**Matemática das redes**

Análise de redes de coautoria científica



Bruno Moreira A82981

Filipe Viana A85948

Gerson Júnior A88000

Marco Lima A86030

Pedro Oliveira A75149

Tiago Rodrigues A81839

Departamento de Matemática e Aplicações

Universidade do Minho

Monografia submetida no âmbito de

*Matemática das Coisas*

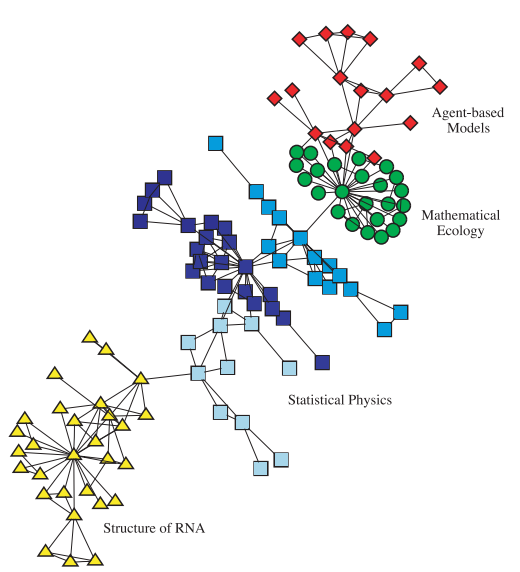
1. **Sumário**

O objetivo deste trabalho foi o de analisar redes de coautoria na comunidade científica. Para tal, recorreu-se às bases de dados Los Alamos e-Print Archive, MEDLINE, SPIRES e NCSTRL, para obtenção dos dados de publicação de artigos científicos. Para a elaboração das redes descreveram-se dois métodos, sendo que o primeiro em que se identifica a inicial do primeiro nome e o apelido do autor, foi o método utilizado.

Numa primeira parte foi feita a representação gráfica da distribuição do número de autores e do número de colaboradores que cada um possuía de modo a verificar se estas apresentavam um comportamento que seguisse uma lei da potência. Posteriormente foram feitas análises para a determinação das percentagens de “*betwenness*” e compararam-se os valores de distância média entre cientistas chegando-se a conclusões que levaram à criação de uma rede pesada da relação entre cientistas para uma melhor análise da relação entre cientistas.

**2. Introdução**

O estudo de redes complexas encontra-se, nas últimas décadas, num período de expansão acelerada, principalmente após a introdução dos modelos da lei da potência e de redes “*scale-free”*, que deu a origem do estudo de diversos fenómenos que de outro modo não poderiam ser estudados.

Um dos objetos de estudo na análise de redes complexas são redes sociais, onde cada vértice será uma pessoa ou entidade e cada aresta será a relação estabelecida entre dois vértices. No caso em concreto de redes de coautoria científica, cada vértice representa um cientista e cada aresta que liga diferentes vértices representa a ligação entre dois cientistas, sendo neste caso a ligação o facto de serem coautores de um ou mais artigos. Um exemplo de uma rede formada desta forma é possível observar-se no artigo “Coauthorship networks and patterns of scientific collaboration”, publicado por M. Newman e pode verificar-se na figura seguinte.

Ainda que grande parte dos artigos científicos sejam escritos “localmente”, ou seja, por cientistas que trabalham na mesma instituição, ou instituições próximas, a ciência tem como um dos objetivos latentes a globalização, já que procura espalhar informação por todo o mundo. Torna-se então importante analisar redes de coautoria científica, não só para uma análise geral da mesma, mas para perceber que autores são os mais vitais nas mesmas, e por isso, mais centrais na difusão da informação científica.

O estudo destas redes permite, portanto, ganhar um maior entendimento acerca do funcionamento da ciência, já que permite identificar algumas características que fazem alguns autores serem mais conhecidos que outros e que fazem alguns indivíduos serem mais vezes escolhidos como colaboradores em artigos de outros cientistas.

**3. Análise de redes de coautoria científica**

Num artigo intitulado “Who is the best connected scientist? A study of scientific coauthorship networks”, Newman recolheu informações de diferentes bases de dados de artigos publicados em diferentes áreas científicas com o intuito de analisar o comportamento das redes de coautoria formadas a partir desses dados.

