máquinas digitais.[1]

A API REST do Ciência-IUL é extensa e feita a pensar em vários cenários de utilização da informação. No caso da construção de rede de coautorias o objetivo é chegar aos autores e suas publicações de modo a construir a rede de coautorias, extraindo apenas as propriedades necessárias. Os autores encontram-se organizados em três conceitos funcionais: escolas, departamentos e unidades de investigação. A API do Ciência-IUL associa a cada autor um único departamento, uma única escola (pode estar omisso) e uma ou mais unidades de investigação. Para se obter a listagem completa de autores é, pois, necessário consumir os endereços de cada departamento e cada unidade de investigação no Ciência-IUL (Tabela 1).

	URL
Departamento	
Departamento de Antropologia	https://ciencia.iscte-iul.pt/api/department/DA

[1] O desenvolvimento da aplicação não seria possível sem o apoio do professor António Luís Lopes, Coordenador do Gabinete de Desenvolvimento de Sistemas de Informação do ISCTE-IUL.

Departamento de Arquitetura e Urbanismo	https://ciencia.iscte-iul.pt/api/department/DAU
Departamento de Ciência Política e Políticas Públicas	https://ciencia.iscte-iul.pt/api/department/DCPPP
Departamento de Ciências e Tecnologias da Informação	https://ciencia.iscte-iul.pt/api/department/DCTI
Departamento de Contabilidade	https://ciencia.iscte-iul.pt/api/department/DC
Departamento de Economia	https://ciencia.iscte-iul.pt/api/department/DE
Departamento de Economia Política	https://ciencia.iscte-iul.pt/api/department/DEP
Departamento de Finanças	https://ciencia.iscte-iul.pt/api/department/DF
Departamento de História	https://ciencia.iscte-iul.pt/api/department/DH
Departamento de Marketing, Operações e Gestão Geral	https://ciencia.iscte-iul.pt/api/department/DMOG
Departamento de Matemática	https://ciencia.iscte-iul.pt/api/department/DM
Departamento de Métodos de Pesquisa Social	https://ciencia.iscte-iul.pt/api/department/DMPS
Departamento de Métodos Quantitativos para Gestão e Economia	https://ciencia.iscte-iul.pt/api/department/DMQGE
Departamento de Psicologia Social e das Organizações	https://ciencia.iscte-iul.pt/api/department/DPSO
Departamento de Recursos Humanos e Comportamento Organizacional	https://ciencia.iscte-iul.pt/api/department/DRHCO
Departamento de Sociologia	https://ciencia.iscte-iul.pt/api/department/DS
Unidade de Investigação	
CEI-IUL - Centro de Estudos Internacionais	https://ciencia.iscte-iul.pt/api/centre/CEI-IUL
CIES-IUL - Centro de Investigação e Estudos de Sociologia	https://ciencia.iscte-iul.pt/api/centre/CIES
CIS-IUL - Centro de Investigação e de Intervenção Social	https://ciencia.iscte-iul.pt/api/centre/CIS
CRIA-IUL - Pólo do ISCTE-IUL do Centro em Rede de Investigação em Antropologia	https://ciencia.iscte-iul.pt/api/centre/CRIA
DINÂMIA'CET-IUL - Centro de Estudos sobre a Mudança Socioeconómica e o Território	https://ciencia.iscte-iul.pt/api/centre/DINAMIA
Instituto de Telecomunicações-IUL	https://ciencia.iscte-iul.pt/api/centre/IT-IUL

ISTAR-IUL - Centro de Investigação em Ciências da Informação, Tecnologias e Arquitetura	https://ciencia.iscte-iul.pt/api/centre/ISTAR-IUL
UNIDE-IUL - Unidade de Investigação em Desenvolvimento Empresarial	https://ciencia.iscte-iul.pt/api/centre/UNIDE

Tabela1 – API REST Ciência-IUL para obter os autores

Após se obter a listagem de autores do ISCTE-IUL, é necessário consumir para cada um o serviço da API que obtém as suas publicações. Se o identificador do autor for XPTO, o serviço a evocar será o *https://ciencia.iscte-iul.pt/api/author/**XPTO***. Por exemplo, para o professor Nuno David obtém-se o identificador numérico 531 e, portanto, as suas publicações encontram-se em *https://ciencia.iscte-iul.pt/api/author/531*, mais concretamente no caminho */publications*.

A API devolve três agrupamentos distintos de publicações: *articles*, a lista de grupos de publicações em revistas científicas; *books*, a lista de grupos de livros (autor ou editor) e capítulos de livros; e *other*, a lista de grupos de outras publicações (conferências, comunicações, relatórios, etc.). Para cada agrupamento, as publicações são divididas de acordo com o seu tipo, presente na propriedade *container*.

Neste trabalho introduziu-se, adicionalmente, uma classificação extra associado ao tipo de publicações (ver Tabela 2). Esta permite que alguns tipos de publicações não sejam contabilizados no cálculo do peso das arestas através da equação de Newman (Equação 1, Secção 3). Por outras palavras, optou-se por restringir o âmbito do que é uma coautoria ao conjunto dos trabalhos que tenham avaliação científica por pares, ou que tenham sido publicados sob controle editorial ou que sejam publicados em veículos de referência. Neste contexto, optou-se ainda por excluir as Comunicações. Note-se ainda que estas são, em grande medida, um reflexo das publicações em atas de conferência, evitando-se assim que o algoritmo processe uma mesma colaboração em duplicado. Esta parametrização é configurável pelo administrador da aplicação, permitindo, por exemplo, alterá-la para avaliar também os padrões de coautoria em termos de colaborações de projetos.

Tipo de Publicação	Descriçao	Classificação
label.type.journal_paper	Artigos em revista	1
label.type.book	Livros (Autoria)	1
label.type.book_editor	Livros (Coordenação Editorial)	1

label.type.architecture	Publicações de projetos de arquitetura em edições de referência	1
label.type.book_chapter	Capítulos de livro	1
label.type.conference_paper	Artigos em atas de conferência (> 12 000 car.)	1
label.type.conference_editor	Atas de conferência (Editor)	1
label.type.preface	Entradas/Posfácios/Prefácios em obras de referência	1
label.type.working_paper	*Working papers* com avaliação científica, com publicação online	1
label.type.non_reviewed_paper	Artigos sem avaliação científica	0
label.type.general_report.international	Relatórios anuais de responsável geral de projeto científico internacional	0
label.type.local_report.international	Relatórios anuais de responsável local de projeto científico internacional	0
label.type.anual_report.national	Relatórios anuais de responsável de projeto científico nacional	0
label.type.final_report.international	Relatórios finais de responsável de projeto científico internacional	0
label.type.final_report.national	Relatórios finais de responsável de projeto científico nacional	0
label.type.scholar_report	Relatórios de coordenação de bolseiros de iniciação à investigação	0
label.type.architecture2	Referências a projetos em publicações temáticas de terceiros (Arquitetura)	0
label.type.recension	Recensões de obras em revistas com avaliação científica	1
label.type.talk	Comunicações	0

Tabela 2 – Tipos de publicação no Ciência-IUL (Ciência-IUL, 2016)

Durante o processamento de uma publicação são adicionadas as suas categorias científicas da *Scimago*, previamente fornecidas pela administração do Ciência-IUL. Esta atribuição é feita através do ISSN (*International Standard Serial Number*), identificador

único atribuído a cada publicação. Em algumas publicações podem não existir categorias *Scimago*.

Dentro de cada publicação existe uma lista de autores em */authors*. Apesar de já existir a lista de autores e suas publicações, esta lista não é redundante, pois é nela que se encontra a propriedade *internal* atribuída a cada autor, permitindo distinguir entre um autor interno e externo à instituição. A solução apenas considera o autor se este tiver for interno.

Foi detetado que a lista de autores de uma dada publicação pode ter autores que já não se encontram ativos no Ciência-IUL; na maioria dos casos tratam-se de antigos funcionários do ISCTE-IUL. A consequência é que estes autores não estão na lista de autores que se obtêm a partir dos departamentos e centros de investigação. Assim sendo, são retirados das publicações os autores que não estão ativos no Ciência-IUL, uma vez que não temos a informação detalhada de cada um deles.

Durante a construção do grafo são também removidas todas as publicações com um único autor, visto que nada acrescentam à rede de coautorias. Analogamente, um autor que não participe em, pelo menos, uma coautoria é também removido, evitando-se a existência de vértices isolados que em nada ajudam a identificação de comunidades.

A API do Ciência-IUL não permite realizar pesquisas temporais das alterações que vão ocorrendo nos seus autores e publicações. Desta forma, a construção da rede de coautorias obriga a pedir sempre a totalidade da informação ao servidor, obrigando também a reconstruir o grafo. Contudo, nem sempre a API responde em tempo útil, originando um erro de *timeout*. Para evitar sobrecarregar o servidor do Ciência-IUL e obter erros de *timeout*, as chamadas à API são intervaladas em cinco segundos. Adicionalmente, foi implementando um mecanismo de repetição da chamada à API, por uma segunda vez, em caso de *timeout*.

5.2. Base de dados

O objetivo da base de dados é guardar a rede de coautorias e as comunidades identificadas ao longo do tempo segundo a modelação apresentada na Figura 9. Dado que uma rede de coautorias é um grafo, existem dois conceitos essenciais, o de vértice (o autor) e o de aresta (a coautoria). A entidade *Author* será o vértice e a entidade *CoAutoriship* será a aresta. Cada coautoria tem necessariamente dois autores e, pelo menos, uma publicação, representada pela entidade *Publication*. Sempre que uma publicação é criada são adicionadas as suas categorias *Scimago* na propriedade *ScimagoCategories*. Uma

publicação pode estar em várias coautorias diferentes. Cada uma destas entidades tem um conjunto de propriedades que a caracteriza no contexto da rede de coautorias.

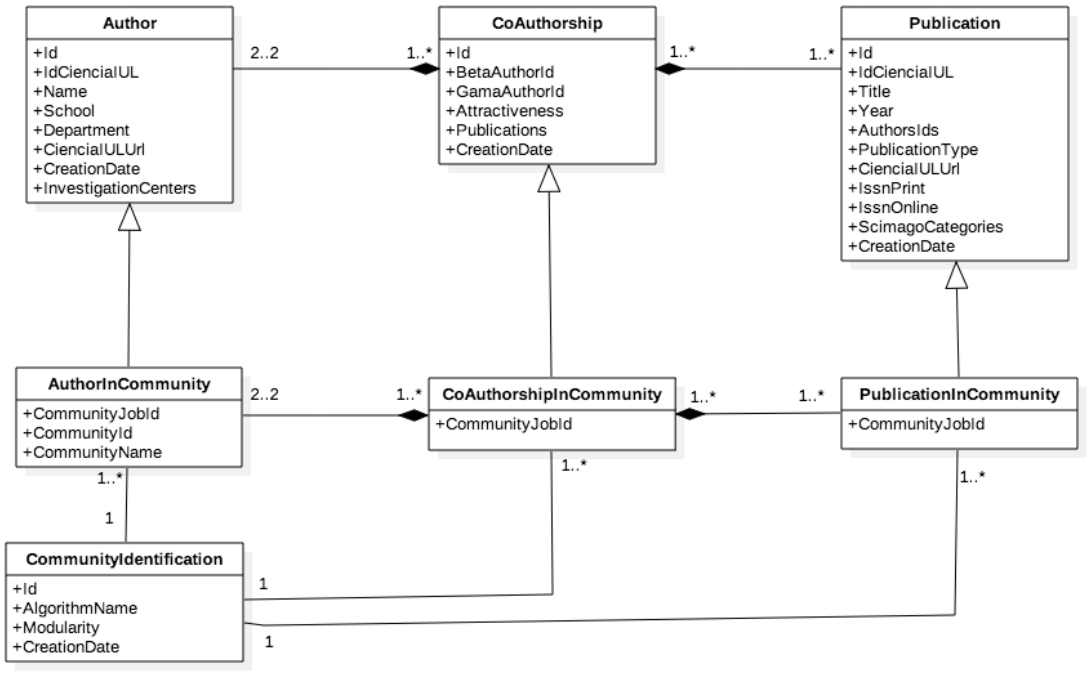

Figura 9 – Modelo da base de dados

Note-se que não se pretende que esta base de dados seja uma cópia do Ciência-IUL. Neste contexto, para se saber mais sobre um autor ou uma publicação terá que se consultar o Ciência-IUL e não esta base de dados. O atributo *CienciaIULUrl* contém o endereço de internet para aceder a mais informação sobre a entidade em causa no Ciência-IUL.

Sempre que são identificadas comunidades, numa determinada data, é copiado o grafo atual e guardado nas entidades derivadas *AuthorInCommunity*, *CoAutorshipCommunity* e *PublicationInCommunity*. O contexto da identificação de comunidades, assim como o algoritmo utilizado e a modularidade do grafo criado, são guardados na entidade *CommunityIdentification*. Isto permite consultar um grafo bem como as comunidades nele identificadas.

No objeto *AuthorInCommunity* será guardado a identificação da sua comunidade e o nome atribuído à mesma. O nome de uma comunidade é atribuído quando se executa um algoritmo de identificação, sendo feita uma recolha dos nomes de todos os departamentos

e centros de investigações presentes na comunidade e, de seguida, selecionadas os três nomes mais frequentes. Os nomes resultantes são concatenados pelo caracter &.[2]

Quando o grafo é atualizado, as entidades *Author*, *CoAutoriship* e *Publication* são apagadas e substituídas por novas entidades do mesmo tipo: toda a rede é reconstruída.

