SENAI CIMATEC

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL E TECNOLOGIA INDUSTRIAL

Mestrado em Modelagem Computacional e Tecnologia Industrial

Dissertação de Mestrado

Redes semânticas baseadas em títulos de artigos científicos

Apresentada por: Marcelo do Vale Cunha
Orientador: Hernane Borges de Barros Pereira
Co-orientador: José Garcia Vivas Miranda

Novembro de 2013

Marcelo do Vale Cunha

Redes semânticas baseadas em títulos de artigos científicos

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Modelagem Computacional e Tecnologia Industrial, Curso de Mestrado em Modelagem Computacional e Tecnologia Industrial do SENAI CIMATEC, como requisito parcial para a obtenção do título de **Mestre em Modelagem Computacional e Tecnologia Industrial**.

Área de conhecimento: Interdisciplinar

Orientador: Hernane Borges de Barros Pereira
SENAI CIMATEC

Co-orientador: José Garcia Vivas Miranda
Universidade Federal da Bahia

Salvador
SENAI CIMATEC
2013

Nota sobre o estilo do PPGMCTI

Esta dissertação de mestrado foi elaborada considerando as normas de estilo (i.e. estéticas e estruturais) propostas aprovadas pelo colegiado do Programa de Pós-graduação em Modelagem Computacional e Tecnologia Industrial e estão disponíveis em formato eletrônico (*download* na Página Web http://ead.fieb.org.br/portal_faculdades/dissertacoes-e-teses-mcti.html ou solicitação via e-mail à secretaria do programa) e em formato impresso somente para consulta.

Ressalta-se que o formato proposto considera diversos itens das normas da Associação Brasileira de Normas Técnicas (ABNT), entretanto opta-se, em alguns aspectos, seguir um estilo próprio elaborado e amadurecido pelos professores do programa de pós-graduação supracitado.

SENAI CIMATEC

Programa de Pós-graduação em Modelagem Computacional e Tecnologia Industrial

Mestrado em Modelagem Computacional e Tecnologia Industrial

A Banca Examinadora, constituída pelos professores abaixo listados, leram e recomendam a aprovação [com distinção] da Dissertação de Mestrado, intitulada "Redes semânticas baseadas em títulos de artigos científicos", apresentada no dia (28) de (Novembro) de (2013), como requisito parcial para a obtenção do título de **Mestre em Modelagem Computacional e Tecnologia Industrial**.

Orientador:

Prof. Dr. Hernane Borges de Barros Pereira
SENAI CIMATEC

Coorientador:

Prof. Dr. José Garcia Vivas Miranda
Instituto de Física, UFBA

Membro interno da Banca:

Prof. Dr. Valter de Senna
SENAI CIMATEC

Membro externo da Banca:

Prof. Dr. Thadeu Josino Pereira Penna
Instituto De Ciências Exatas, UFF

Dedico este trabalho aos meus pais, *Silvanisio Teixeira da Cunha* e *Genilda do Vale Cunha* e às minhas irmãs *Danielle do Vale Cunha* e *Sinara do Vale Cunha*. Vocês são minha verdadeira inspiração de vida.

Agradecimentos

"*A mente que se abre a uma nova idéia jamais voltará ao seu tamanho original*"

Albert Einstein

"*Não me sinto obrigado a acreditar que o mesmo Deus que nos dotou de sentidos, razão e intelecto, pretenda que não os utilizemos*"

Galileu Galilei

Sou extremamente grato a Deus, pelo seu amor incondicional. Agradeço também, aos meus amados pais, Silvio e Genilda, por tornarem esta vitória possível, com ensinamentos, motivação e irrestrito apoio emocional. Às minhas duas irmãs Dani e Sinara por me proporcionarem muitos momentos de paz e alegria. À minha noiva, Renata, pelo carinho, conselhos e por me acompanhar em sonhos, sendo certo que, esta pesquisa nos uniu ainda mais.

Agradeço aos meus orientadores, por serem verdadeiros mestres de vida. A Hernane, pela maestria em orientar esta pesquisa, pela sua dedicação, paciência e pelo constante incentivo a uma produção científica de qualidade.

Agradeço também, ao meu coorientador, Garcia, pela motivação acadêmica que recebi, desde a época da iniciação científica na UFBA, ainda quanto estudante de Física, e por sua preocupação e cuidado com minhas limitações.

Estes professores são para mim, referência em pesquisa científica e atividade docente, pois zelam pela qualidade, inovação, efetividade e ética nos seus trabalhos. Sinto-me honrado por ter sido membro de seus grupos de pesquisa, e também por termos vivenciado tantos momentos de descontração e alegria. Assim, sou grato a Hernane e Garcia, e também a Grilo e Inácio, pela coautoria nos trabalhos publicados.

Aos membros da banca, Thadeu Penna e Valter de Sena, aos colegas do grupo de pesquisa *Fuxicos & Boatos* (Ana, Tereza Kelly, Terezinha, Grilo, Hernane, Inácio, Cláudia e Patrícia) e a Silvia Caldeira, pelas ricas discussões sobre o meu trabalho.

Aos meus amigos, pelo constante apoio, em especial Vinícius Nonato, Thiago Rodrigues e Sérgio Floquet. Aos meus colegas do PPG-MCTI e aos meus colegas de trabalho, pelas trocas de saberes. Aos meus amigos do IFBA (Gustavo, Valter, Valdex, Adriana e Climério) por viabilizarem minhas idas a Salvador.

Agradeço também aos funcionários da secretaria do PPG-MCTI pela eficiência e zelo e a todos, que de alguma forma contribuíram para este processo.

Salvador, Brasil — Marcelo do Vale Cunha

04 de Novembro de 2013

Resumo

Nas últimas décadas tem aumentado os estudos sobre o periódico científico, não só por ser um sistema formal de comunicação científica, mas também pelas redes sociais que fomenta (e.g. redes de coautoria, redes de citação). Entretanto, no âmbito das redes sociais e complexas, pouco se tem estudado sobre o vocabulário comum em publicações de um periódico. O propósito desta pesquisa é diferenciar periódicos científicos a partir das redes de palavras baseadas em títulos de artigos científicos. Duas abordagens são exploradas: a primeira, busca nos títulos uma rede ótima, fenômeno ligado a linguagem humana, que tem sido estudado em redes de discursos orais e escritos. Assim, os periódicos são diferenciados a partir de suas redes críticas. A segunda abordagem sugere diferenciar épocas diferentes de um mesmo periódico, através de sua rede de títulos variando no tempo. Os resultados para a primeira abordagem mostram a existência de rede crítica em redes de títulos e, também, propõem formas de diferenciar periódicos a partir de suas configurações críticas. Neste estudo, um dos experimentos realizados envolveu a construção das redes com títulos escolhidos aleatoriamente em um periódico. Resultados desse processo apontam dependência temporal entre as redes de títulos de épocas diferentes. Neste sentido, a segunda abordagem da pesquisa investiga a rede de títulos da Nature ao longo do tempo (1999 a 2008). A abordagem tem sua base teórica em Grafos que variam no Tempo, i.e. Time-Varying Graphs (TVG). O modelo proposto aqui analisa a evolução de índices clássicos de redes, a partir de janelas de observação de tempo fixo que varrem todo o tempo de vida do TVG. Os resultados desta etapa revelam tendências ao longo do tempo para os índices. Além disso, foi verificado correlações nas séries temporais, com o uso do método DFA. Com a verificação do fenômeno da rede crítica em redes de títulos, bem como a existência de correlações persistentes nas séries temporais, novas possibilidades se abrem para o estudo da colaboração científica. Dessa forma, os resultados desta pesquisa podem fomentar estudos mais precisos que discutam as tendências nas séries temporais, bem como os padrões modulares apresentados nas redes críticas, a fim de observar quais temáticas são mais ou menos abordadas em publicações de um dado periódico.

PALAVRAS CHAVES: Redes semânticas; Grafos variáveis no tempo; Redes de cliques; Colaboração científica; Redes sociais e complexas.

Abstract

In the last few decades has increased studies on the scientific journal, not only because it is a formal system of scientific communication, but also by social networks that encourages (eg networks of co-authorship, citation networks). However, in the context of social networks and complex, little has been studied about the common vocabulary in a periodical publications. The purpose of this research is to differentiate scientific journals from the networks of words based on titles of scientific articles. Two approaches are explored: first, the search titles optimal network, a phenomenon linked to human language, which has been studied in networks of oral and written discourse. Thus, journals are differentiated from their critical networks. The second approach suggests differentiate different times in the same journal, through its network of bonds varying in time. The results for the first approach to show the existence of critical network networking securities and also propose ways to differentiate from its periodical critical settings. In this study, one of the experiments involved the construction of networks with randomly chosen titles in a journal. Results of this process show temporal dependence between networks titles difententes times. In this regard, the second approach of the study investigates the network of bonds over time Nature (1999 to 2008). The approach has its theoretical basis in graphs that vary in time, ie Time - Varying Graphs (TVG). The model proposed here analyzes the evolution of indices classical networks, from the observation windows of a fixed time which sweep the entire lifetime of the TVG. The results of this step show trends over time in the contents. Moreover, it was verified correlations in time series, the method using DFA. With the verification of the phenomenon of network critical networks titles as well as the existence of persistent correlations in time series, open up new possibilities for the study of scientific collaboration. Thus, the results of this research can foster more precise studies that discuss the trends in time series, and the patterns presented in modular networks critical in order to see which themes are more or less covered in publications of a given journal.

Sumário

Lista de Tabelas

Lista de Figuras

Lista de Siglas

AFE	*Agricultural and Forest Entomology*
ARJG	*Antipode: A Radical Journal of Geography*
APPL	*Applied Psycholinguistics: Psychological and Linguistic studies Across Languages and Learning*
CB	*Chemistry and Biology*
DFA	*Detrended Fluctuation Analysis*
HR	*Human Relations: Towards the integration of the Social Sciences*
IF	*Incidencia-fidelidade*
NAT	*Nature*
PPGMCTI ..	Programa de Pós-graduação em Modelagem Computacional e Tecnologia Industrial
PRA	*Physical Review A*
PRB	*Physical Review B*
PRC	*Physical Review C*
PRD	*Physical Review D*
PRE	*Physical Review E*
PRL	*Physical Review L*
PEM	*Probabilistic Engineering Mechanics*
SHI	*Sociology of Health and Illness*
TVG	*Grafos variaveis no tempo*
WWW	World Wide Web

Parte I

Introdução

Capítulo Um

Introdução

1.1 Considerações iniciais

O conjunto de trabalhos que compõe uma revista de publicação científica faz parte de um sistema formal de comunicação. Este sistema expressa - em palavras, diagramas, imagens e equações - o conhecimento de atividades de pesquisa, não só para contribuir com o avanço científico da humanidade, mas também para reforçar os laços de comunicação entre cientistas e da ciência com a sociedade em geral.

Garvey (1979) define a comunicação científica como o conjunto de atividades associadas à produção, disseminação e uso da informação. Este processo ocorre desde o momento em que o cientista concebe uma ideia para pesquisar, até o momento em que a informação acerca dos resultados é aceita como constituinte do conhecimento científico. Para Ziman (1979), este sistema formal de comunicação científica está caracterizado de forma tradicional por uma literatura periódica que é editada, fragmentada e derivada, que é construída por etapas de trabalhos anteriores, e se constitui em fundamento para trabalhos que virão a seguir. Esta literatura, hoje conhecida como Periódico Científico, foi criada em 1665 e desde então transformou-se, de um veículo cuja finalidade consistia em publicar notícias científicas, em um veículo de divulgação do conhecimento, que se origina a partir das atividades de pesquisa de cientistas e colaboradores da ciência (MIRANDA; PEREIRA, 1996). Dessa forma, os relacionamentos sociais no meio acadêmico são fortemente influenciáveis por esse sistema literário, que é regido por regras e condutas éticas exigidas pelos seus membros. Esses relacionamentos são também motivados por ideias e questionamentos presentes nas mentes dos indivíduos pertencentes às comunidades científicas, que leem, publicam e citam trabalhos de outros cientistas. Vanz e Stumpf (2010) reforçam que o cientista que intenciona colaborar com este sistema de comunicação precisa entrar em acordo com o parceiro coautor, quanto à visão de ambos sobre uma determinada pesquisa.

Uma maneira simples de modelar este sistema é utilizando a teoria de redes sociais e complexas. Neste contexto, um periódico científico pode servir de palco para vários tipos de relações sociais (e.g. coautoria, citações, vocabulário comum, etc.). Por exemplo, pode ser visto como um conjunto de artigos que representam um conhecimento comum de uma comunidade de cientistas, que publica, lê e cita artigos desta mesma comunidade.

Fadigas et al. (2013) esperam que o uso da teoria de redes sociais contribua de forma relevante para mapear a colaboração entre pesquisadores pertencentes a uma mesma comunidade científica. No contexto desta pesquisa, a rede social formada tem como relaci-

onamento entre seus atores o vocabulário comum entre os autores utilizado para compor os títulos das publicações.

O uso de redes semânticas baseadas em títulos de artigos científicos pode auxiliar no entendimento da integração de coautores de um mesmo periódico. Desta forma, pode-se apontar evidências de como autores decidem os títulos de seus artigos científicos (FADIGAS; PEREIRA, 2013). Os títulos possuem um papel fundamental em um documento científico, pois é a primeira parte a ser lida. Ele é composto por palavras selecionadas pelos autores, na busca de uma representação sintética e fidedigna das ideias que serão apresentadas no corpo do trabalho. Através das palavras contidas nesses títulos, pode-se construir redes de palavras a fim de se perceber a relação de um trabalho com outro, de um campo do saber com outro e de um grupo de cientistas com outro.

A abordagem estática para este tipo de rede foi estudada por Fadigas et al. (2009) e Pereira et al. (2011). Estes trabalhos propuseram regras para tratamento das palavras e formação das redes de títulos. E, para a construção das redes, parte da premissa de que um título deve ser considerado como uma clique[1], onde os vértices são as palavras que o compõe.

A partir desta premissa, qualquer texto escrito pode ser transformado em uma rede de palavras. Caldeira (2005) foi um dos primeiros autores a considerar as palavras de uma sentença de um texto como vértices de uma clique. Sua premissa se baseia na ideia de que a sentença (i.e. a clique) é a menor unidade de significado de um texto. A cada palavra, um diferente significado pode emergir, de acordo com sua vizinhança, ou seja, os vértices da clique a qual faz parte. Conforme este raciocínio, a adição de vértices em redes de textos escritos se dá pela adição de cliques. Fadigas e Pereira (2013) investigaram propriedades e ressignificaram índices para redes exclusivamente formadas por cliques (e.g. redes de títulos, redes de co-autoria, redes de atores de filmes, etc.).

Assim, nesta dissertação, utiliza-se a premissa de que o conjunto de títulos de artigos de uma revista científica (cliques) e suas interações (processos de união das cliques) fazem emergir o que se pode chamar de "discurso da revista". Este conjunto de palavras com suas interações pode ser entendido como um sistema complexo e sua modelagem aqui será feita através do uso de redes complexas.

A complexidade, enquanto novo paradigma epistemológico, não analisa um sistema complexo sob uma perspectiva reducionista, pois leva em conta as relações entre as partes do sistema. Teixeira (2007) argumenta que o fenômeno da linguagem pode ser considerado como um sistema complexo, por se encaixar em algumas de suas propriedades de um

[1] Clique é um conjunto de n vértices mutuamente conectados. O elemento básico de uma rede de cliques não é o vértice e sim a clique (FADIGAS; PEREIRA, 2013).

sistema adaptativo aberto - em que novos elementos vão sendo agregados, fazendo com que o sistema se modifique e se auto-organize. Uma boa forma de visualizar um sistema complexo, seus elementos e as relações entre eles é através das redes complexas. Rede semântica é nome dado a rede de relacionamentos entre palavras ou conceitos e sua análise quantitativa torna-se mais uma contribuição para o estudo da linguagem.

As redes semânticas do trabalho de Teixeira (2007) representam discursos orais de indivíduos, a partir de um "prime" - que é um tema central sobre o qual discursaram. A autora pressupõe, baseada em Sternberg (2011), que uma rede semântica pode ser entendida como uma representação do conhecimento e portanto uma ferramenta que possibilita o acesso à memória do indivíduo. No entanto, as redes de um simples discurso podem conter milhares de vértices. A partir desta limitação, Teixeira et al. (2010) verificaram a existência de um valor crítico para um indicador denominado *incidência-fidelidade*, que é responsável por uma filtragem na rede. Nesta condição, a rede apresenta um comportamento típico de mudança de fase, juntamente com os valores de seus índices, com um ponto crítico bem definido. A rede gerada para esse nível de filtragem foi denominada de *rede crítica* e foi percebida como a rede característica do discurso de um indivíduo. Ela é uma representação razoável para o acesso à memória declarativa deste indivíduo, por possuir uma configuração com o máximo de informação e o mínimo de ruído (TEIXEIRA, 2007).

O mesmo pressuposto e processo de filtragem, com a *incidência-fidelidade*, foi usado por Aguiar (2009) em discursos escritos, quando verificou a existência do fenômeno crítico em redes de palavras de romances de clássicos da literatura - autores diferentes e idiomas diferentes. Além disso, a proximidade entre os discursos dentro de suas categorias: mesmo autor, mesmo idioma, mesmo tema. Ela percebeu que textos de mesma categoria são próximos, de acordo com sua distância euclidiana no espaço de índices de redes complexas.

Isto possibilita novos experimentos no campo da linguagem e no estudo da comunicação científica. Por exemplo, como foi feito aqui, é interessante verificar se esta filtragem em redes de títulos também gera fenômenos críticos. Se assim for, o emaranhado de palavras filtradas em seu estado crítico possuirá o máximo de informação, com o mínimo de resíduos. Outra abordagem interessante é avaliar como as conexões entre as palavras dos títulos variam para épocas diferentes (CUNHA et al., 2013).

A primeira etapa deste trabalho discute o fenômeno da rede crítica para rede de títulos a partir de duas abordagens. Estas abordagens admitem a rede crítica como uma ótima representação de acesso à memoria declarativa para um dado periódico (i.e. uma visão generalizada de seu conjunto de publicações). Uma abordagem desta etapa enfatiza a verificação do fenômeno da *rede crítica* em um determinado conjunto de títulos, escolhidos aleatoriamente, para cada periódico. Outra abordagem evidencia a possibilidade de

diferenciar um periódico de outro a partir de suas redes críticas.

A segunda etapa deste trabalho analisa uma rede semântica de títulos através de uma modelagem que permite visualizar mudanças nas redes ao longo do tempo. Esta forma de modelar um periódico científico tem suas bases teóricas em grafos que variam no tempo (TVG[2]) (CASTEIGTS et al., 2011). Com ela é possível verificar a evolução do comportamento de uma rede ao longo de diferentes épocas. Os resultados envolvem análises estatísticas nas séries temporais dos índices de redes. São realizados testes de Normalidade(SHAPIRO; WILK, 1965), bem como, é usado o *expoente de Hurst* para análise de correlação das séries(PENG et al., 1994).

