DOI: https://doi.org//10.21747/16466195/ling2022v2a10

Visualização da relevância relativa de investigadores a partir da sua produção textual

Luís Trigo
ltrigo@letras.up.pt
CODA-FLUP (Portugal)

Pavel Brazdil
pbrazdil@inesctec.pt
LIAAD (Portugal)

ABSTRACT.

Building a researchers affinity network through the automatic processing of their publications allows us to gain a perspective that goes beyond the networks established through co-authorship. The definition of the importance of each researcher is defined upon their bibliographic production volume, i.e., number of publications, and also upon their centrality in the general network of researchers. In fact, the centrality of a researcher in a network reveals its importance in communication flows with other researchers, thus assuming that communication between researchers is itself a relevant factor for organizational life and in its production.

Both network and centrality concepts are better interpreted in a graphical way. In this study, we explore the workflow that will provide these visualizations and focus in the empirical selection of the most appropriate centrality measure. We also propose a centrality visualization method that facilitates the interpretation of the selected measures.

KEY-WORDS.

Information Retrieval; Social Network Analysis; Centrality Analysis; Affinity Groups.

RESUMO.

A construção de uma rede de afinidade de investigadores através do processamento automático das suas publicações permite obter uma perspetiva que vai para além das redes estabelecidas através da coautoria. A definição da importância de cada investigador parte do seu volume de produção bibliográfica, i.e., número de publicações, e também da sua centralidade na rede geral de investigadores. De facto, a centralidade de um investigador numa rede revela a sua importância nos fluxos de comunicação com os outros investigadores, pressupondo deste modo que a comunicação entre investigadores é, em si própria, um fator relevante para a vida organizacional e na sua produção.

Tanto os conceitos de rede como de centralidade são melhor interpretados de forma gráfica. Neste estudo, exploramos o fluxo de trabalho que proporcionará estas visualizações e focamos na seleção empírica da medida de centralidade mais adequada. Propomos também um método

de visualização da centralidade que facilite a interpretação das medidas selecionadas.

PALAVRAS-CHAVE.
Recuperação de Informação; Análise de Redes Sociais; Análise de Centralidade; Grupos de Afinidade.

1. Introdução

Neste trabalho descrevemos um método para organizar a visualização de investigadores segundo a sua relevância através da sua produção textual. Consideramos cada autor como um documento de texto que compreende o conjunto das suas publicações. O documento pode ser assim caracterizado, não só pelos seus termos mais relevantes, mas também pela semelhança com documentos já conhecidos – esta parte é particularmente importante quando se quer processar uma procura em que não se tem a certeza dos termos importantes para essa procura.

O método pode ser generalizado para outros tipos de documentos (e.g., artigos científicos ou descrições de *curricula* de cursos (Vita *et al.* 2015) que é normalmente iniciada pela pesquisa de palavras-chave. Como a lista de resultados obtidos pode ser grande, fornece-se uma maneira de navegar no espaço das alternativas potenciais, enquanto se exploram afinidades (semelhanças) entre os documentos. Como se tem, de facto, um grafo/rede, os documentos podem ser processados recorrendo a métodos de análise de redes para identificar grupos de afinidade. Cada grupo de afinidade é de facto uma sub-rede da rede original. Assim, depois de o utilizador identificar um potencial documento de interesse, tem a informação sobre o grupo de afinidade correspondente e pode, além disso, inspecionar e seguir os *links* de afinidade existentes. Desta forma, pode-se identificar vários documentos de forma eficaz, revelando as afinidades que estão para lá da coautoria. Para promover uma exploração plena das suas potencialidades, o método requer boas ferramentas de visualização que mostrem as informações de forma rápida e intuitiva. Este esquema foi implementado (Brazdil *et al.* 2015, Trigo *et al.* 2015) e está disponível para uso como um protótipo (AffinityMiner 2015).

2. Metodologia

Neste artigo, fazemos uma descrição sistemática do método para encontrar a organização dos documentos, tendo como base a sua semelhança na produção textual. O nosso estudo é orientado para a área de publicações científicas de investigadores.

1. Processamento de Publicações dos Investigadores;
2. Elaboração de matriz de similaridade e visualização como um grafo;
3. Descoberta de grupos de afinidade;
4. Identificação dos nós importantes (investigadores) no grafo.

2.1. Processamento de Publicações dos Investigadores

No presente método, os títulos de publicações são extraídos para ficheiros de texto simples, cada um representando um determinado autor. Em vez de utilizar apenas os títulos, pode ainda considerar-se outros elementos dos textos como resumos e palavras-chave. Na validação de semelhanças realizamos uma experiência com esses elementos adicionais, concluindo que, no caso de estudo focado num centro de investigação da República Checa, as palavras-chave fornecidas pelos autores aumentavam a qualidade, mas os resumos, não (Trigo *et al*. 2015). Os ficheiros de texto são pré-processados da maneira usual no que concerne à análise de semelhança conceptual do conteúdo.

