Universidade Federal de Juiz de Fora
Instituto de Ciencias Exatas
Bacharelado em Ciência da Computação

Redes Complexas para Análise de Influência entre Pesquisadores

Laércio Pioli Junior

JUIZ DE FORA
NOVEMBRO, 2017

Redes Complexas para Análise de Influência entre Pesquisadores

Laércio Pioli Junior

Universidade Federal de Juiz de Fora

Instituto de Ciencias Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Victor Ströele de Andrade Menezes

JUIZ DE FORA

NOVEMBRO, 2017

Redes Complexas para Análise de Influência entre Pesquisadores

Laércio Pioli Junior

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIENCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Victor Ströele de Andrade Menezes

Alex Borges Vieira

Paulo Alceu dAlmeida Rezende

JUIZ DE FORA

24 DE NOVEMBRO, 2017

Resumo

Nesse trabalho, foi analisado o comportamento de interação entre autores e coautores de uma rede social científica. A rede foi modelada através dos dados disponibilizados pela DBLP. Para auxiliar a pesquisa, foi utilizado técnicas de redes complexas que permitiram analisar a topologia da rede. Dentre as análises realizadas destacam-se a distribuição de grau, influência e medidas de centralidade em redes complexas. A análise topológica possibilitou o entendimento do comportamento gerado pela função de influência que foi utilizado no modelo da rede. As medidas de centralidade permitiram identificar quais eram os pesquisadores mais centrais da rede.

Palavras-chave: Redes Complexas, Análise de Redes Sociais, DBLP, Big Data, Análise de Influência.

Abstract

In this work, it was analyzed the iteration behavior between authors and coauthors of a scientific social network. The network was modeled using the data provided by DBLP. To support the analysis, we used complex network techniques that allowed us to analyze the topology of the network. Among the analyzes carried out, we highlight the distribution of degree, influence and measures of centrality in complex networks. The topological analysis allowed the understanding of the behavior generated by the influence function that was used in the network model. The measures of centrality allowed to identify which were the most central researchers of the network. ”

Keywords: Complex Networks, Social Network Analysis, DBLP, Big Data, Relationship, Analysis Influence.

Agradecimentos

Agradeço a toda minha família pelo apoio e encorajamento. Em especial minha mãe, (Maria Eliane Ribeiro Campos Pioli) pela motivação e orações realizadas e meu pai, (Laércio Pioli) pelo isentivo e apoio nessa etapa tão especial de minha vida.

Agradeço aos meus amigos pelo apoio.

Agradeço minha namorada Julia Erse Witt pelo apoio, paciência e incentivo em todas minhas decisões.

Ao professor Victor Ströele pela orientação, amizade e ensinamento e a todos os professores do Departamento de Ciência da Computação.

Agradeço a UFJF pela infraestrutura concedida e oportunidade de adquirir meus conhecimentos.

"A book is proof that humans are capable of working magic".

Carl Sagan

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

DCC	Departamento de Ciência da Computação
UFJF	Universidade Federal de Juiz de Fora
DBLP	Digital Bibliography & Library Project
OMS	Organização Mundial da Saúde
MIT	Massachusetts Institute of Technology
G(V,E)	Grafo G com V vértices e E Arestas
SFN	Scale-Free-Networks
RN	Random Network
PL	Power Law
SWN	Small-World-Network
CNT	Complex Network Theory
XML	eXtensible Markup Language
SAX	Simple API for XML
NoSQL	Not Only SQL
ETL	Extract Transform Load

1 Introdução

Diversos problemas do mundo real podem ser representados através de redes, onde indivíduos se conectam uns com os outros, representando um determinado domínio. Atualmente, são conhecidos vários tipos de redes, como por exemplo, as redes sociais, redes de cadeias de DNA, redes elétricas, redes aéreas, redes de neurônios entre outras. Diante disso, surge a necessidade de entender como essas redes funcionam. Entendendo o seu funcionamento é possível identificar característica que represente alguma informação.

Levando em consideração o domínio biológico podemos exemplificar uma rede de conexão entre neurônios. Sabe-se que um neurônio é uma célula nervosa que é responsável pela condução de impulsos elétricos em nosso cérebro, mas pouco se sabe como eles interagem entre si. Se analisarmos cada neurônio individualmente, dificilmente entenderemos seu papel como um todo. Analisando o todo, é possível descobrir que o fluxo de eletricidade que atravessa de uma região do nosso cérebro para outra pode ser caracterizado como a memória que possuímos.

Compreender o comportamento de elementos pertencentes a diferentes tipos de redes certamente é muito importante, pois permite responder questões ligadas ao domínio estudado. Foi estudado nesse trabalho uma rede social científica chamada DBLP (Digital Bibliography & Library Project). Algumas análises topológicas foram feitas como: distribuição de grau, influência e análises de centralidade. Dentre as análises de influência realizadas, foi possível identificar pequenos grupos de pesquisadores que estão ligado na maior componente conexa.

Sabe-se que grupos de indivíduos podem representar comunidades. As comunidades científicas são compostas por pessoas que possuem algum interesse em comum. Essas pessoas influenciam e são influenciadas pelos demais indivíduos da rede. A modelagem da rede científica adotada neste trabalho é baseada em um grafo bi-direcional que analisa as iterações de influência entre pesquisadores.

A utilização de redes complexas e teoria das redes sociais tem apresentado benefícios no entendimento de diversas áreas de estudo, sendo ela de caráter científico ou

social. A abordagem proposta nesse trabalho pode contribuir em muitos aspectos existentes que se referem ao modo de como os pesquisadores trabalham e se influenciam.

Foi analisado o comportamento dos pesquisadores que compõem uma rede de iteração entre autores e coautores de uma rede social científica. Nesse trabalho, foi utilizada uma base de dados real, que contém milhares de registros que representam publicações de documentos científicos da área de computação. Através da modelagem proposta, foi construída uma rede de co-autoria que representa as iterações entre os pesquisadores. Para isso, foi utilizada uma métrica que calcula o quão influente um pesquisador é, se comparado com os demais pesquisadores dessa rede.

Para ajudar na compreensão das características presentes dessa modelagem, foram utilizadas técnicas de redes complexas que auxiliaram nas análises feitas sobre a topologia da rede.

Este trabalho tem como objetivo analisar a estrutura da rede social científica DBLP. Dentre as análises realizadas, destacam-se as distribuições de grau, influência e medidas de centralidade em redes complexas.

Este trabalho está organizado da seguinte forma: na seção 2 será apresentado toda a fundamentação teórica necessária, que auxilia no entendimento das atividades realizadas nos capítulos posteriores. É apresentado também os principais tipos de redes existentes, e também algumas medidas de centralidade em redes, a seção 3 apresenta uma introdução sobre a base de dados DBLP, nessa introdução pode-se destacar alguns trabalhos relacionados e também as instruções de um pré-processamento que foi realizado nos dados para posteriormente, na seção 4, ser apresentado um modelo da rede social científica. Nesse modelo será utilizado métricas de análise de redes complexas que permitirão analisar sua topologia, a seção 5 tem como objetivo fazer algumas análises de centralidade na rede. As análises feitas permitiram entender alguns pontos interessantes da rede e, finalmente, na seção 6 são apresentadas as considerações finais deste trabalho.

2 Fundamentação Teórica

Neste capítulo são apresentados os conceitos necessários para o entendimento deste trabalho. Com isso, são abordados os principais conceitos de análise de redes sociais juntamente com conceitos de teoria de redes complexas.

2.1 Redes Complexas e Teoria dos Grafos

Segundo M.E.J.Newman: (2003, p.2), "Uma rede é um conjunto de itens, que chamaremos vértices ou nós, com conexões entre eles, chamadas arestas. Sistemas que tomam a forma de redes (Também chamado de grafos em grande parte da literatura matemática) são abundantes no mundo. Exemplos incluem a Internet, a *World Wide Web*, as redes sociais de convivência ou outras conexões entre indivíduos, redes organizacional, redes de relações comerciais entre empresas, redes neurais, redes metabólicas, redes alimentares, redes de distribuição, como os vasos sanguíneos ou as rotas de entrega postal, redes de citações entre papéis, e muitos outros"(NEWMAN, 2003).

Segundo Réka Albert e Albert-László Barabási (2002, p.2) "A estruturas de Redes complexas descreve uma grande variedade de sistemas de grande importância tecnológica e intelectual. Por exemplo, a célula é melhor descrita como uma complexa rede de produtos químicos ligados por meio de reações químicas; a Internet é uma rede complexa de roteadores e computadores ligados por vários *links* físicos ou sem fio; Ideias espalhadas na rede social cujos nós são seres humanos e arestas representam várias relações sociais; A *World Wide Web* é um enorme rede de páginas virtual conectados por A *hiperlinks*. Estes sistemas representam apenas alguns dos muitos exemplos que Recentemente, levou a comunidade científica para investigar os mecanismos que determinam a topologia de redes complexas"(ALBERT; BARABÁSI, 2002).

As redes complexas são representadas, em geral, por um grafo, que é uma estrutura de dados na qual os nós são conectados através de arestas. Formalmente, temos que um Grafo G(V,E) é uma representação matemática definida por Leonhard Euler em 1736,

quando publicou seu artigo referente as sete pontes de *Königsberg*. Euler se deparou com o problema que havia na cidade de Königsberg, lá existia um rio, *Pregel*, que cortava a cidade em quatro pedaços de terras e possuía uma ilha no centro como mostra na Figura 2.2.

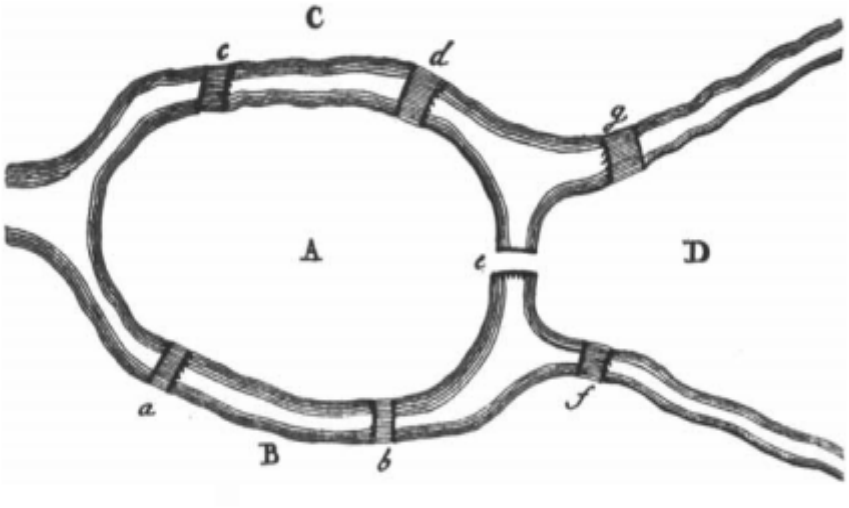


Figura 2.1: Ponte de Königsberg
Fonte: (EULER, 1953)

O problema era saber se era possível uma pessoa sair de um determinado local e retornar ao mesmo local passando por todas as pontes sem que passasse pela mesma ponte mais de uma vez. No total haviam quatro pedaços de terras que eram conectados através de sete pontes. Euler então classificou cada pedaço de terra como A, B, C, D, e provou que não existe um trajeto ao qual se possa partir de uma determinada região e retornar a mesma região sem que se repita ao menos uma ponte. Ao se deparar com esse problema Euler então criou toda a base matemática teórica que conhecemos como Teoria dos Grafos.

Ele modelou o problema da seguinte forma: cada pedaço de terra ele chamou como vértice ou nó (V) e cada ponte que conectava os vértices chamou de Arestas (E). Com isso, ele chamou de Grafo $G(V, E)$, o conjunto de vértices e arestas. Dessa forma conseguiu representar o problema.

O argumento usado por Euler diz que para que uma pessoa percorra o trajeto saindo de um determinado vértice (A) e passando por todas as arestas (E) sem que elas se repitam e no final retorne ao mesmo vértice (A), é necessário que o grau dos vértices

do grafo sejam par, o que não ocorre no caso das pontes de *Königsberg*.

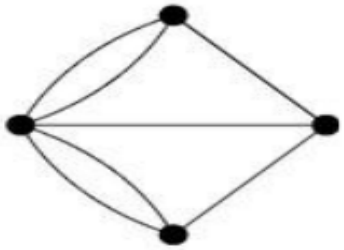


Figura 2.2: Ponte de Königsberg
Fonte: (EULER, 1953)

2.1.1 Redes Aleatórias - *(Random Networks)*

Redes aleatórias, do inglês *(Random Networks)* ou *(RN)*, são redes caracterizadas pela distribuição aleatória de seus elementos, onde cada vértice é conectado com outro vértice ao acaso. Segundo Newman (NEWMAN, 2001) uma rede aleatória é o caso mais simples de rede.

Em uma (RN) todos os vértices da rede possuem aproximadamente a mesma quantidade de intermediações. Nesse tipo de rede é muito raro encontrar vértices centralizadores de poder, (*hubs*), como em outros tipos de redes que serão apresentadas.

É importante frisar que no estudo de redes a distribuição de conectividade é uma característica que diz muito sobre a rede. Em RN essa distribuição é feita de uma forma igualitária, isso quer dizer que não existem vértices com vantagens sobre outros vértices na hora de efetuar conexões. Perante essa característica, as RN seguem a famosa distribuição de Poison, que é uma distribuição probabilística que expressa a probabilidade de uma variável aleatória sem influencias passadas.

O algorítimo que descreve esse tipo de rede, seleciona dois vértices ao acaso e acrescenta uma aresta ligando esses dois vértices. Posteriormente, ele seleciona outros dois vértices e repete o passo anterior até que todos os n vértices estejam conectados. Com isso cada uma das n arestas estão presentes com uma probabilidade p. O número de arestas conectadas por cada vértice seguirá uma distribuição de Poison. A conectividade das arestas não crescem proporcionalmente conforme o número de vértices (V) cresce na rede. Aplicando esse algorítimo a um conjunto finito de vértices, o resultado será uma

rede aleatória com uma distribuição de homogeneidade relativamente alta.

2.1.2 Redes Livre de Escala - $(Scale - Free - Networks)$

As Redes Livre de Escala, do inglês $(Scale - Free - Networks)$ ou SFN, chamaram uma grande atenção dos pesquisadores pois elas possuem características extremamente importantes para o funcionamento e manutenção das redes existentes. Elas se tornaram famosas pois vários tipos de redes do mundo real são desta classificação e representam quase a totalidade das redes sociais existentes.

As SFN seguem uma distribuição de grau conhecida com Lei de Potência (*Power-Law*). A Lei de Potência, Equação 2.1, permite que alguns nós da rede possam ter um alto coeficiente de conectividade enquanto sua grande maioria possui poucas conexões. O fato da rede possuir alguns nós com muitas conexões, explica o porquê redes sociais possuem uma operacionalidade alta. Essa operacionalidade característica em SFN é muito importante porque ela permite que a robustez da rede não seja afetada quando atacada. Redes SFN são redes muito robustas quando atacadas aleatoriamente, mas tornam-se extremamente vulneráveis quando esses ataques são sincronizados.

$$P(k) = K^{-\gamma} \tag{2.1}$$

Como mostra o experimento que Barabási fez em 1995 (BARABÁSI, 2001), ao mapear uma pequena parte da internet de 200 milhões de páginas web ele conseguiu chegar a uma distancia de 16 cliques para dois vértices distintos na rede. Esse número estava totalmente dentro do esperado pois o número tipico de cliques entre quaisquer outros dois nós da internet toda são cerca de 19 cliques. Com base nesse experimento fica provado que a internet satisfaz o conceito de *Small World* que será apresentado na seção seguinte. Isso acontece porque ela possui alguns nós com um alto coeficiente de conectividade, *hubs*, ou seja são nós que estão conectados a muitos outros nós da rede fazendo com que a grande estrutura da rede mantenha-se conectada.

Para exemplificar pensemos que as páginas web são nós de um grande grafo e que os links dessas páginas são conexões existentes entre elas. Tomemos como exemplo uma

página web que dificilmente, ou quase nunca, é acessada. Se por algum motivo essa página parar de funcionar, isso não acarretará nenhum dano a estrutura geral da internet. Por outro lado tomamos como exemplo a página do Google, uma página muito importante que é usada pela maioria das pessoas e consecutivamente possui uma grande quantidade de links que fazem comunicação com a internet. A página do google certamente seria um (*hub*) importante na estrutura da rede. O que aconteceria se ela fosse tirada do ar? Certamente criaria desconexões entre páginas menores causando assim um problema estrutural.

Um sociólogo chamado Robert K. Merton, nascido na Pensilvânia EUA, ao contribuir com estudos sociais formulou, no ano de 1968, um resultado chamado Efeito Matheus ou (*Matthew Effect*)(MERTON et al., 1968). Esse efeito diz que "o rico fica cada vez mais rico e o pobre cada vez mais pobre". Em seus resultados, Robert K. Merton, faz conexões com status sociais, poder aquisitivo, popularidade de pessoas entre outros fatores sociais existentes.

As SFN seguem o conceito de ligação preferencial ou *Matthew Effect*, que se baseia na ideia de que quando um vértice é adicionado a rede ele tende a se conectar com nós já existentes que possuem mais conexões possíveis *hubs*. Isso significa que os nós recém chegados não se conectam aleatoriamente na rede, logo as redes não são formadas por indivíduos com probabilidade iguais de possuírem o mesmo número de conexões.

Esse efeito justifica a existência de hubs em uma rede, se pensarmos que os *hubs* foram os primeiros vértices a serem inseridos na rede. Seguindo essa linha de raciocínio, uma rede será construída adicionando-se os novos elementos e respeitando a probabilidade preferencial dada pela equação:

$$P(k_i) = \frac{k_i}{\sum_{j=1}^{N} k_j} \tag{2.2}$$

A probabilidade do i-ésimo elemento é dada pelo número de conexões que esse i-ésimo elemento faz sobre o número total de elementos que existe na rede.

As SFN são extremamente robustas, pois a remoção aleatória de seus nós tendem

a não alter a topologia da rede. Isso já não acontece para uma RN, ao menos enquanto o coeficiente de conectividade for menor que 3 como demonstrado por Barabási em (BARABÁSI, 2001).

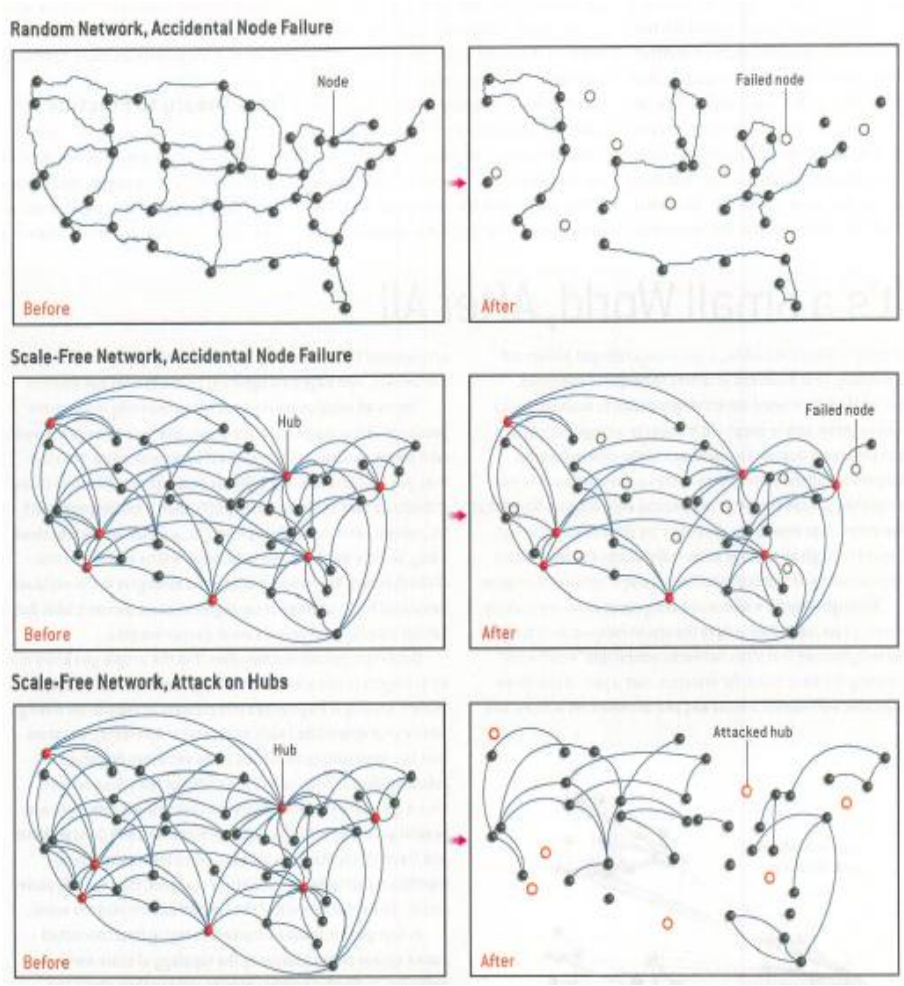


Figura 2.3: Análise da robustez de uma Rede Livre de Escala
Fonte: Scientific American (2003), Barabási, Albert and Bonabeau, Eric, (BARABÁSI; BONABEAU, 2003)

A maioria das redes sociais são classificadas como SFN e sua robustez pode ser associadas ao fato de não satisfazem um conceito físico chamado percolação.

