UMA NOVA CENTRALIDADE PARA REDES MULTIPLEX NÃO DIRECIONADAS

BRUNO FIGUEIRÊDO

UMA NOVA CENTRALIDADE PARA REDES MULTIPLEX NÃO DIRECIONADAS

Tese apresentada ao Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas como requisito parcial para a obtenção do grau de Doutor em Informática.

Orientador: Eduardo Freire Nakamura
Coorientadora: Fabíola Guerra Nakamura

Manaus
Outubro de 2021

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Figueirêdo, Bruno César Barreto de
F475n Uma nova centralidade para redes multiplex não direcionadas / Bruno César Barreto de Figueirêdo . 2021
104 f.: il. color; 31 cm.

Orientador: Eduardo Freire Nakamura
Coorientadora: Fabíola Guerra Nakamura
Tese (Doutorado em Informática) - Universidade Federal do Amazonas.

1. Redes complexas. 2. Redes multiplex. 3. Medidas de centralidade. 4. Detecção de fraude. 5. Detecção de nós relevantes. I. Nakamura, Eduardo Freire. II. Universidade Federal do Amazonas III. Título

FOLHA DE APROVAÇÃO

"UMA NOVA CENTRALIDADE PARA REDES MULTIPLEX NÃO DIRECIONADAS"

BRUNO CÉSAR BARRETO DE FIGUEIREDO

Tese de Doutorado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Eduardo Freire Nakamura - PRESIDENTE

Prof. Carlos Maurício Serodio Figueiredo - MEMBRO INTERNO

Prof. Tiago Eugenio de Melo - MEMBRO EXTERNO

Prof. Leandro Nelinho Balico - MEMBRO EXTERNO

Dr. Vilar Fiuza da Camara Neto - MEMBRO EXTERNO

Manaus, 08 de Outubro de 2021

A minha amada esposa Gardenya pela INESTIMÁVEL ajuda em todas as etapas desta tese. Estando sempre presente e não me deixando desanimar em um só instante me fazendo entender o verdadeiro significado da palavra amor.

Agradecimentos

- a Deus por me permitir existir e me propocionar uma vida que vale à pena ser vivida;
- Aos meus amados pais Carmelo e Sandra, por tudo o que sou;
- Aos meus filhos tão amados Bruno, Bruna e Pedro por todo amor que nos une. Um agradecimento especial a Pedro pela ajuda na revisão do inglês de artigos submetidos e pelo incentivo nas horas difíceis;
- Ao meu querido Tio Edson, pela grande ajuda na formalização matemática da centralidade e pelo amor de pai e filho que sempre nos une;
- A toda a minha família: ao meu querido irmão Lela; meus avós e tios Mimo, Bernadeth, Edson, Paulo, Dada e Marina; a Ellen e Lara; a meus sobrinhos queridos; cunhados, cunhadas e sogros; por todo o suporte, amor e amizade;
- Aos meus orientadores Eduardo e Fabíola Nakamura, por todo conhecimento repassado, incentivo nas horas difíceis e por me proporcionar a possibilidade de realizar o sonho de infância de morar fora do Brasil;
- A professora Dilma da Silva, por me recepcionar na Texas A&M University e me possibilitar lecionar naquela instituição;
- A Maria Carvalho e Thais Almeida pela cessão das bases de dados utilizadas no estudo de caso da coleção de livros de Harry Potter (Carvalho, 2017) e da Lava Jato (Almeida et al., 2017);
- Aos professores do programa PPGI do ICOMP da UFAM, em especial ao professor Feitosa, pela dedicação, responsabilidade e acolhimento;
- À Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM) pelo financiamento de um dos nossos artigos (De Figueirêdo et al., 2021).

"Inútil querer me classificar: eu simplesmente escapulo não deixando"
(Clarice Lispector)

Resumo

Uma questão desafiadora na ciência da informação, em sistemas biológicos e muitos outros campos de pesquisa, é determinar os agentes mais relevantes, ou centrais, em uma rede ou um grafo. Essas redes geralmente descrevem cenários usando nós (objetos) e arestas (as relações entre os objetos). As chamadas medidas de centralidade visam resolver este tipo de desafio, classificando os nós pela sua suposta relevância e elegendo os nós mais relevantes. Esse problema se torna mais desafiador quando uma única rede não é suficiente para representar todo o cenário. Nesses casos, pode-se trabalhar com redes multiplex caracterizadas por um conjunto de camadas de rede, cada uma descrevendo inter-relações que podem mudar dependendo de fatores externos, por exemplo, o tempo. Esta tese propõe uma nova medida de centralidade, a Centralidade Baseada em Grupos para redes multiplex não direcionadas, que tem como objetivo encontrar, de forma eficiente, os nós mais relevantes em uma rede multiplex não direcionada. Utiliza-se três estudos de caso para descrever o uso da centralidade: uma investigação de corrupção brasileira conhecida como "Operação Lava Jato", o conjunto de livros da franquia Harry Potter e a investigação de corrupção brasileira em licitações públicas conhecida como "Operação Licitante Fantasma". Nos três estudos de caso a centralidade proposta supera centralidades bem conhecidas, como: *Betweenness*, *Eigenvector*, *PageRank*, *Closeness* e *Weighted Degree*, e centralidades concebidas para redes multiplex como a *Multiplex PageRank* e a *Cross-Layer Degree Centrality*.

Palavras-chave: redes complexas, redes multiplex, medidas de centralidade, detecção de fraude, detecção de nós relevantes.

Abstract

One challenging issue in information science, biological systems, and many other fields is determining the most central or relevant networked systems agents. These networks usually describe scenarios using nodes (objects) and edges (the objects' relations). The so-called standard centrality measures aim to solve this kind of challenge, ranking the nodes by their supposed relevance and elect the most relevant nodes. This problem becomes more challenging when one single network is not enough to depict the whole scenario. In these cases, we can work with multiplex networks characterized by a set of network layers, each describing interrelationships that can change depending on external factors, e.g., time. This paper proposes a new centrality measure, the **Group-based Centrality for Undirected Multiplex Networks**, to find the most relevant nodes in an undirected multiplex network. We use three case study, to describe the centrality usage: a Brazilian corruption investigation known as the Car Wash Operation, the set of books of the Harry Potter franchise, and the Brazilian corruption investigation of public tenders known as the Ghost Bidder Operation. In these tree analyses, our proposed centrality outperforms well-known centrality methods such as betweenness, eigenvector, PageRank, closeness, weighted degree, and multylayer centralities like Multiplex PageRank cross-layer degree centrality.

Keywords: complex networks, multiplex networks, centrality measures, fraud detection, detection of relevant nodes.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A determinação dos atores mais relevantes em um cenário é um campo de estudo bastante vasto e conta com trabalhos em diversas áreas da computação, e.g., Inteligência Artificial (IA) e Mineração de Dados (Cunha e Bugarin, 2014; Silva e Ralha, 2010; Hu et al., 2013). Esse interesse se justifica pela abrangência e diversidade de aplicações que a tecnologia propicia, indo desde a determinação do indivíduo mais "popular" em uma rede social à descoberta de criminosos e fraudadores.

Uma das áreas do conhecimento que propicia a descoberta dos atores mais importantes num cenário é a das redes sociais. Os primeiros estudos sobre redes sociais datam do século XIX. Entretanto, a área de estudo de Análise de Redes Sociais, como conhecemos hoje, é, normalmente, atribuída ao psiquiatra Jacob Moreno já no século XX (Newman e Newman, 2010). Com o advento da computação e a consequente possibilidade de tratamento de grandes quantidades de informação relacionada às interações humanas, a pesquisa a respeito dessas redes tem crescido continuamente. Os campos do conhecimento interessados são inúmeros, indo desde as ciências sociais e humanidades, da física à ciência da computação; apenas para citar alguns (Fabbri, 2017).

Na Ciência da Computação, uma possibilidade de representação das interações sociais é o uso das chamadas redes complexas. Por meio dessas redes é possível revelar grupos de características das interações humanas (Ball e Newman, 2012). A dinâmica de redes complexas foi estudada em diversos documentos científicos, sendo uma área em constante expansão e com crescente interesse com relação à pesquisa científica (Newman e Newman, 2010).

Uma das conceituações comumente encontradas para as redes complexas as definem como um "grafo com características topológicas não triviais" (Newman e Newman, 2010). A conceituação de trivialidade, nesse caso, diz respeito à capacidade de modelagem de cenários por meio dessas redes, onde os nós representem as entidades e as

arestas as relações entre elas. Segundo Boccaletti et al. (2006), trata-se de um grafo com estrutura topológica irregular, não trivial e que evolui dinamicamente ao longo do tempo. Fabbri (2017) acrescenta a esse conceito a característica de ser esse grafo obrigatoriamente de grandes proporções, com milhares ou milhões de nós, e ter a função de representar um sistema encontrado em observações naturais, reais ou empíricas. Ou seja, a exemplo do citado por Newman e Newman (2010), ele caracteriza essas redes pela sua capacidade de representar cenários, sendo eles reais ou não.

Define-se grafo como uma estrutura composta por um conjunto de itens (nós ou vértices) e um conjunto de arestas, normalmente definidas por relações binárias entre dois ou mais itens. Há mais de um tipo de grafo, dentre eles os não direcionados resultam em matrizes simétricas, sendo uma das representações mais comuns a vértice-aresta, onde cada nó é representado como um ponto sendo as arestas linhas entre os nós os quais elas conectam. Essa representação é importante para ilustração e para orientar intuitivamente a caracterização dos sistemas (Fabbri, 2017) e será utilizado em na centralidade proposta neste trabalho .

Imagine um grande número de objetos (nós) dispostos em um grafo, representando uma rede complexa, interligados por arestas que representam as interações entre eles. Esses objetos podem representar pessoas, empresas, ou qualquer tipo de entidade que se relacione. As arestas podem determinar qualquer tipo de relação entre os objetos, e.g. amizade, comercial, subordinação. Dessa forma, o conjunto dos objetos e suas relações descrevem um ambiente ou contexto em que esses objetos interagem. Classificar esses objetos por relevância num contexto tendo como referência as suas relações é um desafio, uma vez que essas interações podem variar de forma imprevisível. Essa imprevisibilidade diz respeito à dinamicidade intrínseca dessas redes que variam de acordo com a temporalidade, ou seja, à medida que os cenários se alteram ao longo do tempo, as redes, por representá-los, também variam. As medidas de centralidade têm essa missão.

Pode-se entender o termo centralidade como uma ferramenta para quantificar a relevância dos nós em uma rede (Caldarelli, 2020). O estudo das medidas de centralidade teve início nos anos 50, introduzindo o papel dos nós nos padrões de comunicação (Das et al., 2018). Desde então, estudos constantes visam melhorar os resultados da classificação desses nós. As chamadas medidas de centralidade fazem uso de métodos que, por meio de observação de casos específicos, inferem sobre o funcionamento das interações entre os nós, principalmente com relação à disseminação de informações em um grupo (Bavelas, 1950). Dessa forma, as medidas de centralidade, cada uma com uma estratégia, tentam resolver o problema de classificar a relevância dos nós de uma rede. Exemplos de centralidades clássicas incluem: *Betweenness* (Freeman, 1978;

Otte e Rousseau, 2002), *Eigenvector Centrality* (Bonacich, 1972), *PageRank* (Bonacich, 2007), and *Weighted degree* (Beveridge e Shan, 2016).

Normalmente, essas métricas usam a posição topológica dos nós como base para sua classificação, e.g., o número de conexões de nós, as conexões de nós dispostos na vizinhança, o número de caminhadas e os caminhos que cruzam o nó. Métricas diferentes tentam fornecer uma resposta à pergunta: "*quais são os nós mais importantes em uma rede?*" (Brandes e Erlebach, 2005; Liao et al., 2017). A gama de aplicações para esta tecnologia é vasta, por exemplo, epidemiologia (Brandes e Erlebach, 2005; Christakis e Fowler, 2010; Pastor-Satorras et al., 2015), economia (Guimerà et al., 2005; Schweitzer et al., 2009), neurociências (Bullmore e Sporns, 2009), engenharia (Rinaldo et al., 2006) e detecção de fraudes (Kolaczek e Juszczyszyn, 2019).

No campo da auditoria, e.g., um problema comumente encontrado ocorre quando um auditor se depara com muitos suspeitos de fraude, mas não dispõe de recursos suficientes para auditar a todos. Surge então, de forma recorrente, a questão: quais suspeitos devem ser selecionados para auditoria dentre todo o universo de suspeitos? Nesse contexto, ter um sistema que possa indicar quem deve ser prioritariamente auditado, por possuir maiores chances de ser um fraudador, é de grande proveito (Cunha e Bugarin, 2014). Esse exemplo é tratado com detalhes no estudo de caso da "Operação Licitante Fantasma", no capítulo 6.

No campo empresarial, a determinação de um público alvo para uma empresa é uma outra possibilidade de uso da tecnologia. O chamado marketing ativo, onde uma empresa faz uma propaganda direta dos seus produtos a um público alvo, pode ser fortemente aprimorado com o uso de ferramentas que indiquem, ou ranqueiem, os clientes com maiores possibilidades de efetuar uma compra, aumentando, assim, as chances de sucesso na venda dos produtos (Ribeiro, 2016).

Esses são apenas exemplos que servem para ilustrar e dimensionar a diversidade de situações onde se encontra aplicabilidade à tecnologia que permite, mediante um cenário definido e do seu necessário mapeamento, identificar quais são os objetos de maior importância. No entanto, testes em um amplo conjunto de redes demostraram limitações das medidas de centralidade clássicas em encontrar os nós mais relevantes em redes complexas (Sciarra et al., 2018).

Idealmente uma única rede complexa (Newman e Newman, 2010; Boccaletti et al., 2006; Fabbri, 2017) não pode descrever alguns sistemas naturais. Considere-se, por exemplo, uma extensa coleção de livros contendo uma única história principal onde alguns personagens surgem como muito importantes durante a trama, mas perdem relevância ao longo da história. Dessa forma, como obter uma classificação precisa, sendo que as nuances da história mudam a importância das personagens ao longo de

vários livros? Dessa forma, tem-se um cenário (coleção de livros) que deve considerar vários contextos (livros). Qual seria a solução para modelar essa situação real em uma única rede complexa? Como as métricas de centralidade convencionais podem fornecer uma classificação de nós em situações como essa? Esse exemplo é tratado com detalhes no estudo de caso da coleção de livros da saga Harry Potter no capítulo 7.

Nesses casos, uma solução natural é utilizar redes que, analisadas em conjunto, tenham a capacidade de modelar uma situação fática (Almeida et al., 2017). Considerando o exemplo e supondo que cada livro seja modelado como uma rede complexa separada representando cada contexto individualmente. Para obter a classificação de toda a coleção de livros, que representa o cenário, seria necessário adicionar as redes que modelam cada livro em uma única rede e, só então, usar uma métrica de centralidade padrão para obter os nós mais relevantes. No entanto, essa abordagem é falha, pois as redes que compõem todo o cenário podem ter características singulares que mudam com o tempo. Ou seja, uma personagem pode, e.g., ter uma grande importância nos primeiros livros e depois desaparecer nos demais. Dessa forma, a junção de todas as redes que descrevem o cenário (coleção de livros) e a utlização de uma centralidade para classificar os nós mais relevantes poderia apontar, de forma equivocada, essa personagem como sendo muito relevante em todo o contexto, sem considerar, entretanto, a sua saída prematura de cena.

Considere-se agora um cenário de disputa política nos EUA, entre democratas e republicanos, em que os nós representam os deputados e as arestas os encontros entre eles. Como métrica, considera-se como mais relevantes os deputados que se reunirem mais vezes com o maior número de deputados do partido político oposto. Quatro redes descrevem esse cenário, cada uma sendo um contexto, ou ano de compromissos. Agora imagine que, nos primeiros dois anos, um grupo restrito de cinco parlamentares, de ambos os partidos, teve um número enorme de reuniões e, após uma ruptura política, essas reuniões pararam. Nesse cenário, se unirmos as quatro redes em uma rede única, como esses cinco congressistas tiveram um número enorme de reuniões nos primeiros dois anos, e utilizar-se uma centralidade para a classificação dos nós, eles podem aparecer, equivocadamente, como os cinco nós mais relevantes de todo o cenário. Seria uma interpretação errônea, uma vez que não tiveram esse status durante todo o período. Em outras palavras, a combinação de todas as redes que descrevem um cenário em uma rede pode nos levar a uma interpretação incorreta dos fatos que ocorrem em diferentes contextos ao longo do tempo.

É claro que se trata de suposições acerca da influência desses políticos em todo o cenário. Considerando o exemplo, os políticos que tiveram uma grande influência em um curto período poderiam realmente ser os mais influentes, mesmo não estando

presentes no período como um todo. As métricas partem de suposições para classificar os indivíduos mais relevantes sendo que, essas suposições, podem também ser falhas. Logo, não se trata aqui de uma verdade absoluta, mas de tratar o mesmo problema com uma abordagem distinta que, a princípio, é capaz de obter resultados mais precisos.

Os dois exemplos acima apontam lacunas na estratégia de junção de um grupo de redes complexas em uma única rede, combinado com o uso de uma centralidade para indicar os nós mais relevantes, como foi feito em Almeida et al. (2017). Uma vez que uma única rede não consegue representar de forma ideal todos os cenários, as redes multicamadas atraem cada vez mais pesquisadores. Considerando que cada camada represente um contexto, elas podem descrever vários tipos de cenários com interações entre qualquer par de nós (Bianconi, 2018). Considerando o exemplo acima, pode-se dividir os quatro anos de reuniões em quatro camadas (contextos), e ter uma classificação de nós mais precisa, pensando em cada ano como uma parte independente de um cenário.

No entanto, as centralidades clássicas não funcionam com esta estrutura de rede. Assim, propostas de medidas de centralidade, incluindo as métricas que visam trabalhar com redes multicamadas, surgem como a extensão natural das métricas padrão (Bródka et al., 2012; Tu et al., 2018; Solé-Ribalta et al., 2016); sendo, em alguns casos, uma melhoria dos algoritmos originais (Tu et al., 2018; Solé-Ribalta et al., 2016). Neste trabalho, são abordadas as redes multiplex, um tipo particular de redes multicamadas, descrita na Seção (2.4).

O tema central desta tese é a proposta de uma nova medida de centralidade intitulada: *Group-based Centrality for Undirected Multiplex Networks (GCMN)* (De Figueirêdo et al., 2021), descrita no (Capítulo 3). Essa medida é aplicável a situações que requerem, para descrição de seu cenário, uma rede complexa de múltiplas camadas. Ela propõe uma abordagem alternativa para as métricas de centralidade definidas pelo passeio aleatório (*random walk*) (seção 2.5.3) em redes multicamadas não direcionadas (Tu et al., 2018; Solé-Ribalta et al., 2016), propondo uma nova estratégia baseada na interseção das camadas como um parâmetro para atribuir peso aos nós.

1.1 Motivação

Uma das abordagens possíveis no estudo das redes complexas é o da descoberta dos nós mais relevantes, para isso são utilizadas as chamadas medidas de centralidade. Um exemplo muito conhecido é o da medida *PageRank* (Bonacich, 2007), utilizada pelo engenho de busca da Google na tarefa de ranquear páginas na Web. O uso das

redes complexas tem se tornado uma alternativa promissora para classificação de dados pela captura das relações semânticas existentes entre os atores em um cenário, como as formações de padrões; bem como na captura das relações espaciais, topológicas e funcionais dos dados (Carneiro, 2016).

Dessa forma, as medidas de centralidade quantificam a relevância dos nós em redes complexas sob aspectos distintos (Almeida et al., 2017). Cada centralidade tem uma estratégia e, consequentemente, um algoritmo próprio. Como exemplos de centralidades conhecidas tem-se: a *Betweenness centrality*, que representa o número de caminhos mínimos que passam por um nó (Otte e Rousseau, 2002); as centralidades *Eigenvector* e *PageRank* que consideram o número de links de um nó para classificá-lo (Bonacich, 2007); a distribuição do grau ponderado (*Weighted Degree*) que corresponde à soma dos pesos das arestas em um nó v e está relacionada à "popularidade" desse nó (Beveridge e Shan, 2016). Propostas de novas centralidades surgem naturalmente, muitas vezes buscando o preenchimento de lacunas quanto à eficácia ou propondo estratégias inovadoras, sempre tendo como foco a tarefa de buscar as entidades mais relevantes (Yang et al., 2020; Bródka et al., 2012; Agneessens et al., 2017; Ziberna, 2020; Fire e Guestrin, 2020; De Figueirêdo et al., 2021; Tu et al., 2018).

A aplicabilidade do uso das centralidades, a exemplo de outras tecnologias já mencionadas, é ampla e generalista, podendo ir desde a descoberta de pessoas influentes num contexto social, políticos e empresários corruptos em um esquema de ilicitude (Almeida et al., 2017; De Figueirêdo et al., 2020) a personagens mais relevantes de um conjunto de livros (Carvalho, 2017). A condição para que isso ocorra é de que todos esses cenários sejam representados por meio de redes complexas, com os nós sendo as entidades a serem avaliadas, ou descobertas, e as arestas indicando as relações entre essas entidades.

Entretanto, problemas surgem quando o cenário a ser estudado é descrito pela junção de diversas redes complexas. Nessas situações uma abordagem comum é a de somar as redes em uma única grande rede complexa que represente todo o contexto (Almeida et al., 2017). Essa abordagem pode trazer problemas interpretativos quanto à análise final dos nós mais relevantes, uma vez que redes complexas diferentes, embora tratem do mesmo assunto, podem ter características singulares que mudam com o tempo. Esses problemas são abordados na introdução deste trabalho.

Dessa forma, as medidas de Centralidade encontram limitações na tarefa de determinar os nós mais relevantes em cenários descritos por mais de uma rede complexa. Apesar de já existirem trabalhos que visam o preenchimento dessa lacuna (Yang et al., 2020; Bródka et al., 2012; Agneessens et al., 2017; Ziberna, 2020; Fire e Guestrin, 2020; Tu et al., 2018) entende-se que há espaço para novas abordagens (De Figueirêdo et al.,

2021) sendo essa a principal motivação para a elaboração deste trabalho.

1.2 Objetivo

O objetivo central deste trabalho é propor e demonstrar a eficácia de uma nova centralidade voltada à detecção de nós relevantes num cenário descrito por redes multiplex não direcionadas, a ***Group-based Centrality for Undirected Multiplex Networks (GCMN)***. A centralidade proposta se baseia no agrupamento hierárquico dos nós (Capítulo 3), sendo uma estratégia alternativa às métricas de centralidade definidas pelo passeio aleatório (*random walk*) em redes multicamadas não direcionadas (Tu et al., 2018; Solé-Ribalta et al., 2016).

Essa estratégia visa resolver o problema de interpretações equivocadas na determinação dos nós mais relevantes em cenários descritos por redes com multiplas camadas, citadas em dois exemplos na seção 1. Além disso, também visa ofertar uma centralidade de execução eficiente, demonstrada por meio de sua complexidade algorítmica (seção 3.2) com uma eficácia compatível ou superior a outras centralidades, sendo esses resultados demonstrados por meio de estudos de caso (seção 8.1).

Como objetivos específicos, tem-se:

1. demonstrar que a proposta de hierarquização da classificação dos nós, presente na centralidade GCMN, elimina os problemas de superdimensionamento da relevância de nós descrito em dois exemplos na seção 1. Essa demonstração pode ser encontrada nas seções: 5.4.1, 6.4.1 e 7.4.1;

2. constatar que a centralidade GCMN é capaz de classificar nós relevantes não detectados por outras centralidades em um volume maior que todas as demais centralidades analisadas. Essa constatação é encontrada nas seções: 5.4.4 e 7.4.3;

3. provar que a centralidade GCMN é capaz de apresentar um desempenho compatível ou superior, quando comparada a outras centralidades, mediante a aferição de métricas estabelecidas como: *Precision, Recall, Accuracy* e F_1 e das correlações propostas por Pearson e Spearman. Essa prova é encontrada nas seções: 5.4.5, 6.4.3, 6.4.4, 7.4.4 e 7.4.5;

4. provar a viabilidade e eficiência da implementação da centralidade GCMN por meio da análise de sua complexidade algorítmica. Essa prova é encontrada na seção 3.2.

1.3 Justificativa

Áreas, como a de autoria, trabalham fortemente com espaços amostrais como forma de atenuar a limitação de recursos de auditagem em face de grandes demandas. Órgãos de controle como os Tribunais de Contas e a Controladoria Geral da União e dos Estados sofrem com essa questão, onde o estabelecimento de critérios sobre o que auditar tem um impacto significativo na sua produtividade e resolutividade (Rocha, 2002). Dessa forma, ferramentas que apontem quem deve ser auditado, propiciando uma maior chance de sucesso na auditoria, em cenários compostos por uma grande gama de objetos auditáveis, são de grande valia. Por exemplo: tomando-se o estudo de caso abordado no Capítulo 6, a "Operação Licitante Fantasma", tem-se um escândalo de corrupção envolvendo uma grande quantidade de empresas a serem auditadas, mais precisamente 1465. O uso de uma ferramenta que aponte quais são as empresas com maiores chances de estar participando de esquemas de ilicitude é imprescindível à viabilidade e consequente sucesso do processo de auditoria.

O uso de centralidades clássicas na detecção de nós relevantes em cenários descritos por redes com mais de uma camada é passível de problemas como os relatados na introdução deste trabalho, deixando lacunas importantes quanto à fidedignidade das classificações apresentadas. Situações como a dos livros da saga Harry Potter (Capítulo 7) podem levar as centralidades clássicas a imprecisões na análise feita pela junção de redes complexas. As medidas de centralidade clássicas, e.g. *Betweenness* (Freeman, 1978; Otte e Rousseau, 2002), *Eigenvector centrality* (Bonacich, 1972), *PageRank* (Bonacich, 2007) e *Weighted Degree* (Beveridge e Shan, 2016); não são capazes de analisar cada uma dessas redes de forma separada e indicar qual o resultado único do ranqueamento das personagens para toda a coleção de livros. Sendo assim, a propositura de uma nova centralidade que venha a preencher essa lacuna é, dessa forma, relevante dada a sua ampla aplicabilidade.

No caso da centralidade ora proposta, ela trata cada camada da rede complexa de forma restrita ao seu escopo, não permitindo que a relevância de entidades adstritas a uma rede ou a um subconjunto delas, extrapole sua suposta importância de forma irreal. Para tanto, cada rede complexa, que represente um contexto, é transformada numa camada de uma rede multiplex, que representa todo o cenário, sendo preservadas todas as relações entre os nós das redes complexas originais.

Faz-se também necessária a formalização da centralidade proposta por meio da especificação de um modelo matemático e da proposição de algoritmos para a sua implementação. Uma vez formalizado o modelo matemático, haverá a sua aplicação a três situações reais para que se possa atestar sua validade de forma pragmática descritas

nos capítulos 5, 6 e 7.

Dessa forma, justifica-se a elaboração deste trabalho pela relevância de classificar-se nós em redes que representem contextos de forma a preencher as lacunas interpretativas das centralidades clássicas apontadas na introdução desta tese. Além disso, o uso de uma nova estratégia que utilize algoritmos novos amplia o escopo de pesquisa propondo novas estratégias, sendo relevante na medida que propõe uma nova forma de pensar o problema da classificação.

1.4 Organização do Trabalho

- Capítulo 2 — Referencial Teórico. O capítulo traz a definição de conceitos básicos tais como: Teoria dos Grafos (seção 2.1), Redes Complexas (seção 2.2), Notação Tensorial (seção 2.3), Redes Multiplex (seção 2.4) e Medidas de Centralidade (seção 2.5). Com relação às medidas de centralidade há o detalhamento daquelas que servirão como parâmetro de aferição da eficácia da centralidade GCMN nos estudos de caso nos capítulos 5, 6 e 7. Essas medidas são descritas em detalhes nas subseções da seção 2.5;

- Capítulo 3 — *The Group-based Centrality for Multiplex Networks (GCMN)*. Apresenta a centralidade GCMN, objetivo desta tese. Formaliza o modelo matemático, exemplifica uma situação genérica para melhor entendimento do modelo proposto (seção 3.1), propõe algoritmos para composição da classificação e recupareração da classificação de um nó (seção 3.2), expõe casos particulares e fragilidades da proposta (seção 3.3) e traz as considerações finais na seção 3.4;

- Capítulo 4 — Materiais e Métodos. Expõe os materias e métodos utilizados na obtenção dos resultados. A seção 4.1 traz alguns conceitos gerais necessários ao entedimento dos três capítulos de estudos de caso (capítulos 5, 6 e 7), a seção 4.2 apresenta os recursos técnicos utlizados na obtenção e tratamento dos dados, tais como: fontes de dados, liguagens de programação, softwares aplicativos e bancos de dados e a seção 4.3 faz as considerações finais do capítulo;

- Capítulos 5, 6 e 7 — Estudos de Caso. Apresentam três estudos de caso, a saber: Operação Lava Jato (capítulo 5), Operação Licitante Fantasma (capítulo 6) e Coleção de Livros da Personagem Harry Potter (capítulo 7). A estrutura dos três capítulos é semelhante, contendo uma introdução, os trabalhos relacionados, uma métrica específica de avaliação proposta para cada estudo de caso, os resultados e a discussão e considerações finais;

- Capítulo 8 — Conclusões. Discute-se as principais contribuições da tese (seção 8.1), as limitações da proposta apresentada (seção 8.2), propõe trabalhos futuros (seção 8.3) e traz a lista as publicações obtidas a partir desde trabalho (seção 8.4).

Capítulo 2

Referencial Teórico

Este capítulo traz a definição de conceitos básicos tais como: Teoria dos Grafos (seção 2.1), Redes Complexas (seção 2.2), Notação Tensorial (seção 2.3), Redes Multiplex (seção 2.4) e Medidas de Centralidade (seção 2.5). Com relação às medidas de centralidade há o detalhamento daquelas que servirão como parâmetro de aferição da eficácia da centralidade GCMN nos estudos de caso nos capítulos 5, 6 e 7. Essas medidas são descritas em detalhes nas subseções da seção 2.5.

2.1 Teoria dos Grafos

A teoria dos grafos é um ramo da matemática que trata dos objetos e das relações entre eles. Os objetos são representados como vértices (ou nós), e a suas relações como arestas. Formalmente tem-se que um grafo G sobre um conjunto não vazio de vértices V e um conjunto pares não ordenados de V denominados arestas E é dado por $G(V, E)$. Dessa forma cada nó é, usualmente, identificado por número natural ordenado $i = 1, 2, ..., n$; sendo a representação da conexão (aresta) entre dois nós i e j dada por (i, j), logo tem-se que $E \subseteq (i, j)|i, j \in V$.

A depender da caraterística do grafo eles podem ser classificados de diversas formas:

1. Grafos não direcionados, onde uma aresta (i, j) apresenta uma relação de simetria, ou seja, implica na existência de (j, i), com o mesmo significado semântico;

2. Grafos ponderados, onde há a associação de valores numéricos às suas arestas. Esses valores determinam um peso associado a cada aresta que pode ser representado por meio de uma função $\omega : E \to \mathbb{R}$, que atribui a cada aresta $(i, j) \in E$ um peso $\omega(i, j)$.

Embora exista uma variedade de outros tipos de grafo, esta tese restringe-se a tratar com grafos não direcionados e grafos ponderados. As estruturas que podem ser representadas por esses grafos são inumeráveis, e.g., a estrutura topológica de uma rede de computadores, as rotas aéreas de uma empresa de aviação, as relações de amizade numa rede social. Nesse último exemplo, os nós da rede seriam associados às pessoas, enquanto as conexões, ou arestas, representariam as relações de amizades entre esses indivíduos (Carneiro, 2016). Em termos de representação gráfica (Figura 2.1), num grafo não direcionado, geralmente são utilizados círculos simbolizando os nós e retas como sendo as arestas (Biggs et al., 1986, pp. 1736-1936).

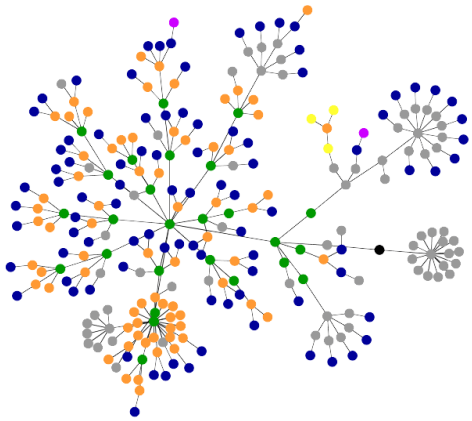


Figura 2.1. Representação gráfica de um grafo.