Tuzzi (2010) testou várias abordagens de *Bag of Words,* tendo obtido os melhores resultados na sua exploração de agrupamento de textos em *corpora* com a inclusão de todas as palavras no *corpus* de discursos de fim de ano de presidentes italianos e a inclusão de apenas palavras lexicais no *corpus* de notícias. Importa referir que, no âmbito de trabalhos em que importa ter em consideração elementos estilísticos (e.g., deteção de autoria), palavras funcionais (que consideraremos, no nosso âmbito, *stop words*) e pontuação serão também relevantes. Sari *et al*. (2018) acrescentam ainda outros atributos de interesse no caso da autoria, como o tamanho médio das palavras e frases.

No nosso estudo, utilizamos a representação *Bag of Words*, removendo-se previamente números, *stopwords* (palavras muito frequentes e palavras funcionais), sinais de pontuação e outros elementos espúrios.

2.2. Elaboração de Matriz de Semelhança

Para demonstrar o método, recorremos a um exemplo simplificado com uma amostra de 11 investigadores do Laboratório de Inteligência Artificial e Análise de Dados (http://www.liaad.up.pt/). A lista de documentos é transformada numa matriz documento-termo com uma frequência ponderada *tf-idf* (*term frequency – inverse document frequency*) (Feldman & Sanger 2007). Esta ponderação dá mais peso aos termos mais frequentes de um documento, penalizando aqueles que são comuns aos vários domínios, isto é, que funcionam como *stopwords*. A Tabela 1 recolhe um pequeno subconjunto dos termos que caracterizam esses investigadores.

TABELA 1 – Excerto da matriz documento-termo para 11 membros do LIAAD

	classification	learning	mining	regression	decision	market	cooperation	flow	genetic	scheduling	dynamics	immune	cancer
AJ	0,01	0,03	0,02	0	0	0	0	0	0	0	0	0	0
LT	0,02	0,04	0	0,09	0	0,02	0	0	0	0	0	0	0
PB	0,04	0,05	0,01	0	0	0	0,01	0	0	0	0	0	0
JG	0,01	0,05	0,03	0,01	0,02	0	0	0	0	0	0	0	0
DF	0	0	0	0	0,02	0	0	0,07	0,05	0,01	0	0	0
JFG	0	0	0	0	0	0	0	0,02	0,15	0,08	0	0	0
JV	0	0	0	0	0	0	0	0,01	0,02	0,16	0	0	0
PC	0	0	0	0	0	0	0,07	0	0	0	0,01	0,02	0,07
AP	0	0	0	0	0	0,02	0	0	0	0	0,02	0	0,04
LMF	0	0	0	0	0	0	0	0	0	0	0,03	0,02	0
BO	0	0	0	0	0	0	0	0	0	0	0,01	0,02	0

Os termos da matriz caracterizam cada um dos investigadores. É de frisar, no entanto, que se trata de uma amostra exemplificativa de mais de 1500 termos extraídos. A matriz documento-termo é, por natureza, esparsa, isto é, com a maior parte das células iguais a zero – em termos relativos, ainda mais do que a tabela de amostra. Optou-se nesta amostra por incluir apenas palavras com os valores *tf-idf* (*term frequency – inverse document frequency*) mais elevados, ou seja, os que têm maior poder discriminativo.

A matriz contém deste modo uma representação vetorial – cada documento é um vetor com o número de dimensões correspondente ao número de termos – que será usada para gerar a matriz de similaridade

cosseno. Desta métrica de similaridade importa destacar que não tem em consideração o tamanho dos documentos e varia os seus valores entre 0 e 1.

Esta matriz pode ser visualizada sob a forma de uma tabela simétrica, por isso podemos simplificá-la na forma de representação triangular da Tabela 2. O valor da semelhança da diagonal é logicamente 1, por se tratar do próprio autor.

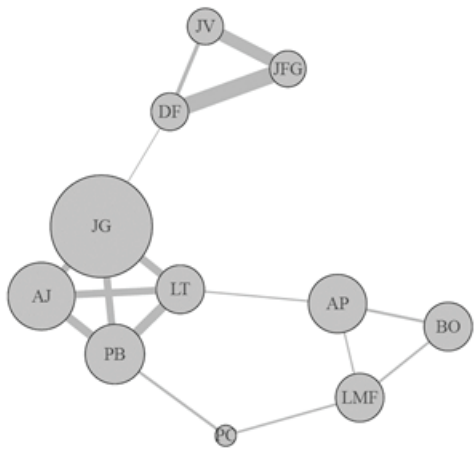
TABELA 2 – Matriz de semelhança cosseno para 11 membros do LIAAD