Dessa forma, esta dissertação abre lacunas para o estudo da comunicação científica a partir das palavras do vocabulário comum de cientistas que publicam em um determinado periódico e visa contribuir não só para quem estuda sobre os vetores da difusão do conhecimento humano, como também para pesquisadores que estudam linguagem - mais especificamente redes semânticas - e Redes Complexas em geral, que podem variar no tempo.

1.2 Motivação da Pesquisa

Redes de títulos de artigos científicos já foram analisadas sob algumas diferentes perspectivas. Grupos de periódicos em diferentes idiomas foram caracterizados de acordo com similaridades em seus campos de estudos, jargões técnicos e perfis de pesquisadores que publicam, independente do idioma científico selecionado (FADIGAS et al., 2009; FADIGAS, 2011; PEREIRA et al., 2011; FADIGAS; PEREIRA, 2013).

Pouco antes destes trabalhos, Caldeira (2005), Teixeira (2007) e Aguiar (2009) assumiram uma premissa que diz que as palavras que ocorrem juntas em uma mesma sentença teriam sido evocadas de forma associativa na construção de uma ideia a ser apresentada. Com isso, pôde-se construir uma rede onde as palavras são representadas como os vértices e as arestas são criadas entre pares de palavras que ocorrem em uma mesma sentença em um discurso escrito. A sentença é vista como uma clique e a menor unidade de significado de um texto. Assim, cada palavra pode ter um significado diferente a depender das palavras que estejam ao seu redor.

Este pressuposto também vem sendo empregado em redes de títulos, inclusive no presente trabalho, que foi motivado pela possibilidade de comparação com estes trabalhos supracitados. Outros trabalhos que utilizam a mesma base de dados ou mesmo método de construção de Redes Semânticas com outras bases, estão sendo desenvolvidos em paralelo

[2] *Time-Varying Graphs*

por outros pesquisadores e colaboradores desta pesquisa. Desta forma, este trabalho é uma pequena parte de um estudo maior sobre redes semânticas. Sua motivação também se deve por ser engrenagem fundamental no estudo da difusão da informação em periódicos científicos a partir da linguagem humana.

1.3 Importância da pesquisa

Os autores citados na Seção anterior (1.2) deixaram lacunas para pesquisas futuras no mesmo campo de pesquisa que esta dissertação está inserida:

- As propriedades das redes de títulos estudadas indicam estar relacionadas com a diversidade do vocabulário das revistas. Entretanto, faltam análises que incluam métodos capazes de diferenciar cada periódico quanto à sua (multi)disciplinaridade, a partir de uma rede de palavras com menos vértices, porém significativa (PEREIRA et al., 2011);
- Comparação de índices e topologias de redes oriundas de campos de conhecimento distintos, porém com mesmo número de títulos (FADIGAS, 2011);
- Investigação da dinâmica de coesão[3] das cliques na formação de redes de títulos (FADIGAS, 2011; FADIGAS; PEREIRA, 2013);
- Análise de um texto com janelas de co-ocorrência de palavras (CALDEIRA, 2005);
- Busca de um comportamento universal que possa definir um núcleo topológico para a associação semântica e para o complexo mecanismo da linguagem (TEIXEIRA, 2007);
- Verificar a similaridade entre textos previamente agrupados em suas configurações iniciais e nas configurações de rede crítica (AGUIAR, 2009).

1.4 Definição do problema

Apresenta-se então, enquanto problema norteador deste estudo a seguinte questão:

Os periódicos científicos permitem uma classificação a partir dos títulos de seus artigos?

[3] Esta perspectiva mostra o quão um índice de rede clássico varia em relação a seu respectivo na configuração inicial (i.e. cliques isoladas).

1.5 Hipóteses

Um título busca sintetizar as principais ideias de um trabalho em uma única frase. De acordo com a modelagem proposta aqui, um determinado conjunto de títulos de artigos de um dado periódico pode ser visto como o "discurso da revista". A partir deste discurso escrito, pode-se modelar uma rede semântica, onde as palavras de cada título representam os vértices da rede e as arestas representam a ocorrencia de um par de palavras em um mesmo título.

Dessa forma, espera-se que a caracterização destas redes quanto à topologia, existência do fenômeno *rede crítica* e quanto as palavras mais frequentes de um dado conjunto de títulos possa apontar indícios de como diferenciar um periódico de outro e até determinar inclinações sobre sua (multi)disciplinaridade em relação a outros periódicos, como também à ele mesmo em diferentes épocas.

Abaixo segue algumas hipóteses que são testadas nestaa pesquisa:

1. A rede crítica é um mecanismo intrínseco da linguagem, especialmente em rede de títulos;
2. O surgimento de um conjunto de títulos em uma revista depende de títulos anteriores.

1.6 Objetivos

De maneira geral este estudo visa caracterizar as redes de títulos dos periódicos que compõe a base de dados (ver Seção 5.1) quanto à topologia, tanto com o fenômeno da *rede crítica*, quanto com a variação das redes no tempo.

Os objetivos específicos, por sua vez, são:

- Encontrar a rede crítica (se existir) para rede semântica de títulos de artigos científicos;
- Verificar se diferentes periódicos exibem propriedades topológicas em suas redes críticas que permitam diferenciá-los;
- Verificar os itens anteriores para uma quantidade fixa de títulos de cada periódico, obtidos de maneira aleatória;

- Verificar se as redes de títulos de diferentes épocas exibem propriedades topológicas que permitam diferenciá-las;
- Descrever como o padrão de conexão das palavras mudam a cada número de um periódico divulgado;
- Investigar a influência do vocabulário da revista nas tendências observadas para os valores dos índices de redes ao longo do tempo.

1.7 Limites da Pesquisa

Este trabalho possui um caráter interdisciplinar, pois admite investigações em diversas áreas do conhecimento (e.g Matemática, Estatística, Física, Computação, Educação e Difusão do Conhecimento). Os limites para a coleta, metodologia e análise pertencem a um horizonte muito distante, haja vista a pluralidade dos fenômenos envolvidos. Os limites podem ser classificados em: (1) Na escolha dos periódicos; (2) a escolha da Modelagem utilizada e (3) na escolha dos aspectos metodológicos empregados.

Os principais limites impostos podem ser melhor entendidos de acordo com os ítens subsequentes:

1. A escolha dos 15 periódicos da base deveu-se a possibilidade de comparação com trabalhos anteriores: Fadigas et al. (2009), Fadigas (2011), Pereira et al. (2011) e Fadigas e Pereira (2013). A escolha da Nature para a abordagem de *redes variáveis no tempo* se deu pela sua alta frequência de publicação (semana a semana).
2. A modelagem dos títulos dos artigos como redes semânticas de Cliques foi também motivada pelos trabalhos supracitados e por considerar a premissa proposta por Caldeira (2005), segundo a qual um texto pode ser modelado por uma rede de conexão de palavras, dispostas em forma de cliques.

Não obstante as condições de contorno estarem bem definidas, durante o desenvolvimento da pesquisa surgiram algumas limitações. Estas por sua vez, motivaram a inclusão de outros métodos de análise, não previstos inicialmente, na tentativa de responder as questões principais. As limitações da pesquisa estão no Apêndice A, por conter termos que só serão definidos nas próximas seções.

1.8 Organização da Dissertação de Mestrado

Esta dissertação está estruturada em 6 partes - 8 capítulos e 1 apêndice. A primeira parte contém o 1º capítulo (Introdução). Nele é apresentado as considerações iniciais do estudo, o contexto do campo de pesquisa inserido e as lacunas presentes no referido campo. Diante disso, o capítulo expõe a pergunta que norteia este trabalho, descreve os objetivos gerais e específicos, a motivação, as hipóteses e relata os limites encontradas ao longo do tempo em que esta pesquisa foi realizada.

A Parte 2 contém a fundamentação teórica do trabalho, associado ao campo de pesquisa em que o problema está inserido. São apresentados as redes complexas (Capítulo 2); alguns conceitos e definições da *teoria dos grafos* (Capítulo 3); e *modelos de redes* (Capítulo 4) que apresenta modelos já consagrados da literatura, como *redes aleatórias*, *redes de mundo pequeno*, *redes livres de escala*, *redes de cliques* e *redes semânticas*, com alguns índices de redes relevantes para este estudo, e.g. índices clássicos de redes, *incidência-fidelidade* e os índices para redes de cliques.

Na Parte 3, Capítulo 5, tem-se o método proposto para gerar as redes semânticas de títulos, bem como os modelos de análise empregados. O capítulo está orientado com base em uma figura que sintetiza a metodologia utilizada em todo o trabalho. A primeira seção descreve a coleta de dados; a segunda apresenta a modelagem do problema, com subseções que indicam os tratamentos manual e computacional adotados, bem como o processo de construção das redes; a terceira seção explica como é feita a mineração dos dados. Esta mineração foi dividida em duas etapas: (a) a quarta seção, que trata sobre o uso da *Incidência Fidelidade* na construção de redes críticas; e (b) a quinta seção, referente ao método *Time-Varying Graphs* (*TVG*) aplicado em redes de títulos.

A Parte 4 apresenta os resultados e as discussões a partir da utilização do índice *incidência-fidelidade* em redes de títulos (Capítulo 6), bem como os resultados frutos da aplicação do método *TVG* (Capítulo 7).

A Parte 5 contém as considerações finais do trabalho, as conclusões e perspectivas de futuros trabalhos que surgiram a partir deste estudo (Capítulo 8).

O Apêndice A (Parte 6) relata as principais limitações encontradas durante esta pesquisa.

Parte II

Fundamentação Teórica

Capítulo Dois

Redes Complexas

Cientistas de diversas áreas do conhecimento nas últimas décadas aumentaram o interesse pelo estudo de sistemas naturais e sociais que contém elementos que se relacionam entre si. De acordo com esta perspectiva, uma rede seria a abstração destes elementos (chamados de vértices) e suas relações entre si (arestas). Como motivação, este capítulo mostra alguns exemplos de sistemas naturais que podem ser modelados na forma de uma rede complexa

Existem inúmeros exemplos de sistemas naturais que podem ser modelados na forma de uma rede, que já foram e continuam sendo amplamente estudados em diversas áreas do conhecimento humano:

- Redes biológicas;
- Redes tecnológicas;
- Redes sociais;
- Redes de informação.

No campo da Biologia, existem vários exemplos de sistemas que podem ser modelados a partir do uso de redes:

- O cérebro humano pode ser visto de maneira simplificada como um conjunto de células nervosas conectadas por axônios. Uma possível rede seria: neurônios estimulados como nós e as arestas os possíveis caminhos associativos (OLIVEIRA-FILHO, 2012);
- As células em geral, podem ser vistas como uma rede de moléculas conectadas por reações bioquímicas (BARABáSI A. L., 2003). A Figura 2.1 mostra a rede de conexões entre proteínas. A Figura 2.2, por sua vez, representa o exemplo de uma rede de diferenciação celular (GALVãO et al., 2010);
- As relações predatórias entre os animais, conhecida como cadeia ou teia alimentar, podem ser visualizadas a partir de uma rede complexa, em que os vértices representam espécies em um ecossistema e uma aresta direcionada indica que uma espécie (nó de origem) ataca outra (nó destino) (NEWMAN, 2003).

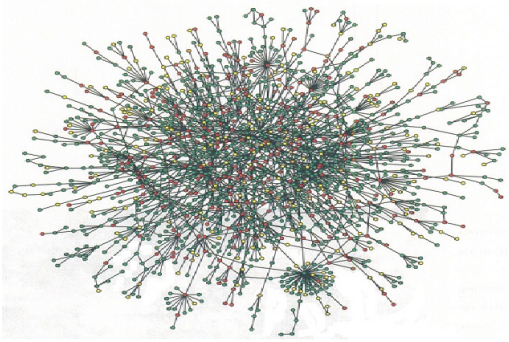


Figura 2.1: Rede de interação entre proteínas. As arestas evidenciam as reações químicas entre pares de proteínas (nós da rede) FONTE: (BARABáSI A. L., 2003, p.59).

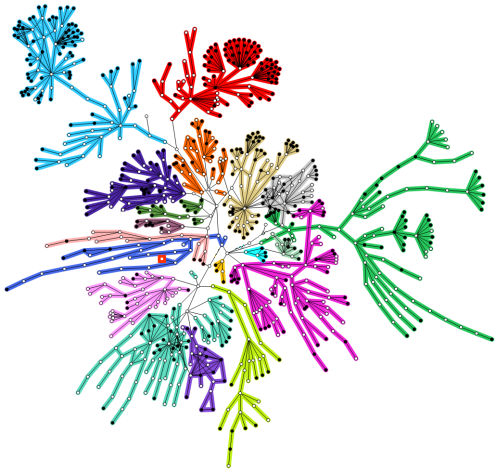


Figura 2.2: Representação da rede complexa do processo de diferenciação celular. FONTE: (GALVãO et al., 2010).

No campo da tecnologia, as redes tecnológicas são estruturas artificiais projetadas tipicamente para distribuição de alguma mercadoria ou recurso, como a electricidade ou informação (NEWMAN, 2003). A rede de energia elétrica é um bom exemplo, Figura 2.3 (NASCIMENTO, 2012).

As redes sociais, por sua vez, são redes formadas por pessoas ou organizações (denominados atores) que podem estar ligadas à diversos tipos de relações, eg. parentesco, laços profissionais, localização geográfica, etc. A Figura 2.4 exibe a rede de colaboração científica entre pesquisadores de um mesmo programa de pós graduação.

No caso de uma rede de informação ou de conhecimento, é preciso ter como base algum conhecimento formal, e.g. as citações de artigos científicos, a WWW (World Wide Web), os registros de patentes, a estrutura das linguagens, etc. (ANGELIS, 2005). Nelas, os vértices

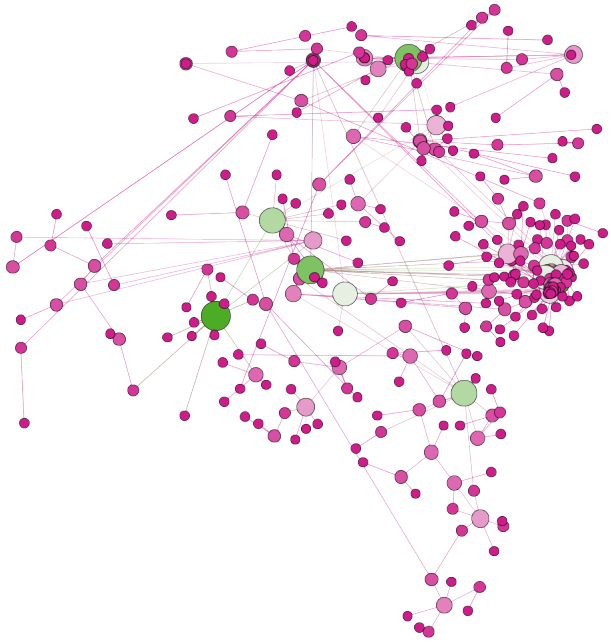


Figura 2.3: Rede de subestações elétricas do estado da Bahia. Os nós são as subestações e as arestas representam os cabos de distribuição de energia elétrica que conectam os postes. FONTE: (NASCIMENTO, 2012, p.42).

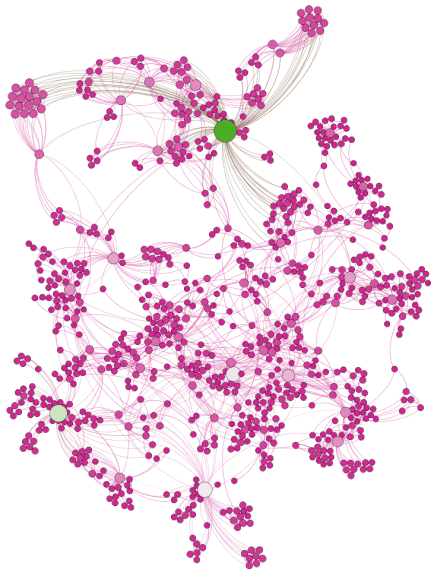


Figura 2.4: Rede de colaboração científica entre pesquisadores de um programa de Pós-Graduação. FONTE: (ANDRADE, 2013, p.79).

podem representar conceitos ou informações e as arestas a difusão de uma informação ou interação entre dois conceitos. Como exemplo, a Figura 2.5 contém uma rede a partir de palavras evocadas por um indivíduo em um discurso oral (TEIXEIRA et al., 2010).

Boa parte de sistemas que podem ser representados através de redes complexas fazem

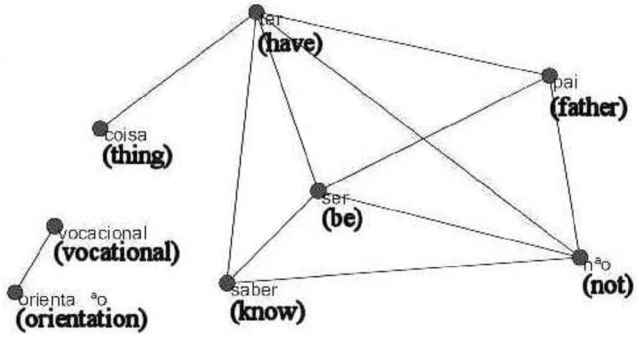


Figura 2.5: Subrede do discurso oral de um indivíduo. Os vértices são palavras presentes no discurso e as arestas representam a ocorrência dos pares na mesma sentença. Este grafo representa os pares de palavras de maiores do discurso deste indivíduo FONTE: Teixeira et al. (2010).

parte de uma classe de sistemas muito especial, os sistemas complexos. Nestes sistemas, elementos ou partes do sistema interagem entre si no tempo e no espaço, em geral com respostas não-lineares, fazendo emergir padrões e propriedades que vão além da soma das partes que constituem o sistema, já que o deixa (NUSSENSVEIG, 2008).

A transição de fase é um outro fenômeno recorrente em Sistemas Complexos e foi enfaticamente estudada por Bak, Tang e Wiesenfeld (1988). Eles formularam uma teoria geral para o ruído $1/f$ e acabaram explicando a razão física para o surgimento dos fractais de Mandelbrot (1983) e, principalmente, para o aparecimento da *criticalidade auto-organizada*. Pela sua teoria, tanto o ruído $1/f$ como os fractais podem surgir na natureza sem a necessidade de intervenção externa (BAK; TANG; WIESENFELD, 1988).

Bak, Tang e Wiesenfeld (1988) desenvolveram o conceito de *criticalidade auto-organizada*. O termo *auto organizada* se refere ao fato de que nada externo opera o sistema, ele por si só sustenta seu padrão. Nesse contexto, o termo *criticalidade* se deve ao fato que o sistema é equilibrado em um ou mais pontos críticos entre ordem e desordem.