2.2 Redes Complexas

As redes complexas lidam com conceitos de estatística, teoria dos grafos, sistemas dinâmicos, despertando interesse em inúmeras áreas do conhecimento científico, como: física, matemática, biologia e sociologia. Isso se dá graças as suas aplicações envolvendo uma vasta gama de problemas, que incluem: redes socais, biológicas, Internet e redes elétricas (Newman e Newman, 2010; Newman, 2003; Boccaletti et al., 2006; Fortunato, 2010). A literatura possui diversas medidas e modelos para caracterizar a estrutura e topologia dessas redes. Essas medidas são frequentemente usadas para analisar as propriedades estatísticas que descrevem a estrutura e o comportamento de sistemas descritos por redes complexas, enquanto que a criação de modelos de rede está relacionado ao entendimento do significado dessas propriedades.

Uma rede complexa consiste em um grafo onde os nós representam as entidades e as arestas os relacionamentos entre elas. Sua estrutura topológica é, em geral, a representação, ou modelagem, de um cenário real ou fictício, que pode evoluir ao longo

do tempo (Boccaletti et al., 2006), sendo compostas de estruturas que não são completamente regulares ou aleatórias (Carneiro, 2016) (seção 2.2). Em outras palavras, a sua estrutura é mutável ao longo do tempo, por representar a descrição de cenários também mutáveis que evoluem de forma temporal, não sendo regulares ou aleatórias justamente pro representar tais cenários.

No que se refere aos tipos de redes complexas, os mais conhecidos são: Pequeno Mundo e as Redes Livres de Escalas. No primeiro caso, as redes são caraterizadas por um elevado agrupamento de nós, sendo, o segundo, regido por uma lei de potência, onde dois escalares x e y possuem uma relação que pode ser escrita na forma $y = ax^k$, onde a (a constante de proporcionalidade) e k (o expoente) são constantes (Guerriero, 2012). De forma diversa a outros de tipos de redes, que são descritos por meio de mecanismos de crescimento, alguns problemas em redes complexas são solucionados ou entendidos por meio de otimizações de uma meta ou mesmo por meio das características dos seus nós e arestas, o que é chamado de otimização estrutural de redes, e.g., redes que representam rotas aéreas (Newman e Newman, 2010).

A possibilidade de segmentar uma rede complexa em grupos de nós que tenham características comuns levou ao estudo da modularidade. Conceitualmente, a modularidade Q é capaz de quantificar quão pertinente é a divisão da rede proposta. É uma medida de varia entre 0, como sendo a aleatoriedade total da divisão, a 1, indicando a existência de comunidades de nós. A modularidade foi proposta por Clauset et al. (2004), tornando-se um elemento essencial de diversos algoritmos de agrupamento. A modularidade é a mais conhecida e utilizada função para o agrupamento de nós que existe, incorporando todos os itens essenciais e questões, a partir de definição de comunidade, com o intuito da escolha de um modelo nulo para a expressão da "força" de comunidades e partições (Fortunato, 2010).

Formalmente define-se a modularidade como:

$$Q = \frac{1}{2E} \sum_{u,v} \left[e_{uv} - \frac{k_u k_v}{2E} \right] \delta(c_u, c_v), \tag{2.1}$$

onde E corresponde ao número total de arestas na rede, e_{uv} diz respeito à fração de arestas que conectam nós entre u e v na comunidade, k_u é o grau do nó u, c_u e c_v são os valores escalares atribuídos aos vértices e $\delta(c_u, c_v)$ representa o delta de Kronecher, o qual produz 1 caso $c_u = c_v$ e 0, caso isso não ocorra.

De uma forma simplificada, a modularidade é responsável por calcular a fração de conexões que ocorrem dentro dos grupos, subtraída pela quantidade esperada de conexões caso elas sejam distribuídas aleatoriamente. Dessa forma, considerando uma determinada divisão de nós da rede, a modularidade reflete a concentração de nós

dentro dos módulos, comparada com uma distribuição aleatória de ligações entre todos os vértices, independentemente dos módulos (Newman, 2006).

Um exemplo de rede complexa onde foi utilizada a modularidade para a detecção de comunidades, separadas por cores, pode ser encontrada na Figura 2.2.

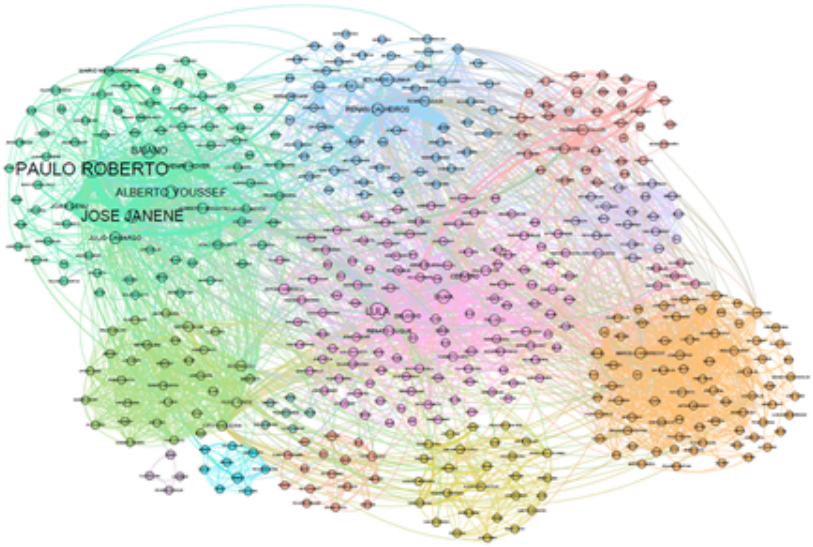


Figura 2.2. Exemplo de Rede Complexa, com separação de comunidades.

2.3 Notação Tensorial

Nesta tese será utilizada a notação tensorial representando matrizes adjacentes usando álgebra de ordem superior (De Domenico et al., 2013). Uma vantagem significativa em usar o formalismo dos tensores é sua compactação. Pode-se escrever uma matriz de adjacência, ou tensor, usando uma notação compacta que é muito útil para a generalização de descritores de rede para multicamadas ou, no caso específico da Centralidade GCMN, redes multiplex (Seção 2.4).

Em notação tensorial, um vetor linha $i \in \mathbb{N}$ é dado por um vetor co-variante $i_\alpha(\alpha = [1, N])$. Seu vetor contravariante correspondente i^α (i.e., seu vetor duplo) é um vetor coluna no espaço euclidiano. Um vetor canônico é atribuído a cada nó e um tensor de adjacência representa a rede multicamadas interconectada correspondente. Neste caso, um tensor $A_{ij}^{\alpha\beta}$ pode representar a intensidade da relação (que pode não ser simétrica) de um nó i na camada α para um nó j na camada β.

Na formalização da centralidade GCMN, será utilizado um tensor de adjacência intra-camada, ou tensor de segunda ordem A_{ij}^{α}. Esse tipo de tensor indica as relações

entre os nós i e j numa camada α. Na seção 2.4, será utilizada a notação tensorial para referenciar uma rede fictícia, que será utilizada como exemplo.

2.4 Redes Multiplex Não Direcionadas

Redes multiplex são um tipo específico de redes multicamadas em que cada nó aparece em camadas diferentes e cada camada contem apenas arestas de um determinado tipo, ou seja, arestas que tenham a mesma representação semântica, e.g. relações de amizade, relações comerciais. Esses nós não podem ter conexões com outros nós em outras camadas. Uma matriz tridimensional de tamanho $(V \times V) \times L$, na qual V representa os vértices (nós) e L as camadas, ou dimensões; onde entradas do tipo A_{ij}^{α}, são suficientes para representar a estrutura do sistema (Mucha et al., 2010; Nicosia et al., 2013). Usando a notação tensorial (De Domenico et al., 2013), matrizes de adjacência são indicadas por tensores de adjacência multiplex A_{ij}^{α} para codificar conexões entre nós $\{i, j \mid i, j \in V\}$ na camada $\{\alpha \mid \alpha \in L\}$.

Um tensor é um objeto algébrico que descreve uma relação multilinear entre conjuntos de objetos algébricos. No caso tratado nesta tese, esses objetos são as matrizes de adjacência indicadas pelos nós em V e as camadas em L. Os valores A_{ij}^{α} serão iguais a um quando houver uma aresta entre os nós i e j na camada α, e zero caso contrário.

Assim, pode-se representar uma UMN (*Undirected Multiplex Network*) como uma tripla (V, E, L), com V sendo o conjunto de nós, L as camadas, e E um conjunto de entradas $\{A_{ij}^{\alpha} \mid i, j \in V, \alpha \in L, i \neq j\}$; para quaisquer duas entradas $\{A_{ij}^{\alpha}, A_{xy}^{\beta} \in E\}$, se $i = x$ e $j = y$, então $\alpha \neq \beta$. Observe-se que se estar lidando com redes não direcionadas, então uma entrada A_{ij}^{α} é equivalente a A_{ji}^{α} (Bródka et al., 2012).

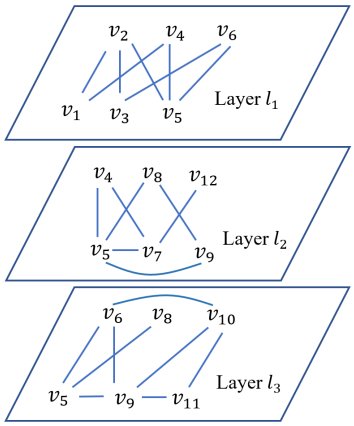


Figura 2.3. Exemplo de rede multiplex não direcionada.

A Figura 2.3 traz um exemplo de uma UMN em que: $V = \{v_1, ...v_{12}\}$, $L = \{l_1, l_2, l_3\}$ e E é um conjunto de matrizes de adjacência que representam as conexões entre os nós pertencentes a V nas camadas em L. Cada matriz de adjacência $A_{ij}^{l_1}$ representa as conexões entre os nós $\{i, j \mid i, j \in V\}$ na camada l_1, com os componentes são iguais a um quando há uma aresta entre i e j, e zero caso contrário.

2.5 Medidas de Centralidade

As medidas de centralidade, ou simplesmente centralidades, são, comumente, relacionadas à análise das propriedades estatísticas que descrevem a estrutura e o comportamento de uma rede complexa. A centralidade GCMN, proposta neste trabalho, enquadra-se nesse tipo de abordagem. Elas são oriundas de estudos em diversos campos do conhecimento, e.g., estatística, matemática, sistemas complexos, sistemas não lineares (Newman, 2003). Uma das características mais relevantes das medidas de centralidade trata da sua capacidade de ranquear os nós de uma rede complexa (Almeida et al., 2017; Carvalho, 2017).

As centralidades fazem uso de métodos que, por meio de observação de casos específicos, inferem sobre o funcionamento das interações entre os nós, principalmente com relação à disseminação de informações em um grupo (Bavelas, 1950). Dessa forma, as medidas de centralidade, cada uma com uma estratégia, tentam resolver o problema de classificar a relevância dos nós de uma rede. Exemplos de centralidades clássicas incluem: *Betweenness* (Freeman, 1978; Otte e Rousseau, 2002), *Eigenvector centrality* (Bonacich, 1972), *PageRank* (Bonacich, 2007) e *Weighted Degree* (Beveridge e Shan, 2016).

Embora existam inúmeras medidas de centralidade, serão apenas detalhadas a seguir aquelas que servirão de parâmetro para aferição da eficácia da Centralidade GCMN, proposta nesta tese.

2.5.1 Betweenness Centrality

Trata-se de uma medida que indica a centralidade de nós em uma rede (Otte e Rousseau, 2002; Freeman, 1978). Tomando-se uma rede G composta por um conjunto de vértices V, que se "comunicam" por meio de troca de mensagens entre os seus nós e que, para tanto, utilizam o algoritmo do menor caminho (caminho geodésico). Dessa maneira, o grau de intermediação b_i de um nó i é equivalente ao número de caminhos geodésicos (η), que possuam o nó i em seu percurso. Em outras palavras, avalia-se

a presença do nó i no menor caminho possível de um nó u para outro nó v, tal que $u, v \in V - i$. Formalmente tem-se que o *Betweeness* de um nó i, é dado por:

$$b_i = \sum_{u,v \in V-i} \eta^i_{uv}. \tag{2.2}$$

Considerando ainda a possibilidade de existir mais de um caminho ótimo, ou menor caminho, entre u e v, denomina-se ng_{uv} a quantidade de menores caminhos entre os nós. Considerando ainda que o nó i também pode vir a participar de caminhos existentes em ng_{uv}, denomina-se ng^i_{uv} como sendo o número de caminhos em ng_{uv} que passam por i. Assim, η^i_{uv} é definido como:

$$\eta^i_{uv} = \frac{ng^i_{uv}}{ng_{uv}}. \tag{2.3}$$

Dessa forma, a complexidade algorítmica para a criação de uma classificação utilizando a Betweenness Centrality é da ordem de $O(|V||E|)$, onde $|V|$ representa o conjunto dos vértices e E o conjunto dos vértices da rede complexa.

2.5.2 Eigenvector Centrality

A *Eigenvector Centrality* visa aferir a influência que um nó exerce sobre uma rede. A centralidade ranqueia os nós da rede baseando-se na premissa de que as conexões com os nós mais bem ranqueados têm um peso maior para indicar a relevância de um nó que esteja sendo avaliado do que as conexões aos nós de baixa pontuação.

Formalmente tem-se que, para um grafo $G = (V, E)$ com $|V|$ nós e $E = (e_{v,t})$, a matriz de adjacência, ou seja, $e_{v,t} = 1$ caso os nós v e t estejam conectados, e $e_{v,t} = 0$ caso contrário. O cálculo da centralidade do nó v é dado por:

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} e_{v,t} x_t, \tag{2.4}$$

onde λ é um valor constante e $M(v)$ diz respeito ao conjunto de vizinhos de v. Em geral, haverá muitos valores diferentes de λ para os quais existe uma solução de *eigenvector* diferente de zero.

Normalmente existirão muitos valores possíveis de λ para os quais uma solução da *Eigenvector Centrality* existe. Entretanto, todas as entradas do Eigenvector Centrality só serão positivas caso apenas os maiores resultados sejam considerados. Dessa forma, pela premissa adotada, apenas os valores mais significativos terão influência na determinação da pontuação do nó avaliado (Newman e Newman, 2010).

A complexidade algorítmica para a criação de uma classificação utilizando a *Eigenvector Centrality* é da ordem de $O(|V||E|)$, onde $|V|$ representa o conjunto dos vértices e E a matriz de adjacência contendo todos as arestas da rede complexa.

2.5.3 PageRank Centrality

A PageRank Centrality, utilizada pelo engenho de busca Google, é uma das variantes da *Eigenvector Centrality* (Bonacich, 2007). A medida de centralidade *PageRank* baseia-se em passeios aleatórios (*random walk*).

A teoria do *random walk* descreve um caminho que consiste em uma sucessão de passos aleatórios que defende que as variações nos valores das ações independem uma da outra, tendo a mesma distribuição. Segundo a teoria, as ações seguem caminhos aleatórios e imprevisíveis, e.g. o caminho de um animal à procura de alimento. A teoria tem aplicações em muitas áreas científicas, incluindo ecologia, psicologia, ciência da computação, física, química, e biologia, e também para a economia. Os passeios aleatórios explicam os comportamentos observados em muitos processos desses campos, e, assim, servem como um modelo fundamental para o registro de atividades estocásticas (Wirth et al., 2016).

De forma intuitiva, a centralidade pode ser interpretada como um modelo de caminhamento aleatório em um grafo. Dessa forma, ao alcançar um nó o algoritmo seleciona, de forma aleatória, uma das arestas daquele nó para dar prosseguimento ao caminhamento no grafo. Além disso, mediante uma probabilidade, o algoritmo permite a seleção de um outro nó qualquer do grafo, independentemente desse nó ter conexões com o nó atual, fazendo com que o agente "salte" para esse nó.

Essa caraterística faz com o que o caminhamento não se prenda a um subgrafo sem conexões de saída, evitando os problemas de convergência conhecidos como *spider-trap* e *dead-end*, que serão comentadas logo abaixo. O valor atribuído de *PageRank* a um nó é dado pela probabilidade de o agente estar naquele nó. Em outras palavras, a relevância de um nó é dada pelo número de arestas daquele nó. Além disso, se os nós que têm conexões com um dado nó possuem uma grande relevância, esse nó também possuirá uma grande relevância (Bonacich, 2007).

Formalmente PageRank é definido como:

$$PR_j^{(t+1)} = \sum_{i \to j} \beta . \frac{PR_j^{(t)}}{k_i^{out}} + (1+\beta)\frac{1}{n}, \tag{2.5}$$

onde $PR_j^{(t)}$ é o *PageRank* do nó i, k_i^{out} é o grau de saída do nó i, n é dado pelo número de nós na rede, β é a probabilidade de saltos aleatórios, e i indica a conexão entre

um nó i e um nó j. Preliminarmente, os valores de PR podem ser definidos de forma aleatória segundo a condição $\sum_j PR_j = 1$, sendo a equação calculada iterativamente até que se obtenha a convergência.

O *PageRank* por ser entendido como uma forma de caminhada aleatória, de forma que, em um tempo t, o agente caminhador encontra-se em um nó i, e, no tempo $t+1$, em um dos nós diretamente conectados a i, dessa forma:

$$p^{(t+1)} = M.p^{(t)}, \tag{2.6}$$

sendo p uma distribuição de probabilidade e M uma matriz de adjacência estocástica, definida como:

$$M_{ji} = \begin{cases} \frac{1}{k_i^{out}} & \text{se } i \to j, \\ 0 & \text{caso contrario.} \end{cases} \tag{2.7}$$

onde a soma de quaisquer das colunas é sempre 1. Uma vez supondo que a caminhada aleatória alcança um estado estacionário, i.e., $p(t+1) = p(t)$, chega-se a uma distribuição estacionária para a caminhada.

É necessário ainda lidar com duas condições possíveis: *spider-trap* e *dead-end*. A *spider-trap* ocorre quando todas as conexões de saída estão dentro de mesmo grupo e o *dead-end* refere-se à quando não há conexões de saída a partir de um dado nó.

O *PageRank* lida com essas situações de convergência com saltos aleatórios, onde, a partir de uma probabilidade β, o agente segue uma das conexões de saída do nó atual, e com a probabilidade $1-\beta$ ele desloca-se para um nó aleatório do grafo. Dessa forma, e sendo equivalente à Equação 2.6, a formulação do *PageRank* também pode ser simplificada para:

$$p^{(t+1)} = \beta M.p^{(t)} + \left[\frac{1-\beta}{n}\right]. \tag{2.8}$$

A complexidade algorítmica para a criação de uma classificação utilizando a *PageRank Centrality* é da ordem de $O(|V||E|)$, onde $|V|$ representa o conjunto dos vértices e E a matriz de adjacência contendo todos as arestas da rede complexa.

2.5.4 Weighted Degree Centrality

A *Weighted Degree Centrality* pode ser definida como similar à *Degree Centrality*, que considera o grau de um nó como parâmetro de sua classificação, sendo este a soma de suas arestas, dada por:

$$D(v) = \sum_{j=1}^{n} a_{vi}. \tag{2.9}$$

Na *Weighted Degree Centrality* a relevância de um nó é dada pela soma dos pesos das arestas que adjacentes a esse nó, dessa forma:

$$f(u) = \sum_{v \in V, v \neq u} \frac{D(v)}{d(u,v)}, \tag{2.10}$$

onde $D(v)$ é o grau do nó v, $d(u,v)$, sendo a menor distância do caminho entre u e v. A ideia central consiste em medir a centralidade de um nó u pela soma dos graus de todos os outros nós com pesos decrescentes com a distância de u (Beveridge e Shan, 2016).

A complexidade algorítmica para a criação de uma classificação utilizando a *Weighted Degree Centrality* é da ordem de $O(|V|(2|E|))$, onde $|V|$ representa o conjunto dos vértices e E a matriz de adjacÊncia contendo todos as arestas da rede complexa.

2.5.5 Closeness Centrality

A medida de centralidade *closeness* considera o inverso do caminho mínimo (distância geodésica) médio entre um nó e os demais nós da rede (Freeman, 1978). Formalmente é defina como:

$$AC_i = \frac{n-1}{\sum_{j=1}^{n} d(i,j)}, \tag{2.11}$$

onde $AC_i \in [0,1], d(i,j)$ define o caminho mínimo entre os nós i e j; e n corresponde ao número de nós na rede. A *closeness* média é definida como:

$$AC = \frac{1}{n} \sum_{j=1}^{n} AC_i. \tag{2.12}$$

A complexidade algorítmica para a criação de uma classificação utilizando a Weighted Degree Centrality é da ordem de $O(|V||E|)$, onde $|V|$ representa o conjunto dos vértices e E a matriz de adjacência contendo todos as arestas da rede complexa.

2.5.6 Novel Multiplex PageRank in Multilayer Networks

Esta centralidade é uma adaptação da *PageRank* (Subseção 2.5.3) aplicável a redes multiplex. Considerando uma rede multiplex direcionada $G = (G_1, G_2, ..., G_M)$ com M camadas e N nós em cada camada, onde $g_{ij}^{(L)}$ com $L = 1, 2, ..., M$ indicando as conexões

de um nó i com um nó j na camada L, onde $g_{ij}^{(L)} = 1$ se existe a conexão e $g_{ij}^{(L)} = 0$, caso contrário. Note-se que essa definição de rede multiplex não contém conexões de nós entre diferentes camadas, sendo a mesma utilizada na proposição da centralidade GCMN.

Esta medida de centralidade depende do conjunto de valores do parâmetro L, que representa a distribuição de coeficiente de acoplamento, assim cada nó obtém sua centralidade em cada camada. As centralidades do nó em diferentes camadas irão agregar a formar uma centralidade única de forma acoplada, considerando os links internos às camadas. A centralidade é posteriormente atribuída a cada camada do nó pela distribuição de acoplamento coeficiente L. Dessa forma, a centralidade do PageRank Multiplex Si mistura a centralidade do nó i em todas as camadas, e então entrega de volta ao nó em cada camada que entregará para seus vizinhos externos na mesma camada na próxima etapa (Tu et al., 2018).

A definição formal da *Multiplex PageRank* é dada por:

$$S_i(t+1) = \sum_L \sum_j g_{ji}^{(L)} \frac{\alpha_L S_i(t)}{H_j^{(L)}} + \left(1 - \sum_L \alpha_L\right) v_i, \tag{2.13}$$

onde $H_j^{(L)} = \sum_r g_{jr}^{(L)} + \delta\left(0, \sum_r g_{jr}^{(L)}\right)$. Sendo L o conjunto das camadas da rede, $g_{ji}^{(L)}$ as arestas do nó i, S_i a $MultiplexPageRank$ do nó i e $H_j^{(L)}$ o delta de Kronecher aplicado ao somatório das arestas.

A complexidade algorítmica para a criação de uma classificação utilizando a Novel Multiplex PageRank in Multilayer Networks Centrality é da ordem de $O(|L||V||E|))$, onde $|L|$ é o conjunto de camadas da rede multiplex, onde $|V|$ representa o conjunto dos vértices e E a matriz de adjacência contendo todos as arestas da rede complexa.

2.5.7 Cross-layer Degree Centrality - CLDC

A CLDC (Bródka et al., 2012), assim como a GCMN, proposta nesta tese, é uma centralidade aplicável a redes multiplex não direcionadas, que utiliza a notação tensorial (Seção 2.3) na sua formalização. A CLDC é calculada para um determinado nó x pela razão entre o número de nós conectados a ele e o número total de todos os nós da rede (diminuido em um). A Centralidade de grau de camada cruzada (CLDC) é definida como uma soma dos pesos das arestas de entrada e sainda do nó x para a vizinhança em várias camadas $MN(x, \alpha)$, dividido pelo número de camadas multiplicado pelo número total de membros da rede.

A definição formal da CLDC é dada por:

$$CLDC(x,\alpha) = \frac{\sum_{y \in MN(x,\alpha)} w(x,y,l) + \sum_{y \in MN(x,\alpha)} w(y,x,l)}{(m-1)|L|}, \tag{2.14}$$

onde $w(x,y,l)$ representa o peso da aresta l que conecta os nós x e y.

A complexidade algorítmica para a criação de uma classificação utilizando a Cross-layer Degree Centrality Centrality é da ordem de $O(|L||E|)$, onde $|L|$ é o conjunto de camadas da rede multiplex e E a matriz de adjacência contendo todos as arestas da rede complexa.

2.6 Considerações Finais

A revisão realizada neste capítulo oferece subsídios para a apresentação da centralidade GCMN no capítulo 3. Foram vistos os conceitos básicos que permeiam a teoria necessária ao entendimento das redes complexas e multiplex, tendo sido dado um enfoque especial às medidas de centralidade que serão objeto de análise e discussão nos estudos de caso (capítulos 5, 6 e 7).

No próximo capítulo a Centralidade GCMN será matematicamente formalizada. Serão propostos dois algoritmos, o primeiro para a formatação do ranqueamento da centralidade e o segundo para a recuperação da classificação de um nó além de serem discutidos casos particulares e fragilidades da proposta.

Capítulo 3

The Group-based Centrality for Multiplex Networks

A premissa inicial da Centralidade GCMN é a existência de uma rede com múltiplas camadas, e, dentro dessas camadas, a existência de nós interconectados por arestas não direcionadas. Os nós representam elementos (e.g. pessoas, empresas) e as arestas não direcionadas os seus relacionamentos (e.g. amizade, contratos). As camadas representam os diferentes contextos nos quais esses nós podem ou não estar relacionados entre si. O conjunto de camadas de rede representa todo o cenário da análise. Deve-se observar que, como se lida com redes multiplex, as arestas podem representar relacionamentos de diferentes naturezas em cada camada, ou seja, e.g. em uma camada as arestas poderão representar uma relação de amizade entre os nós e, em outra camada, uma relação comercial, considerando, em ambas as camadas, o mesmo conjunto de nós. Assim, a estratégia de classificação dos nós da centralidade GCMN está fortemente baseada na topologia da rede. A centralidade considera a presença de nós em cada camada como primeiro critério de classificação e, como segundo critério, o grau de cada nó. A centralidade não explora nenhuma outra informação inserida em um nó, nem a direção das conexões entre os nós.

As redes com múltiplas camadas podem ser utilizadas para mapear cenários em situações onde uma única rede complexa seja insuficiente. Uma aplicação típica de desse tipo de rede é o mapeamento das relações de amizade entre indivíduos em redes sociais (e.g. Facebook, Instagram ou Twitter). Nesse cenário, cada camada mapeia os relacionamentos em uma rede social, sendo os indivíduos representados pelos nós e relações de amizades pelas arestas. Dessa forma, cada camada pode ter os mesmos indivíduos (nós) com conexões diferentes (arestas) a depender da rede social que cada camada represente.

Como a centralidade GCMN funciona com redes multiplex não direcionadas (UMN), foi especificada uma tripla (V, E, L), com um conjunto V, de nós, e um conjunto L, de camadas. Assim, foi definido o conjunto E (de arestas) representando as relações entre os nós em V em cada camada de L, como:

$$E = \{A_{ij}^{\alpha} \mid i, j \in V \wedge \alpha \in L \wedge i \neq j\}. \tag{3.1}$$

A centralidade GCMN divide os nós em grupos. O critério para este agrupamento é o número de camadas $\alpha \in L$ em que um nó $\{i \mid A_{ij}^{\alpha} \in E\}$ está presente. Essa propriedade foi chamada de peso W de um nó. Assim, a definição da função W é:

$$W_i = \mid \{\alpha \mid A_{ij}^{\alpha} \in E \} \mid . \tag{3.2}$$

Foi definido o grupo G de nós i, que têm o mesmo peso, como:

$$G_w = \{i \mid W_i = w\}, \tag{3.3}$$

em que w é o peso dos nós no grupo G_w. A premissa é que o peso w de um nó é proporcional à sua importância, e, como consequência, um grupo G_w deve ter nós mais relevantes do que um grupo G_{w-1}, sendo que essa premissa traz vantagens (seções 5.4.1, 6.4.1 e 7.4.1) e limitações (seção 8.2). Foram considerados como relevantes apenas os grupos com peso $w \geq 2$, i.e. os grupos contendo nós que aparecem em pelo menos duas camadas.

Note-se que a condição $w \geq 2$ é apenas uma sugestão, podendo ser ajustada segundo às necessidades de cada situação fática analisada. Ou seja, caso a rede multiplex disponha de um número muito elevado de camadas, possivelmente $w \geq 2$ irá representar uma eliminação muito tímida dos nós menos relevantes. Assim a restrição para que os nós sejam considerados relevantes é dada por $w \geq i$, onde i é o número de camadas mínimo em que um nó deve estar presente para que seja considerado como relevante.

Outro conceito essencial é o grau D de um nó, que é a quatidade de arestas conectadas ao longo de todas as camadas (Battiston et al., 2014). Assim, define-se o grau de um nó como:

$$D_i = \mid \{A_{ij}^{\alpha} \mid A_{ij}^{\alpha} \in E \wedge (i = n \vee j = n)\}, \tag{3.4}$$

onde o tensor A_{ij}^{α} tem todas as componentes iguais a um, para todas as arestas existentes. O ranking de centralidade GCMN considera dois critérios simultaneamente: o grupo G de um nó i e seu grau D. A classificação do GCMN considera o grupo como o primeiro critério e o grau como o segundo. Dessa forma, os nós que aparecem apenas

em um pequeno subconjunto de camadas não serão classificados como significativos, mesmo que tenham um valor alto para o seu grau, resolvendo, assim, o problema de distorção encontrado na avaliação dos nós mais relevantes em um cenário composto por múltiplas redes complexas, visto na introdução desta tese.

Para que um nó i seja considerado relevante, ele deve satisfazer a dois critérios simultaneamente: estar presente em um número significativo de camadas, maximizando W_i, e ter um grau significativo D_i. Como o primeiro critério se sobrepõe ao segundo, um nó que apareça em um pequeno subconjunto de camadas não pode ser considerado relevante no contexto como um todo, mesmo que tenha um valor alto para o seu grau. Dessa forma, cada nó estará limitado ao grupo G_w correspondente ao número de camadas nas quais esse nó esteja presente.

É necessário, portanto, garantir que nenhum nó em um grupo G_w tenha uma classificação superior a qualquer outro nó de um grupo G_{w-1}. Consegue-se isso definindo a formulação geral da classificação R de um nó i como:

$$R_i \;=\; \varphi(W_i) \;+\; D_i, \tag{3.5}$$

onde φ deve garantir que $W_i > \; W_j \rightarrow R_i > R_j$. Observe-se que a função φ retorna o mesmo valor para todos os nós de um grupo, considerando que esses nós têm o mesmo peso W (Equação 3.3), então o grau D classifica esses nós em seus grupos.

O mínimo da função φ ocorre quando $W_i \;=\; 2$, uma vez que o ranking de centralidade GCMN considera nós com peso associado, no mínimo, igual a dois. Considerando a Equação 3.5, para garantir que $W_i > \; W_j \rightarrow R_i > R_j$, tem-se que $\varphi(2) > \max_{z \in V} D_z$. Ou seja, a função φ deve garantir que, em seu pior caso, φ exceda o grau mais alto D.

Assim, o modelo nos permite definir qualquer função φ, desde que a condição acima seja respeitada. Considerando que $D_z \in \mathbb{N}$, e que $W_i \geq 2$, tem-se que $\max_{z \in V} D_z + 1$ é o menor valor possível para φ. Portanto, propomos a função φ como

$$\varphi(W_i) = \; \left(\max_{z \in V} D_z + 1 \right) (W_i - 1)\,. \tag{3.6}$$

Então, como o mínimo de φ ocorre quando $W_i \;=\; 2$, tem-se que $\varphi(2) \;=\; (\max_{z \in V} D_z + 1)\,(2 - 1) \;>\; \max_{z \in V} D_z$. Observe-se que esta é apenas uma proposta para a função φ, que será utilizada nos estudos de caso no Capítulo 4; outras funções φ também são válidas desde que respeitem a condição $W_i > \; W_j \rightarrow R_i > R_j$.

O impacto do uso de outras funções para R_i e $\varphi(W_i)$ (Equações 3.5 e 3.6) é proposto como trabalhos futuros na seção 8.3.

3.1 Exemplo de Aplicação da Centralidade GCMN a uma rede multiplex hipotética

O exemplo a seguir demonstra o funcionamento da Centralidade GCMN numa rede multplex hipotética. Aplicar-se-á a centralidade GCMN à rede multiplex proposta na Figura 3.1. A Figura traz um exemplo de uma UMN em que: $V = \{v_1, ...v_{12}\}$, $L = \{l_1, l_2, l_3\}$ e E é um conjunto de matrizes de adjacência que representam as conexões entre os nós pertencentes a V nas camadas em L. Cada matriz de adjacência $A_{ij}^{l_1}$ representa as conexões entre os nós $\{i, j \mid i, j \in V\}$ na camada l_1, com os componentes são iguais a um quando há uma aresta entre i e j, e zero caso contrário.

A aplicação da equação 3.2 no leva aos valores de peso para cada um dos nós da rede. Por exemplo: $W_{v4} =\mid \{\alpha \mid A_{v_4 j}^{\alpha} \in E\} \mid = 2$ (Figura 3.2). O que indica que o nó v_4 tem peso dois por possuir arestas nas camadas l_1 e l_2.