	AJ	LT	PB	JG	DF	JFG	JV	LMF	PC	AP	BO
AJ	1										
LT	0,17	1									
PB	0,2	0,22	1								
JG	0,17	0,18	0,17	1							
DF				0,03	1						
JFG					0,39	1					
JV					0,09	0,29	1				
LMF								1			
PC			0,06					0,05	1		
AP		0,04						0,04	0	1	
BO								0,05	0	0,06	1

A tabela de semelhança (Tabela 2) serve para definir a matriz de adjacência que será usada para construir o grafo de visualização da rede de afinidades, patente na Figura 4. Cada investigador é representado por um círculo. O tamanho do nó é proporcional ao número de publicações de cada investigador na base de dados *Authenticus* (2014). A espessura das arestas representa o valor de similaridade entre os pares de investigadores. Quanto mais espessa a ligação, mais similar é o par de investigadores unidos por esta ligação. Para simplificar todas as ligações, as semelhanças abaixo de um determinado limiar foram consideradas irrelevantes e removidas. O valor do limiar de 0,03 de semelhança foi escolhido na base de uma análise experimental, tendo sido o valor que nos apresentava os resultados mais claros na visualização. O valor utilizado na aplicação que serviu de protótipo foi 0,05, como descrito por Martinez *et al*. (2017). No mesmo trabalho, estes autores utilizaram a otimização do valor da modularidade na geração de comunidades para chegar a um valor mais objetivo para este limiar. Deve considerar-se, contudo, que este valor pode variar em função

do conjunto de dados. Por outro lado, o limiar inferior de visualização de semelhanças poderá ser usado como um dos parâmetros de ajustamento da visualização da rede.

FIGURA 1 – Rede de afinidade de 11 investigadores do LIAAD



Os agrupamentos podem ser de algum modo aferidos através deste grafo. Ainda assim, o agrupamento central não tem a sua separação muito evidenciada relativamente ao agrupamento no lado inferior direito. Deste modo, recorrer-se-á a um método automático para definir os grupos de afinidade.

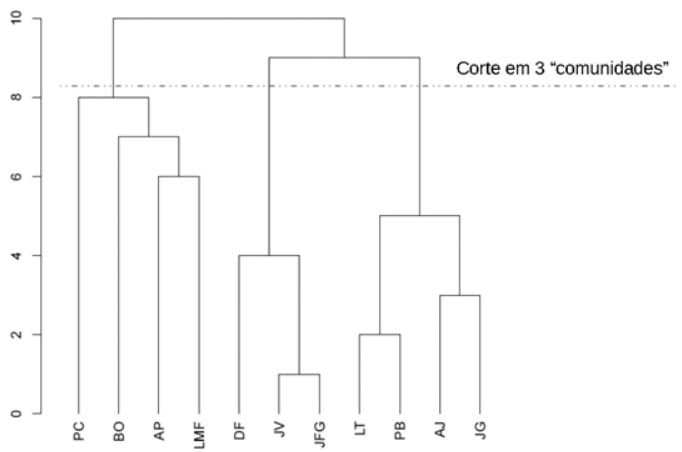
2.3. Deteção de Grupos de Afinidade

Depois de transformar a matriz de similaridade em formato gráfico, utilizamos o algoritmo de descoberta de comunidades *Walktrap* (Pons & Latapy 2005). Esta técnica encontra subgrafos densamente ligados, também definidos como comunidades, através de passeios aleatórios, presumindo que caminhos aleatórios curtos tendem a significar a pertença à mesma comunidade. Segundo Pons e Latapy (2005), em várias áreas, nomeadamente

nas redes de colaboração social, encontram-se grafos globalmente esparsos, mas localmente densos. Estes subgrafos recebem, então, o nome de *comunidades*. No início, todos os nós são tratados como comunidades hipotéticas, mas, à medida que o processo de análise avança, os nós vão-se juntando, até formarem comunidades estacionárias. Embora os seus autores tenham apenas considerado as relações de adjacência, admitem que se pode utilizar a ponderação do peso dessas relações.

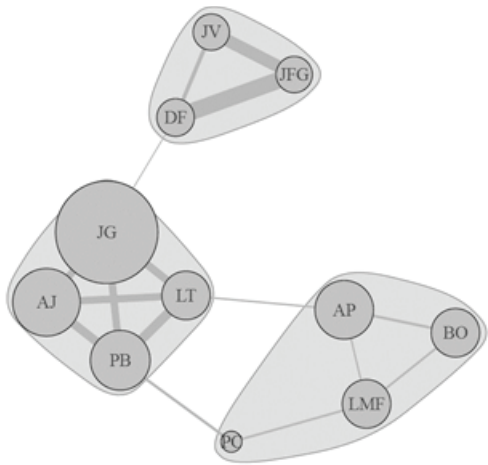
As diferentes comunidades descobertas (grupos de afinidade) podem ser identificadas no grafo de maneiras diferentes. No nosso caso, usamos a cor para destacá-las no grafo. Deste modo, e à semelhança do agrupamento hierárquico tradicional, também este método permite a representação hierárquica com dendrograma. No dendrograma gerado para os nossos 11 investigadores de exemplo, podem visualizar-se as três partições principais que resultam dos passeios aleatórios realizados pelo algoritmo *Walktrap*.