Estudos recentes sobre linguagem revelam fenômenos críticos também para sistemas compostos por palavras. Em discursos escritos de romances literários, Aguiar et al. (2007) observaram o fenômeno crítico para a rede de conexão entre as palavras e constataram que esta rede exibe um padrão característico deste tipo de discurso.

Teixeira et al. (2010) verificaram o mesmo fenômeno em redes oriundas de discursos orais. As duas abordagens verificam estruturas de palavras categorizadas nas redes, chamadas de módulos. Entretanto, se todas as palavras de um destes discursos forem embaralhadas, quebrando as estruturas sintáticas que existem entre elas, então as redes não reproduzem o mesmo comportamento modular e o fenômeno crítico desaparece.

A matemática deu início ao estudo das redes a partir do conceito de *grafo* (Capítulo 3), proposto por Euler em 1736.

Capítulo Três

Teoria dos grafos

3.1 Motivação Histórica

Em 1736 o Matemático Suiço Leonhard Euler publicou a solução para um problema clássico da época das pontes de Königsberg, hoje Caliningrado - Russia. O problema consiste em encontrar um caminho que percorresse todos bairros e a ilha de Kneiphof (A, B, C e D - Figura 3.1 à esquerda) da cidade prussiana, passando por todas as pontes, sendo que uma única vez em cada ponte. Euler modelou a cidade de Königsberg através de um esquema a qual chamou de grafo (Figura 3.1 à direita), onde os bairros seriam os vértices e as pontes seriam as arestas (ver seção 3.2). Esta representação simplificada fez Euler concluir que era impossível tal percurso.

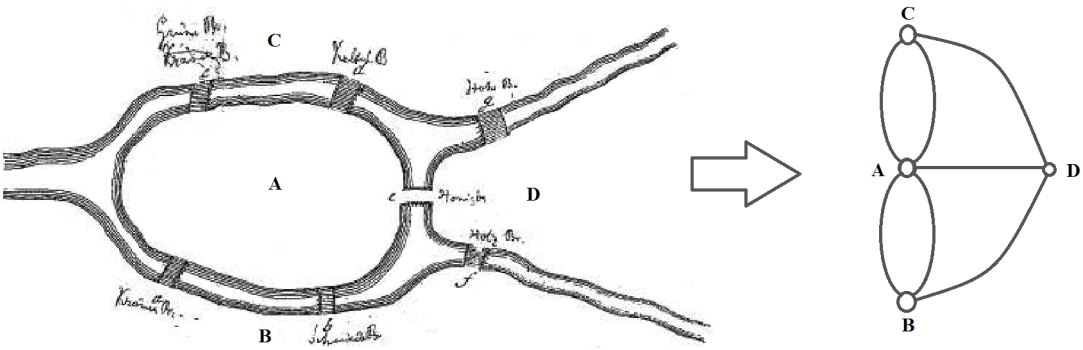


Figura 3.1: à esquerda: Esquema da cidade de Königsberg, seus bairros e pontes interligando-os; à direita: Representação esquemática da cidade como um grafo, proposta por Euler FONTE: (EULER, 1736).

O teorema de Euler dizia que se todos os bairros possuírem um número par de pontes, ou seja, se todos os nós possuirem um número par de arestas, é condição suficiente para que o passeio completo pela cidade possa ser realizado passando somente uma vez por cada ponte, retornando ao ponto de partida (EULER, 1736). Segundo Newman (1953), este caminho foi chamado de *caminho euleriano*. Como o grafo de Königsberg tinha quatro nós com número ímpar de pontes, não era possível encontrar este caminho. De acordo com Euler (1736), este caminho só existe se não houver no grafo nós com número ímpar de arestas ou se houver exatamente 2 nós com esta condição.

A existência de um número par de caminhos se dá por que é preciso um caminho para "entrar" e outro para "sair". Os dois pontos com caminhos ímpares referem-se ao início e ao fim do percurso, pois estes não necessitam de um caminho para entrar e outro para sair. Quando não há nós com número ímpar de caminhos, deve-se iniciar e terminar o

trajeto no mesmo ponto, podendo esse ser qualquer ponto do grafo. Isso não é possível quando temos dois pontos com números ímpares de caminhos, sendo obrigatoriamente um o início e outro o fim.

O aspecto mais importante da prova de Euler em sua época é que a existência do caminho independe de qualquer esforço para encontrá-lo. Trata-se, mais exatamente, de uma propriedade do grafo. Após 1875, uma nova ponte foi criada, entre B e C, o que incrementou a quantidade de links desses dois nós para quatro. Assim, apenas os nós A e D tiveram número ímpar de arestas e o *caminho euleriano* tornou-se possível nesta nova configuração.

Existem situações semelhantes que poderíamos visualizar o *caminho euleriano*, como o desafio enfrentado pelos carteiros de percorrer ruas para fazer as entregas e voltar ao seu posto de trabalho, evitando passar por uma mesma rua. Na coleta de lixo também pode-se pensar semelhante. O caminhão tem que sair do depósito e percorrer o seu trajeto com o mínimo de repetição de ruas. Apesar da possibilidade de uma modelagem utilizando grafos de Euller, estas situações encaixam-se melhor em uma classe específica de problemas do tipo "problema do caxeiro viajante".

Dessa forma, os grafos (ou redes, como serão chamados ao longo desta dissertação) servem como modelagem para inúmeros problemas práticos. Atualmente, são largamente utilizados para representar sistemas naturais e verificar propriedades individuais e coletivas que emergem das interações entre os elementos do sistema. Para melhor entendimento, as próximas seções são dedicadas à algumas métricas envolvidas no estudo do grafo como um todo, seus elementos e as relações entre eles.

3.2 Conceitos e Definições

Feofiloff, Kohayakawa e Wakabayashi (2007) em seu trabalho, sintetizaram conceitos importantes para uma introdução ao estudo de teoria dos grafos. Abaixo segue alguns, importantes para esta pesquisa.

Seja $V^{(2)}$ o conjunto de todos os pares não ordenados de elementos de um certo conjunto V. Sua cardinalidade corresponde a $\binom{n}{2} = \dfrac{n(n-1)}{2}$ em que n é a cardinalidade de V. Assim, os elementos de $V^{(2)}$ possuem cardinalidade $|V^{(2)}| = 2$ e terão a forma (u, w), sendo u e w dois elementos distintos de V.

Um grafo $G(V, \mathcal{E})$ é um par ordenado de conjuntos disjuntos e não vazios, em que V é um conjunto arbitrário de pontos e representa objetos reais. $\mathcal{E}$ é subconjunto de $V^{(2)}$ e

representa as ligações (relações) entre os elementos de V, chamadas arestas. O conjunto $V(G)$ =$\{v_1,...,v_n\}$ é chamado conjunto de vértices e o conjunto $\mathcal{E} = \{(e_1,..., e_m\}$, sendo e_k = $\{(v_i, v_j)\}$ (com $i \neq j$[1] e $i, j = 1, ..., n$) o conjunto das arestas. A Figura 3.2 é exemplo de uma representação gráfica de um grafo, evidenciando seus nós e arestas.

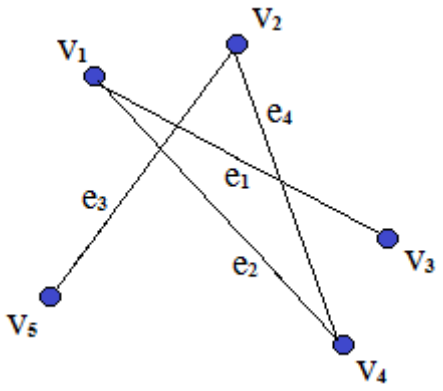


Figura 3.2: Exemplo de grafo G_α com 5 vértices e 4 arestas.

Na Figura 3.2, o conjunto $V(G_\alpha) = \{v_1, v_2, v_3, v_4, v_5\}$ representa os vértices da rede e o conjunto $\mathcal{E}(G_\alpha)$ =$\{(v_1, v_3), (v_1, v_4), (v_2,v_5), (v_2,v_4)\}$ representa as relações entre os vértices desta rede, ou seja suas arestas. A cardinalidade do conjunto V ($|V|$= 5) representa a ordem do grafo (chamaremos de $n(G)$ ou simplesmente n) e a cardinalidade do conjunto $\mathcal{E}$ ($|\mathcal{E}| = 4$) representa o tamanho do grafo (que chamaremos de $m(G)$ ou m). Considere a aresta $e_3 = \{(v_2, v_5)\}$. Diz-se que os vértices v_2 e v_5 são vizinhos ou adjacentes, por serem as pontas da aresta e_3.

3.3 Subgrafos

Um subgrafo $S(V_S, \mathcal{E}_S)$ é um grafo que pode ser obtido retirando arestas ou retirando vértices de um grafo $G(V, \mathcal{E})$. Dessa forma $V_S \subseteq V$ e $\mathcal{E}_S \subseteq \mathcal{E}$.

Uma vez definido o conceito e as condições de existência de uma *grafo* e sabendo que o conceito de *redes* é apropriado a este modelo, de agora em diante será usado o termo *redes* para tratar de redes ou grafos, visto que neste trabalho eles representam a mesma coisa.

[1]Existem situações em que uma aresta tem um único vértice como origem e alvo dela. Este tipo de conexão é chamado de laço e não será considerado nesta pesquisa, como também não será admitido aqui a existencia arestas "paralelas", ou seja, arestas diferentes com o mesmo par de pontas.

3.4 Índices de Redes Complexas

Para entender melhor o comportamento de um grafo ou rede complexa, tanto do ponto de vista quantitativo quanto do ponto de vista qualitativo, utilizamos índices matemáticos da *teoria dos grafos*. Existem dezenas de índices e para cada estudo utiliza-se um conjunto deles para assim caracterizar a rede estudada, entender o fluxo de informação e a dinâmica dos vértices e das relações entre eles.

Estes índices, evidenciam a capacidade da rede, como sistema complexo, de ultrapassar em muito a mera soma de seus elementos, não obstante os índices serem calculados utilizando-se apenas números de arestas e de vértices.

No presente trabalho, os índices mais relevantes são: *grau*, *grau médio*, *caminho mínimo médio*, *densidade*, *diâmetro* e *coeficiente de aglomeração médio*. Utilizaremos também o índice *incidência-fidelidade*, criado para estudar alguns tipos de redes semânticas, que será detalhado na Seção 4.5.2.

Para grafos variáveis no tempo existem um cojunto de indicadores úteis para analisar sua evolução estrutural. Este conjunto é divido em *índices atemporais* e *índices temporais* (e.g.*jornada*, *distância temporal*, *excentricidade.*) (AMBLARD et al., 2011). Consta aqui apenas a abordagem de índices atemporais. Esta abordagem está dividida em *índices clássicos* e *índices de redes de cliques*.

3.4.1 Grau e grau médio

O grau k_i de um vértice i nos informa a quantidade de arestas m_i que incidem nele, ou seja a quantidade de relações que ele faz. O grau médio $< k >$ da rede representa a média desses valores. Na Figura 3.2 o grau do vértice v_1 é $k_1 = 2$, o grau do vértice v_2 é $k_2 = 1$ e o grau médio da rede é $< k >= 1,6$. Note que o mesmo valor poderia ser obtido com a a equação:

$$< k >= \frac{2m}{n} \tag{3.1}$$

já que o dobro do número de arestas é igual a soma de todos os graus médios, ou seja:

$$2m = \sum_{i=1\epsilon V}^{n} k_i \tag{3.2}$$

A Equação 3.2 traduz o primeiro teorema da *teoria dos grafos* proposta por Euler (NEWMAN, 1953).

3.4.2 *Distribuição de Graus*

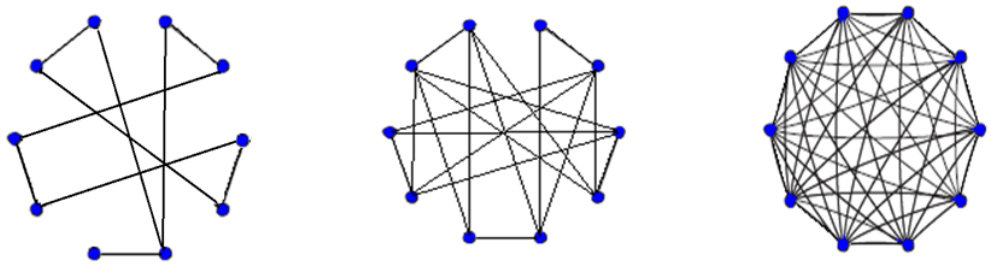
A distribuição das frequências dos graus existentes em uma determinada rede pode muito informar sobre a dinâmica de conexão de seus vértices. Os casos mais considerados na literatura são os de Redes Livres de Escala (*scalle-free*); Redes de mundo pequeno (*small-world*) e redes aleatórias.

Para o caso das *redes aleatórias* (ERDOS; RENYI, 1960), onde os vértices se conectam aleatoriamente, com uma dada probabilidade fixa, a distribuição de graus é Normal. Ou seja, a rede possui uma média bem representativa para o conjunto dos graus.

Já as redes complexas do tipo *livre de escala* ($P(k) \propto k^{\gamma}$), a distribuição de graus possui forma de uma lei de potência. O coeficiente γ é suficiente para caracterizar a topologia da rede em Livre de Escala (BARABASI; ALBERT; JEONG, 1999). As redes do tipo Mundo Pequeno podem apresentar tanto uma distribuição em lei de potência quanto uma distribuição normal (WATTS; STROGATZ, 1998). O Capítulo 4 detalha mais estes modelos de redes.

3.4.3 *Densidade*

A densidade de uma rede não dirigida é uma medida do "poder de relacionamento" dos vértices que compõe a rede. Ou seja, quanto mais arestas a rede tiver mais densa ela será. A figura 3.3 a seguir mostra uma rede de 10 vértices em três estágios de densidades diferentes.



(a) Rede com densidade baixa: 10 vértices e 10 arestas.

(b) Rede com 10 vértices e 19 arestas

(c) Rede com densidade máxima: 10 vértices e 45 arestas.

Figura 3.3: Exemplos de redes em ordem crescente de densidade.

Na Figura 3.3(a) ve-se que os vértices se relacionam muito pouco ($\Delta = 0,22$) , na Figura 3.3(b) um relacionamento intermediário ($\Delta = 0,42$)e na Figura 3.3(c) todos os vértices se relacionam entre si ($\Delta = 1,00$), formando o que chamamos de clique, ou seja quando

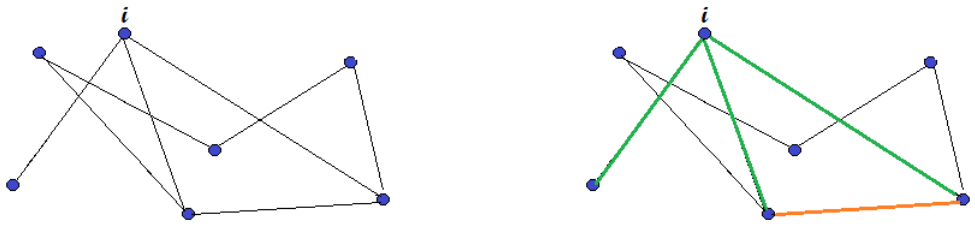
todos os vértices da rede se relaciona com todos os outros. A expressão que determina a densidade de uma rede não dirigida é dada pela Equação 3.3:

$$\triangle = \frac{m}{\frac{n(n-1)}{2}} \tag{3.3}$$

Na Equação 3.3 m é o número de arestas do grafo e n o número de vértices da rede e $\frac{n(n-1)}{2} = \binom{n}{2}$ representa o número de arestas possíveis da rede. Dessa forma, o seu cálculo trata-se da razão entre o número de arestas existentes e o número de arestas possíveis da rede. Seus valores variam desde 0, não existem arestas no grafo, até 1, quando todas as arestas possíveis estão presentes.

3.4.4 Coeficiente de aglomeração médio

O coeficiente de de clusterização ou aglomeração, C_i, de um vértice i é uma medida da densidade da rede formada pelos seus vizinhos, ou seja pelos vértices que se ligam ao vértice i, ver Equação 3.4.



(a) vértice i de um grafo qualquer. (b) vizinhos de i e ligações ente eles.

Figura 3.4: Aglomeração do vértice i e aglomeração média da rede.

$$C_i = \frac{m_i}{\frac{k_i(k_i-1)}{2}} \tag{3.4}$$

Na equação 3.4 m_i significa o número de arestas existentes no subgrafo formado pelos vizinhos de i e k_i é o grau do vértice i e nesta expressão está representando o número de vizinhos de i. Percebe-se que quanto maior for a quantidade de ligações entre os vizinhos de i, maior será a sua aglomeração. Caso um vértice esteja ligado a somente 1 vértice, sua aglomeração é 0.

Se o coeficiente de aglomeração for 0 significa que os nós adjacentes a i não estão conec-

tados entre si. mas se for 1, significa que todos os vértices adjacentes a i estão conectados entre si.

Como se pode ver na Figura 3.4(b) o vértice i possui três vizinhos, indicado pelas arestas de cor verde, e seus vizinhos possuem apenas 1 ligação entre eles, então a aglomeração deste vértice é calculada como sendo $C_i = \frac{1}{3} = 0,33$.

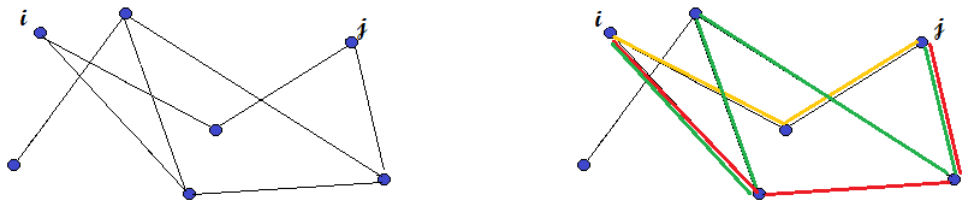
O Coeficiente de Aglomeração médio (Equação 3.5) de uma rede, $\langle C \rangle$, é a média das aglomerações de seus vértices, ou seja:

$$\langle C \rangle = \frac{1}{n} \sum_{i}^{n} C_i \tag{3.5}$$

Para o grafo da figura 3.4(a) a média das clusterizações é $\langle C \rangle = 0,14$.

3.4.5 *Caminho mínimo médio*

Sejam dois vértices v_i e v_j de uma rede. Existe um caminho entre eles se e somente se existir uma (ou mais de uma) sequência de vértices $\{v_i, v_i + 1, v_i + 2, ..., v_j\}$, em que v_k está conectado com $v_k + 1$, sendo $i \leq k \leq j$. O caminho entres estes dois nós da rede é uma sequência de arestas que ligam a sequência de vértices, sem repetição de vértice, do extremo v_i ao extremo v_j. O caminho mínimo entre dois vértices é o menor caminho que ligam estes vértices. Para quantificar este índice, adimite-se que cada aresta ao longo do caminho que ligam dois vértices tem comprimento igual a 1. Como exemplo, considere a Figura 3.5.