A teoria da percolação diz que se possuirmos uma rede e começarmos a excluir nós da rede de uma forma randômica, e a uma certa frequência x, a rede resultante será constituída apenas de subgrupos pequenos e incomunicáveis Figura 2.3.

Pode-se notar esse acontecimento na Figura 2.3, mais especificamente no segundo quadrado, que representa uma falha em alguns nós de uma rede aleatória. É fácil notar que a estrutura da rede ficou bem danificada e totalmente desconexa para a RN.

Isso não é verdade para Redes Livres de escala, claramente conseguimos notar,

nessa mesma Figura 2.3, que mesmo quando *hubs* são desconectados da rede, boa parte de sua comunicação com os outros nós não deixam de existir. É notório, na Figura 2.3, que o número de componentes desconexas é menor nas SFN do que nas RN.

Segundo Barabási, (BARABÁSI, 2001), cerca de 80% dos nós de uma rede foram removidas e a rede resultante ainda permitia a conexão da maioria dos seus nós.

2.1.3 Redes Small World - *(Método de Watts e Strogatz)*

O conceito de redes de pequeno mundo tomou proporção maiores depois que o psicólogo Stanley Milgran e Jeffrey Travers realizaram um experimento na cidade de *Nebraska* (MILGRAM, 1967). Eles queriam responder questões como: "Qual é a probabilidade de duas pessoas selecionadas ao acaso no mundo conhecerem umas as outras?". "Dentre duas pessoas selecionadas aleatoriamente no mundo, quantas pessoas intermediarias são necessárias para que se estabeleça uma conexão entre elas?". Milgram e Travers chegaram a conclusão que duas pessoas escolhidas aleatoriamente poderiam ser ligadas em termos de seus conhecidos intermediários.

Buscando resposta para tal, eles deram inicio ao seu experimento. Primeiramente, eles escolheram seus grupos iniciais de participantes e chamaram de população. Cada uma dessas populações possuíam características diferentes, que o permitia compara-las a fim de responder suas duvidas. A primeira população foi um grupo de 100 pessoas que eram acionistas de uma empresa em *Nebraska*. A ideia com a escolha desse grupo, era saber se, por serem acionistas eles possuíam fácil acesso para completar o experimento. A segunda população possuía 96 pessoas e também era um grupo de *Nebraska*, mas era um grupo de pessoas escolhidas aleatoriamente. A terceira e ultima população era uma população da cidade de Boston com 100 voluntários, essas pessoas também foram escolhidas aleatoriamente. Elas foram escolhidas, pois Milgram queria confirmar se a distância geográfica realmente influenciaria nos resultados. Logo o experimento empírico foi baseado nesses três grupos de pessoas.

O experimento consistia em distribuir documentos para cada integrante da população. Essas pessoas seriam responsáveis por inicializar todo o processo de enviar o documento para uma pessoa alvo. A pessoa alvo era um acionista, Sharom Massachu-

setts, que vivia um bairro de *Boston*. Cada pessoa deveria entregar o documento a uma outra pessoa com o proposito de faze-lo chegar ao acionista em *Boston*. Com esse objetivo traçado, existia a possibilidade de algumas pessoas não conhecerem o alvo de *Boston*, logo a pessoa que não conhecia o acionista foi aconselhada a enviar o documento para uma pessoa que achava que estaria mais próxima do acionista.

Dentro do documento existiam várias instruções a fim de informar os participantes como a pesquisa estava sendo feita. Foram solicitados para os participantes informações como: Nome, sexo, idade, ocupação, etc. Esses dados permitiam analisar toda a trajetória do documento na rede. Existia também uma lista que informava todos os outros participantes que já tinham participado até aquele momento. O intuito era fazer com que pessoas repetidas não fossem re-escolhidas novamente. No final foi possível obter os caminhos específicos que o documento percorreu durante o experimento. Esses caminhos eram importantes pois, a análise do perfil de cada participante permitia inferir certas afirmativas.

Das 296 pessoas que pertenciam as 3 populações e que foram convidadas a participar do experimento, 217 delas realmente enviaram o documento, as outras por algum motivo, não deram continuidade no experimento.

No final do Processo, 64 documentos chegaram ao destino final, o correspondente a 29%. Para cada população foi calculado a quantidade de caminhos intermediários necessários para o documento chegar ao destinatário.

Para a população que foi escolhida aleatoriamente, em *Nebraska*, o comprimento médio necessário para que o documento chegasse a Boston foi de 5,7, isso significa que com menos de 6 pessoas intermediarias de ligação foi possível entregar um documento a uma distância de 1300 Milhas.

Para a população de *Nebraska*, que era acionista de uma empresa, a quantidade de intermediários necessários foi de 5,4 pessoas. Esse resultado não foi muito satisfatório já que população foi escolhida levando em consideração que eram assessores. Esperava-se mais de uma população com esse perfil.

E a ultima população foi a que se localizava exatamente na mesma região para onde o documento deveria ser enviado. A media de intermediários necessários foi de

4,4. Esse resultado comprova que a posição geográfica de seus indivíduos influência no resultado para se atingir uma certa pessoa na rede.

Esse resultado ficou conhecido como *"Small Word Effect"* que diz, que duas pessoas selecionadas ao acaso podem ser conectadas através de uma cadeia curta de conhecidos intermediários, que estão conectados, mesmo se a população da rede for muito grande.

Posteriormente ao experimento de Milgram, o matemático Steven Strogatz e o sociologista Duncan J. Watts em (WATTS; STROGATZ, 1998) propuseram um modelo para o *"Small Word Effect"* conforme mostra a Figura 2.4. Eles modelaram o problema da seguinte maneira. Primeiramente começaram com uma rede toda regular, onde cada vértice da rede possuía conexão com seus vizinho e vizinhos dos vizinhos. Posteriormente, com uma probabilidade p, que varia de 0 a 1, é escolhido aleatoriamente pares de nós, e posteriormente é alterado a conexão desses vértices. Quando a probabilidade p, de aleatoriedade, é igual a 1, isso quer dizer que será alterado aleatoriamente todos os pares de nós da rede, o que resulta em uma rede toda aleatória. Dada uma certa probabilidade p, acontece o chamado *"Small Word Effect"*, onde as redes possuem um caminho mínimo médio baixo e a conexão entre as comunidades da rede são pequenas. Isso possibilita que a comunicação entre esses vértices se torne muito rápido.

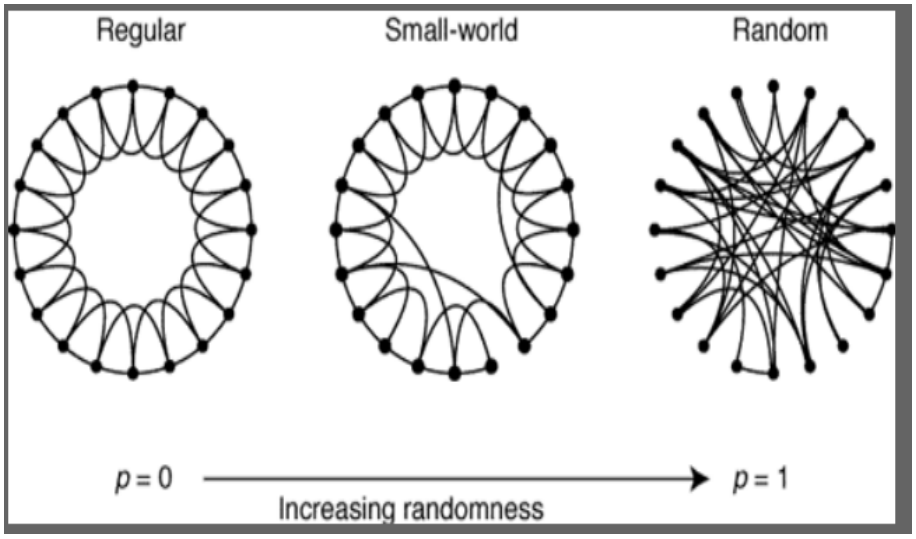


Figura 2.4: Modelo de Watts e Strogatz
Fonte: (WATTS; STROGATZ, 1998)

2.1.4 Redes Sociais

A Teoria de Redes Sociais pode parecer uma área de estudo nova devido a recente disseminação da internet e redes sociais virtuais, mas o fato é que ela já despertava interesse de pesquisadores em meados do século XIX. Uma contribuição muito importante para o desenvolvimento dos estudos partiu do Sociólogo Jacob Moreno, que em 1933 publicou pelo jornal *(New York Times)* um Sociograma, Figura 2.5, onde explorava a sociometria de cada individuo e todas as inter-relações na rede.

Esse Sociograma, que de uma forma simplificada pode ser vista como um Grafo, representava a rede de amizades de uma escola do ensino primário. Na representação direcionada criada por Moreno J. os meninos eram representados por triângulos e as meninas por círculos. As setas que partem de um indivíduo para o outro significam que, quem recebeu era considerado amigo de quem enviou. E as setas que possuem um traço no meio significava que ambos se consideravam amigos. Sociólogos queriam saber qual era o indivíduo que possuía o maior poder de influência sobre outros membros de uma rede de amigos.

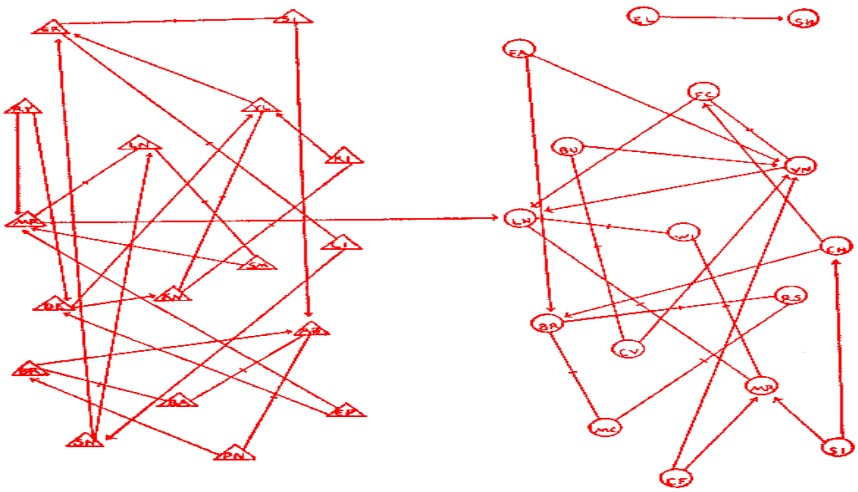


Figura 2.5: Jacob Sociogram
Fonte: (MORENO, 1933)

Pensando nisso, uma pessoa era classificada com muito influente quando possuía duas características. A primeira era a quantidade de indivíduos com quem ela se conectava e a segunda era quem possuía uma centralidade maior na rede.

Ao estudarmos essas redes sociais e outros tipos de redes existentes, como redes de

colaboração, redes de conhecimento, redes de organismos vivos, etc. observamos que essas redes não seguem um padrão regular, elas possuem características próprias que podem ser observadas de diferentes pontos de vista a fim de obter informações especificas.

Essas Redes são chamadas de Redes Complexas e é um ramo de pesquisa altamente ascendente devido a grande necessidade de se relacionar objetos e tentar entender por que essas relações acontecem.

Atualmente grandes áreas de pesquisa como matemática, física, química, computação entre outras, utilizam da teoria de redes complexas para modelagem de seus problemas.

Uma Rede complexa pode ser classificada como um tipo de grafo especial que possui características que o diferem de grafos convencionais. Algumas características como *(Small-World-Network)*, *(Scale-Free-Network)*, *(Power Law)*, *(Random Networks)* e serão apresentadas neste trabalho.

2.2 Métricas de Centralidade

O estudo de centralidades se iniciou na década de 1940, quando Alex Bavelas, na Universidade do *(M.I.T - Massachusetts Institute of Technology)*, estava procurando compreender como funcionava a comunicação entre indivíduos de pequenos grupos de pessoas. Ele chegou a uma conclusão que os indivíduos mais centrais desses grupos, eram os que possuíam um poder de influência maior do que os outros e indicavam que a centralidade estava relacionada com a eficiência do grupo para resolver problemas (BAVELAS, 1948).

A ideia principal do estudo da centralidade em uma Rede Social, é descobrir qual é o individuo mais poderoso pertencente a rede, quais são os lideres da sociedade, quais ajudam na propagação de uma informação ou doença, ou seja qual a importância desse autor dentre todos os outros presentes e consequentemente o quanto ele consegue influenciar os outros indivíduos que mantém contato com ele.

A dificuldade encontrada no estudo de centralidade, oriunda da falta de definição que essa métrica possui. Para alguns cientistas importantes no estudo de redes e suas propriedades como, (SHAW, 1954), (CZEPIEL, 1974), (NIEMINEN, 1973) entre outros, a centralidade pode ser definida de acordo com grau do nó na rede. Para esses pesquisadores,

esse é o fator principal e essencial na definição de centralidade, conforme citado por (FREEMAN, 1978/79).

Vamos analisar a rede da Figura 2.6. Conseguimos observar nessa figura que existe um vértice que possui uma grande quantidade de conexões, o vértice *(V10)*. Ele é adjacente a outros 10 pontos, como mostrado na Figura 2.6. Os Adjacentes do vértice *(V10)* são: *Adj(10)=(1,2,3,4,5,6,7,8,9,11)*. Supondo que os vértices *(V12),(V13)* não fossem adjacentes ao vértice *(V11)*, teríamos uma rede desconexa e com isso poderíamos analisar cada rede separadamente.

Analisando somente a rede do lado esquerdo, observa-se que ela possui uma estrutura semelhante a uma estrela. Sabe-se que para redes do tipo estrela com n vértices, o maior grau possível de um vértice seria *(n-1)* conexões. É possível observar também que o vértice *(V10)*, além de possuir um alto grau de conectividade, ele consegue atingir todos os outros vértices com apenas um passo. Se o vértice *(V10)* fosse uma pessoa infectada com alguma doença contagiosa, e as arestas significassem algum tipo de relação entre as pessoas desse grupo, certamente todas elas iriam se infectar muito rapidamente.

Agora analisemos a rede do lado direito do vértice (v11). É possível notar que essa rede segue uma estrutura de árvore binária. Para redes desse tipo tem-se que o grau de todos os vértices é menor ou igual a dois. Analisando essa árvore binária e levando em consideração somente a centralidade de grau, fica completamente difícil apontar, qual é o vértice mais central dessa rede, já que vários vértices podem ter o mesmo grau.

Como proceder então quando essas duas redes se unem e formam a rede da Figura 2.6. Por mais que o vértice *(V10)* possua um grau muito elevado, algo parece dizer que o vértice *(V11)* também é muito importante na estrutura da rede apresentada. Devido a dificuldade de responder questões desse tipo, surge a necessidade da criação de um novo conceito de centralidade, levando em consideração fatores estruturais presentes na rede.

O vértice *(V11)* possui uma centralidade que será apresentada nas próximas seções, chamada *Betweenness*. A ideia principal é que o vértice *(V11)* está localizado estrategicamente na rede. Esse tipo de centralidade mede o quanto uma pessoa está localizada em um menor caminho de comunicação que liga outros dois vértices da rede.

A centralidade nos diz o quão importante um nó é em uma rede. O conceito de

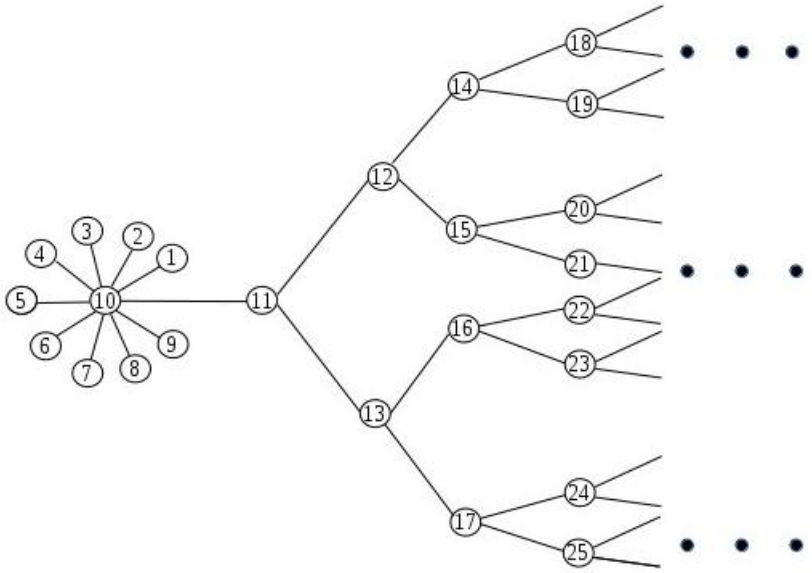


Figura 2.6: Centralidade em redes sociais

centralidade pode ser aplicado a qualquer tipo de rede que esteja sendo estudada. Por exemplo, em uma rede comercial, quais são os atores que mais exportam ou importam produtos de um determinado estado? Os caminhos que estão sendo trafegados são os melhores em custo beneficio para a empresa? E em uma Rede Biológica, quais são os genes dominantes responsável pelas distribuições das características de seus antecedentes? Para responder perguntas desse tipo, deve-se identificar qual a estrutura dominante na rede e de acordo com a estrutura identificar quais são os vértices mais centrais.

Para conseguirmos responder essas perguntas acima, será apresentado algumas métricas de centralidade que, permitem analisar a posição geográfica de um vértice dentro de uma rede social de acordo com sua localidade.

2.2.1 Centralidade de Grau

Talvez a centralidade de Grau seja o mais simples, mas não menos importante, atributo, que distingua o quão central um nó se localiza em uma rede. Olhando a grosso modo para um vértice v_k de um Grafo *(G=(V,E))*, a primeira característica que conseguimos verificar sem a necessidade de efetuarmos cálculos complexos é quantas arestas são incidentes nesse nó, ou seja, com quantas pessoas o nó v_k se comunica diretamente.

Conforme citado por (FREEMAN, 1978/79), a ideia de centralidade pontual foi proposta por vários pesquisadores como ((SHAW, 1954), (CZEPIEL, 1974), (NIEMINEN,

1973)) entre alguns outros que contribuíram para o desenvolvimento da métrica. Cada um deles tentava desenvolver uma definição concreta mas não era possível, pois os autores criavam a definição dentro do contexto estudado. Alguns desses contextos eram conceitos matemáticos para curvas, outros para fins de frequência de estatística outros deles eram para ajudar na derivação da matemática e alguns deles chegaram a ser campos de estudos restritos.

Segundo Freeman em seu artigo (FREEMAN, 1978/79), a definição geral do que viria a ser centralidade de grau partiu de Nieminen(1974). Ele faz a contagem de todas as arestas incidentes em cada vértice da rede e essa quantidade é definida como grau de um vértice v_k.

Para sociólogos, o grau de um nó é um atributo muito importante na análise de redes sociais, pois ele esta intrinsecamente ligado com a capacidade de propagação de uma informação ou uma doença.

Abaixo conseguimos ver a definição matemática que Newman faz para a centralidade de grau.

Definição: Dado um Gravo *(G=(V,E))* onde *(V)* é o conjunto dos vértices e *(E)* é o conjunto de arestas, o grau de um vértice *(K)*, denotado por k_i, é o número de arestas conectadas a ele. Matematicamente falando, temos que n é a quantidade de vértices existentes, e $A_{i,j}$ é a matriz de adjacência que contabiliza o valor 1 para cada incidência existente no vértice k_i. (NEWMAN, 2010)

$$k_i = \sum_{j=1}^{n} A_{i,j} \tag{2.3}$$

Para facilitar o entendimento, vamos analisar esse pequeno grafo da Figura 2.7 em função do grau de seus nós. Tem-se então uma rede não direcionada, composta pelo conjunto de vértices *V = (1),(2),(3),(4),(5),(6),(7)* e conjunto de arestas *E = (1,4),(2,4),(3,4),(4,5),(5,6),(5,7)*. Consegue-se verificar facilmente que o vértice com o maior número de conexões na rede é o vértice *(4)*. Além do vértice *(4)* possuir o maior número de conexões com seus adjacentes, ele permite uma maior comunicação e disseminação de informação. Com isso é possível afirmar que, em termos de centralidade de

grau, o vértice *(4)* é certamente o mais central.