É importante observar que os nós $v_1, v_2, v_3, v_7, v_{10}, v_{11}$ e v_{12}, apesar de esterem presentes nas camadas l_1, L_2 e l_3, não têm peso W associado. Isso ocorre por que esses nós têm peso igual a um e, presumivelmente, não são relevantes. Para os nós v_4, v_5, v_6, v_8 e v_9 os pesos indicam o número de vezes que esses nós aparecem nas camadas l_1, L_2 e l_3 de forma conjunta (Figura 3.2).

A Centralidade GCMN divide os nós em grupos e só então aplica o segundo critério de raqueamento que é o grau D do nó. Isso significa que o primeiro critério para apontar os nós relevantes é o grupo e, depois disso, o ranqueamento dos nós naquele grupo. Em outras palavras, o último nó ranqueado (de menor grau D) em um grupo G_i é, segundo a premissa da centralidade GCMN, mais relevante que o nó mais bem raqueado no grupo G_{i-1}. Logo, para que um nó seja considerado relevante, não é apenas necessário apenas que ele tenha um grau alto, mas, primordialmente, que ele esteja presente no maior número de camadas de rede possível.

Na Tabela 3.1, tem-se a aplicação das funções: W, φ, D, e R (Equações 3.2, 3.4, 3.5 e 3.6) para todos os nós em V e todas as camadas em L. Como o $\max_{i \in V} D_i = 10$, tem-se que $\varphi(W_i) = (10 + 1) \,.\, (W_i - 1)$ (Equação 3.6), e dois grupos: $G_2 = \{v_4,\ v_6,\ v_8,\ v_9\}$ e $G_3 = \{v_5\}$ (Equação 3.3 e Figura 3.2).

A classificação final R (ou sua versão normalizada $\hat{R}$) mostra o nó v_5 como o primeiro colocado, seguido por v_9, v_6, v_4 e v_8. A versão normalizada visa colocar todos os resultados dos ranqueamento num intervalo $[0, 1]$, para tento foi utilizada a metodologia $Min - Max$, onde $X_{normalizado} = \frac{X - X_{min}}{X_{max} - X_{min}}$, sendo X os valores da cada ranqueamento obtido, X_{max} o valor da maior classificação e X_{min} o valor da menor classificação. Este é o resultado esperado, pois W_{v5} atingiu o maior valor para os nós

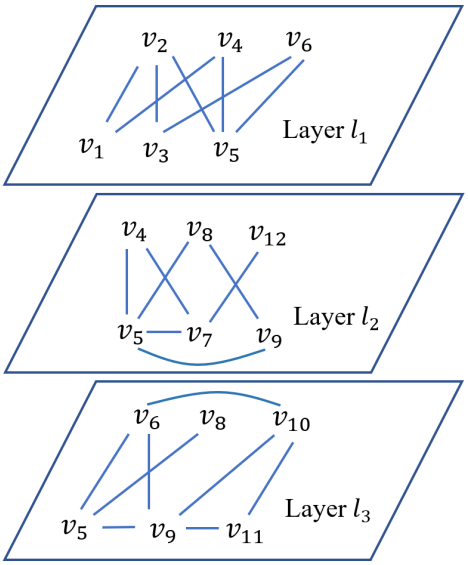


Figura 3.1. Exemplo de rede multiplex não direcionada.

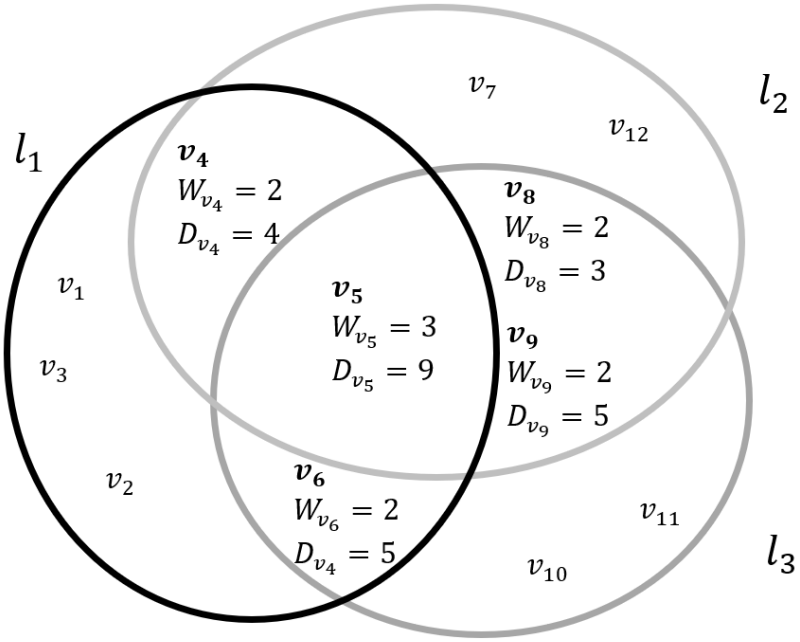


Figura 3.2. Exemplo de aplicação da centralidade GCMN para três redes N_0, N_1, N_2 com seus nós $v_1, ..., v_{12}$ e seus pesos associados $W(v_x)$.

em V. Como os outros nós têm o mesmo peso W, suas classificações são baseadas em seus graus D. Como os nós v_1, v_2, v_3, v_7, v_{10}, v_{11} e v_{12} têm peso igual a um; de acordo com a centralidade do GCMN, eles não são relevantes e não terão uma classificação.

Tabela 3.1. Exemplo de aplicação da Centralidade GCMN.

	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}	v_{12}
W	1	1	1	2	3	2	1	2	2	1	1	1
φ				11	22	11		11	11			
D				4	9	5		3	6			
R				15	31	16		14	17			
$\hat{R}$				0.468	1.000	0.500		0.437	0.531			

3.2 Algoritmo e Complexidade de Tempo de Execução

A entrada do primeiro algoritmo é uma rede complexa multiplex $N : (V, E, L)$, e a saída é um conjunto X com a classificação de todos os nós. Para facilitar a leitura do algoritmo, considera-se que a rede N é visível para as funções W e D. O algoritmo 1 constrói a classificação de centralidade GCMN. Este procedimento realiza operações primitivas com complexidade igual a $O(1)$; um loop externo sobre todos os vértices em V, com dois loops internos independentes sobre todas as arestas em E, ao aplicar as Equações 3.2 e 3.4. Portanto, o algoritmo 1 é executado em $O(|V||E|)$.

O algoritmo varre todos o nós (linha 3), utilizando as variáveis auxiliares x, z e Y (linhas 4 a 6) para armazenar os nós, arestas e camadas, respectivamente. Após inicializar o valor de $x.node$ com o nó i, presente no laço (linha 7), e zerar o valor do peso a ser calculado para esse nó (linha 8), é implementada a equação 3.2 (linhas 11 a 17), onde são percorridas todas as arestas da rede e, caso haja uma aresta conectando o nó i, é acrescentado 1 ao peso de i (linha 14) e a camada onde essa aresta esteja pesente é retirada da análise (linha 15), para que não haja uma contagem em duplicidade das arestas no cálculo do peso de i. Logo após é implementada a equação 3.4 (linhas 20 a 24), que calcula o grau do nó i. Nessa implementação são percorridas, outra vez, todas as arestas da rede e é somado 1 a cada ocorrência de i em uma aresta (linha 22). Finalmente acrescenta-se o nó i ao conjunto dos nós com seus pesos e graus calculados

X (linha 25) e o algoritmo retorna esse conjunto com seu resultado (linha 27).

Algorithm 1: Construção da base de dados para o cálculo do Ranqueamento da Centralidade GCMN

```
   Result: X // conjuntos de nós ranqueados
 1 Input: N : (V, E, L);
 2 Let X = {};
 3 for i ∈ N.V do
 4 |   Let x : (node, weight, degree)
 5 |   Let z : (source, target, layer)
 6 |   Let Y : N.L; // Y recebe o conjunto de todas as camadas da rede
   |    multiplex;
 7 |   x.node = i;
 8 |   x.weight = 0;
 9 |   // Implementacao da Equacao 3.2;
10 |   // Varrendo todas as arestas de todas as Camadas N.L, ;
11 |   for z ∈ N.E do
12 |   |   if (i=z.source ∨i = z.target) ∧ z.layer ∈ Y then
13 |   |   |   // Caso i esteja em uma das camadas àinda não consideradas;
14 |   |   |   x.weight = x.weight + 1; // acrescenta 1 ao peso de i;
15 |   |   |   Y = Y - z.layer; // retira a camada z.layer de Y para que o peso
   |   |   |    de i não seja computado duas vezes para a mesma camada;
16 |   |   end
17 |   end
18 |   // Implementacao da Equacao 3.4;
19 |   // Varrendo todas as arestas de todas as Camadas N.L;
20 |   for z ∈ N.E do
21 |   |   if i = z.source ∨ i = z.target then
22 |   |   |   x.degree = x.degree + 1; // Conta a quantidade de arestas de i;
23 |   |   end
24 |   end
25 |   X = X ∪ {x};
26 end
27 return X;
```

A entrada do segundo algoritmo é o nó i para o qual deseja-se a classificação em um conjunto X de nós, já classificados. Este conjunto X é a saída do Algoritmo 1. O algoritmo 2 recupera o índice de um nó. Assumindo uma estrutura de dados de hash

Tabela 3.2. Análise comparativa das Complexidades Algorítmicas das Centralidades.

Centralidade	Complexidade Algorítmica
GCMN	$O(\|V\|\|E\|)$
Betweenness	$O(\|V\|\|E\|)$
Eigenvector	$O(\|V\|\|E\|)$
PageRank	$O(\|V\|\|E\|)$
Weighted Degree	$O(\|V\|\|E\|)$
Closeness	$O(\|V\|\|E\|)$
Novel Multiplex PageRank in Multilayer Networks	$O(\|L\|\|V\|\|E\|)$
Cross-layer Degree Centrality	$O(\|L\|\|E\|)$

para X e a classificação proposta na equação 3.6, pode-se calcular todos os $D_z \mid z \in V$ necessários para φ enquanto popula-se X. Portanto, não é necessário atravessar a rede novamente no Algoritmo 2, essencialmente permitindo a recuperação de classificação em $O(1)$.

Algorithm 2: Recuperação do ranqueeamento de um nó

Result: $r \in \mathbb{N}$ // a classificação no nó i

1 Input: i, X:{(node, weight, degree)};

2 Let x in $X \mid i = x$.node;

3 $r = \varphi(x.\text{weight}) + x.\text{degree}$; // Equações 3.5 e 3.6

Considerando as complexidades algorítmicas das demais centralidades analisadas neste trabalho (Tabela 3.2), verifica-se que a complexidade para montagem da classificação proposta pela centralidade GCMN encontra-se entre as menores. Um fato a ser ressaltado é que, apesar de ser uma centralidade para redes multiplex, a criação da classificação independe da quantidade de camadas da rede, de forma diversa da *Novel Multiplex PageRank in Multilayer Networks* e da *Cross-layer Degree Centrality*. Para a análise das complexidades algorítmicas considera-se que L é o conjunto de camadas da rede multiplex, V representa o conjunto dos vértices e E o conjutno das arestas, ou seja, a matriz de adjacência contendo todos as arestas das redes complexas.

Com relação à recuperação das classificações de cada nó, uma vez que os raqueamentos já foram previamente calculados, se consdera que todos serão da ordem $O(1)$, como pode ser visto no algoritmo 2.

3.3 Casos Particulares e Fragilidades da Proposta

Uma vez que a centralidade GCMN tem como premissa a criação de grupos compostos pelos nós presentes em cada camada, um caso extremo seria o que todos os nós estivessem presentes em todas as camadas. Considerando essa hipótese, a hierarquia de grupos não seria aplicável e todos os nós estariam no mesmo grupo. Portanto, o único critério de classificação seria o grau do nó, o que levaria a uma classificação equivalente à da *DegreeCentrality*, discutida na seção 2.5.4. Esse caso pode ser apontado como uma fragilidade da centralidade proposta.

Outro caso a ser esclarecido é o de nós isolados sem conexões com outros nós de uma camada. Esses nós são considerados para classificação pelo GCMN, no entanto, assume-se que a probabilidade de um nó participar isoladamente em um grupo significativo de camadas é mínima. Essa suposição advém do fato de que nós isolados repesentariam entidades com pouca interação com as demais, sendo que, dessa forma, dificilmente esse tipo de nó teria uma participação mais efetiva em um número significativo de camadas que comporiam a modelagem do cenário. Naturalmente, deve-se ressalvar que cada situação fática tem suas nuances onde essa suposição pode não ser válida, entretanto essa possibilidade não foi o evidenciada nos estudos de caso deste trabalho (Capítulos 5, 6 e 7).

3.4 Considerações Finais

O capítulo apresentou a defínção formal da *Group-based Centrality for Multiplex - GCMN*, além de um breve exemplo generalista de uso da Centralidade, tomando como base a rede proposta na Figura 2.3, na seção 2.4. A seção 3.2 propôs dois algoritmos, o primeiro para formatação do ranqueamento da centralidade e o segundo para a busca da classificação de um nó.

No próximo capítulo serão expostos os materiais e métodos utilizados como prova do conceito, funcionamento e viabilidade dos resultados do estudo por meio de sua aplicação em três estudos de caso, que serão objeto dos capítulos 5, 6 e 7.

Capítulo 4

Materiais e Métodos

Este capítulo traz os conceitos comuns aos estudos de caso discutidos nos capítulos 5, 6 e 7. A organização do capítulo é a que segue:

- a seção 4.1 traz algumas premissas e formalizações que são comuns aos três estudos de caso, sendo necessárias aos seus entendimentos, tais como: a proposta de separações de nós entre os grupos propostos pela Centralidade GCMN (seção 4.1.1) para centralidades que não trabalham com o conceito de grupos, ou ainda conceitos fundamentais de estatística, tais como as heurísticas conhecidas como a análise de *Precision, Recall,* F_1 *e Accuracy* (seção 4.1.2) e dos coeficientes das correlações de Pearson e Spearman (seção 4.1.3). No caso específico dos conceitos de estatística, deu-se preferência em abordá-los nesse capítulo, a fazê-lo no capítulo 2, por se tratar de um assunto que diz respeito apenas ao entendimento dos estudos de caso e que não tem relação com a teoria das redes complexas ou com a formulação da proposta da Centralidade GCMN, objeto desta tese.

- a seção 4.2 descreve a metodologia utilizada na elaboração desde trabalho, especificando softwares, fontes de dados, bancos de dados e a metodologia de aquisição de dados para a elaboração das redes multiplex utilizadas nos estudos de caso;

- a seção 4.3 traz as considerações finais do capítulo.

4.1 Conceitos Comuns aos Estudos de Caso

Esta seção traz alguns conceitos que são comuns a todos os três estudos de caso analisados nesta tese. A seção 4.1.1 demonstra como se deu a distribuição de nós pelos grupos propostos pela Centralidade GCMN e como essa distribuição afetou a forma de

comparação de resultados do ranqueamento entre a Centralidade GCMN e as demais centralidades. A seção 4.1.2 traz a noção estatística de *Accuracy, Precision, Recall* e F_1 *Score* uma vez que essas métricas serão utilizadas na demonstração dos resultados dos estudos de caso. De forma semelhante, a seção 4.1.3 discorre acerca das correlações entre sequencias numéricas propostas por Pearson e Spearman que também serão utilizadas das discussões dos resultados obtidos.

4.1.1 Distribuição de nós por grupos

A GCMN é uma centralidade que divide os nós em grupos de acordo com sua relevância. As outras medidas de centralidade com as quais a GCMN será comparada, nos três estudos de caso a seguir (capítulos 5, 6 e 7), não têm esse conceito. Desta forma, para efeitos de comparação, será considerado o número de nós em cada grupo da GCMN e o mesmo número de nós classificados pelas demais medidas de centralidade, preservando sua classificação. A tabela 4.1 ilustra como é dado o processo de distribuição de nós nos grupos, sendo i o maior peso atingido e j o número de medidas de centralidade comparadas.

Tabela 4.1. Distribuição de nós por grupos para medidas de centralidade.

Group $G(i)$				
	Centrality 1	Centrality 2	...	Centrality j
Node 1	ranking for node 1	ranking for node 1	...	ranking for node 1
...	...	...	...	...
Node n	ranking for node n	ranking for node n	...	ranking for node n
Group $G(i-1)$				
	Centrality 1	Centrality 2	...	Centrality j
Node $n+1$	ranking for node $n+1$	ranking for node $n+1$	...	ranking for node $n+1$
...	...	...	...	...

4.1.2 *Accuracy, Precision, Recall* e F_1 *Score*

Uma maneira típica de quantificar a qualidade da classificação e agrupamentos é usando métricas como *Precision* e *Recall* (Carneiro, 2016; Hilary, 2015). A *Precision* corresponde à fração de elementos relevantes entre os elementos recuperados, enquanto *Recall* avalia a fração da quantidade total de elementos relevantes que foram recuperados. Ambas métricas ajudam a medir a relevância dos nós classificados. A F_1 *Score* é a média harmônica da *Precision* e da *Recall* e resume a qualidade do agrupamento em um valor.

Para calcular a *Precision* e o *Recall*, deve-se dividir o universo de elementos em quatro grupos: TP *the true positives* (detectados corretamente), FP - *the false positives*

(detectados incorretamente), FN - *the false negatives* (não detectados incorretamente) e TN - *true negatives* (não detectados corretamente). A *Precision* é dada por

$$precision = \frac{TP}{TP + FP} \tag{4.1}$$

e o *Recall* por

$$recall = \frac{TP}{TP + FN}. \tag{4.2}$$

A *Accuracy* indica o grau de concordância que há entre o resultado da medição e o dito valor verdadeiro (aquele que é aceito, desde que estabelecido por uma definição ou consenso) da grandeza. Ela é dada por

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4.3}$$

As métricas *Accuracy, Precision, Recall* e F_1 *Score* serão utilizadas nos três estudos de caso descritos nos capítulos 5, 6 e 7.

4.1.3 Correlações de Pearson e Spearman

O coeficiente de Pearson (Pearson, 1905) r mede o grau de correlação entre duas variáveis de escala métrica na estatística descritiva. O uso da correlação de Pearson deve obedecer ao seguinte: a escala de medição deve ser uma escala ou razão de intervalo; a distribuição das variáveis deve ser aproximadamente uniforme; a associação deve ser linear e não deve haver *outliers* nos dados.

Para entender a correlação de Spearman (Spearman, 1904) (r_s) é necessário saber o que é uma função monotônica. Uma função monotônica preserva (ou inverte) a relação de ordem. O coeficiente de correlação de Spearman pode analisar a intensidade e a direção de uma relação monotônica entre duas variáveis contínuas ou ordinais. Em um relacionamento monotônico, as variáveis tendem a se mover na mesma direção relativa, mas não necessariamente a uma taxa linear. Essas relações monotônicas podem ser estritamente crescentes ($\forall x, y \in A, (x > y \Rightarrow f(x) > f(y))$ ou decrescentes ($\forall x, y \in A, (x > y \Rightarrow f(x) < f(y))$.

Os coeficientes de correlação de Pearson e Spearman são medidas estatísticas da força de uma relação entre dados pareados e têm valores na faixa de -1 a 1. O sinal de cada coeficiente indica o sentido da relação. Se ambas as variáveis tendem a aumentar ou diminuir juntas, o coeficiente é positivo; caso contrário, se uma variável aumenta à medida que a outra diminui, o coeficiente é negativo. Assim, quanto mais perto r ou r_s se tornam de 1 ou -1, mais significativo é o relacionamento.

Dessa forma, o uso das correlações de Pearson e Spearman acrescenta uma análise relevante ao ranqueamento fornecido pelas centralidades analisadas na medida que avalia as correlações entre essas classificações e métricas propostas para os estudos de caso (Carneiro, 2016; Hilary, 2015) disponíveis nos Capítulos 5, 6 e 7.

A análise dos coeficientes de Pearson e Spearman deve também considerar a significância (valor p). Para determinar se a correlação entre as variáveis é significativa, comparando-se o seu valor com seu nível de significância. Em geral, um nível de significância (denotado como α) de 0,05 é aceitável, indicando que o risco de concluir que existe uma correlação, quando na verdade não existe correlação, é de 5%. O valor p indica se o coeficiente de correlação é significativamente diferente de 0. (Um coeficiente de 0 indica que não há relacionamento linear). Se o valor p-value $\leq \alpha$, a correlação é estatisticamente significante.

4.2 Metodologia

A centralidade proposta neste trabalho foi formalizada utilizando-se expressões regulares e a teoria dos conjuntos (Capítulo 3). Foram propostos dois algoritmos para a implementação deste modelo matemático (Seção 3.2). Foi realizada a implementação dos algoritmos propostos em linguagem de programação Java (Arnold et al., 2000) e, para persistência, foi utilizado o SGBD PostgreSQL (www.postgresql.org). Para os cálculos relativos às redes complexas, tais como: modularidade e das medidas de centralidade clássicas foi utilizado o software Gephi, versão 0.9.2 (www.gephi.org). Para as centralidades não disponíveis no Gephi, utilizou-se a linguagem Java ou a PLPGSQL, nativa do PostgreSQL, para as suas implementações. Para os cálculos das métricas *precision, recall, accuracy e* F_1 e das correlações propostas por Pearson e Spearman, foi utilizado o software livre JASP, versão 0.14.1.

Quanto à obtenção dos dados dos estudos de caso, foram utilizadas as seguintes fontes:

- capítulo 5 — Operação Lava Jato — redes complexas utilizadas em Almeida et al. (2017), cedidas pelos autores. Nesse caso foram utilizadas as redes complexas prontas, conforme cedidas, para a criação da rede multiplex, sendo que cada rede complexa relativa a cada testemunho foi utilizada como uma camada da rede multiplex. A base de dados do estudo de caso encontra-se disponível em https://data.mendeley.com/datasets/28xd6jz46j/draft?aD 726 45e496ce-b8a8-4601-9ffa-c487526a6327;

- capítulo 6 — Operação Licitante Fantasma — extração de dados por meio de uma API disponibilizada pelo Governo Federal (http://compras.dados.gov.br/docs/home.html). Nesse caso a rede multiplex foi formada a partir das informações colhidas, contendo 1467 nós, respresentando as empresas que participaram em licitações na área de saúde, sendo cada ano de licitações relativo a cada camada da rede, proporcionando uma análise temporal dos dados. A base de dados do estudo de caso encontra-se disponível em https://data.mendeley.com/datasets/s8p55kp2dp/draft?a=922eefee-28ed-40ea-b9f4-fda46123c08e;

- capítulo 7 — Personagens da Saga Harry Potter — redes complexas utilizadas em Carvalho (2017), cedidas pelos autores. A exemplo do estudo de caso do Capítulo 5, foram utilizadas as redes complexas prontas, conforme cedidas, para a criação da rede multiplex, sendo que cada rede complexa relativa a cada livro. A base de dados do estudo de caso encontra-se disponível em https://data.mendeley.com/datasets/sxhgk7m4tg/draft?a=0c022f3b-7ba3-4a8db44d- d437fbf29033

Todos os algoritmos, scripts e estrutura de dados utilizada estão disponíveis mediante solicitação ao autor.

4.3 Considerações Finais

Este capítulo apresentou alguns conceitos comuns aos estudos de caso descritos nos capítulos 5, 6 e 7 tais como: a proposta de separações de nós entre os grupos propostos pela Centralidade GCMN para centralidades que não trabalham com o conceito de grupos, ou ainda conceitos fundamentais de estatística. Todas essas premissas são comuns aos estudos de caso e não faria sentido repeti-las estudo a estudo, dessa forma elas foram sintetizadas na seção 4.1. No caso específico dos conceitos de estatística, deu-se preferência em abordá-los nesse capítulo, a fazê-lo no capítulo 2, por se tratar de um assunto que diz respeito apenas ao entendimento dos estudos de caso e que não têm relação com a teoria das redes complexas ou com a formulação da proposta da Centralidade GCMN, objeto desta tese.

Além disso, foi descrita a metodologia utilizada (seção 4.2) na obtenção das fontes de dados e tratamento desses dados. Nessa seção são relacionados os softwares, linguagens, bancos de dados e fontes de informação utilizados na elaboração deste trabalho.

Nos próximos três capítulos (capítulos 5, 6 e 7), serão descritos três estudos de caso onde a centralidade proposta nesta tese é aplicada e comparada a outras centralidades, tanto as ditas "clássicas" quanto centralidades multiplex.

Capítulo 5

Estudo de Caso - Operação Lava Jato

5.1 Introdução

O estudo de caso trata de uma investigação de corrupção brasileira, denominada Operação Lava Jato. Os seus resultados foram preliminarmente publicados em De Figueirêdo et al. (2021), dessa forma, para que se haja coerência com os resultados obtidos na publicação, algums gráficos e tabelas serão mantidos em inglês.

A operação teve início em 2009, pela Polícia Federal do Brasil, que investiga a prática de crimes financeiros e desvio de dinheiro público. Para incentivar a colaboração de criminosos nas investigações, o Ministério Público Federal firmou acordos de leniência, ou seja, em troca de informações criminais úteis à investigação por benefícios aos criminosos condenados que puderam ter as suas penas reduzidas ou até extintas. Nesse contexto, as informações dos testemunhos foram vitais para a identificação de entidades influentes (indivíduos), auxiliando as autoridades responsáveis na aplicação da lei e no direcionamento de suas investigações.

O estudo de caso trata de cinco testemunhos prestados por criminosos condenados pela Operação Lava Jato (Almeida et al., 2017). Esses testemunhos explicam em detalhes o mecanismo de corrupção que existe nos mais altos escalões da política no Brasil. Os condenados tiveram suas penas reduzidas ou obtiveram outros benefícios como o cumprimento da pena de prisão domiciliar, como recompensa por colaborar com a justiça.

Como a centralidade do GCMN lida com redes multiplex, a escolha natural foi dividir os cinco testemunhos em cinco camadas (L) de uma rede multiplex, em que

os nós (V) se referem aos indivíduos citados por cada condenado, e as arestas (E), a ocorrência conjunta desses indivíduos nos trechos dos testemunhos (Equação 3.1).

A próxima etapa foi determinar o peso (W) de cada nó (Equação 3.2), e criar os grupos (G) de nós (Equação 3.3) de acordo com seus pesos. Como resultado, foram obtidos os seguintes valores: $\mid G_2 \mid = 62$, $\mid G_3 \mid = 29$, $\mid G_4 \mid = 16$, and $\mid G_5 \mid = 5$. É importante observar que, como a rede é composta por cinco camadas, o número de grupos considerados relevantes é quatro, ou seja, apenas os grupos com nós i cujo peso é $W_i \geq 2$, conforme formalizado na introdução do capítulo 3. A última etapa consistiu em aplicar as Equações 3.4, 3.5 e 3.6, a todos os nós dos grupos G_2 a G_5 para obter o ranqueamento da centralidade GCMN. No caso específico deste estudo de caso, o $\max_{z \in V} D_z = 95$, então, a função φ será $\varphi(W_i) = (95 + 1) \,.\, (W_i - 1)$ (Equação 3.6).

Vale ressaltar que, apesar de tratar-se de uma operação que envolve um forte viés político-partidário, em especial pelo momento de acirramento do embate entre a esquerda e a direita no Brasil. A análise feita se porta de uma forma absolutamente isenta, sem ter sido contaminada por nenhuma convicção doutrinária.

As subseções a seguir comparam a classificação fornecida pela Centralidade GCMN com as classificações das medidas de centralidade clássicas: *weighted degree* (WD)(Beveridge e Shan, 2016), *Betweenness* (BC)(Freeman, 1978; Otte e Rousseau, 2002),*Eigenvector* (EV)(Bonacich, 1972) e *Closeness* (CL) (Freeman, 1978); do o ANN SCORE que representa a média geométrica normalizada das métricas (BC, EV e WD) (Almeida et al., 2017) e duas medidas de centralidade multicamadas: *Cross-layer degree centrality* (CLDC)(Bródka et al., 2012) e a *The Multiplex PageRank* (MPR)(Tu et al., 2018).

5.2 Trabalhos Relacionados

O estudo de caso aborda um tipo específico de rede chamado redes multiplex, um tipo específico de rede multicamadas em que cada nó aparece em camadas diferentes e cada camada descreve todas as bordas de um determinado tipo. Esses nós não podem ter conexões com outros nós em outras camadas, e uma matriz tridimensional de tamanho $(V \times V) \times L$, na qual V representa os vértices (nós), e L as camadas, ou dimensões (Mucha et al., 2010; Nicosia et al., 2013). O uso de medidas de centralidade padrão (Freeman, 1978; Otte e Rousseau, 2002; Bonacich, 1972, 2007; Beveridge e Shan, 2016) teve que ser revisado para cobrir as múltiplas camadas desta nova estrutura de rede.

Essa revisão levou à proposição de extensões das medidas clássicas de centralidade como sendo um caminho natural a ser seguido. Algumas propostas para adaptar

essas medidas de centralidade surgiram, como *Novel Multiplex PageRank in Multilayer Networks* (Tu et al., 2018), *Random walk centrality in interconnected multilayer networks* (Solé-Ribalta et al., 2016) e *Random Walks on Multiplex Networks: Supplementary Information for Navigability of Interconnected Networks under Random Failures* (De Domenico et al., 2014), por exemplo. Entretanto, este tipo de extensão das centralidades clássicas, embora válido, não apresenta novas estratégias de classificação, sendo adaptações de estratégias anteriores, para lidar com redes de multiplas camadas.

Outras estratégias surgiram com um foco específico nessas novas estruturas de rede multicamadas, como *The CLDC* (Bródka et al., 2012), por exemplo. Na medida de centralidade CLDC, a classificação do nó x é calculada como uma razão entre o número de nós conectados ao nó x e o número total de todos os nós na rede (reduzido em um). Assim, o CLDC é uma soma dos pesos das bordas de entrada e saída do nó x em direção à vizinhança de várias camadas dividida pelo número de camadas e o número total de membros da rede. O CLDC é uma centralidade de camada cruzada que considera as bordas de entrada e saída através das camadas de rede para classificar os nós.

As medidas de centralidade em geral podem ter desvantagens ao apontar os nós mais relevantes de uma rede complexa em cenários de decisão de múltiplos critérios, onde as arestas podem representar diversos tipos de relacionamento, inclusive podendo utilizar multiplas camadas para seccionar essa tipificação. Estratégias baseadas em grupo mostraram-se uma opção viável nessas situações. Um estudo de caso apresenta uma abordagem baseada em k-means para classificar as empresas de capital de risco no mercado de investimento chinês como uma estratégia alternativa. A abordagem apresenta critérios de avaliação baseados em grupo para classificar os nós (Yang et al., 2020). Outra proposta baseada em agrupamento diz respeito a uma forma complementar à modelagem de blocos generalizada para decomposição hierárquica, usando o método k-means para decompor uma rede social em grupos de nós tendo, como critério, a existência de perfis congruentes de dissimilaridades com outros nodes (Hsieh e Magee, 2010). Portanto, propostas baseadas em grupos de nós, com uma hierarquia associada, podem ajudar a classificar os nós de acordo com sua relevância, classificando aqueles presentes no núcleo das redes como os de maior interesse na pesquisa e apontando os nós associados da periferia da rede como os de menor interesse (Borgatti e Everett, 2007).

A descoberta de entidades de acordo com a sua relavância é objeto de interesse de diversas áreas de pesquisa na computação. A detecção de entidades que sejam caracterizadas como *outliers* (pontos fora da curva) em uma coleção de dados é um dos problemas considerados basilares e, consequentemente, mais discutidos em mineração

de dados. Um *outlier* pode ser entendido como uma entidade que, pelas suas características, difere das outras em um conjunto de dados. A detecção de *outliers* é útil na descoberta de dados imprevisíveis e não identificados em áreas específicas, como a detecção de fraudes no uso de cartões de crédito, cartões telefônicos, intrusão de computadores e comportamento criminoso, entre outros (Bansal et al., 2016).

Uma outra abordagem na identificação de *outliers* faz uso de lógica fuzzy para detecção de fraudes no uso de serviços públicos. Trata-se de um método de identificação de consumidores que adulteram a leitura de medidores de consumo de energia com a intenção de ter suas contas reduzidas (Medvedeva e Komotskiy, 2016). Dessa forma, a identificação desses potencais fraudadores indica quais consumidores têm maiores chances de estar cometendo ilícitos. Esse processo auxilia a auditoria, uma vez que seria inviável a auditagem de todos os consumidores de uma empresa fornecedora de energia elétrica, pela sua grande quantidade.