(a) vértices i e j de um grafo qualquer. (b) Caminhos entre os vértices i e j.

Figura 3.5: Caminho mínimo médio entre i e j.

A Figura 3.5(a) representa uma rede com 7 elementos, dentre eles os vértices i e j. Existem três caminhos para se ir do vértice i para o vértice j, como mostrado na 3.5(b). Observe que a sequência em amarelo representa o menor destes três caminhos e portanto, o caminho mínimo entre os vértices i e j vale 2.

Se fixarmos i e fizermos este mesmo processo para os outros nós da rede e tirarmos uma média dos valores dos menores caminhos para se ir de i para qualquer outro vértice, teremos o caminho mínimo médio do vértice i, que representa em média a distância do vértice i (em termos de arestas) para qualquer outro vértice da rede. No exemplo da Figura 3.5 o caminho mínimo do vértice i ou para o vértice i é de 1,83.

A média dos caminhos mínimos médios de todos os vértices da rede é chamada de caminho mínimo médio ($\langle \ell \rangle$) da rede, ver Equação 3.6. Este indicador representa, em média, qual o menor caminho entre dois nós quaisquer da rede. Para exemplificar, considere a rede da Figura 3.5. O caminho mínimo médio da rede é de $\langle \ell \rangle = 1,86$.

$$\langle \ell \rangle = \frac{1}{n(n-1)} \sum_{i \neq j} \ell_{ij} \tag{3.6}$$

Na Equação 3.6, ℓ_{ij} representa algum caminho que conecte os dois vértices.

No presente trabalho, para efeito de cálculo, iremos considerar como *zero* o valor do *caminho* entre dois vértices que não se conectam, ou seja que pertencem a componentes diferentes da rede.

3.4.6 Diâmetro

Chama-se diametro de uma rede o valor de seu maior caminho mínimo. O diâmetro do grafo (Figura 3.4(a)) tem valor 4 e pode ser ilustrado na Figura 3.6.

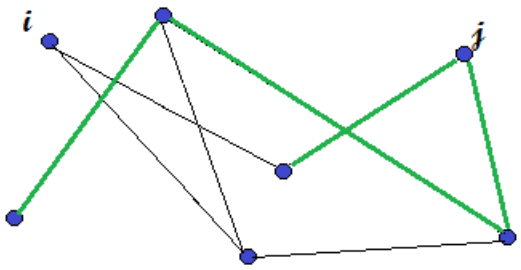


Figura 3.6: Diâmetro de um grafo

Para as redes deste trabalho, em geral desconectadas (mais de um componente), o *caminho mínimo médio* e o *diâmetro* tem valor próximo do que o valor destes mesmos índices para o maior componente da rede. Dessa forma, quando o maior componente tem um número

de vértices muito próximos da rede toda, o cálculo destes índices é feito como descrito nas equações acima. Nas redes em que o maior componente tem um uma parcela significativa de nós a menos, estes índices são calculados para o maior componente.

Existem inúmeros outros indicadores úteis, porém os supracitados são suficientes para as primeiras análises desta pesquisa. Estes indicadores são fundamentais para observar características em redes complexas com muitos nós, mesmo utilizano conceitos e métricas simples. As maiores dificuldades estão relacionadas à cálculos de índices em redes muito grandes (e.g. bilhões, trilhões de nós). Dessa forma, estamos limitado apenas à capacidade de processamento de nossos computadores e de algoritmos de varredura e busca de padrões mais inteligentes.

A próxima seção mostra como estes índices de redes ajudam na caracterização de uma rede, quando se percebe padrões na dinâmica de conexão dos nós.

Capítulo Quatro

Modelos de redes

Conforme foi visto, rede complexa é uma maneira de descrever um sistema natural, tecnológico ou social, que evidencia os relacionamentos entre as partes do sistema (e.g. internet e redes *www*; neurônios; palavras de um discurso oral ou escrito; cadeias alimentares; relações sociais entre pessoas; epidemias). De acordo com a seção anterior, a estrutura de um grafo é adequada para esta modelagem.

A mera inspeção visual em grafos pequenos é suficiente para revelar a importância de um ator (i.e. nó da rede) em relação a seus vizinhos ou ao grafo todo. Entretanto, essa maneira de analisar um grafo ou um nó não é viável para sistemas com grande quantidade de elementos (e.g. milhões, bilhões de nós).

Assim, para entender melhor um sistema é preciso muito mais do que apenas estudar suas partes menores. É preciso levar em conta os relacionamentos entre as partes do sistema. As tentativas em não se fazer isto certamente levarão o pesquisador a resultados incompletos, ao se deparar com questionamentos como - o comportamento de uma parte menor do sistema pode depender da relação de seus elementos com elementos de outras partes.

Surge então uma necessidade crescente em estudar sistemas como um todo, transcendendo abordagens reducionistas. A física estatística tem contribuído muito no estudo de redes, pois consegue descrever de maneira razoável o comportamento de sistemas com grande quantidade de elementos (eg. um gás no interior de um recipiente). Um dos caminhos para uma análise holística mais confiável é saber de que forma os elementos de um sistema se relacionam entre si.

Inicialmente, a descrição de sistemas modelados através de redes era feita a partir de modelos de grafos aleatórios (Seção 4.1). Neste tipo de estrutura, os elementos interagem entre si a partir de uma dada probabilidade. Hoje se sabe que a grande maioria dos sistemas modelados a partir de redes possuem elementos que não se relacionam randômicamente.

Torna-se fundamental a existência de modelos que representem bem as topologias das redes de sistemas naturais e que possuam ferramentas quantitativas e qualitativas que sejam capazes de caracterizar de maneira razoável os princípios que governam os relacionamentos dos seus nós. Felizmente, recentes descobertas neste campo estão relacionadas à maneira como as redes do mundo real diferem de redes aleatórias (NEWMAN, 2003).

> "*Deve haver uma nítida diferença nas regras que governam a localização de* links *nas várias redes que encontramos na natureza. Descobrir um modelo para descrever todos esses diferentes sistemas parece, a primeira vista, um desafio intransponível* (BARABáSI, 2002, p.15).
> "

Outros modelos, além do modelo aleatório (também listado abaixo) se consagraram na literatura científica por se adequar à grande maioria dos sistemas naturais modelados a partir de Redes Complexas[1]:

- O Modelo $G(n, p)$, para grafos aleatórios (ERDOS; RENYI, 1960). Apesar de não se adequar à maioria dos sistemas naturais, este modelo pode servir de referência para comparação com redes reais;
- O Modelo $small-world$ mostra que, mesmo para sistemas com muitos elementos, os caminhos entre os quaisquer dois nós são relativamente curtos (WATTS; STROGATZ, 1998);
- O Modelo $scale-free$ é capaz de descrever inúmeros sistemas naturais, e se caracteriza pela distribuição do número de conexões dos seus elementos seguir uma lei de potência;
- O Modelo *rede de cliques* (*network of cliques*) possui métricas específicas para sistemas que evoluem a partir da entrada de grupos de nós mutuamente conectados, chamados cliques (eg. redes de coautoria, rede de atores de filmes) (FADIGAS; PEREIRA, 2013);
- A modelagem através de uma *rede semântica* (*Semantic Networks*) é útil para sistemas em que os elementos são dotados de significados linguísticos (STERNBERG, 2011).

4.1 Redes aleatórias

Uma rede aleatória é obtida a partir de um conjunto $V(G) = v_1, ..., v_n$ de vértices e adicionando-se arestas entre eles aleatoriamente. Esse processo de adição aleatório de arestas pode ser dado por uma probabilidade fixa p para cada vértice da rede se conectar com outro.

Neste caso, trata-se do modelo conhecido como $G(n; p)$ (ERDOS; RENYI, 1960). O parâmetro n determina o número de vértices da rede, rotulados de 1 a n. O parâmetro p determina a probabilidade de uma determinada aresta ser incluída na rede.

[1] Uma rede complexa, entretanto, pode ser descrita por mais de um modelo.

Para exemplificar imagine uma situação em que uma rede aleatória é construída com 200 vértices e probabilidade $p = \frac{1}{6} = 0,17$ de conexão entre os vertices. Diante disto é feito o seguinte procedimento: Para cada par de vértices joga-se um dado, se cair o número 6 estes vértices serão conectados, se não cair 6, se escolhe outro par de vértices e o procedimento é repetido até que todos os $\binom{200}{2} = 19900$ pares de vértices tenham tido a chance de se conectar. O resultado é um grafo $G(200; 0,17)$, como o mostrado na Figura 4.1.

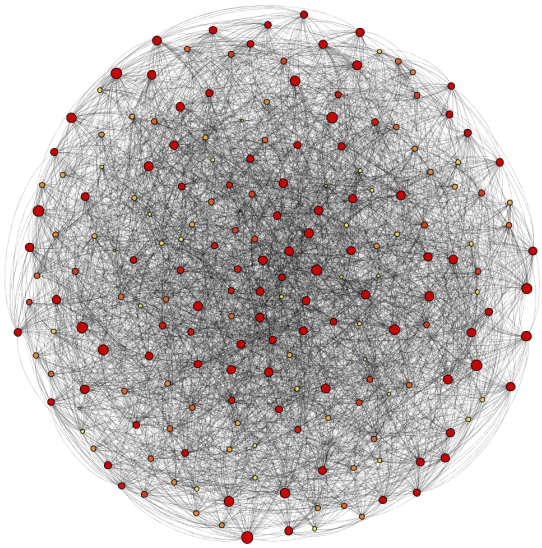


Figura 4.1: Rede aleatória formada a partir do modelo $G(n,p)$, com $n = 200$ e $p = 0,17$. Os tamanhos dos nós são porporcionais à seus graus.

Redes que seguem este modelo possuem, em geral, distribuição de graus que se ajustam à uma curva normal (Figura 4.2). Neste modelo, cada possível aresta é incluída de forma independente das outras. No entanto, o modelo não determina a estrutura da rede, e sim o processo aleatório que irá gerar essa estrutura.

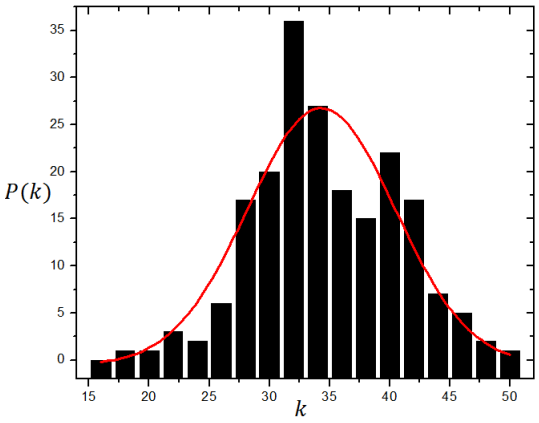


Figura 4.2: Distribuição de Graus de uma rede aleatória formada a partir do modelo $G(n,p)$, com $n = 200$ e $p = 0,17$ (Figura 4.1). A linha vermelha é a curva de ajuste para para uma distribuição Gaussiana.

Outra forma de se obter rede aleatória é distribuir um número fixo de arestas M aleatoriamente[2] em um número fixo n de vértices, com igual probabilidade de conexão. Este modelo é conhecido como modelo $G(n; M)$. É como se escolhêssemos aleatoriamente um grafo $G_{(n,M)}$, dentre os $\binom{\binom{n}{2}}{M} = C_{(n;M)}$ grafos possíveis, com igual probabilidade de escolha (ERDOS; RENYI, 1960).

A diversidade de sistemas naturais e sociais e das leis que governam os relacionamentos entre os elementos desses sistemas evidenciam em sua grande maioria redes reais que diferem consideravelmente do modelo $G(n; p)$. Por exemplo, sabemos que muitas redes reais possuem uma distribuição de grau com cauda longa, o que não ocorre com o modelo randômico, cuja distribuição de graus é normal. Da mesma forma, o coeficiente de aglomeração médio em redes reais quase sempre é muito maior (em ordens de grandeza) que o de redes do modelo $G(n; p)$.

Neste sentido, outros modelos surgiram para dar conta da diversidade de formas de conexão entre elementos de sistemas naturais e sociais. Na sua grande maioria, as conexões de nós de uma rede real são feitas por serem mais prováveis ou por minimizar energia (eg. As relações de amizade; relacionamentos de co-autoria; relações predatórias entre animais em uma selva; as reações químicas ao unir elementos em uma mesma célula do corpo humano).

4.2 Redes de mundo pequeno

Redes de mundo pequeno, *small-world* ou como inicialmente foram chamadas - fenômeno dos *seis graus de separação*[3], são modelos que descrevem redes que possuem caminhos curtos entre seus vértices. Por esta característica, este tipo de rede em geral, também possui alta aglomeração entre os vértices. Isto que dizer que, em uma rede social de amizades, por exemplo, é muito frequente pessoas que tenham amigos em comum, o que torna a distância entre desconhecidos menor (baixo valor de $\langle \ell \rangle$). Outra característica muito comum deste fenômeno é que os amigos de uma pessoa da rede se conheçam entre si (Alto $\langle C \rangle$).

Em 1998, Watts e Strogratz propôs um modelo para redes com o fenômeno *small-world* (Figura 4.3). O modelo considera a construção a partir de uma rede regular (i.e. redes que possuem todos os vértices com mesmo grau). Os vértices desta rede, então são reconectados com uma dada probabilidade p. Assim, para $p \to 0$ temos uma estrutura que tende a ser uma rede regular. Para $p \to 1$ temos uma rede que tende a ser aleatória,

[2] $0 \leq M \leq \binom{n}{2}$.

[3] Este nome vem do famoso experimento proposto por Stanley Milgram, que tinha como objetivo determinar a distância entre quaisquer duas pessoas nos Estados Unidos, escolhidas aleatoriamente (MILGRAM, 1967).

com distribuição de graus com curva normal. O fenômeno *small world* acontece, então, para valores intermediários de p (WATTS; STROGATZ, 1998).

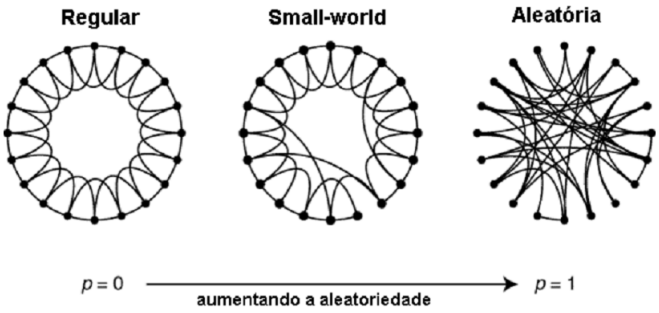


Figura 4.3: O fenômeno *small-world* ocorre a meio caminho do processo de aleatorização de uma rede regular (WATTS; STROGATZ, 1998).

Considere as condições abaixo para uma rede real.

- Não direcionada;
- Não ponderada;
- Sem arestas múltiplas;
- Esparsas;
- Conectadas.

Segundo Watts e Strogatz (1998), uma rede real nestas condições apresenta o efeito small-world se $\langle C \rangle \gg \langle C \rangle_{rd}$ e se $\langle \ell \rangle$ é comparável com $\langle \ell \rangle_{rd}$. Nesta definição, $\langle C \rangle_{rd}$ é o coeficiente de aglomeração médio para uma rede aleatória com mesmo grau médio $\langle k \rangle$ e mesmo número de vértices n. Analogamente, $\langle \ell \rangle_{rd}$ é o caminho mínimo médio para a rede aleatória correspondente.

A distribuição de graus neste tipo de rede pode ser qualquer uma. Entretanto, em geral assemelha-se à distribuição de Gauss. Isto indica que a rede é relativamente homogênea e os vértices tem aproximadamente o mesmo grau (Figura 4.4). Como exemplos de redes assim, temos as redes sociais em geral.

4.3 Redes livres de escala

Ainda que muitas redes reais contenham mecanismos intrínsecos de encurtar o caminho entre dois vértices, evidenciando o fenômeno *small-world*, nem sempre a dinâmica de

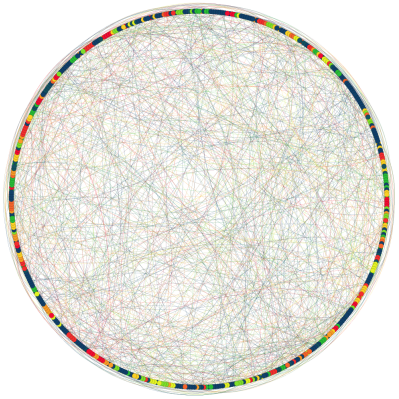


Figura 4.4: Exemplo de rede que apresenta o fenômeno *small-world* de Watts & Strogratz.

conexão de vértices segue esta lógica. Em muitos casos, as conexões são feitas privilegiando nós com mais conexões, de acordo com o dito popular *"O rico fica cada vez mais rico"*.

Em 1999, Barabási e Albert propuseram um modelo de rede em que poucos vértices possuem muitas conexões e muitos vértices possuem poucas conexões (Figura 4.5), fazendo com que a distribuição de graus siga uma lei de potência. Este modelo ficou conhecido como $scale-free$, i.e. rede livre de escala. Os gráficos da Figura 4.6 ilustram a diferença da distribuição de uma rede aleatória para uma rede livre de escala.

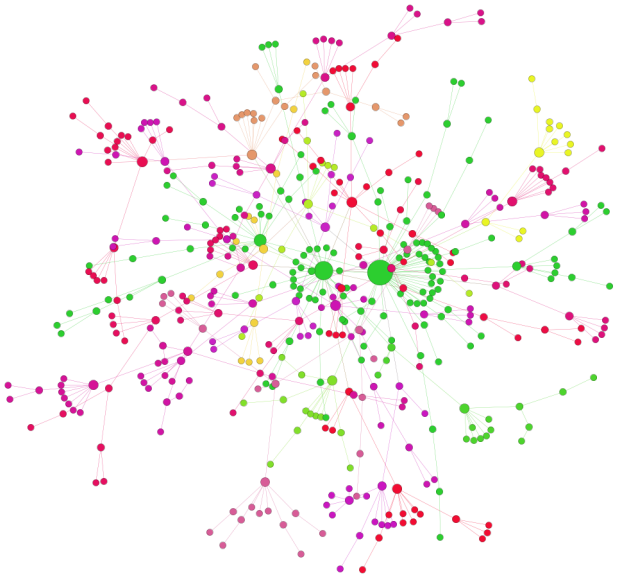


Figura 4.5: Exemplo de rede livre de escala do modelo de Barabási e Albert. O tamanho dos vértices são proporcionais aos seus respectivos graus.