Dada a natureza da informação em causa, a base de dados utilizada tem o paradigma *NoSql*. Neste paradigma, as bases de dados estão preparadas para processar um grande volume de dados e conseguem rapidamente adaptar-se a mudanças no modelo de dados. Elas distinguem-se das bases de dados relacionais pois não guardam a informação em tabelas relacionais, mas em objetos, neste caso no formato *JSON*.

5.3. Tecnologias

A solução é uma aplicação *web* e foi desenvolvida com tecnologias associadas ao desenvolvimento *web*. Foram selecionadas tecnologias que não implicassem custos de licenciamento. Desta forma, garante-se que a solução desenvolvida pode ser replicada em outros contextos e, se necessário, melhorada sem custos associados.

A solução tem uma arquitetura de três camadas: cliente – servidor – base de dados (ver Figura 10). Na Tabela 3 encontra-se o levantamento das tecnologias utilizadas em cada camada, onde se destaca o *software* Node.js para o servidor e a solução da MongoDB para a base de dados. Do lado do cliente, a solução será acedida num navegador *web* comercial, pelo que a solução será desenvolvida com JavaScript, tecnologia amplamente suportadas por todos os navegadores *web* comerciais. De realçar que a linguagem de programação do lado do servidor também é em JavaScript, promovendo-se assim sinergias na construção das camadas de servidor e de cliente. Em ambos são utilizadas bibliotecas externas para auxiliar tarefas como a gestão de ligações HTTP, ligação à API do Ciência-IUL, desenho de grafos, etc.

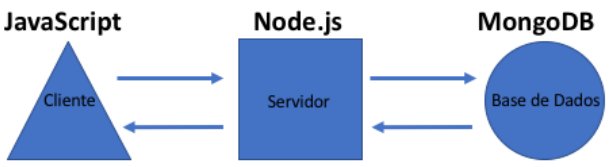

Figura 10 – Arquitetura da solução

[2] É ainda realizada uma limpeza prévia ao nome dos departamentos, retirando o *Department of* do início do seu nome; no caso dos centros de investigação são retiradas do seu nome as siglas iniciais como ISTAR, CEI-IUL ou CIS-IUL, por exemplo.

Tecnologia	Descrição
Node.js	Servidor aplicacional que permite ao cliente aceder à interface gráfica e aos dados guardados na base de dados. Também permite executar os algoritmos de construção da rede de coautoras e identificação de comunidades. É utilizado o *middleware* Express.JS de modo a automatizar as funções de gestão típicas de aplicações *web*. (Node.js, 2017)
MongoDB	Sistema de gestão de base de dados *NoSql*. (MongoDB, 2017)
JavaScript	Linguagem de programação interpretada pelos navegadores *web* comerciais. Os grafos são desenhados recorrendo à biblioteca *vis.js*. (vis.js, 2017)

Tabela 3 – Tecnologias utilizadas na solução

5.4. Funcionalidades

5.4.1. REST API

A solução desenvolvida permite a interação com a mesma através de uma API REST própria. A API REST oferece o que se designa de *endpoints* para obter informação ou então executar alguma funcionalidade (Tabela 4). A troca de informação com os *endpoints* é feita através do formato JSON. A interface gráfica utiliza esta API para disponibilizar as suas funcionalidades.

Uma vez que a solução tem informação privada do Ciência-IUL, nomeadamente as categorias científicas de cada publicação atribuídas pela *Scimago*, foi necessário autenticar a mesma através da plataforma de identificação da Google, que utiliza o protocolo de autenticação OAuth 2 (Google, 2017). Todos os utilizadores do ISCTE-IUL possuem uma conta Google e podem assim autenticar-se com a mesma. A solução apenas recebe o seu email e se o mesmo não for do domínio *iscte.pt* ou *iscte-iul.pt* então não será permitido o acesso do utilizador à aplicação. Desta forma, a gestão de perfis fica exclusivamente à responsabilidade do Google e da gestão informática do ISCTE-IUL.

A validação se um utilizador está autenticado é feita sempre que um endpoint da API REST é invocado. Caso não esteja, o *endpoint* responderá com a mensagem de erro de utilizador não autorizado. Para além da autenticação com o email ISCTE-IUL, a solução tem implementada uma *password* de administração, necessária para efetuar certas operações, e que é gerida pela própria solução.

Endpoint	Função
/api/graph/authors	Obter a lista de autores e suas propriedades.

/api/graph/publications	Obter a lista de publicações e suas propriedades.
/api/graph/coauthorships	Obter a lista de coautorias e suas propriedades.
/api/harvest	Começar o processo de recolha de informação no Ciência-IUL. Requer autenticação de administração.
/api/deleteDB	Apagar a base de dados. Requer autenticação de administração.
/api/communitydetection	Começa o processo de identificação de comunidades, utilizando todos os algoritmos configurados na solução. Requer autenticação de administração.
/api/graph/getcommunitiesids	Obter a lista de identificadores de grafos com comunidades identificadas.
/api/graph/authorsincommunity	Obter a lista de autores e suas propriedades para um grafo com comunidades identificadas.
/api/graph/publicationsincommunity	Obter a lista de publicações e suas propriedades para um grafo com comunidades identificadas.
/api/graph/coauthorshipsincommunity	Obter a lista de coautorias e suas propriedades para um grafo com comunidades identificadas.
/api/deletecommunities	Apagar na base de todos os grafos com comunidades identificadas. Requer autenticação de administração.

Tabela 4 – API REST da solução

5.4.2. Interface Gráfica

A interface gráfica foi feita recorrendo a HTML, CSS e JavaScript, tecnologias amplamente utilizadas em desenvolvimento *web*. Foi utilizado como base do HTML e CSS um *template* disponibilizado por Matt Brown em 2010 (Sample Resume Template, 2017). O mesmo sofreu várias alterações para ser adaptado a este trabalho. Para desenhar o grafo, com os seus nós e vértices, foi utilizada a biblioteca *vis.js* (vis.js, 2017).

A solução desenvolvida tem cinco páginas: a inicial (*index.html*), a informativa (*about.html*), a administrativa (*admin.html*), a analítica (*analytics.html*) e a de ajuda (*help.html*). Uma vez que a API REST utilizada pela interface necessita de autenticação, sempre que um *endpoint* retorna o erro de utilizador não autorizado, o mesmo é reencaminhado para uma página de autenticação (*login.html*).

A página administrativa permite executar as quatro funções essenciais para a manutenção e utilização da solução: apagar toda a base de dados, apagar apenas a informação de

comunidades na base de dados, executar o processo de construção do grafo a partir da informação presente no Ciência-IUL e executar o processo de identificação de comunidades a partir do grafo mais recentemente criado e guardado na solução. Todas estas funções requerem autenticação de administração e no fundo da página aparece *feedback* sobre o processo em curso, caso seja necessário.

A página informativa descreve a motivação, enquadramento e objetivos da solução. A página analítica apresenta um conjunto de estatísticas gerais da rede de coautorias, nomeadamente, a listagem textual de todas as coautorias com os respetivos autores e força de atração, a listagem de todos os autores presentes no grafo com os seus totais de publicações e coautorias e a listagem de autores por escola, departamentos e unidades de investigação. A página de ajuda contem itens de ajuda genérica de como utilizar e interagir com a área de apresentação dos grafos desenhados na página inicial da aplicação.

A página inicial é a parte central da solução, sendo a que permite visualizar a rede de coautorias, os grafos com as comunidades identificadas, um autor em particular com as suas coautorias e propriedades, uma comunidade em particular com as suas propriedades e alguma informação analítica geral, como por exemplo, o total de autores por escolas, por departamentos e por unidades de investigação.

No topo da página inicial encontram-se os formulários que permitem a construção dos grafos com autores e coautorias (ver Figura 11). Com o botão *Compute Coauthorship Network* é construída e desenhada a rede de coautorias na área de visualização. Na escolha múltipla *Select an Algorithm for Community Detection* é possível escolher os grafos de coautorias com as comunidades identificadas, sendo depois desenhadas na área de visualização. O nome destes grafos é composto pelo nome do algoritmo e a data em que o mesmo foi executado, não existindo limites ao número de quantos podem existir. Na escolha múltipla *Refine Search - Select a Community* é possível escolher uma comunidade em particular do grafo, no âmbito do algoritmo selecionado na escolha múltipla anterior.

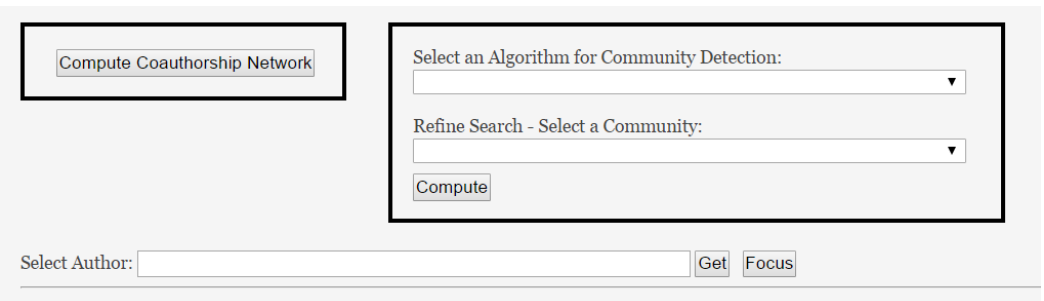

Figura 11 – Formulários de pesquisa da página inicial

Sempre que existir um grafo desenhado na área de visualização é possível procurar por um autor em particular presente no mesmo através da caixa de texto *Select Author* e carregando no botão *Get*.[3] Se o grafo desenhado for um com comunidades identificadas, a ação anterior desenha a comunidade a que o autor pertence (ver Figura 15). Este comportamento é também obtido quando se carrega duas vezes com o rato no vértice de um autor no grafo. O botão *Focus* coloca o autor no centro da área de visualização, não afetando o grafo desenhado.

Na área de apresentação do grafo existem várias zonas funcionais (Figura 12). No topo esquerdo da área de visualização encontra-se o título do grafo desenhado e no topo direito os botões de voltar atrás e abrir a página de ajuda. Na zona onde o grafo é desenhado é possível fazer *zoom in*, *zoom out* e movimentar o grafo quer através dos botões no fundo ou através do rato. É possível carregar numa aresta e ver a lista de publicações associadas à mesma, assim como carregar num autor em particular e navegar para o mesmo (ver Figura 13). Com o botão de voltar para trás é possível desfazer esta ação. Ao passar o rato por cima de um autor é apresentada uma legenda do mesmo, com o seu nome, escola e departamento; caso se esteja no contexto de um grafo com comunidades, é apresentado também o nome da comunidade a que o autor pertence.

[3] Após preenchimento é necessário carregar no botão *Get* para o autor ser desenhado no centro da área de visualização, com os respetivos coautores à volta caso o grafo seja o grafo geral desenhado sem identificação de comunidades (ver Figura 13).

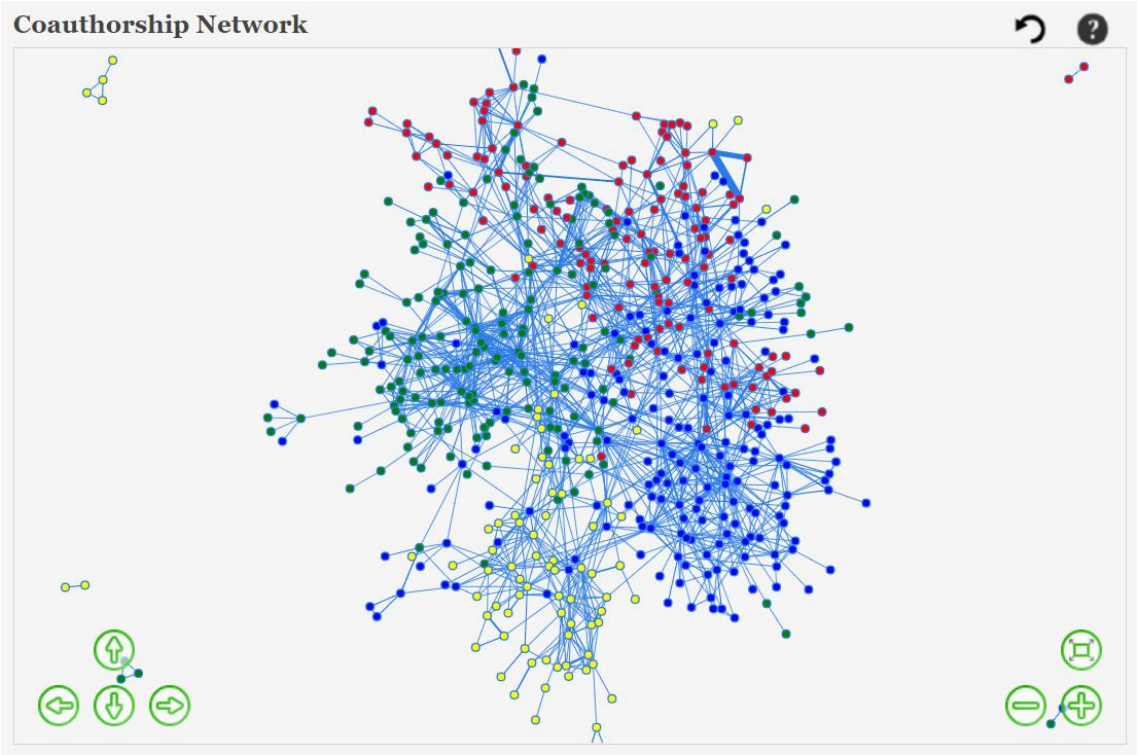

Figura 12 – Área de visualização do grafo na página inicial

Os vértices representativos dos autores têm todos o mesmo tamanho e estão divididos por quatro cores que representam as escolas do ISCTE-IUL: *ISCTE Business School* (vermelho), *School of Social Sciences* (azul), *School of Sociology and Public Policy* (verde) e *School of Technology and Architecture* (amarelo). Esta escala de cores apenas não se aplica quando é desenhado um grafo com comunidades identificadas. O tamanho dos vértices dos autores é sempre igual, ao contrário das arestas que têm a espessura proporcional à força de atratividade da coautoria, calculada de acordo com a equação de Newman. Ao carregar duas vezes com o rato num vértice são apresentadas as publicações (nome e URL), que são parte integrante da coautoria que o mesmo representa.