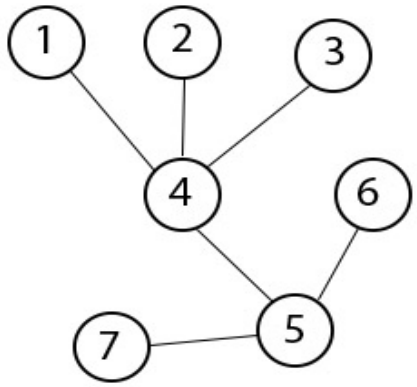


Figura 2.7: Grau de um vértice

Na rede exemplificada pela Figura 2.7, é possível visualizar que o vértice *(4)* se destaca também por ser um vértice que conecta conjuntos de vértices. Além disso, o vértice *(4)* está localizado no menor caminho entre vértices dos dois grupos. Esse conceito foi utilizado no resultado que será apresentado na próxima seção, chamado de centralidade de intermediação ou *Betweenness*.

2.2.2 Centralidade de Intermediação (*Betweenness*)

Dando continuidade com o estudo de centralidade, percebeu-se que, somente a quantidade de arestas incidentes em um vértice *(vk)* não é suficiente para classifica-lo como mais central da rede.

A medida de centralidade por intermediação foi implantada por Linton C. Freeman (FREEMAN, 1978/79), embora alguns outros cientistas já haviam explorado a ideia de medir a centralidade de um vértice de acordo com contagens intermediarias.

Foi observado que muitas vezes alguns vértices tinham o poder de controlar a informação que por ele passava. Esses vértices estão em posições geograficamente importantes para a comunicação dos outros vértices, fazendo com que os outros vértices da rede possuam um certo tipo de dependência desse vértice.

Tome como exemplo a rede representada pela Figura 2.8. Em uma rede dessa estrutura fica completamente difícil indicar qual vértice é mais central utilizando centralidade de grau, já que todos os vértices possuem quase a mesma quantidade de conexões.

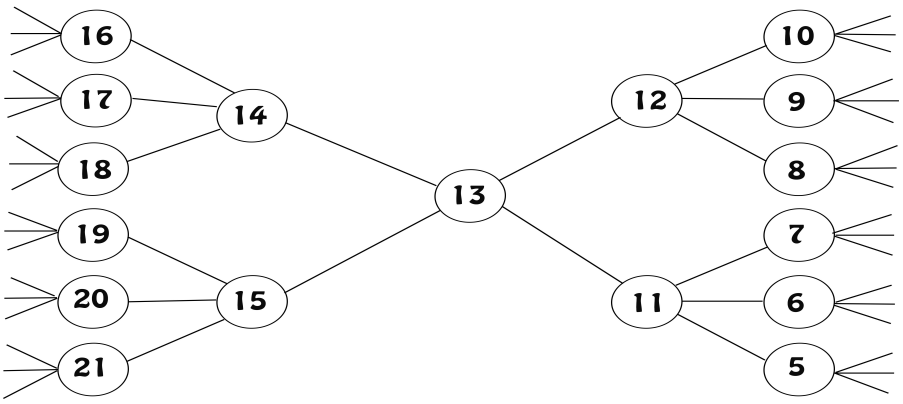


Figura 2.8: Centralidade em redes sociais(2)

Suponha que os vértices do lado esquerdo queiram se comunicar com os vértices do lado direito. Necessariamente, terão que passar pelo vértice *(V13)* para poder chegar ao outro lado. Isso caracteriza que toda a subárvore a esquerda depende do vértice *(V13)* para poder chegar ao lado direito e todos os vértices do lado direito também dependem do vértice *(V13)* para poder chegar do lado esquerdo. Então é possível afirmar que o vértice *(V13)* é o vértice mais central da rede. Isso ocorre por que o vértice *(V13)* está no menor caminho que conecta qualquer outros dois vértices da rede.

Para definir a centralidade de intermediação, primeiramente Freeman implementou o conceito de *Betweenness* parcial, que efetua alguns cálculos somente para um determinado vértice v_k da rede. Posteriormente generalizou o conceito de *Betweenness* para todos os vértices que compõe a rede.

Considere dois vértices v_i e v_j pertencentes a uma rede, se não existe um caminho que conecte diretamente esses vértices e não existe um nó v_k que faça a intermediação entre eles, então tem-se que o *Betweenness* do vértice v_k para esse caminho em especifico é igual a zero.

Podemos ver um desenho dessa rede Figura 2.9, onde os vértices v_i e v_j não conseguem trocar informação, ou seja, não existe uma aresta na rede que permita que ambos se comuniquem.

Se v_i e v_j não estão conectados diretamente, Figura 2.10, mas possuem caminhos mínimos que permitam que troquem informações através de outros nós, como por exemplo v_k, a probabilidade do vértice v_k estar entre v_i e v_j através de uma seleção aleatória é

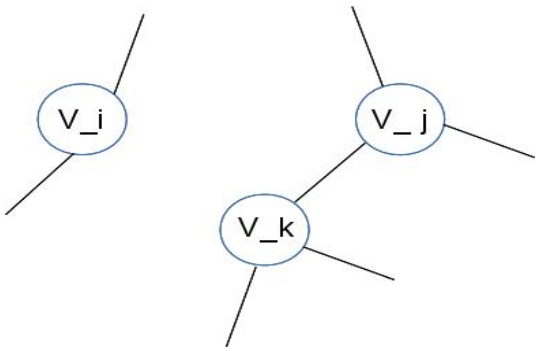


Figura 2.9: Rede onde o vértice "V_i" não é acessível

dada por:

$$\frac{1}{g_{ij}} \tag{2.4}$$

Onde (g_{ij}) é o número de caminhos mínimos existentes entre v_i e v_j na rede. A probabilidade de v_k cair em um caminho mínimo que conecta v_i a v_j dado uma seleção randômica é dada pela Equação 2.5.

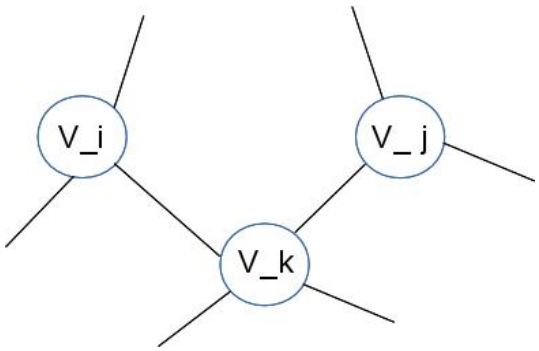


Figura 2.10: Rede onde o vértice "V_i" é acessível

$$b_{ij}(v_k) = \frac{g_{ij}(v_k)}{g_{ij}} \tag{2.5}$$

Na Equação 2.5, $g_{ij}(v_k)$ é igual ao número de caminhos mínimos que vão de v_i até v_j que passam por v_k.

Contudo, para calcularmos a centralidade final do vértice v_k, deve-se fazer o somatório de todos os valores parciais desses caminhos para todos os pares de pontos não ordenados, onde i$\neq j \neq k$.

$$C_B(v_k) = \sum_{i<j}^{N} \sum_{i<j}^{N} = \frac{g_i j(v_k)}{g_i j} \tag{2.6}$$

A Equação 2.6 apresenta a centralidade de intermediação ou *Betweenness*. Essa medida é muito importante em análise de redes complexas, pois ela mede o quanto no meio de um caminho que conecta outros dois vérticesv_i até v_j um vértice v_k está.

2.2.3 Centralidade de Proximidade *(Closeness)*

A centralidade de Proximidade foi uma medida de centralidade muito importante para o desenvolvimento dessa área de pesquisa. Como o próprio nome diz, a centralidade de Proximidade, (*Closeness*), indica o quão próximo um vértice se localiza geograficamente de outro.

Como diz Freeman em seu estudo (FREEMAN, 1978/79), essa medida sofreu uma grande quantidade de contribuições feitas por: (BAVELAS; BARRETT, 1951) , (BEAUCHAMP, 1965), (SABIDUSSI, 1966), (MOXLEY; MOXLEY, 1974) e (ROGERS, 1974), mas foi a partir da Definição de (SABIDUSSI, 1966) que definiu a medida, fazendo uma relação com o tempo de comunicação para com todos os outros nós da rede.

Foi visto anteriormente que se um vértice possuir a centralidade de intermediação, *(Betweenness)*, um valor alto, isso caracteriza que esse vértice tem poder de controle para com o fluxo de conexões que passam através dele na rede.

A centralidade de Proximidade *(Closeness)* tenta evitar que esse controle de comunicação seja atribuído a um vértice. Se um determinado vértice v_i não precisa se conectar a um vértice v_k para se conectar a v_j isso significa que o vértice v_i se conecta diretamente com o vértice v_j caracterizando então sua independência em relação a v_k, logo ele não contribui para que o vértice v_k aumente seu potencial de controle.

Para calcular o *Closeness* de um vértice v_i pertencente a uma rede, deve-se medir a distância dos menores caminhos que vão do próprio v_i a todos os outros vértices pertencentes na rede.

Foi definido por Newman em (NEWMAN, 2010), que o caminho mais curto existente ao longo de uma rede que interliga dois vértices v_i e v_j é chamado de caminho

geodésico desses dois vértices. Logo Newman definiu *li* sendo 1 sobre a quantidade de vértices multiplicado pelo somatório das distancias geodésica média dos vértices v_i e v_j para todos os vértices da rede.

Uma curiosidade a ser observada é que d_{ij} significa a quantidade de arestas existentes entre os vértices v_i e v_j.

$$l_i = \frac{1}{n}\sum_{j} d_{ij} \tag{2.7}$$

É sabido da teoria dos grafos, que ao medir a distância entre dois vértices, normalmente, estamos interessado em obter um número pequeno de passos para que a conexão seja efetuada. Pode-se notar isso, analisando o algorítimo de Dijkstra (SKIENA, 1990) que tenta encontrar o menor caminho entre um par de nós na rede.

Pensando em rede social, um indivíduo que possui uma média de menores caminhos com um valor pequeno, pode atingir outros elementos de sua comunidade com um tempo menor que uma pessoa que possui uma media de menores caminhos com um valor maior.

Isso permite afirmar que a Equação 2.7 possuí uma característica diferente das outras medidas de centralidade apresentadas, pois quanto mais central um vértice v_k é menor é o seu valor de proximidade, e quanto menos central o vértice v_k é maior é o seu valor. Com isso para chegar-se a um padrão na apresentação dos resultados, os pesquisadores definiram como Centralidade de Proximidade ou (*Closeness*), como sendo o inverso de *li* conforme mostra a definição a seguir.

Definição: Dado uma vértice v_k pertencente a uma Grafo G, a *Centralidade de Proximidade* do vértice v_k é dada fazendo o inverso da soma das *menores* distâncias de v_k para todos os outros nós pertencentes ao Grafo G.

$$C_c(v_k) = \frac{1}{\sum_{j=1}^{n} dist(v_j, v_k)} \tag{2.8}$$

A centralidade de proximidade é uma medida que pode ser deduzida naturalmente quando se olha para uma rede social. Sociólogos utilizam com frequência os resultados obtidos pelo cálculo dessa medida, pois o mesmo permite analisar o quão próximo um indivíduo está de outro, o que permite inferir na transmissão de doenças e vírus.

Entretanto, ela apresenta alguns problemas que dificultam a análise dos resultados. Conforme apresentado por Newman (NEWMAN, 2010) os valores calculados variam em um intervalo muito pequeno. Esses valores que representam as distâncias mínimas entre dois vértices tendem a ser pequenos. Esses valores pequenos dificultam à análise de vértices mais centrais para vértices menos centrais. Normalmente, a diferença acontece em casas decimais ou até centesimais nesse tipo de centralidade.

O segundo problema tem a ver com a forma como é calculado a centralidade de proximidade quando a rede possui mais de uma componente conexa. Suponha uma rede R é desconexa, e que o cálculo da *Centralidade de Proximidade* selecione um vértice v_i e um vértice v_j que não estão localizados na mesma componente conexa. Quando isso ocorrer, o cálculo do caminho geodésico d_{ij} tende ao *infinito*, o que altera o resultado final do *Closeness* para zero.

Nesse exemplo é mostrado o cálculo do *Closeness* da Rede representada pela Figura 2.11, utilizando a equação 2.8.

$C_c(1)= 1 \rightarrow 2 = 1;\ 1 \rightarrow 3 = 2;\ 1 \rightarrow 4 = 1;\ 1 \rightarrow 5 = 2;\ 1 \rightarrow 6 = 2;\ = 1/8 = 0{,}125$

$C_c(2)= 2 \rightarrow 1 = 1;\ 2 \rightarrow 3 = 1;\ 2 \rightarrow 4 = 2;\ 2 \rightarrow 5 = 3;\ 2 \rightarrow 6 = 3;\ = 1/10 = 0{,}1$

$C_c(3)= 3 \rightarrow 1 = 2;\ 3 \rightarrow 2 = 1;\ 3 \rightarrow 4 = 1;\ 3 \rightarrow 5 = 2;\ 3 \rightarrow 6 = 2;\ = 1/8 = 0{,}125$

$C_c(4)= 4 \rightarrow 1 = 1;\ 4 \rightarrow 2 = 2;\ 4 \rightarrow 3 = 1;\ 4 \rightarrow 5 = 1;\ 4 \rightarrow 6 = 1;\ = 1/6 = 0{,}167$

$C_c(5)= 5 \rightarrow 1 = 2;\ 5 \rightarrow 2 = 3;\ 5 \rightarrow 3 = 2;\ 5 \rightarrow 4 = 1;\ 5 \rightarrow 6 = 2;\ = 1/10 = 0{,}1$

$C_c(6)= 6 \rightarrow 1 = 2;\ 6 \rightarrow 2 = 3;\ 6 \rightarrow 3 = 2;\ 6 \rightarrow 4 = 1;\ 6 \rightarrow 5 = 2;\ = 1/10 = 0{,}1$

Analisando os cálculos obtidos para cada nó pertencente a rede, observa-se que os vértices possuem uma distribuição de *Closeness* quase uniforme, diferindo-se somente a partir da segunda casa decimal após a virgula. Esse fato é um dos problemas apresentado nos resultados da execução dessa métrica. Conforme dito acima, isso dificulta na análise dos resultados.

Os vértices (5, 6) possuem ambos resultado de $C_c = 0{,}1$. Se observarmos seus

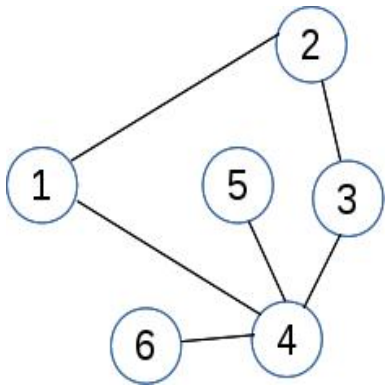


Figura 2.11: Closeness da Rede

graus, ambos possuem também grau um, logo esse resultado é esperado. Os vértices *(1, 2, 3)* apresentaram o cálculo do C_c como (0,125; 0,1; 0,125) respectivamente, e nota-se que os vértices possuem uma mesma quantidade na distribuição de grau, para todos os três vértices.

Consegue-se observar também, que o vértice *(2)* possui *Closeness* igual ao *Closeness* dos nós *(5, 6)*. Porém o grau do vértice *(2)* é igual a 2 e o grau dos vértices *(5, 6)* é igual a 1. Podemos afirmar nesse caso que, por mais que o vértice *(2)* possa parecer ser mais central devido ao fato de possuir grau maior, levando em consideração o *Closeness*, eles são exatamente iguais.

O vértice *(4)*, certamente é o vértice mais central da rede. Para uma rede qualquer, o máximo de conexões que um vértice pode ter é $n - 1$ conexões, o grau do vértice (4) seria *(n-1)* em seu máximo, porém nesse caso, seu grau é igual a *(n-2)*. Se compararmos as proporções da rede, esse valor é bastante alto. Nota-se também que o *Closeness* desse vértice é o maior de todos os outros apresentados *(0,167)*. Esse valor representa aproximadamente 25% de toda a rede. Ele indica que 1/4 da rede está centralizada nesse único vértice *(4)*. Esse vértice pode ser facilmente identificado como um *hub*.

Se estivéssemos observando uma rede social que representasse a propagação de uma doença, certamente se a contaminação começasse com o vértice *(4)*, ela se propagaria muito mais rápido, comprometendo então toda a rede.

3 Base de Dados

Neste capítulo será apresentado algumas informações pertinentes à base de dados da DBLP. Poderá ser verificado os trabalhos relacionados além de uma pequena estatística que informa a carga anual de dados. É possível verificar também o formato como os registros são disponibilizados além de uma pequena validação que foi feita na base.

3.1 DBLP

DBLP é uma base de dados oriunda da Universidade de Trier - *(Universität Trier)* localizada no oeste da Alemanha. Esta base originou-se de um pequeno projeto experimental que visava armazenar os registros de pesquisadores que colaboraram com a evolução da área da Ciência da Computação.

A partir dos registros apresentados e utilizando técnicas de análise de redes complexas em conjunto com teorias de redes sociais é possível verificar a rastreabilidade dos trabalhos dos pesquisadores, possibilitando extrair importantes informações sobre como os indivíduos se relacionam para com a comunidade científica.

Todos os registros foram disponibilizados em um documento XML que serviu como base para o estudo apresentado nesse trabalho. Nas próximas seções vamos explorar os dados presentes nessa base, para melhor compreender as manipulações feitas e entender os resultados obtidos.

3.2 Trabalhos Relacionados

Em maio de 2006, a DBLP possuía aproximadamente 750.000 publicações e aproximadamente 450.000 autores. Atualmente, sabe-se que a DBLP possuí mais de 3.5 Milhões de publicações e mais de 1.1 Milhão de autores.

Conforme mostrado em (BIRYUKOV; DONG, 2010), o processo de aquisição e manutenção dos registros encontrados na DBLP foram armazenados através do processo

ETL *(Extract Transform Load)*. Os dados são extraídos de fontes externas que armazenam registros de publicações escritas por pesquisadores, possivelmente são transformados para se adequarem as regras de negócio da construção da DBLP e depois carregados na DBLP.

Nesse documento é proposto uma medida de similaridade com base em uma rede de co-autoria modelada através da *Teoria dos Grafos* onde os autores são representados através de vértices e suas conexões são representadas através das arestas.

Um outro estudo interessante que foi feito levando em consideração o interesse em estudar a DBLP ocorreu em 2003 e a quantidade de dados existentes na base era muito diferente do que temos hoje.

A quantidade de registros utilizado no estudo realizado por (ELMACIOGLU; LEE, 2005) era de 38773 publicações com 32689 pesquisadores, um número razoavelmente pequeno comparado com os dias de hoje, mas os resultados obtidos foram realmente muito interessantes e merecem ser mencionados.

Conforme dito, em 1967 o pesquisador Stanley Milgran fez o experimento *"six degrees of separation"*, onde afirmava que quaisquer duas pessoas no mundo poderiam ser conectadas com somente seis passos de intermediação. Isso motivou os pesquisadores a calcular qual seria a distância media entre todos os pesquisadores da DBLP (ELMACIOGLU; LEE, 2005). O resultado obtido, com base nos últimos 15 anos que antecedem 2003, foram de 6 passos de intermediação.

Ainda segundo (ELMACIOGLU; LEE, 2005), observa-se que somente alguns poucos pesquisadores publicavam uma grande quantidade de artigos, enquanto a grande maioria de pesquisadores publicavam poucos artigos.

São mostradas também algumas estatísticas que se referem aos autores da base, por exemplo, a quantidade de pesquisadores novos que eram adicionados a DBLP por ano. Foi constatado que a partir de 1985 até 2003, foram adicionados novos pesquisadores a uma taxa de 10% ao ano conforme mostra a Figura 3.1.

Foi analisado também, o quão ativo eram os pesquisadores que compunham a DBLP conforme mostra a Figura 3.2. Para isso foi definido como pesquisador ativo, aquele pesquisador que conseguia contribuir com pelo menos um artigo por ano. Percebeu-se, que no ano de 2003 dos mais de 32 mil pesquisadores existentes no banco, somente 6 mil eram

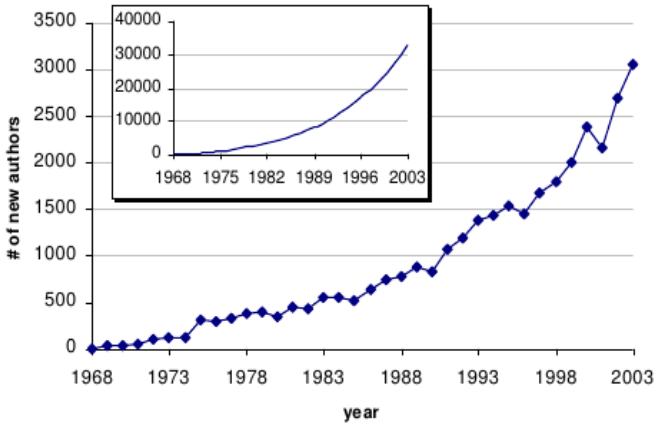


Figura 3.1: Número de novos pesquisadores adicionados à DBLP de 1968 - 2003
Fonte: (ELMACIOGLU; LEE, 2005)

pesquisadores que contribuíam anualmente com a ciência. Desses 6 mil pesquisadores que contribuíam com a ciência, foram identificados que 3 mil eram novos pesquisadores que estavam ingressando em estudos científicos e somente 3 mil eram pesquisadores antigos que estavam ativos.