FIGURA 2 – Dendrograma extraído com algoritmo *Walktrap* para os 11 investigadores



A representação dos grupos de afinidade (comunidades emergentes) resulta no grafo seguinte.

FIGURA 3 – Representação dos grupos de afinidade extraídos com o algoritmo *Walktrap*



Recupera-se agora a tabela documento-termo apresentada no início deste capítulo (Tabela 1), para mostrar que a definição dos grupos de afinidade já estava latente. Os grupos de afinidade gerados estão identificados por diferentes tons de cinzento na tabela que se segue. Também se podem identificar os termos que caracterizam cada grupo de afinidade pelo critério *tf-idf,* assim como os termos que servem de ligação entre os grupos.

TABELA 3 – Excerto da matriz documento-termo com destaque de grupos de afinidade

	classification	learning	mining	regression	decision	market	cooperation	flow	genetic	scheduling	dynamics	immune	cancer
AJ	**0,01**	**0,03**	**0,02**	**0**	0	0	0	0	0	0	0	0	0
LT	**0,02**	**0,04**	**0**	**0,09**	0	0,02	0	0	0	0	0	0	0
PB	**0,04**	**0,05**	**0,01**	**0**	0	0	0,01	0	0	0	0	0	0
JG	**0,01**	**0,05**	**0,03**	**0,01**	0,02	0	0	0	0	0	0	0	0
DF	0	0	0	0	0,02	0	0	**0,07**	**0,05**	**0,01**	0	0	0
JFG	0	0	0	0	0	0	0	**0,02**	**0,15**	**0,08**	0	0	0
JV	0	0	0	0	0	0	0	**0,01**	**0,02**	**0,16**	0	0	0
PC	0	0	0	0	0	0	0,07	0	0	0	**0,01**	**0,02**	**0,07**
AP	0	0	0	0	0	0,02	0	0	0	0	**0,02**	**0**	**0,04**
LMF	0	0	0	0	0	0	0	0	0	0	**0,03**	**0,02**	**0**
BO	0	0	0	0	0	0	0	0	0	0	**0,01**	**0,02**	**0**

2.4. Identificação de Elementos Importantes através de Centralidades

Nesta secção, procede-se à descrição de algumas das medidas de centralidade mais relevantes para identificar elementos importantes no seio de uma rede. Estas medidas serão aplicadas e empiricamente estudadas nas secções 2.5 e 2.6.

a) Centralidade de Grau

A Centralidade de Grau define-se pelo número total de nós que se ligam a determinado nó. O valor mais elevado (Figura 4.a) está localizado no grupo de afinidade identificado como o grupo na parte inferior direita, associado à *Modelação Matemática* (e que tem, em média, os valores mais altos) e no grupo associado à *Investigação Operacional* (parte superior). A força das afinidades que os elementos mais preponderantes têm com os seus vizinhos no grafo é muito importante neste resultado, como evidencia o valor do nó JFG.

Esta medida de centralidade pode ser criticada, de acordo com Feldman e Sanger (2007), porque apenas considera as ligações diretas de uma entidade (um ator) em vez de considerar as relações indiretas com todas as outras entidades (os outros atores). Uma entidade pode estar diretamente ligada a um grande número de outras entidades que podem estar isoladas da rede. Tal entidade apenas é central na sua vizinhança de rede e, por isso, desenvolveram-se outras métricas de que falaremos de seguida.

b) Centralidade de Proximidade

Esta medida baseia-se no cálculo das distâncias geodésicas (menor distância entre dois pontos) entre a entidade e todas as outras entidades na rede (Wasserman & Faust 1994). O princípio subjacente a esta medida é que a centralidade de um ator em termos comunicacionais depende da rapidez do seu acesso a todos os outros atores, isto é, não precisa de muitos intermediários, pois dispõe de caminhos mais curtos para comunicar a sua informação a qualquer outro envolvido na resolução de um dado problema. Uma das suas limitações é não poder ser calculada quando existem subgrafos desligados.

Os valores mais elevados para a Centralidade de Proximidade encontram-se

no grupo de *Machine Learning* (parte inferior esquerda), uma vez que os seus elementos se encontram mais perto de todos os outros. Considerando cada um dos grupos, verificamos também que são os membros que servem de ponte com os grupos externos que possuem os valores mais elevados – DF, AP e PC nos grupos de *Investigação Operacional* e *Modelação Matemática*.

c) Centralidade de Intermediação

A Centralidade de Intermediação mede, segundo Feldman e Sanger (2007), a eficácia com que um vértice se liga às várias partes da rede. Entidades que se localizam em muitos caminhos geodésicos entre outros pares de entidades são mais poderosos, uma vez que controlam o fluxo de informação entre os pares.