Estudos de detecção de fraudes utilizam técnicas de mineração de dados aliadas a redes bayesianas na tarefa de classificação de grupos de risco e de criação de árvores de decisão, para a criação de modelos que descrevam cada um desses grupos. Em geral, esses tipos de abordagem geram um modelo baseado em dados históricos e utilizam esses modelos na classificação de novas instâncias. Essa estratégia permite uma análise precisa de dados de treinamento, mas se torna inviável para lidar com uma grande entrada de dados contínua (Bhowmik, 2008). Essa dificuldade em lidar com novos fluxos de dados é uma limitação no uso da mineração de dados aliadas a redes bayesianas em casos que exigem detecção imediata de fraude. Uma das premissas das centralidades em geral, e a GCMN não é uma excessão à regra, é a do cálculo do ranqueamento dos nós de forma estática, sendo assim, a questão de lidar como novos fluxos de dados também se torna um problema para a tecnologia.

A centralidade GCMN combina aspectos de várias medidas de centralidade. É uma nova estratégia, baseada no agrupamento de nós, para classificar nós em redes multicamadas. O agrupamento de nós ocorre hierarquicamente permitindo a classificação baseada em grupos. Assim, a centralidade GCMN pode apontar os nós mais relevantes de uma rede multiplex em cenários de decisão que envolvam múltiplos critérios. Além disso, a GCMN, como as demais centralidades, age como um detector de *outliers* quando detecta os nós mais revelantes apontando para aqueles que se destacaram.

5.3 Métrica de avaliação proposta

Considerando a distribuição de nós por grupos proposta na seção 4.1.1 e aplicando a classificação ao estudo de caso, tem-se que, para o grupo G_5, com cinco nós, foram considerados os cinco primeiros nós mais bem classificados em todas as outras medidas de centralidade. Desta forma, foi possível comparar o grupo G_5 com os nós melhor classificados em todas as outras medidas de centralidade. O grupo G_4 possui dezesseis nós, e foram considerados os nós classificados entre seis e vinte e um em todas as outras medidas. Este processo foi o mesmo para os grupos G_3 e G_2.

Para efeito de análise, um parâmetro relevante é a situação legal de cada indivíduo apontado nos testemunhos. Portanto, é coerente a proposta de uma métrica que considere essas situações legais como parâmetro para medir o sucesso ou não da classificação proposta por cada centralidade. Em outras palavras, quão grave foi a situação legal de um indivíduo mais importante ele será considerado na análise.

Desta forma, a relevância dos grupos será aferida de acordo com as situações legais dos indivíduos apontados pelas classificações de cada centralidade, verificando se os grupos mais relevantes conseguiram apontar os indivíduos com situações legais de maior severidade. Lembrando que a premissa é que o valor escalar que indica o peso (W_i) de um nó i (Equação 3.2) é proporcional à sua importância, e que, consequêntemente, um grupo G_w deve ter nós mais relevantes do que um grupo G_{w-1} (Equação 3.3).

No cenário Lava Jato, serão considerados cinco valores possíveis para as situações legais S de um nó v, como $S(v)$ onde

$$S : s \rightarrow \exists! s | s \in \{NotInvestigated, Investigated, Denounced, Defendant, Convicted\}, \tag{5.1}$$

cada um com sua característica e grau de importância.

- $NotInvestigated$: não houve investigação, ou a investigação concluiu que o indivíduo era inocente, sendo a menor classificação da nossa escala proposta, englobando os chamados falsos positivos;

- $Investigated$: há investigação em andamento, mas ainda sem resultado. A análise não pode considerar esses indivíduos como relevantes;

- $Denounced$: o Ministério Público apresentou uma denúncia formal e os indivíduos são relevantes para a análise;

- *Defendant*: houve o acatamento da denúncia e o indivíduo irá a julgamento por algum crime relacionado à Operação Lava Jato. Esse indivíduo tem alta relevância para a análise;

- *Convicted*: Houve um julgamento formal e o indivíduo foi condenado por crimes relacionados à Operação Lava Jato, sendo o grupo mais relevante para a análise.

De acordo com o rito legal do processo de indiciamento de um indivíduo no Brasil, desde a investigação até a eventual condenação, é composto das seguintes situações legais: $NotInvestigated$ e $Investigated$ que referem-se apenas a suspeitas policiais; já as situações: $Denounced$, $Defendant$, $Convicted$; envolvem a participação de um promotor de justiça e/ou um juiz de direito. Assim, entende-se como razoável dividir as cinco situações em dois conjuntos distintos: *most relevant legal status* — $MRS = \{Denounced, Defendant, Convicted\}$ — e *least relevant legal status* — $LRS = \{NotInvestigated, Investigated\}$. Assim, a relevância do nó é classificada como MRS ou LRS, de acordo com sua situação legal, formalmente definido como $S(v)$ (Equação 5.1).

Foi definida uma métrica chamada *Relevance Index* de um Grupo para determinar a precisão dos resultados de um grupo como

$$RI(G(n)) = \frac{|\{v_i | S(v_i) \in MRS \wedge v_i \in G(n)\}|}{|G(n)|} \times 100, \tag{5.2}$$

que corresponde à porcentagem de nós desse grupo cuja situação legal pertence à MRS, onde G_n é o grupo em análise e v_i é um nó que pertence a G_n.

Como uma extensão do *Relevance Index*, calcula-se a *general relevance*, definida como:

$$GR = \frac{|\{v_i | S(v_i) \in MRS|}{|\{v_j | S(v_j) \in (MRS \cup LRS)\}|}, \tag{5.3}$$

isto é, a soma dos nós no MRS de G_2 a G_5, dividido pelo total de nós, ou seja $MRS \cup LRS$.

Esta abordagem pretende atestar o desempenho da centralidade do GCMN na determinação dos nós mais relevantes numa rede multiplex e a eficácia do peso como critério de classificação desses nós. A coerência desse critério será demonstrada verificando-se que o crescimento do peso é diretamente proporcional à relevância dos nós selecionados.

Passa-se agora a comparar os resultados da aplicação da centralidade GCMN com as centralidades utilizadas em Almeida et al. (2017): a *Betweenness centrality*

(BC), *Eigenvector* (EV), *Weighted Degree* (WD) e a ANN SCORE definida como $\sqrt[3]{BC \times EV \times WD}$; demostrando que a centralidade GCMN é capaz apontar os nós mais relevantes de forma mais efetiva do que aquelas medidas de centralidades.

Complementarmente, foi adicionada à analise uma outra centralidade clássica, a *closeness* (CL), e duas medidas mais recentes: a *Cross-layer degree centrality* (CLDC)(Bródka et al., 2012) e a *Multiplex PageRank* (MPR)(Tu et al., 2018). O acréscimo dessas últimas duas medidas deu-se afim de tornar a análise mais precisa uma vez que as centralidades foram projetadas para redes multiplex, como é o caso da GCMN.

5.4 Resultados e Discussão

Os resultados e sua discussão considerarão dois parâmetros: o uso de métricas conhecidas como *Accuracy, Precision, Recall and* F_1 *Score* (seção 5.4.5) e o uso da métrica proposta de forma específica para este estudo de caso (seção 5.3) que avalia o desempenho da GCMN em três aspectos: "O peso como um parâmetro de agrupamento" (Subseção 5.4.1), "A Relevância por grupo" (seções 5.4.2 e 5.4.3) e a "Análise qualitativa dos grupos G_3 a G_5" (seção 5.4.4).

A Figura 5.1 traz a rede complexa original, relativa à junção dos cinco testemunhos, cedida pelos autores de Almeida et al. (2017). Nela pode-se verificar visualmente a existência de núcleos de pessoas, sugerindo associações de indivíduos. A escolha natural para a segmentação desses testemenhos foi a alocação de cada um deles como um camada de uma rede multiplex não direcionada, sendo essas camadas relativas aos depoimentos dos seguintes condenados: Paulo Roberto Costa, Alberto Youssef, Nestor Cervero, Delcídio do Amaral e Cláudio Melo; contando, cada uma, com: 152, 138, 125, 178 e 128 nós, respectivamente; onde cada nó refere-se aos indivíduos mencionados nos testemenhos e as arestas à ocorrência conjunta desses indivíduos nos trechos de depoimentos.

A Figura 5.2 aponta os nós em comum entre duas das cinco camadas da rede multiplex, representando os depoimentos de Paulo Roberto Costa e Nestor Cerveró, essa quantidade serve como base para o cálculo do peso dos nós (equação 3.2) e na consequente formação dos grupos (equação 3.3). A Figura 5.3 traz uma visão ampla da rede expondo a quantidade de nós em comum entre as cinco camadas da rede multiplex, sendo que cada uma corresponde a um dos testemunhos analisados. É possível observar que todas as cinco camadas têm nós em comum com todas as demais camadas. Essa análise demonstra que há um grupo de suspeitos presente em todos

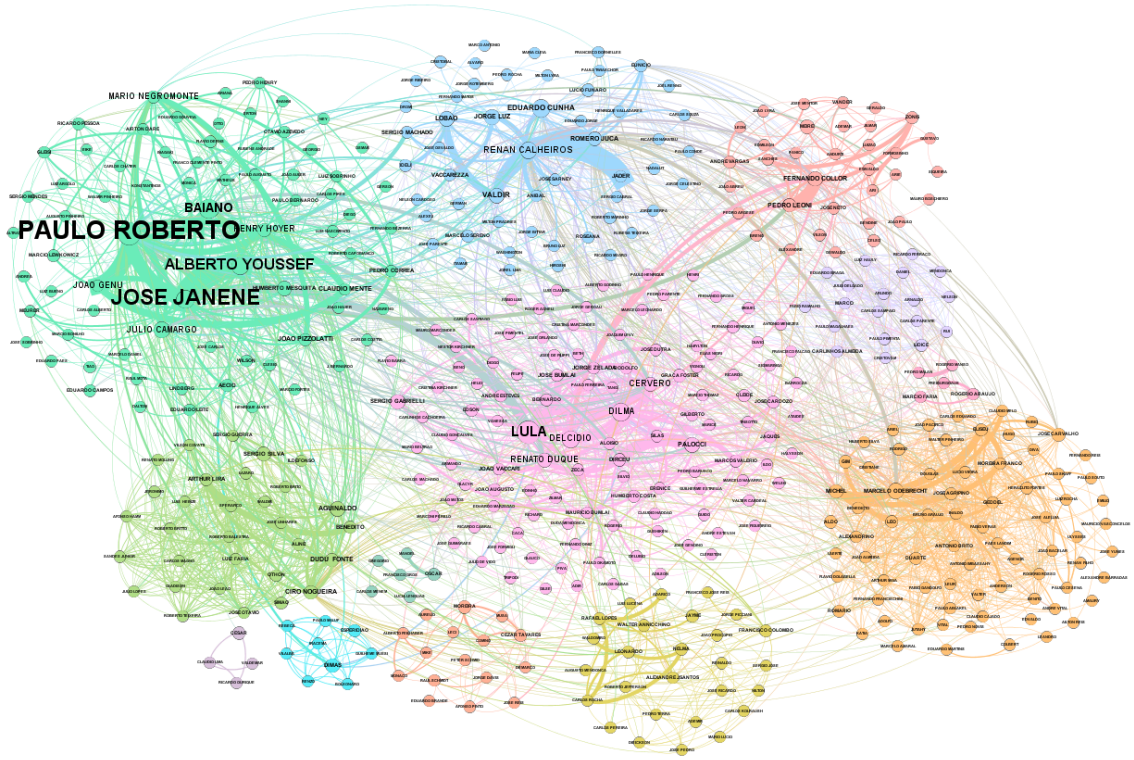


Figura 5.1. Rede complexa do estudo de caso: Operação Lava Jato.

os depoimentos, pertencentes ao grupo G_5, relacionados na Tabela 5.1. Além disso, verifica-se que a quantidade de indivíduos em comum entre as camadas é similar, isso demonstra a coerência dos testemunhos por citarem, conjuntamente, uma quantidade significativa de indivíduos suspeitos em comum.

O objetivo da centralidade GCMN em particionar essa rede original em camadas e definir grupos de nós de acordo com essa divisão, é garantir a consistência e efetividade do uso dessa centralidade ao longo dos testemunhos, excluindo qualquer possível super-dimensionamento indevido de relevância, proporcionando um melhor ranqueamento. A divulgação dos nomes mais relevantes na análise (Tabela 5.1) objetiva apenas o relato do resultado do ranqueamento da centralidade GCMN. É importante deixar claro que não há nenhum juízo de valor acerca da conduta ou culpabilidade dos indivíduos nela relacionados.

5.4.1 O peso como um parâmetro de agrupamento

A Figura 5.4 mostra a distribuição normalizada dos indivíduos por situação legal e grupo, considerando o agrupamento proposto pela centralidade GCMN. Percebe-se que o número de indivíduos na situação legal mais relevante é consistente com o grau

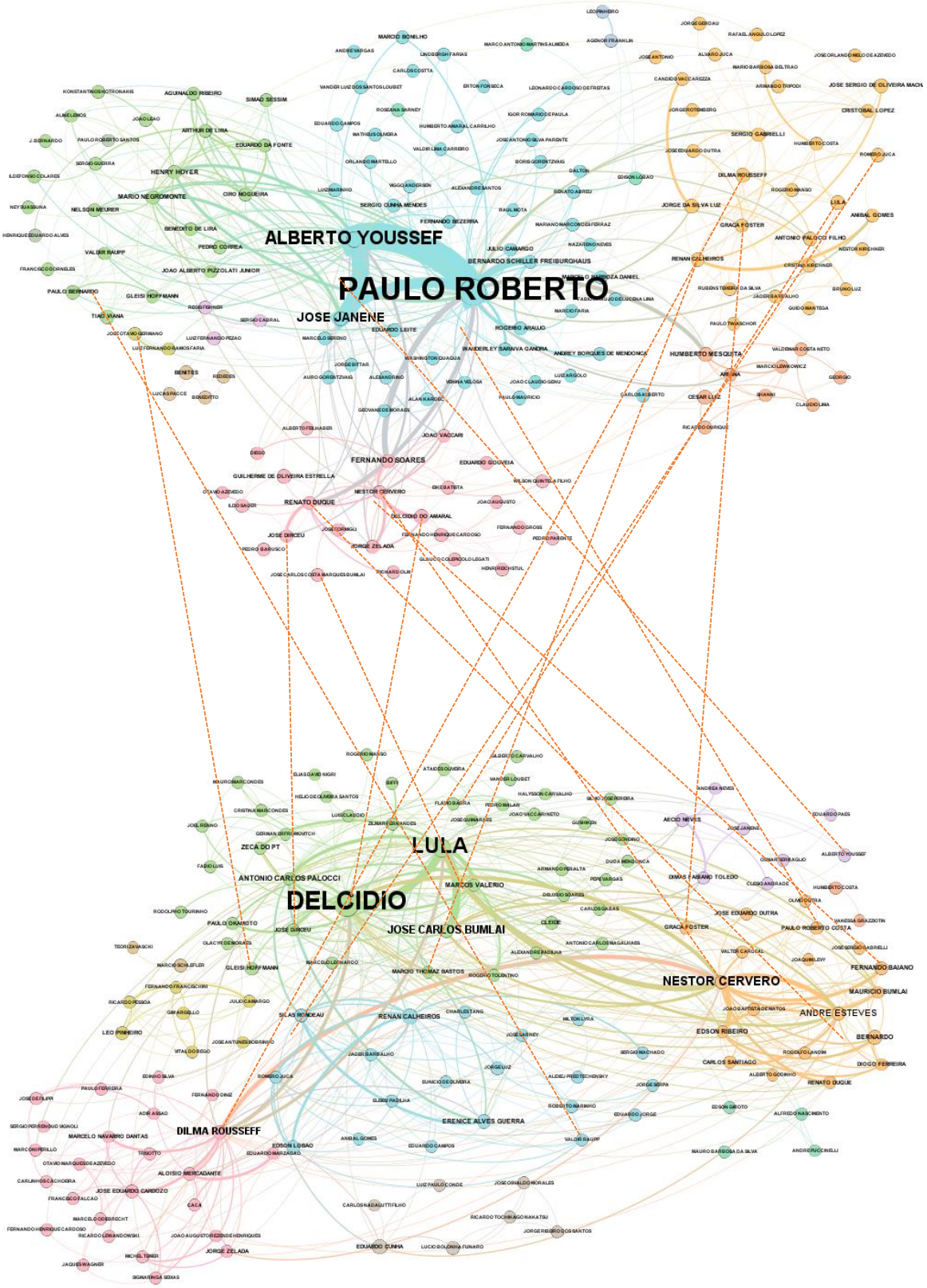


Figura 5.2. Nós em comum entre duas camadas da Rede Multiplex, representando os depoimentos de Paulo Roberto Costa e Nestor Cerveró (Estudo de Caso - Lava Jato).

Tabela 5.1. Ranqueamento de Suspeitos da Centralidade GCMN - Operação Lava Jato.

Grupo G_5	Grupo G_4
RENAN CALHEIROS	PAULO ROBERTO COSTA
LUIZ INACIO LULA DA SILVA	DELCIDIO DO AMARAL
DILMA ROUSSEFF	NESTOR CERVERO
ANTONIO CARLOS PALOCCI	JOSE JANENE
GRACA FOSTER	ROMERO JUCA
	EDUARDO CUNHA
	JULIO CAMARGO
	RENATO DUQUE
	EDSON LOBAO
	VALDIR RAUPP
	JORGE LUZ
	JOSE DIRCEU
	JOSE CARLOS BUMLAI
	MARCIO FARIA
	LEO PINHEIRO
	VANDER LUIS DOS SANTOS LOUBET

de relevância dos grupos de G_2 a G_5.

Analisando os resultados de cada grupo e considerando o *Relevance Index* (Equação 5.2), tem-se a seguinte distribuição por grupos de nós:

- G_2 (62 nós): este grupo traz 18 indivíduos em LRS ($NotInvestigated$ e $Investigated$) e 44 em MRS ($Denounced$, $Defendant$ e $Convicted$). O *Relevance Index* do grupo é 71%;

- G_3 (29 nós): aplicando os mesmos critérios do grupo anterior, encontra-se um *Relevance Index* de 86%. É importante enfatizar que existem apenas quatro indivíduos com o status menos relevante (LRS);

- G_4 (16 nós): o *Relevance Index* cresce novamente para 94%. Neste grupo, destaca-se a inexistência de indivíduos com condição de $Investigated$ ou $Denounced$. Assim, exceto por um indivíduo, todos os indivíduos do grupo têm o status $Defendant$ ou $Convicted$;

- G_5 (5 nós): o *Relevance Index* do grupo atinge a porcentagem máxima de 100%. Isso significa que todos os indivíduos encontrados são relevantes para a análise.

Tabela 5.2. Análise numérica de relevância entre grupos.

Group	Relevance Criterion	BC	EV	WD	ANN SCORE	GCMN	CLDC	MPR	CL
G_2	MRS	32	35	27	33	44	32	28	28
	LRS	30	27	35	29	18	30	30	26
G_3	MRS	20	26	26	23	25	24	14	24
	LRS	9	3	3	6	4	5	14	4
G_4	MRS	15	14	15	15	15	14	14	13
	LRS	1	2	1	1	1	2	2	3
G_5	MRS	4	4	4	4	5	4	5	5
	LRS	1	1	1	1	0	1	0	0

Após essa análise, verifica-se que o crescimento do peso associado aos nós é diretamente proporcional ao grau de severidade relacionado ao seu estado legal. Esse fato demostra que o peso é uma excelente escolha como parâmetro para construir grupos.

5.4.2 Relevância por Grupo

A Figura 5.5 compara a distribuição de nós por grupo e a situação legal dos indivíduos. Análise considera a centralidade GCMN e as demais centralidades presentes em Almeida et al. (2017), além de três outras centralidades (CLDC, MPR e CL).

Considerando o número de nós de cada grupo (G_2 a G_5) e analisando-os pelo critério de relevância (LRS e MRS), percebe-se que a centralidade GCMN tem um desempenho geral superior quando comparada às demais medidas de centralidade. A centralidade GCMN alcançou os melhores resultados nos grupos G_2 e G_5, empatou com as centralidades BC, WD e ANN SCORE no grupo G_4, e perdeu das medidas EV e WD, por apenas um indivíduo, no grupo G_3 (Tabela 5.2). Esse resultado demonstra que quando a centralidade GCMN perde, como o caso do grupo G_3, o faz por uma pequena margem, o mesmo ocorre quando supera as demais centralidades no grupo G_5, isso indica uma equivalência entre os resultados, bem como uma consistência entre eles. A única exceção é o grupo G_2, onde a GCMN teve uma performance bem superior às demais, entretanto, esse resultado não muito significativo, uma vez que o grupo é o de menor relevância.

A centralidade GCMN atingiu o melhor desempenho no *Relevance Index* (Equação 5.2) nos grupos G_2, G_4 e G_5; e o segundo melhor no grupo G_3. A centralidade

Tabela 5.3. Análise do *relevance index* (Equação 5.2) e *general relevance* (Equação 5.3).

Group	BC	EV	WD	ANN SCORE	GCMN	CLDC	MPR	CL
G_2	52	56	44	53	71	52	48	52
G_3	69	90	90	79	86	83	50	86
G_4	94	88	94	94	94	88	88	81
G_5	80	80	80	80	100	80	100	100
General Relevance	63	71	64	67	79	66	57	68

GCMN também atingiu uma *general relevance* superior, ou seja, 80 contra 68 da segunda métrica melhor classificada, a CL (Equação 5.3) (Tabela 5.3). Como critério de agrupamento, a utilização do peso teve influência não só para a centralidade GCMN, mas também para outras centralidades cuja relevância foi, em geral, consistente com esta abordagem (Tabela 5.3). A única exceção significativa a esta regra foi o grupo G_5. Porém, deve-se considerar o número reduzido de nós neste grupo, o que resulta em uma significância maior no *Relevance Index* por elemento.

Também é importante observar que no grupo G_5 os resultados da centralidade GCMN são consistentes, uma vez que não há nós relacionados às duas situações legais menos relevantes ($NotInvestigated$ e $Investigated$), o que ocorre em BC, EV, WD, ANN SCORE e CLDC.

5.4.3 Relevância por Grupo, uma análise cumulativa

A Figura 5.6 traz a distribuição de nós por grupo e situações legais cumulativamente. Desta forma, o grupo G_5 é o mesmo da Figura 5.5, e, para os demais grupos, percebe-se o aumento da relevância de uma forma gradual e cumulativa, proporcionando uma análise da evolução da relevância de forma agrupada e uma visão geral dos resultados. Assim como na Seção 5.4.2, foi feita uma comparação entre as medidas de centralidade e a centralidade GCMN, com os resultados destacados na Tabela 5.4.

Para todos os grupos, o número de indivíduos da centralidade GCMN em LRS é o menor e, consequentemente, o número de nós em MRS é o maior (Tabela 5.4). A centralidade GCMN apresenta o maior número de nós associados à situação legal mais relevante ($Convicted$) do grupo G_2 ao G_4. Apesar do grupo G_5 ser uma exceção, não se trata de um problema ou deficiência da GCMN considerando que o grupo possui apenas cinco nós e que apenas um elemento equivale a 20% de todo o grupo, reduzindo a possibilidade de uma análise estatística mais precisa.

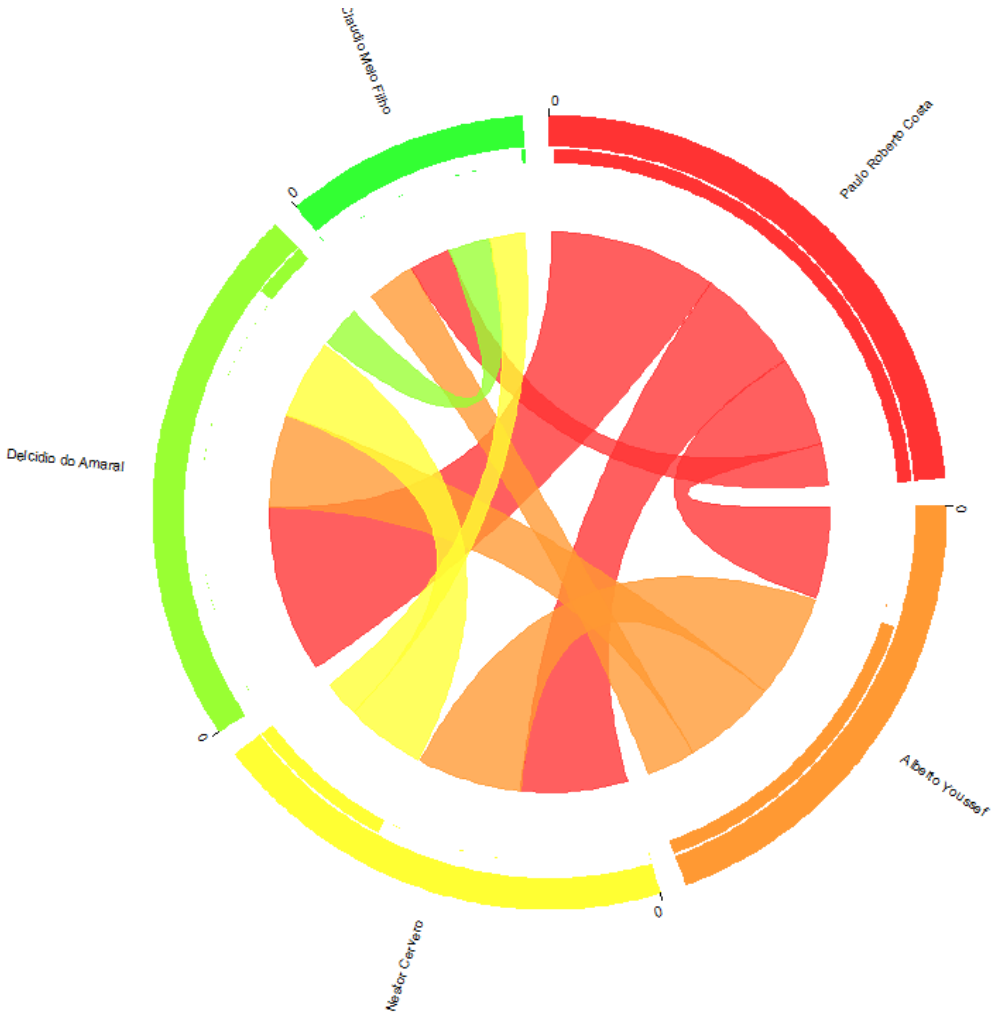


Figura 5.3. Quantidade de nós em comum entre as camadas da Rede Multiplex (Estudo de Caso - Lava Jato).

Tabela 5.4. Distribuição de indivíduos por situação legal e medidas de centralidade.

Legal Status	BC	EV	WD	ANN SCORE	GCMN	CLDC	MPR	CL
Convicted	27	31	24	31	32	22	23	26
Defendant	24	25	24	25	28	24	21	25
Denounced	20	23	24	19	29	28	17	19
Total	71	79	72	75	89	74	61	70
Investigated	7	4	13	8	4	11	13	8
NotInvestigated	34	29	27	29	19	27	33	25
Total	41	33	40	37	23	38	46	33

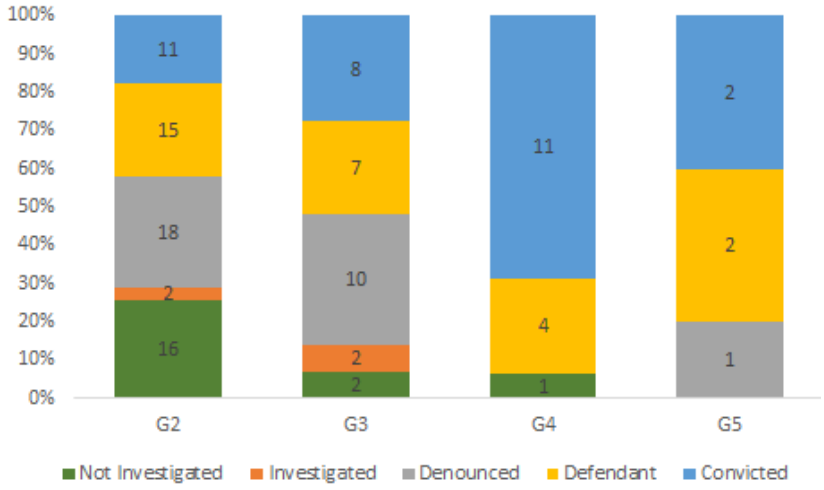


Figura 5.4. O peso como um parâmetro de agrupamento de nós.

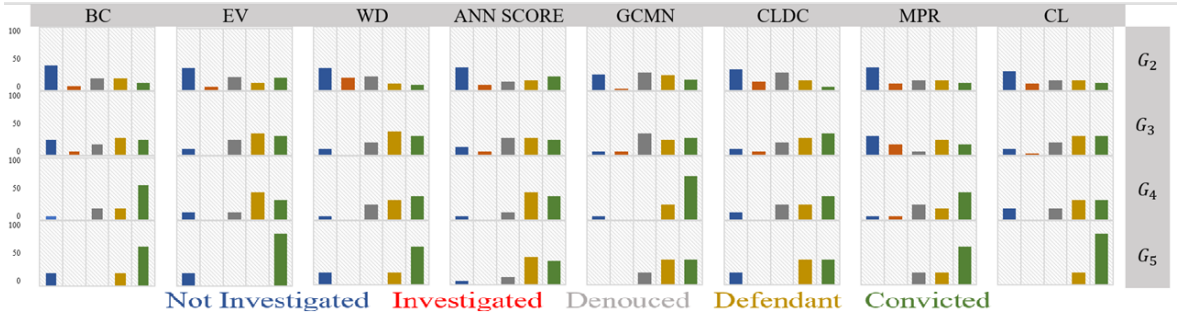


Figura 5.5. Análise comparativa do número de nós por grupo/situação legal.

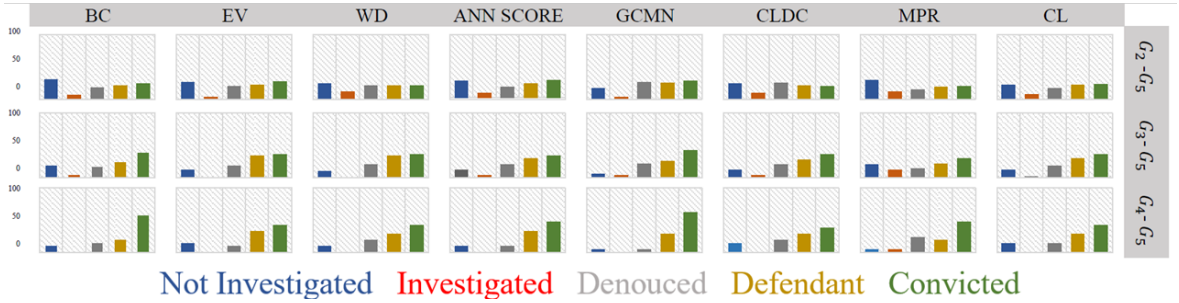


Figura 5.6. Análise comparativa cumulativa do número de nós por grupo/situação legal.

5.4.4 Análise qualitativa dos grupos G_3 a G_5

A análise qualitativa pretendeu verificar a capacidade de cada métrica em apontar "novidades". Esta análise é particularmente útil para a descoberta de nós até então não detectados.

A Tabela 5.5 mostra que a centralidade GCMN foi capaz de apontar dezessete

Tabela 5.5. Análise qualitativa dos grupos G_3 a G_5 (MRS).

	BC	EV	WD	GCMN	CLDC	MPR	CL
LRS	6	1	1	4	4	3	1
MRS	4	3	4	13	8	8	3
Total	10	4	5	17	12	11	4

nós não detectados pelas outras medidas de centralidade, ou seja, 42% a mais que a segunda colocada, a Centralidade do CLDC. Além disso, treze dos dezessete nós apontados tinham as situações jurídicas mais relevantes (MRS). Comparando este resultado com todas as outras centralidades (cinquenta e cinco nós), verifica-se que a centralidade GCMN apontou para 31% dos nós não detectados por outras centralidades e, ao mesmo tempo, relevantes.

Atribui-se esse resultado ao uso, pela GCMN, de uma estratégia completamente diferente daquelas usadas por medidas de centralidade como *Betweenness* (BC), *Eigenvector* (EV), *Closeness* (CL) e *Weighted Degree* (WD) que, em algum ponto, têm premissas baseadas em conceitos semelhantes (Bonacich, 1972, 2007; Otte e Rousseau, 2002; Freeman, 1978; Beveridge e Shan, 2016). Como esperado, a centralidade CLDC e MPR alcançaram um resultado superior às centralidades clássicas, uma vez que sua classificação foi baseada em uma estratégia de múltiplas camadas.

É fundamental esclarecer que a ausência de novos nós no ANN SCORE é o resultado esperado, uma vez que ela representa a média geométrica normalizada das outras três métricas (BC, EV e WD) (Almeida et al., 2017).