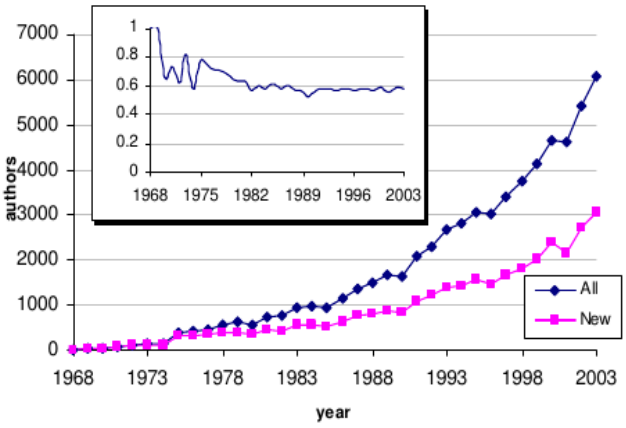


Figura 3.2: Número de pesquisadores ativos na DBLP
Fonte: (ELMACIOGLU; LEE, 2005)

O documento monstra também, o número médio de artigos publicados anualmente por autores na Figura 3.3. Os valores mostram que, a quantidade de artigos publicados por ano, mais especificamente em 1969, aumentaram em 0,2% em cima do valor inicial. Com esse aumento a média passou a ser de 1,2. Posteriormente esse valor foi diminuindo até o ano de 1980 onde se estabilizou por aproximadamente 0,3 artigos publicados por pesquisadores.

Foi analisado também, conforme mostra a Figura 3.5, o número de colaboradores

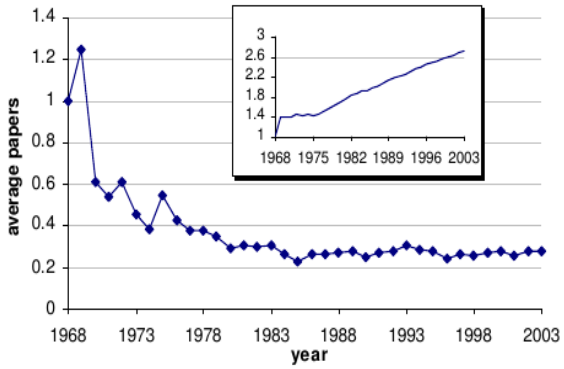


Figura 3.3: Média de artigo publicado por autor ao passar dos anos.
Fonte: (ELMACIOGLU; LEE, 2005)

que cada autor possui. O estudo realizado entre 1968 à 2003, mostra que o número de colaboradores por autor era da media de 3.93 e tende a aumentar de forma constante. Se analisarmos a curvatura do gráfico que mostra os colaboradores em função do ano, podemos ver que essa curva tende a crescer com o passar dos anos. Isso mostra que mais e mais pesquisadores estão contribuindo com os estudos científicos.

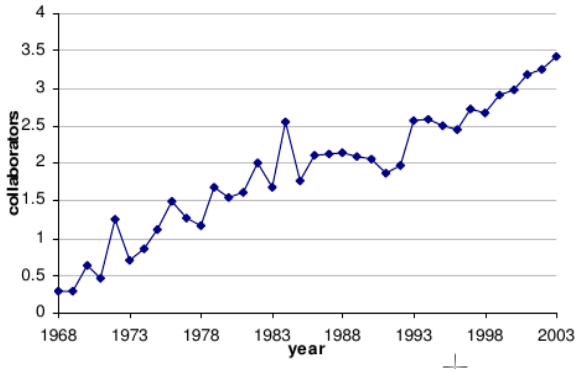


Figura 3.4: Média do número de colaborador por autor a cada ano.
Fonte: (ELMACIOGLU; LEE, 2005)

Além dos pesquisadores que contribuem com auxílio de co-autores, existem também os pesquisadores que publicam artigos sozinhos, ou seja, sem ajuda de co-autores conforme mostra a Figura 3.5. Esse número chegou a 3073 pesquisadores no ano de 2003 o que correspondia à 9.4% da DBLP naquela época.

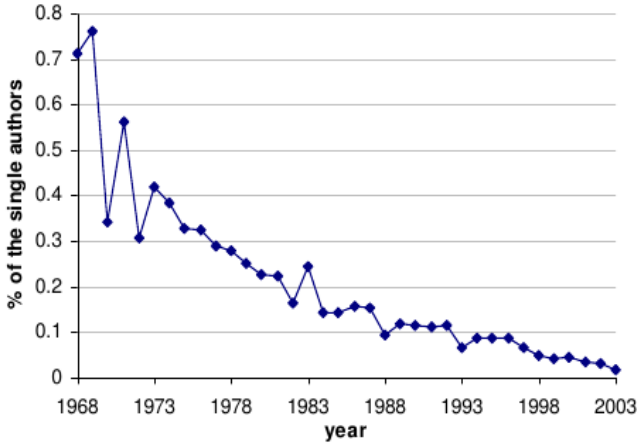


Figura 3.5: Porcentagem de autores sozinhos por documento.
Fonte: (ELMACIOGLU; LEE, 2005)

3.3 Análise da Base

Até o Março de 2017, foi constatado que a DBLP possui cerca de 3.751.554 publicações cadastradas. Entre essas publicações existem vários tipos de registros que se diferenciam uns dos outros como: artigos (jornais, revistas, conferências), teses de mestrado, páginas web e alguns outros tipos de publicações da área da ciência da computação. Em nosso estudo analisaremos somente as iterações entre pesquisadores que contribuíram para a comunidade científica através de artigos, pois a grande quantidade de registros que compõe a DBLP impossibilita o processamento dos dados.

Dentre esses artigos incluem-se: artigos publicados em jornais, revistas, transações, e boletins informativos. Além desses tipos específicos foram inseridos alguns artigos de conferências *"conf"* e artigos de pessoas *"persons"*.

De acordo com (LEY, 2009) a base de dados da DBLP já foi utilizada em mais de 400 publicações que visavam retirar informações importantes sobre a comunidade científica.

Os dados disponibilizados são de livre acesso, qualquer pessoa interessada em estuda-la poderá baixa-la e iniciar seus estudos. Uma das vantagens de se analisar uma base de dados como a DBLP é que essa base possui um alto números de registros reais. Por se tratar de uma base real é notório a existência de alguns desafios, que serão tratados para apresentar o melhor resultado possível.

Ao observarmos a Figura 3.6 é possível visualizar a quantidade de dados que

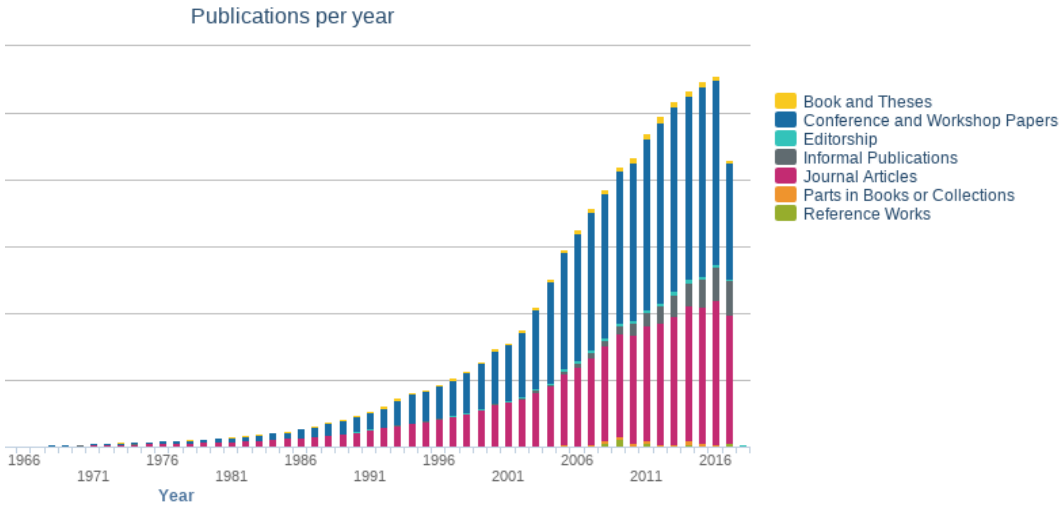


Figura 3.6: Estatística - Carga anual de dados
Fonte: http://dblp.uni-trier.de/statistics/newrecordsperyear

foram inseridos desde 1968 até o presente momento. Dentre os tipos de registros que foram adicionados a base, observa-se a existência de diferentes tipos de dados como: artigos de jornais, livros e teses, publicações informais, documentos de conferencias e *workshops* entre outros.

3.4 Formato dos dados

Os dados disponibilizados pela Universidade de Trier são estruturados em uma sintaxe mundialmente padronizada e muito fácil de compreender, a linguagem de marcação XML. O documento é composto por uma Tag raiz dblp/dblp e dentro dela, é possível encontrar todos os registros científicos, que foram disponibilizados de diferentes regiões do planeta, juntamente com seus respectivos atributos.

Dentro do XML é possível encontrar vários tipos de objetos *(Article, Inprocedings, Procedings, Book, Incollection, WWW)*. Como dito no tópico anterior, foram analisadas nesse trabalho somente as relações entre os objetos do tipo Artigo. Os dados foram armazenados em um banco de dados NoSQL, o Neo4j, que é um banco de dados orientado a grafos. Esse tipo de banco de dados foi escolhido, pois entende-se que ele atende aos requisitos no processo de representação de uma rede social científica.

Na Figura 3.7, temos um exemplo real de um registro da DBLP, é possível ver que esse registro possui alguns atributos e características importantes que serão discutida

a seguir.

```
<article key="journals/cacm/Szalay08"
    mdate="2008-11-03">
  <author>Alexander S. Szalay</author>
  <title>Jim Gray, astronomer.</title>
  <pages>58-65</pages>
  <year>2008</year>
  <volume>51</volume>
  <journal>Commun. ACM</journal>
  <number>11</number>
  <ee>http://doi.acm.org/10.1145/
    1400214.1400231</ee>
  <url>db/journals/cacm/
    cacm51.html#Szalay08</url>
</article>
```

Figura 3.7: Representação de um registro do tipo Artigo
Fonte: http://dblp.uni-trier.de/xml/docu/dblpxml.pdf

Todo registro contido no XML possuí um número identificador único, que no nosso caso é caracterizado pelo atributo *key*. A chave de cada registro é uma sentença separada por barras, onde, cada palavra traz informações importantes sobre o registro.

Na Figura 3.7 podemos ver um exemplo de um objeto do tipo *"artigo"* que contém como chave *"journals/cacm/Szalay08"*. Para esse registro em específico, temos que a primeira palavra que compõe a chave, *"journals"*, significa que o referente registro é um artigo que foi publicado em *jornais, revistas, transações ou boletins informativos*.

A segunda parte da chave de um registro, normalmente, designa a série de conferência ou periódico que o artigo apareceu. A última parte faz referência ao nome do autor, juntamente com o ano de publicação do artigo.

O atributo *"mdate"* faz referência à data da última atualização do artigo, isso é importante pois em aplicações que deseja-se analisar somente artigos atuais, pode-se utilizar desse atributo para efetuar as pesquisas desejadas.

A Tag *"author"* faz referência ao nome de cada autor que contribuiu para a escrita do artigo. O foco desse trabalho está em entender, utilizando técnicas de redes complexas, como esses autores se relacionam entre si.

Devido ao fato da base DBLP ser utilizada mundialmente e sabendo que a escrita do nome de cada autor pode ser expressa de várias maneiras diferentes, verificou-se a existência de alguns desafios que devem ser analisados cuidadosamente.

Sabe-se que nomes ocidentais normalmente começam com nomes seguidos de

sobrenomes familiares. Estes diferenciam-se de nomes islandeses que possuem por características junções de nomes e sobrenomes em uma única palavra. A não padronização dos nomes é um dos diversos desafios que foram expostos pelos criadores da DBLP (LEY, 2009).

Para exemplificar a dificuldade de trabalhar com os nomes contidos na DBLP, o autor *"Alexander S. Szalay"* poderia ter sido inserido na DBLP de diferentes formas: *A.S. Szalay, Alexander. S. Szalay, Alexander S. S., etc.* Nesse trabalho, caso o mesmo autor possua uma variação em seu nome como exemplificado acima, foi considerado um autor diferente para cada nome, ou seja, foi considerado uma pessoa diferente, mesmo quando eles representam uma só entidade.

3.5 Validando registros

A base DBLP que está sendo utilizada para a execução desse trabalho possui cerca de 3.751.554 registros. No total foram contabilizados 1.614.210 artigos inseridos no Neo4j através de um algoritmo que utiliza uma biblioteca SAX para manipulação do XML.

Levando em consideração a quantidade de registros que está sendo estudado, torna-se necessário fazer uma breve validação dos dados extraídos do XML. Verificando os registros inseridos, foi constatado o aparecimento de alguns artigos que não serão úteis para o desenvolvimento deste trabalho, pois tratam de artigos que, por algum motivo, não possuem os nomes dos autores que os escreveram.

```
<article mdate="2013-02-13" key="dblpnote/duplicate" publtype="informal publication">
<title>(duplicate entry was deleted)</title>
</article>
<article mdate="2013-02-10" key="dblpnote/withdrawn" publtype="informal publication">
<title>(paper withdrawn)</title>
</article>
<article mdate="2014-05-26" key="dblpnote/error" publtype="informal publication">
<title>(error)</title>
</article>
<article mdate="2013-02-12" key="dblpnote/ellipsis" publtype="informal publication">
<title>&#8230;</title>
</article>
<article mdate="2013-02-12" key="dblpnote/neverpublished" publtype="informal publication">
<title>(was never published)</title>
</article>
<article mdate="2017-01-13" key="dblpnote/retracted" publtype="withdrawn">
<title>paper retracted</title>
</article>
```

Figura 3.8: Objetos excluídos (1)

O número de artigos inutilizáveis chega a 8.716. Estes artigos foram excluídos do Neo4j, pois eles não forneciam seus respectivos autores o que não condiz com os objetivos desse trabalho. As Figuras 3.8 e 3.9, exemplificam a ausência desse atributo nos registros.

Figura 3.9: Objetos excluídos (2)

É possível notar que mais de 40% da DBLP é composta por artigos. Isso permite criarmos uma grande rede de iteração entre seus autores. Após a criação da rede foi possível aplicarmos as métricas conhecidas, para inferirmos resultados.

Após os artigos serem validados, sua representação no banco se mostra igualmente aos dados fictícios criados na Figura 3.10. É possível visualizar as iterações entre os artigos inseridos (*rosa*) e objetos do tipo autor (*verde*). Cada artigo está conectado com no mínimo um autor pela aresta *escrito*, e cada autor deve ter contribuído com no mínimo um artigo. Com essa modelagem é possível analisar os artigos que foram escritos pelos respectivos autores.

Após a inserção, foi constatado 1.605.494 artigos e 4.449.870 autores inseridos. Dentre esses autores, existem alguns autores que foram inseridos mais de uma vez no banco. Para solucionar esse problema foram removidas todas as instâncias repetidas dos autores no banco, logo após, foram conectados seus respectivos artigos à um único autor conforme mostra a Figura 3.11.

Após a inserção dos registros no banco e sua validação, constatou que mais de 70% dos autores eram autores repetidos, que estavam presentes nas publicações de vários artigos. Com a remoção de todos os autores duplicados, a rede ficou com 985.672 autores, onde cada um representa uma única entidade. Esses autores foram utilizados como a base da rede que será apresentada no capítulo seguinte.

Figura 3.10: Ilustração da inserção de registros no Neo4j

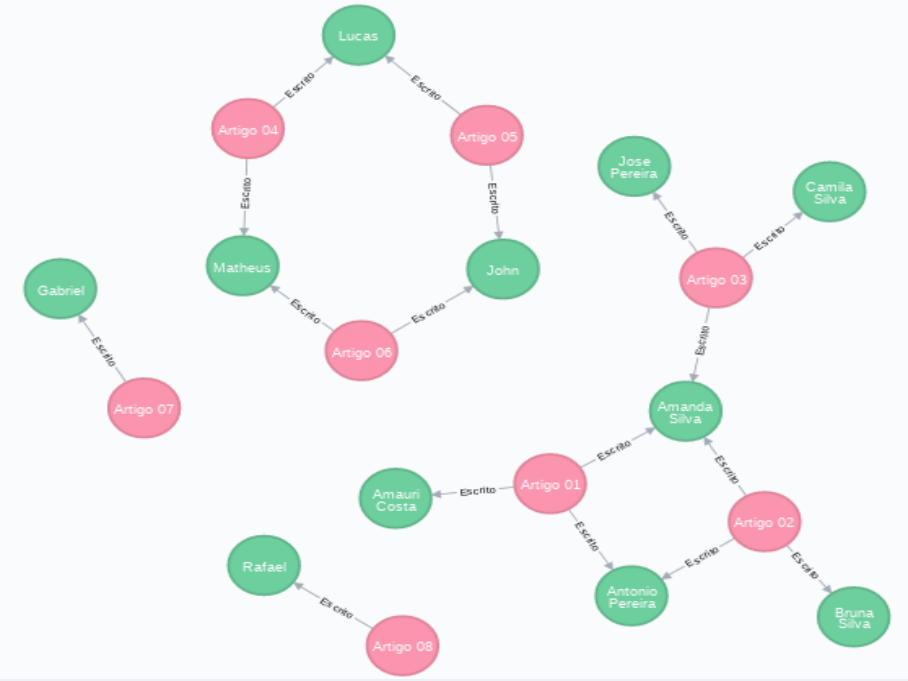


Figura 3.11: Removendo autores repetidos do Neo4j

4 Análise da Rede Científica

Neste capitulo será apresentado a modelagem de uma rede social científica que utiliza a métrica proposta por (STROELE et al., 2017). Após modelada a rede, foi apresentado uma análise topológica da rede. É possível constatar também que foram identificadas as componentes conexas que compõe a rede. Finalmente é possível verificar que foi calculado a influência para cada pesquisador, posteriormente é apresentado a sua distribuição de influência juntamente com suas devidas análises.

4.1 Análise da Influência entre os Pesquisadores

A inserção dos registros no Neo4j caracteriza a criação de uma rede que é representada por um grafo direcional, Figura 3.11. Nessa rede existem três tipos de objetos, dois deles são representados através de nós, e um através de arestas. Entre os nós, possuímos artigos e autores, e entre as arestas possuímos relações que partem de artigos e chegam em autores através do rótulo *Escrito*.

Para analisarmos a influência entre os pesquisadores, torna-se necessário a criação de mais um objeto do tipo aresta. Segundo Newman pesquisadores estão relacionados se ambos contribuíram para a publicação de um artigo científico (NEWMAN, 2001).

Como a Figura 3.11 representa a modelagem feita entre pesquisadores e artigos, é possível identificarmos quais são os pesquisadores relacionados na rede. Se dois pesquisadores A e B estão relacionados, o modelo de rede científica adotado cria arestas direcionais nos dois sentidos, que representarão a influência entre eles, conforme a Equação 4.1.

Pelo fato da aresta ser direcionada é possível visualizarmos os pesquisadores influenciadores e também os pesquisadores influenciados.

Baseado em (STROELE et al., 2017) o lado esquerdo da equação 4.1 representa o valor de influência que um pesquisador B exerce sobre um pesquisador A. O numerador da fração simboliza o número de publicações em que o pesquisador B e o pesquisador A são co-autores, ou seja, o número de publicações que publicaram juntos. O denominador da

fração simboliza o módulo do total de publicações do pesquisador A. Esse cálculo é sempre executado nos dois sentidos da relação e também para todos os pares de pesquisadores que se relacionam na rede.

$$IP_{AB} = \frac{||P_A \cap P_B||}{||P_A||} \tag{4.1}$$

Após o cálculo ser realizado para todos os pesquisadores que estão relacionados, a representação da rede social científica se assemelha com a Figura 4.1

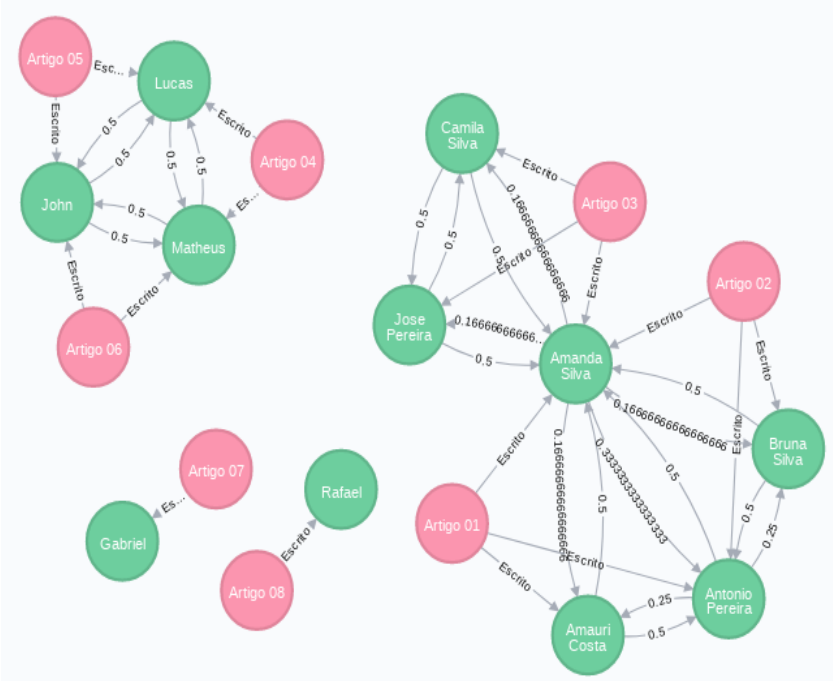


Figura 4.1: Cálculo de Influência para a rede de pesquisadores

É possível notar na Figura 4.1 a existência de alguns elementos que já não são mais necessários, como por exemplo os *Artigos* e as arestas de rótulo *Escrito*. Esses elementos serviram para fazer a relação entre os pesquisadores e os artigos. Após a exclusão desses objetos, a rede social cientifica se assemelha com a Figura 4.2

De acordo com a Figura 4.2 temos a modelagem da rede social científica que caracteriza a iteração entre os pesquisadores baseado no cálculo da influência. Ao analisarmos essa rede, podemos notar que as arestas que se conectam em um pesquisador possuem dois sentidos.