Para tal foram recolhidos dados das bases de dados Los Alamos e-Print Archive (uma base de dados de artigos de Física que permite a categorização dos artigos segundo a área específica da Física sobre a qual o artigo incide, entre elas encontram-se por exemplo a Física da matéria condensada ou a Física teórica das altas energias), MEDLINE (uma base de dados que contem o registo minucioso de artigos científicos no âmbito de estudos biomédicos, que representam uma grande parte dos estudos científicos mundiais), SPIRES (uma base de dados de artigos de Física de altas energias tanto teóricos como experimentais, que normalmente apresentam uma lista extensa de coautores) e NCSTRL (uma base de dados de artigos de ciência computacional, cujos artigos são submetidos por instituições parceiras à base de dados, o que pode indicar uma falta de artigos publicados presente na mesma).

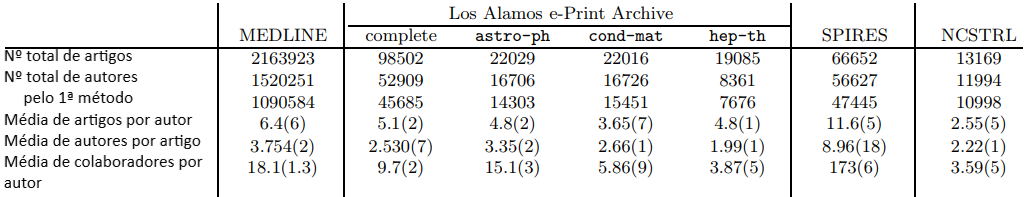
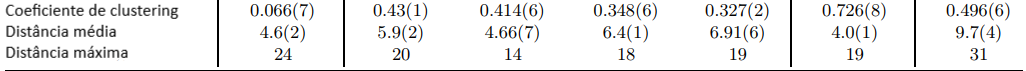
As redes foram formadas a partir dos artigos publicados desde 1995 até 1999, já que, ainda que algumas bases de dados, como a MEDLINE possuíssem informação extensa sobre artigos publicados desde 1961, o mesmo não era verdade para outras bases de dados, pelo que se escolheu um período de tempo para o qual a informação presente em todas as bases de dados era fidedigna. Desta forma, garante-se uma maior confiança nas conclusões que se podem tirar da análise dos dados, além de que permite uma melhor comparação entre as redes obtidas a partir de cada base de dados.

É ainda de notar que um mesmo autor pode publicar diferentes artigos segundo “diferentes nomes”, uma vez que, por exemplo no caso de Newman, este pode publicar artigos segundo M. E. J. Newman, Mark. E. J. Newman, M. Newman, entre outras possibilidades. Há também a possibilidade de existirem autores diferentes que publiquem segundo o mesmo nome. Assim, será necessário um método para a formação das redes de coautoria tendo em conta a existência destes possíveis problemas.

Num primeiro método pode-se ter conta apenas a inicial do primeiro nome e o último do nome do autor, o que garantirá que todos os artigos de um mesmo autor serão atribuídos a este, contudo também se gera a possibilidade de atribuir a autoria de artigos a um autor que não contribui para estes, simplesmente porque partilha a inicial do primeiro nome e o último nome com um outro autor.

Num segundo método tem-se em conta não só o que se tinha em conta no primeiro método, como todas as iniciais que surgem no meio da primeira inicial e do último nome. Este método reduz a probabilidade da atribuição da autoria de artigos a cientistas indevidos, contudo está mais propenso a criar diversos autores de artigos referentes a um mesmo autor que escreveu o nome de forma diferente na publicação de diferentes artigos.

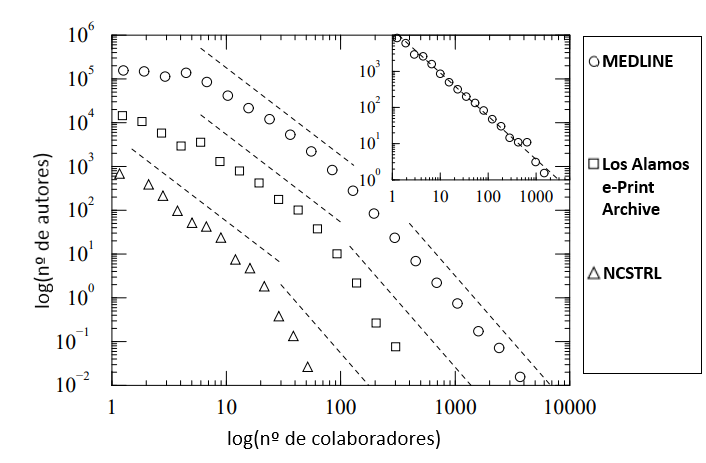
Através dos dois métodos serão obtidas duas redes diferentes, sendo que nenhuma delas representará exatamente a realidade, uma vez que é provável que a primeira condense a informação de diferentes cientistas debaixo de um mesmo nomes e que a segunda espalhe a informação de um mesmo cientista debaixo de diferentes nomes. No entanto a análise destas redes ainda se prova bastante útil, uma vez que estas apresentam um limite inferior e superior para a realidade.