Como seria de esperar, o maior valor da Centralidade de Intermediação dá-se nos elementos que servem de ligação entre os diversos grupos de afinidade, com mais preponderância para os que se encontram no grupo de *Machine Learning* (parte inferior esquerda) e que perfazem o caminho mais curto entre todos os grupos.

d) Centralidade de Valores Próprios

Esta medida de centralidade parte do princípio de que a importância de um nó decorre da importância das ligações dos seus vizinhos. Os valores desta centralidade correspondem ao vetor principal de valores próprios da matriz de adjacência do grafo.

Na Centralidade de Valores Próprios ganham preponderância os elementos do grupo de *Machine Learning* (parte inferior esquerda), aquele que tem afinidade maior e com ligações mais fortes com elementos já de si importantes.

e) Centralidade de PageRank

Segundo Mihalcea (2004), o PageRank, concebido como método para analisar a proeminência das páginas web através da análise das suas hiperligações, integra as ligações externas que referenciam cada página e as

ligações internas de cada página que referenciam outras páginas, de modo a produzir um conjunto de pontuações.

Como seria de esperar, sendo a Centralidade de PageRank uma variação da centralidade de valores próprios, em média, os valores mais elevados encontram-se no grupo de *Machine Learning* (parte inferior esquerda). Tendo como pressuposto passeios aleatórios, repara-se que a força das ligações existente no grupo de *Investigação Operacional* (parte superior), torna os valores entre estes dois grupos relativamente menos discrepante do que acontecia na Centralidade de Valores Próprios.

f) Centralidade Laplaciana

A Centralidade Laplaciana é outra medida que considera mais informação ambiental de um nó, captando não só as ligações diretas, como a importância dos vizinhos dessas ligações. Deste modo, a importância de um nó prende-se com a capacidade da rede em responder à desativação desse nó na rede.

A força das ligações vai definir os níveis de energia de *Laplace* e, considerando a reduzida dimensão do grafo, vai deste modo resultar que a ordenação da Centralidade Laplaciana seja semelhante à da Centralidade de Grau.

2.5. O Cálculo dos Valores de Centralidade

Nesta secção fazemos uma análise empírica para melhor compreensão dos conceitos associados às centralidades, recorrendo para este efeito aos 11 investigadores do LIAAD que foram introduzidos anteriormente.

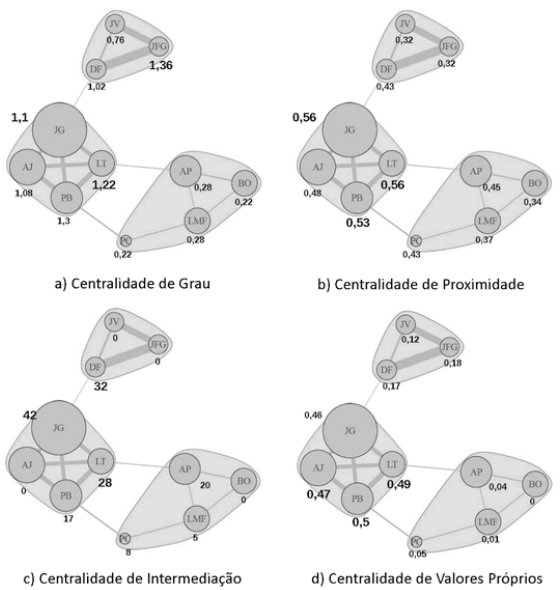
A tabela seguinte mostra um exemplo das centralidades tendo como base o grafo desses investigadores. A cinzento, encontram-se os três valores mais altos para cada uma das centralidades consideradas. A última linha da tabela corresponde ao número de publicações, uma medida da importância do investigador, visível nos grafos anteriores através do tamanho dos nós dos grafos.

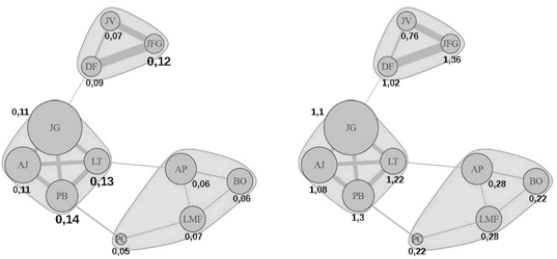
TABELA 4 – Indicadores de centralidade e número de publicações para 11