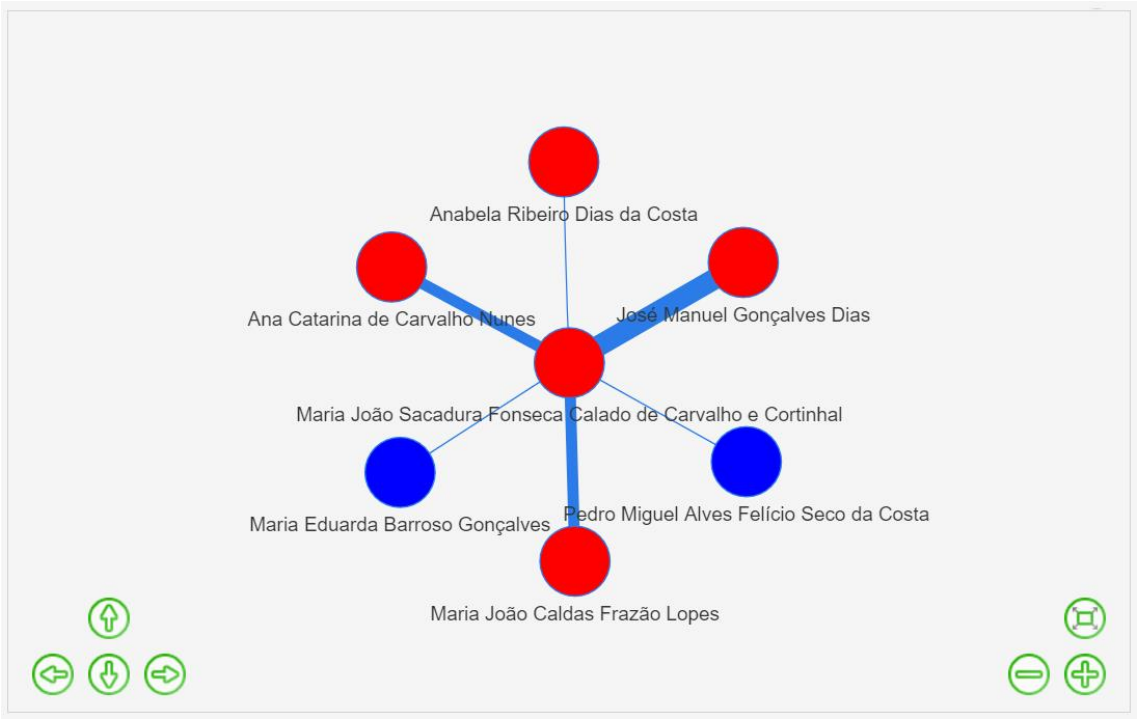

Figura 13 – Grafo de autor selecionado e suas coautorias

Na representação de um grafo com comunidades identificadas (Figura 14) os autores não têm a cor da escola a que pertencem, mas sim a cor da sua comunidade. Dada a natureza dinâmica e imprevisível das comunidades, as cores são atribuídas dinamicamente sempre que o grafo é desenhado. Ao selecionar um autor em particular é apresentada a sua comunidade (Figura 15), sendo a cor dos vértices representativa da escola a que cada um dos autores pertence. De realçar que as comunidades identificadas com menos de quatro autores não são incluídas na apresentação e que cada autor apenas pode aparecer numa comunidade. Também é possível que existam comunidades diferentes com o mesmo nome.

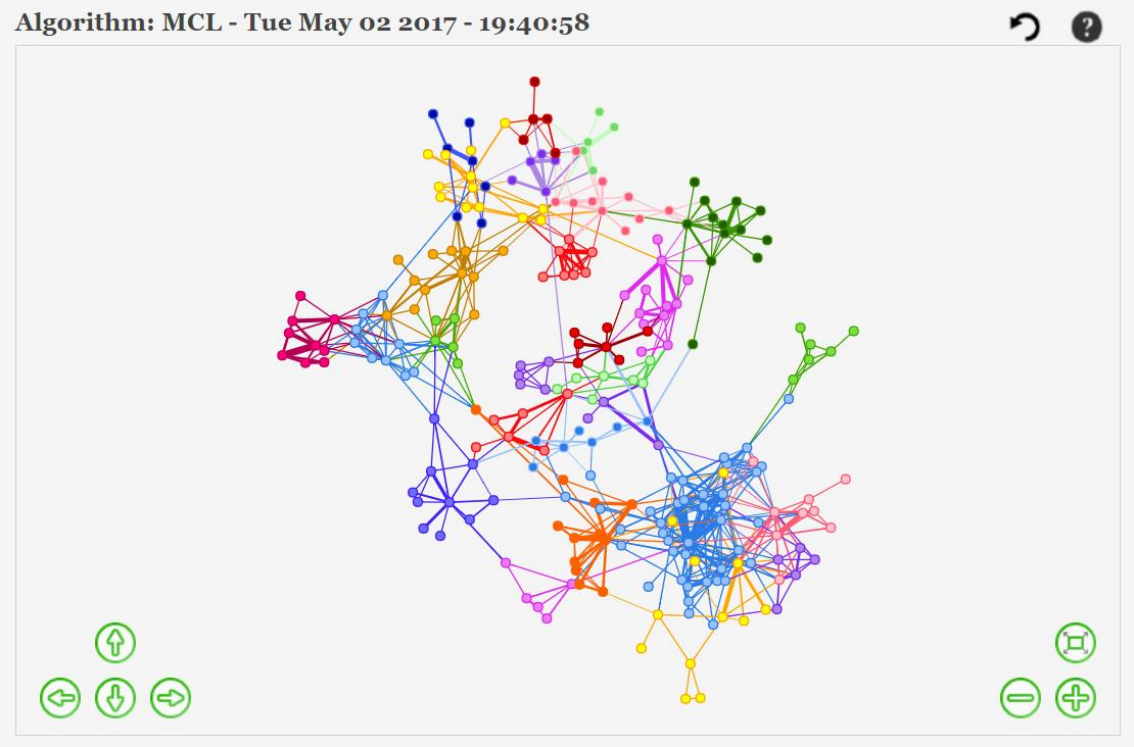

Figura 14 – Grafo do algoritmo MCL com várias comunidades identificadas

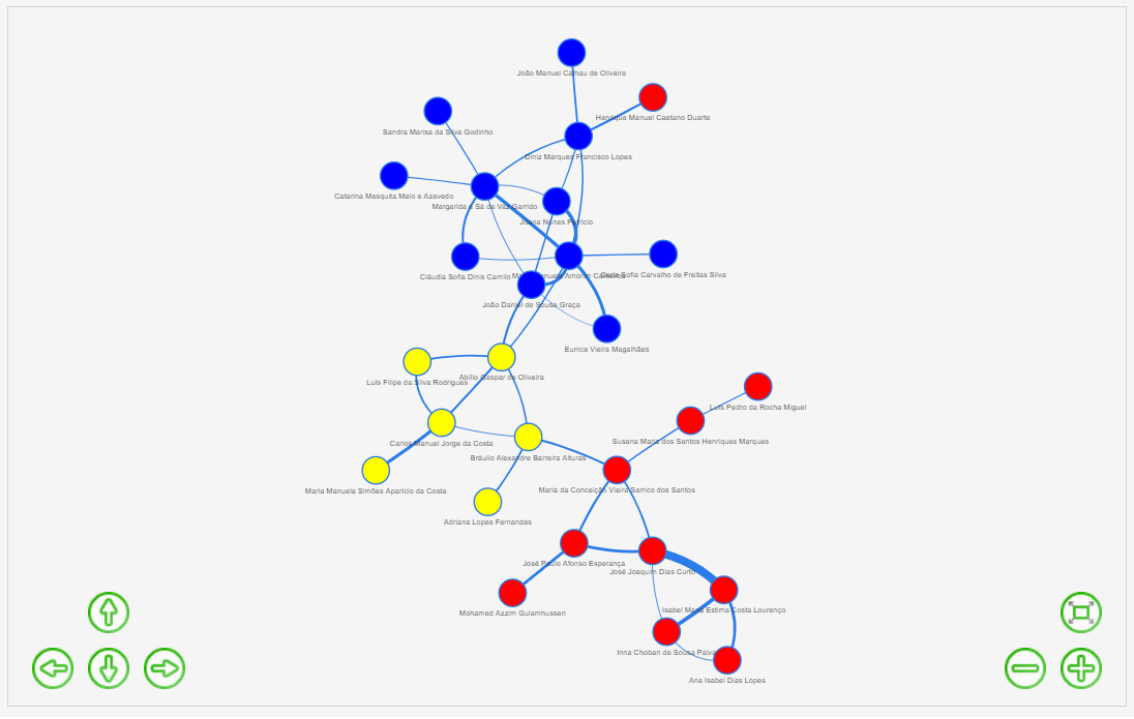

Figura 15 – Comunidade de um autor selecionado

Seja quando são apresentadas as coautorias imediatas de um autor (Figura 13), seja quando é apresentada a comunidade de um autor (Figura 15), é apresentado também um quadro resumo com os autores presentes, as suas escolas, os seus departamentos, os seus centros de investigação e o seu grau (número de coautores a que está ligado). O nome de

cada autor tem a ligação para o perfil público do autor no Ciência-IUL. Caso seja selecionado um contexto de comunidade aparece também o nome da comunidade e as cinco categorias *Scimago* que mais ocorrem nas publicações da comunidade (Figura 16). Adicionalmente são apresentados os totais de autores por escola, departamento e unidades de investigação, assim como uma lista detalhada de escolas, departamentos e unidades de investigação representadas.

Community Name: Information Sciences, Technologies and Architecture Research Center & Information Science and Technology & Telecomunicações-IUL
Scimago Categories (5 more representative): Computer Science (miscellaneous) , Software , Computer Science Applications , Hardware and Architecture , Signal Processing

Author	School	Department	Investigation Centers	Degree
Nuno Manuel Mendes Cruz David	School of Social Sciences	Department of Information Science and Technology	DINÂMIA'CET-IUL - Centre for Socioeconomic Change and Territorial Studies	1
Luís Miguel Martins Nunes	School of Technology and Architecture	Department of Information Science and Technology	Instituto de Telecomunicações-IUL ; ISTAR-IUL - Information Sciences, Technologies and Architecture Research Center	3
José Miguel de Oliveira Monteiro Sales Dias	School of Technology and Architecture	Department of Information Science and Technology	ISTAR-IUL - Information Sciences, Technologies and Architecture Research Center	4
Joaquim José Gonçalves Marques	School of Technology and Architecture		ISTAR-IUL - Information Sciences, Technologies and Architecture Research Center	1
Filipe Alexandre Gonçalves Gaspar	School of Technology and Architecture		ISTAR-IUL - Information Sciences, Technologies and Architecture Research Center	1
David Walter Figueira Jardim	School of Technology and Architecture		ISTAR-IUL - Information Sciences, Technologies and Architecture Research Center	2
Carlos José Corredoura Serrão	School of Technology and Architecture	Department of Information Science and Technology	ISTAR-IUL - Information Sciences, Technologies and Architecture Research Center	2
Ana Maria Carvalho de Almeida	School of Technology and Architecture	Department of Information Science and Technology	ISTAR-IUL - Information Sciences, Technologies and Architecture Research Center	2

Total authors: 8
Total coauthorships: 8
Total publications: 44
Total communities: 1

Authors by Schools:
School of Social Sciences (blue): 1
School of Technology and Architecture (yellow): 7

Authors by Departments:
Department of Information Science and Technology: 5

Authors by Investigation Centers:
DINÂMIA'CET-IUL - Centre for Socioeconomic Change and Territorial Studies: 1
Instituto de Telecomunicações-IUL: 1
ISTAR-IUL - Information Sciences, Technologies and Architecture Research Center: 7

Figura 16 – Quadro resumo de um autor selecionado

Após selecionar-se um grafo de comunidades existe também a opção de *Export Communities Report*, de modo a permitir exportar em formato HTML um relatório textual e geral das comunidades presentes no grafo. Nesse relatório estão presentes os nomes das

comunidades, os seus totais de autores, os seus totais de coautorias, os seus autores com maior grau e as suas categorias *Scimago*.