Devido ao mecanismo inerente de ligação preferencial, os vértices que possuem maior grau (chamados de Hubs) terão maior probabilidade de realizar novas conexões. Assim, a forma da distribuição dos graus não varia com o tempo e é descrita por uma lei de potência $P(k) \propto k^{-\gamma}$, em geral com $2 \leq \gamma \leq 3$, onde k é o número de conexões feitas por um ou mais vértices da rede e $P(k)$ representa o número de vértices com grau k

(BARABASI; ALBERT, 1999).

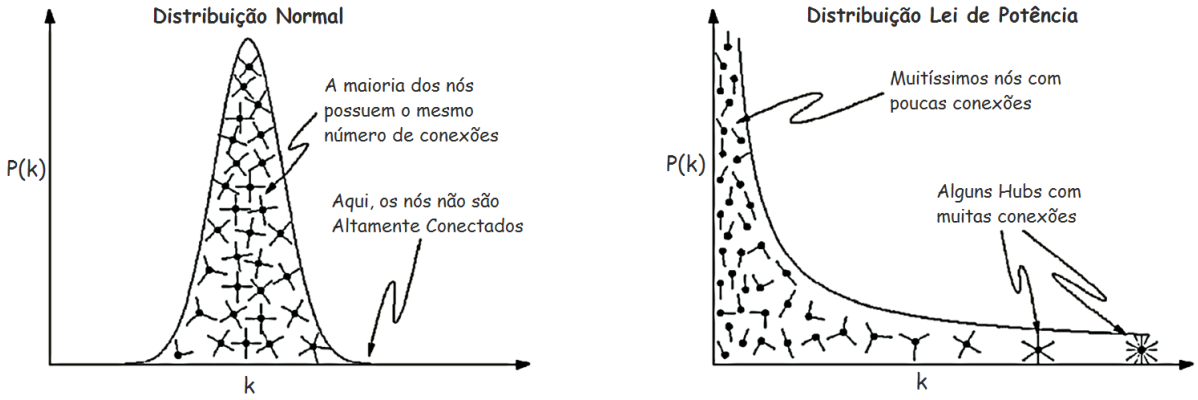


Figura 4.6: As redes aleatórias possuem uma distribuição de graus que segue uma curva normal (à esquerda), com um grau médio representativo; nas redes livre de escala a distribuição dos graus segue uma Lei de Potência (à direita). O grau médio para este tipo de rede não é representativo. Entretanto, é comum a existência de fenômenos raros. Adaptado de Barabasi (2007).

Um ponto negativo para redes livre de escala é que são vulneráveis a ataques (a remoção de um hub pode fazer com que muitos vértices se desconectem, o que torna o diâmetro da rede muito maior). Nas redes Aleatórias, um ataque não desestrutura a rede, já que a maioria dos vértices possuem valores de grau médio muito próximos. Entretanto, se houver uma falha em uma rede Livre de Escala de maneira aleatória, será muito raro um Hub ser removido. Portanto esse tipo de redes é resistente à falhas (BARABASI; ALBERT; JEONG, 1999).

> "Quando começamos a mapear a Web, esperávamos que os nós seguissem uma distribuição em forma de sino, como no caso da altura das pessoas. Em vez disso, descobrimos alguns nós que desafiavam explicações evidentes (quase como se tivéssemos tropeçado em uma quantidade significativa de pessoas de 100 metros de altura), o que nos levou a criar o termo *livre de escala* (BARABáSI A. L., 2003, p.53). "

Redes no mundo real, em sua grande maioria, são abertas a entrada de novos vértices e novas arestas. A dinâmica de crescimento, com as ligações preferenciais, são mecanismos extremamente comuns em muitos sistemas naturais e sociais (BARABASI; ALBERT, 1999).

Além destes modelos, já bastante conhecidos na literatura - redes *scale free* (BARABASI; ALBERT; JEONG, 1999), redes *small world* (WATTS; STROGATZ, 1998) e *redes aleatórias* (ERDOS; RENYI, 1960), existem os modelos de *redes de cliques*, e de *redes semânticas*, que podem ou não se ajustar aos três primeiros. Estes dois modelos são pouco difundidos na literatura. Por isso, são abordados nas próximas seções.

4.4 Rede de cliques

Redes de cliques são redes formadas apenas por um conjunto de cliques unidas, parcialmente unidas ou não unidas. Uma clique é definida como uma rede ou subrede que contém todos os seus vértices conectados entre si (ERDOS, 1966). Isto implica que os valores dos índices *densidade, coeficiente de aglomeração, caminho mínimo médio* e *diâmetro* para uma clique terão valor igual a 1. Ao unir cliques, constrói-se o que denomina-se *rede de cliques*. Este tipo de rede têm inúmeras aplicações:

- Redes de coautoria: Todos os autores de um mesmo trabalho são conectados entre si. Neste tipo de rede, cada autor é um vértice na rede. Existe aresta entre dois vértices se estes autores são coautores de um mesmo trabalho. Cada publicação que é inserida na rede representa um conjunto de autores coautores, ou seja vértices conectados entre si (cliques);

  Segundo Vanz e Stumpf (2010) a coautoria é apenas uma faceta da colaboração científica, pois ela não mede a colaboração em sua totalidade e complexidade. Newman (2001) mostra através da distribuição de graus dessas redes que em todos os campos estudados a maioria dos autores possuem poucos coautores, no entanto existem poucos autores que possuem centenas, até milhares de coautores. Ele ressalta que este tipo de rede é mais "social" do que redes de atores de filmes, pois nesta ultima os atores não necessariamente se conhecem;

- Redes de atores de filmes: As cliques representam os atores que participaram de um mesmo filme. Da mesma forma, a cada filme corresponde a uma clique de atores conectados. Alguns pesquisadores estudaram este tipo de redes e concluíram que elas podem se apresentar como small-world, scale free ou até as duas topologias (BARABASI; ALBERT, 1999; WATTS; STROGATZ, 1998);

- Redes de discursos orais ou escritos: Os vértices da clique são palavras pertencentes à uma mesma sentença (CALDEIRA, 2005; TEIXEIRA et al., 2010; AGUIAR, 2009). Informações sobre estes tipos encontram-se nas Seções 4.5.4 e 4.5.3;

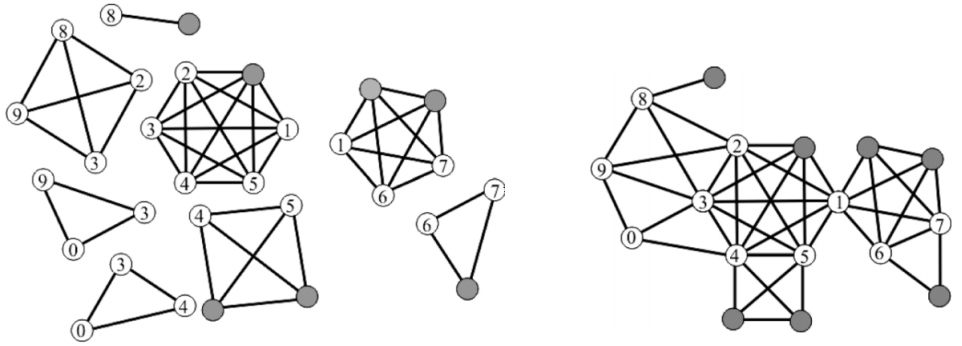
- Redes de títulos: Cada título comporta-se como uma clique e os vértices da clique são as palavras que compõe o título (FADIGAS et al., 2009; PEREIRA et al., 2011; FADIGAS et al., 2013).

4.4.1 Processos de formação de rede de cliques

O elemento básico em uma rede deste tipo não é o vértice, mas sim um conjunto deles mutuamente conectados, ou seja a clique. Existem duas maneiras de se conectar duas

cliques: *justaposição* ou *sobreposição*. O primeiro processo se dá pela união de duas cliques a partir de um único vértice comum. O segundo, pela junção de duas cliques com dois ou mais vértices em comum (FADIGAS; PEREIRA, 2013).

A Figura 4.7 mostra uma rede formada por justaposição e sobreposição de suas cliques, bem como, o aspecto inicial antes das cliques serem unidas.



(a) Estado inicial das cliques isoladas

(b) Rede de cliques formada por justaposição e/ou sobreposição das cliques da Figura 4.7(a)

Figura 4.7: Estado inicial de cliques isoladas e uma possível configuração para redes de cliques. Fonte: Fadigas e Pereira (2013, p.2577)

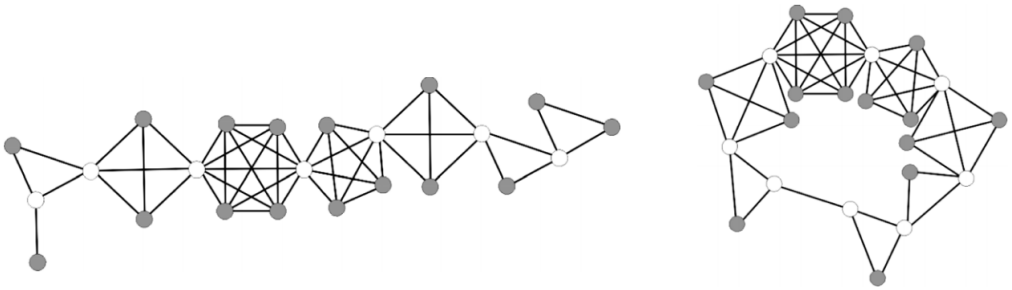
4.4.2 Classificação de uma rede de cliques

Esta pesquisa utiliza alguns índices de coesão para redes de cliques, propostos por Fadigas e Pereira (2013). Quando tem-se uma configuração inicial de cliques desconectadas, o número de arestas m_0 é dado pela soma da quantidade de arestas em cada clique. Designa-se n_q o número de cliques, q_i o tamanho (número de vértices) da i-ésima clique e n_0 o número total de vértices do estado inicial das cliques isoladas.

Fadigas e Pereira (2013) apresentam estruturas de cliques minimamente conectadas: *estrela, círculo, camada e Llinha*. Uma rede de cliques minimamente conectada é um conjunto de cliques unidas, que formam um único componente, em que quaisquer duas cliques foram unidas por apenas 1 vértice, Figura 4.8.

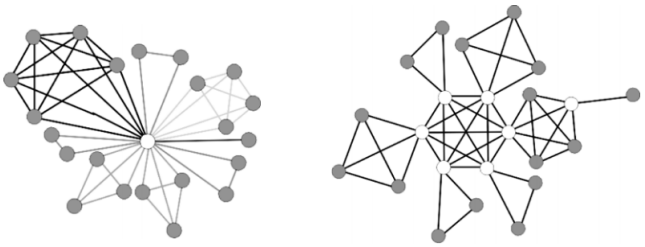
Para comparar uma Rede de cliques real com estas quatro estruturas teóricas, faz-se o uso do *diâmetro de referência* normalizado em escala logarítmica (Equação 4.1):

No encadeamento tipo linha a rede é conectada clique a clique. Isto torna a rede pouco eficiente, com *caminhos mínimos médios* e *diâmetro* muito altos. No tipo estrela o vértice



(a) Rede de cliques minimamente conectadas com encadeamento *Linha*

(b) Rede de cliques minimamente conectadas com encadeamento *Círculo*



(c) Rede de cliques minimamente conectadas com encadeamento *Estrela*

(d) Rede de cliques minimamente conectadas com encadeamento *Camada*

Figura 4.8: Exemplo de estruturas teóricas para redes de cliques minimamente conectadas através do processo de Justaposição . Fonte: Fadigas e Pereira (2013, p. 2578)

D^*_{ref}	Estrutura Teórica de Cliques
$0.00 - 0.25$	Layout Estrela
$0.26 - 0.75$	Layout Círculo ou Camada
$0.76 - 1.00$	Layout Linha

$$D^*_{ref} = \frac{ln(D/2)}{ln(n_q/2)} \tag{4.1}$$

Tabela 4.1: Classificação de uma rede de cliques através do Diâmetro de referência. Fonte: Fadigas e Pereira (2013, p. 2581)

mais importante da rede conecta todas as cliques.

4.4.3 Índices para redes de cliques

Pode-se reescrever os índices *clássicos* (ou *atemporais*) em função dos parâmetros acima citados para redes de cliques reais, considerando todas as justaposições e sobreposições, e o quanto estes valores diferem dos mesmos índices para a mesma rede, só que com cliques desconectadas.

O **Grau médio** de uma rede de cliques desconectadas é dado pela Equação 4.2:

$$\langle k_{q0} \rangle = \frac{\sum_{i=1}^{n_q} q_i(q_i - 1)}{n_0} \tag{4.2}$$

Para quantificar a **variação do grau médio** da rede de cliques, comparada com a mesma rede, só que desconectada, é dada pela Equação 4.3:

$$\upsilon(\langle k \rangle) = \frac{\langle k \rangle - \langle k_{q0} \rangle}{\langle k_{q0} \rangle} \tag{4.3}$$

A **densidade** de uma rede de cliques desconectada é dada pela Equação 4.4:

$$\Delta_{q0} = \frac{2m}{n_0(n_0-1)} = \frac{\sum_{i=1}^{n_q} q_i(q_i-1)}{n_0(n_0-1)} \tag{4.4}$$

A **variação de densidade** é dada pela Equação 4.5:

$$\upsilon(\Delta) = \frac{\Delta - \Delta_{q0}}{\Delta_{q0}} \tag{4.5}$$

De acordo com Fadigas e Pereira (2013), a expressão 4.5 mede o "adensamento"da rede, em relação ao seu estado inicial (rede de cliques desconectadas).

Para rede de cliques reais vale a aproximação $n \gg 1 \Rightarrow n-1 \simeq n$ e $n_0 \gg 1 \Rightarrow n_0 - 1 \simeq n_0$.

A expressão da variação da densidade pode também ser escrita em função do número de arestas e vértices da rede:

$$\upsilon(\Delta) = \frac{m}{m_0} \cdot \frac{n_0}{n} \cdot \frac{n_0-1}{n-1} - 1 \simeq \frac{(n_0/n)^2}{(m_0/m)} - 1 \tag{4.6}$$

Assim, a variação na densidade é diretamente proporcional ao quadrado da razão entre o número de vértices e inversamente proporcional à razão entre o número de arestas. Isto foi possível graças à aproximação $n \gg 1 \Rightarrow n-1 \simeq n$ e $n_0 \gg 1 \Rightarrow n_0 - 1 \simeq n_0$.

4.5 Redes semânticas

O elemento principal que aparece relacionado em uma rede semântica é a palavra. Esta forma de representação adequa-se a uma variedade de métodos computacionais e tem sido cada vez mais estudada por cientistas de diversas áreas. Uma rede semântica é, então, a rede de um conjunto de elementos - palavras, conceitos ou entidades - interconectados, que estão relacionados através dos significados (i.e. símbolos linguísticos) (STERNBERG, 2011).

Atualmente, as Redes semânticas vêm sendo representadas de acordo com a *teoria dos grafos*, em que cada vértice da rede representa uma palavra e as arestas representam ligações entre essas palavras. Com este formato é possível visualizar como cada conceito

está sendo definido nos termos de sua posição na rede de relacionamento entre eles. Por exemplo, considere o seguinte texto:

> *Hoje me sinto bem. Ajudei o meu filho a ter uma casa. Chamei minha mulher no sofá. Perguntei a meu bem sua opinião. Minha mulher concordou. Me desfiz de bens materiais, casa de praia e carro. Com o dinheiro comprei uma casa. Meu filho e sua mulher ficaram bem felizes e vão se casar em dezembro.*

Observe que a palavra "bem" possui diferentes significados, a depender do contexto em que está inserida. Pode representar felicidade, vantagem, virtude; algo bom, que está de acordo com a moral; referência à pessoa a quem se dedica afeição; ou até significar haveres, propriedades.

Uma estrutura possível para representar os conceitos contidos neste trecho e os seus relacionamentos está na Figura 4.9. Através da rede semântica é possível visualizar como os diferentes significados da palavra "bem" se relacionam, ou até, inferir interpretações para a proximidade entre dois conceitos - na figura pode-se ver (em vermelho) o caminho entre as palavras "concordar" e "desfazer".

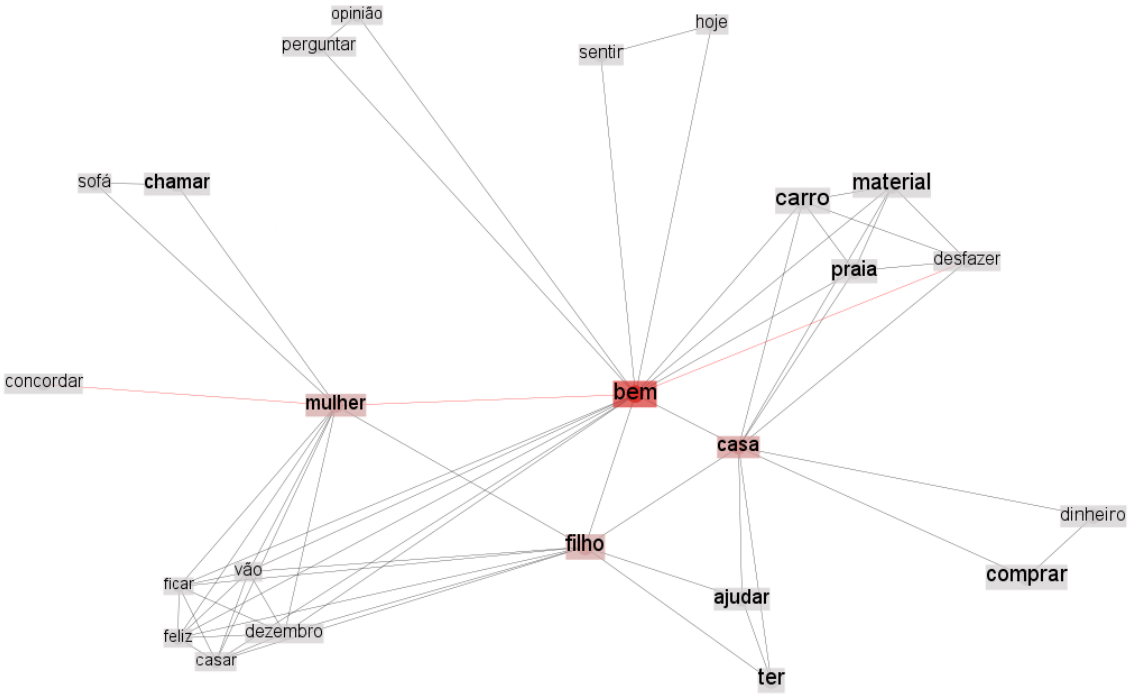


Figura 4.9: Rede semântica formada por cliques, de acordo com o tratamento proposto por Caldeira (2005). Observe a palavra *bem* em seus diferentes contextos. As ligações em vermelho representam o menor caminho (ℓ) entre as palavras "concordar" e "desfazer".