	AP	AJ	BO	DF	JG	JV	JFG	LMF	LT	PB	PC
C. Grau	0,28	1,08	0,22	1,02	1,1	0,76	1,36	0,28	1,22	1,3	0,22
C. Proximidade	0,45	0,48	0,34	0,43	0,56	0,32	0,32	0,37	0,56	0,53	0,43
C. Intermediação	20	0	0	32	42	0	0	5	28	17	8
C. Valores Próprios	0,04	0,47	0	0,17	0,46	0,12	0,18	0,01	0,49	0,5	0,05
C. Page Rank	0,06	0,11	0,06	0,09	0,11	0,07	0,12	0,07	0,13	0,14	0,05
C. Laplace	0,03	0,49	0,02	0,58	0,48	0,33	0,93	0,03	0,59	0,66	0,02
# Publicações	73	97	50	30	223	28	28	50	50	76	9

Para melhor interpretação e localização destes valores, recorremos ao grafo gerado na etapa de deteção de grupos de afinidade. Os grafos da Figura 4 apresentam os valores das centralidades na parte inferior dos nós, conforme os valores na Tabela 4, permitindo comparar com o número de publicações representado pelo tamanho dos nós. O grupo na parte superior corresponde ao grupo de *Investigação Operacional*, o grupo na parte inferior esquerda corresponde ao de *Machine Learning* e o grupo na parte inferior direita a *Modelação Matemática*. Os três valores mais relevantes em cada figura foram destacados na Tabela 4, para uma melhor identificação.

FIGURA 4 – Grafos com valores para diversas medidas de centralidade





e) Centralidade de PageRank f) Centralidade Laplaciana

2.6. Seleção de Valores de Centralidade

Nesta subsecção, procede-se a uma análise mais aprofundada dos indicadores de importância (centralidades e número de publicações) mencionados anteriormente. Para tal, abordamos a questão de correlação entre os diferentes indicadores. A análise recai tanto sobre o grafo reduzido com 11 nós, como sobre o grafo com os 104 elementos de cinco unidades do INESC TEC, uma vez que a própria dimensão do grafo pode ter influência sobre as correlações entre os indicadores de importância.

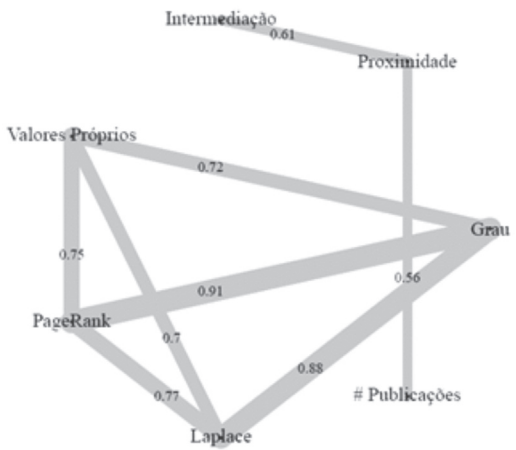
Como a dimensão de uma das amostras é reduzida (11 elementos), recorreu-se ao coeficiente não-paramétrico *tau de Kendall*. Esta medida tem em consideração a ordenação dos elementos da amostra para inferir a correlação entre duas variáveis, considerando o número de inversões e as suas posições (Lesot & Rifqi 2010). Deste modo, duas medidas com poucas inversões podem ser consideradas mais próximas e, de forma inversa, duas medidas podem ser consideradas menos próximas se as inversões ocorrem com mais frequência. Esta medida pode tomar valores entre -1 e 1, atestando a correlação negativa ou positiva (Newson 2002).

A Tabela 5 mostra a matriz com os valores de correlação superiores a 0,5 (considerados de algum modo significativos) destacados a cinzento. Esta tabela pode ser representada pelo grafo circular da Figura 5, onde se explicitam melhor as relações entre as várias medidas com valores de correlação acima de 0,5. Para conseguir um maior contraste entre os menores e os maiores valores na visualização, representa-se a espessura das arestas da Figura 5 elevando o seu valor ao quadrado.

TABELA 5 – Correlação *tau de Kendall* para valores de diferentes medidas de centralidade e número de publicações (11 investigadores)

C. Grau	1						
C. Proximidade	0,3	1					
C. Intermediação	0,1	0,61	1				
C. Valores Próprios	0,72	0,45	0,21	1			
C. Page Rank	0,91	0,34	0,13	0,75	1		
C. Laplace	0,88	0,21	0,08	0,7	0,77	1	
# Publicações	0,21	0,56	0,2	0,25	0,28	0,11	1
	C. G.	C. P.	C. I.	C. V. P	C. P. R	C. L.	# Pub.