Dado uma aresta que parte de um pesquisador A e chega a um pesquisador B, é possível tirarmos duas informações dessa relação. Como a relação parte do pesquisador A

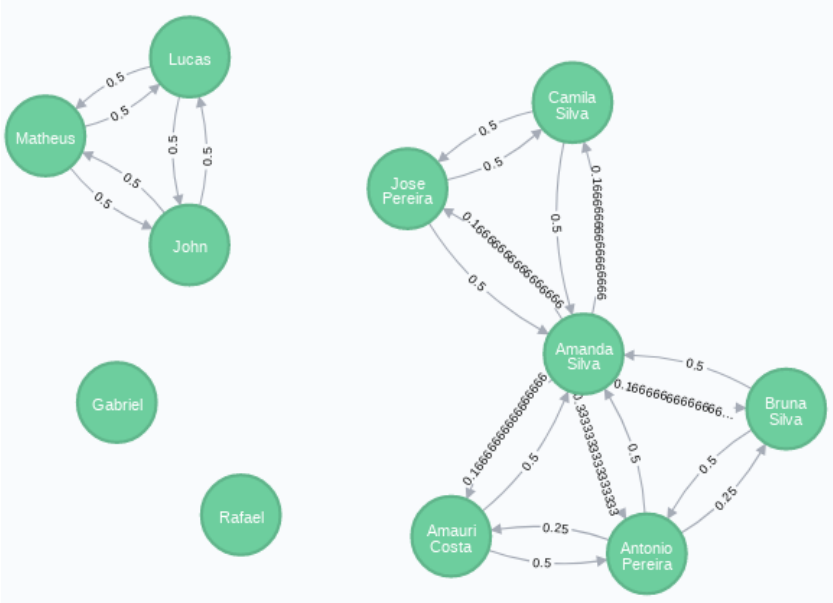


Figura 4.2: Rede de influência entre pesquisadores

para o pesquisador B, isso indica o quão dependente o pesquisador A é do pesquisador B. Tomando a mesma relação, se analisarmos do ponto de vista do pesquisador B, a relação indica o quão influente o pesquisador B é sobre o pesquisador A, conforme a equação 4.1 sintetiza.

4.2 Análise Topológica da Rede

Após a construção da rede é possível iniciarmos o estudo de sua topologia. Com esse estudo ficará mais fácil entender o porquê alguns fenômenos acontecem na rede. Porém para entendermos esses fenômenos, devemos primeiramente entender sua estrutura, que fornecerá elementos essenciais para essa análise.

Por exemplo, suponhamos que a OMS *(Organização Mundial da Saúde)* queira entender o porquê várias pessoas de uma cidade estão morrendo muito rápido. Para descobrir isso, seria necessário mapear todas as pessoas da cidade e suas possíveis conexões. Após mapear as conexões, seria possível ver quais eram as pessoas que mais se conectavam umas com as outras e, a partir disso caracterizar um grau de conexões. É possível deduzir que se uma pessoa com poucas conexões ficar com uma doença ela poderá transmitir essa doença para menos pessoas do que uma pessoa que possui muitas conexões. Logo, através da análise do grau de cada pessoa é possível inferir que as pessoas com grau de contato

mais alto, poderiam infectar mais pessoas, o que poderia causar mais morte, do que as pessoas com grau pequeno.

Ao analisarmos a topologia da DBLP, é possível responder perguntas não triviais que a caracterizam. Isso facilita entender como a rede se comporta perante fatores internos e externos.

Para darmos início a análise, vamos primeiramente calcular a distribuição de grau com base nas conexões de influência de cada pesquisador da rede. Vale ressaltar que essa rede possuí uma grande quantidade de vértices 1.175.971 e relacionamentos 9.141.438. O resultado é apresentado pela Figura 4.3.

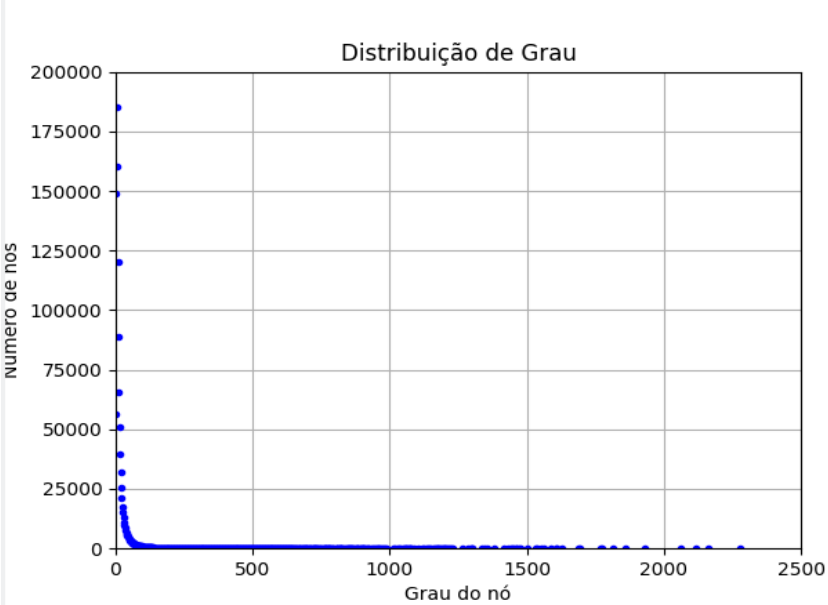


Figura 4.3: Distribuição de Grau da DBLP

A Figura 4.3 mostra a distribuição de grau da DBLP e a Tabela 4.1 mostra a distribuição dos 50 graus mais recorrentes da rede.

Conforme podemos observar na distribuição dos 50 graus mais recorrentes da rede mostrado na Tabela 4.1, constatamos que existem 148924 vértices com grau 2 e 184269 vértices com grau 4.

O fato do valor da quantidade de nós com grau 2 ser menor que a quantidade de nós com valor 4 é uma prova que essa rede não satisfaz o conceito da lei de potência.

Apesar da rede não satisfazer a lei de potência, e consequentemente indicar não ser uma rede livre de escala, ela é semelhante para valores onde o grau é maior que 2.

Sabe-se da robustez que as redes livre de escala possuem, elas são capazes de

Tabela 4.1: Distribuição dos 50 graus mais recorrentes da rede

Grau Total do nó	Número de nós com esse grau	Grau Total do nó	Número de nós com esse grau
0	56491	50	3775
2	148924	52	3528
4	185269	54	3186
6	160400	56	2911
8	119989	58	2705
10	88788	60	2447
12	65375	62	2442
14	50880	64	2143
16	39464	66	2063
18	32268	68	1923
20	25439	70	1805
22	21074	72	1693
24	17691	74	1585
26	15029	76	1428
28	13066	78	1347
30	11173	80	1346
32	9833	84	1132
34	8607	86	1104
36	7447	88	969
38	6865	90	953
40	6226	92	975
42	5487	94	941
44	5059	96	836
46	4754	98	786
48	3983	100	742

suportar uma porcentagem de falhas sem que comprometa sua estrutura e consequentemente seu funcionamento. Há indícios, que rede de influência calculada possa possuir essa robustez, isso se dá, pois seu quantitativo maior está nos vértices de baixo grau.

Podemos notar na Tabela: 4.1 que existem alguns nós de grau 0. Esses nós são pessoas que não possuem relação de influência com ninguém. Possivelmente são novos pesquisadores que entraram na rede e publicaram artigos sozinhos.

Nota-se que os nós com um grau pequeno são nós mais recorrentes na rede. Isso significa que a DBLP possui muitos pesquisadores com poucas relações (baixo grau) e poucos pesquisadores com muitas relações (alto grau). Pode-se notar na Figura 4.3 que poucos pesquisadores possuem o grau acima de 1500 e muitos pesquisadores possuem grau menor que 100.

4.3 Identificação de Componentes Conexas

Conforme podemos observar na Figura 4.2, essa rede foi calculada através da Equação 4.1 e resulta em um grafo com várias componentes conexas. Foi feito um algoritmo que calcula e contabiliza a quantidade de componentes na rede conforme apresentado a seguir:

Algoritmo 1: Algoritmo que identifica as componentes conexas.

```
   Entrada: Rede de influência, (rede).
   Saída: componentes conexas
 1 início
 2     ler rede;
 3     cria dicionarioAdjacencia;
 4     para todo no na rede faça
 5         se no ∉ dicionarioAdjacencia então
 6             dicionarioDeAdjacencia ← no  ;
 7         senão
 8             dicionarioDeAdjacencia[no] ←vizinhos no;
 9         fim se
10     fim para
11     cria valorComponente = 0;
12     cria vetorComponente;
13     para todo no em dicionarioDeAdjacencia faça
14         se no e seus vizinhos ∉ vetorComponente então
15             vetorComponente ← no And vizinhos  ;
16             remova no e vizinhos do dicionarioDeAdjacencia  ;
17             para cada no vetorComponente volte a linha 14 ;
18         senão
19             va para o proximo no em vetorComponente;
20         fim se
21     fim para
22     marque com valorComponente todos valores de vetorComponente;
23     valorComponente ← valorComponente + 1;
24     vetorComponente ← Null;
25     volte a linha 13 até dicionarioDeAdjacencia ser todo percorrido;
26 fim
```

Após a execução do Algoritmo 1, foi observado que a rede de influência da DBLP criada anteriormente possui várias componentes independentes. Um total de 99986 componentes foram encontradas conforme apresentado na Tabela 4.2.

Pode-se notar, que as componentes de tamanho 1 são pessoas que não possuem relação de influência com ninguém. Como dito anteriormente provavelmente são novos pesquisadores que entraram na rede publicando sem coautores.

Pode-se observar também que, foram encontradas 22705 componentes de tamanho

2. Se multiplicarmos as quantidades de componentes com quantidade de pesquisadores que compõem essas componentes, somam-se 45410 pesquisadores. Esses pesquisadores formam pequenos grupos de co-autoria.

Analisando os valores obtidos na tabela, nota-se que existe um certo padrão entre a quantidade de componentes e a quantidade de nó existente na componente. Quanto mais pesquisadores existem na componente, menor é a quantidade de componentes existentes com esse tamanho e analogamente, quanto menos pessoas existem na componente, mais componentes existem desse tamanho.

Tomando essa breve análise, nota-se que existe uma componente que se destaca das demais. O valor da maior componente conexa é de 985.672 pesquisadores. A maior componente conexa possui um valor muito alto se comparado com a segunda maior componente conexa de 44 pesquisadores.

Utilizaremos daqui para frente como objeto de estudo a maior componente conexa. Deseja-se analisar uma rede de tamanho expressivo, que possua características interessantes que possam ser exploradas. Procura-se entender como os pesquisadores dessa maior componente estão interagindo, quais são os pesquisadores mais influentes e como funciona sua topologia.

Tabela 4.2: Componentes Conexas da Rede

Nº de Nós	Nº de Componentes	(Nº de nós) X (Nº de componentes)
1	56491	56491
2	22705	45410
3	10365	31095
4	4779	19116
5	2319	11595
6	1306	7836
7	670	4690
8	460	3680
9	266	2394
10	187	1870
11	117	1287
12	89	1068
13	69	897
14	35	490
15	31	465
16	23	368
17	17	289
18	10	180
19	7	133
20	13	260
21	5	105
22	5	110
23	3	69
24	4	96
29	2	58
31	1	31
32	1	32
33	2	66
37	2	74
44	1	44
985672	1	985672
Total:	99986	1175971

4.4 Maior componente conexa

Para continuarmos as análises vamos fazer a distribuição de grau para cada pesquisador da maior componente conexa da rede. Ela pode ser verificada através da Figura 4.4.

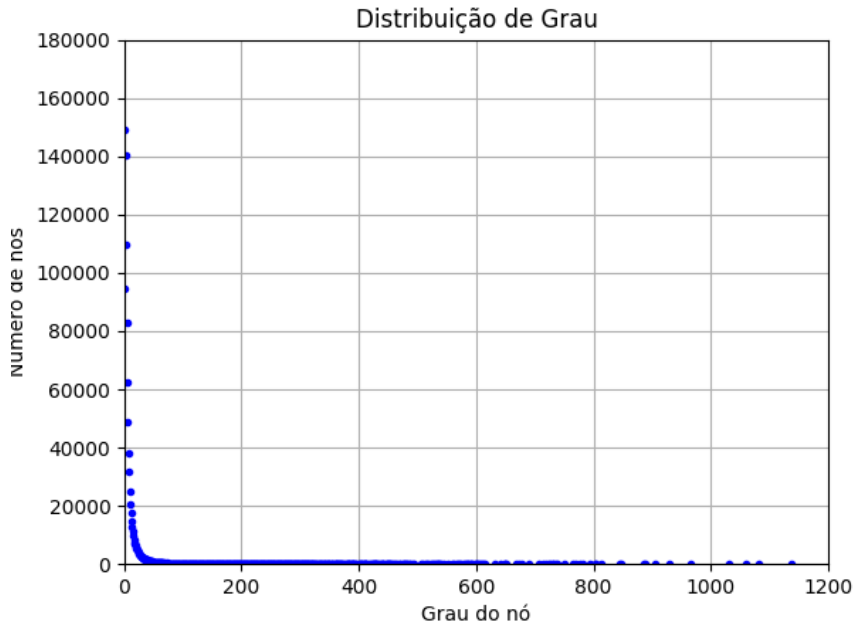


Figura 4.4: Distribuição de Grau da maior componente conexa da DBLP

O resultado obtido está de acordo com o esperado, trata-se de uma distribuição que se assemelha à distribuição de toda a rede mostrado na Figura 4.1. Isso faz sentido pois essa rede é uma sub-rede da rede de distribuição de grau total e o resultado não invalida a característica dominante apresentada anteriormente.

Na Tabela: 4.3, pode-se verificar os pesquisadores que possuem os maiores graus na rede, ou seja, levando em consideração a centralidade de grau, esses são os pesquisadores mais centrais da rede.

Tabela 4.3: Distribuição de Grau dos 40 Maiores Pesquisadores

Pesquisadores	Grau Entrada/Saída	Pesquisadores	Grau Entrada/Saída
Wei Zhang	1138	Yang Li	738
Wei Li	1082	Min Chen	736
Jun Wang	1059	Xin Li	732
Wei Wang	1031	Bo Zhang	730
Yang Liu	965	Yan Li	727
Yu Zhang	930	Jun Li	723
Jing Li	906	Hui Li	720
Wei Chen	888	Jun Zhang	715
Li Zhang	885	Tao Wang	707
Athanasios V. V.	846	Yan Zhang	690
Ying Zhang	844	Xi Chen	690
Lei Wang	814	Li Li	677
Jing Wang	804	Yan Wang	672
Yong Zhang	803	Bin Li	669
Jian Zhang	794	Ying Li	651
Yi Wang	781	Jian Li	649
Yong Wang	776	Jing Zhang	648
Lei Zhang	769	Ying Wang	642
Jian Wang	765	Tao Li	633
Xin Wang	751	Yong Liu	615

4.5 Cálculo da influência para cada pesquisador

De acordo com a Figura 4.2, nota-se que um valor de influência pode ser atribuído para cada pesquisador da rede. Para isso foi considerado que a influência (I_A) de um pesquisador A, é igual ao somatório dos valores de cada aresta incidente no pesquisador A dividido pelo número e_{ij} de arestas incidentes em A, conforme mostrado na Equação 4.2.

$$I_A = \frac{\sum A_{\hookleftarrow e_{ij}}}{|A_{\hookleftarrow e_{ij}}|} \tag{4.2}$$

Na rede apresentada pelo exemplo da Figura 4.5, tem-se o valor da influência que foi calculado para cada pesquisador pertencente a maior componente conexa.

Figura 4.5: Distribuição de Influência em cada pesquisador

4.6 Distribuição de influência e suas análises

O valor de influência atribuído para cada pesquisador apresentado pela Equação 4.2 varia no intervalo de 0 a 1. Com isso, foi calculado a distribuição de grau da rede, levando em consideração o valor de influência anteriormente calculado.

O resultado é apresentado conforme mostra a Figura 4.6.

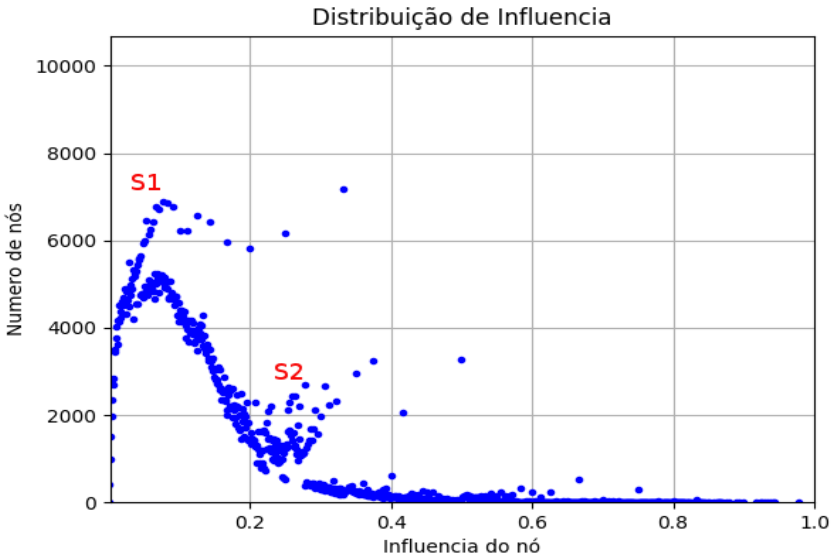


Figura 4.6: Distribuição de Influência na maior componente conexa

Analisando a Figura 4.6 nota-se alguns aspectos interessantes que merecem ser explorados.

Claramente consegue-se ver que existem alguns aspectos que caracterizam essa distribuição. A maioria dos pesquisadores que pertencem a rede tendem a seguir a distribuição de uma curva parabólica que chega a um máximo quando possui um valor de influência aproximadamente igual a 0,1, porém existem dois conjuntos que não seguem o padrão dessa distribuição.

Primeiramente identificamos esses dois conjuntos de pontos, chamamos de S1 e S2. Ambos possuem padrões diferente da distribuição dominante. Para descobrirmos o porquê desse comportamento é interessante que seja identificado alguns valores desses conjuntos.

Para melhorar a visualização dos valores de influência de S1, o conjunto S1 foi ampliado conforme mostra a Figura 4.7. Após fazer a identificação dos valores de influência

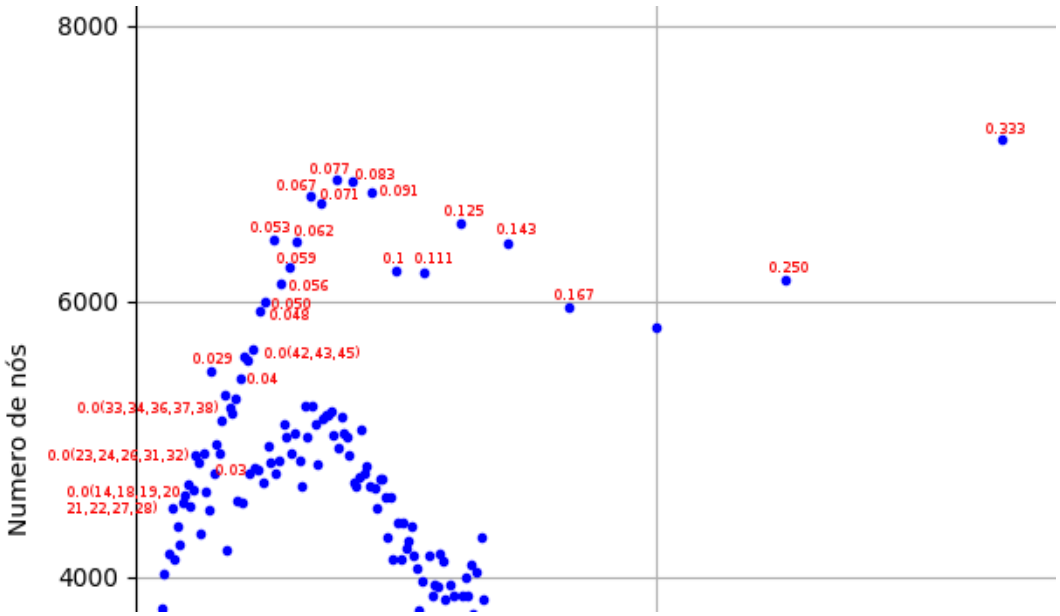


Figura 4.7: Identificação de valores de Influência do Conjunto S1.

de cada pesquisador, estamos aptos a analisar as iterações entre esses pesquisadores, para descobrirmos o porquê eles possuem tal comportamento.