Observar um conceito através de uma rede semântica é identificar o mais fiel possível a posição, importância e significados de uma palavra dentro texto, já que é possível

visualizar os contextos nos quais a palavra associada ao conceito observado está inserida, principalmente se o texto for muito grande.

A técnica da modelagem de um texto através de redes semânticas oferece um meio empírico de acesso à organização mental do conhecimento (ALBUQUERQUE; PIMENTEL, 2004). A proposta clássica dos pioneiros neste estudo sugere modelos gráficos baseados em redes semânticas para representação do conhecimento, a fim de demonstrar como o conhecimento humano é estruturado através dos conceitos evocados na fala ou escrita, a partir de diferentes estímulos ou perspectivas. Para uma completa descrição dos autores clássicos que iniciaram este estudo, sugere-se a leitura de Sternberg (2011).

A Psicologia Cognitiva, área interdisciplinar da Psicologia, tem utilizado as redes semânticas para investigar a memória, mais precisamente a *memória declarativa*. É a partir desta memória que um indivíduo externaliza o conhecimento sobre o mundo que o cerca (STERNBERG, 2011). O conhecimento declarativo do indivíduo é então evocado através de signos linguísticos (palavras). Assim, a linguagem, oral ou escrita, é um instrumento utilizado no processo de acesso à memória através de signos linguísticos e pode ser representada através de redes semânticas, que podem indicar o que está na memória declarativa de quem faz o discurso.

Aparentemente as pessoas possuem liberdade com as palavras evocadas em um discurso. Entretanto, existem algumas regularidades na linguagem humana, que podem ser percebidas a partir da estatística que relaciona as palavras evocadas com a frequência de aparição delas no texto. Curiosamente, a distribuição obedece a uma lei de potência e foi observada por Zipf (1972). Isto sugere que a linguagem humana, pelo menos do ponto de vista da memória declarativa, comporta-se como um sistema complexo.

Uma propriedade importante de um sistema complexo é que o todo é sempre maior que a soma de suas partes, porque se os elementos do sistema interagem, é comum se perceber propriedades que emergem destas interações, o que seria impossível se analisarmos o elemento sozinho. Por exemplo, um *formigueiro* é um bom exemplo de um sistema complexo. Cada formiga age de acordo com regras pré-determinadas, a depender de sua vizinhança, e as interações entre elas geram um formigueiro inteligente que busca comida para o grupo, resolvem problemas geométricos, etc (NUSSENSVEIG, 2008).

Da mesma maneira, uma palavra por si só pode apresentar diversos significados a depender de quais palavras ela tem como vizinhas. Um conjunto de palavras com organização sintática tem significado próprio. Segundo Caldeira (2005), a sentença é a menor unidade de significado de um texto. Este significado é uma propriedade que emerge do conjunto de palavras que compõe a sentença. Estas palavras se relacionam segundo regras sintáticas.

De acordo com Aguiar (2009), a linguagem humana pode ser vista como um fenômeno em que signos linguísticos com significados próprios são organizados de forma a gerar uma estrutura com significado diferente da soma de cada unidade linguística. Assim, este sistema de signos linguísticos, que surge de um processo mental dinâmico, complexo e associativo na memória declarativa, pode ser modelado como uma Rede, em que os vértices são representados pelas palavras evocadas e as arestas são as associações entre estas palavras.

Existem várias maneiras de configurar os relacionamentos entre palavras de um texto na modelagem utilizando redes complexas. Tudo depende de como a vizinhança de cada palavra é vista pelo pesquisador. Por exemplo, para uma rede não dirigida, a configuração onde os vizinhos de uma palavra são as duas palavras que se encontram imediatamente antes e depois dela é denominada como rede semântica em formato de *linha*. Outra configuração semelhante é a rede semântica em formato de *círculo*, onde a primeira e última palavra de cada sentenças são consideradas vizinhas. Existe uma configuração, em que se considera como a vizinhança de uma palavra todas as outras palavras da mesma sentença que ela. Essa configuração denomina-se rede semântica em formato de *cliques*.

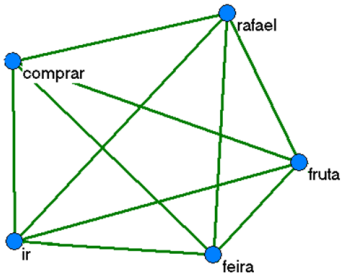
4.5.1 Redes semânticas de cliques

Para este tipo de rede, acredita-se que a inclusão de uma nova ideia em qualquer texto é dada pela inclusão de uma sentença neste texto. Cada sentença contém todas as suas palavras relacionadas entre si e, se uma ou mais palavras forem comuns em duas ou mais sentenças do texto, então elas conectam dois significados presentes no texto. O compartilhamento de palavras entre cliques pode fazer emergir propriedades semânticas no texto, que dificilmente seriam vistas se analisássemos palavras ou frases isoladas.

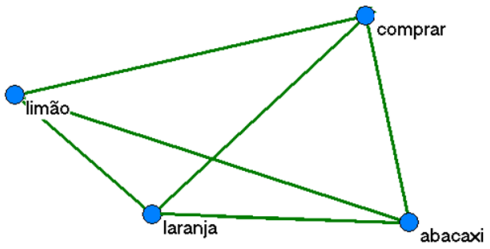
A rede da Figura 4.9, vista anteriormente é um exemplo de uma rede semântica de Cliques. A seguir, apresenta-se o procedimento usado para gerar estas redes. Para esta construção é preciso antes realizar um tratamento manual e computacional das palavras, que vai ser explicado nas Seções 5.2.1 e 5.2.2. Por enquanto, é desejável apenas que se perceba o processo de junção das cliques, pelo compartilhamento das palavras comuns. Considere o seguinte trecho:

"Rafael foi à feira comprar frutas. Comprou laranja, limão e abacaxi. Em casa Rafael fez um suco com o que comprou."

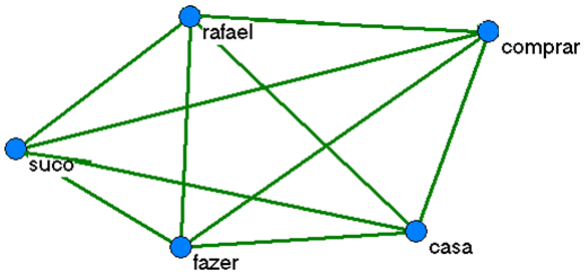
As três sentenças que compõe o trecho acima poderão ter seus significados representados pelos seguintes grafos, na figura 4.10. Entretanto em cada sentença de um texto podem



(a) Sentença 1: Rafael foi à feira comprar frutas.



(b) Sentença 2: Comprou laranja, limão e abacaxi.



(c) Sentença 3: Em casa Rafael fez um suco com o que comprou.

Figura 4.10: Sentenças de um discurso em forma de redes semânticas.

existir palavras comuns ou palavras de mesma forma canônica das derivadas de uma dada palavra em outras sentenças. Assim, os grupos de sentenças interagem por conta destas palavras em comum, formando assim um único sistema, i.e. rede semântica de cliques, Figura 4.11:

4.5.2 Incidência-fidelidade

O conceito de *incidência-fidelidade* (ou *força-fidelidade* como foi inicialmente chamado) advém do conceito de *força* proposto por Nelson, McEvoy e Schreiber (1998), que mede em um texto com N_S sentenças a probabilidade ρ de ocorrência de um par, que apareceu em n sentenças do texto. A medida é dada pela Equação 4.7):

$$\rho_{PAR} = \frac{n}{N_S} \tag{4.7}$$

Nelson, McEvoy e Schreiber (1998) investigaram como é organizado o conhecimento na memória humana, a partir do uso da linguagem oral. Posteriormente, o trabalho de Teixeira (2007) - que trata da análise de redes geradas por discursos orais de indivíduos - propõe um novo método para construção de redes semânticas, baseado no mesmo conceito

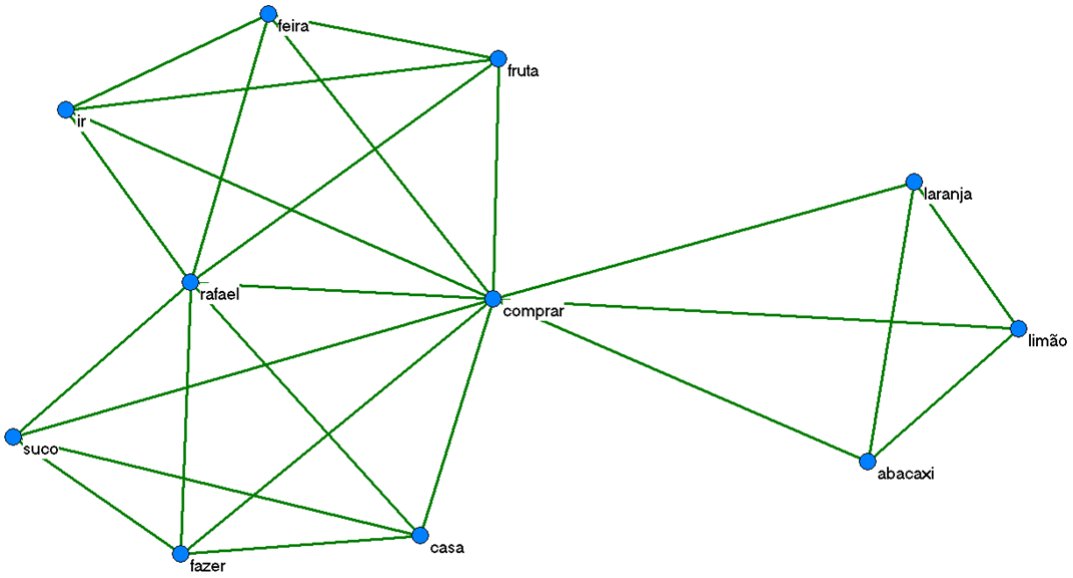


Figura 4.11: Exemplo de rede semântica.

de *Força* proposto por Nelson, McEvoy e Schreiber (1998).

Para isso, ela desenvolveu o índice *força-fidelidade*, que mais tarde Teixeira et al. (2010) chamaram de *incidência-fidelidade*. Basicamente, o nome *força* mudou para o nome *incidência* por uma questão conceitual da Física, já que este ultimo trabalho foi publicado em um periódico no campo da Física, que já existe o conceito "força"(newtoniano) bem definido. O termo *Incidência* (I) tem o mesmo significado do termo *Força* da equação 4.7 e pode ser calculado de acordo com a Equação 4.8:

$$I_{(\Psi,\Omega)} = \frac{|C_\Psi \bigcap C_\Omega|}{|\bigcup_{i=1}^{N_S} C_i|} \tag{4.8}$$

O termo *fidelidade* (F) representa a probabilidade de um par de palavras, que aparece em uma mesma sentença, ocorra em um universo de sentenças que pelo menos uma das duas palavras ocorre, e é representado pela Equação 4.9. Ou seja, mede a probabilidade de ocorrência de um par no universo de sentenças que o contextualiza.

$$F_{(\Psi,\Omega)} = \frac{|C_\Psi \bigcap C_\Omega|}{|C_\Psi \bigcup C_\Omega|} \tag{4.9}$$

A *incidência-fidelidade* ($IF_{(\Psi,\Omega)}$), para um par de palavras Ψ e Ω, é dada pelo produto dos Índices $I_{(\Psi,\Omega)}$ e $F_{(\Psi,\Omega)}$ (Equação 4.10) e será um importante índice na construção das redes semânticas deste trabalho. Para cada par de palavras, seus valores atuam como

peso nas arestas das redes de cada texto.

$$IF_{(\Psi,\Omega)} = I_{(\Psi,\Omega)} \cdot F_{(\Psi,\Omega)} \tag{4.10}$$

Nas Equações 4.8 e 4.9, C é um conjunto das sentenças do texto e C_i é um subconjunto de C, que representa o conjunto de sentenças em que a palavra i faz parte. Nesse contexto Ψ e Ω compõem um par de palavras arbitrário, Sendo C_Ψ o conjunto formado pelas sentenças em que a 1ª palavra do par ocorre e C_Ω o conjunto formado pelas sentenças em que a 2ª palavra do par ocorre.

Assim sendo, considere $C_p \equiv C_\Psi \bigcap C_\Omega$ como sendo o conjunto formado pelas sentenças em que ocorreu o par de palavras Ψ e Ω. Dessa forma, pode-se reescrever a Equação 4.10 para 4.11.

$$IF_{(\Psi,\Omega)} = \frac{|C_\Psi \bigcap C_\Omega|}{|\bigcup_{i=1}^{N_p} C_i|} \times \frac{|C_\Psi \bigcap C_\Omega|}{|C_\Psi \bigcup C_\Omega|} = \frac{|C_p|}{|C|} \times \frac{|C_p|}{|C_\Psi| + |C_\Omega| - |C_p|} \tag{4.11}$$

A Figura 4.12 mostra o diagrama em conjunto das sentenças de um texto e seus subconjuntos - sentenças que ocorrem as palavras de um determinado par.

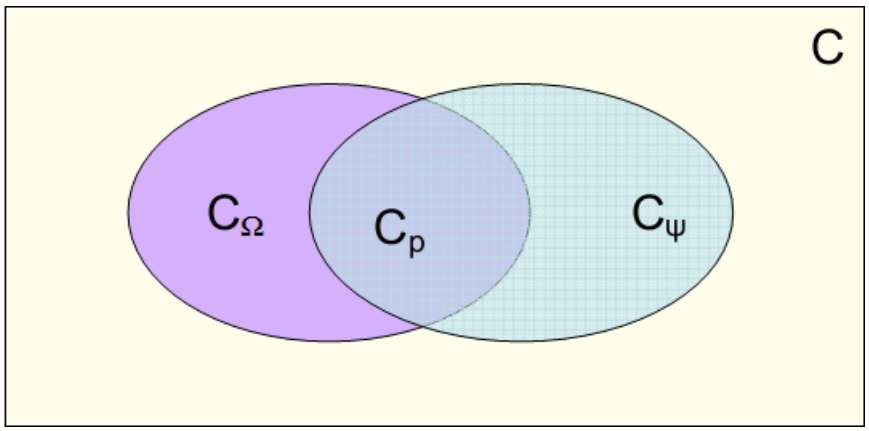


Figura 4.12: Diagrama dos conjuntos de sentenças de um texto. FONTE: (TEIXEIRA, 2007, p. 65).

Considere $S_i = |C_i|$ como sendo a cardinalidade do conjunto C_i, ou seja, o seu número de elementos. Com isso pode-se calcular e entender melhor o índice *incidência-fidelidade* a partir da Equação 4.12:

$$IF_{(\Psi,\Omega)} = \frac{S_p}{N_S} \times \frac{S_p}{S_\Psi + S_\Omega - S_p} = \frac{(S_p)^2}{N_S(S_\Psi + S_\Omega - S_p)} \tag{4.12}$$

O termo N_S na Equação 4.12 representa o número de sentenças do texto. Isto mostra

que o índice IF depende do tamanho do texto. Para os trabalhos de Teixeira (2007) e Teixeira et al. (2010) este fato não comprometeu os resultados, visto que os textos analisados diferem muito pouco entre si do número total de sentenças (Figura 4.13), o que não ocorre em redes de títulos (Figura 4.14).

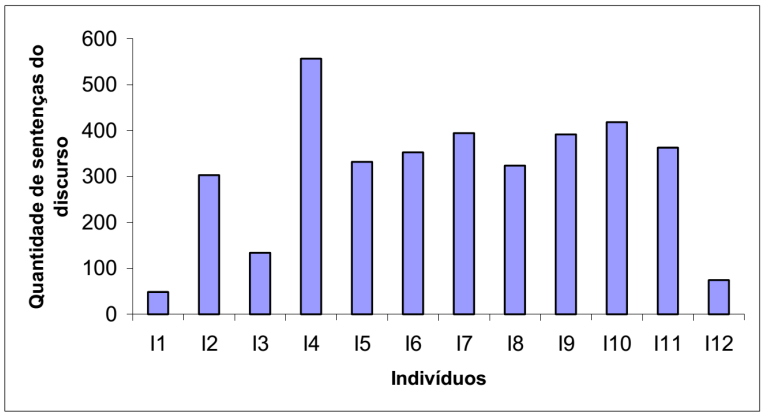


Figura 4.13: Quantidade de Sentenças pronunciadas por cada indivíduo em um discurso oral. Os dois indivíduos com menor quantidade de sentenças, I_1 e I_{12} possuem esquizofrenia. Fonte:Teixeira (2007, p.88)

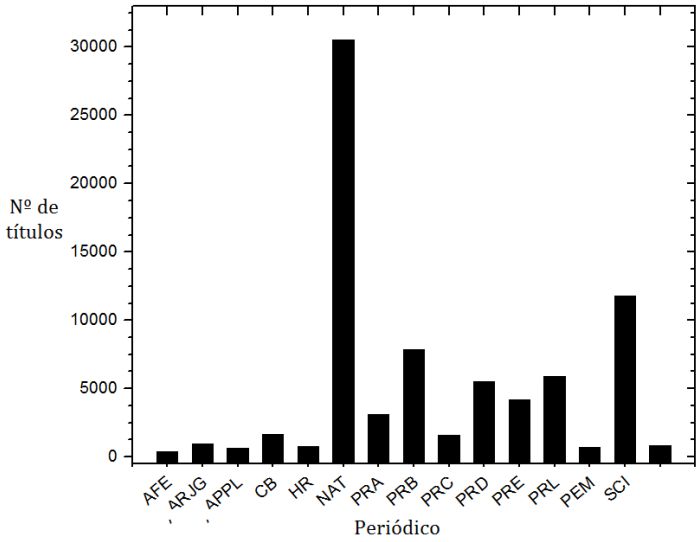


Figura 4.14: Quantidade de títulos por periódico. Os valores desta base diferem consideravelmente entre si. Diferentemente dos discursos orais da Figura 4.13, que diferem muito menos entre si.

Entretanto, Aguiar (2009) - que caracterizou autores de romances de idiomas diferentes através de seus "discursos escritos", na busca de comportamentos críticos das suas redes percebeu que o índice *incidência-fidelidade* não era adequado, pois a diferença entre os tamanhos dos textos (número de sentenças) dos romances eram consideráveis. Então foi necessário propor um ajuste no índice e o chamou de *força fidelidade normalizada*[4],

[4]Este trabalho não entende este processo como normalização e sim um reescalonamento do índice, a fim de que ele esteja sempre entre 0 e 1. Portanto, em alguns momentos do texto em que se precisar diferenciar os dois tipos de *incidência-fidelidade*, não será utilizado o termo "normalizada".