5.4.5 *Accuracy, Precision, Recall* e F_1 *Score*

Considerando o exposto na seção 4.1.2 e tomando o estudo de caso, considera-se os indivíduos detectados em MRS como $thetruepositives$; os indivíduos detectados em LRS como $thefalsepositives$; o não detectado em MRS como $thefalsenegatives$; e o não detectado em LRS como $thetruenegatives$ (Tabela 5.6). Essa decisão de montagem da "tabela de confusão" teve como base critérios de relevância dos indivíduos e a sua detecção, ou não, pela centralidade. Dessa forma, aqueles indivíduos que pertencem ao MRS foram considerados como os mais relevantes ou, positivos, sendo esse um conjunto disjunto daqueles pertencentes ao LRS, dos indivíduos menos relevantes, ou negativos. Os indivíduos detectados ou não, foram considerados os como os verdadeiros ou falsos, respectivamente, uma vez que esse critério determina se a centralidade em análise foi capaz de apontá-lo. Ressalta-se que, por se tratar de um sistema dinâmico,

esses resultados podem mudar ao longo do tempo, à medida que as situações legais dos indivíduos venha a alterar-se, sendo necessário a contínua análise dos dados de parte das centralidades.

Tabela 5.6. Agrupamento de Indivíduos para a Análise da *Precision* e do *Recall*.

	MRS	LRS
Detected (G_3 to G_5)	$truepositives$	$falsepositives$
Not Detected (G_2)	$falsenegatives$	$truenegatives$

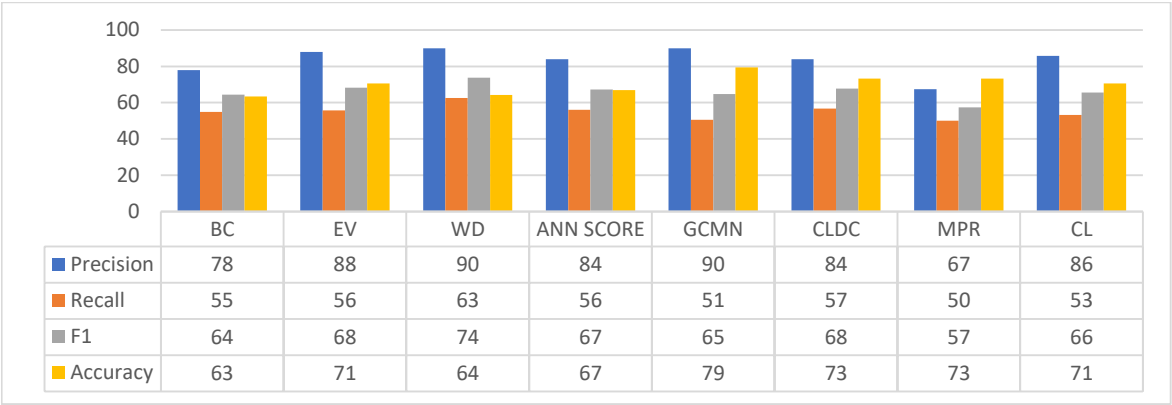
A análise da *Precision* indica qual percentual de indivíduos relevantes foi atingido, considerando como universo apenas os detectados. Assume-se que os indivíduos em MRS como os nós relevantes. Essa análise pode indicar qual centralidade pode ser utilizada em uma situação real, e.g. um processo de auditoria. Ou seja, qual centralidade é capaz de apontar o maior número de indivíduos para auditar e encontrar irregularidades e ao menor número de indivíduos em que o processo de auditoria não levará a nenhum resultado. O GCMN e o *Weighted Degree* (WD) alcançaram uma *Precision* de 90%, sendo as duas centralidades mais bem posicionadas.

A análise de *Recall* indica qual centralidade pode apontar para um número mais significativo de indivíduos relevantes, considerando todos os indivíduos que devem ser detectados. Assim como a análise da *Precision*, ela também é útil para uma equipe de investigação ao escolher qual centralidade utilizar. A centralidade GCMN atingiu 51% de *Recall*, sendo o segundo valor mais baixo dentre todas as centralidades analisadas. Apesar do resultado pouco satisfatório, deve-se considerar a proximidade entre os valores, cuja mediana é 55,5%, ou seja, apenas 2,5% acima do valor obtido pela GCMN, o que indica uma pequena variação (Figura 5.7). O *Recall* é usado em uma situação em que os Falsos Negativos são considerados mais prejudiciais que os Falsos Positivos. Por exemplo, o modelo deve de qualquer maneira encontrar todos os pacientes doentes, mesmo que classifique alguns saudáveis como doentes (situação de Falso Positivo) no processo. Ou seja, o modelo deve ter alto recall, pois classificar pacientes doentes como saudáveis pode ser uma tragédia. Essa situação não se aplica ao estudo de caso, uma vez que classificar como suspeito alguém que não tem uma situação legal gravosa não se constitue em fato grave. Dessa forma, o resultado ruim no *Recall* não se constitue em um problema grave, sendo a precisão, no caso em análise, a medida mais relevante a ser considerada.

A análise F_1 traz uma visão geral do desempenho das duas métricas (*Precision* e Recall), mostrando que a centralidade GCMN atingiu um bom resultado geral, atin-

gindo 65%. (Figura 5.7).

Considerando o estudo de caso, a *Accuracy* se refere ao grau de conformidade de uma quantidade calculada com um valor real. A *Accuracy* está intimamente relacionada à *Precision*, mas não é um sinônimo. Um resultado é considerado de alta *Accuracy* quando corresponde a um alvo específico. Neste estudo de caso, o alvo são os indivíduos com situação legal com alto grau de gravidade. Nesse quesito, a Centralidade GCMN alcança um alto grau de precisão (90 %) com alta acurácia (79%), siginificando que o ranqueamento proposto pela centralidade GCMN foi capaz de apontar os indivíduos de forma precisa, atingindo a meta com alta acurácia, sendo o melhor desempenho entre todas as outras medidas de centralidade (Figura 5.7).



	BC	EV	WD	ANN SCORE	GCMN	CLDC	MPR	CL
Precision	78	88	90	84	90	84	67	86
Recall	55	56	63	56	51	57	50	53
F1	64	68	74	67	65	68	57	66
Accuracy	63	71	64	67	79	73	73	71

Figura 5.7. Análise comparativa entre *Accuracy, Precision, Recall* e F_1.

5.5 Considerações Finais

Considerando o estudo de caso Operação Lava Jato, a centralidade GCMN provou ser superior a quatro medidas de centralidade clássicas: *Weighted Degree* (WD) (Beveridge e Shan, 2016), *Betweenness* (BW) (Freeman, 1978; Otte e Rousseau, 2002), *Closeness* (CL) (Freeman, 1978) e *Eigenvector* (EV) (Bonacich, 1972); e duas medidas de centralidade destinadas a redes multiplex: CLDC (Bródka et al., 2012), e *Multiplex PageRank* (Tu et al., 2018), na detecção de indivíduos denunciados, réus ou condenados. Esta análise foi feita de forma segmentada por grupos de nós — G_2 a G_5 —, nos quais a centralidade GCMN apresentou os seguintes resultados:

- O ranking de centralidade GCMN alcançou 90% de *Precision* e 79% de *Recall*; na detecção de indivíduos com o situação legal mais relevante (MRS). As demais medidas de centralidade obtiveram resultados inferiores nesta análise (seção 5.4.5);

- Foi utilizado o peso (Equação 3.2) como critério para distribuir os nós em grupos (Equação 3.3). A análise na seção 5.4.1 demostrou que esta foi a escolha acertada uma vez que o grau de severidade associado à situação legal dos indivíduos, distribuído nos grupos, teve um crescimento consistente (de G_2 a G_5) para todas as medidas de centralidade avaliadas;

- A centralidade GCMN foi capaz de apontar mais "novidades" do que todas as outras centralidades juntas. Isso significa que a GCMN atingiu nós mais significativos não apontados por nenhuma outra centralidade (seção 5.4.4);

- Finalmente, foi feita a análise da relevância por grupo de forma individual e cumulativa. Com relação à análise individual dos grupos mais importantes (G_3 a G_5) de indivíduos e, considerando como parâmetro de relevância o grau de severidade da sua situação legal(MRS), a centralidade GCMN foi capaz apontar os indivíduos mais relevantes. A centralidade GCMN obteve também o melhor desempenho geral na análise cumulativa, tanto para os grupos principais G_3 a G_5 quanto para a totalidade dos grupos G_4 a G_5, apontando os indivíduos com situação legal mais severa (seções 5.4.2 e 5.4.3).

No próximo capítulo haverá a discussão de mais um estudo de caso, a Operação Licitante Fantasma, que tratará da descoberta de empresas participantes de esquemas de conluio em licitações públicas. O capítulo terá a mesma estrutura encontrada neste capítulo.

Capítulo 6

Estudo de Caso - Operação Licitante Fantasma

6.1 Introdução

A Operação Licitante Fantasma foi desencadeada pela Polícia Federal e desmantelou um esquema de fraude e formação de cartel em licitações públicas entre um grupo de empresas de suprimentos médicos e hospitalares entre os anos de 2013 e 2016 (Federal, 2019). O esquema beneficiava um grupo de empresas que se revezavam como fornecedoras de insumos ao Governo do Estado de Mato Grosso do Sul. Os resultados desse estudo de caso foram preliminarmente publicados em De Figueirêdo et al. (2020) e Figueiredo et al. (2020), com uma versão preliminar da centralidade GCMN intitulada de modelo NDNS (*Nodes Detection using Network Science*).

A Lei de acesso à informação (Lei no 12.527, de 18/11/2001) exige que todos os dados de licitações públicas estejam disponíveis ao público em geral. Em cumprimento a essa norma, o site de transparência de compras do Governo Federal disponibilizou uma API (*Application Programming Interface*) para permitir de consultas às bases de dados de forma automatizada (http://compras.dados.gov.br/docs/home.html). Utilizou-se essa API na extração dos dados de todas as licitações federais de medicamentos e materiais médico-hospitalares no Estados de Mata Grosso do Sul no período de 2013 a 2016. Esses dados subsidiaram a construção de uma rede multiplex que analisamos neste estudo de caso. Uma descrição mais detalhada da obtenção desses dados encontra-se na seção 4.2.

Como a centralidade do GCMN trata de redes multiplex, a escolha natural foi dividir as licitações de quatro anos em quatro camadas (L), em que os nós (V) referem-

se às empresas que participaram das licitações e as arestas (E), representam a participação conjunta dessas empresas na mesma licitação (Equação 3.1). Essa divisão, considerando o ano fiscal, tem como fundamento a Lei Orçamentária Anual (LOA) é uma lei elaborada pelo Poder Executivo que estabelece as despesas e as receitas que serão realizadas no próximo ano. Ou seja, o orçamento de um ano refere-se às despesas do ano seguinte, e, dessa forma, essa previsão faz com que a divisão dessa despesa anualmente coincida com as tendências e forças políticas que a possam influenciar de alguma forma.

A próxima etapa foi determinar o peso (W) de cada nó (Equação 3.2), e criar os grupos (G) de nós (Equação 3.3) de acordo com seus pesos. Como resultado, foi alcançado $\mid G_2 \mid = 157$, $\mid G_3 \mid = 30$ e $\mid G_4 \mid = 5$. Observe-se que, como tem-se quatro camadas, o número de grupos considerados relevantes é três, ou seja, grupos com nós i cujo peso é $w_i \geq 2$.

A última etapa foi aplicar as Equações 3.4, 3.5 e 3.6, a todos os nós dos grupos G_2 a G_4 para obter o ranqueamento da centralidade GCMN . Neste estudo de caso, o $\max_{z \in V} D_z = 184$, então, a função $\varphi(W_i) = (184 - 1) . (W_i - 1)$ (Equação 3.6). As equações referenciadas neste parágrafo encontram-se formalizadas no capítulo 3.

As seções a seguir comparam a classificação de centralidade GCMN com as classificações das medidas de centralidade clássicas: *Weighted Degree* (WD)(Beveridge e Shan, 2016), *Betweenness* (BW)(Freeman, 1978; Otte e Rousseau, 2002), *PageRank* (PR)(Bonacich, 2007) e *Eigenvector* (EI)(Bonacich, 1972); e com as medidas de centralidade que trabalham com redes multicamadas como: CLDC (Bródka et al., 2012), *Random-walk occupation centrality* (RW)(Solé-Ribalta et al., 2016) e *Multiplex PageRank* (MPR)(Tu et al., 2018).

6.2 Trabalhos Relacionados

Estratégias para determinar as entidades mais relevantes em um cenário, em particular na detecção de fraudes em licitações públicas, são temas recorrentes em estudos científicos. A lista vai de técnicas de mineração de dados, inteligência artificial, redes complexas, entre outros.

A seguir serão discutidos trabalhos relacionados à detecção de fraudes por meio da detecção das entidades mais relevantes, podendo ser essas entidades empresas ou indivíduos suspeitos de fraude. Complementarmente há uma discussão da centralidade GCMN frente a essas tecnologias.

A mineração de dados refere-se à extração (*mining*) de conhecimento a partir de

uma grande quantidade de dados. A Lei de Newcomb-Benford é uma ferramenta de mineração de dados concebida para trabalhar em parceria com a curva ABC contribuindo para uma análise de possíveis sobrepreços nos itens de uma licitação pública (Carvalho e Ramos, 2002). A curva de experiência ABC, também chamada de análise de Pareto ou regra 80/20, é um método de categorização de estoques, cujo objetivo é determinar quais são os produtos mais importantes de uma empresa. Foi desenvolvido pelo consultor de qualidade romeno-americano Joseph Moses Juran, que verificou que 80% dos problemas são geralmente causados por 20% dos fatores. O nome "Pareto" é uma homenagem ao economista italiano Vilfredo Pareto, que em um estudo observou que 80% das riquezas são concentradas nas mãos de 20% da população, sendo que boa parte do entendimento da curva ABC se deve a esse estudo de Pareto (Machado, 2012). A Lei de Newcomb-Benford propõe que a frequência na qual o primeiro dígito aparece em um grande e genuíno banco de dados decresce do dígito um ao dígito nove. Ou seja, o dígito um aparece em 30% dos dados, enquanto o dígito nove não alcança 5% .

Considerando a Lei de Newcomb-Benford, apenas 30% dos valores empregados na reforma do estádio de futebol do Maracanã para a copa do mundo de 2014 correspondiam ao esperado. O experimento permitiu identificar dezessete serviços realizados que não se enquadraram com o que preconiza a Lei, representando 71.54% do valor total gasto, alcançando o valor total aproximado de 64 milhões de dólares (Cunha e Bugarin, 2014; Carvalho, 2017). Dessa forma, foram identificadas as prováveis empresas fraudulentas que auxiliaram um processo de auditoria do Tribunal de Contas da União.

O uso de heurísticas e de agentes inteligentes, associado a técnicas de mineração de dados, tem sido testado com sucesso em bases de dados voltadas à auditoria como uma ferramenta de combate à formação de carteis em licitações públicas (Cunha e Bugarin, 2014; Silva e Ralha, 2010). São encontradas abordagens semelhantes no desenvolvimento de Agentes de Mineração de Dados (AGMI) que, utilizando base de dados reais da Controladoria Geral da União (CGU), atuam como uma ferramenta de predição de corrupção e formação de carteis, conseguindo, nessa tarefa, uma acurácia de 90% de casos detectados (Ghedini Ralha e Sarmento Silva, 2012).

Uma outra abordagem de sucesso trata do emprego de mineração de dados associada a regras de auditoria. O uso de tabelas de preços padrão, onde discrepâncias que extrapolem um percentual quantificável de acordo com a as caraterísticas de cada licitação, pode indicar a existência de associações ilícitas em licitações públicas. É a prática do chamado *overpricing*, ou extrapolação de preços, típica dos carteis de fornecedores (Silva e Ralha, 2010; Costa e Aparicio, 2011).

A CGU faz uso de ontologias e de redes Bayesianas como instrumento na preven-

ção de fraudes no setor público (Hu et al., 2013). As ontologias levam em consideração as informações semânticas (e.g. relações comerciais ou sociais entre indivíduos ou empresas). Essa mesma abordagem é também encontrada em outros trabalhos com objetivos similares (Balaniuk et al., 2013). Em todos os casos a análise da informação semântica levaram a resultados superiores dos que os encontrados em outras abordagens que fazem uso de ontologias sem essa característica.

O uso de heurísticas permite a um auditor detectar, de forma mais eficaz, comportamentos fraudulentos. Um experimento de campo envolvendo vinte e quatro auditores, parceiros em empresas internacionais de auditoria, os convidou a relatar situações reais ocorridas em empresas que tinham participado de esquemas de fraude financeira (Johnson et al., 2001). O experimento funcionou num esquema de mesa redonda, onde os casos eram relatados e, então, eram formadas pequenas equipes de auditoria com o objetivo de detectar as empresas envolvidas nessas fraudes.

Enquanto muitos auditores falharam em detectar os procedimentos fraudulentos, um pequeno número de auditores, utilizando heurísticas, foi bem-sucedido nessa tarefa. O estudo justifica o sucesso no uso das heurísticas por elas terem sido concebidas a partir da experiência obtida com erros no trabalho cotidiano de auditores. Essas heurísticas são baseadas na interpretação de inconsistências detectadas que consideram os objetivos e as possíveis ações dos fraudadores. Elas usam táticas desenvolvidas a partir da experiência com falhas em várias situações de auditoria, estando disponíveis para o uso em situações específicas, como a auditoria das demonstrações financeiras de uma empresa, por exemplo (Johnson et al., 2001).

Embora fraudes em empresas públicas tenham, em geral, um grande impacto financeiro, a sua detecção quase sempre ocorre apenas quando há uma denúncia formal a um órgão de controle ou fiscalização. Não existe, portanto, um trabalho investigativo sistematizado e proativo que monitore as ações dos gestores e das empresas contratadas de forma a detectar comportamentos anômalos mesmo que não haja denúncia nesse sentido.

Uma tática para detecção de fraudes de forma proativa e de *outliers* é o exame textual de processos e de editais de licitações disponibilizados por empresas públicas. Essa abordagem obtém uma acurácia superior a 88%, fazendo uso de um conjunto de palavras derivado de forma empírica (Skillicorn e Purda, 2012). Fazendo uso de redes neurais, o processo de detecção de fraudes utiliza dados históricos de padrões de texto e expressões que denotem suspeita de fraude (Silva, 2016). Entretanto essas análises se restringem a casos que envolvam análise textual, não sendo aplicáveis a bancos de dados relacionais convencionais.

Silva (2016) introduz a ideia de "velocidade normal", que trata do tempo estatis-

ticamente presumido para que cada uma das fases de uma licitação pública ocorra. Ou seja, caso em alguma das fases de uma licitação pública ocorra uma velocidade considerada anormal (muito rápida ou lenta), o trabalho é capaz de apontar a ocorrência indícios de fraude. O trabalho utiliza *deep learning* na proposição de técnicas de reconhecimento de imagens para determinar se o estágio de evolução de uma obra pública corresponde ao reportado por meios documentais. Entretanto, o escopo de trabalho restringe-se à análise de licitações públicas, não sendo um indício de fraude aplicável a situações generalistas.

A centralidade GCMN, ao contrário de outros trabalhos cuja aplicabilidade é restrita a situações específicas, sendo chamados de "sistemas especialistas" (Balaniuk et al., 2013; Hu et al., 2013; Costa e Aparicio, 2011; Cunha e Bugarin, 2014; Silva, 2016), é útil para situações onde seja possível a utilização de uma rede multiplex não direcionada no mapeamento do cenário, tendo, portanto, a vantagem de ser aplicável a uma vasta gama de circunstâncias. Outro aspecto é que alguns trabalhos lidam apenas com tipos específicos de dados, por exemplo, informações textuais (Skillicorn e Purda, 2012) ou indicadores de dados (Bhowmik, 2008; Virdhagriswaran e Gordon, 2006), já a centralidade GCMN é capaz de gerenciar qualquer tipo de informação que possa ser modelada em uma rede, tornando-a mais flexível.

6.3 Métrica de avaliação proposta

A Centralidade GCMN baseia-se na distribuição dos nós por grupos de acordo com sua relevância. As outras medidas de centralidade com as quais compara-se a GCMN não têm esse conceito. Desta forma, para efeito de comparação, será utilizada a metodologia descrita na seção 4.1.1. Optou-se pelo uso de medidas de centralidade que não fossem basedas em grupos para que a definição dos grupos a serem comparadados fosse dada utilizando-se a metodologia proposta pela GCMN. Isso possibilita uma análise comparativa mais precisa, na medida que estar-se-á tratando dos mesmos grupos de nós, com as mesmas quantidades, para todas as medidas em análise.

Aplicando-se a classificação ao estudo de caso, tem-se que, para o grupo G_4, com cinco nós, toma-se também os cinco primeiros nós mais bem classificados nas outras medidas de centralidade. Desta forma, é possível comparar o grupo G_4 com os nós mais bem classificados em todas as outras medidas de centralidade. O grupo G_3 possui trinta nós, e considera-se, para este grupo, os nós classificados entre seis e trinta e cinco em todas as outras medidas. Este processo é o mesmo para o grupo G_2.

Para efeito de análise, considerando que o estudo de caso trata da participação

de empresas em licitações, cujo objetivo final, por óbvio, é o lucro, um parâmetro relevante é o ganho monetário de cada empresa. Portanto, considera-se coerente a proposta de uma métrica que considere os valores auferidos na participação das empresas nos processos licitatórios como balizador de sua importância. Desta forma, a relevância dos grupos será medida de acordo com os valores obtidos pelas empresas pertencentes àqueles grupos, verificando assim se os grupos mais relevantes conseguiram apontar as empresas com ganhos também mais significativos. É importante frisar que o valor escalar que indica o peso (W_i) de um nó i (Equação 3.2) é proporcional à sua importância, e que, consequentemente, um grupo G_w deve ter nós mais relevantes do que um grupo G_{w-1} (Equação 3.3).

6.4 Resultados e Discussão

Na análise dos resultados e sua discussão consideraram dois parâmetros: o uso de métricas conhecidas como *Accuracy, Precision, Recall* e F_1 *Score* e as Correlações de Pearson e Spearman (seções 6.4.3 e 6.4.4) e o uso de uma métrica proposta nesta tese (seção 6.3). Essa métrica considerou avaliou o desempenho do CGMN sob dois aspectos: o peso como um parâmetro de agrupamento (seção 6.4.1) e a relevância por grupo (seção 6.4.2).

A Figura 6.1 traz a rede complexa relativa ao ano de 2013, cujo processo de aquisição de dados encontra-se descrito na seção 4.2. Nela pode-se verificar visualmente a existência de núcleos de empresas, sugerindo associações. A escolha natural para a segmentação dessa rede em camadas foi a alocação de cada um dos anos de licitações como um camada de uma rede multiplex não direcionada, cada uma, com: 671, 358, 297 e 388 nós, respectivamente; onde os nós referem-se às empresas participantes de licitações ano-a-ano, e as arestas à participação conjunta dessas empresas na mesma licitação.

A Figura 6.2 aponta os nós em comum entre duas das quatro camadas da rede multiplex, representando os anos de 2013 e 2014, essa quantidade serve como base para o cálculo do peso dos nós (equação 3.2) e na consequente formação dos grupos (equação 3.3). A Figura 6.3 traz uma visão ampla da rede expondo a quantidade de nós em comum entre as quatro camadas da rede multiplex, onde cada uma corresponde a um ano de licitações analisadas. É possível observar que todas as quatro camadas têm nós em comum com todas as demais, sendo essas quantidades similares. Esses nós são relativos aos pertencentes ao grupo G_4, relacionados na Tabela 6.1. Essa análise é útil uma vez que demonstra que o volume de licitações envolvendo as empresas sob

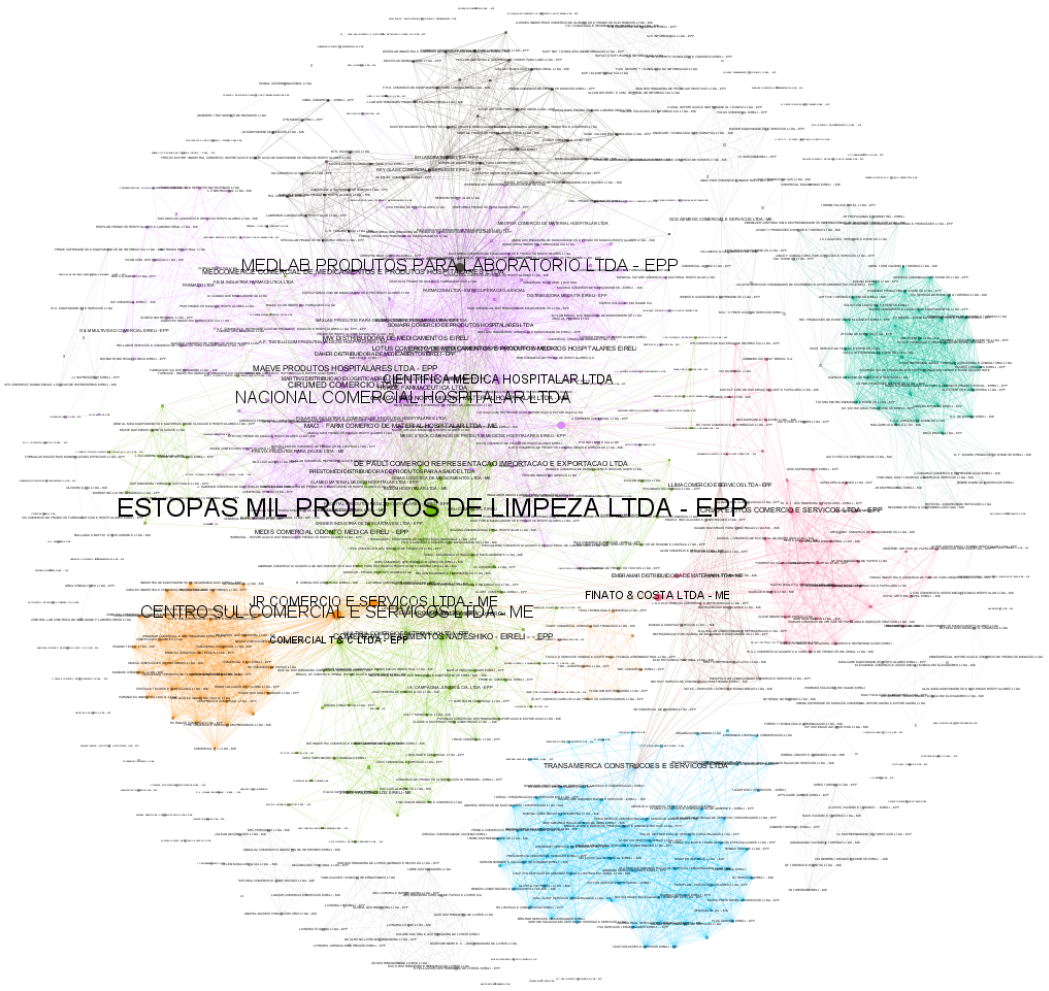


Figura 6.1. Rede complexa do ano de 2013 - Operação Licitante Fantasma.

análise se manteve coerente ao longo do período, ou seja, que a quantidade de empresas analisadas ano-a-ano é semelhante, não havendo um ano, ou período, que detenha um volume de empresas significativamente maior ou menor que os demais.

A divulgação dos nomes das empresas mais relevantes na análise, relativas ao grupo G_4 (Tabela 6.1), objetiva apenas o relato do resultado do ranqueamento da centralidade GCMN. É importante deixar claro que não há nenhum juízo de valor acerca da conduta ou culpabilidade dessas empresas.

6.4.1 O peso como um parâmetro de agrupamento

Considerando o agrupamento de nós proposto pela Centralidade GCMN, a Figura 6.4 mostra a distribuição normalizada dos rendimentos corporativos por ano e grupos.

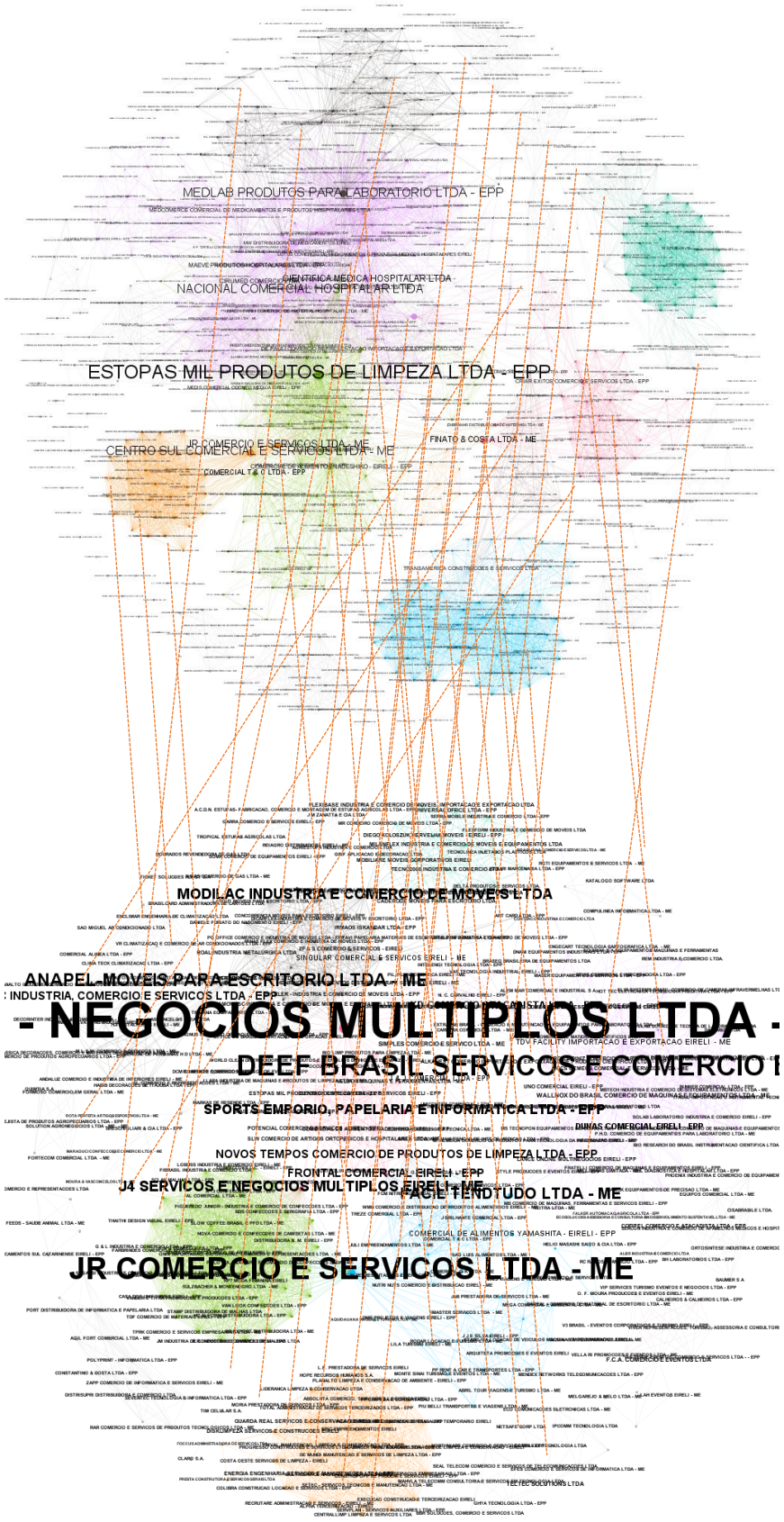


Figura 6.2. Nós em comum entre duas camadas da Rede Multiplex, representando os anos de 2013 e 2014 (Estudo de Caso - Licitante Fantasma).