5.4.3. Exportação para GraphML

De modo a permitir que a informação da aplicação seja trabalhada por ferramentas externas é oferecida, através da interface gráfica, a possibilidade de exportação da rede de coautorias no formato GEXF - *Graph Exchange XML Format* (GEXF Working Group, 2017). Este formato foi criado em 2007 pela *Gephi*, uma das referências na construção de ferramentas informáticas para análise de grafos, e tem sido adotado por várias entidades. A rede exportada não tem qualquer comunidade identificada na mesma, tendo apenas os seus vértices, arestas e peso (força de atratividade) associado a cada aresta. O formato é transversal a várias ferramentas que o utilizam como protocolo de importação e exportação de grafos.

6. Avaliação dos Resultados

Durante o mês de março de 2017, a solução foi testada em três ambientes computacionais na recolha da informação do Ciência-IUL e posterior construção do grafo: Windows 10 Intel i5 2.3 GHz 16 GB RAM, macOS Sierra Intel i5 1.6 GHz 8 GB RAM e Amazon Linux (Elastic Cloud Computing) 2vcpu 3.3 GHz 4 GB RAM. Foram detetados 613 autores internos do ISCTE-IUL com, pelo menos, uma publicação partilhada com outro autor interno. No total, os 613 autores partilham entre si 3766 publicações e estão ligados por 1718 coautorias. Este processo demorou cerca de 1 hora e 10 minutos a ser executado nos três ambientes computacionais referidos. De realçar que 482 autores foram descartados porque não tinham qualquer coautoria e que 940 publicações foram ignoradas porque ou tinham apenas autores externos ao ISCTE-IUL ou apenas 1 autor interno.

A distribuição dos 613 autores pelas diferentes escolas, departamentos e unidades de investigação é a seguinte:

- Autores por escolas:
    - *School of Social Sciences*: 195
    - *School of Sociology and Public Policy*: 185
    - *ISCTE Business School*: 141
    - *School of Technology and Architecture*: 92
- Autores por departamentos:
    - *Department of Information Science and Technology*: 59
    - *Department of Social and Organizational Psychology*: 46
    - *Department of Sociology*: 45
    - *Department of Marketing, Operation and Management*: 34
    - *Department of Quantitative Methods for Management and Economics*: 22
    - *Department of Human Resources and Organizational Behavior*: 22
    - *Department of Architecture and Urbanism*: 19
    - *Department of Political Science and Public Policy*: 18
    - *Department of Political Economy*: 17
    - *Department of Accounting*: 16
    - *Department of Economics*: 15
    - *Department of Social Research Methods*: 13
    - *Department of Finance*: 12
    - *Department of Anthropology*: 11

- *Department of Mathematics*: 7
- *Department of History*: 7

- Autores por unidades de investigação:
  - *CIES-IUL - Centre for Research and Studies in Sociology*: 166
  - *BRU-IUL - Business Research Unit*: 99
  - *CIS-IUL - Centre for Social Research and Intervention*: 99
  - *DINÂMIA'CET-IUL - Centre for Socioeconomic Change and Territorial Studies*: 82
  - *ISTAR-IUL - Information Sciences, Technologies and Architecture Research Center*: 55
  - *Instituto de Telecomunicações-IUL*: 32
  - *CRIA-IUL - Centre for Research in Anthropology - IUL*: 17
  - *CEI-IUL - Center for International Studies*: 17
  - *Instituto de Plasmas e Fusão Nuclear*: 1

A rede de coautorias (Figura 17) encontra-se dividida pelas quatro escolas do ISCTE-IUL: *ISCTE Business School* (vermelho), *School of Social Sciences* (azul), *School of Sociology and Public Policy* (verde) e *School of Technology and Architecture* (amarelo).

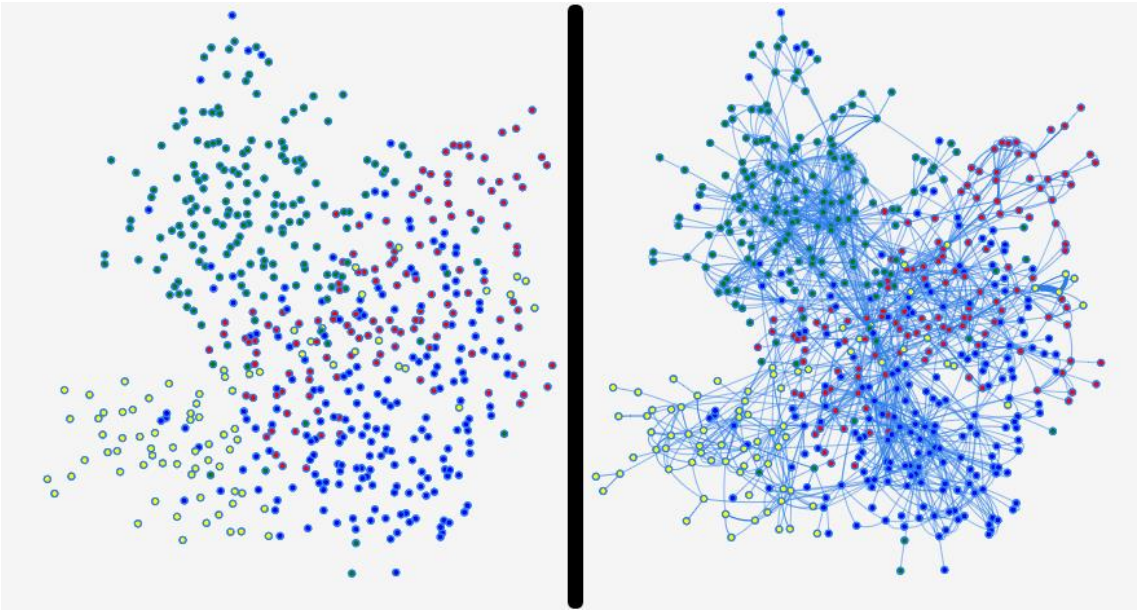

Figura 17 – Rede de coautorias, à esquerda apenas com os autores, à direita com autores e coautorias.

Uma análise visual da rede apresentada na Figura 17 permite observar que a *School of Technology and Architecture,* a *School of Sociology and Public Policy* e a *School of Social Sciences* encontram-se, de modo geral, separadas em blocos distintos. A *School of*

Sociology and Public Policy apresenta um grande número de coautorias dentro da própria escola que faz dela um bloco coeso. Já os autores da *ISCTE Business School* partilham muitas ligações com as outras 3 escolas, o que os faz não formarem um bloco denso entre si, mas antes uma ponte que se estende ao longo das outras escolas. Desta estrutura, pode teorizar-se que os autores desta escola exercem uma partilha de conhecimento tendencialmente mais transversal ao ISCTE-IUL, com os demais autores a fazerem-no de forma mais localizada.

Dentro da rede de coautorias com 613 autores existem alguns que apresentam características que são de realçar, como sejam os 5 autores com mais publicações (Tabela 5) e os 5 autores com mais coautorias (Tabela 6).

Autor	Nº Publicações	Percentagem do Total
Manuel Alberto Martins Ferreira (Departamento de Matemática)	199	5%
Gustavo Alberto Guerreiro Seabra Leitão Cardoso (Departamento de Sociologia)	153	4%
José António Candeias Bonito Filipe (Departamento de Matemática)	129	3%
António Manuel Hipólito Firmino da Costa (Departamento de Sociologia)	99	3%
Maria Luísa Soares Almeida Pedroso de Lima (Departamento de Psicologia Social e das Organizações)	90	2%

Tabela 5 – Autores com mais Publicações.

Autor	Nº Coautorias	Percentagem do Total
António Manuel Hipólito Firmino da Costa (Departamento de Sociologia)	40	2%
Patrícia Durães Ávila (Departamento de Métodos de Pesquisa Social)	36	2%
Maria Luísa Soares Almeida Pedroso de Lima (Departamento de Psicologia Social e das Organizações)	32	2%
Helena Maria Barroso Carvalho (Departamento de Métodos de Pesquisa Social)	31	2%

Margarida e Sá de Vaz Garrido (Departamento de Psicologia Social e das Organizações)	29	2%

Tabela 6 – Autores com mais Coautorias.

Deve-se realçar que uma coautoria pode ter várias publicações, pelo que o número de publicações é sempre superior, ou igual, ao número de coautorias. Também interessante de analisar são as 5 coautorias com maior força de atratividade, calculada de acordo com a equação de Newman (Tabela 7).

Autor	**Autor**	**Atratividade**
Manuel Alberto Martins Ferreira (Departamento de Matemática)	José António Candeias Bonito Filipe (Departamento de Matemática)	109
Manuel Alberto Martins Ferreira (Departamento de Matemática)	Marina Alexandra Pedro Andrade (Departamento de Matemática)	69
Isabel Maria Estima Costa Lourenço (Departamento de Contabilidade)	José Joaquim Dias Curto (Departamento de Métodos Quantitativos para Gestão e Economia)	24
Sofia Maria Lopes Portela (Departamento de Métodos Quantitativos para Gestão e Economia)	Rui Manuel Campilho Pereira de Menezes (Departamento de Métodos Quantitativos para Gestão e Economia)	24
António Manuel Hipólito Firmino da Costa (Departamento de Sociologia)	Fernando Luís Machado (Departamento de Sociologia)	20

Tabela 7 – Coautorias com mais força de atratividade.

Pela Tabela 7 observa-se que as coautorias com maior força de atratividade são entre autores do mesmo departamento. A única exceção é entre a autora Isabel Maria Estima Costa Lourenço (Departamento de Contabilidade) e o autor José Joaquim Dias Curto (Departamento de Métodos Quantitativos para Gestão e Economia), realçando que ambos os departamentos são da *ISCTE Business School*.

Foram utilizados dois algoritmos para identificação de comunidades, o ABCD e o MCL. Os algoritmos foram testados em três ambientes computacionais: Windows 10 Intel i5 2.3 GHz 16 GB RAM, macOS Sierra Intel i5 1.6 GHz 8 GB RAM e Amazon Linux (Elastic Cloud Computing) 2vcpu 3.3 GHz 4 GB RAM. Nos três casos os resultados obtidos para o mês de março de 2017 foram semelhantes. O MCL demorou cerca de 5 segundos a ser executado e o ABCD cerca de 2 minutos. Esta diferença entre os tempos de execução é explicada pelo facto do MCL ser compilado para a arquitetura da máquina

onde se encontra a ser executado e o ABCD ter sido implementado em JavaScript, uma linguagem que exige o seu processamento e interpretação das suas instruções em tempo de execução, com prejuízo da sua performance.

Os algoritmos ABCD e MCL são determinísticos e serão obtidos os mesmos resultados para o mesmo grafo da rede de coautorias sempre que forem executados com os mesmos dados de entrada. Contudo, deve realçar-se que basta adicionar um único autor na rede de coautorias e os resultados poderão ser significativamente diferentes, uma vez que a constituição da rede em termos de probabilidades de transições e pesos nas arestas muda; os algoritmos utilizados têm elevada sensibilidade a mudanças na rede de coautorias.

Como consequência do funcionamento dos algoritmos utilizados, o mesmo autor não pode pertencer a diferentes comunidades. Os autores de uma comunidade encontram-se todos ligados entre si e, assim sendo, num grafo com comunidades identificadas é sempre possível a partir de um vértice chegar a outro vértice. Por uma questão de legibilidade, na denominação de comunidades apenas foram consideradas as 5 categorias *Scimago* que mais aparecem associadas às publicações que a integram.

Saliente-se que a rede que serviu de suporte para a aplicação dos algoritmos para deteção de comunidades foi a rede de coautorias completa, ou seja, com 613 autores, 3766 publicações e 1718 coautorias. Contudo, verificou-se que muitas das comunidades eram constituídas por 4 ou menos autores. Sendo assim, para simplificar a análise que se segue essas comunidades não foram consideradas.

6.1. Identificação de Comunidades com o ABCD

Por execução do algoritmo ABCD foram identificadas 25 comunidades, compostas por 263 autores, 754 coautorias e 1671 publicações (ver Figura 18 e Tabela 8). É de realçar que estes valores correspondem a 43%, 44% e 44%, respetivamente, dos valores apresentados para a rede de coautorias. Isto reflete desde logo a existência de uma elevada percentagem de autores que, ou por terem um número reduzido de publicações ou por terem poucas colaborações científicas com outros autores do ISCTE-IUL, acabam por ser parte de pequenas sub-redes muito isoladas.

Por autor mais influente entende-se o que tem maior grau dentro da comunidade, ou seja, o que tem o maior número de ligações (arestas do grafo que representam coautorias) com os outros autores da mesma comunidade. Relembremos que uma coautoria pode ser composta por uma ou mais publicações. É comum existirem dentro duma mesma

comunidade diferentes autores com o mesmo grau, tal como se pode observar na coluna "Autor mais influente".