FIGURA 5 – Grafo de correlação *tau de Kendall* para valores de diferentes medidas de centralidade e número de publicações (11 investigadores)

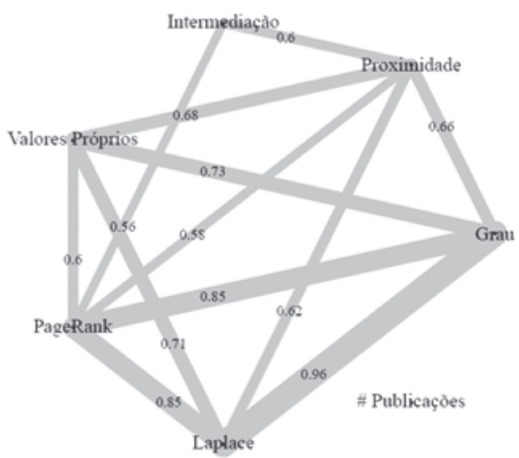


A expectativa à partida era que deveria haver uma correlação significativa entre a Centralidade de Valores Próprios e a de PageRank, assim como entre a Centralidade de Proximidade e a de Intermediação, considerando que os nós com caminhos mais próximos em relação aos outros nós também se devem encontrar nos caminhos mais curtos. Tratando-se de um grafo pequeno, ganha alguma preponderância a correlação estabelecida entre a Centralidade de Grau, que considera apenas a vizinhança imediata, e as Centralidades de Valores Próprios, de PageRank e de Laplace, que capturam critérios de ponderação de vizinhança mais complexos.

A avaliação experimental da correlação entre medidas de importância dos nós na Tabela 5 confirma a expectativa teórica. A outra medida de importância que utilizamos para medir a correlação foi o número de publicações. Todos os valores de correlação do número de publicações com as medidas de centralidade apresentam valores não significativos, à exceção da Centralidade de Proximidade, a qual apresenta, ainda assim, um valor mediano. A interpretação que podemos tirar daqui pode ser circunstancial a este grafo. De qualquer forma, faz sentido dizer que quanto mais publicações uma pessoa tiver, maior probabilidade terá de se encontrar próxima de qualquer outra pessoa na rede.

Procede-se agora à análise de correlação para 104 elementos do INESC TEC (Figura 6), uma vez que a dimensão do grafo pode ter influência sobre os resultados da correlação. Sendo que a Centralidade de Proximidade implica a existência de um único componente, eliminou-se desta análise dois nós isolados que estavam desligados do componente principal.

FIGURA 6 – Grafo de correlação *tau de Kendall* para valores de diferentes medidas de centralidade e número de publicações (102 investigadores)

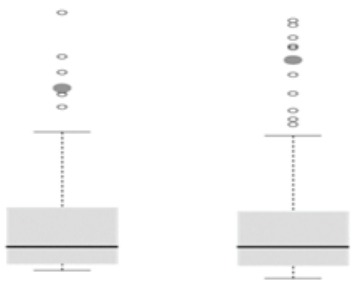


A análise da Figura 6 evidencia que a correlação do número de publicações com qualquer uma das outras medidas de centralidade não tem significância, isto é, a sua correlação com a Centralidade de Proximidade verificada anteriormente deixa de ser significativa, o que confirma de algum modo que essa correlação foi acidental. É ainda de sublinhar que a correlação entre as medidas de centralidade ligadas à vizinhança (centralidades de Grau, Valores Próprios, Page Rank e Laplaciana) preservam um alto grau de correlação, como estava patente nos valores de correlação registados anteriormente. Tendo em consideração os resultados dos valores de correlação entre as medidas de importância analisadas, podemos selecionar algumas destas para integrar no nosso método. A Centralidade Laplace apresenta-se muito correlacionada com a Centralidade de Grau, apesar de ser computacionalmente mais complexa. O mesmo se poderá dizer da Centralidade de PageRank apesar de estar um pouco menos correlacionada. Contudo, consideramos também que devemos descartar a Centralidade de Grau por ser muito simples na forma como considera a sua vizinhança imediata – sem ponderação da importância de cada um dos vizinhos. Apesar da correlação entre a Centralidade de Intermediação e a Centralidade de Proximidade não ser muito elevada, a última medida padece do grave problema de não permitir componentes desligados. Deste modo, as medidas de importância que consideramos mais relevantes para o nosso método são as Centralidades de Intermediação e a de Valores Próprios, assim como o número de publicações. Serão estas medidas de centralidade que merecem especial tratamento na próxima subsecção. Uma vez que os valores das centralidades têm pouco significado por si próprios, adotou-se uma medida normalizada, usando os seus valores em percentis. A representação gráfica em *diagrama de bigodes* dá uma perceção imediata do posicionamento ou importância de um autor relativamente aos restantes.

A Figura 7 mostra dois diagramas de bigodes representando a distribuição da centralidade para os 102 investigadores do INESC TEC. A posição relativa de um autor (selecionado aleatoriamente) encontra-se no ponto com preenchimento. A caixa, parte sombreada, também denominada intervalo interquartil, representa a distribuição entre o segundo e o terceiro quartil, dividido assimetricamente pelo segmento da mediana. Os restantes pontos para além dos bigodes, isto é, mínimo e máximo, são considerados

valores extremos ou *outliers* (Massart & Smeyers-verbeke 2005). No caso em apreço, o investigador selecionado encontra-se perto do valor máximo de centralidade.