As análise foram iniciadas por pesquisadores que possuem o valor de influência de 0.250. Acredita-se que esse valor possua uma característica dominante por estar devidamente afastado dos demais.

Conforme apresentado pela Figura 4.8, a apresentação dos pesquisadores foi dividida em três faixas P1, P2 e P3 onde cada faixa separa a profundidade de cada pesquisador para com o próximo pesquisador.

Os pesquisadores da faixa P1, são os pesquisadores que possuem a influência de 0.250. Esses pesquisadores se conectam com outros pesquisadores da faixa P2. Analisando os pesquisadores da faixa P2, consegue-se visualizar que esses pesquisadores formam um grupo fechado com os pesquisadores da faixa P3.

Na Figura 4.8 somente os pesquisadores das faixas P1 e P2 foram explorados, isso quer dizer que a rede continua se conectando pelos pesquisadores da faixa P3.

Na Figura 4.9 é possível visualizar os pesquisadores das três faixas com suas devidas conexões de influência. É possível confirmar também, a existência dessa mesma característica apresentada na Figura 4.8.

Ao analisarmos vários pesquisadores no Neo4j com os valores de influência identificados na distribuição da Figura 4.7, foi confirmado que em todos os casos observados,

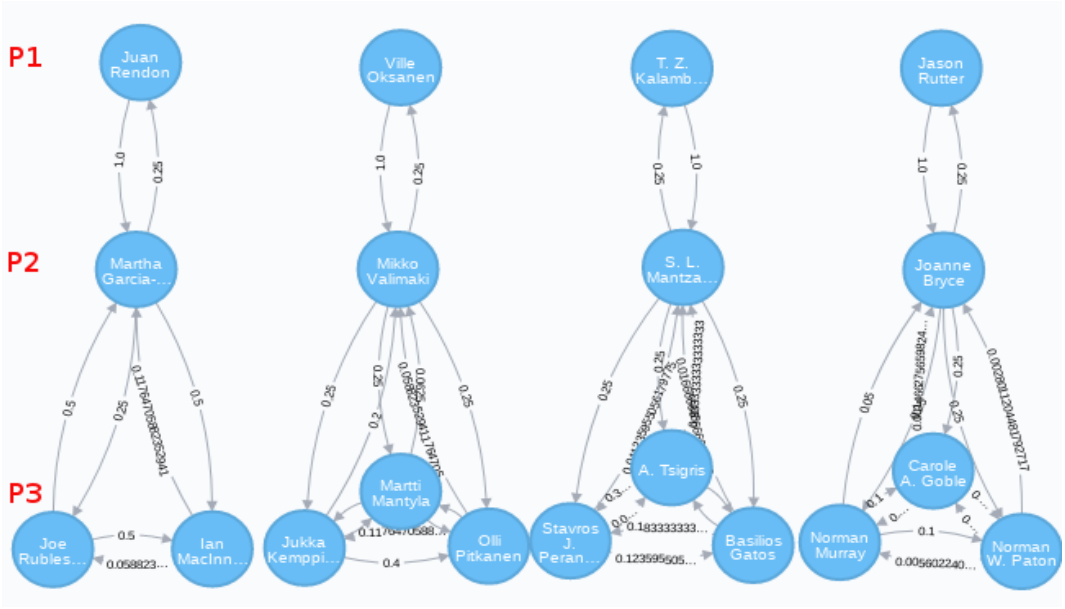


Figura 4.8: Pesquisador com influência de 0.250 pertencente ao grupo S1 e suas conexões de influência.

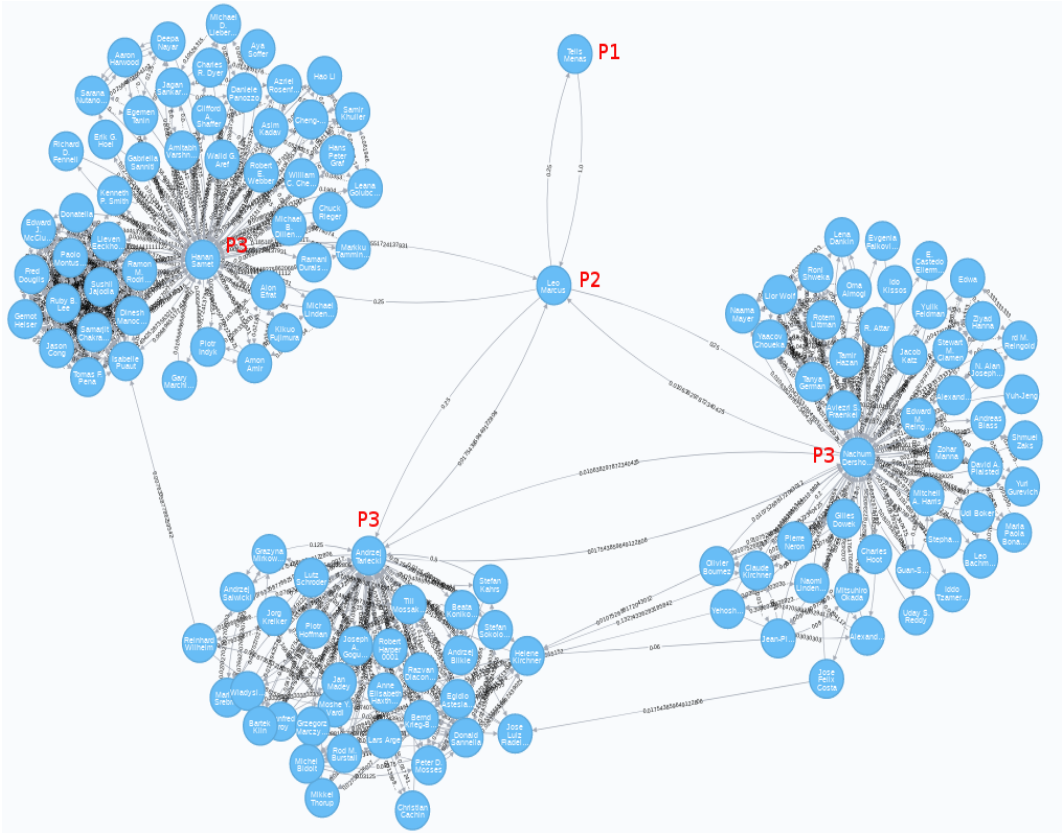


Figura 4.9: Pesquisador com influência de 0.250 pertencente ao grupo S1 e suas conexões de influência.

esses vértices da faixa P1 possuíam grau de chegada igual 1 e se conectavam com pesquisadores que pertenciam a um grupo fortemente conectado.

Analisando esse termo em seu contexto, pode-se dizer que os pesquisadores que

compõem o grupo S1, são pesquisadores que em algum momento publicaram artigos em coautoria com no máximo uma pessoa. Esses pesquisadores podem também ter publicados artigos sozinhos, contanto que eles tenham uma pessoa com quem já tenham publicado algum trabalho.

Além disso o pesquisador pertencente a profundidade P2, é um pesquisador pertencente a um grupo de pesquisadores conectados entre si.

Na Figura 4.10 pode-se conferir um exemplo real de um pesquisador que possui influência 0.333 e que segue esse padrão.

Nota-se que nessa sub-rede, existe um pesquisador que está exatamente entre dois grupos cíclicos *"Byung kwan Lee"*. Esse pesquisador que está exatamente nessa posição exerce um papel fundamental para com a conectividade entre esses outros pesquisadores dos outros grupos. Além dele pertencer a dois grupos de publicações distintas, ele influencia em 100% as pesquisas do pesquisador "Eun-Hee Jeong".

Essa informação pode ser confirmada ao consultarmos esse pesquisador folha no banco. *"Eun-Hee Jeong"* possui exatamente e unicamente dois artigos escritos com *"Byung kwan Lee"* intituladas *"An IP Traceback Protocol using a Compressed Hash Table, a Sinkhole Router and Data Mining based on Network Forensics against Network Attacks."* e *"A Black Hole Detection Protocol Design based on a Mutual Authentication Scheme on VANET."*

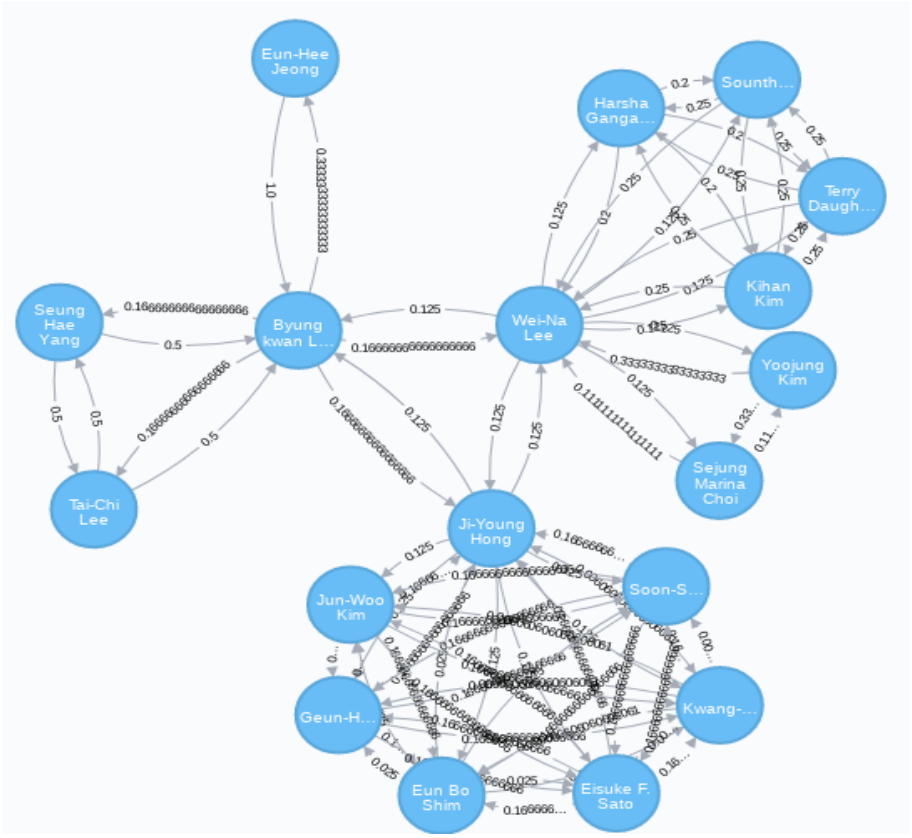


Figura 4.10: Pesquisador com influência de 0.333 pertencente ao conjunto S1 e suas conexões de influência.

Para confirmar essa premissa, foi identificado o número modular desses pesquisadores no banco. Foi constatado 94744 de pesquisadores e posteriormente foram removidos no cálculo da distribuição de influência.

Após essa remoção foi plotado, novamente, a distribuição de grau levando em consideração a influência dos pesquisadores. O resultado é apresentado pela Figura 4.11. Ele está conforme o esperado, caracterizando assim os pesquisadores que formavam o conjunto S1.

O próximo passo a ser executado é entender o comportamento dos pontos em S2. Para isso foi feitor o mesmo processo executado para a resolução do problema do conjunto S1.

Alguns dos valores de influência que compõem essa distribuição são apresentados na Figura 4.12.

Inicialmente foi analisado um ponto em específico, pertencente a S2. O valor de 0.375 foi escolhido por se tratar de um ponto que possui um valor dominante em comparação aos pontos mais nebulosos pertencentes a S2.

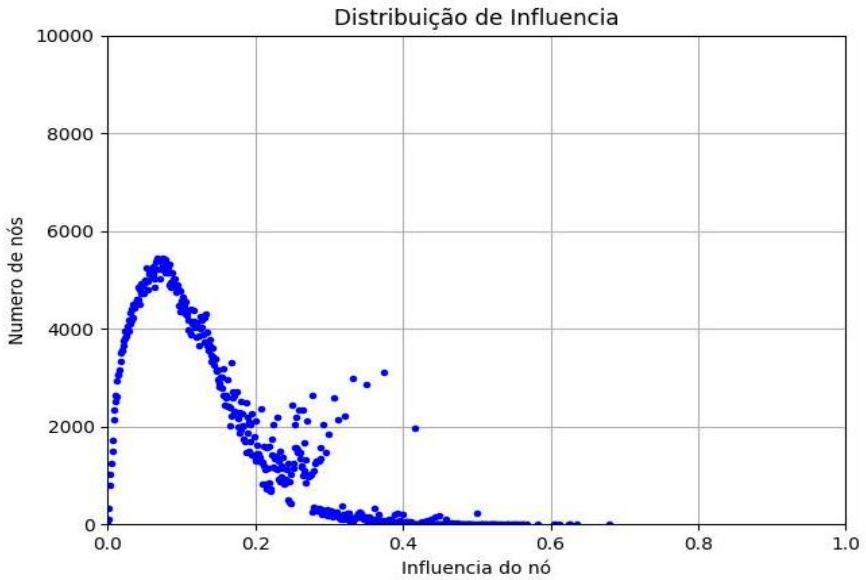


Figura 4.11: Distribuição de influência dos pesquisadores que possuem grau de chegada maior que 1

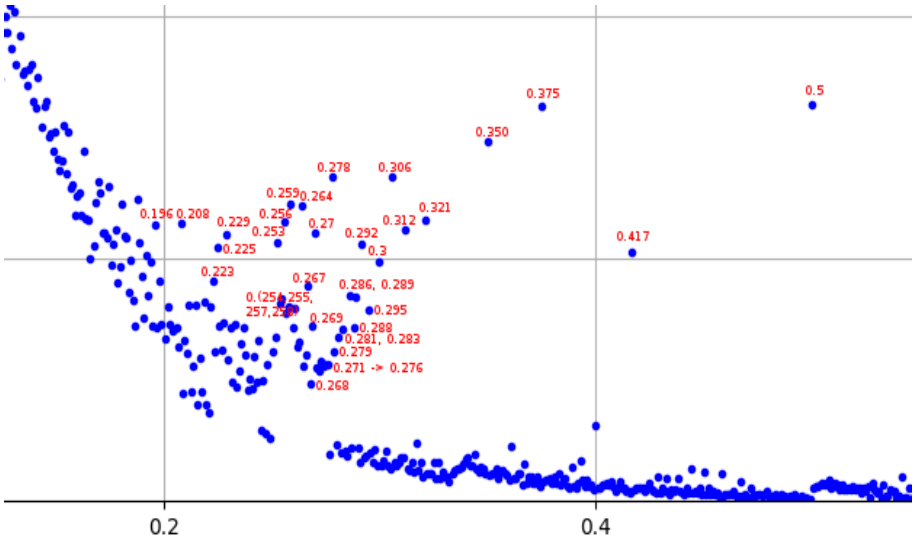
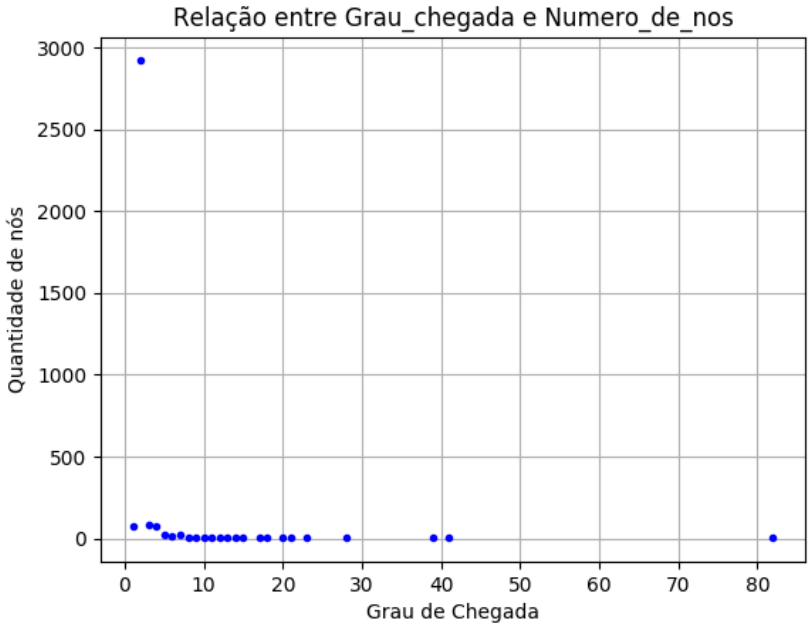


Figura 4.12: Identificação de valores de Influência do conjunto S2.

Torna-se interessante saber como os pontos que caracterizam a influência de 0.375 se comportam em relação a sua distribuição de grau. Para tal, foi feito um gráfico que faz a distribuição de grau de todos os nós que possuem a influência igual a 0.375 conforme apresentado pela Figura 4.13



Grau_Chegada	Qtd.
2	2922
3	80
4	77
1	70
5	25
7	18
6	16
8	7
10	6
11	4
20	4
9	4
14	3
13	3
18	3
12	2
15	2
23	1
41	1
17	1
82	1
28	1
21	1
39	1

Figura 4.13: Distribuição de grau dos pesquisadores com influência de 0.375

Como pode-se observar na distribuição apresentada pela Figura 4.13, a maioria dos pesquisadores que possuem a influência de 0.375 possuem também grau de chegada igual a 2.

O grau de chegada ser igual a 2, significa que o pesquisador influencia outros 2 pesquisadores. Há indícios que o grau de chegada ser igual a 2, é a característica responsável pela localização do valor de influência 0.375 no conjunto S2.

A característica mostrada acima está presente para alguns outros valores analisados que compõem a distribuição de influência caracterizada pelo conjunto S2 conforme mostra a Tabela 4.4.

Tabela 4.4: Porcentagem de pesquisadores com grau de chegada 2

Influência	Total de pesquisadores	Quantidade de pesquisadores com grau de entrada igual a 2	Porcentual
0.417	2050	1734	84,59 %
0.375	3253	2922	89,82 %
0.350	2969	2702	91,00 %
0.321	2316	2065	89,16 %
0.312	2233	1926	86,25 %
0.306	2678	1584	59,14 %
0.300	1980	1490	75,25 %
0.295	1573	1267	80,55 %
0.292	2125	1658	78,02 %
0.288	1435	1096	76,38 %

Mas o fato de ele possuir grau de chegada igual a 2 não é o suficiente. Analisando individualmente esses pesquisadores, notamos que eles se influenciam em 50% de um para com o outro e ambos dependem de um terceiro pesquisador também em 50%, formando assim um grupo fechado de pesquisadores que se relacionam conforme apresentado na Figura 4.14.

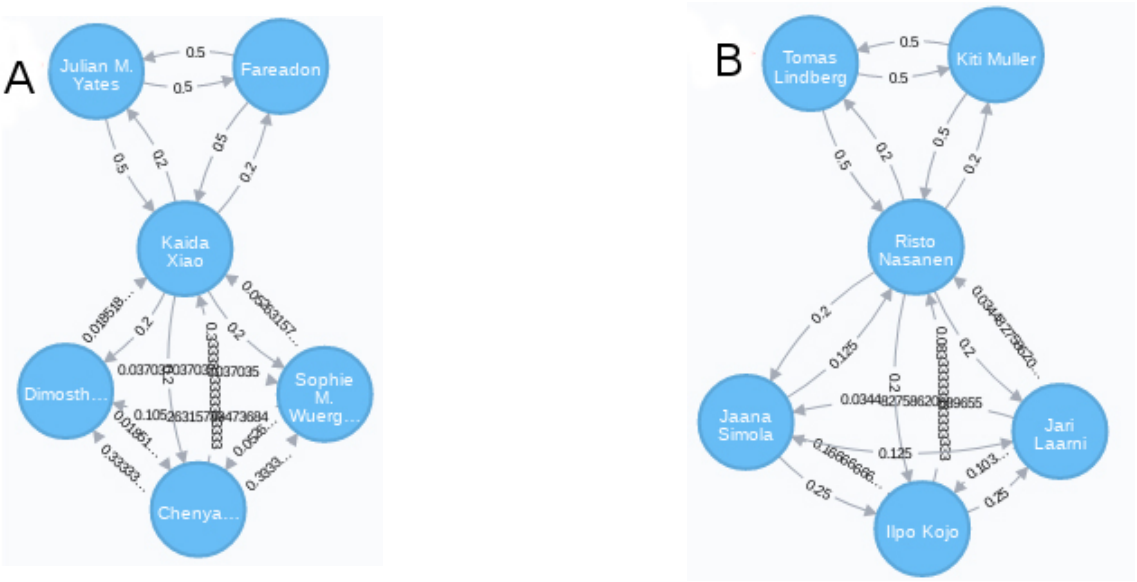


Figura 4.14: Pesquisadores com influência de 0.375

Os pesquisadores que estão na direção das letras A, B, C e D Figuras 4.14, 4.15 são pesquisadores pertencentes a S2 porém com valores de influência diferente. Esses pesquisadores se influenciam ambos em 50%, e também dependem em 50% de um terceiro pesquisador.