#	Comunidade	Nº Autores	Nº Coautorias	Autor mais influente	Categorias *Scimago*
1	Business Research Unit & Marketing, Operation and Management & Quantitative Methods for Management and Economics	9	11	Elisabeth de Azevedo Reis, João Carlos Rosmaninho de Menezes, Catarina Maria Valente Antunes Marques	Tourism, Leisure and Hospitality Management , Geography, Planning and Development , Computer Science (miscellaneous) , Speech and Hearing , Psychology (miscellaneous)
2	Research and Studies in Sociology & Political Science and Public Policy & Social Research Methods	17	22	Nuno Alexandre de Almeida Alves	Social Sciences (miscellaneous) , Sociology and Political Science , Computer Science (miscellaneous) , Life-span and Life-course Studies
3	Information Sciences, Technologies and Architecture Research Center & Socioeconomic Change	5	5	Pedro Cláudio de Faria Lopes, Vasco Nunes da Ponte Moreira Rato	Architecture , Conservation , Visual Arts and

	and Territorial Studies & Information Science and Technology				Performing Arts , Computer Science (miscellaneous)
4	Socioeconomic Change and Territorial Studies & Political Economy & Architecture and Urbanism	9	11	Pedro Miguel Alves Felício Seco da Costa	Political Science and International Relations , Sociology and Political Science , Social Sciences (miscellaneous) , Environmental Science (miscellaneous) , Computer Science (miscellaneous)
5	Business Research Unit & Quantitative Methods for Management and Economics & Economics	6	8	Rui Manuel Campilho Pereira de Menezes	Condensed Matter Physics , Statistics and Probability , Aerospace Engineering , Applied Mathematics , Control and Systems Engineering

6	Socioeconomic Change and Territorial Studies & Political Economy & Finance	11	11	Maria Eduarda Barroso Gonçalves	Environmental Science (miscellaneous) , Philosophy , Economics and Econometrics , Law , Computer Science (miscellaneous)
7	Research and Studies in Sociology & Sociology & Political Science and Public Policy	7	6	Maria Teresa de Morais Sarmento Patrício	Social Sciences (miscellaneous) , Education , Sociology and Political Science , Public Health, Environmental and Occupational Health , Arts and Humanities (miscellaneous)
8	Social Research and Intervention & Research and Studies in Sociology & Social and Organizational Psychology	29	37	Maria Luísa Soares Almeida Pedroso de Lima	Sociology and Political Science , Social Sciences (miscellaneous) , Education ,

					Social Psychology , Public Health, Environmental and Occupational Health
9	Social Research and Intervention & Social and Organizational Psychology & Sociology	19	28	Sónia Gomes da Costa Figueira Bernardes	Psychology (miscellaneous) , Education , Social Psychology , Anesthesiology and Pain Medicine , Medicine (miscellaneous)
10	Human Resources and Organizational Behavior & Business Research Unit & Social Research and Intervention	6	8	António Caetano	Computer Science (miscellaneous) , Public Health, Environmental and Occupational Health , Safety Research , Safety, Risk, Reliability and Quality , Business and International Management
11	Information Sciences, Technologies and	9	9	Joaquim António Marques dos Reis,	Computer Science

	Architecture Research Center & Information Science and Technology & Architecture and Urbanism			Sara Eloy Cardoso Rodrigues	(miscellaneous) , Artificial Intelligence , Industrial and Manufacturing Engineering , Software
12	Marketing, Operation and Management & Business Research Unit & Information Sciences, Technologies and Architecture Research Center	9	10	Paulo Miguel Rasquinho Ferreira Rita	Tourism, Leisure and Hospitality Management , Computer Science (miscellaneous) , Marketing , Artificial Intelligence , Business, Management and Accounting (miscellaneous)

13	Research and Studies in Sociology & Sociology & Economics	6	7	Gustavo Alberto Guerreiro Seabra Leitão Cardoso	Sociology and Political Science , Computer Science (miscellaneous) , Business and International Management , Development , Social Sciences (miscellaneous)
14	Socioeconomic Change and Territorial Studies & Business Research Unit & Political Economy	17	20	Sérgio Miguel Chilra Lagoa	Business and International Management , Management of Technology and Innovation , Strategy and Management , Social Sciences (miscellaneous) , Economics and Econometrics

15	Research and Studies in Sociology & Social Research Methods & Sociology	11	18	António Manuel Hipólito Firmino da Costa	Sociology and Political Science , Social Sciences (miscellaneous) , Computer Science (miscellaneous) , Geography, Planning and Development , Anthropology
16	Information Sciences, Technologies and Architecture Research Center & Business Research Unit & Information Science and Technology	15	16	Abílio Gaspar de Oliveira	Psychology (miscellaneous) , Arts and Humanities (miscellaneous) , Sociology and Political Science , Education , Human-Computer Interaction
17	Research and Studies in Sociology & Political Science and Public Policy	7	7	Luís Manuel Antunes Capucha, Alexandra Isabel Francisco Duarte	Social Sciences (miscellaneous) , Sociology and Political Science , Computer

					Science (miscellaneous) , Education , Political Science and International Relations
18	Research and Studies in Sociology & History & Research in Anthropology	9	11	Maria João Mendes Vaz	Computer Science (miscellaneous)
19	Research and Studies in Sociology & Political Science and Public Policy & Sociology	5	7	Jorge Manuel Leitão Ferreira	
20	Information Sciences, Technologies and Architecture Research Center & Socioeconomic Change and Territorial Studies & Information Science and Technology	6	6	Maria João Marques de Oliveira	
21	Information Sciences, Technologies and Architecture Research Center & Information Science and Technology & Telecomunicações-IUL	8	8	José Miguel de Oliveira Monteiro Sales Dias	Computer Science (miscellaneous) , Software , Electrical and Electronic Engineering , Computer Science Applications , Hardware

					and Architecture
22	Research and Studies in Sociology & Sociology	8	13	Maria das Dores Horta Guerreiro	Social Sciences (miscellaneous) , Sociology and Political Science , Decision Sciences (miscellaneous) , Engineering (miscellaneous) , Strategy and Management
23	Social Research and Intervention & Social and Organizational Psychology & Human Resources and Organizational Behavior	10	14	Margarida e Sá de Vaz Garrido	Psychology (miscellaneous) , Sociology and Political Science , Developmental and Educational Psychology , Education , Social Psychology
24	Research and Studies in Sociology & Quantitative Methods for Management and Economics & Business Research Unit	20	20	José Manuel Gonçalves Dias	Sociology and Political Science , Social Sciences (miscellaneous) , Computer

					Science (miscellaneous) , Marketing , Developmental and Educational Psychology
25	Business Research Unit & Accounting & Quantitative Methods for Management and Economics	5	7	Isabel Maria Estima Costa Lourenço	Business, Management and Accounting (miscellaneous) , Economics and Econometrics , Accounting , Statistics and Probability , Statistics, Probability and Uncertainty

Tabela 8 – Comunidades identificadas com o ABCD

As 25 comunidades são compostas em média por 11 autores e 13 coautorias. De realçar também que a categoria *Scimago, Computer Science,* está presente em 14 das 25 comunidades, o que revela a sua transversalidade e interdisciplinaridade.

A maior comunidade é a *Social Research and Intervention & Research and Studies in Sociology & Social and Organizational Psychology* com 29 autores e 37 coautorias, sendo o seu autor com maior grau, a professora Maria Luísa Soares Almeida Pedroso de Lima. O seu grau, o número de coautorias, é de 7. As categorias *Scimago* que a caracterizam são as *Sociology and Political Science, Social Sciences (miscellaneous), Education, Social Psychology, Public Health e Environmental and Occupational Health.* De realçar a correspondência entre o nome da comunidade, deduzido dos seus

departamentos e centros de investigação, e as categorias *Scimago*, identificando esta comunidade como sendo da área da Sociologia.

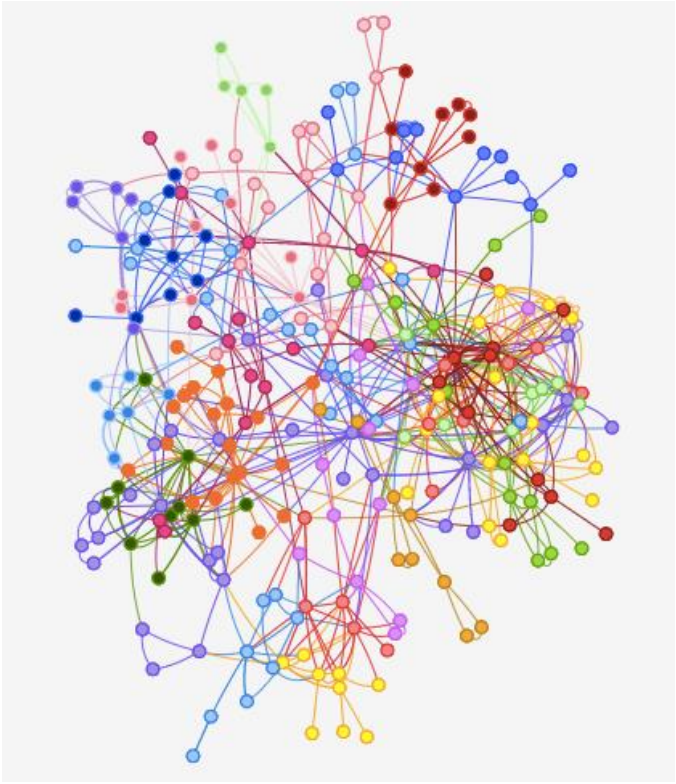

Figura 18 – Comunidades identificadas com o ABCD

Do ponto de vista visual é possível identificar distintos padrões de comunidade (ver Figura 19): desde redes altamente ligadas a redes muito ramificadas.

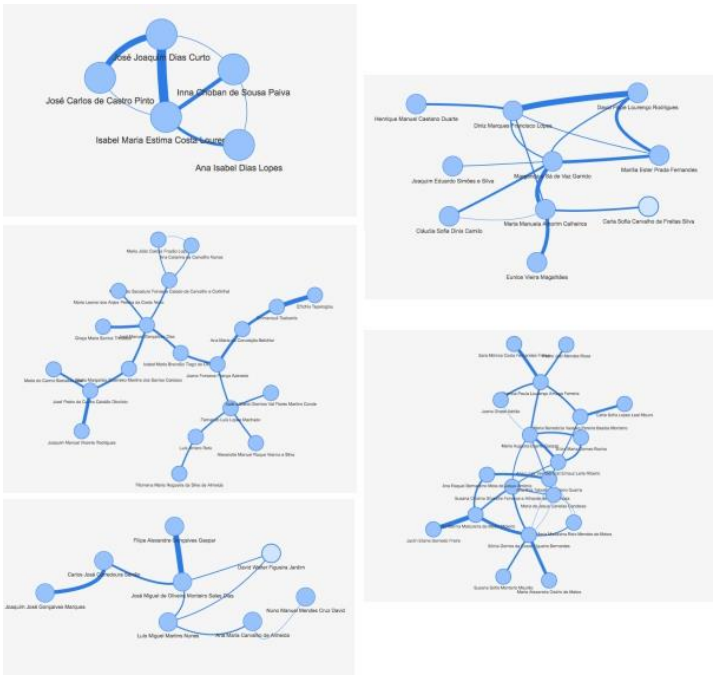

Figura 19 – Configurações de várias comunidades identificadas com o ABCD

6.2. Identificação de Comunidades com o MCL

Executando o algoritmo MCL identificam-se 26 comunidades (Figura 20 e Tabela 9), com um total de 247 autores, 703 coautorias e 1263 publicações (Fernandes, 2017). Se comparados com os valores da rede de coautorias, estes valores correspondem a 40%, 33% e 40%, respetivamente.