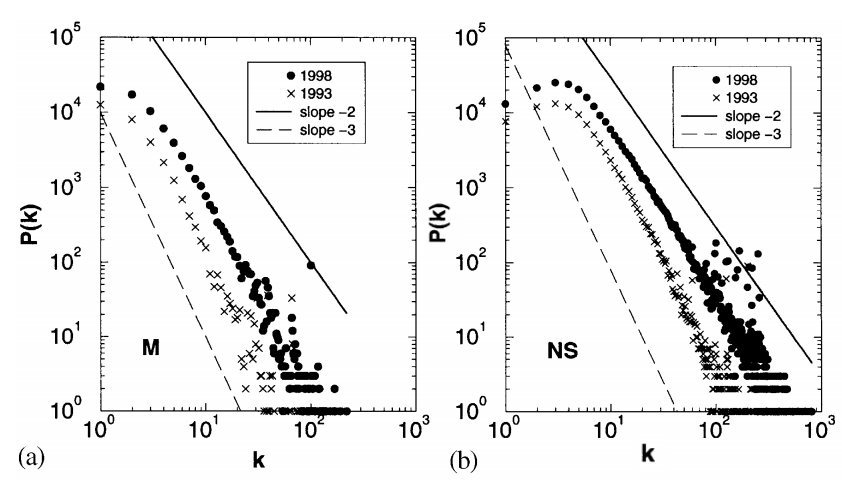
Para efeitos de análise serão utilizadas as redes obtidas a partir do primeiro método. Através das informações obtidas a partir das bases de dados, é então possível obter-se a tabela seguinte.

Na tabela anterior encontra-se a informação geral de cada base de dados, contudo pode-se fazer uma análise mais a fundo sobre cada rede, a partir da qual se podem tirar algumas conclusões sobre a forma como se faz a comunicação entre cientistas.

--------------------------------------------------------------------------------------------------------------------------------

Numa primeira análise foi feita a representação gráfica do número de autores e do número de colaboradores que cada um possuía, ambos colocados em escalas logarítmicas, de modo a verificar se estas apresentavam, como era esperado de uma rede complexa, um comportamento que seguisse uma lei da potência.



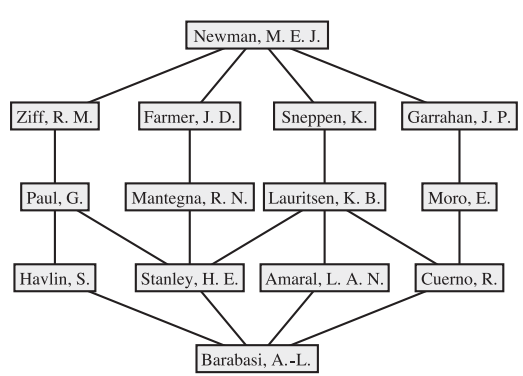
Como verificado na figura, a distribuição do número de autores pelo seu número de colaboradores aparenta seguir não uma lei da potência, mas sim duas leis da potência, para todas as bases de dados com exceção da SPIRES. Este comportamento da distribuição já era esperado por Barabási, que no artigo “Evolution of the social network of scientific collaborations” faz o estudo do desenvolvimento e evolução de 2 redes de coautoria científicas no período de 1991 a 1998, na área da Matemática e da Neurociência. Neste artigo, Barabási, apresenta a distribuição de grau obtida para as redes obtidas para os dados nos períodos de 1991-1993 e 1991-1998.

Nesta distribuição é possível verificar que a distribuição segue a lei das potências de exponencial -2 no inicio, mas passa a ser proporcional a um exponencial -3 no final.

Sabendo então que este comportamento era esperado foram traçadas linhas a tracejado com declive -2 e -3 para comparar os mesmos com o comportamento da distribuição. Daí é possível verificar que os mesmos se adequam bastante bem à distribuição obtida a partir dos dados da MEDLINE, contudo, ainda que o declive de -2 se pareça adequar à distribuição inicial dos dados da Los Alamos e da NCSTRL, o mesmo não se pode dizer do declive -3, já que para as duas bases de dados o declive no final da distribuição aparenta ser maior em módulo do que 3.