(Equação 4.13). Ao longo deste trabalho, será usado a denominação *incidência-fidelidade* (IF ao invés de IF_N.) para este índice.

$$IF_{N(\Psi,\Omega)} = (I_N) \times (F_N) = \frac{I_{(\Psi,\Omega)} - I_{Min}}{I_{Max} - I_{Min}} \times \frac{F_{(\Psi,\Omega)} - F_{Min}}{I_{Max} - I_{Min}} \tag{4.13}$$

Na Equação 4.13, I_{Max} e I_{Min} representam respectivamente o maior e o menor valor de *incidência* que ocorreu no texto. Da mesma forma, F_{Max} e F_{Min} representam respectivamente o maior e o menor valor de *fidelidade* ocorridos no referido texto. Os dois índices são essenciais para a construção de redes filtradas pelos pesos das arestas - representados pelo parâmetro IF.

A Seção 5.5 explica de que forma, a partir de uma rede semântica, obtém-se uma subrede crítica. O índice *incidência-fidelidade* atua como um filtro que elimina informação da rede original (i.e. pares de palavras com IF mais baixos que um valor limite IF_L) até chegar em uma configuração crítica. Nesta configuração, a rede possui pares de palavras muito frequentes no texto e no contexto em que apareceram, como também pares não tão frequentes, mas que são úteis para deixar o discurso consideravelmente conectado. Em outras palavras, a rede crítica de um texto possui o máximo de informação com o mínimo de ruído.

O IF proposto por Teixeira et al. (2010) é mais adequado para filtrar e comparar textos quando não existe diferença considerável de tamanhos entre eles(quantidade de sentenças). O IF proposto por Aguiar (2009) possui a mesma utilidade, independente das quantidades de sentenças que os textos possuem, já que este índice não depende do tamanho do texto.

4.5.3 Redes de discursos escritos

A rede de um discurso escrito é composta por palavras de um texto escrito por uma ou mais pessoas. Na sua grande maioria, os textos escritos são precedidos por um raciocínio prévio bem elaborado, ao contrário da maioria dos discursos orais, no qual a comunicação é mais espontânea.

O estudo da linguagem escrita numa abordagem estatística começou por volta da década de 20, 30. O trabalho de Zipf (1949) teve importância fundamental neste processo. Sua abordagem está baseada na contagem de palavras do texto e na correlação entre essa quantidade e o ranking das frequências de aparição das palavras. Com isso, ele propôs duas leis que são capazes de relacionar palavras em grupos de baixa e alta frequência de aparição no texto.

Mais tarde, em 1972, Zipf pôde demonstrar que as distribuições de frequências de palavras em um texto seguia uma lei de potência. Esta foi uma das primeiras evidências de um sistema complexo. Desde então, este método tem sido amplamente aplicado por outros pesquisadores em uma grande diversidade de textos e sistemas naturais.

Montemurro e Zanette (2002) estudaram textos literários em Inglês utilizando uma nova abordagem da mecânica estatística, diferentemente de Zipf (1972). Eles demonstram que há uma relação quantitativa entre as palavras do texto e a entropia da informação de Shannon[5], definida sobre uma distribuição de probabilidade apropriada. A partir deste estudo é possível não apenas avaliar quantas vezes uma palavra é usada, mas também agrupar certas palavras de acordo com sua função específica no texto, independente de se ter ou não conhecimento prévio da estrutura sintática da língua.

Em seu trabalho, Caldeira (2005) analisou 312 textos dos seguintes estilos: técnicos e literários; escritos por homens, mulheres e de autoria coletiva; com menos de 1000 sentenças e maiores; do idioma inglês e português. Nesse estudo, a autora faz uma comparação da rede semântica gerada com o aparelho psíquico de Freud que contém o modelo de representações-objeto, além de caracterizar a dinâmica de crescimento dos textos escritos. Suas redes apresentaram topologia livre de escala e mundo pequeno.

Aguiar (2009) analisou a linguagem verbal escrita utilizando como base de dados textos literários clássicos escritos em quatro idiomas distintos. Todas as redes de palavras dos textos originais exibiram um comportamento crítico bem definido, quando submetidas a um processo de filtragem através do índice *incidencia fidelidade*. A autora compara os índices de rede e as distribuições de conectividade das redes, entre os textos originais e seus respectivos textos aleatórios, e observa diferenças quantitativas consideráveis. Na segunda parte de seu trabalho, foi proposto um método de se calcular distâncias entre pares de texto, utilizando os índices de redes, através da *distância euclidiana* no espaço dos índices de redes. A partir desta análise, observa-se que a estrutura topológica das redes revela, de forma significativa, as diferenças entre os textos de autores distintos, apesar de não ser sensível a diferentes idiomas e conteúdos. Isto é, os textos que foram escritos pelo mesmo autor tem redes semânticas topologicamente semelhantes.

4.5.4 Redes de discursos orais

Uma forma de se adquirir uma rede de discurso oral é capturar o discurso de alguém utilizando um gravador e depois transcrever o discurso. Então, pode-se utilizar algum método de criação de redes (e.g. rede semântica de linha, clique ou círculo) para gerar a rede de palavras do discurso. Teixeira et al. (2010) construiram redes semânticas a

[5]Esta entropia tem o intuito de quantificar a "informação".

partir do modelo de cliques. Neste trabalho, eles entrevistaram doze indivíduos, escolhidos aleatoriamente entre dois grupos de estudantes de cursos universitários distintos. Os entrevistados foram estimulados a partir de um "Prime" (i.e. tema central de seus discursos).

Neste trabalho, os autores caracterizaram as relações existentes entre as palavras que emergiam durante os discursos, em um método que utiliza conceitos e propriedades de redes complexas e da teoria de conjuntos, para identificar a rede que melhor representa a estrutura de associação entre os conceitos do discurso. Para tal, elaborou-se um conceito de força associativa e força fidelidade, que representa a probabilidade da ocorrência dos pares das palavras na mesma sentença no discurso inteiro. Estes conceitos foram utilizados para construir a rede de associações semânticas de cada discurso. As redes geradas apresentam comportamentos típicos de redes livres de escala e mundo pequeno.

A partir do conceito *incidência-fidelidade* foi possível encontrar uma rede crítica que possui um comportamento típico de mudança de fase e, curiosamente com topologias similares, indicando a possibilidade de que tal comportamento e topologia crítica sejam características intrínsecas do mecanismo da linguagem humana. Segundo a autora, as redes geradas apresentam padrões de associações promissores no estudo de novos métodos psicométricos para estudos de acesso à memória. A nossa capacidade de associação simbólica (que dá origem a linguagem) exibe um comportamento que é representado pela emergência de uma rede de associações simbólicas, as vezes com características topológicas bem definidas.

4.5.5 Redes baseadas em títulos de artigos científicos

O título de um artigo científico transmite ao leitor as principais ideias do trabalho, ou seja, carrega consigo um significado. Este significado emerge através das palavras que o compõe, na tentativa dos autores de passar para o leitor a(s) principal(is) contribuição(ões) do trabalho. Um periódico científico dispõe de um conjunto de títulos que, se dispostos em forma de rede semântica, pode nos revelar as principais contribuições do referido periódico para a comunidade científica e como elas estão relacionadas entre si. De maneira análoga às redes de discurso oral ou escrito, este emaranhado de palavras pode ser capaz de nos revelar aspectos semânticos intrínsecos de uma revista, já que esta se constitui como um sistema de comunicação científica.

Em teoria, cada artigo científico de uma revista contempla um conjunto de ideias comuns aos pesquisadores que o escreveram. Essas ideias são fundamentais para aumentar a fronteira do conhecimento humano. O título, então, funciona como um chamariz para a leitura do trabalho por outros pesquisadores que anseiam em agregar novos conhecimentos,

bem como usar destes conhecimentos para produzir mais ideias que podem ser alocadas na mesma revista em que se leram outros artigos.

As palavras que compõe um título provavelmente são cuidadosamente selecionadas pelos cientistas que publicam um determinado trabalho. Algumas delas são palavras chamadas neste estudo de *curingas*, ou seja, que aparecem com muita freqüência em títulos de trabalhos de diversas temáticas e assim não representam conceitos capazes de diferenciar um conjunto de artigos de outro. Por exemplo, as palavras "research", "be"[6], "study´´, etc. aparecem frequentemente em títulos de trabalhos de temáticas completamente diferentes.

Entretanto, existem palavras que são muito comuns em um determinado grupo de artigos científicos, mas não tão comuns em outro grupo de artigos de um mesmo periódico. Por exemplo, as palavras "gene", "dna", "human", "United States", "cell", "biology", "physics", "health", "protein", etc. são muito freqüentes em determinados grupos de artigos da revista Nature, em que cada grupo trata de uma temática específica. Isto por que cada palavra deste tipo agrupa títulos de temáticas parecidas às dela, já que não se trata de uma palavra *curinga*.

O uso das redes complexas é útil para a identificação destas palavras, suas frequências de ocorrência e as épocas em que cada uma se torna mais ou menos relacionada com outras palavras de mesma temática ou de temáticas diferentes. Algumas delas são fundamentais para manter a estrutura de conexão da rede semântica associada à ela, mesmo que tenha pouca frequência de ocorrência nos títulos. Este tipo de rede semântica, formada por títulos de artigos de periódicos científicos, foi há pouco tempo estudada pelos autores Fadigas et al. (2009), Pereira et al. (2011). Parece ser que esses autores são os pioneiros neste tipo de estudo.

No primeiro trabalho sobre redes de títulos, Fadigas e colaboradores discutem sobre a divulgação científica em educação matemática a partir de um diagnóstico quantitativo e qualitativo fundamentado em *redes sociais e complexas*. Como resultado, a pesquisa mostrou que é possível agrupar as redes semânticas de palavras usadas nos títulos dos periódicos em basicamente dois grupos distintos, se a análise for fundamentada em índices de Redes Complexas. Se a análise for baseada em índices de Redes Sociais, outros dois grupos de revistas são observados, porém de composições diferentes dos dois primeiros. Com isso, ele contribui ao oferecer suporte para definição de estratégias que captem mais leitores, ajudando no processo de difusão do conhecimento em campos específicos.

No segundo trabalho, os autores procederam da mesma forma, com adição de revistas

[6]Este verbo (em português: ser/estar) encontra-se em nossa base de dados na sua forma canônica "be", que é variação dele em formas flexionadas, como "está", "ser", "foi", "é" etc. Esta transformação é necessária para algumas palavras e faz parte do tratamento aplicado, a fim de extrair o sentido semântico mais adequado de cada palavra. Este processo é explicado na Metodologia, Seção 5.2.2

internacionais[7]. O autor caracteriza grupos diferentes de periódicos a partir da densidade e do caminho mínimo e compara os resultados com periódicos em português, do trabalho de Fadigas et al. (2009). Os resultados mostram que as redes têm topologias semelhantes, independente da língua em que se encontra o periódico. Estas topologias são basicamente *livre de escala* e *mundo pequeno*. Assim, os resultados dos principais índices de Redes Complexas utilizados nesse trabalho (i.e. densidade, distribuição de graus e caminho mínimo médio) e os modelos de rede provocam reflexões importantes na contribuição para a difusão do conhecimento.

A análise dos índices de redes complexas permite caracterizar, diferenciar os periódicos e dar um caminho para o entendimento de como funciona a difusão do conhecimento das publicações da revista. Figura 5.3 apresenta um exemplo da rede formada com um conjunto de títulos da revista Nature. Esta configuração foi obtida com apenas 300 títulos da revista e é também resultado de um processo de filtragem, obtido pelo índice *incidência-fidelidade*. Ela exibe um comportamento crítico de mudança de fase[8].

Diante o escopo teórico, exposto na parte II, as próximas seções se dedicam à metodologia aplicada nesta pesquisa.

[7]Estas revistas são as mesmas utilizadas neste trabalho e as informações sobre elas e o motivo de escolhê-las podem ser vistos na Metodologia, Seção 5.1.

[8]Mais detalhes sobre esse processo, na Metodologia, Seção 4.5.2.

Parte III

Trabalho teórico e prático

Capítulo Cinco

Metodologia da pesquisa

Esta seção apresenta a metodologia do presente estudo. A Figura 5.1 sintetiza, passo a passo, a metodologia empregada nesta pesquisa. As seções subsequentes detalham cada etapa da metodologia ilustrada pela Figura 5.1.

O problema de pesquisa, os objetivos e limitações foram descritos especificamentes no Capítulo 1.1 (Introdução), nas Seções 1.4, 1.6 e 1.7. Em síntese, a questão que norteia esta pesquisa é:

Os periódicos científicos permitem uma classificação a partir dos títulos de seus artigos?

5.1 Aquisição dos dados

Os títulos dos artigos científicos foram obtidos em periódicos de circulação internacional (idioma em inglês), das mais abrangentes áreas do conhecimento. A Tabela 5.1 mostra informações básicas sobre os periódicos usados na pesquisa.

A base de dados desta pesquisa é a mesma utilizada por Pereira et al. (2011) - são os periódicos: *Agricultural and Forest Entomology* (AFE); *Antipode: A Radical Journal of Geography* (ARJG); *Applied Psycholinguistics: Psychological and Linguistic studies Across Languages and Learning* (APPL); *Chemistry and Biology* (CB); *Human Relations: Towards the integration of the Social Sciences*(HR); *Nature* (NAT); *Physical Review A* (PRA); *Physical Review B* (PRB); *Physical Review C* (PRC); *Physical Review D* (PRD); *Physical Review E* (PRE); *Physical Review Letter* (PRL); *Probabilistic Engineering Mechanics* (PEM); *Science* (SCI); *Sociology of Health and Illness* (SHI). No estudo supracitado, os autores ressaltam que os critérios utilizados para selecionar as revistas científicas publicadas em inglês são:

- Fator de impacto maior do que 1;
- As revistas devem estar disponíveis na internet;
- Cada revista deve representar da melhor forma possível uma área do conhecimento, incluindo campos interdisciplinares.

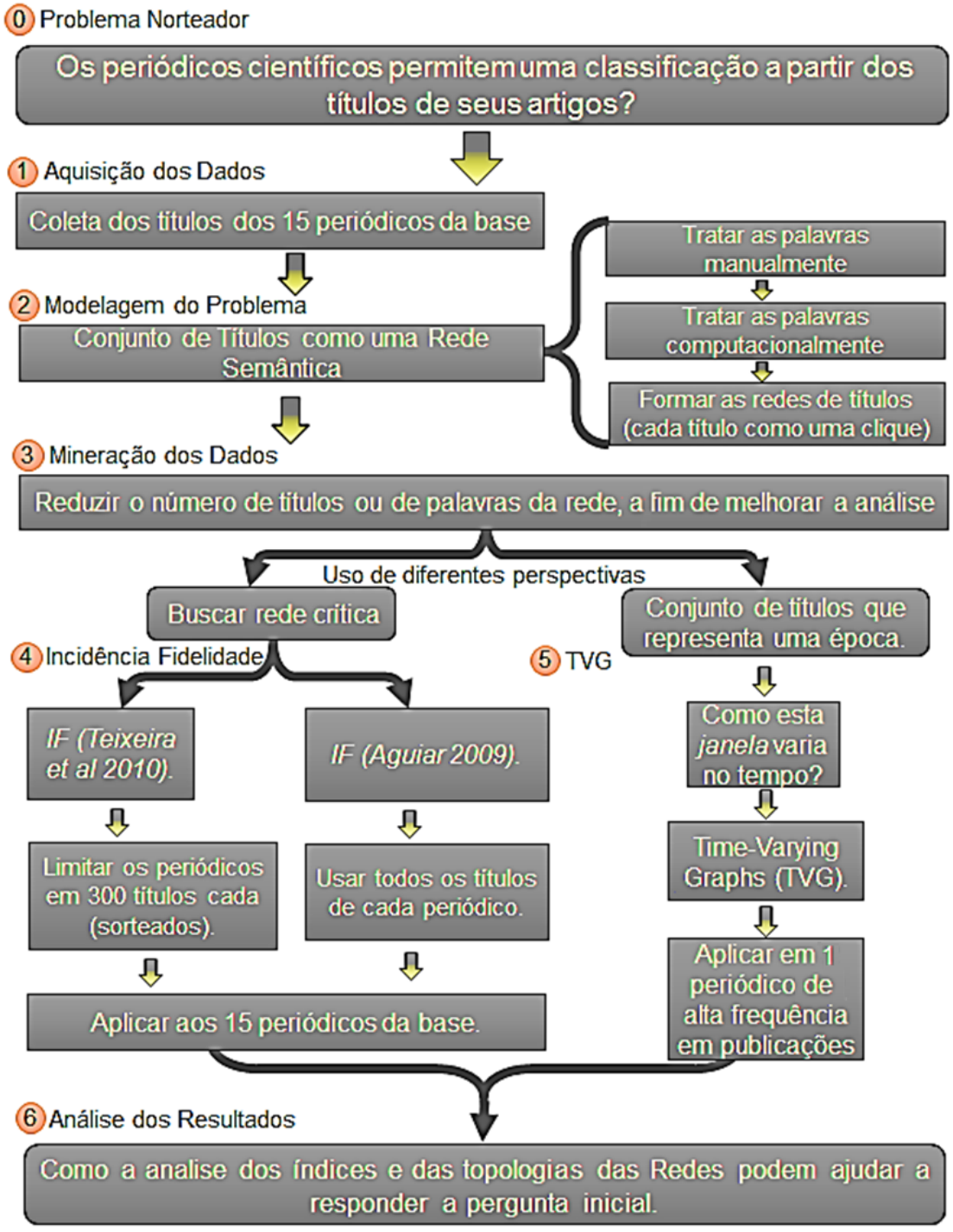


Figura 5.1: Aspectos gerais da metodologia da pesquisa.

5.2 Modelagem do problema

O título de um artigo científico transmite ao leitor a ideia principal do trabalho. Um conjunto de títulos pode ser visto como um discurso escrito, sendo cada título uma sentença do discurso. Partindo desta premisa, um conjunto de títulos de um determinado periódico pode ser visto como o "discurso" da revista durante o intervalo de tempo em

Periódico	*Nome nesta Pesquisa*	*Frequência de Publicações*	*Período coletado*	*Número de títulos*
Agricultural Forest Entomology	AFE	Anual	1999 a 2008	371
Antipode: A Radical Journal of Geography	ARJG	5 por ano	1969 a 2008	971
Applied Psycholinguistics	APPL	Trimestral	1980 a 2009	658
Chemistry & Biology	CB	Mensal	1994 a 2008	1643
Human Relations	HR	Anual	1997 a 2008	738
Nature	NAT	Semanal	1999 a 2008	30490
Physical Review A	PRA	Mensal	2007 e 2008	3089
Physical Review B	PRB	Mensal	2007 e 2008	7847
Physical Review C	PRC	Mensal	2007 e 2008	1572
Physical Review D	PRD	Mensal	2007 e 2008	5527
Physical Review E	PRE	Mensal	2007 e 2008	4157
Physical Review Letters	PRL	Mensal	2007 e 2008	5929
Probabilistic Engineering Mechanicis	PEM	Anual	1986 a 2009	703
Science	SCI	Semanal	1999 a 2008	11798
Sociology Health and Illness	SHI	Bimestral	1979 a 2008	845

Tabela 5.1: Principais informações sobre os dados adotados nesta pesquisa.

que os trabalhos que correspondem a estes títulos foram publicados.