Finalmente foi efetuado um teste na rede para garantir a veracidade da afirmação

Figura 4.15: Pesquisadores com influência de 0.417

acima.

Após analisar o comportamento desses pesquisadores, foram removidos da rede de influência todos os sub grupos de pesquisadores que se influenciavam entre si. Os grupos removidos variavam de um a cinco pesquisadores uniformemente conectados, pois devido as análises feitas acredita-se que esse é um número bem recorrente entre grupos fechados na rede.

Posteriormente foi feita a distribuição de influência para essas redes em cada um dos casos. Ao final da distribuição é possível analisar a curva que rege essas novas distribuições da rede. Os resultados são mostrados pelas Figuras 4.16, 4.17, 4.18, 4.19.

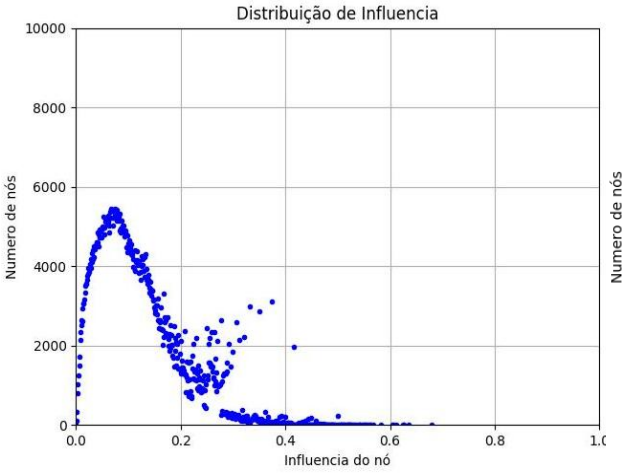


Figura 4.16: Exclusão de pesquisadores de grau 1

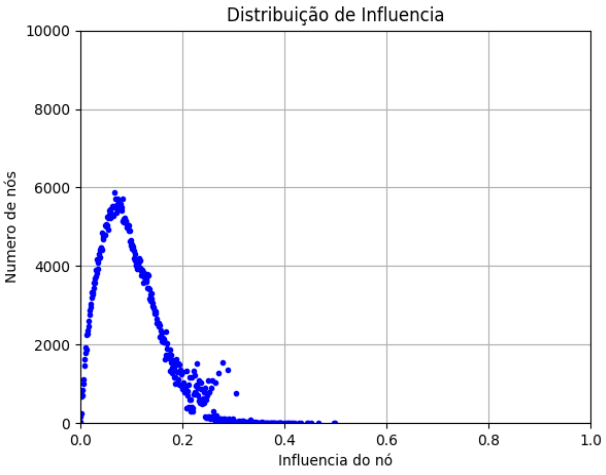


Figura 4.17: Exclusão de pesquisadores de grau 2

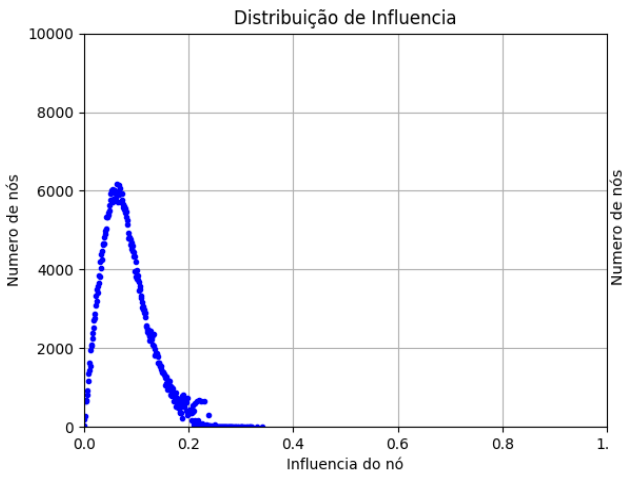


Figura 4.18: Exclusão de pesquisadores de grau 3

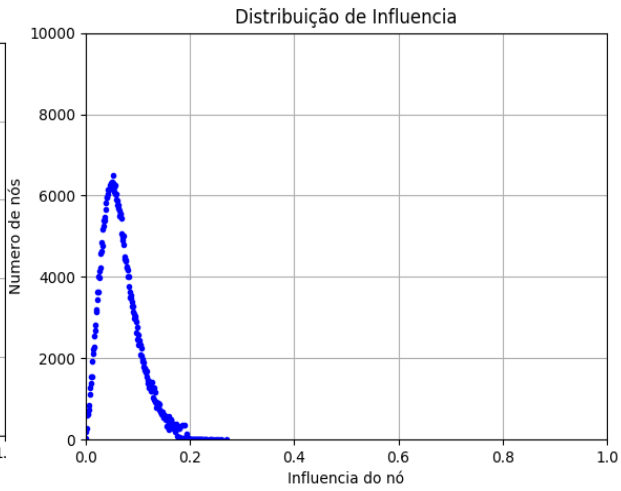


Figura 4.19: Exclusão de pesquisadores de grau 4

Como podemos ver, a medida que removemos os grupos de pesquisadores do banco, fomos conseguindo uma distribuição mais limpa e definida.

O resultado apresentado pelas Figuras 4.16, 4.17, 4.18, 4.19 apontam para a veracidade da hipótese acima, inferindo que o conjunto S2 possivelmente é composto por pares de pesquisadores que se influenciam em 50% de um para com o outro e ambos dependem de um terceiro pesquisador em 50%, formando assim um grupo fechado de pesquisadores que se relacionam. Esses pesquisadores se influenciam em mesma proporção e são dependentes, também em mesma proporção, de um terceiro pesquisador.

4.7 Considerações Finais do Capítulo

Neste capítulo, foi apresentado uma modelagem de uma rede que utilizou uma métrica de influência apresentada por (STROELE et al., 2017). Após sua criação, foi apresentado a distribuição de influência dos pesquisadores que compunham a rede. A distribuição de influência apresentou algumas disfunções que não seguiam o comportamento padrão da maioria dos pesquisadores que pertenciam a rede.

Para entender o porquê do resultado apresentado pela distribuição de influência, primeiramente, foi feito uma análise topológica da rede.

Analisando essa topologia, constatou-se que a rede de influência calculada, não era uma rede livre de escala. Apesar de não ser uma SFN, ela apresentava algumas

características que se pareciam com as SFN. foi avaliado que, apesar da distribuição de grau não satisfazer a lei de potência para o todo, ela era satisfeita para um certo intervalo.

Foi desenvolvido um algoritmo para que identificássemos o número total de componentes que existiam na rede. Após a execução, foram encontrados 99986 componentes. Através desse algoritmo foi possível detectar a maior componente conexa.

Dentre todas as componentes que foram encontradas, nota-se um certo padrão entre o número de pesquisadores pertencente a componente e o número de componentes com a quantidade de pesquisadores. Nesse processo foi mostrado que existe uma componente conexa dominante que possui 985672 pesquisadores conectados através de 8.824.668 arestas.

Posteriormente, foi feita a distribuição de grau da maior componente conexa. foi verificado que, falando-se em número de conexões, poucos possuíam muito e muitos possuíam pouco (MERTON et al., 1968).

Logo depois foi calculado a influência bruta de cada pesquisador, onde foi atribuído um valor fazendo uma média entre a soma dos valores de influência sobre o total de relações.

Posteriormente, foi feito uma distribuição de grau baseando-se nesse novo valor de influência e constatou o aparecimento de dois conjuntos que não seguiam o comportamento de distribuição da curva dominante.

Foi concluído, que os valores pertencentes aos conjuntos eram valores referentes a características topológicas da rede.

Tudo indica, que o primeiro conjunto possuía tal característica porque os pesquisadores que a compunham, fizeram publicações em coautoria com no máximo mais um pesquisador. Isso o caracterizava como uma folha na rede, e normalmente o outro pesquisador com quem ele se conectava, era integrante de um grupo de pesquisadores conectados entre si.

Para confirmar esse ponto, foram removidos todos os pesquisadores que satisfaziam essa característica, e foi feito mais uma vez a distribuição de grau, onde tal conjunto deixou de existir.

Para o conjunto S2, foi feito um processo parecido com o anterior mas, possivel-

mente, a característica que causava a aparição do conjunto S2, era que sempre existiam grupos de pesquisadores conectados entre si que possuíam grau dois. Esses pesquisadores também se influenciavam em mesma proporção e dependiam também em mesma proporção de um terceiro componente do grupo que ele pertencia.

Identificado essa característica, foram feitos as remoções desses grupos gradativamente, apresentado as distribuições da influencia para cada caso. Como mostrado, o segundo conjunto S2 também deixou de existir, caracterizando assim os valores desse conjunto.

5 Análise de Centralidade

Nesse capítulo será apresentado uma representação gráfica da rede de influência entre os pesquisadores, além de algumas análises de centralidade que foram feitas. Dentre essas análises, é possível identificar a localização dos pesquisadores que possuem os cem maiores Graus e *Closeness* da rede.

5.1 Representação Gráfica

Para fazermos algumas análises de centralidade, torna-se interessante fazer uma representação gráfica da rede. Essa representação foi feita utilizando um software chamado *Gephi*. Gephi é um software *open source* que foi desenvolvido por estudantes da *University of Technology of Compiègne - França*. Ele é muito utilizado por cientistas para representações que demandam análises, além de ser utilizado para vários projetos de pesquisa científica, pois da suporte na execução de cálculos estatísticos além de representar muito bem a rede desejada.

Alguns parâmetros foram definidos para ser possível montar a representação da rede. Como existem muito mais arestas na rede do que nós *(8824668 Arestas e 985672 Nós)* normalmente fica difícil fazer uma representação devido ao volume desproporcional de arestas para com vértices.

Pensando nessa desproporção de valores, os valores de influência que variavam de 0 a 0,0977 foram representados pela cor azul. Esse número foi definido pois 0,0977 é exatamente 10% do valor da maior influência que um pesquisador da rede, "John R. Pugh", possuí. Como definido anteriormente, todos os pesquisadores que possuem um valor de influência que esteja nesse intervalo, será representado pela cor azul.

Foram constatados que existiam 461.661 pesquisadores que pertencem a este intervalo, isso equivale a 46,84% de todos os nós pertencentes a rede.

Como podemos perceber, a porcentagem de pesquisadores que possuem valor de influência menor ou igual a 10% é muito alta e isso é exatamente normal nesse tipo de

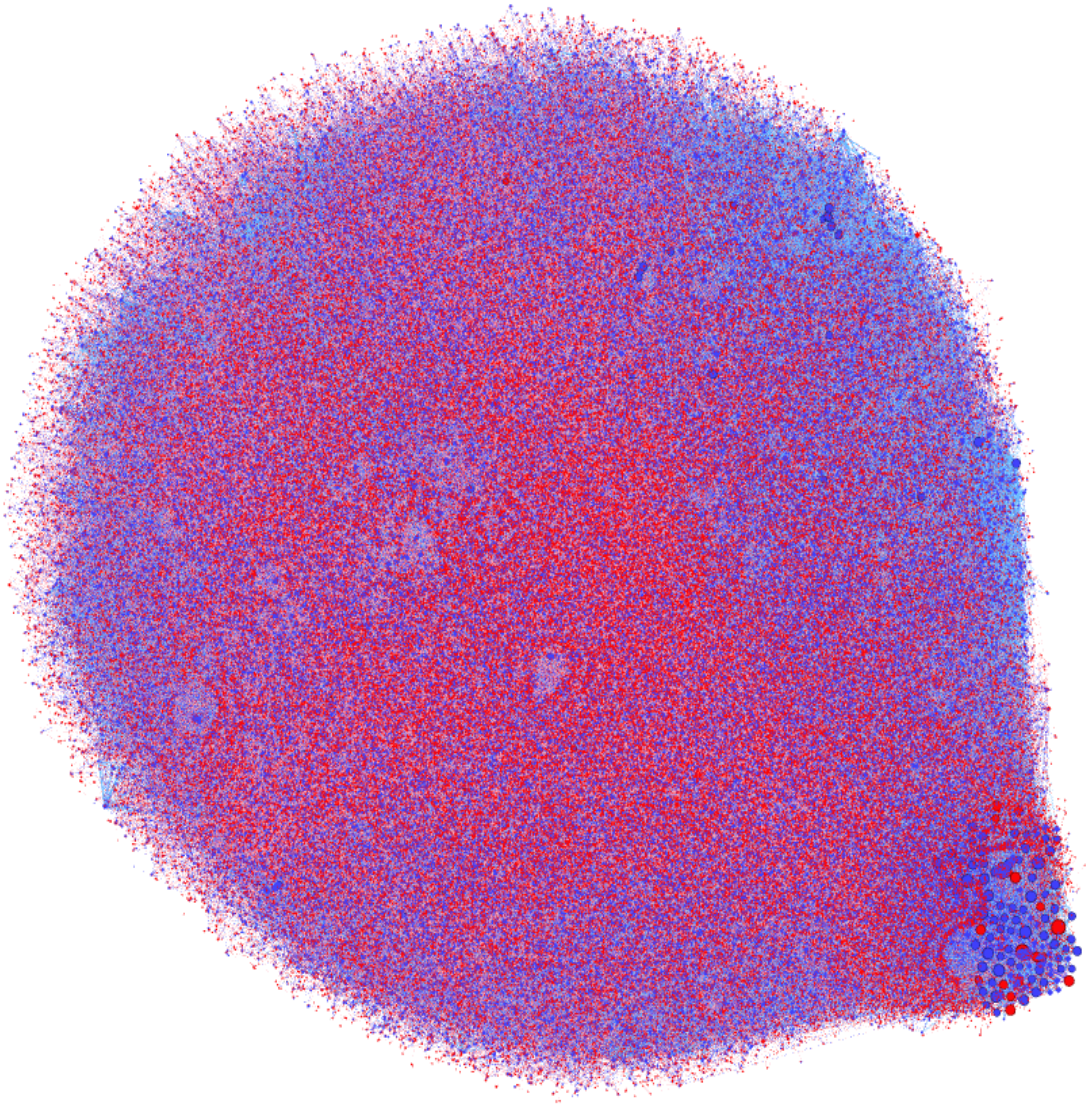


Figura 5.1: Rede de influência representada

rede.

Consequentemente, os pesquisadores pertencentes aos outros 53,16% da rede foram pesquisadores que possuíam a influência entre 0,0977 e 0,977. Esse pesquisadores, que somam 524011, foram representados pela cor vermelha.

Montada a representação da rede, estamos aptos a fazer uma análise dos pesquisadores que possuem os maiores graus da rede. Para isso vamos analisar os pesquisadores que possuem os 100 maiores graus conforme mostra a Tabela 5.1. É possível destacar na rede esse pesquisadores apresentados na Tabela 5.1.

Analisando a Figura 5.2, conseguimos observar que esses pesquisadores que possuem os maiores *graus_entrada* da rede, também possuem um alto índice de conectividade

Tabela 5.1: Pesquisadores com os 100 maiores Grau_chegada da rede de influência

Pesquisador	Grau_Chegada	Pesquisador	Grau_Chegada
Wei Zhang	1138	Gang Wang	594
Wei Li	1082	Arthur W. Toga	592
Jun Wang	1059	Qing Li	590
Wei Wang	1031	Tao Zhang	587
Yang Liu	965	Qi Wang	586
Yu Zhang	930	Xin Liu	584
Jing Li	906	Chao Wang	581
Wei Chen	888	Xin Chen	581
Li Zhang	885	Li Wang	579
Athanasios V. Vasilakos	846	H. Vincent Poor	575
Ying Zhang	844	Wei Xu	572
Lei Wang	814	Bin Wang	569
Jing Wang	804	Fan Zhang	566
Yong Zhang	803	Xiaodong Wang	562
Jian Zhang	794	Xiang Li	560
Yi Wang	781	Wei Liu	553
Yong Wang	776	Jie Chen	544
Lei Zhang	769	Rui Zhang	539
Jian Wang	765	Jian Chen	537
Xin Wang	751	Jian Yang	537
Yang Li	738	Ying Liu	536
Min Chen	736	Yang Zhang	536
Xin Li	732	Fang Liu	535
Bo Zhang	730	Li Chen	532
Yan Li	727	Wei Wu	530
Jun Li	723	Hui Zhang	529
Hui Li	720	Yan Chen	526
Jun Zhang	715	Hong Liu	524
Tao Wang	707	Lei Liu	519
Yan Zhang	690	Peng Wang	517
Xi Chen	690	Paul M. Thompson	513
Li Li	677	Bo Li	509
Yan Wang	672	Qian Zhang	506
Bin Li	669	Witold Pedrycz	495
Ying Li	651	Hao Li	495
Jian Li	649	Qiang Li	492
Jing Zhang	648	Yan Liu	488
Ying Wang	642	Wei Wei	486
Tao Li	633	Ewan Birney	486
Yong Liu	615	Christopher J. Mungall	486
Jing Liu	612	Bo Wang	483
Hui Wang	612	Meng Wang	481
Jie Zhang	610	Judith A. Blake	479
Peng Li	610	Hong Zhang	478
Feng Liu	607	Qian Wang	477
Yi Zhang	605	Ping Wang	475
Yu Liu	603	Hui Liu	471
Dong Wang	601	Gang Li	470
Jie Li	600	Jie Liu	470
Yang Yang	598	Jun Yang	466

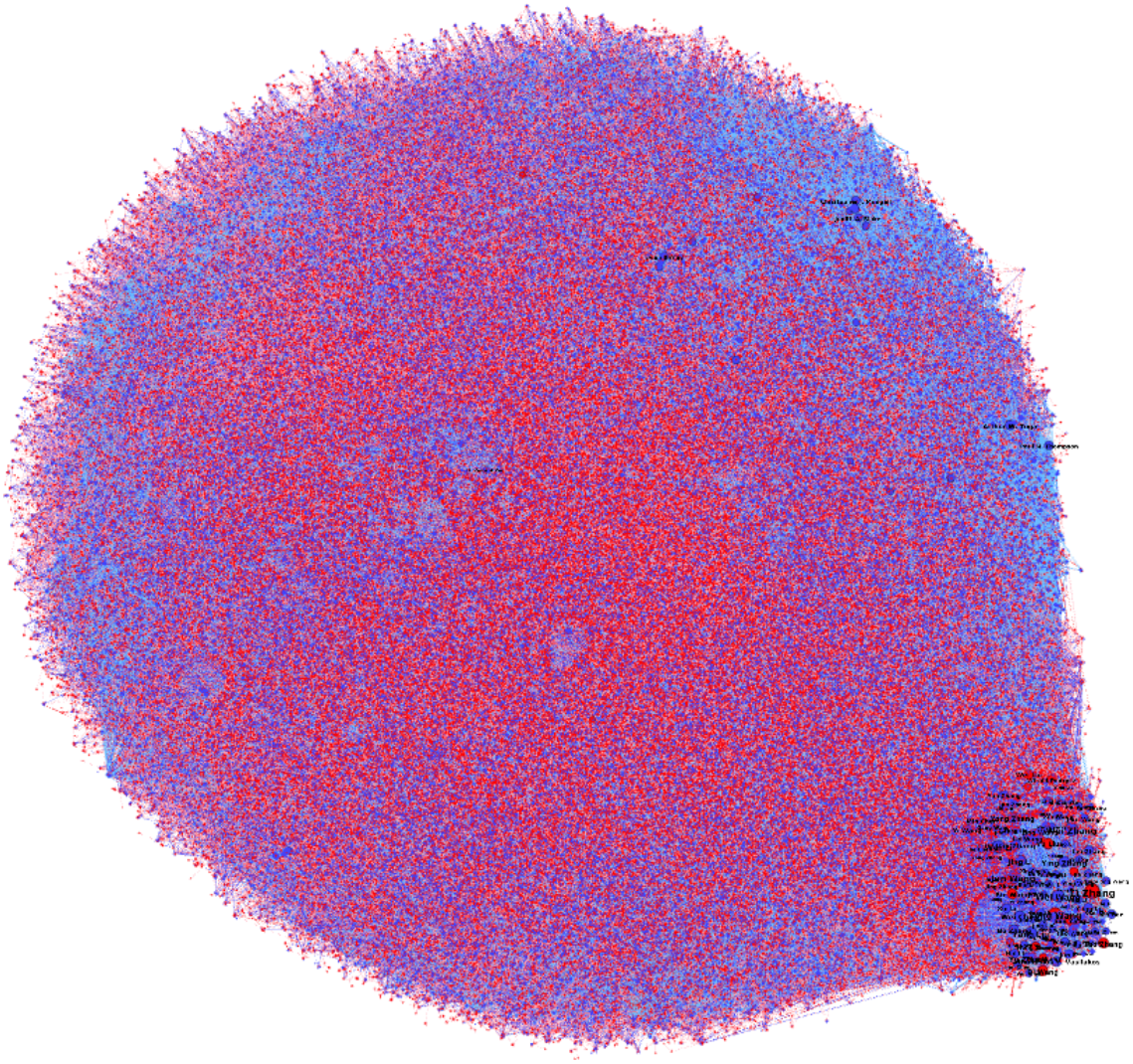


Figura 5.2: Rede com Pesquisadores identificados de acordo com 100 maiores Graus

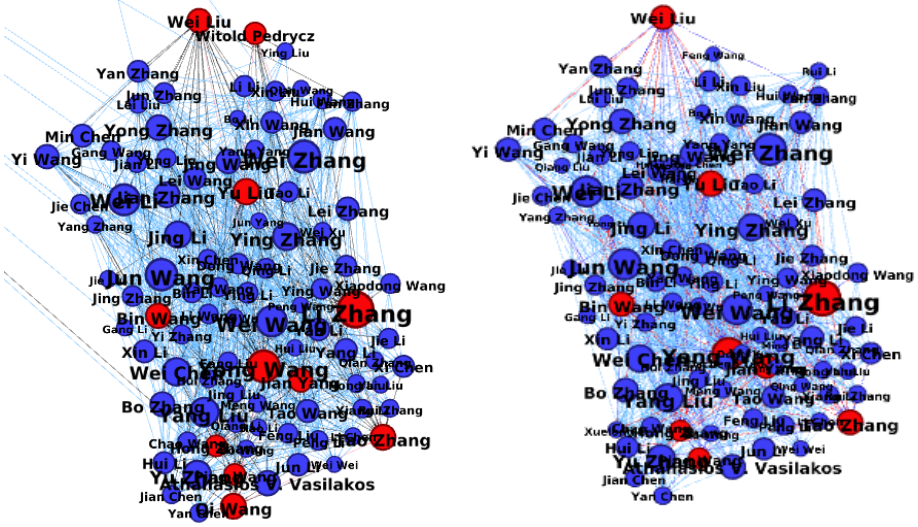


Figura 5.3: Grupo à Esquerda extraído da (Figura 5.2), Grupo à Direita extraído da (Figura 5.4)

entre si. Pela Figura 5.3 conseguimos identificar que os pesquisadores com alto grau da rede, em sua maioria, não possuem nem 10% do maior valor de influência conforme indicado pelas cores.