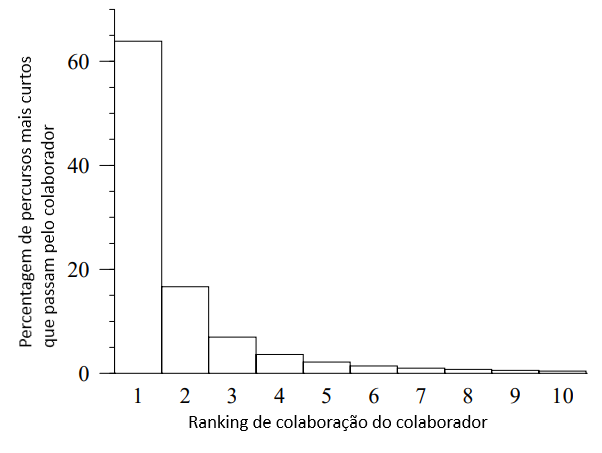
No artigo “Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality” Newman faz a análise de valores estatísticos não locais das redes obtidas, como o percurso mais curto, a distância média entre autores e a *“betweenness”,* que se define como o número de percursos mais curtos que passam por um determinado autor.

Quanto ao percurso mais curto, tal como o nome indica, é definido como o percurso que conecta dois vértices diferentes passando pelo menor número possível de arestas. Este percurso pode ser único ou podem existir diversos percursos que requerem o mesmo número de arestas para conectar dois vértices, dependendo de, para o caso específico, quantas ligações os cientistas em causa (vértices que se pretende conectar) estabeleceram com outros cientistas (grau do vértice).

Um exemplo de um caso onde existem diversos percursos mais curtos entre dois cientistas é apresentado seguidamente, onde se pretende estabelecer o percurso mais curto entre Newman e Barabási, dois cientistas que estudam redes complexas.

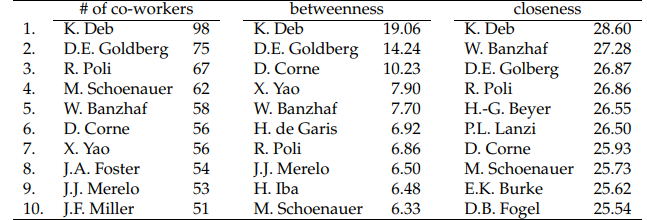
Como se pode observar, devido à notoriedade dos autores em causa, são diversos os percursos mais curtos que se podem estabelecer, uma vez que cada um dos cientistas tem ligações a diversos outros autores, incluindo alguns que não trabalham na área das redes. Este é o caso de Garrahan, cuja ligação a Newman se deve a um artigo publicado acerca de imanes de spin desorientado, ou *spin glasses*.

Quanto à “*betwenness”*, depois de analisada foi possível encontrar uma propriedade interessante nas redes de coautoria que se encontra ilustrada na figura seguinte.



Por análise da figura, verifica-se que ao serem analisados os percursos mais curtos entre um determinado cientista, chamar-lhe-emos A, e todos os outros, em média, cerca de 64% desses percursos irão passar pelo cientista que mais vezes colaborou com A, ou seja, ao se ter em conta apenas os percursos mais curtos entre A e todos os outros autores, o cientista que mais vezes colaborou com A terá uma percentagem de “*betweenness*” de cerca de 64%. Além disso, verifica-se que cerca de 98% dos percursos mais curtos para A passam por um dos top 10 colaboradores de A.

A existência de “super-estrelas sociométricas”, como foram chamadas por Milgram, em redes de coautoria foi novamente obtido no artigo “The Complex Network of Evolutionary Computation Authors: an Initial Study” de Carlos Cotta e Juan-Julián Merelo, onde é avaliada a rede de coautoria de artigos acerca de computação evolucionária.

Neste estudo foi obtida a tabela seguinte:

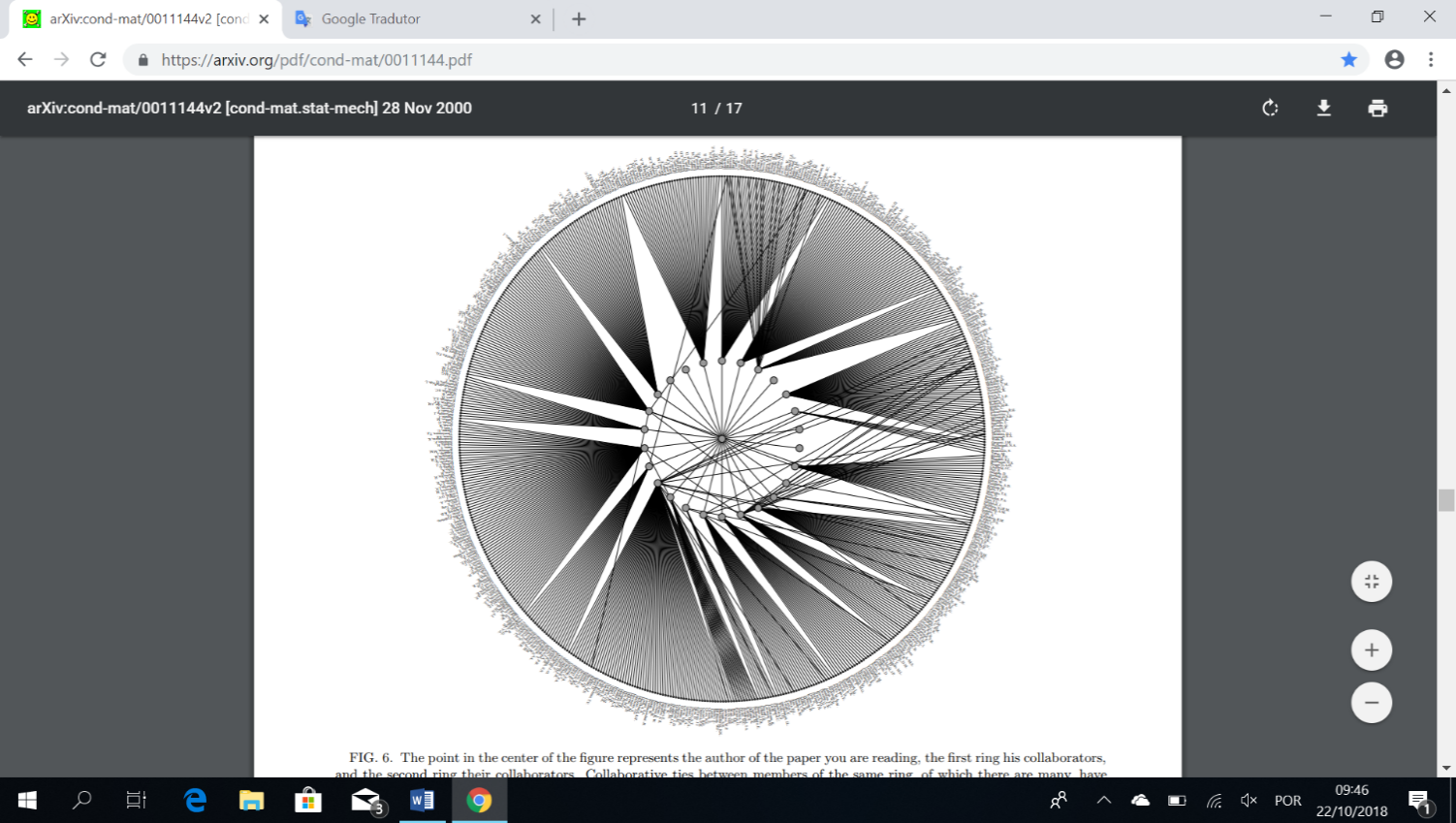
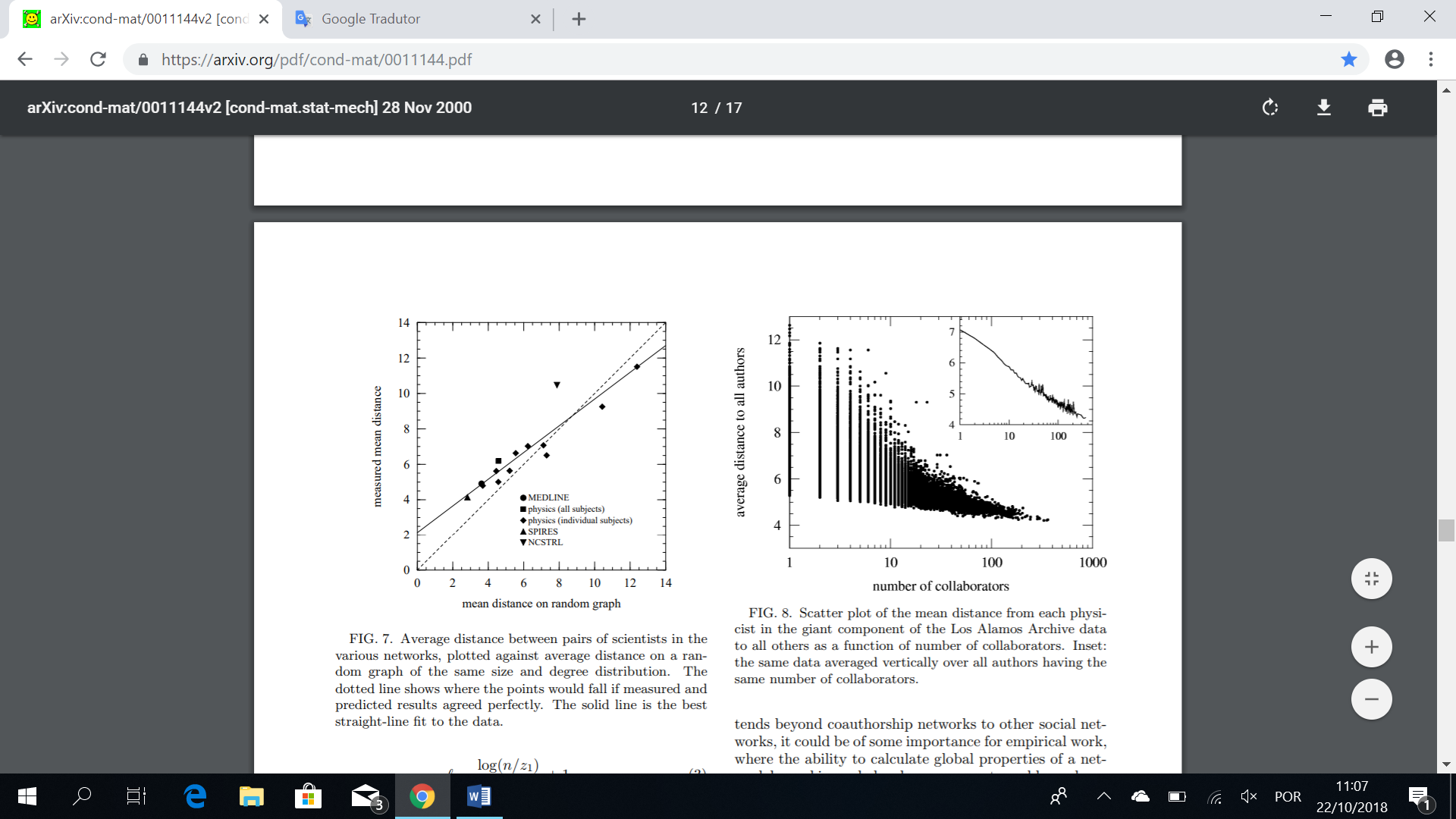
Nela é possível observar que os níveis de *“betweeness”* relativos de autores como D. E. Goldberg, autor de um dos livros mais proeminentes sobre o assunto e Kalyanmoy Deb, outro autor reconhecido de livros na área teórica de computação evolucionária são bastante superiores ao de outros autores, em especial o de K. Deb.