#	Comunidade	Nº Autores	Nº Coautorias	Autor mais influente	Categorias Scimago
1	Business Research Unit & Marketing, Operation and Management & Quantitative Methods for Management and Economics	7	8	Álvaro Augusto da Rosa	Computer Science (miscellaneous) , Computer Science Applications , Law , Library and Information Sciences , Social Sciences (miscellaneous)
2	Research and Studies in Sociology & Sociology & Social Research and Intervention	39	121	Patrícia Durães Ávila	Sociology and Political Science , Social Sciences (miscellaneous) , Strategy and Management , Business and International Management , Computer Science (miscellaneous)
3	Research and Studies in Sociology & Sociology & Socioeconomic Change and Territorial Studies	5	7	José Soares da Silva Neves, Maria João Soares Almeida Pedroso de Lima	Political Science and International Relations , Sociology and Political Science , Social Sciences (miscellaneous)

4	Information Sciences, Technologies and Architecture Research Center & Architecture and Urbanism & Information Science and Technology	5	7	Alexandra Cláudia Rebelo Paio	Biomedical Engineering , Biotechnology , Cultural Studies , Health (social science) , History and Philosophy of Science
5	Socioeconomic Change and Territorial Studies & Political Economy & Social Research Methods	7	9	Fátima Suleman, Ricardo Nuno Ferreira Paes Mamede	Management of Technology and Innovation , Sociology and Political Science , Organizational Behavior and Human Resource Management , Strategy and Management , Social Sciences (miscellaneous)
6	Accounting & Business Research Unit & Finance	8	13	Inna Choban de Sousa Paiva	Business, Management and Accounting (miscellaneous) , Gender Studies , Computer Science (miscellaneous)
7	Research and Studies in Sociology & Sociology & Social Research Methods	9	14	Pedro António da Silva Abrantes	Computer Science (miscellaneous) , Social Sciences (miscellaneous) , Education , Sociology and Political Science , Development

8	Research and Studies in Sociology & Sociology	12	28	Gustavo Alberto Guerreiro Seabra Leitão Cardoso	Computer Science (miscellaneous) , Communication , Social Sciences (miscellaneous) , Sociology and Political Science , Biomedical Engineering
9	Socioeconomic Change and Territorial Studies & Architecture and Urbanism & Research and Studies in Sociology	7	7	Mafalda Gambutas Teixeira de Sampaio	Architecture , Mathematics (miscellaneous) , Visual Arts and Performing Arts
10	Social Research and Intervention & Social and Organizational Psychology	5	6	Thomas Wolfgang Schubert	Social Psychology , Developmental and Educational Psychology , Sociology and Political Science , Agricultural and Biological Sciences (miscellaneous) , Biochemistry, Genetics and Molecular Biology (miscellaneous)
11	Business Research Unit & Human Resources and Organizational Behavior & Social Research and Intervention	13	27	Ana Margarida Soares Lopes Passos	Applied Psychology , Organizational Behavior and Human Resource Management , Strategy and Management , Social Psychology , Business and

					International Management
12	Social Research and Intervention & Social and Organizational Psychology	11	23	Sónia Gomes da Costa Figueira Bernardes	Medicine (miscellaneous) , Social Psychology , Psychology (miscellaneous) , Geriatrics and Gerontology , Applied Psychology
13	Business Research Unit & Marketing, Operation and Management & Information Sciences, Technologies and Architecture Research Center	12	20	Hélia Maria Gonçalves Pereira	Marketing , Business and International Management , Computer Science (miscellaneous) , Tourism, Leisure and Hospitality Management , Artificial Intelligence
14	Socioeconomic Change and Territorial Studies & Political Economy & Architecture and Urbanism	11	21	Pedro Miguel Alves Felício Seco da Costa	Arts and Humanities (miscellaneous) , Geography, Planning and Development , Urban Studies , Political Science and International Relations , Sociology and Political Science
15	Social Research and Intervention & Social and Organizational Psychology	13	23	Francisco Gomes Esteves	Psychology (miscellaneous) , Developmental and Educational Psychology ,

					Education , Health (social science) , Public Health, Environmental and Occupational Health
16	Social Research and Intervention & Social and Organizational Psychology & Political Science and Public Policy	8	19	Carla Marina Madureira de Matos Moleiro, Marta dos Santos Nogueira Gonçalves Pimenta de Brito, Jaclin Elaine Semedo Freire, Sandra Gaspar Roberto	Psychology (miscellaneous) , Computer Science (miscellaneous) , Public Health, Environmental and Occupational Health , Life-span and Life-course Studies , Religious Studies
17	Telecomunicações-IUL & Information Science and Technology & Information Sciences, Technologies and Architecture Research Center	8	20	Sancho Moura Oliveira, Maria João Marques de Oliveira	Computational Mathematics , Agricultural and Biological Sciences (miscellaneous) , Biochemistry, Genetics and Molecular Biology (miscellaneous) , Medicine (miscellaneous) , Artificial Intelligence
18	Information Sciences, Technologies and Architecture Research Center & Architecture and Urbanism &	12	20	Sara Eloy Cardoso Rodrigues	Computer Science (miscellaneous)

	Information Science and Technology				
19	Information Science and Technology & Telecomunicações-IUL	9	21	Francisco António Bucho Cercas, Pedro Joaquim Amaro Sebastião, Nuno Manuel Branco Souto	Electrical and Electronic Engineering , Aerospace Engineering , Computer Science Applications , Computer Science (miscellaneous) , Applied Mathematics
20	Business Research Unit & Information Sciences, Technologies and Architecture Research Center & Accounting	7	8	Raul Manuel Silva Laureano	Strategy and Management , Business and International Management , Public Administration , Computer Science (miscellaneous) , Economics, Econometrics and Finance (miscellaneous)
21	Research and Studies in Sociology & Sociology & Anthropology	5	7	Lígia Sofia Alves Passos Ferro	Sociology and Political Science , Political Science and International Relations , Computer Science (miscellaneous) , Geography, Planning and Development , Urban Studies

22	Research and Studies in Sociology & History & Research in Anthropology	12	16	Maria Luísa Macedo Ferreira Veloso	Sociology and Political Science , Education , Social Sciences (miscellaneous) , Computer Science (miscellaneous) , Economics, Econometrics and Finance (miscellaneous)
23	Socioeconomic Change and Territorial Studies & Political Economy & Quantitative Methods for Management and Economics	6	8	Cristina Maria Paixão de Sousa	Management of Technology and Innovation , Business and International Management , Applied Psychology , Environmental Science (miscellaneous) , Renewable Energy, Sustainability and the Environment
24	Research and Studies in Sociology & Research in Anthropology & Sociology	6	8	Elsa Beatriz Padilla, Erika Masanet Ripoli	Medicine (miscellaneous) , Communication , Linguistics and Language , Management, Monitoring, Policy and Law , Public Administration
25	Social and Organizational Psychology & Social	5	7	Diniz Marques Francisco Lopes	Psychology (miscellaneous) , Social Psychology

	Research and Intervention & Human Resources and Organizational Behavior				, Developmental and Educational Psychology , Sociology and Political Science , Arts and Humanities (miscellaneous)
26	Social Research and Intervention & Social and Organizational Psychology	5	6	Maria Manuela Amorim Calheiros	Developmental and Educational Psychology , Education , Sociology and Political Science , Social Work , Psychology (miscellaneous)

Tabela 9 – Comunidades identificadas com o MCL

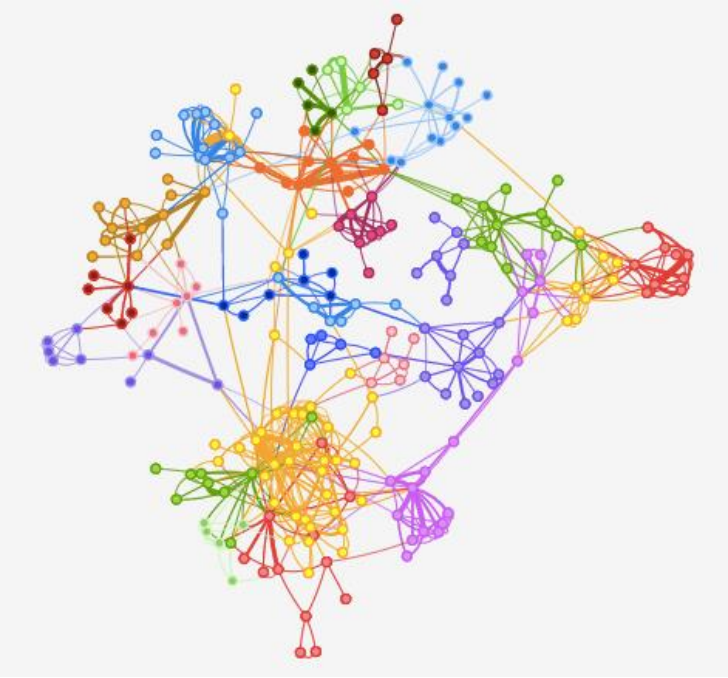

Figura 20 – Comunidades identificadas com o MCL

As 26 comunidades são compostas em média por 10 autores e 18 coautorias. Na Figura 21 pode-se observar as diferentes configurações das comunidades detetadas com o MCL.

A maior comunidade foi a *Research and Studies in Sociology & Sociology & Social Research and Intervention* com 39 autores e 121 coautorias. A autora mais influente foi a professora Patrícia Durães Ávila e a comunidade é caracterizada pelas categorias *Scimago*: *Sociology and Political Science, Social Sciences (miscellaneous), Strategy and Management, Business and International Management, Computer Science (miscellaneous).* De notar que as categorias *Scimago* são todas das áreas de Gestão e Sociologia, exceto a de *Computer Science*. Esta última repete-se 13 vezes nas 26 comunidades.

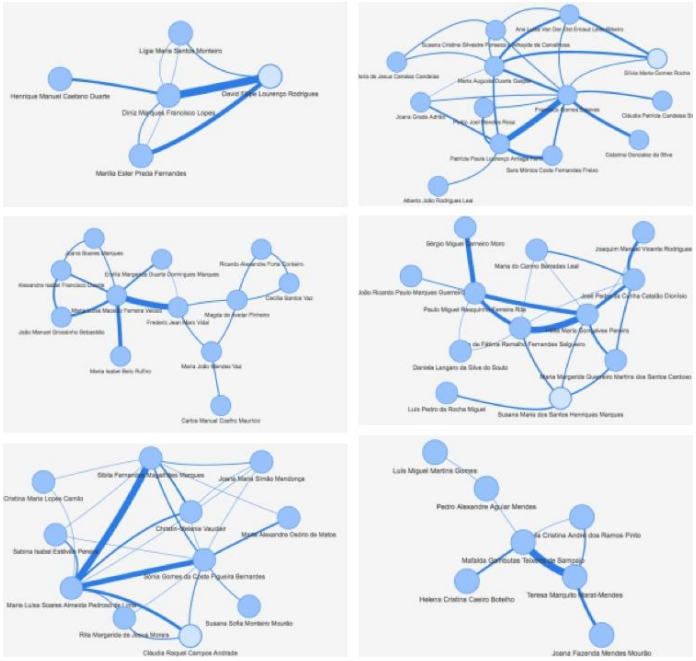

Figura 21 – Configurações de várias comunidades identificadas com o MCL

6.3. Análise Comparativa entre o MCL e o ABCD

Ambos os algoritmos, MCL e ABCD, e apesar de utilizarem diferentes metodologias identificaram sensivelmente o mesmo número de comunidades: 26 o MCL e 25 o ABCD. Em ambos os casos, e tendo em conta que não foram tidas em conta comunidades com menos de 4 autores, também o número de autores e coautorias foi semelhante, com 263 autores e 754 coautorias para o ABCD e 247 autores e 703 coautorias para o MCL. Saliente-se que estes valores indicam que da totalidade de autores considerados na rede de coautorias, apenas cerca de 40% dos mesmos foram agrupados em comunidades com, pelo menos, 4 autores. Uma possível explicação é a fraca força de atratividade entre os vértices expressa pelo peso das arestas e que foi calculada segundo a equação de Newman Foi possível verificar que das 1755 coautorias, 1227 (70%) apresentavam uma força de

atratividade inferior a 1. O valor 1, conceptualmente, representa que entre dois autores apenas existe 1 coautoria. Seja no MCL, seja no ABCD, a atratividade entre os vértices é central, pelo que quanto menor, menos comunidades serão identificadas.

Nas Figura 22 e 23, apresenta-se a distribuição do número de autores e coautorias para cada uma das comunidades obtidas através dos algoritmos ABCD e MCL, respetivamente.

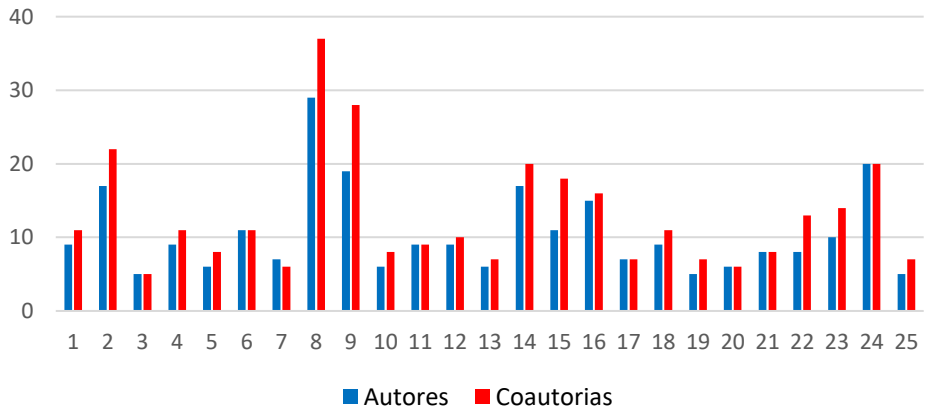

Figura 22 – Algoritmo ABCD, número de autores e coautorias por comunidade

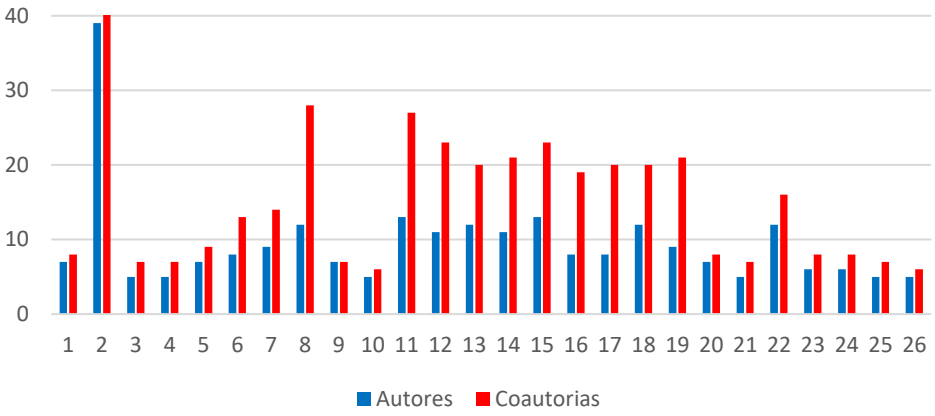

Figura 23 – Algoritmo MCL, número de autores e coautorias por comunidade

Como se pode observar na Figura 22, o número de coautorias nas 25 comunidades por aplicação do algoritmo ABCD é, na maioria das vezes, superior ao número de autores. Apenas em 6 das 25 comunidades é igual (comunidades 2,6,17,20,21 e 24) e em uma (comunidade 7) é inferior. Pode ainda verificar-se que cerca de 68% (17 em 25) das comunidades identificadas tem 10 ou menos autores. Já em relação a coautorias, essa percentagem baixa para os 48% (12 em 25).