Podemos fazer também a mesma representação tomando em consideração o *Closeness* de cada pesquisador. Abaixo na Tabela 5.2, podemos observar os 100 pesquisadores que possuem os 100 maiores *Closeness* da rede. Fazendo o mesmo procedimento adotado anteriormente, verificamos que a maioria dos pesquisadores, mostrados no grupo mais à esquerda na Figura 5.3, fazem parte do grupo dos pesquisadores que possuem os maiores valores para o *Closeness* conforme mostra a imagem mais à direita da Figura 5.3.

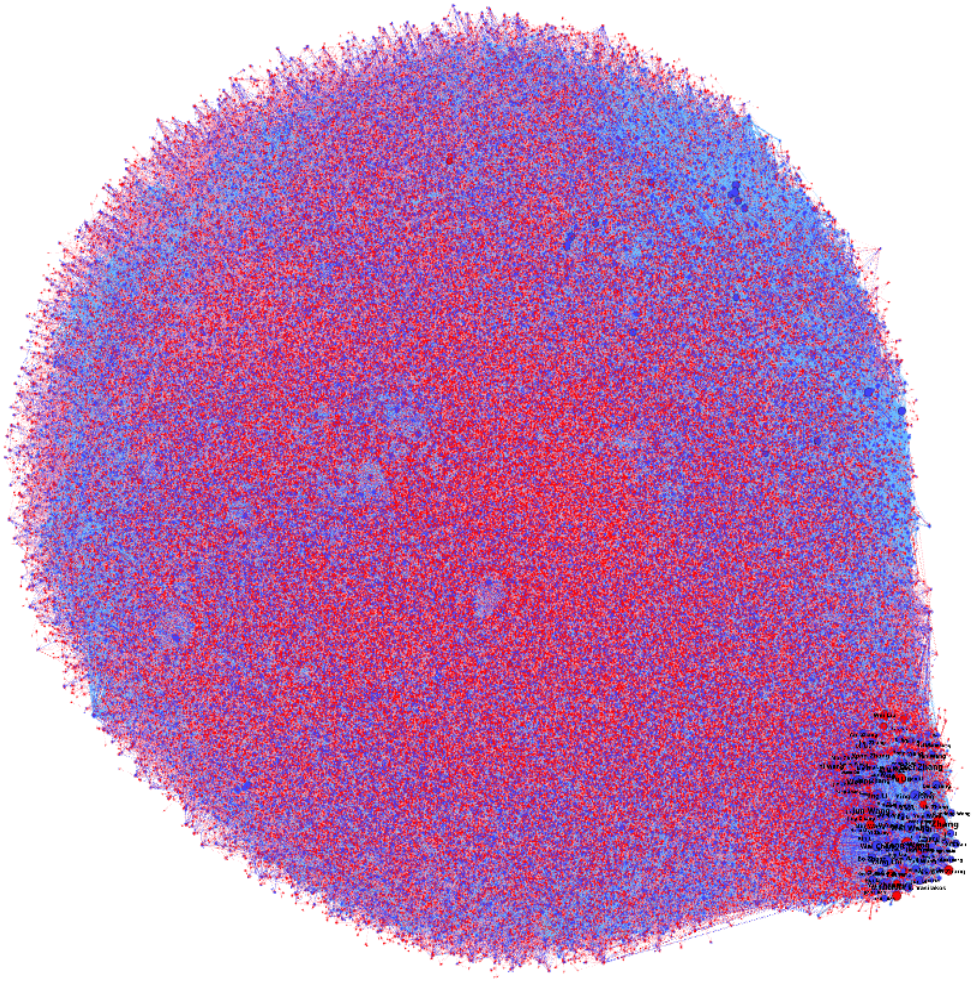


Figura 5.4: Rede com Pesquisadores identificados de acordo com 100 maiores *Closeness*

Como foram identificados os pesquisadores que estão no conjunto dos 100 maiores Graus da rede e também foram identificados os pesquisadores que estão no conjunto dos 100 maiores *Closeness* da rede, torna-se interessante fazer uma intersecção desses dois conjuntos conforme mostra a Figura 5.5

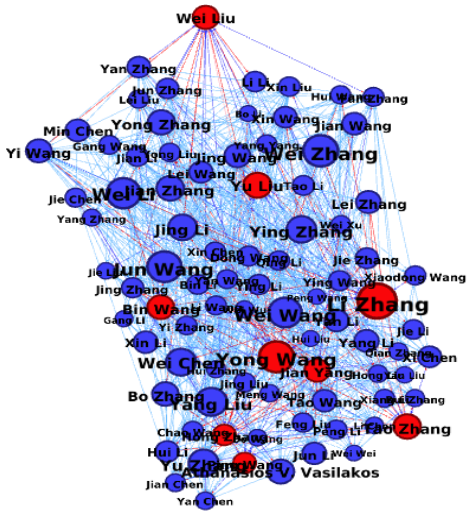


Figura 5.5: Intersecção entre os pesquisadores com 100 maiores Graus e *Closeness* da rede.

Através da centralidade de Grau e da centralidade de *Closeness*, é possível identificarmos os pesquisadores mais centrais da rede na Figura 5.5. Foram identificados 86 pesquisadores que estão no conjunto dos 100 maiores graus e 100 maiores *Closeness* da rede. Conseguimos verificar também que a maioria dos pesquisadores que compõe esse grupo, são pesquisadores que não possuem um valor de influência maior que 10% do maior valor de influência.

5.2 Considerações Finais do Capítulo

Nesse capitulo, foi apresentado uma maneira para ser feito uma representação gráfica da rede de influência entre os pesquisadores. Após fazer essa representação foram identificados os pesquisadores mais centrais na rede de acordo com as medidas de centralidade de grau e centralidade de *closeness*.

Após feita a identificação desses pesquisadores na rede, foi verificado que dos pesquisadores que possuem os 100 maiores graus da rede e dos pesquisadores que possuem os 100 maiores *closeness* da rede, 86 deles estavam presentes nesses dois conjuntos consecutivamente. Isso indica que esses pesquisadores são os pesquisadores mais centrais da rede e também da a entender que são pesquisadores que possuem um certo grupo de conexão onde possivelmente publicam trabalhos juntos.

Tabela 5.2: Closeness dos maiores 100 Pesquisadores da rede de influência

Pesquisador	*Closeness*	Pesquisador	*Closeness*
Wei Li	0.25434852786007295	Hui Wang	0.2454428509958445
Athanasios V. Vasilakos	0.2533054415885959	Yong Liu	0.24543551706072297
Wei Zhang	0.2525301151726976	Li Li	0.24539873176634028
Wei Chen	0.2518940526516737	Hui Zhang	0.24528307051579393
Wei Wang	0.2515029510428736	Yang Zhang	0.2451751410911647
Yu Zhang	0.2514617585493008	Jian Yang	0.24509375354335877
Jun Wang	0.2512449186214744	Xiang Li	0.2448710221883473
Yang Liu	0.25123332761540007	Hai Jin	0.24473975306802734
Li Zhang	0.25098735120980126	Dong Wang	0.24468841462960528
Xin Wang	0.2503910808667338	Jie Li	0.24458270709434787
Jing Li	0.2503661493584576	Qian Zhang	0.24457736646171718
Jian Zhang	0.24993920595753724	Wei Liu	0.24416048630411843
Jing Wang	0.2495606544925151	Xuelong Li	0.2439093255190309
Ying Zhang	0.24916220944650083	Jian Chen	0.24381182822695327
Lei Zhang	0.2491481017601642	Yu Liu	0.24379801839335422
Min Chen	0.2488322313488356	Bo Wang	0.2437877675635247
Jie Zhang	0.24879756086910318	Bin Wang	0.2437792057663875
Yi Wang	0.2485194165761302	Meng Wang	0.24377480451316935
Lei Wang	0.2485021236879637	Bo Li	0.24365295896142497
Jun Zhang	0.24830255260115724	Hui Liu	0.243630073788197
Xin Li	0.24828998059870452	Yang Yang	0.2435586755968376
Tao Wang	0.24825202212850198	Rui Zhang	0.24353219787048214
Bo Zhang	0.2481785145158614	Hong Liu	0.24352064575758348
Yan Wang	0.2478318163863464	Wei Xu	0.24351312543975462
Jun Li	0.24765739911125137	Wei Wu	0.2434797408372521
Yong Zhang	0.24744973913707116	Hong Zhang	0.2434146220998432
Peng Li	0.24742669420228336	Xin Chen	0.24325867126560916
Hui Li	0.24728547473083048	Yan Liu	0.24313572167701242
Qing Li	0.2471433009457515	Ming Li	0.24307360412287846
Yong Wang	0.2471288633041577	Li Wang	0.24306982772779126
Yan Zhang	0.2470730494955245	Yong Li	0.2429564706410128
Yang Li	0.2470361432255293	Li Chen	0.2429268307109756
Ying Wang	0.24695518622402002	Qing Wang	0.24279380959794114
Jian Wang	0.24687570443172085	Peng Wang	0.24270401452179477
Xin Liu	0.24673579243645058	Dong Xu	0.24268232301359402
Yan Li	0.24673468069735574	Rui Li	0.24261159858844483
Xi Chen	0.2466858975706593	Gang Li	0.24236826109364734
Jian Li	0.24660942717727355	Wei Wei	0.24232929136080616
Jing Zhang	0.24654496720307678	Hao Wu	0.2422222784701875
Tao Li	0.2462473468359895	Ping Wang	0.24217776228240032
Bin Li	0.246140903019329	Yang Wang	0.24214159005502575
Chao Wang	0.2461392434460408	Lei Liu	0.2420537627120642
Yi Zhang	0.246108883351394	Jie Chen	0.24203610979534737
Ying Li	0.2459693478788169	Xiaodong Wang	0.24202939404192877
Feng Liu	0.245956213165994	Qiang Liu	0.24188649037579033
Fan Zhang	0.24590761465340152	Jie Liu	0.241812194935263
Gang Wang	0.2458988419687355	Hsiao-Hwa Chen	0.2417483207476784
Jing Liu	0.24566778766121722	Feng Wang	0.24170794974523915
Yan Chen	0.24566637937867034	Yi Liu	0.2416569865367331
Tao Zhang	0.2456191806389476	Dacheng Tao	0.2416455523970978

6 Considerações Finais

Neste trabalho foi proposto a utilização de uma métrica (STROELE et al., 2017) para análise da iteração de influência entre pesquisadores em uma base de dados científica chamada DBLP que armazena publicações científicas desde o ano de 1968. Existem vários tipos de registros que compõem a DBLP *(Article, Inprocedings, Procedings, Book, Incollection, www)* mas nesse trabalho foram utilizados somente os objetos do tipo *artigo*.

Para isso, foi proposto a modelagem de uma rede social científica, onde os elementos *(nós)* representaram os pesquisadores e suas ligações *(arestas)* representam relacionamentos de coautoria. Feita essa modelagem, foi executado a distribuição de influência para cada pesquisador que compõe a rede.

Analisando a distribuição de influência, percebeu-se que existiam dois conjuntos, nomeados de S1 e S2, que apresentavam um comportamento diferente da maioria dos elementos da distribuição principal. Verificamos que a maioria dos pesquisadores que caracterizavam tais valore da influência, eram pesquisadores que estavam conectados a maior componente conexa através de ligações fracas.

O conjunto superior S1 possuía tal comportamento pois era caracterizado por pesquisadores que possuíam grau de entrada igual a 1 e se conectavam com pesquisadores pertencentes a grupos cíclicos. Isso quer dizer que esses pesquisadores tinham esses valores pois, eram pesquisadores que publicaram artigos somente em coautoria com um único pesquisador que pertencia a um grupo cíclico de pesquisadores.

O conjunto inferior S2 possuía tal comportamento pois era caracterizado por pesquisadores que possuíam grau de chegada igual a 2. Além disso, esses pesquisadores se influenciavam em uma mesma proporção de um para com o outro em 50%, e ambos dependiam, também em mesma proporção, de 50%, de um terceiro pesquisador, formando assim um grupo fechado de pesquisadores.

A distribuição de grau permitiu identificar alguns pesquisadores com alto nível de conectividade e também pesquisadores que publicam sozinhos. A distribuição de influência permitiu identificar grupos de pesquisadores conectados a maior componente

conexa através de ligações fracas.

Foi apresentado também que os 100 pesquisadores que possuem os maiores grau de chegada da rede, também possuem um alto índice de conectividade. Além disso foi mostrado que a maioria desses pesquisadores possuíam os 100 maiores *Closeness da rede*.

6.1 Dificuldades encontradas

O objetivo do trabalho foi, estudar as interações entre pesquisadores que pertencem a DBLP, porém algumas dificuldades foram encontradas na execução do mesmo. A não padronização dos nomes dos autores que compõem a DBLP é um dos pontos que merecem mais atenção.

Sabe-se que o pré processamento completo desses elementos para execução desse trabalho é inviável, devido aos diferentes tipos e variações de nomes de pessoas que existem no mundo. Conforme vimos em (BIRYUKOV; DONG, 2010), as dificuldades encontradas no processamento desses atributos já são estudadas por outros pesquisadores há algum tempo e mesmo assim não se tem solução perfeita.

Outro ponto que dificultou a execução desse trabalho, refere-se a quantidade de dado que foi analisado. Devido ao grande volume, todo trabalho simples se tornou mais complicado e principalmente demorado.

Por exemplo, a inserção de todos os dados no banco demorou mais de 3 semanas de processamento com o computador ligado 24 horas por dia. A execução de tarefas simples, como a identificação de todas as componentes conexas que foi feito nesse trabalho, não pode seguir os algoritmos padrões tradicionais, pois os métodos tradicionais não resolviam o problema em um tempo simples.

Devido a grande quantidade de registros que compõe a DBLP foi utilizado nesse estudo somente os artigos. A alta demanda de processamento necessário para cálculo da métrica *Betweenness* impossibilitou de ser executada.

6.2 Trabalhos Futuros

Como indicações para trabalhos futuros, seria interessante executar o cálculo de influência para todos os pesquisadores que constituem a DBLP e considerar todos os tipos de registros disponibilizados. Esse fato não foi permitido devido as limitações tecnológicas existentes. Acredita-se que a execução das técnicas de SNA em toda a DBLP poderá resultar em trabalhos muito bem fundamentados e interessantes que podem contribuir com a comunidade científica.

Um outro trabalho que pode ser executado futuramente é a aplicação das técnicas de *data mining* nessa rede social científica anteriormente criada. Certamente a detecção de comunidades científicas através das diferentes formas de clusterização incrementariam o estudo atual.

A execução da métrica de *Betweenness* é um outro cálculo que pode ser feito, ela permitiria acrescentar uma grande quantidade de conhecimento sobre como os pesquisadores se relacionam na rede.

Bibliografia

ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. *Reviews of modern physics*, APS, v. 74, n. 1, p. 47, 2002.

BARABÁSI, A.-L. The physics of the web. *Physics World*, IOP Publishing, v. 14, n. 7, p. 33, 2001.

BARABÁSI, A.-L.; BONABEAU, E. Scale-free networks. *Scientific American*, Nature Publishing Group, v. 288, n. 5, p. 50–59, 2003.

BAVELAS, A. A mathematical model for group structures. *Human organization*, Society for Applied Anthropology, v. 7, n. 3, p. 16–30, 1948.

BAVELAS, A.; BARRETT, D. *An experimental approach to organizational communication*. [S.l.]: American Management Association, 1951.

BEAUCHAMP, M. A. An improved index of centrality. *Systems Research and Behavioral Science*, Wiley Online Library, v. 10, n. 2, p. 161–163, 1965.

BIRYUKOV, M.; DONG, C. Analysis of computer science communities based on dblp. *Research and advanced technology for digital libraries*, Springer, p. 228–235, 2010.

CZEPIEL, J. A. Word-of-mouth processes in the diffusion of a major technological innovation. *Journal of Marketing Research*, JSTOR, p. 172–180, 1974.

ELMACIOGLU, E.; LEE, D. On six degrees of separation in dblp-db and more. *ACM SIGMOD Record*, ACM, v. 34, n. 2, p. 33–40, 2005.

EULER, L. Leonhard euler and the königsberg bridges. *Scientific American*, v. 189, n. 1, p. 66–70, 1953.

FREEMAN, L. C. Centrality in social networks conceptual clarification. *Social Networks*, v. 1, p. 215–239, 1978/79.

LEY, M. Dblp: some lessons learned. *Proceedings of the VLDB Endowment*, VLDB Endowment, v. 2, n. 2, p. 1493–1500, 2009.

MERTON, R. K. et al. The matthew effect in science. *Science*, Washington, v. 159, n. 3810, p. 56–63, 1968.

MILGRAM, S. The small world problem. *Psychology today*, New York, v. 2, n. 1, p. 60–67, 1967.

MORENO, J. Emotions mapped by new geography. *New York Times*, v. 3, p. 17, 1933.

MOXLEY, R. L.; MOXLEY, N. F. Determining point-centrality in uncontrived social networks. *Sociometry*, JSTOR, p. 122–130, 1974.

NEWMAN, M. *Networks: an introduction*. [S.l.]: Oxford university press, 2010.

NEWMAN, M. E. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 98, n. 2, p. 404–409, 2001.

NEWMAN, M. E. The structure and function of complex networks. *SIAM review*, SIAM, v. 45, n. 2, p. 167–256, 2003.

NIEMINEN, U. On the centrality in a directed graph. *Social Science Research*, Elsevier, v. 2, n. 4, p. 371–378, 1973.

ROGERS, D. L. Sociometric analysis of interorganizational relations: Application of theory and measurement. *Rural Sociology*, Rural Sociological Society, etc., v. 39, n. 4, p. 487, 1974.

SABIDUSSI, G. The centrality index of a graph. *Psychometrika*, Springer, v. 31, n. 4, p. 581–603, 1966.

SHAW, M. E. Group structure and the behavior of individuals in small groups. *The Journal of psychology*, Taylor Francis, v. 38, n. 1, p. 139–149, 1954.

SKIENA, S. Dijkstra's algorithm. *Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica, Reading, MA: Addison-Wesley*, p. 225–227, 1990.

STROELE, V. et al. Redes sociais científicas: análise topológica da influência dos pesquisadores. In: *SBBD Technical Session 5: Data Analytics (B3)*. [S.l.: s.n.], 2017.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world'networks. *nature*, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998.