A explicação apontada para a existência destas entidades na rede passa desde logo pela sua excelência académica, que, ainda que difícil de medir, tem um bom indicativo no número de colaboradores do cientista. Como se pode observar na tabela, os dois autores que possuem maior valor de *“betweeness”* são também aqueles que possuem maior número de colaboradores. A quantidade de colaboradores oferece ainda informação quanto à eficiência que o cientista tem em transmitir informação para o resto da comunidade, porque um maior número de colaboradores permite uma difusão mais rápida de informação. Além disso, personalidades com elevada excelência académica serão também mais procuradas para a supervisão de teses e de projetos de investigação de diversas pessoas, aumentando ao mesmo o número de colaboradores e a *“betweeness”* das mesmas. A *“betweenness”* torna-se então, neste contexto, uma medida do controlo que um individuo tem sobre o fluxo de informação, já que um elevado valor desta indica que grande parte da partilha de informação passa por este individuo. Além disso, também é possível concluir que estes indivíduos possuem grande importância nas redes em que se encontra, uma vez que a sua retirada implicaria o aumento da distância de muitos percursos entre outros vértices.

Recuperando agora os valores de distância média entre cientistas apresentados na penúltima linha tabela 1, observa-se que para todas as redes se obtém valores relativamente pequenos, principalmente quando comparados com as dimensões das redes.

A variação destes valores está no intervalo de 4.0 para a rede obtida através da SPIRES e 9.7 para a NCSTRL, o que, como referido anteriormente, pode ser consequência de uma cobertura insuficiente de artigos publicados no âmbito da ciência computacional. No entanto, mesmo para o caso da rede da NCSTRL observa-se que a distância média entre cientistas é muito menor que a dimensão da rede, o que mostra que as redes científicas possuem características que permitem a rápida difusão de informação para todos os interessados. É ainda de notar que mesmo as distâncias máximas entre cientistas para as diferentes redes são todas relativamente pequenas, havendo um máximo de distância de 31 para o caso da NCSTRL, que pode, uma vez mais, ser explicada por uma cobertura insuficiente dos artigos publicados.