As comunidades obtidas por aplicação do algoritmo MCL apresentam uma estrutura semelhante (Figura 23): cerca de 68% (17 em 26) das comunidades identificadas tem 10 ou menos autores e, em relação a coautorias, a percentagem é 48% (12 em 26). Contudo, comparando as Figuras 22 e 23 verifica-se que com o algoritmo MCL foram obtidas comunidades com maior discrepância entre o número de autores e o número de coautorias. Para complementar esta análise comparativa, na Tabela 10 apresenta-se um conjunto de medidas descritivas.

	Algoritmo ABCD		**Algoritmo MCL**	
	Nº de autores	**Nº de coautorias**	**Nº de autores**	**Nº de coautorias**
Média	10,5	12,8	9,5	18,2
Mínimo	5	5	5	6
Máximo	29	37	39	121
Desvio padrão	5,9	6,6	6,6	22,2
Moda	9	5	5	7
1º quartil	8	10	7	16
Mediana	11	11	11	21
3º quartil	19	22	13	121

Tabela 10 – Medidas descritivas para o número de autores e de coautorias das comunidades obtidas pelos algoritmos ABCD e MCL

Comparando os resultados apresentados na Tabela 10 no que a número de autores diz respeito podemos observar que a média bem como a moda de autores por comunidade são superiores no algoritmo ABCD se comparados com o algoritmo MCL. Contudo o desvio-padrão é superior no algoritmo MCL, facto que pode ser explicado pelo número máximo de autores por comunidade também ser superior. Comparando agora os três quartis, verifica-se que apenas no 3º quartil a diferença é significativa: no ABCD o valor é 19 enquanto que no MCL é apenas 13. Isto indica-nos que a diferença no número de autores por comunidade nas 50% comunidades centrais é maior no ABCD do que no MCL (19-8=11 > 6=13-7). Poderemos assim dizer que existe uma tendência para obter comunidades mais homogéneas, no que a número de autores diz respeito, com o algoritmo MCL do que com o ABCD. Fazendo uma comparação análogo, mas agora em relação ao número de coautorias podemos concluir que neste caso o algoritmo MCL produz comunidades mais homogéneas.

Ao analisar as 5 categorias *Scimago* que mais aparecem na categorização das comunidades para cada algoritmo chega-se à conclusão que existem 3 categorias que aparecem nas duas contagens: *Sociology and Political Science*, *Computer Science (miscellaneous)* e *Social Sciences (miscellaneous).* (ver Tabela 11 e Tabela 12)

Categoria Scimago	Total
Sociology and Political Science	19
Computer Science (miscellaneous)	18
Education	11
Social Sciences (miscellaneous)	11
Communication	9

Tabela 11 – Categorias Scimago para o algoritmo MCL

Categoria Scimago	Total
Computer Science (miscellaneous)	18
Social Sciences (miscellaneous)	16
Sociology and Political Science	14
Strategy and Management	10
Business and International Management	9

Tabela 12 – Categorias Scimago para o algoritmo ABCD

A categoria *Computer Science* da *Scimago* está representada em mais de 50% das comunidades quer do ABCD quer do MCL, sendo a categoria que mais vezes se repete. Este dado reflete a transversalidade da área da computação, hoje presente em qualquer área científica, seja de forma substantiva ou instrumental. Razão pela qual quase todos os veículos de publicação classificados numa dada área da *Scimago*, estarem também classificados na área de *Computer Science*. Com efeito, vale lembrar que uma mesma publicação pode estar classificada em várias áreas *Scimago*. Isto não significa pois que os conteúdos das publicações nestas comunidades sejam necessários e efetivamente relacionados com *Computer Science*, exceção feita às comunidades com autores da Escola de Tecnologias e Arquitetura.

6.4. Inquérito a Utilizadores da Aplicação

De modo a avaliar preliminarmente e qualitativamente a aplicação desenvolvida, foi construído um questionário e enviado via internet a um grupo selecionado de autores do

ISCTE-IUL, cuja a amostragem refletiu todos os departamentos presentes na rede de coautorias. Foram feitas sete questões de resposta Sim/Não e um campo de texto de livre preenchimento para observações genéricas. O objetivo foi validar se a aplicação pode ser utilizada e disponibilizada à comunidade do ISCTE-IUL, respeitando os requisitos mínimos de usabilidade. Foi assim validado se um autor consegue aceder à rede de coautorias, pesquisar pelo seu nome e aceder às publicações que partilha com um coautor em particular. É, pois, assim validada também a informação recolhida do Ciência-IUL e se a mesma foi corretamente inserida na aplicação. Foram respondidos 19 inquéritos, de modo confidencial, sendo que as respostas a cada questão do questionário encontram-se mapeadas na Tabela 13.

Questão	Sim	Não
Consegue fazer Login na aplicação?	16 (84%)	3 (16%)
Consegue visualizar a rede de coautorias (opção Compute Network)?	14 (73%)	5 (27%)
Consegue pesquisar autores através do campo de pesquisa de autores?	13 (68%)	6 (32%)
Consegue ver o seu grafo pessoal de coautorias após pesquisar o seu nome no campo de pesquisa de autores?	12 (63%)	7 (37%)
No seu grafo de coautorias, ao clicar na ligação entre si e outro autor consegue ver a lista de publicações que partilham?	8 (42%)	11 (58%)
Existe alguma publicação que lhe esteja atribuída e na qual não tenha participado?	3 (16%)	16 (84%)
Existem publicações em que tenha participado e que não lhe estejam atribuídas?	5 (26%)	15 (74%)

Tabela 13 – Perguntas e respostas dos inquéritos

De realçar que à medida que as operações se tornam mais complexas, mais difícil é chegar à área funcional desejada. O primeiro passo, entrar na aplicação, tem uma taxa de sucesso de 84%, sendo que o segundo passo, visualizar a rede de coautorias apresentada por defeito n a página inicial, desce para 73%. Por fim, no terceiro passo, pesquisar na caixa de autores, a taxa de sucesso é apenas 68%. Este padrão pode ser explicado pela complexidade da página inicial, com área de visualização e várias opções de pesquisa; o utilizador poderá sentir-se algo perdido na mesma. Contudo, com uma taxa de sucesso na ordem dos 70% a realizar estas três ações, pode concluir-se pela robustez na interface desenvolvida em termos de disponibilização de funcionalidades.

O ponto mais negativo refere-se à menor capacidade para um autor ver a lista de publicações que partilha com outro coautor, uma ação em que apenas 42% dos inquiridos tiveram sucesso em executar. Após análise da interface, este resultado foi atribuído ao facto de a lista aparecer por baixo da área do grafo sem nenhum aviso especial; o inquirido pode nem sequer se ter apercebido que a mesma já estava disponível.[4] No entanto, os que conseguiram visualizar e analisar a lista de publicações concluíram que as mesmas eram representativas do seu trabalho: 84% não identificaram a existência de qualquer publicação que lhe estivesse atribuída e na qual não tivesse participado.

Na caixa de observações de texto foram feitos vários comentários pertinentes para avaliar a qualidade da aplicação. Por duas vezes foi feito o comentário que foi impossível realizar a autenticação na aplicação, pelo que não conseguiram responder ao questionário. Para entrar na aplicação é necessário utilizar a conta de email do ISCTE-IUL. Contudo, como o utilizador pode ter várias contas do Google autenticadas ao mesmo tempo (e a conta do ISCTE-IUL é gerida pela Google), poderá ser necessário sair de todas para de seguida se autenticar apenas com a conta do ISCTE-IUL.[5]

Em três comentários distintos foi também referido a dificuldade do autor para encontrar as suas publicações. Isto deve-se ao facto de que para as ver ser necessário carregar numa aresta do grafo desenhado, mais concretamente numa coautoria entre o autor e um coautor. Apenas dessa forma é visualizada a lista de publicações, não existindo uma visão global de publicações de um determinado autor. Associada a esta necessidade, também em duas respostas foi mencionado a necessidade de mais ferramentas de extração de informação já presente na ferramenta. Ou seja, diferentes perspetivas de tratamento da informação já existente na ferramenta.

6.5. Rede de Coautorias com Autores Externos

Os resultados anteriormente reportados foram obtidos tendo em conta apenas os autores internos ao ISCTE-IUL. Contudo, e a título experimental, foi construída uma rede de coautorias com autores externos de modo a avaliar a rede obtida. Foi mantida a regra de descartar publicações com apenas 1 autor e autores sem nenhuma coautoria. Foram obtidos, e em comparação com a rede apenas com autores internos, 6652 autores

[4] Para corrigir esta situação poderá introduzir-se na interface um *focus* para a lista de publicações assim que a mesma for colocada na sua área de visualização.

[5] Por vezes o Google não oferece a possibilidade de selecionar a conta a utilizar e seleciona automaticamente uma conta que não é a do ISCTE-IUL, i.e., a aplicação rejeita a autenticação porque o email não é o institucional.

(+1086%), 9491 publicações (+252%) e 32682 coautorias (+1903%). Em termos absolutos, o volume de informação obtido é muito superior ao da rede apenas com autores internos.

Dado que a equação de Newman apenas depende das publicações, a mesma pode ser utilizada para calcular o peso das arestas entre vértices que representem quer autores internos quer autores externos. O problema que se coloca é que a metainformação disponível para autores externos não é a mesma que para autores internos. A API do Ciência-IUL devolve para os autores externos apenas um identificador único e o seu nome. Ao contrário do que acontece com os autores internos, para os quais é possível identificar a escola, departamento, centros de investigações, URL internet para obter a informação, etc. Esta escassez de informação não permite analisar e categorizar as comunidades identificadas, pelo que apesar de se saber que existem, não sabemos as razões. De realçar que temos as categorias *Scimago* podem ser associadas às publicações e pode-se categorizar a comunidade com base nas mesmas. Contudo, sem nada mais saber sobre os autores, a categorização é incompleta: categorizamos uma relação entre autores, mas que relação é essa?

7. Conclusão

O Ciência-IUL e os seus repositórios de produção científica permitem construir uma rede de coautorias modelada como um grafo. É possível recorrer a algoritmos genéricos como o ABCD e o MCL para identificar comunidades presentes no mesmo. A identificação é automática e genérica, pelo que *a priori*, sem demais interpretação o resultado das comunidades é imprevisível. Para uma adequada interpretação por parte dos utilizadores das razões das mesmas, foi necessário adornar as comunidades com modelos vários de informação, desde a pertença a escolas, departamentos ou áreas científicas.

A força de atratividade entre autores segundo a equação de Newman revelou-se uma boa métrica para atribuir pesos às coautorias, o que possibilitou a utilização de algoritmos como o ABCD e o MCL.

Na rede de coautorias foram detetados 613 autores internos do ISCTE-IUL com, pelo menos, uma publicação partilhada com outro autor interno. No total, os 613 autores partilham entre si 3766 publicações e estão ligados por 1718 coautorias. Em termos de identificação de comunidades, quer o MCL, quer o ABCD, descobriram sensivelmente o mesmo número de comunidades, 26 o MCL, 25 o ABCD. Em ambos os casos também o número de autores e coautorias foi semelhante, com 263 autores e 754 coautorias para o ABCD e 247 autores e 703 coautorias para o MCL. Contudo, isto significa que apenas cerca de 40% dos autores da rede de coautorias foram agrupados em comunidades com, pelo menos, 4 autores.

A solução tecnológica desenvolvida, com *Node.js* e *MongoDB*, mostrou-se eficaz na recolha da informação do Ciência-IUL e construção da rede de coautorias, assim como na execução dos algoritmos de identificação de comunidades. A sua arquitetura modelar permitiu a realização de várias experiências com diferentes módulos e abordagens, o que permite concluir que a sua futura manutenção também será possível. A base de dados permitiu persistir a rede de coautorias e os grafos com comunidades para análises profundas e metódicas ao longo do tempo através da interface gráfica.

Na avaliação por inquérito a um conjunto de autores do ISCTE-IUL sobre a primeira versão da aplicação, identificaram-se um conjunto de questões em termos de funcionalidade da interface que ajudam na sua revisão para uma melhor utilização da aplicação desenvolvida. No entanto, os resultados foram satisfatórios. Em termos da qualidade e correção da informação recolhida do Ciência-IUL, e da informação resultante

do processamento da aplicação, a taxa de sucesso foi superior a 80% em várias questões do inquérito. Concluímos que existe confiança na aplicação desenvolvida e nas possibilidades futuras da utilização da mesma para análise das colaborações científicas do ISCTE-IUL.