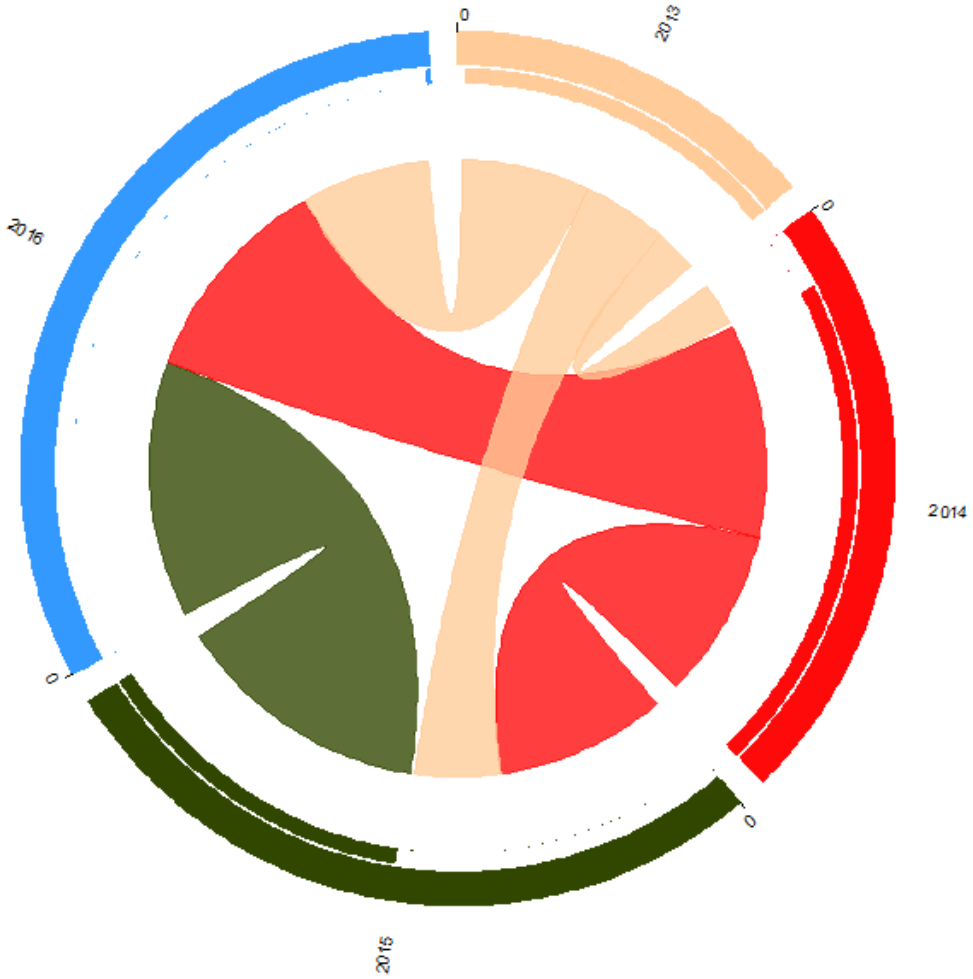
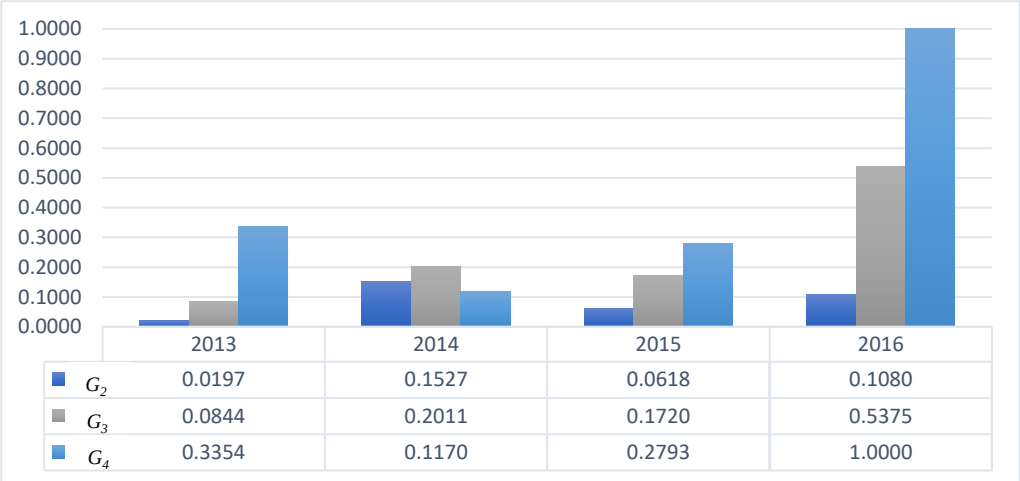


Figura 6.3. Quantidade de nós em comum entre as camadas da Rede Multiplex (Estudo de Caso - Licitante Fantasma).

Tabela 6.1. Ranqueamento de Suspeitos da Centralidade GCMN - Operação Licitante Fantasma.

Grupo G_4
PGA Servicos Terceirizados Eireli - Epp
Lideranca Limpeza e Conservacao Ltda
Total Administracao de Servicos Terceirizados Ltda - Epp
Planalto Limpeza e Conservacao de Ambiente - Eireli - Epp
Clima Teck Climatizacao Ltda - Epp

Percebe-se que o crescimento dos valores é consistente com o grau de relevância dos grupos G_2 a G_4. A única exceção ocorre no grupo G_4 para o ano de 2014.



	2013	2014	2015	2016
G_2	0.0197	0.1527	0.0618	0.1080
G_3	0.0844	0.2011	0.1720	0.5375
G_4	0.3354	0.1170	0.2793	1.0000

Figura 6.4. Valores Normalizados da Centralidade GCMN por Grupo.

Considerando todas as medidas de centralidade (GCMN, WD, BW, PR, EI, CLCD, RW e MPR), a Figura 6.5 traz a distribuição normalizada dos ganhos das empresas para todo o período. Como esperado, o grupo G_4 possui os maiores valores que diminuem para os grupos G_3 e G_2, respectivamente.

É importante ressaltar que o lucro das empresas é medido em valores absolutos, não sendo um valor médio auferido pelas empresas em cada grupo (Figura 6.5). Isso significa que o grupo G_4, mesmo tendo um número muito menor de empresas, atingiu valores absolutos superiores aos grupos G_3 e G_2. A mesma análise é válida para o grupo G_3 em relação ao grupo G_2.

Percebe-se que a separação das empresas em grupos demonstrou ser uma proposta coerente, sendo aplicável não apenas para a Centralidade GCMN, mas também para análise conjunta de todas as centralidades.

6.4.2 Relevância por Grupo

A Figura 6.6 traz uma análise comparativa baseada em grupos da relevância para todas as medidas de centralidade. A métrica proposta considera os ganhos da empresa como parâmetro de relevância, assim, a soma dos resultados das empresas em cada grupo indica a importância desse grupo. Consequentemente, a relevância leva em consideração o ganho total de cada grupo. A figura traz a relevância por grupo (G_2 a G_4) e uma análise cumulativa ($G_3 + G_4$ e $G_2 + G_3 + G_4$).

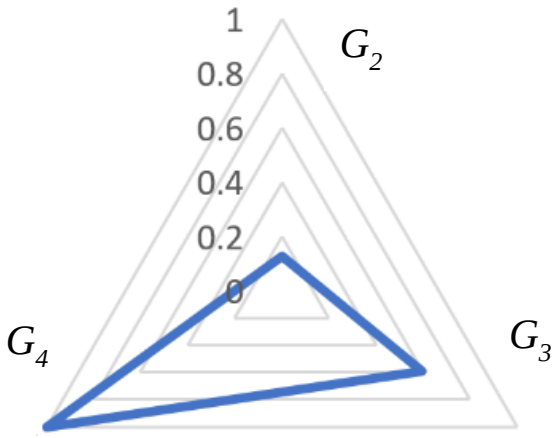


Figura 6.5. Ganhos das Empresas por Grupos.

Observa-se que para os grupos G_3 e G_4 a centralidade GCMN foi capaz de apontar as empresas mais relevantes. Para o grupo G_2, a centralidade *Multiplex PageRank* (MPR) foi capaz de apontar as empresas com maiores ganhos. No entanto, como o grupo G_2 deve conter, supostamente, as empresas menos relevantes, não se pode considerar este um resultado ruim. Ou seja, as empresas menos relevantes deveriam ter ganhos menores, não os maiores. Diante disso, a distribuição desejada dos valores é justamente essa; ou seja, as empresas mais relevantes (ganhos mais altos) estão nos grupos mais relevantes — G_3 e G_4 — e as empresas menos relevantes (ganhos mais baixos) estando no grupo menos relevante(G_2). Assim, o fato da centralidade GCMN apontar para empresas com menor ganho no grupo G_2 é o resultado esperado.

A análise cumulativa mostra que a centralidade GCMN foi capaz de apontar as empresas mais relevantes em todo o cenário. Um resultado esperado é a superioridade das medidas de centralidade que lidam com redes multiplex (CDML, RW e MPR) em comparação com as medidas de centralidade clássicas (BW, PR e EI). A única exceção é a centralidade *Weighted Degree* (WD), que obteve resultados gerais compatíveis com as medidas de centralidade multiplex.

A Figura 6.6 traz uma análise cumulativa baseada em grupos. Percebe-se que a centralidade GCMN foi capaz de apontar as empresas mais relevantes (ganhos mais significativos), obtendo o melhor desempenho global.

Comparando os resultados da centralidades clássicas com as centralidades multicamadas, percebe-se a importância de uma modelagem adequada à realidade, como é o caso das centralidade multiplex nestes casos. Dessa forma, conclui-se ser imprescindível o uso de estruturas adequadas de mapeamento dos cenários para a obteção de

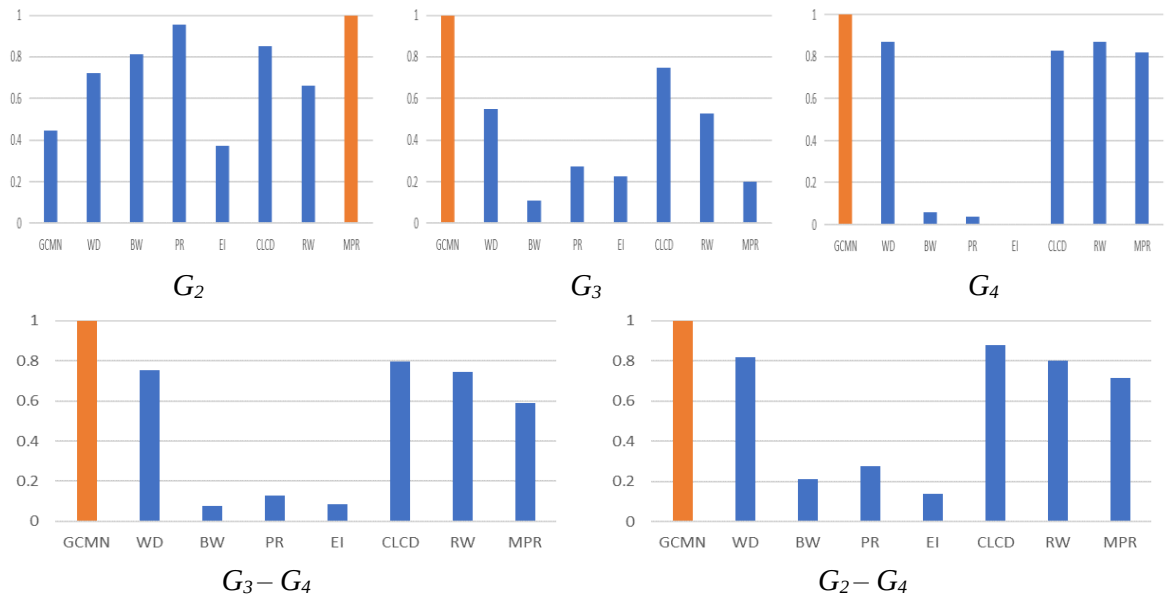


Figura 6.6. A relevância baseada em grupos dos nós.

melhores resultados. Isso é comprovado pela comparação dos resultados obtidos entre os dois grupos de centralidades.

6.4.3 *Accuracy, Precision, Recall* e F_1 *Score*

Os conceitos de *Accuracy, Precision, Recall* e F_1 *Score* são métricas estatísticas que já foram discutidas na seção 4.1.2. Esses conceitos serão utilizados nesta seção deste estudo de caso.

Para fins de análise as empresas serão divididas segundo dois critérios:

- detectadas/não detectadas: são aquelas que foram ou não apontadas pela centralidade em análise, podendo ter ou não sido condenadas;
- condenadas/não condenadas: são aquelas que sofreram ou não um processo legal de denúncia e foram apontadas como culpadas de fraude.

Essa decisão de montagem da "tabela de confusão" teve como base critérios de relevância das empresas e a sua detecção, ou não, pela centralidade. Dessa forma, aquelas empresas que pertencem ao conjunto das condenadas e foram considerados como os mais relevantes ou, positivos, sendo esse um conjunto disjunto daquelas que não tiveram condenações, das empresas menos relevantes, ou negativos. As empresas detectados ou não, foram consideradas os como os verdadeiros ou falsos, respectivamente,

uma vez que esse critério determina se a centralidade em análise foi capaz de apontá-la. Ressalta-se que, por se tratar de um sistema dinâmico, esses resultados podem mudar ao longo do tempo, à medida que as situações legais (condenações) das empresas venha a alterar-se, sendo necessário a contínua análise dos dados de parte das centralidades.

Tomando o estudo de caso, considera-se as empresas condenadas detectadas como $thetruepositives$; as empresas detectadas não condenadas como $thefalsepositives$; as empresas não detectadas e condenadas como $thefalsenegatives$; e as empresas não detectadas e não condenadas como $thetruenegatives$ (Tabela 6.2).

Tabela 6.2. Análise da *Precision* e da *Recall* por Grupos de Empresas.

	Condenadas	Não Condenadas
Detectadas	$truepositives$	$falsepositives$
Não Detectadas	$falsenegatives$	$truenegatives$

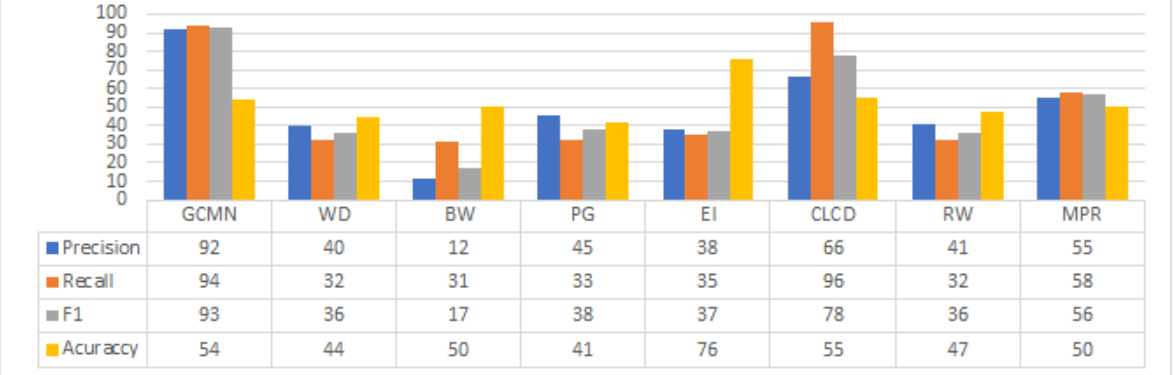
A análise da *Precision* mostra qual o percentual de empresas relevantes descobertas no universo que considera apenas as detectadas pela centralidade. Assume-se que as empresas condenadas como sendo os nós relevantes, uma vez que o objetivo do estudo de caso é detectar empresas fraudulentas. Dessa forma a *Precision* pode indicar a qual centralidade deva ser utilizada em uma situação real, e.g. um processo de auditoria. Em outras palavras, qual a centralidade que é capaz de conduzir ao número mais significativo de empresas a auditar e apurar irregularidades. A Centralidade GCMN alcançou 92% de *Precision* contra 66% da segunda centralidade melhor posicionada, o CLCD.

A análise do *Recall* é essencial para mostrar qual centralidade pode apontar para um número mais significativo de empresas relevantes, considerando todas as empresas e não apenas as detectadas. O *Recall*, a exemplo da *Precision*, é útil para uma equipe de investigação ao escolher qual centralidade utilizar. A centralidade GCMN atingiu o segundo melhor resultado nesta métrica (94%), atrás apenas da centralidade CLCD, por uma diferença de 2%. Em relação às demais medidas de centralidade, a diferença foi significativa, tendo o terceiro melhor colocado um *Recall* abaixo de 60% (Figura 6.7).

A análise F_1 traz uma visão geral do desempenho das duas métricas, mostrando que a centralidade GCMN alcançou um resultado geral melhor, atingindo 93%, contra 78% da segunda medida de centralidade mais bem posicionada, o CLCD (Figura 6.7).

A análise da acurácia demonstrou que a centralidade GCMN atingiu o percentual de 54%, sendo o terceiro melhor, atrás da CLCD (55%) e da EI (76%). Analisando esse

resultado em conjunto com a *Precision* verifica-se que a centralidade GCMN foi capaz de alcançar um alto grau de *Precision* (92%) com uma boa acurácia (54%) (Figura 6.7).



	GCMN	WD	BW	PG	EI	CLCD	RW	MPR
Precision	92	40	12	45	38	66	41	55
Recall	94	32	31	33	35	96	32	58
F1	93	36	17	38	37	78	36	56
Acuraccy	54	44	50	41	76	55	47	50

Figura 6.7. Análise Comparativa entre *Accuracy, Precision, Recall* e F_1.

A Figura 6.8 mostra a classificação normalizada das medidas de centralidade, considerando o vetor composto pelas medidas de *Precision* e *Recall* de cada centralidade $c = (r, p)$ até o ponto de desempenho máximo $m = (1, 1)$. Assim, quanto mais próximo o ponto de medida c de uma centralidade estiver do ponto de máximo m, melhor será seu desempenho. Considerando a norma do vetor composto pelos pontos $z = (0, 0)$ e m como $d_{zm} = \sqrt{(1-0)^2 + (1-0)^2}$, sendo este o maior vetor possível, subtrai-se deste valor máximo a norma obtida por cada vetor composto pelas coordenadas de cada centralidade c e m, assim $d_{cm} = \sqrt{(1-r)^2 + (1-p)^2}$ alcançando a classificação por *Precision* e *Recall* dada por $d_{zm} - d_{cm}$.

As figuras 6.7 e 6.8 demonstram que a centralidade GCMN atingiu o melhor desempenho geral em comparação com as outras sete medidas de centralidade. Como esperado, as medidas de centralidade CLCD e MPR, cuja concepção contempla as redes multiplex, superaram as medidas de centralidade clássicas.

6.4.4 Análise das Correlações de Pearson e Spearman

No estudo de caso, o uso da análise estatística é uma forma de demonstrar a coerência do uso da distribuição por grupos para classificar os nós de acordo com sua suposta relevância. Pretende-se demonstrar a correlação entre os grupos propostos na centralidade GCMN e o faturamento das empresas, verificando como o lucro tende a crescer com a significância dos grupos. A força de uma relação entre essas duas variáveis é capaz de demonstrar a coerência do agrupamento dos nós proposto pela centralidade GCMN.

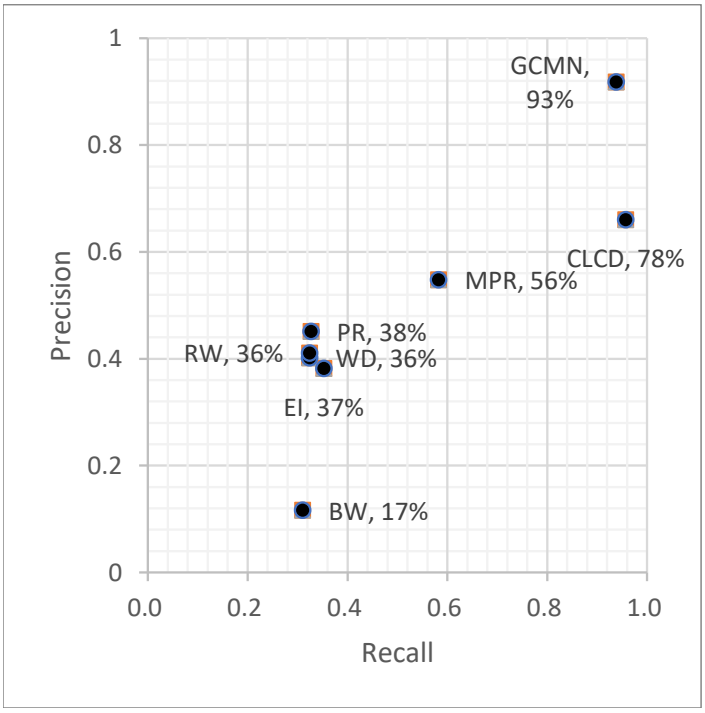


Figura 6.8. Classificação Normalizada das Medidas de Centralidade segundo a *Precision* e o *Recall*.

Uma vez que a correlação proposta por Spearman (Spearman, 1904) trata de funções monotônicas ela é aplicável ao estudo de caso, por se tratar de uma função monotônica de crescimento estrito, ou seja, $\forall x, y \in A, (x > y \Rightarrow f(x) \geq f(y))$, onde x pode representar o faturamento das empresas e y o ranqueamento fornecido pelas centralidades analisadas. Dessa forma, o coeficiente de correlação de Spearman é capaz analisar a intensidade e a direção dessa relação monotônica. Ressalta-se que em um relacionamento monotônico, as variáveis tendem a se mover na mesma direção relativa, mas não necessariamente a uma taxa linear.

Como os coeficientes de correlação de Pearson e Spearman são medidas estatísticas da força de uma relação entre dados pareados e, considerando que, nesse estudo de caso, lida-se com duas variáveis que deveriam, na situação ideal, ter uma correlação linear e monotônica, ou seja, à medida que uma cresce, a outra também deveria crescer linear e continuamente, utiliza-se ambos os coeficientes para avaliar essas relações.

A tabela 6.3 traz os coeficientes de correlação de Pearson e Spearman para cada centralidade, considerando os ganhos das empresas distribuídos por grupo. Observa-se que, tanto nos métodos de Pearson quanto no de Spearman, a centralidade GCMN alcançou as melhores pontuações na avaliação da correlação entre o lucro da empresa e a relevância dos grupos. Essa correlação demonstra que a centralidade proposta aloca

Tabela 6.3. A análise de correlações de Pearson e Spearman.

Centralidade	Pearson		Spearman	
	r	p-value	r_s	p-value
GCMN	0.999	0.025	1.000	0.333
WD	0.938	0.226	1.000	0.333
BW	-0.882	0.312	-1.000	0.333
PR	-1.000	0.013	-1.000	0.333
EI	-0.759	0.442	-0.500	1.000
SC	-0.515	0.656	-0.500	1.000
CLCD	0.983	0.118	1.000	0.333
RW	0.938	0.225	1.000	0.333
MPR	0.926	0.247	1.000	0.333

as empresas mais relevantes nos grupos também mais relevantes.

Tomando-se a Figura 6.9, verifica-se, com relação à centralidade GCMN, que a distribuição normal dos valores faturados cresce de acordo com a relevância dos grupos de G_2 para G_4. Observa-se que outras centralidades, como WD, CLCD e RW; tiveram comportamentos semelhantes, mas alcançando valores mais baixos nas correlações de Pearson e Spearman. Dois casos interessantes são as centralidades BW e PR que tinham correlações de Spearman iguais a -1. Isso significa que elas se comportaram na direção oposta à esperada; ou seja, eles tiveram uma concentração de valores inversamente proporcional à relevância dos grupos.

Considerando que a análise dos coeficientes de Pearson e Spearman deve considerar a significância (valor p) para validar a análise das correlações, verifica-se que, no estudo de caso, a correlação de Pearson atingiu um p-value $= 0,025$; ou seja, a correlação é estatisticamente significativa. A correlação de Spearman, com p-value $= 0,333$, indica que não é estatisticamente significativa. No entanto, optou-se por manter esta análise como informação complementar uma vez que ambas correlações atingiram valores significativos (r e r_s).

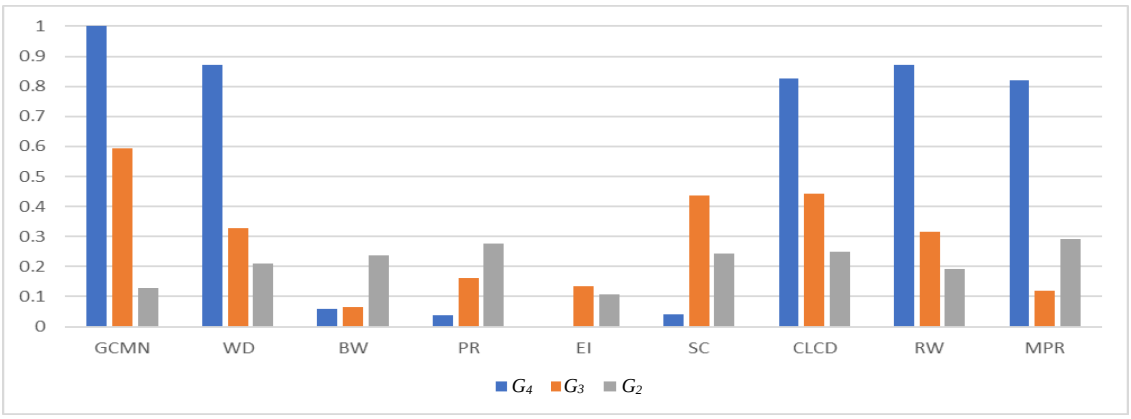


Figura 6.9. Valores normalizados por centralidade/grupo.

6.5 Considerações Finais

Considerando o estudo de caso Operação Licitante Fantasma, a centralidade GCMN provou ser superior a quatro medidas de centralidade clássicas: *weighted degree* (WD) (Beveridge e Shan, 2016), *betweenness* (BW) (Freeman, 1978; Otte e Rousseau, 2002), *PageRank* (PR) (Bonacich, 2007) e *eigenvector* (EI) (Bonacich, 1972); e três medidas de centralidade multicamadas: CLDC (Bródka et al., 2012), *Random-walk occupation centrality* (Solé-Ribalta et al., 2016) e *Multiplex PageRank* (Tu et al., 2018), na detecção de empresas fraudulentas. Esta análise foi feita de forma segmentada por grupos de nós - G_2 a G_4 -, nos quais a centralidade GCMN apresentou os seguintes resultados

- O ranking de centralidade GCMN alcançou 92% de *Precision*, 94% de *Recall* e 93% de F_1; na detecção de valores fraudulentos, contra 66% de *Precision*, 96% de *Recall* e 78% de F_1, da segunda medida de centralidade mais bem posicionada, o CLCD. As outras medidas de centralidade obtiveram resultados inferiores nesta análise (seção 6.4.3).

- Foi utilizado o peso (Equação 3.2) como critério para distribuir os nós em grupos (Equação 3.3). A análise na seção 6.4.1 demonstrou que esta foi a escolha acertada uma vez que os ganhos das empresas, distribuídos por grupos, tiveram um crescimento consistente nos grupos de G_2 a G_4 para todas as medidas de centralidade avaliadas, atingindo os maiores valores no grupo G_4 e os menores no grupo G_2, conforme esperado.

- A centralidade GCMN obteve o melhor resultado na análise de correlações de Pearson e Spearman, quando comparada com as demais centralidades, mostrando

ser capaz de distribuir as empresas nos grupos de forma que seus ganhos fossem proporcionais à relevância dos grupos (seção 6.4.4).

- Finalmente, a análise da relevância por grupos, que foi feita de forma individual (G_2, G_3 e G_4) e cumulativa (G_3 - G_4 e G_2 - G_4) demonstrando que, com relação aos grupos mais importantes (G_3 a G_4), a centralidade CGML foi capaz de apontar para as empresas com os maiores ganhos. O MPR obteve o melhor desempenho no grupo menos relevante (G_2), mas uma vez que este grupo dever ter, supostamente, as empresas menos importantes, encontrar os maiores valores neste grupo equivale a dizer que as empresas menos importantes são as principais. Assim, não ser a centralidade mais bem posicionada nesse grupo (G_2) e ser nos grupos mais relevantes (G_3 e G_4) é o melhor resultado possível. Na análise cumulativa, somaram-se os ganhos das empresas nos grupos G_3 a G_4 e G_2 a G_4 para se obter uma visão geral de desempenho das centralidades em todos os grupos. Nessa análise, tanto para os grupos principais (G_3 a G_4) quanto para a totalidade dos grupos (G_2 a G_4), a centralidade GCMN obteve os melhores resultados, em ambos os cenários, sendo capaz de apontar as empresas com os maiores ganhos (seção 6.4.2).

No próximo capítulo haverá a discussão de mais um estudo de caso, a Coleção de Livros da Personagem Harry Potter, que tratará da descoberta das personagens mais relevantes daquela obra. O capítulo terá a mesma estrutura encontrada neste capítulo.

Capítulo 7

Estudo de Caso - Coleção de Livros da Personagem Harry Potter

7.1 Introdução

O estudo de caso toma como base uma rede multiplex [1] que representa os relacionamentos entre as personagens do conjunto de livros [2] que reportam a estória do protagonista "Harry Potter", todos de autoria de J. K. Rowling. O mapeamento dos sete livros é dado nas sete camadas da rede, cada camada L representando um livro, com arestas não direcionadas, onde os nós V representam as personagens, as arestas E a ocorrência conjunta dessas personagens em um mesmo parágrafo e o peso das arestas é dado pelo número de vezes que esse fato ocorre.

A próxima etapa foi determinar o peso (W) de cada nó (Equação 3.2), e criar os grupos (G) de nós (Equação 3.3) de acordo com seus pesos. Como resultado tem-se a seguinte distribuição de nós por grupo: $\mid G_2 \mid = 65$, $\mid G_3 \mid = 38$, $\mid G_4 \mid = 18$, $\mid G_5 \mid = 12$, $\mid G_6 \mid = 13$, and $\mid G_7 \mid = 26$. Observa-se que, como tem-se uma rede multiplex com sete camadas, o número de grupos considerados relevantes é seis, ou seja, grupos com nós i cujo peso é $W_i \geq 2$. A última etapa foi aplicar as Equações 3.4, 3.5 e 3.6; a todos os nós dos grupos G_2 a G_7 para obter a classificação dos nós proposta pela Centralidade GCMN. Neste estudo de caso, o $\max_{z \in V} D_z = 509$, então, a função φ será $\varphi(W_i) = \ (509 + 1) \, . \, (W_i - 1)$ (Equation 3.6). As equações referenciadas neste parágráfo encontram-se formalizadas no capítulo 3.

[1] A base de dados que deu origem à rede multiplex foi cedida pelos autores de Carvalho (2017).

[2] Harry Potter e a Pedra Filosofal, Harry Potter e a Câmara Secreta, Harry Potter e o Prisioneiro de Azkaban, Harry Potter e o Cálice de Fogo, Harry Potter e a Ordem da Fênix, Harry Potter e o Enigma do Príncipe e Harry Potter e as Relíquias da Morte.

As seções a seguir comparam a classificação de centralidade GCMN com as classificações das medidas de centralidade: *Weighted Degree* (WD)(Beveridge e Shan, 2016), *Betweenness* (BW)(Freeman, 1978; Otte e Rousseau, 2002), *PageRank* (PR)(Bonacich, 2007), *Eigenvector* (EI)(Bonacich, 1972) e *Closeness* (CL) (Freeman, 1978).

7.2 Trabalhos Relacionados

Estratégias para determinar as entidades mais relevantes em um cenário são aplicáveis a uma vasta gama de situações onde seja útil se apontar os elementos que se destaquem. Sob essa ótica, a detecção dos chamados *outliners* tem relação direta com a descoberta dessas entidades, como é o caso das personagens mais relevantes da coleção de livros do Harry Potter.

A detecção de entidades que sejam caracterizadas como *outliners* (pontos fora da curva) em uma coleção de dados é uma das questões recorrentemente tratadas em mineração de dados. Define-se *outliner* como uma entidade que, pelas suas características, difere das outras em um conjunto de dados. A detecção de *outliners* é útil na descoberta de dados imprevisíveis e não identificados (Bansal et al., 2016). Considerando os dados do estudo de caso, os outliners podem ser entendidos como as personagens que mais de destacaram, sendo, portanto, teoricamente as mais relevantes neste contexto.

A Inteligência Artificial, em particular estudos abordando a lógica Fuzzy e de redes bayesianas, têm também como proposta a detecção de *outliners*. Em Medvedeva e Komotskiy (2016) encontra-se uma tentativa de identificação de consumidores que adulteram a leitura de medidores de consumo de energia com a intenção de ter suas contas reduzidas. Dessa forma, a identificação desses potencais fraudadores indica quais consumidores têm maiores chances de estar cometendo ilícitos. Esse processo auxilia a auditoria, uma vez que seria inviável a auditagem de todos os consumidores de uma empresa fornecedora de energia elétrica, pela sua grande quantidade.

Virdhagriswaran e Gordon (2006) propuseram uma estratégia interessante na descoberta de fraudes contábeis, por meio da detecção de *outliners*, que faz uso de mineração de dados e de redes bayesianas, tendo duas fases distintas. A primeira trata de determinar as características consideradas normais dos dados contábeis, com o objetivo de criar classificadores que sirvam como critério para agregação e classificação desses dados. A segunda infere acerca do comportamento futuro desses dados, fazendo uso de redes bayesianas. Os resultados obtidos apontam para uma detecção de fraudes com um percentual de precisão variando entre 50% e 72%. Essas informações baseiam-se na aplicação da estratégia a um conjunto de empresas públicas e privadas, num

período de quatro anos.

O uso das medidas de centralidade tem como objetivo a descoberta dos atores mais relevante em um contexto. Entretanto, as medidas de centralidade clássicas apresentam desvantagens na tarefa de identificar os nós mais significativos de uma rede complexa nos casos de cenários em que as decisões são baseadas em vários contextos, propiciando o surgimento de novas propostas. Um exemplo de estratégia que considera essa nova premissa é o trabalho apresentado por Yang et al. (2020) que aborda a existência de vários contextos, sendo baseada no $k - means$ ponderado para classificar as empresas de capital de risco no mercado de investimento chinês. A proposta apresenta um critério de ranqueamento de nós baseado na avaliação por grupos de empresas como uma alternativa às medidas de centralidade convencionais.