FIGURA 7 – Diagramas de bigodes representando a Centralidade de Intermediação e a Centralidade de Valor Próprio do investigador JG em termos relativos (102



3. Discussão e conclusão

A visualização gráfica é um poderoso auxiliar na exploração de *corpora*. No presente estudo, o nosso *corpus* é constituído pela produção textual de investigadores. A partir deste *corpus*, desenvolvemos um método para produzir objetos visuais que permitem uma análise rápida da importância relativa dos investigadores no âmbito de uma organização ou domínio. A identificação de grupos de afinidade permite um melhor enquadramento das várias medidas de centralidade, tanto ao nível da identificação de elementos centrais a cada grupo como dos elementos que servem de ligação entre grupos. O grupo de afinidade parte do conceito de comunidade, que se define pelo agrupamento gerado com base nas ligações – este agrupamento subdivide a rede em partes mais compreensíveis, tendo como pressupostos uma certa homogeneidade intragrupo e heterogeneidade relativamente aos outros grupos.

Numa perspetiva prática, a visualização das afinidades e centralidades no âmbito de um grafo/rede permite estudar aspetos organizacionais para além da estrutura formal das instituições, perspetivando e gerindo colaborações

potenciais através da visualização das afinidades entre investigadores.

Para além da perspetiva relacional, procuramos uma forma visual que nos permitisse ter uma visão relativa de um individuo de uma forma mais quantitativa e direta, i.e., sem estar dependente da estrutura relacional do grafo que nos dá uma perceção mais qualitativa. Para tal, recorremos a diagramas de bigodes para apresentar as medidas de centralidade que selecionamos. A escolha das medidas produziu-se através de um estudo empírico nos nossos dados que nos revelou aquelas que poderiam maximizar a nossa informação sem redundâncias.

REFERÊNCIAS

Brazdil, P., Trigo, L., Cordeiro, J., Sarmento, R., & Valizadeh, M. (2015). Affinity mining of documents sets via network analysis, keywords and summaries. *Oslo Studies in Language*, *7*(1), 183-207.

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

AffinityMiner. (2015). *Affinity Miner Online Prototype* [Software]. http://affinityminer.com/

Lesot, M. & Rifqi, M. (2010). Order-Based Equivalence Degrees for Similarity and Distance Measures. In E. Hüllermeier, R. Kruse, F. Hoffmann (Eds.), *Computational Intelligence for Knowledge-Based Systems Design*. IPMU 2010. Lecture Notes in Computer Science (Vol. 6178, pp. 19-28). Springer.

Martinez, A., Brazdil, P, Sarmento, R., Trigo, L., Silva, F., & Bugla, S. (2017). Analysis of Publications of Academic Staff of FEP and Their Affinities, *FEP working papers*, 599, Faculdade de Economia da Universidade do Porto.

Massart, D. L., & Smeyers-verbeke, A. J. (2005). Practical Data Handling - Visual Presentation of Data by Means of Box Plots. *LC- GC Europe*, *18*(4), 215-218.

Mihalcea, R. (2004). Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions* (pp. 170-173). Association for Computational Linguistics.

Newson, R. (2002). Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *The Stata Journal*, *2*(1), 45-64.

Pons, P., & Latapy, M. (2005). Computing Communities in Large Networks Using Random Walks. In P. Yolum, T. Güngör, F. Gürgen, & Can Özturan (Eds.) *Proceedings of the 20th International Conference on Computer and Information Sciences* (pp. 284-293). Springer-Verlag.

Sari, Y., Stevenson, M., & Vlachos, A. (2018). Topic or style? Exploring the most useful features for authorship attribution. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 343-353). Association for Computational Linguistics.

Trigo, L., Vita, M., Sarmento, R., & Brazdil, P. (2015). Retrieval, visualization and validation of affinities between documents. In A. Fred, J. L. G. Dietz, D. Aveiro, K. Liu, J. Filipe (Eds.). *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015)* (Vol.

3, pp. 452-459). Springer.

Tuzzi, A. (2010). What to put in the bag? Comparing and contrasting procedures for text clustering. *Italian Journal of Applied Statistics/Statistica Applicata*, *22*(1), 77-94.

Víta, M., Komenda, M., & Pokorná, A. (2015, September). Exploring medical curricula using social network analysis methods. *In 2015 Federated Conference on Computer Science and Information Systems (FedCSIS)* (pp. 297-302). IEEE.

Wasserman, S., & Faust, K. (1994). *Social network analysis: methods and applications*. Cambridge University Press.