7.1. Trabalho Futuro

Para lá da utilização de diferentes algoritmos no futuro, também se podem utilizar outras funções de atratividade, diferentes da equação de Newman. Utilizar outras funções de atratividade pode levar à identificação de comunidades diferentes.

A API do Ciência-IUL devolve tipos distintos de publicações para cada autor. Neste trabalho considerou-se que algumas colaborações são incluídas e outras não, de acordo com determinados critérios, não sendo ponderado um peso específico para cada um. Para trabalho futuro poderia seria interessante atribuir-se pesos diferentes a cada tipo de publicação (uma coautoria num livro é mais importante que uma coautoria numa conferência?) de modo a obterem-se atratividades diferentes para as mesmas coautorias, com as mesmas publicações, mas com diferentes ponderações sobre as mesmas, o que poderá culminar em resultados significativamente diferentes dos obtidos.

Nos algoritmos de identificação de comunidades tem-se que ter em consideração o esforço computacional e o tempo despendido na identificação das mesmas, nomeadamente através da análise da complexidade assintótica (Aaron Clauset, 2004). A rede de coautorias atuais não apresenta uma grande dimensão, em termos de números totais de autores, publicações e coautorias, pelo que a questão da complexidade não se colocou. Contudo, no futuro, o número de autores, assim como de coautorias e publicações, irá inevitavelmente aumentar e a problemática do esforço computacional terá que ser equacionada e tida em conta.

A partição de um grafo em agrupamentos não necessita ser plana, ou seja, de um nível, podendo ter uma estrutura hierárquica (Schaeffer, 2007). Cada partição global pode desdobrar-se em vários agrupamentos até um determinado nível caso seja aplicada um algoritmo que o permita fazer. Estes algoritmos normalmente criam um dendrograma com a estrutura hierárquica das comunidades identificadas. A identificação de comunidades dentro das comunidades numa comunidade de redes de coautorias permitirá fazer uma análise mais granular dos padrões de partilha de conhecimento.

Na identificação de comunidades foi feita uma restrição ao número de comunidades que um autor pode pertencer: apenas uma. Seria também interessante analisar a possibilidade um mesmo autor estar presente em diferentes comunidades simultaneamente.

Em termos de avaliação da solução, não foram validadas as comunidades identificadas em testes de replicação. Um trabalho futuro possível será aceder à base de dados da solução desenvolvida e, com uma implementação diferente, identificar comunidades com os mesmos algoritmos de modo a validar as comunidades identificadas neste trabalho.

Referências Bibliográficas

Aaron Clauset, M. E. (2004). Finding community structure in very large networks. *Phys. Rev. E*, 70:066111.

Cervera, A. M. (2010). *Discovering and analyzing scientific communities using conference network.* Final Project, Computer Science, Universidad Catolica Nuestra Senora de la Asuncion.

Ciência-IUL. (11 de 2016). *Documentação da API Pública*. Obtido de Ciência-IUL: https://ciencia.iscte-iul.pt/api/doc

Deepjyoti Choudhury, A. P. (2013). Community Detection In Social Networks: An Overview. *International Journal of Research in Engineering and Technology*, 83-88.

Dongen, S. v. (2000). *Graph Clustering by Flow Simulation.* PhD thesis, University of Utrecht.

Dongen, S. v. (01 de 2017). *MCL - a cluster algorithm for graphs*. Obtido de http://micans.org/mcl/

Fernandes, D. (2017). Identification and Evaluation of Communities in Co-authorship Networks.

GEXF Working Group. (05 de 03 de 2017). *GEXF File Format*. Obtido de GEXF File Format: https://gephi.org/gexf/format/

Google. (4 de 2017). *Google Sign-In for Websites*. Obtido de Google Sign-In for Websites: https://developers.google.com/identity/sign-in/web/devconsole-project

ISCTE-IUL. (12 de 2016). Obtido de ISCTE-IUL: www.iscte-iul.pt

Lab, M. M. (26 de 12 de 2016). *Visualizing Clinton, Podesta, and the DNC's Wikileaks E-mail Networks*. Obtido de https://clinton.media.mit.edu/

Leonard Richardson, S. R. (2007). *RESTful Web Services.* O'Reilly Media.

MongoDB. (6 de 2 de 2017). *MongoDB*. Obtido de MongoDB: https://www.mongodb.com

Newman, M. E. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 5200-5205.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 8577-8582.

Node.js. (6 de 2 de 2017). *Node.js*. Obtido de Node.js: https://nodejs.org

Pinheiro, D. F. (2016). *Identificação de Comunidades no ISCTE-IUL usando Bases de Dados e Técnicas de Análise de Redes Sociais.* Dissertação de Mestrado, ISCTE-IUL.

Ruifang Liua, S. F. (2014). Weighted graph clustering for community detection of large social networks. *2nd International Conference on Information Technology and QuantitativeManagement, ITQM*, 85-94.

Ruohonen, K. (2013). *Graph Theory.* Self-publishing.

Sample Resume Template. (01 de 02 de 2017). *Sample Resume Template*. Obtido de Sample Resume Template: http://sampleresumetemplate.net/

Schaeffer, S. E. (2007). Graph clustering. *Computer Science Review*, 27-64 .

Schulz, C. (2016). *Graph Partitioning and Graph Clustering in Theory and Practice.* Lecture Notes, Institute for Theoretical Informatics - Karlsruhe Institute of Technology.

SCImago. (11 de 2016). *Scimago Journal & Country Rank.* Obtido de Scimago Journal & Country Rank: http://www.scimagojr.com/

vis.js. (01 de 02 de 2017). *vis.js - A dynamic, browser based visualization library. .* Obtido de vis.js - A dynamic, browser based visualization library. : http://visjs.org

Anexos

Anexo A – Instalação da aplicação em macOS

A solução é uma aplicação web com três camadas: cliente – servidor – base de dados. Para instalar a solução numa máquina é necessário instalar as camadas de servidor e base dados. A camada de cliente é acedida por um navegador *web* comercial. As seguintes instruções de instalação foram feitas num sistema operativo macOS e servem como referência para instalar em outros sistemas operativos:

1. Para instalar o Node.js basta ir à página oficial (nodejs.org) e fazer download da aplicação. Ao instalar a aplicação irá receber a mensagem que irá também instalar o software *npm*. O *npm* é o gestor oficial de pacotes e bibliotecas desenvolvidas em código aberto do Node.js.
2. Para instalar o MongoDB é necessário primeiro instalar o Homebrew, um gestor de pacotes do macOS:

   ```
   /usr/bin/ruby -e "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/master/install)"
   brew install mongodb
   ```

3. Para o MongoDB funcionar é necessário criar a pasta onde as bases de dados irão ser guardadas no sistema de ficheiros:

   ```
   mkdir -p /data/db
   sudo chown -R `id -u` /data/db
   ```

4. Colocar o MongoDB a ser executado:

   ```
   mongod
   ```

5. Para criar a base de dados com o nome *redecoautorias* e as coleções onde os objetos irão ser guardados no MongoDB:

   ```
   mongo
   use redecoautorias
   db.createCollection('authors')
   db.createCollection('publications')
   db.createCollection('coauthorships')
   db.createCollection('authorsInCommunity')
   db.createCollection('publicationsInCommunity')
   db.createCollection('coauthorshipsInCommunity')
   db.createCollection('communityJobs')
   ```

O código da solução encontra-se nas pastas *core* e *model* e foi desenvolvido em *JavaScript*, a linguagem interpretada pelo servidor Node.js. Já na pasta *public* encontra-se o código que será utilizado pelo cliente no seu navegador *web*. Para a correr na máquina onde se instalou o Node.js e o MongoDB deve-se colocar o código fonte numa diretoria à escolha. Na raiz da diretoria encontra-se o ficheiro *package.json*. Neste ficheiro encontra-se a discrição geral da aplicação e as dependências de bibliotecas externas. Para as mesmas serem instaladas na pasta *node_modules* tem-se de executar o comando:

```
npm install
```

O ficheiro *server.js* é definido como o ponto de entrada da aplicação no ficheiro *package.json*. Basta executar o seguinte comando para correr a aplicação:

```
npm start
```

Para o algoritmo de identificação de comunidades MCL funcionar é necessário compilar o código da aplicação externa utilizada no computador onde a solução está a ser executada. Na pasta *core\MCL* encontra-se o código fonte num arquivo, *mcl-latest.tar.gz*. Após o descomprimir, deve-se seguir as instruções presentes no ficheiro *INSTALL* de modo a instalar a aplicação. Após instalado, o mesmo fica acessível a nível de sistema através do comando *mcl*.

Anexo B – Instalação da aplicação na Amazon AWS

A *Amazon Web Services* (AWS) é uma plataforma da empresa Amazon que oferece serviços na nuvem, nomeadamente poder computacional e bases de dados. Para demonstrar o funcionamento da solução foi criado um servidor na Amazon AWS através do produto *Amazon Elastic Cloud Computing* (EC2). Este servidor tem a camada aplicacional (Node.js) e a base de dados (MongoDB), estando exposto à internet. Este serviço tem custos associados que devem ser considerados e ponderados.

1. Criar uma conta na Amazon AWS (aws.amazon.com);
2. Selecionar o serviço EC2 (*Elastic Compute Cloud*) e de seguida selecionar a opção *Launch Instance*;
3. Na configuração do servidor, escolher o sistema operativo *Amazon* Linux e a configuração de máquina *t2.medium*;
4. Depois de configurado será pedido para criar um par de chaves de segurança que permitirem aceder ao servidor;

5. Na consola de gestão do serviço EC2 é possível obter o endereço público DNS para aceder ao servidor;
6. Aceder pelo protocolo SSH através da linha de comando ao servidor respeitando as instruções presentes na consola de gestão do serviço EC2;
7. Atualizar o software do servidor com: *sudo yum update*;
8. Instalar o Node.js e o *npm*:

```
sudo yum install gcc-c++ make
sudo yum install openssl-devel
sudo yum install git
git clone https://github.com/nodejs/node.git
cd node
git checkout tags/v7.9.0
./configure
make
sudo make install
sudo vi /etc/sudoers
        - encontrar a linha secure_path = /sbin:/bin:/usr/sbin:/usr/bin e acrescentar no final :/usr/local/bin
        - carregar ESC e escrever :w!
git clone https://github.com/npm/npm.git
cd npm
sudo make install
```

9. Executar o comando *node* na consola para validar se a instalação correu bem;
10. Instalar a base de dados MongoDB (ficando a correr como serviço, mesmo em caso de *reboot* do servidor):

```
echo "[10gen]
name=10gen Repository
baseurl=https://repo.mongodb.org/yum/amazon/2013.03/mongodb-org/3.3/x86_64/
gpgcheck=0" | sudo tee -a /etc/yum.repos.d/10gen.repo
echo "[10gen]
name=10gen Repository
baseurl=http://downloads-distro.mongodb.org/repo/redhat/os/x86_64
gpgcheck=0" | sudo tee -a /etc/yum.repos.d/10gen.repo
sudo yum -y install mongo-10gen-server mongodb-org-shell
sudo yum -y install sysstat
sudo mkdir /var/lib/mongo/data
sudo mkdir /var/lib/mongo/log
sudo mkdir /var/lib/mongo/journal
sudo chown mongod:mongod /var/lib/mongo/data
sudo chown mongod:mongod /var/lib/mongo/log
sudo chown mongod:mongod /var/lib/mongo/journal
sudo chkconfig mongod on
```

```
sudo /etc/init.d/mongod start
```

11. Executar o comando *mongo* na consola para validar se a instalação correu bem;
12. Para criar a base de dados com o nome *redecoautorias* e as coleções onde os objetos irão ser guardados no MongoDB:

```
mongo
use redecoautorias
db.createCollection('authors')
db.createCollection('publications')
db.createCollection('coauthorships')
db.createCollection('authorsInCommunity')
db.createCollection('publicationsInCommunity')
db.createCollection('coauthorshipsInCommunity')
db.createCollection('communityJobs')
```

13. Criar uma pasta para o código da solução na raiz do sistema de ficheiros:

```
mkdir redecoautorias
```

14. Fazer *upload* do código da solução (app.zip) para a pasta criada. Uma possibilidade é utilizar o software *FileZilla* com a opção SFTP;
15. Compilar a aplicação do MCL presente na pasta *core/MCL/ mcl-latest.tar.gz*;
16. No ficheiro de configuração *core/configuration.js* alterar a propriedade *this.googleCallbackURL* para o endereço público HTTP do servidor EC2. Para a autenticação na Google continuar a funcionar este endereço tem de ser adicionado à lista de endereços da aplicação na Google API Console (console.developers.google.com);
17. Na raiz da pasta onde está o código da solução, instalar o *pm2* que permitirá manter o *Node.js* a funcionar mesmo após desligar o terminal:

```
npm install pm2 -g
```

18. Na raiz da pasta onde está o código da solução, instalar todas as dependências da mesma com:

```
npm install
```

19. Iniciar o *Node.js* com:

```
sudo pm2 start server.js
```

20. O *Node.js* encontra-se a correr na porta definida no ficheiro server.js. No caso deste guia encontra-se na porta 80. É necessário na consola de gestão do serviço EC2 permitir que o servidor seja acedido através da porta 80 via internet (*Network & Security ->Security Groups -> Edit inbound rules*).

21. Ao aceder através do endereço DNS público através de um navegador web será possível ver a página inicial da solução.