Outro novo modelo de geração de redes, baseado em dados do mundo real em larga escala, é o proposto por Fire e Guestrin (2020). Essa abordagem realiza a junção de 38.000 redes do mundo real e 2,5 milhões de grafos, para analisar a acensão e queda na a relevância de vértices em sistemas dinâmicos, considerando, dessa forma, inúmeros contextos para classificar os nós mais significativos de todo o cenário. O trabalho é aplicável a sistemas dinâmicos na natureza e na sociedade e utiliza uma abordagem baseada no tempo, onde o modelo é capaz de entender como a relevância dos nós aumenta e diminui nas redes sendo aplicável a sistemas dinâmicos na natureza e na sociedade.

Ziberna (2020) apresenta uma proposta de algoritmo baseado em k-means para modelagem por blocos de redes vinculadas e também discute as vantagens de lidar com uma coleção de redes em vez de uma. As chamadas redes vinculadas são uma coleção de redes disjuntas e conectadas. Exemplos de redes vinculadas incluem redes as multiníveis, ou multicamadas.

Em Hsieh e Magee (2010) encontra-se uma proposta de raqueamento utilizando grupos e estruturas hierárquicas. O trabalho utiliza o agrupamento de nós (método k-means) para decompor uma rede social tendo como critério a existência de perfis congruentes e dissimilaridades com outros nós. Portanto, as propostas baseadas em grupos com uma hierarquia associada podem ser úteis para classificar os nós de acordo com sua relevância, ranqueando os nós presentes no núcleo das redes como sendo de maior relevância do que os nós associados à periferia das redes (Borgatti e Everett, 2007). Dessa forma, o uso das medidas de centralidade clássicas (Freeman, 1978; Otte e Rousseau, 2002; Bonacich, 1972, 2007; Beveridge e Shan, 2016) teve que ser revisto para cotemplar as múltiplas camadas presentes nessa nova propositura de estrutura de rede.

As redes multi camadas, dentre elas as redes multiplex, foram o fator catalizador

para o surgimento de novas centralidades tais como a *Novel Multiplex PageRank in Multilayer Networks*(Tu et al., 2018) (seção 2.5.6), *Random walk centrality in interconnected multilayer networks* (Solé-Ribalta et al., 2016) e a *Random Walks on Multiplex Networks: Supplementary Information for Navigability of Interconnected Networks under Random Failures* (De Domenico et al., 2014); sendo todas elas extensões de centralidades clássicas. Complementarmente surgiram novas estratégias pensadas de forma exclusiva para essas redes como a a CLDC (Bródka et al., 2012), o modelo NDNS (De Figueirêdo et al., 2020) e a próprima GCMN (De Figueirêdo et al., 2021), proposta nesta tese.

Também encontram-se propostas de estratégias inéditas de classificação voltadas a objetivos específicos como as redes de mobilidade urbana (Nanni et al., 2020). Para esse caso de uso os autores localizam lugares em uma cidade projetando grafos direcionados ponderados cujos nós denotam localizações de cidades e as arestas ponderadas representam o número de viagens entre eles. Os atributos dos nós indicam características socioeconômicas em um determinado local da cidade e combinam essas informações com "postos" de diferentes tipos de atividades socioeconômicas.

7.3 Métrica de avaliação proposta

A Centralidade GCMN baseia-se na distribuição dos nós por grupos de acordo com sua relevância. As outras medidas de centralidade com as quais compara-se a GCMN não têm esse conceito. Desta forma, para efeitos de comparação, será utilizada a metodologia de agrupamento de nós descrita na seção 4.1.1.

Tomando-se então o agrupamento de nós proposto pela Centralidade GCMN e aplicando-se a metodologia de agrupamento de nós ao estudo de caso (seção 4.1.1), tem-se para o grupo G_7 vinte e seis nós, toma-se então os primeiros vinte e seis nós mais bem classificados para todas as outras medidas de centralidade. Desta forma, é possível comparar o grupo G_7 com os nós melhor classificados em todas as outras medidas de centralidade. O grupo G_6 tem treze nós, considera-se então os nós classificados entre vinte e sete e trinta e nove em todas as outras medidas como sendo os seus respectivos grupos G_6. Este processo é o mesmo para os grupos G_2 a G_5.

Determinar uma métrica que mostre qual personagem seria mais relevante do que outro foi um desafio, pois trata-se de uma percepção empírica e dependente de fatores subjetivos como a empatia. No entanto, esta determinação foi necessária como parâmetro comparativo entre a classificação de centralidade GCMN e as medidas de centralidade.

Para tanto, foi assumido como premissa que a popularidade das personagens é diretamente proporcional à classificação fornecida pelo site *ranker.com*. Trata-se de um site especializado na classificação de diversas atividades e assuntos como entretenimento, música, esportes e história. É considerado um dos 10 melhores no ranking mundial de sites dessa categoria, com milhares de acessos diários. Criado em 2009, possui cerca de 250 milhões de opiniões registradas sobre cerca de um milhão de itens (SimilarWeb, 2021). No caso específico das personagens de Harry Potter, as avaliações de popularidade variaram de 1 a 99, sendo 1 a personagem mais relevante e 99 o de menor relevância. Embora o número total de personagens seja muito superior (468), as demais personagens não se encontram ranqueadas, sendo consideradas pela nossa análise como de menor relevância.

Para fins de comparação, define-se a medida de significância de um Grupo $SG(G_i)$, como a relevância de um grupo G_i de acordo com a classificação fornecida por uma centralidade. Ou seja, toma-se todo o conjunto de nós $\{v_j | v_j \in G_i\}$ para cada grupo G_i construído a partir das classificações fornecidas por cada uma das centralidades e, dessa forma, se obtém a significância de cada grupo para cada centralidade. A significância é definida como:

$$SG(G_i) = 1 - \frac{\sum S_v}{99 \times |G_i|} | v \in G_i, \tag{7.1}$$

onde S_v é o ranqueamento de um nó v, fornecido pelo site *ranker.com*. Não se deve confundir esse ranqueamento com o fornecido por cada centralidade, dado R_v (Equação 3.5). Dessa forma, a significância do grupo G_i, $SG(G_i)$, é dada pela razão entre a soma das classificações $\sum S_v$ e o valor máximo possível da soma das classificações dos nós do grupo G_i. Ou seja, o número de nós do grupo $|G_i|$ multiplicado pelo maior ranqueamento possível, no caso, 99. É fundamental esclarecer que um ranqueamento inferior representa o nó mais relevante, ou seja, um nó v com ranqueamento $S_v = i$ é melhor avaliado que um nó j com ranqueamento $S_j = i + 1$, dessa forma a Equação 7.1 utiliza o complemento de um para determinar a significância do grupo G_i.

Considerando essa métrica proposta, realizar-se-á um conjunto de análises comparando o ranqueamento proposto pela centralidade GCMN com os fornecidos das demais centralidades (Seção 7.4).

7.4 Resultados e Discussão

Na análise dos resultados e sua discussão considerarão alguns parâmetros: o uso de uma métrica proposta nesta tese (seção 7.3) para avaliar o desempenho do CGMN em dois

aspectos: o peso como um parâmetro de agrupamento (seção 7.4.1) e a relevância por grupo (seção 7.4.2); a análise qualitativa dos grupos (seção 7.4.3) e o uso de métricas conhecidas como *Accuracy, Precision, Recall* e F_1 *Score* e as Correlações de Pearson e Spearman (seções 7.4.4 e 7.4.5).

A Figura 7.1 traz a rede complexa relativa à junção de todos os sete livros da coleção de estórias do Harry Potter. A rede foi cedida pelos autores de Carvalho (2017) e nela pode-se verificar visualmente a existência de um núcleo central em torno da personagem "Harry". A escolha natural para a segmentação dessa rede em camadas foi a alocação de cada um dos livros como um camada de uma rede multiplex não direcionada.

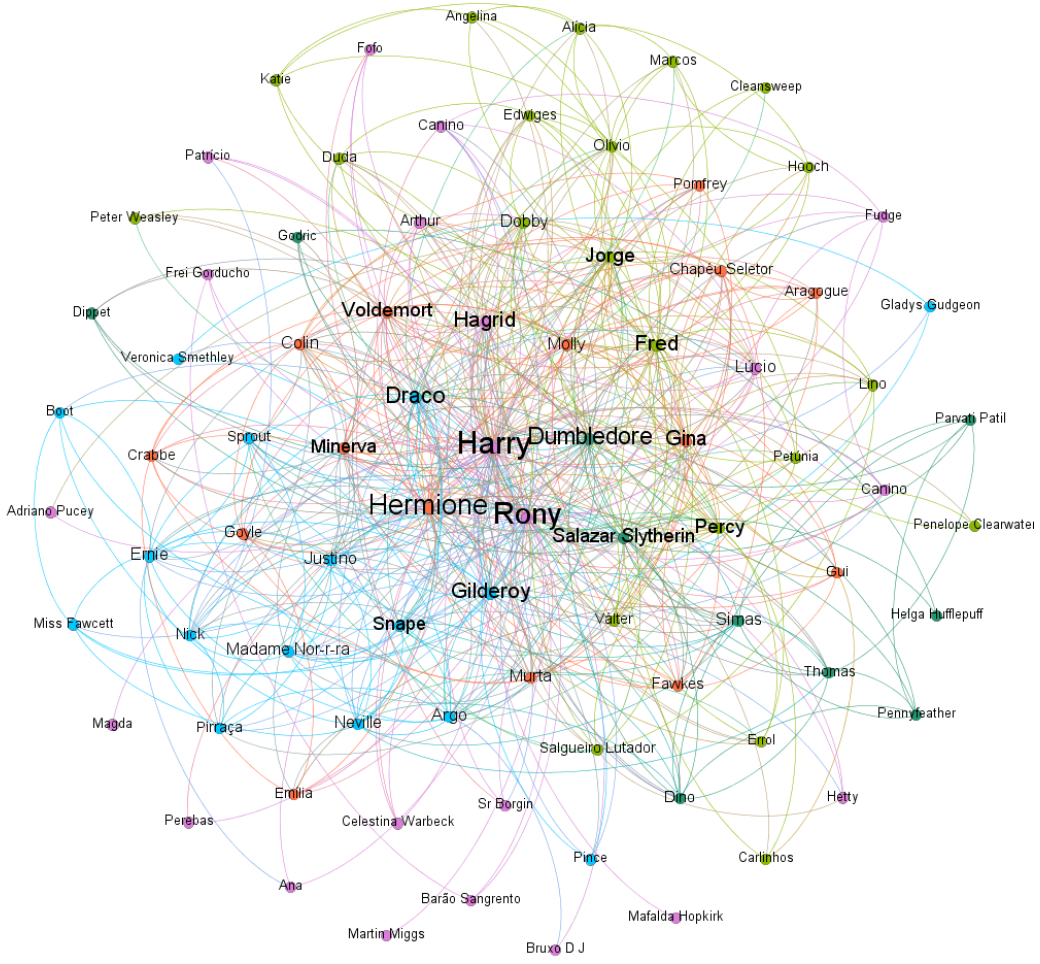


Figura 7.1. Rede complexa do estudo de caso: coleção de livros do Harry Potter.

A Figura 7.2 aponta os nós em comum entre duas das sete camadas da rede multiplex, representando os livros "Harry Potter e a Pedra Filosofal" e "Harry Potter

e a Câmara Secreta", essa quantidade serve como base para o cálculo do peso dos nós (equação 3.2) e na consequente formação dos grupos (equação 3.3). A Figura 7.3 traz uma visão ampla da rede expondo a quantidade de nós em comum entre as sete camadas da rede multiplex, cada uma correspondendo a um dos livros da coleção. Observa-se que todas as camadas têm nós em comum com todas as demais o que era esperado uma vez que um grupo de personagens encontra-se presente em todos os livros da coleção. Essas personagens pertencem ao ao grupo G_7 e encontram-se relacionadas na Tabela 7.1. Além disso verifica-se que a quantidade de personagens em comum é variável quando as camadas são tomadas em pares, e.g. a quantidade de nós em comum nos livros 1 e 2 não é a mesma da dos livros 2 e 3, e assim por diante. Essa variação ocorre pela dinamicidade com que as personagens participam de algumas etapas da estória e não atuam em outras, restringindo-se a alguns livros.

A divulgação dos nomes das personagens mais relevantes na análise, relativas ao grupo G_7 (Tabela 7.1), objetiva proporcionar uma análise comparativa entre o raqueamento fornecido pelo site *ranker.com* e o calculado pela centralidade GCMN, que levou a um desvio padrão de 2,27, o que demonstra haver coerência e concordância entre as duas classificações.

7.4.1 O peso como um parâmetro de agrupamento

Quanto à significância (Equação 7.1) média associada a cada grupo, calculada como a média das significâncias obtidas por todas as centralidades em cada grupo (Tabela 7.2), verifica-se que, como esperado, ela é a mais baixa para os grupos G_2 e G_3, com valores semelhantes; crescendo, como previsto, para os grupos G_4 e G_5 e obtendo o seu maior valor para o grupo G_7. A única exceção é o grupo G_6 que teve uma significância abaixo do esperado, sendo inferior a G_4 e G_5.

Após essa análise, verifica-se que, apesar da exceção encontrada no grupo G_6, o crescimento à importância associada aos nós é diretamente proporcional ao grau de relevância das personagens considerando-se todas as centralidades analisadas e não apenas a GCMN. Esse fato demostra que o peso é uma escolha coerente como parâmetro para construir grupos, a exemplos das conclusões já obtidas nos estudos de caso anteriores (seções 5.4.1 e 6.4.1).

7.4.2 Relevância por Grupo

Considerando o agrupamento de nós proposto pela Centralidade GCMN, a Tabela 7.2 traz uma análise comparativa das centralidades e dos grupos (Equação 7.1). a

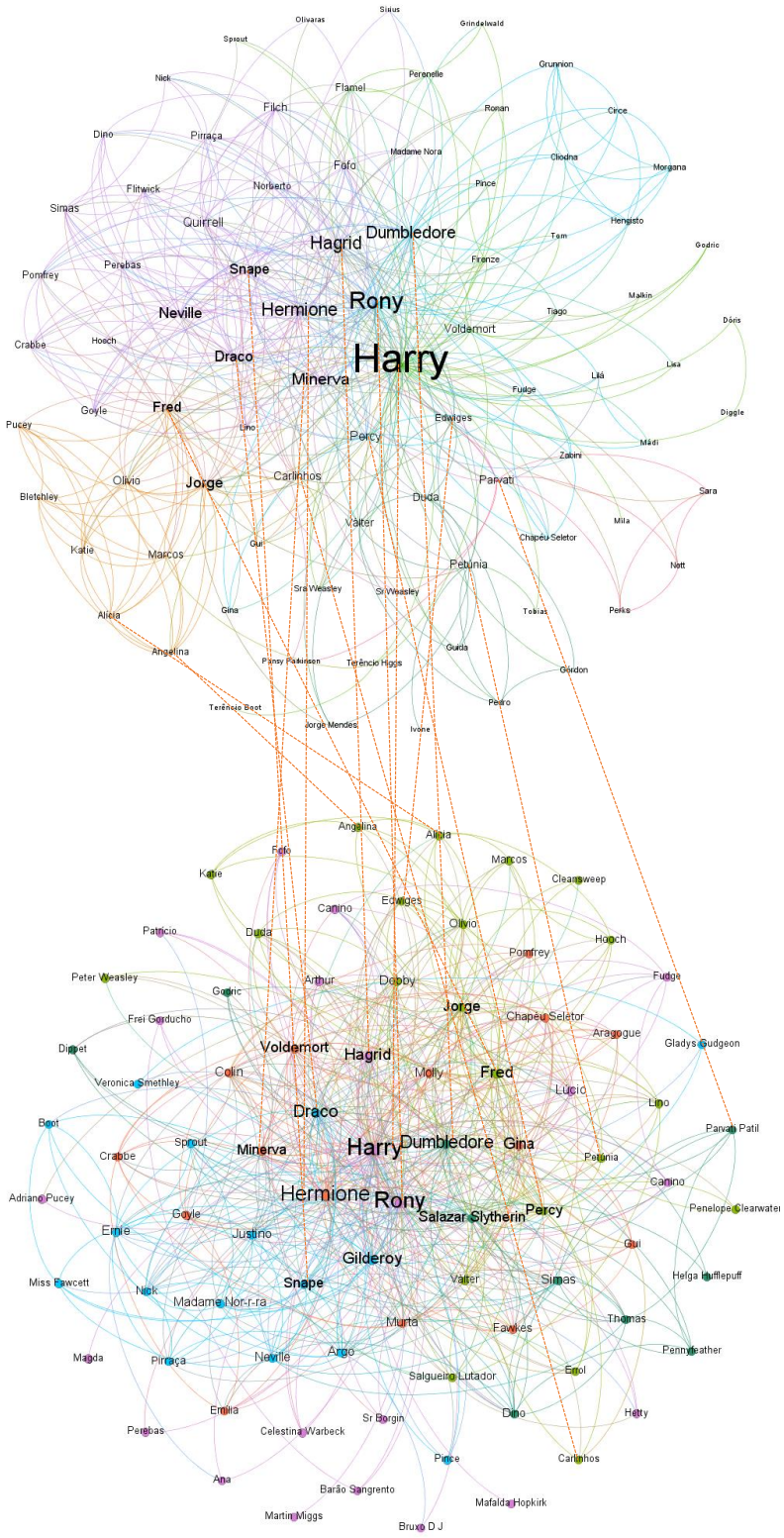


Figura 7.2. Nós em comum entre duas camadas da Rede Multiplex, representando os livros "Harry Potter e a Pedra Filosofal" e "Harry Potter e a Câmara Secreta" (Estudo de Caso - Harry Potter).

Tabela 7.1. Ranqueamento das personagens do grupo G_7 Centralidade GCMN x *ranker.com*.

Personagem	*ranker.com*	GCMN
Hermione	1	3
Voldemort	2	4
Snape	3	5
Draco Malfoy	4	6
Harry Potter	5	1
Hagrid	6	7
Nevile Longbottom	7	10
Ginny Weasley	8	8
Fred	9	9
Rony	10	2
Minerva McGonagall	11	12
Sirius Black	12	11
Remo Lupin	13	14
Percy	14	15
Molly	15	15
Simas Finnigan	16	17
Dino Thomas	17	16
Petunia	18	21
Lino Jordan	19	19
Goyle	20	18
Charlie Weasley	21	23
Vernon Dursley	22	20
Duda Dursley	23	22
Peeves	24	24
Nick	25	25
Sorting Hat	26	26

Tabela 7.2. Relevância por Centralidade/Grupo.

Centralidade/Grupos	G_2	G_3	G_4	G_5	G_6	G_7
GCMN	0.353	0.459	0.833	0.729	0.448	0.712
WD	0.432	0.447	0.582	0.707	0.383	0.692
CL	0.419	0.430	0.613	0.710	0.512	0.644
BW	0.420	0.425	0.616	0.651	0.539	0.708
PR	0.409	0.421	0.494	0.494	0.529	0.663
EI	0.518	0.382	0.760	0.665	0.548	0.626
$\frac{\sum SG(G_i)}{6}$	42.9	41.1	65.0	65.9	49.3	67.4

Centralidade GCMN obteve desempenho superior nos grupos G_3, G_4, G_5 e G_7. A centralidade de *Eigenvector* (EI) obteve o melhor desempenho nos grupos G_2 e G_6. O resultado mostra que a Centralidade GCMN atingiu os nós mais significativos em quatro dos seis grupos, obtendo o melhor desempenho geral.

7.4.3 Análise Qualitativa dos Grupos

A análise qualitativa pretende verificar qual métrica é capaz de apontar "novidades", ou seja, nós que foram apontados por apenas uma centralidade. Esta análise é particularmente útil para que os nós, até agora, "não descobertos", possam ser alcançados.

A Tabela 7.3 mostra que a centralidade GCMN foi capaz de apontar a maior quantidade (sete) de nós não detectados pelas outras medidas de centralidade, representando 53% do total. Além disso, foi a única centralidade capaz de apontar novidades entre as personagens ranqueadas entre os 99 apontados pelo site *ranker.com* (2 nós).

A exemplo da seção 5.4.4, atribui-se esse resultado ao uso, pela GCMN, de uma estratégia completamente diferente daquelas usadas por medidas de centralidade como *betweenness* (BC), *eigenvector* (EI), *closeness* (CL) e *weighted degree* (WD) que, em algum ponto, têm premissas baseadas em conceitos semelhantes (Bonacich, 1972, 2007; Otte e Rousseau, 2002; Freeman, 1978; Beveridge e Shan, 2016).

7.4.4 *Accuracy, Precision, Recall* e F_1 *Score*

Os conceitos de *Accuracy, Precision, Recall* e F_1 *Score* são métricas estatísticas que já foram discutidas na seção 4.1.2. Esses conceitos serão utilizados nesta seção deste estudo de caso.

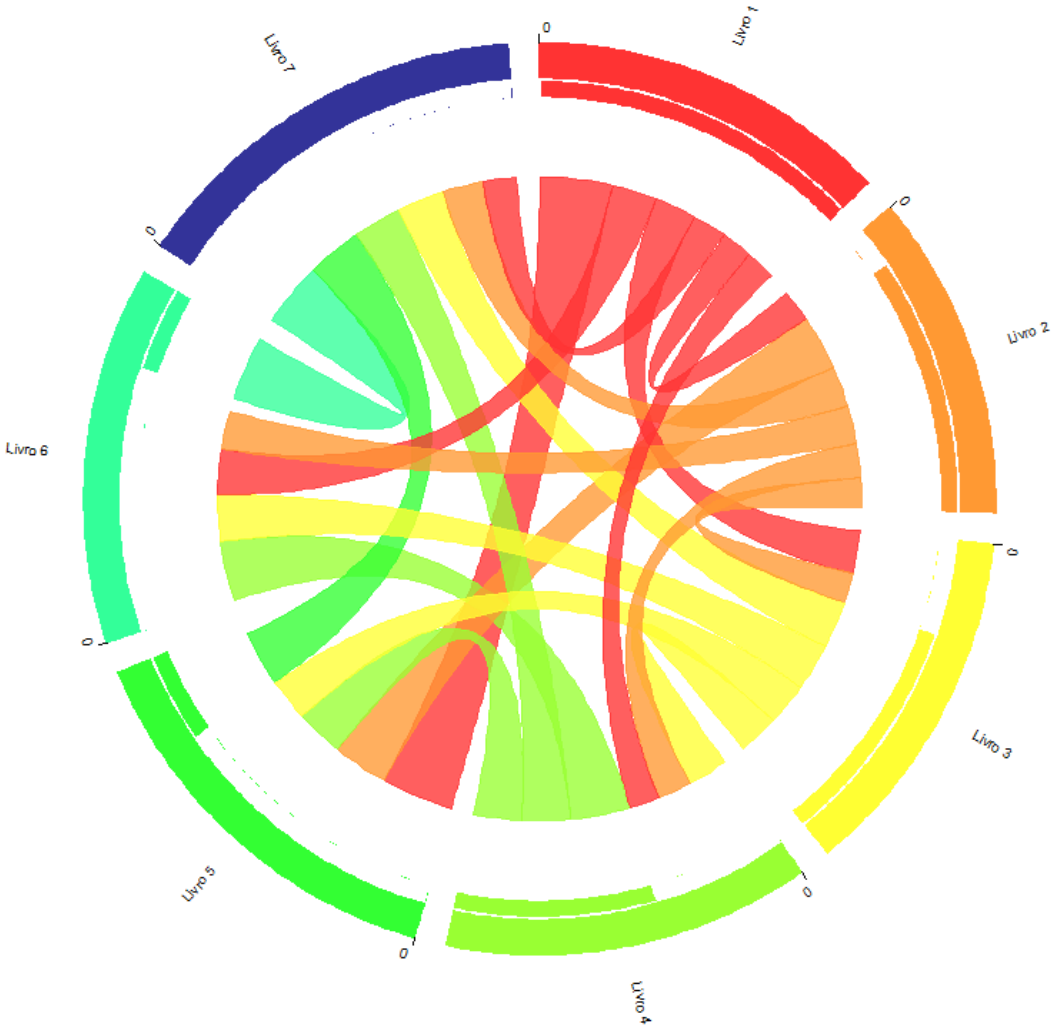


Figura 7.3. Quantidade de nós em comum entre as camadas da Rede Multiplex (Estudo de Caso - Harry Potter).

Tabela 7.3. Análise qualitativa dos grupos.

	GCMN	WD	CL	BW	PR	EI
$Rank < 100$	2	0	0	0	0	0
$Rank \geq 100$	5	1	0	1	3	1
Total	7	1	0	1	3	1

Considerando o estudo de caso, tem-se um total de 468 personagens, sendo 99 classificadas pelo site *ranker.com* e a mesma quantidade por cada uma das centralidades em análise. Tomando cada uma das centralidades, considera-se as personagens classificadas pelo site e pela centralidade como $thetruepositives$; as personagens não classificados pelo site, mas classificados pela centralidade como $thefalsepositives$; as personagens classificados pelo site, mas não classificados pela centralidade como $thefalsenegatives$ e as personagens não classificadas pelo site nem pela centralidade como $thetruenegatives$ (Tabela 7.4).

Tabela 7.4. Agrupamento de indivíduos para análise de precisão e recall.

	ranker.com	
Centralidade	Ranqueamento ≤ 99	Não Ranqueado
Ranqueamento ≤ 99	$truepositives$	$falsepositives$
Não Ranqueado	$falsenegatives$	$truenegatives$

A análise da *Precision* mostra o percentual de indivíduos detectados por cada centralidade e que são relevantes para a análise. Assume-se como relevantes apenas os indivíduos classificados por *ranker.com*. Dessa forma a *Precision* pode indicar qual centralidade deve ser usada para atingir o melhor resultado, uma vez que mostra percentualmente a interseção entre os dois conjuntos, ou seja, os nós tanto detectados pela centralidade quanto classificados pelo site *ranker.com*. Na análise o *Weighted Degree* (WD) atingiu uma *Precision* de 64%, seguido pela centralidade GCMN com 60%, sendo essas as duas centralidades mais bem posicionadas (Figura 7.4).

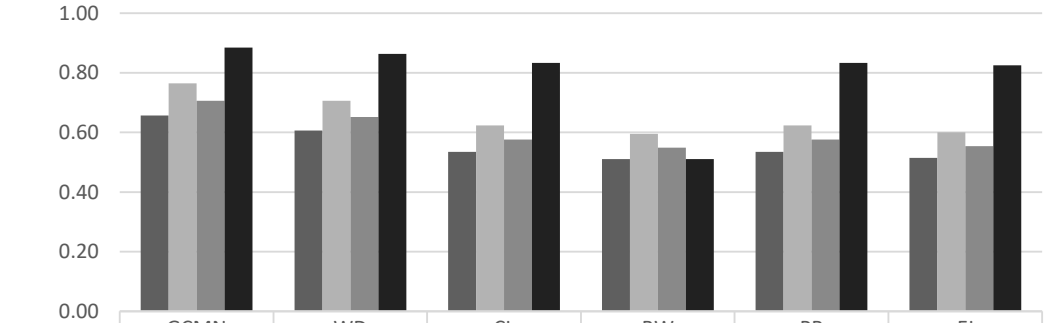
A análise de *Recall* mostra qual centralidade foi capaz de apontar para um número mais significativo de personagens relevantes, considerando todos os indivíduos e não apenas aqueles classificados pelo site *ranker.com*. Assim, como a análise da *Precision*, a *Recall* também é útil ao escolher qual centralidade utilizar. Na análise a GCMN atingiu 76%, contra 71% da *Weighted Degree* (WD), que foi a segunda melhor classificada (Figura 7.4).

A análise do F_1 traz uma visão geral do desempenho das duas métricas (*Precision* e *Recall*), mostrando que a centralidade GCMN alcançou um melhor resultado geral. A GCMN foi a centralidade melhor posicionada, atingindo 71%, contra 65% da segunda medida de centralidade mais bem posicionada, o *Weighted Degree* (WD) (Figura 7.4).

A análise da *Acuraccy* demonstrou que a centralidade GCMN atingiu o percentual de 88%, sendo a melhor posicionada, seguida por WD (86%) e, empatadas em terceira lugar, a CL e a PR com 83%. Um resultado é considerado de boa acurácia quando

consegue atingir um alvo específico. No estudo de caso, o alvo são as personagens classificadas por *ranker.com*.

Numa análise geral a Centralidade GCMN atingiu um bom grau de *Precision* (66 %) e *Recall* (76 %) com alta *Acuraccy* (88 %), fazendo com que o ranqueamento proposto pudesse apontar para os personagens de forma precisa, alcançando a meta com alta acurácia, sendo o melhor desempenho dentre todas as medidas de centralidade analisadas (Figura 7.4).



	GCMN	WD	CL	BW	PR	EI
Precision	0.66	0.61	0.54	0.51	0.54	0.52
Recall	0.76	0.71	0.62	0.60	0.62	0.60
F1	0.71	0.65	0.58	0.55	0.58	0.55
Accuracy	0.88	0.86	0.83	0.51	0.83	0.82

Figura 7.4. Análise comparativa de *Accuracy Precision Recall* e F_1.

A Figura 7.5 traz uma análise comparativa entre a F_1 e a *Accuracy*. Como F_1 é uma visão geral do desempenho de *Precision* e *Recall*, a análise pode mostrar o desempenho geral das centralidades. O gráfico mostra a classificação normalizada das medidas de centralidade, de acordo com a distância entre os pontos cartesianos de cada centralidade até o ponto de desempenho máximo $m = (1, 1)$. Assim, quanto mais próximo o ponto de medida de centralidade $c = (x, y)$ estiver do ponto de desempenho máximo m, melhor será sua performance. Considerando a distância entre $z = (0, 0)$ e m como $d_{zm} = \sqrt{(1-0)^2 + (1-0)^2}$ a distância máxima possível no plano cartesiano, subtrai-se deste valor máximo a distância obtida por cada centralidade, assim $d_{cm} = \sqrt{(1-x)^2 + (1-y)^2}$ alcançando a classificação por *Accuracy* e F_1 dada por $d_{zm} - d_{cm}$.

As figuras 7.4 e 7.5 demonstram que a centralidade GCMN atingiu o melhor desempenho geral em comparação com as outras cinco medidas de centralidade, sendo o Grau Ponderado (WD), a segunda centralidade mais bem posicionada. Considera-se este um bom resultado, mostrando que a centralidade proposta pode atingir um desempenho competitivo quando comparada com outras centralidades.

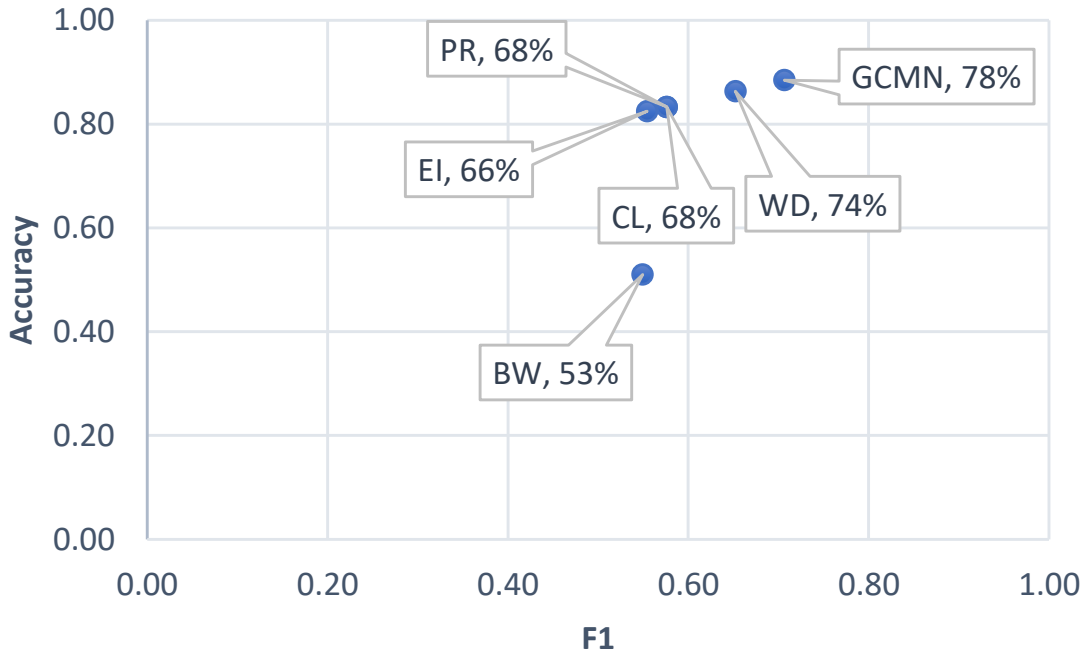


Figura 7.5. Análise comparativa - *Accuracy* e F_1.

7.4.5 Análise das Correlações de Pearson e Spearman

No estudo de caso o uso da análise estatística é uma forma de demonstrar a coerência do uso da distribuição por grupos, para classificar os nós de acordo com sua suposta relevância. Pretende-se demonstrar a correlação entre os grupos propostos na centralidade GCMN e a relevância das personagens, verificando como essa relevância tende a crescer com a significância dos grupos. A força de uma relação entre essas duas variáveis é capaz de demonstrar a coerência do agrupamento dos nós proposto pela centralidade GCMN.