Estas propriedades são um indicativo de que ao lidar-se com redes de coautoria se está, de facto, a lidar com redes do tipo “small world”, onde em poucos passos é possível conectar quaisquer dois vértices da rede. Uma figura ilustrativa desta conclusão é apresentada seguidamente, onde se apresenta as ligações entre um cientista central, neste caso Newman, e os seus colaboradores diretos, e as ligações destes com os seus colaboradores diretos.

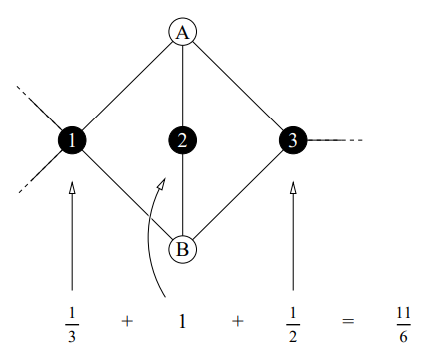
Pode-se ainda fazer a representação gráfica da distribuição de distâncias médias entre um determinado cientista e todos os outros em função do seu número de colaboradores, o que para o caso dos cientistas presentes na base de dados Los-Alamos será a distribuição representada seguidamente.

Como se pode observar, há uma tendência clara para a distância média entre um determinado cientista e todos os outros diminuir com o aumento do número de colaboradores deste. Esta tendência é corroborada pelos dados obtidos na tabela 1, nas 6ª e 8ª linhas. Ao analisar os dados presentes nessa tabela é também possível verificar que as áreas experimentais, como a astrofísica e a física das altas energias experimental (SPIRES) são aquelas que possuem maior número de colaboradores, como seria de esperar, já que há uma tendência para artigos teóricos serem publicados por um grupo restrito de pessoas e artigos experimentais serem publicados por grupos maiores de cientistas que contribuíram com diferentes ensaios para um mesmo artigo.

Contudo, isto levanta um problema, porque enquanto que é plausível considerar que a maioria dos grupos reduzidos de pessoas que escreveram um artigo juntos se conhecem razoavelmente bem, num artigo com muitos coautores é improvável que cada indivíduo tenha estabelecido uma relação com todos os outros. É, portanto, necessário obter uma medida de proximidade entre cientistas que tenha em conta a probabilidade de cientistas estabelecerem uma relação considerando o tamanho do grupo a que pertenciam.

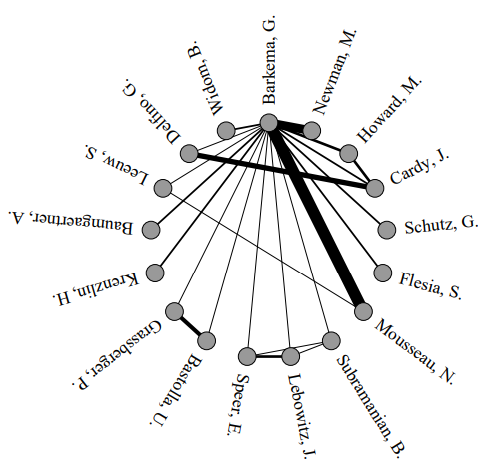
Tendo em conta este problema, desenvolve-se então um método para a criação de redes pesadas, que tenham em conta não só o número de colaboradores de cada autor, como a quantidade de vezes que estes colaboraram.

Para isso, atribui-se um peso à relação estabelecida entre dois colaboradores num projeto inverso ao número de colaboradores que participaram nesse projeto, ou, por outras palavras, num projeto com n coautores, a ligação entre 2 coautores será de . Este método é uma aproximação à realidade, já que na vida real cada pessoa não dedica exatamente a mesma proporção de tempo a cada colaborador e fatores como a personalidade e interesses comuns influenciam o grau de proximidade criado a partir de cada interação. No entanto, é a melhor aproximação que se pode fazer com a quantidade limitada de dados que é possível obter através de cada base de dados.

O peso total da relação entre dois cientistas será então a soma dos pesos de cada interação estabelecida em diferentes trabalhos, como se ilustra na figura seguinte, onde o autor A e o autor B colaboraram juntos em 3 projetos (1, 2 e 3), onde participaram 4, 2 e 3 cientistas, respetivamente.