A exemplo do que foi analisado na seção 6.4.4, considerando que a correlação proposta por Spearman (Spearman, 1904) trata de funções monotônicas, ela é aplicável também ao estudo de caso, por se tratar de uma função monotônica de crescimento estrito, ou seja, $\forall x, y \in A, (x > y \Rightarrow f(x) \geq f(y))$, onde x pode representar a relevâncias das personagens e y o ranqueamento fornecido pelas centralidades analisadas. Dessa forma, o coeficiente de correlação de Spearman é capaz analisar a intensidade e a direção dessa relação monotônica, lembrando que, em um relacionamento monotônico, as variáveis tendem a se mover na mesma direção relativa, mas não necessariamente a uma taxa linear.

De forma semelhante ao tratado na seção 6.4.4, como os coeficientes de correlação de Pearson e Spearman são medidas estatísticas da força de uma relação entre dados

Tabela 7.5. A análise de correlações de Pearson e Spearman.

Centrality	Pearson		Spearman	
	r	p-value	r_s	p-value
GCMN	0.999	0.025	1.000	0.333
WD	0.938	0.226	1.000	0.333
BW	-0.882	0.312	-1.000	0.333
PR	-1.000	0.013	-1.000	0.333
EI	-0.759	0.442	-0.500	1.000
SC	-0.515	0.656	-0.500	1.000
CLCD	0.983	0.118	1.000	0.333
RW	0.938	0.225	1.000	0.333
MPR	0.926	0.247	1.000	0.333

pareados e, considerando que, nesse estudo de caso, lida-se com duas variáveis que deveriam, na situação ideal, ter uma correlação linear e monotônica. Ou seja, à medida que uma cresce, a outra também deveria crescer linear e continuamente. Utiliza-se ambos os coeficientes para avaliar as relações.

A tabela 7.5 traz os coeficientes de correlação de Pearson e Spearman para cada centralidade. Considerando a classificação fornecida por *ranker.com* e as classificações das seis centralidades analisadas, a centralidade GCMN obteve as melhores pontuações na avaliação dos métodos.

A análise é confirmada visualmente na Figura 7.6. Nela é possível verificar que a distribuição e a concentração dos nós em torno da reta que significaria a distribuição ideal e linear, é bem mais clara na Centralidade GCMN que nas demais.

Considerando que a análise dos coeficientes de Pearson e Spearman deve considerar a significância (valor p) para validar a análise das correlações. Verifica-se que, para o estudo de caso, a GCMN atingiu, em ambas as correlações, um p-value $= 0,001$, ou seja, isso indica serem ambas as correlações estatisticamente significativas.

7.5 Considerações Finais

Considerando o estudo de caso da Coleção de Livros do Harry Potter, a centralidade GCMN provou ser superior a cinco medidas de centralidade: *Weighted Degree*

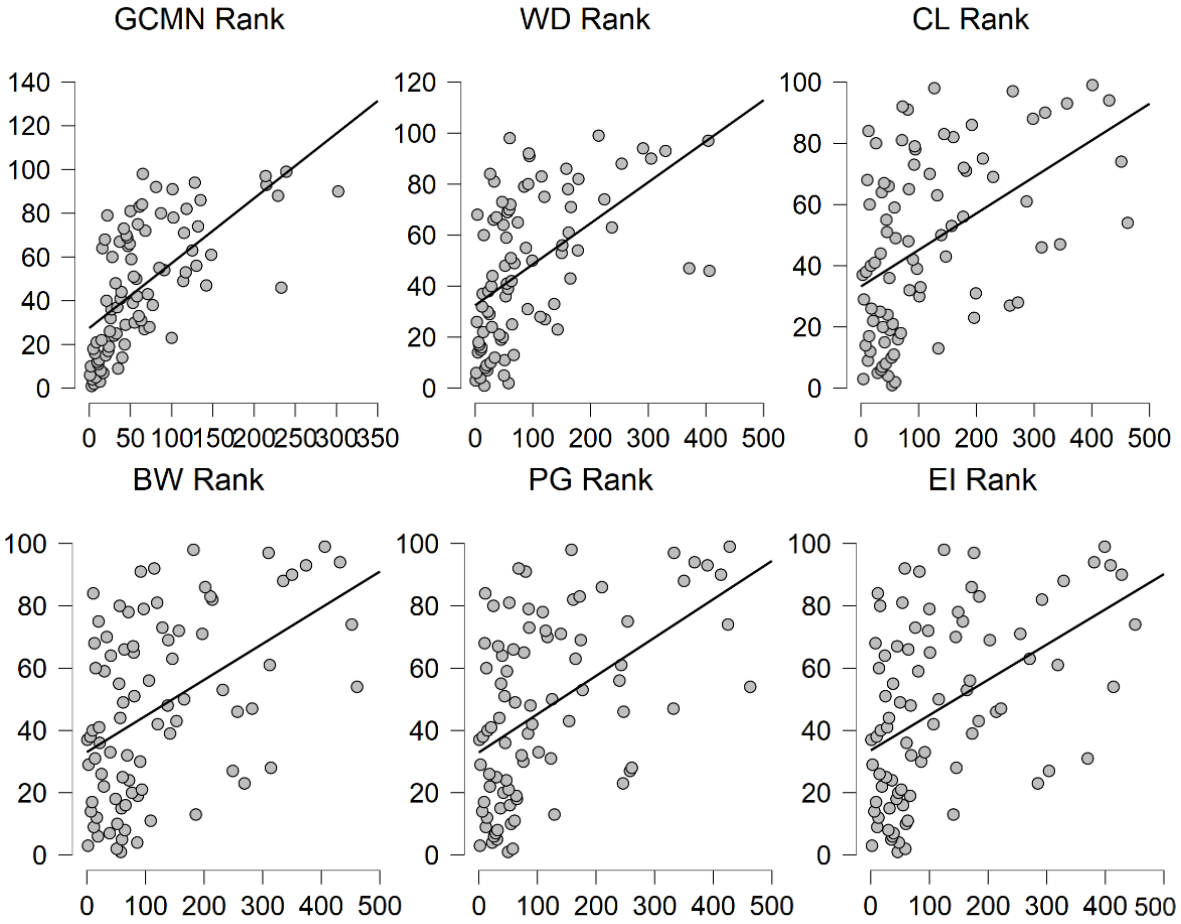


Figura 7.6. Distribuição dos nós por Centralidade.

(WD)(Beveridge e Shan, 2016), *Betweenness* (BW)(Freeman, 1978; Otte e Rousseau, 2002), *PageRank* (PR)(Bonacich, 2007), *Eigenvector* (EI)(Bonacich, 1972) e *Closeness* (CL) (Freeman, 1978); na detecção de personagens relevantes. Esta análise foi feita de forma segmentada por grupos de nós — G_2 a G_7 —, cada um representando um livro da coleção, nos quais a centralidade GCMN apresentou os seguintes resultados:

- Foi utilizado peso (Equação 3.2) como critério para distribuir os nós em grupos (Equação 3.3). A análise na seção 7.4.1 demostrou que esta foi a escolha certa uma vez que a relevância dos nós cresce à medida que os grupos também aumentam sua suposta relevância (de G_2 para G_7);

- Quanto à análise da relevância por grupo, a GCMN atingiu os nós mais significativos em quatro dos seis grupos, obtendo o melhor desempenho geral (seção 7.4.2);

- Quanto à descoberta de "novidades", a análise qualitativa demonstrou que a Centralidade GCMN foi capaz de apontar 53% de todos os nós não alcançados por outras centralidades, sendo a única a encontar esses nós dentre os raqueados pelo site *ranker.com* (seção 7.4.3);

- O ranqueamento proposto pela Centralidade GCMN alcançou 66% de *Precision*, 76% de *Recall*, 71% de F_1 e 88% de *Acuraccy*; na detecção dos caracteres mais relevantes considerando o conjunto de livros, sendo a centralidade com a melhor performance dentre as avaliadas (seção 7.4.4);

- Por fim, a centralidade GCMN obteve o melhor resultado na análise de correlações de Pearson e Spearman. Comparada com as outras centralidades, foi capaz de trazer a correlação mais coerente com a classificação fornecida por *ranker.com* (Seção 7.4.5).

O próximo capítulo trará as conclusões e contribuições da tese, as limitações da proposta apresentada a proposta de trabalhos futuros a lista as publicações obtidas a partir desde trabalho.

Capítulo 8

Conclusões

Como principal contribuição, esta tese propôs a Centralidade GCMN, que trata de uma nova proposta de classificação de nós em redes complexas multiplex. A nova centralidade visa ser uma alternativa às medidas de centralidade clássicas e voltadas às redes multiplex, na detecção dos nós mais relevantes em determinado contexto. A seguir, é apresentado um resumo das principais contribuições, limitações da proposta, oportunidades de futuras investigações e uma lista das publicações advindas do trabalho desenvolvido.

8.1 Principais Contribuições

Primeiramente, a Centralidade GCMN, proposta nesta tese, não foi concebida com base em nenhuma proposta de trabalho anterior, não sendo uma extensão ou melhoria de uma centralidade. Dessa forma a proposta abre uma linha pensamento alternativa às propostas anteriores, sendo, portanto, um campo a ser explorado e melhorado.

Foi proposto um novo conceito de classificação de nós baseado no agrupamento desses nós utilizando dois critérios de forma hieráquica (Capítulo 3). De forma mais específica, a tese desmonstrou a viabilidade do uso da separação de nós em grupos formados de acordo com a incidência desses nós nas camadas de uma rede complexa multiplex, para efeitos de sua classificação. Dessa forma, a separação em grupos utiliza como primeiro critério o peso do nó (Equação 3.2) que nada mais é qua a quantidade de vezes que um nó v apararece em cada uma das camadas de uma rede complexa multiplex e como segundo critério de classificação dentro do grupo, o grau do nó (Equação 3.4). A validade dessa premissa foi comprovada em três estudos de caso (seções 5.4.1, 5.4.2, 5.4.3, 6.4.1, 6.4.2, 7.4.1 e 7.4.2), onde, em situações distintas, foi demonstrado ser essa uma estratégia viável de classificação.

Em segundo lugar, por ser uma proposta não atrelada a tecnologias já propostas, a GCMN se mostrou capaz de encontrar nós relevantes de uma forma diversa às centralidades estudadas. Ou seja, por tratar de uma forma completamente diferente e abordar o tema de encontrar nós relevantes, a nova centralidade também aponta nós não detectados por outras centralidades (seção 2.5). Exemplos desse tipo de análise podem ser encontrados nas seções 5.4.4 e 7.4.3. Ressalte-se que a seções tratam da mesma análise em dois estudos de caso distintos, aferindo, dessa forma, ser essa uma característica intrínseca à centralidade GCMN. A análise das tabelas 5.5 e 7.3 demonstra que a centralidade GCMN foi capaz de sozinha, detectar essa chamadas "novidades" em 46% dos casos, quando comparadas às demais centralidades em dois estudos de caso distintos.

Finalmente, foram utilizadas métricas estatisticamente comprovadas como as medidas de *Accuracy, Precision, Recall*, F_1 *Score* e as correlações propostas por Pearson e Spearman como parâmetros para aferição da performance da centralidade proposta. Foi realizada uma comparação de sua classificação com as fornecidas por outras centralidades em três estudos de caso (seções 5.4.5, 6.4.3, 6.4.4, 7.4.4 e 7.4.5). Em todas as análises a Centralidade GCMN ou apontou os melhores resultados ou esteve entre as melhor avaliadas.

A figura 8.1 traz a análise consolidada da *Accuracy, Precision, Recall* e F_1 *Score* para os três estudos de caso: Operação Lava Jato, Operação Licitante Fantasma e Coleção dos livros de Harry Poter, respectivamente.

Analisando a figura 8.1 verifica-se que a centralidade GCMN demonstrou ser a mais precisa dentre todas as centralidades, obtendo resultados de 90%, 92% e 60%, ou seja, foi a primeira colocada nos dois primeiros estudos de caso e manteve-se na média no último. Esses resultados demonstram que a centralidade GCMN foi capaz de apontar o maior percentual de nós relevantes, considerando como universo apenas os nós considerados relevantes. A precisão deve ser usada em situações em que os falsos positivos são considerados mais prejudiciais que os falsos negativos, ou seja, quando apontar, de forma errônea um nó como relevante, sendo isso uma falsa afirmação é pior que não apontar um relevante. Essa métrica é extremamente importante, em particular nos estudos de caso que envolvem questões legais (Operação Lava Jato e Licitante Fantasma), uma vez que o objetivo é apontar os culpados de forma mais precisa possível, sendo muito ruim apontar alguém como suspeito que não tenha relevância.

Quanto ao *Recall*, a GCMN obteve um resultado fraco no primeiro estudo de caso (51%) e foi a segunda colocada nos dois seguintes (94% e 68%). O *recall* pode ser usada em uma situação em que os falsos negativos são considerados mais prejudiciais que os falsos positivos. Dessa forma, considerando-se os estudos de caso "Operação

Lava Jato" e "Operação Licitante Fantasma", como apontar-se, de forma errônea, um suspeito como inocente, apesar de grave, tem um impacto menor que colocar-se como suspeito um inocente, o *recall* torna-se menos importante que a *precision*. A análise do F_1 *Score* traduz a superioridade da centralidade quanto à análise conjunta de *Precision* e *Recall*, colocando-a em primeiro lugar nas duas primeiras análises (65% e 93%) e na média na última (63%).

Finalmente, quanto à *Accuracy*, a centralidade foi, respectivamente primero, terceiro e segundo colocada (79%, 55% e 85%). A *Accuracy* é uma boa indicação geral de como a centralidade GCMN performou. Porém, pode haver situações em que ela é enganosa. Por exemplo, na criação de um modelo de identificação de fraudes em cartões de crédito, o número de casos considerados como fraude pode ser bem pequeno em relação ao número de casos considerados legais. Para colocar em números, em uma situação hipotética de 280000 casos legais e 2000 casos fraudulentos, um modelo simplório que simplesmente classifica tudo como legal obteria uma acurácia de 99,3%. Ou seja, você estaria validando como ótimo um modelo que falha em detectar fraudes. Nos estudos de caso trazidos à tese esse problema não ocorreu uma vez que os critérios de aferição de falsos, tanto negativos como positivos, na "tabela de confusão", foram adotados tomando-se critérios objetivos claros (seções 5.4.5, 6.4.3 e 7.4.4) o que leva a uma *Accuracy* válida.

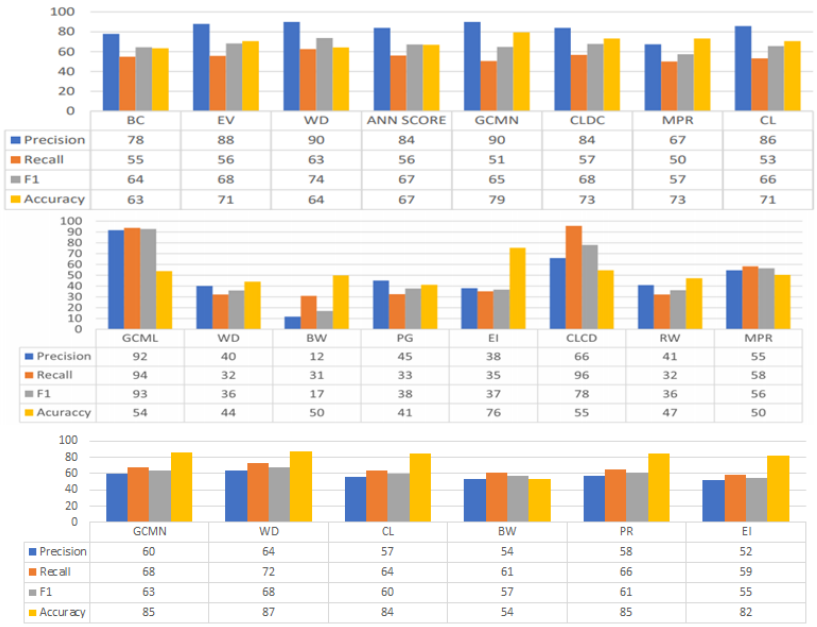


Figura 8.1. Análise consolidade de *Accuracy, Precision, Recall* e F_1 *Score*.

Quanto às correlações de Pearson e Spearman, no estudo de caso "Operação Licitante Fantasma", a centralidade GCMN alcançou as melhores pontuações na avaliação

da correlação entre o lucro da empresa e a relevância dos grupos. Essa correlação demonstra que a centralidade proposta aloca as empresas mais relevantes nos grupos também mais relevantes. Quanto ao estudo de caso da "Coleção de Livros do Harry Potter", a GCMN obteve o segundo melhor resultado, nas correlações de Pearson e o melhor na de Spearman. Dessa forma, considerado os dois estudos de caso, a GCMN obteve a melhor performance, sendo a centralidade que ofertou um ranqueamento mais coerente com relação ao crecimento das grandezas propostas nas métricas de avaliação.

8.2 Limitações da Centralidade GCMN

Por tratar-se de um proposta generalista, a GCMN tem algumas limitações inerentes à essa condição. Ou seja, dificilmente seria possível propor uma centralidade que funcionasse sempre de forma superior a todas as outras propostas de determinar os atores mais relevantes em um cenário, em especial aquelas pensadas para situações específicas. Vislumbram-se, portanto, as seguintes limitações:

- Por ser uma proposta concebida para tratar de uma rede multiplex que tem múltiplas camadas e, a partir dessas camadas, determinar os nós mais relevantes, ela é também restrita a essa condição. Ou seja, apesar de ter uma proposta generalista, a centralidade não é útil a qualquer tipo de situação, sendo apenas aplicável a cenários mapeados por uma rede complexa multiplex. Dessa forma a centralidade tem a sua atuação e aplicabilidade mais restrita que as medidas de centralidade clássicas ou mesmo, a depender da aplicação, que soluções dedicadas a situações específicas;

- Como a GCMN trabalha com grupos compostos pelos nós presentes em cada camada de uma rede complexa multiplex, um caso extremo seria que todos os nós estivessem presentes em todas as camadas dessa rede. Nesse caso, a hierarquia de grupos não seria aplicável e todos os nós estariam no mesmo grupo. Nesse caso específico o único critério de classificação seria o grau do nó, o que levaria a uma classificação possivelmente ruim;

- Outro caso a ser esclarecido é o de nós isolados sem conexões com outros nós da mesma camada. Esses nós são considerados para classificação pelo GCMN. Essa situação pode representar uma distorção na classificação, uma vez que nós isolados poderiam, a princípio, ser considerados relevantes a depender do número de camadas em que se fizessem presentes. No entanto, assume-se que a probabilidade de um nó participar isoladamente em um grupo significativo de

camadas é mínima. Essa suposição advém do fato de que nós isolados repesentariam, em tese, entidades com pouca interação com as demais, sendo que, dessa forma, dificilmente esse tipo de nó teria uma participação mais efetiva em um número significativo de camadas que comporiam a modelagem do cenário. Naturalmente, deve-se ressalvar que cada situação fática tem suas nuances onde essa suposição pode não ser válida, entretanto essa possibilidade não foi o evidenciada nos estudos de caso deste trabalho (Capítulos 5, 6 e 7);

- A proposta de ranqueamento trazida pelo modelo prioriza a presença dos nós na maior quantidade de camadas possível de uma rede complexa multiplex. Esse aspecto, apesar de ter se mostrado eficaz nos estudos de caso, pode redundar em análises equivocadas dos nós mais relevantes em situações específicas onde a presença em uma quantidade grande de camadas seja pouco relevante frente a outras possíveis estratégias de classificação.

8.3 Oportunidades de Trabalhos Futuros

A seguir, são apresentadas oportunidades de trabalhos futuros derivadas das contribuições desta tese.

- A aplicação da Centralidade GCMN a outros estudos de caso para atestar sua eficácia e comparar a sua performance com medidas de centralidade baseadas em agrupamentos de nós (Agneessens et al., 2017; Borgatti e Everett, 2007; Hsieh e Magee, 2010);

- Desenvolvimento de um framework que implemente a Centralidade GCMN. Essa implementação poderia ser feita, inclusive, na forma de extensão a um software já existente. Uma sugestão seria a plataforma Gelphi (gelphi.org), que, por se tratar de um software OpenSource, viabiliza essa possibilidade (https://github.com/gephi/gephi);

- Utilização de outros estudos de caso, com rede complexas de grande porte, com milhares ou milhões de nós, para aferir a validade da Centralidade GCMN em grandes massas de dados;

- Novas propostas de ranqueamento que considerem os pesos das arestas como critério de classificação dos nós dentros dos grupos;

- Verificar o impacto do uso de outras funções para R_i e $\varphi(W_i)$ (Equações 3.5 e 3.6), segundo as possibilidades e restrições comentadas no final da introdução do Capítulo 3;
- Propor uma variação da centralidade GCMN considerando a possibilidade de agrupamento de camadas/grupos de nós, para estudos de casos onde o número de camadas seja muito grande;
- Verificar o nível de significância dos resultados obtidos utilizado a análise estatística dos erros do tipo 1 e 2;
- Uma análise comparativa de performance da Centralidade GCMN frente a sistemas especialistas voltados a soluções específicas.

Dessa forma considera-se relevante a contribuição da centralidade proposta esperando que trabalhos futuros possam ser desenvolvidos para o seu aprimoramento.

8.4 Publicações

- B. C. B. De Figueirêdo, F. G. Nakamura and E. F. Nakamura, "A Group-Based Centrality for Undirected Multiplex Networks: A Case Study of the Brazilian Car Wash Operation", in IEEE Access, vol. 9, pp. 81946-81956, 2021, doi: 10.1109/ACCESS.2021.3086027.
- FIGUEIREDO, Bruno; NAKAMURA, Fabiola; FELIX, Gardenya; NAKAMURA, Eduardo. Usando análises sociais na identificação de nós relevantes em um cenário multi-redes: Operação Licitante Fantasma, um estudo de caso. Em: Anais do VIII Workshop de Computação Aplicada em Governo Eletrônico (WCGE), 2020, Cuiabá. Sociedade Brasileira de Computação. pp. 108-119. doi: https://doi.org/10.5753/wcge.2020.11262.
- De Figueirêdo, B. C. B., Nakamura, F. G., Felix, G. S., e Nakamura, E. F. Usando análises sociais na identificação de nós relevantes em um cenário multi-redes: operação licitante fantasma, um estudo de caso, in Brazilian Journal of Development, 2020, 6(9). DOI:10.34117/bjdv6n9-280.

Referências Bibliográficas

Agneessens, F., Borgatti, S. P., e Everett, M. G. (2017). Geodesic based centrality: Unifying the local and the global. *Soc Networks*.

Almeida, T., Nakamura, F. G., e Nakamura, E. F. (2017). Uma abordagem baseada em redes complexas para análise de depoimentos legais. Em *XXXVII Congresso da Sociedade Brasileira de Computação*, pgs. 2482–2491.

Arnold, K., Gosling, J., e Holmes, D. (2000). *The Java Programming Language*. Addison Wesley Publishing Company.

Balaniuk, R., Bessiere, P., Mazer, E., e Cobbe, P. (2013). Collusion and corruption risk analysis using naive bayes classifiers. *Tweedale J.W., Jain L.C. (eds) Advanced Techniques for Knowledge Engineering and Innovative Applications. Communications in Computer and Information Science*, 246.

Ball, B. e Newman, M. E. J. (2012). Friendship networks and social status. *CoRR*, abs/1205.6822.

Bansal, R., Gaur, N., e Singh, S. N. (2016). Outlier detection: Applications and techniques in data mining. Em *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, pgs. 373–377.

Battiston, F., Nicosia, V., e Latora, V. (2014). Structural measures for multiplex networks. *Physical Review E*, 89(3).

Bavelas, A. (1950). Communication patterns in task-oriented groups. *J Acoust Soc Am.*

Beveridge, A. e Shan, J. (2016). Network of thrones. *Math Horizons.*

Bhowmik, R. (2008). Data mining techniques in fraud detection. *Journal of Digital Forensics, Security and Law*, 3, Article 3.

Bianconi, G. (2018). *Multilayer Networks: Structure and Function*. Oxford University Press.

Biggs, N., Lloyd, E., e Wilson, R. (1986). *Graph Theory*. Oxford University Press.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., e Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics Reports*, 424(4):175–308.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *J Math Sociol*.

Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Soc Networks*.

Borgatti, S. P. e Everett, M. G. (2007). Models of core/periphery structures. *Soc Networks*.

Brandes, U. e Erlebach, T. (2005). Network analysis - methodological foundations - introduction. *Netw Anal Found*.

Bródka, P., Skibicki, K., Kazienko, P., e Musial, K. (2012). A degree centrality in multi-layered social network. *CoRR*, abs/1210.5184.

Bullmore, E. e Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci*.

Caldarelli, G. (2020). Scale-free networks: Complex webs. *Nature and Technology*.

Carneiro, M. G. (2016). *Redes complexas para classificação de dados via conformidade de padrão, caracterização de importância e otimização estrutural*. PhD thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos.

Carvalho, J. e Ramos, T. (2002). *Logistica na Saúde*. 3rd ed. Lisbon: Sílabo.

Carvalho, M. C. M. (2017). Utilização de redes complexas para análise social de redes de personagens. Master's thesis.

Christakis, N. A. e Fowler, J. H. (2010). Social network sensors for early detection of contagious outbreaks. *PLoS One*.

Clauset, A., Newman, M. E. J., e Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70(6).

Costa, C. e Aparicio, M. (2011). Using data mining to help auditors. Em *17th International Business Information Management Association Conference, IBIMA 2011; Milan; Italy; 14 November 2011 through 15 November 2011; Code 106712*, volume 4, pgs. 1864–1868.

Cunha, R. e Bugarin, M. S. (2014). Lei de benford e auditoria de obras públicas: Uma análise de sobrepreço na reforma do maracanã. *Revista TCU*, pgs. 48–53.

Das, K., Samanta, S., e Pal, M. (2018). Study on centrality measures in social networks: a survey. *Social Network Analysis and Mining*, 8.

De Domenico, M., Solé-Ribalta, A., Cozzo, E., Kivelä, M., Moreno, Y., Porter, M. A., Gómez, S., e Arenas, A. (2013). Mathematical formulation of multilayer networks. *Phys. Rev. X*, 3:041022.

De Domenico, M., Sole-Ribalta, A., Gomez, S., e Arenas, A. (2014). Navigability of interconnected networks under random failures. *Proceedings of the National Academy of Sciences*, 111(23):8351–8356.

De Figueirêdo, B. C. B., Nakamura, F. G., Felix, G. S., e Nakamura, E. F. (2020). Usando análises sociais na identificação de nós relevantes em um cenário multi-redes operação licitante fantasma, um estudo de caso. *Brazilian Journal of Development*, 6(9).

De Figueirêdo, B. C. B., Nakamura, F. G., e Nakamura, E. F. (2021). A group-based centrality for undirected multiplex networks: A case study of the brazilian car wash operation. *IEEE Access*, 9:81946–81956.

Fabbri, R. (2017). Redes complexas para redes sociais: Introdução, aspectos críticos e software. *Impulso*.

Federal, M. P. (2019). A lava jato em números no paraná. in: Brazilian fed. public prosec. Access date: 7 may. 2019.

Figueiredo, B., Nakamura, F., Felix, G., e Nakamura, E. (2020). Usando análises sociais na identificação de nós relevantes em um cenário multi-redes: Operação licitante fantasma, um estudo de caso. Em *Anais do VIII Workshop de Computação Aplicada em Governo Eletrônico*, pgs. 108–119, Porto Alegre, RS, Brasil. SBC.

Fire, M. e Guestrin, C. (2020). The rise and fall of network stars: Analyzing 2.5 million graphs to reveal how high-degree vertices emerge over time. *Inf Process Manag.*

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75–174.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Soc Networks*.

Ghedini Ralha, C. e Sarmento Silva, C. V. (2012). A multi-agent data mining system for cartel detection in brazilian government procurement. *Expert Systems with Applications*, 39(14):11642–11656.

Guerriero, V. (2012). Power law distribution: Method of multi-scale inferential statistics. *Journal of Modern Mathematics Frontier (JMMF)*, 1(1):21–28.

Guimerà, R., Mossa, S., Turtschi, A., e Amaral, L. A. N. (2005). The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proc Natl Acad Sci U S A*.

Hilary (2015). *Centrality measures in multilayer networks*. PhD thesis, University of Oxford.

Hsieh, M. H. e Magee, C. L. (2010). A new method for finding hierarchical subgroups from networks. *Soc Networks*.

Hu, B., Carvalho, N., Laera, L., Lee, V., Matsutsuka, T., Menday, R., e A, N. (2013). Applying semantic technologies to public sector: A case study in fraud detection. *Takeda H., Qu Y., Mizoguchi R., Kitamura Y. (eds) Semantic Technology. JIST 2012. Lecture Notes in Computer Science*, 7774.

Johnson, P. E., S, G., Jamal, K., e Berryman, R. G. (2001). Detecting deception: Adversarial problem solving in a low base-rate world. *Cognitive Science*, 25(3):355–392.

Kolaczek, G. e Juszczyszyn, K. (2019). Complex networks monitoring and security and fraud detection for enterprises. Em *Proceedings - 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises*, pg. 402.

Liao, H., Mariani, M. S., Medo, M., Zhang, Y. C., e Zhou, M. Y. (2017). Ranking in evolving complex networks. *Phys. Rep.*

Machado, S. (2012). *Gestão da Qualidade*. Inhumas/GO: e-Tec Brasil.

Medvedeva, M. A. e Komotskiy, E. I. (2016). About usage of data mining methods for fraud detection in the sphere of communal services. Em *AIP Conference Proceedings*, volume 1738.

Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., e Onnela, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328:876–878.

Nanni, M., Tortosa, L., Vicent, J. F., e Yeghikyan, G. (2020). Ranking places in attributed temporal urban mobility networks. *PloS one*, 15:10.

Newman, M. e Newman, M. E. J. (2010). *Mathematics of Networks*. Oxford Scholarship Online.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582.

Nicosia, V., Bianconi, G., Latora, V., e Barthelemy, M. (2013). Growing multiplex networks. *Physical Review Letters*.

Otte, E. e Rousseau, R. (2002). Social network analysis: A powerful strategy, also for the information sciences. *J Inf Sci*.

Pastor-Satorras, R., Castellano, C., Van Mieghem, P., e Vespignani, A. (2015). Epidemic processes in complex networks. *Rev Mod Phys*.

Pearson, K. (1905). On the general theory of skew correlation and non-linear regression. *Dulau and Co*.

Ribeiro, A. C. P. A. (2016). Deteção de outliers e previsão de vendas numa empresa de distribuição farmacêutica em portugal. *Universidade Portucalense*.

Rinaldo, A., Banavar, J. R., e Maritan, A. (2006). Trees, networks, and hydrology. *Water Resour. Res.*

Rocha, C. A. A. (2002). *Técnicas de Amostragem para Auditorias*. Tribunal de Contas da União.

Schweitzer, F., Fagiolo, G., Sornette, D., Vega-Redondo, F., Vespignani, A., e White, D. R. (2009). Economic networks: The new challenges. *Science*, pg. 80.

Sciarra, C., Chiarotti, G., Laio, F., e Ridolfi, L. (2018). A change of perspective in network centrality. *Scientific Reports*, 8(1):2045–2322.

Silva, C. V. S. e Ralha, C. G. (2010). Utilização de técnicas de mineração de dados como auxílio na detecção de cartéis em licitações. pgs. 1–14. XXX Congresso Da Sociedade Brasileira de Computação.

Silva, L. A. D. (2016). Utilização de técnicas de mineração de dados como auxílio na detecção de cartéis em licitações. *Revista TCU*, pgs. 18–23.

SimilarWeb (2021). Accessed: 2021-04-10.

Skillicorn, D. e Purda, L. (2012). Detecting fraud in financial reports. Em *2012 European Intelligence and Security Informatics Conference*, pgs. 7–13.

Solé-Ribalta, A., De Domenico, M., Gómez, S., e Arenas, A. (2016). Random walk centrality in interconnected multilayer networks. *Physica D: Nonlinear Phenomena*, 323-324:73–79. Nonlinear Dynamics on Interconnected Networks.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Tu, X., Jiang, G., Song, Y., e Zhang, X. (2018). Novel multiplex pagerank in multilayer networks. *IEEE Access*, 6:12530–12538.

Virdhagriswaran, S. e Gordon, D. (2006). Camouflaged fraud detection in domains with complex relationships. pg. 941–947. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.

Wirth, E., Szabó, G., e Czinkóczky, A. (2016). Measure landscape diversity with logical scout agents. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B2:491–495.

Yang, H., Luo, J. D., Fan, Y., e Zhu, L. (2020). Using weighted k-means to identify chinese leading venture capital firms incorporating with centrality measures. *Inf Process Manag.*

Ziberna, A. (2020). k-means-based algorithm for blockmodeling linked networks. *Soc Networks.*