Como se pode ver, para o caso apresentado, o peso da conexão entre os cientistas A e B será de , ou seja, , onde é o peso da ligação entre A e B. Ao seguir este método, é fácil de perceber que o grau de um vértice será então:

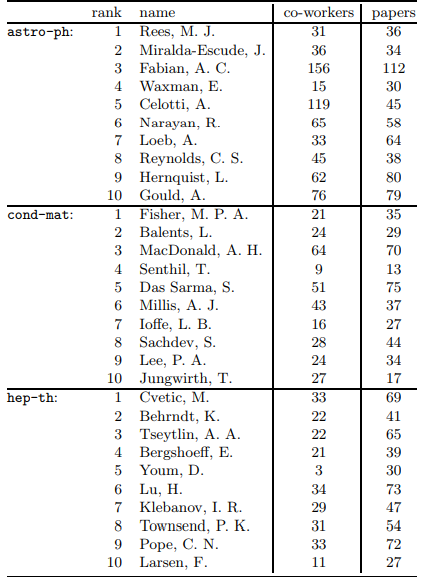
Onde, é o peso da ligação entre os vértices i e j, é o número de coautores do trabalho k e e são deltas de Kronecker, que assumem o valor de 1 se o cientista i (ou j) participaram no trabalho k e 0 se não participou.

Ao generalizar esta definição para a rede de colaboradores do cientista Gerard Barkema obtida a partir da Los Alamos, obtém-se a rede seguinte.

Nela é possível observar que, no caso de Barkema, os cientistas com os quais colabora mais serão Newman e Mousseau, como se pode observar pelo peso dessas arestas que é ilustrado através da largura das mesmas.

Utilizando agora o peso das arestas para avaliar a distância entre cientistas, pode-se definir a distância entre cientistas como inversamente proporcional ao peso da aresta que os une, o que para a rede mostrada anteriormente significa que a distância na rede entre Barkema e Mousseau seria muito menor que entre Barkema e Lebowitz, por exemplo.

Ao calcular agora a distância média de cada um dos cientistas a todos os outros através das distâncias calculadas da forma descrita em cima, para o caso dos cientistas presentes na base Los Alamos, foi possível obter um top 10 de cientistas mais bem conectados (com distâncias médias mais curtas) em cada uma das áreas presentes nesta base de dados, como apresentado na tabela seguinte.



É então interessante observar que, ao contrário da avaliação de conectividade anterior, onde os autores de artigos experimentais dominavam completamente o topo graças ao grande número de colaboradores que participava em cada trabalho, estão agora presentes no top 10 das diversas áreas autores que possuem muito poucos colaboradores, como é o caso de Senthil (9 colaboradores) e, em particular, Youm (apenas 3 colaboradores).

Pode-se então observar que o critério apresentado, ainda que não seja exatamente adequado à realidade, é uma boa aproximação, já que tem em conta não só o número de colaboradores que um cientista tem, como o número de vezes que estes colaboraram, levando a que um autor como Youm, com apenas 3 colaboradores com os quais trabalha regularmente, possa estar no top 5 de cientistas mais bem conectados na área da física das altas energias.

**5. Conclusões**

Em suma, e primeiramente encontraram-se dois métodos de criação de redes, que organizavam os autores de cada artigo pelos seus nomes, para que limitar, superior e inferiormente, o valor da realidade. De seguida, para fazer a análise dos resultados obtidos através destes dois métodos, foi feita uma representação gráfica, em escala logarítmica, destes resultados e observou-se que a distribuição do número de autores pelo seu número de colaboradores seguem duas leis da potência, e portanto decidiu-se abordar este assunta através do método do percurso mais curto.

Através deste método, usaram-se as relações de cada autor para determinar a relação de dois autores através das suas relações com outros autores, observando-se que os autores que possuem maior valor de *“betweeness”* são também aqueles que possuem maior número de colaboradores, e como a quantidade de colaboradores oferece ainda informação quanto à eficiência que o cientista tem em transmitir informação para o resto da comunidade científica, a *“betweenness”* passa a ser um medidor do controlo sobre o fluxo de informação.

Porém, e depois de obtidos vários resultados através deste método, nota-se que existe uma diferença significativa dos valores de *“betweenness”* entre artigos com números elevados de autores e os artigos em que o número de autores é mais reduzido, o que era de se esperar. Cria-se então um método para a criação de redes pesadas, que tem em conta a “força” entre a relação que cada autor tem com todos os outros coautores, ou seja, tem-se em conta o número de artigos que dois ou mais autores têm em comum, para criar uma rede entre todos os autores, de onde se pode observar que os resultados obtidos são os que mais se aproximam da realidade de entre os resultados obtidos nos outros métodos.

**6. Referências**