Assim, as topologias das redes destes conjuntos de títulos podem ser capazes de diferenciar um periódico de outro ou duas épocas distintas de um mesmo periódico. Para este fim, será utilizada a modelagem a partir de redes semânticas de cliques (Seção 4.5.1). O processo de formação destas redes envolveram três etapas, resumidas na Figura 5.1 e detalhadas nas seções que seguem.

5.2.1 Tratamento manual das palavras

Este procedimento consiste em tratar as palavras dos títulos de acordo com as necessidades do pesquisador. Este tratamento ainda é impossível de ser automatizado e então tem sido feito manualmente a cada conjuntode títulos que é incorporada a base de dados. Fadigas et al. (2009) e Pereira et al. (2011) proporam algumas regras para o tratamento manual de palavras, de forma que as mesmas não percam seus significados quando estiverem nas redes.

O termo "Rio de Janeiro", por exemplo possuem muitas possibilidades de conexões, caso sejam dispostas separadamente em uma rede. As palavras "Rio" e "Janeiro" podem existir em outros títulos e não faz sentido conectá-las com as palavras que pertencem aos

mesmos títulos da palavra "Rio de Janeiro". Sendo assim, a melhor configuração para este termo em uma rede é "Riodejaneiro". O mesmo vale para "United States" que é convertido para "unitedstates". O procedimento foi o mesmo para todos os títulos dos 15 periódicos da base.

Antes de construir a rede semântica associada aos títulos de seus artigos, foi necessário construir um discurso escrito baseado nos conceitos existentes dentro destes títulos. Esta etapa consiste na organização dos títulos em um arquivo de texto, de formato *.txt*, em que cada linha contem um único título, como sentenças de um discurso escrito. A partir daí foi necessário realizar o *tratamento manual* das palavras. Modificou-se cada uma, se necessário, de acordo com as regras propostas por Fadigas et al. (2009) e Pereira et al. (2011), como apresentado no resumo abaixo:

1. Títulos em outra língua foram traduzidos para a linguagem de análise (e.g., um artigo publicado em uma revista cuja principal língua é o Português, com um título em outro idioma devem ser traduzidos para o Português);
2. Nomes próprios, palavras compostas e sequências de palavras que têm um significado próprio devem formar uma única palavra (e.g. Albert Einstein, Difusão do Conhecimento, e-mail e Educação Matemática devem ser convertidos para alberteinstein, difusaodoconhecimento email e educaçãomatematica, respectivamente);
3. Números cardinais e números ordinais foram escritos na forma textual, por exemplo: primeiro, décimosegundo ao invés de 1º, 12º;
4. Palavras repetidas no mesmo título são excluídas,restando apenas uma ocorrência;
5. Sinais gráficos, como o travessão, ponto e vírgula, ponto de interrogação, ponto de exclamação e reticências foram eliminados;
6. Palavras incorretamente grafadas, foram corrigidas;
7. Linguagem especializada foi mantida o máximo possível.

5.2.2 Tratamento computacional das palavras

Nesta etapa da modelagem, o texto foi submetido a um tratamento computacional[1] que classifica, modifica e elimina palavras quando necessário. A etapa de classificação das palavras distingue palavras *lexicais* de palavras *gramaticais*. De acordo com Caldeira (2005) e Martins (2008):

[1] O detalhamento completo dos softwares utilizados nesta etapa pode ser encontrado em Caldeira (2005).

- As palavras lexicais têm significação por si mesmas. Elas despertam a representação de algo que está fora da língua e que faz parte do mundo físico, psíquico ou social. São os substantivos, os adjetivos, os advérbios a eles derivados, os verbos que exprimem ação (os verbos auxiliares e os de ligação são palavras *gramaticais*);
- As palavras gramaticais são também chamadas palavras vazias, instrumentos gramaticais ou não-palavras. Sua função é relacionada com a organização do texto, a estruturação da frase e o ato de enunciação. São as preposições, os pronomes, os artigos, numerais, advérbios, conjunções e interjeições.

A classificação foi feita com a utilização do conjunto de pacotes do software livre *UNITEX*[2]. A etapa final do tratamento foi realizada com o programa Ambisin que foi desenvolvido por Caldeira (2005).

- O pacote *UNITEX* (PAUMIER, 2002) é um software livre que contém recursos linguísticos como dicionários eletrônicos e tábuas *léxico-gramaticais*. Ele trata textos de idiomas distintos (e.g. inglês, francês, português, grego, italiano, russo, espanhol e tailandês) e é capaz de classificar as palavras em *lexicais* e *gramaticais*;
- O programa *Ambisin* é utilizado para eliminar as ambiguidades, eliminar palavras *gramaticais* e converter todos os verbos para sua forma canônica (e.g. initiating → initiate), ou seja no infinitivo. As palavras que não são encontradas no dicionário não são eliminadas.

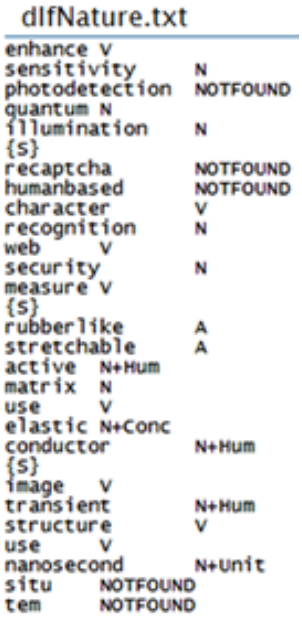
As palavras classificadas pelo *UNITEX* de acordo com seu significado léxico (i.e. pronome, advérbio, adjetivo e substantivo) são mantidas. As palavras *gramaticais* como artigos e preposições, que quando isoladas não possuem significado semântico relevante para a construção das redes, são eliminadas. Os verbos foram reduzidos à sua forma canônica e, por conterem significado léxico, também são mantidos.

5.2.3 Construção das Redes de Títulos

Para um conjunto de títulos de um dado periódico, o produto final dos tratamentos, manual e computacional supracitados, é um arquivo de texto na forma *dlfnomedoperiodico.txt*, como exemplificado na Figura 5.2.

Como se pode ver no exemplo da Figura 5.2, existe um marcador { S } que indica o fim de um título e início de outro. Ao lado das palavras, já tratadas e contidas em

[2]Disponibilizado pela Rede Relex Brasil. Disponível em: http://ladl.univ-mlv.fr/brasil/Ferramentas/Unitex.html.


dlfNature.txt

```
enhance V
sensitivity     N
photodetection  NOTFOUND
quantum N
illumination    N
{S}
recaptcha       NOTFOUND
humanbased      NOTFOUND
character       V
recognition     N
web     V
security        N
measure V
{S}
rubberlike      A
stretchable     A
active  N+Hum
matrix  N
use     V
elastic N+Conc
conductor       N+Hum
{S}
image   V
transient       N+Hum
structure       V
use     V
nanosecond      N+Unit
situ    NOTFOUND
tem     NOTFOUND
```

Figura 5.2: Excerto do arquivo dlfNature.txt, que contém títulos já tratados, da revista Nature.

cada agrupamento entre os marcadores, estão as funções semânticas de cada uma delas, i.e. Nome ou Substantivo(N), Verbo (V), Adjetivo (A), Advérbio (ADV) e Palavras não encontradas no dicionário (NotFound) (PAUMIER, 2002). Em relação à esses marcadores de sentenças, Caldeira (2005) complementa:

> "*Este é um requisito metodológico do nosso estudo da rede dos textos escritos, pois entendemos que a sentença é a menor unidade para análise dos significados expressos nos textos. Cada palavra isoladamente pode adquirir um significado que somente será identificado a partir do contexto. Esse contexto, para a nossa pesquisa é a sentença em que a palavra participa* (CALDEIRA, 2005, p. 67). "

No contexto deste trabalho, o título de um artigo científico é visto como uma sentença de um discurso escrito e as palavras de cada título como vértices de uma Clique[3]. Consequentemente, a dinâmica de construção de uma rede de títulos foi dada por justaposição e/ou sobreposição de cliques. De acordo com Fadigas e Pereira (2013), designa-se de justaposição o processo no qual duas cliques são ligadas por apenas um vértice comum. Quando a ligação ocorre com dois ou mais vértices comuns, chama-se o processo de sobreposição.

Observe a construção da rede na Figura 5.3. Na figura, tem-se a ilustração da primeira página de dois artigos de um mesmo periódico, em destaque seus títulos. Todas as palavras de um mesmo título são interligadas, formando uma clique. Contudo, títulos diferentes podem conter palavras iguais ou palavras de mesma forma canônica. Neste caso, as cliques foram unidas, justapondo a palavra comum. Este procedimento, estendido a todos os

[3]Clique é a denominação dada à uma rede ou sub-rede que possui todos os seus vértices conectados entre si. Em nossa análise, as palavras de cada título formam uma clique.

títulos de uma revista gerou a Rede Semântica de seus títulos.

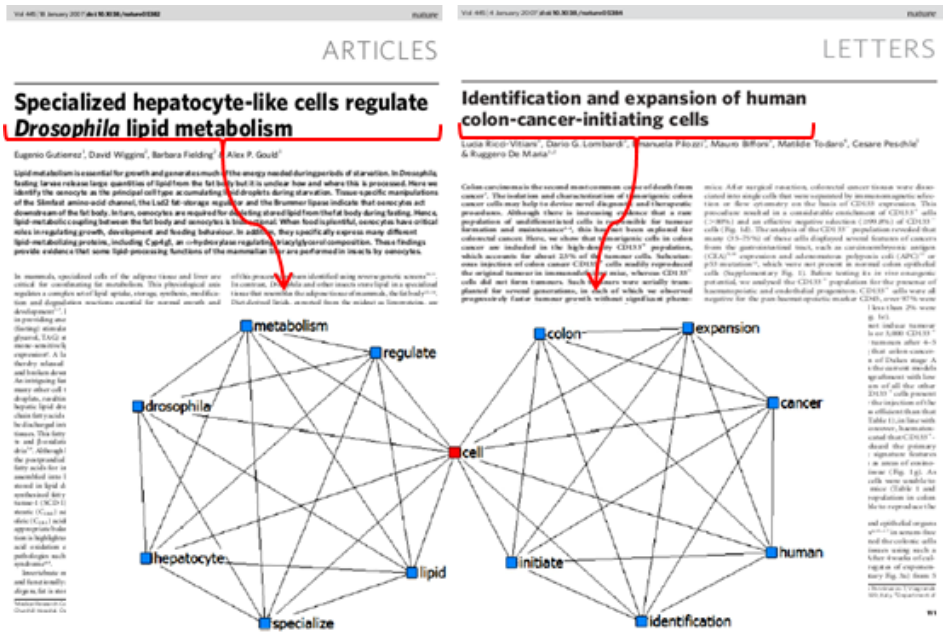


Figura 5.3: Primeira página de dois artigos da Nature. Em destaque, seus títulos e o processo de construção da rede. Adaptado de Pereira et al. (2011).

Portanto, cada arquivo de texto (e.g. *dlfScience.txt*), que contém um conjunto de títulos de um dado periódico foi transformado em uma rede de cliques. Este procedimento foi realizado pelo programa *NetPal*[4]. A visualização das redes foi feita pelos softwares livres *Pajek*[5] e *Gephi*[6]

5.3 Mineração dos Dados

Não obstante o tratamento computacional eliminar palavras gramaticais, as redes ainda continuam com muitos vértices. Muitos deles não são tão importantes para o processo de diferenciar um periódico de outro a partir de seus títulos. Existe uma grande quantidade de vértices que podem atrapalhar a análise da rede de um periódico. Para contornar esta limitação e enriquecer a análise das redes, dois caminhos foram seguidos:

- A busca de uma rede ótima, em cada periódico, assim como nos trabalhos de Aguiar (2009) e Teixeira et al. (2010);

[4]Mais informações em Caldeira (2005).
[5]http://pajek.imfm.si/doku.php, último acesso em 31 de agosto de 2013 às 21:13.
[6]https://gephi.org/, último acesso em 31 de agosto de 2013 às 21:06.

- Estipulou-se um período temporal (janela de tamanho fixo), que contendo um conjunto de títulos dispostos na mesma ordem em que foram publicados. Esta janela avançou no tempo. A investigação consistiu em observar o comportamento das redes na janela em cada época.

O primeiro ítem sugere a busca de uma rede ótima, também chamada de rede crítica. Esta rede é formada tanto por pares de palavras extremamente fiéis quanto por pares razoavelmente fiéis, no que concerne ao título em que estão inseridos. Estes pares são representados por arestas ponderadas pelo índice *incidência-fidelidade*. Este índice foi proposto inicialmente por Teixeira (2007). Posteriormente, ele foi modificado por Aguiar (2009) para que seus valores não dependesse do tamanho do texto[7], o índice então foi reescalonado para apresentar valores entre 0 e 1. Para a rede de títulos, no presente estudo, foi verificado a existência de pontos críticos e construiram-se as redes críticas dos periódicos a partir da *incidência-fidelidade* em suas duas formas matemáticas propostas pelos trabalhos supracitados.

O segundo item sugere que, para as revistas Science e Nature[8], uma janela temporal de 8 semanas seja analisada à medida que se avança no tempo (dado em semanas). Este método possui sua base teórica em *Grafos que variam no tempo* (*TVG*), proposto por Casteigts et al. (2011).

A descrição completa dos aspectos metodológicos para estes dois caminhos estão nas seções que seguem.

5.4 *O uso da* incidência fidelidade *na Construção das redes de títulos*

Independente do índice *incidência-fidelidade* utilizado, reescalonado ou não, o procedimento abaixo foi realizado para todos os 15 periódicos da base, Seção 5.1. Ademais, nas Seções 5.5.1 e 5.5.2, haverá uma descrição da metodologia específica para rede de títulos em cada uma das duas abordagens do *incidência-fidelidade*.

De acordo com a Figura 5.4, o arquivo de entrada é um arquivo de texto, que no exemplo é o "nature.txt". Este arquivo contém o conjunto de títulos da revista, em que cada linha contem um único título, tratados manualmente, ou seja, de acordo com as regras propostas

[7] Vale lembrar que no trabalho de Teixeira et al. (2010) os textos estudados possuíam o quantidades de sentenças próximas umas das outras.

[8] A escolha destas revistas deu-se pela grande quantidade de títulos que possuem na base de dados apresentada aqui. Como também, por possuírem mesma frequência temporal de publicação (1 revista por semana), o que possibilita uma comparação significativa entre as duas.

por Fadigas et al. (2009) e Pereira et al. (2011), mostradas na Seção 5.2.1. O primeiro processo foi realizado pelo *UNITEX*, que fez a classificação gramatical de todas as palavras que contém no texto. Logo após, o Ambisin eliminou palavras gramaticais, minimizou efeitos de ambiguidades e separa as formas canônicas ou flexionadas das palavras do restante dos ítens de classificação gramatical gerados pelo *UNITEX*.

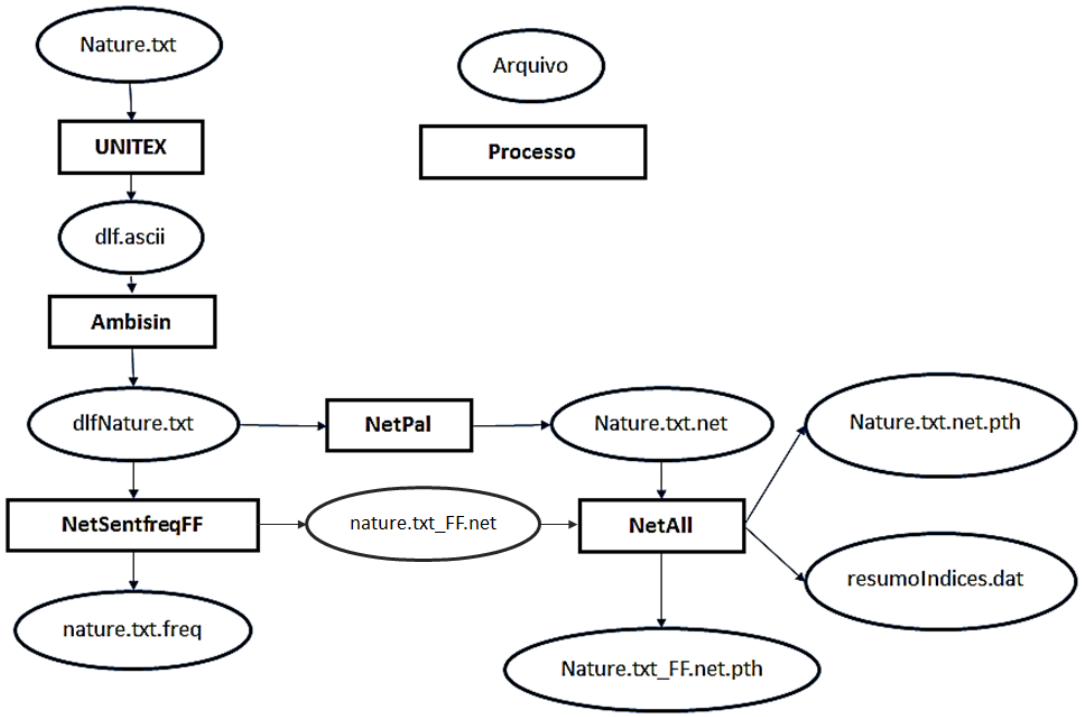


Figura 5.4: Método para gerar arquivos de frequências de palavras, redes e seus respectivos índices para cada valor de IF. Adaptado de: Teixeira (2007, p. 81)

Após isto, o *NetsentFreqFF* utilizou o arquivo dfl.txt como entrada de dados e calculou a *incidência-fidelidade* dos pares de palavras do texto analisado (i.e. arquivo .freq) e a partir destes valores o *NetPal* gerou as redes (arquivos .net). O *NetAl*[9], então calculou os índices de redes complexas para cada rede e os armazenou na forma de tabela em arquivos do tipo "ResumoIndices.dat". Este arquivo pode ser aberto por algum editor de planilhas e assim os gráficos dos índices podem ser gerados.

Nas próximas seções serão apresentados procedimentos metodológicos, a partir do uso dos métodos de Aguiar (2009) e Teixeira et al. (2010), para o cálculo do índice *incidência-fidelidade*.

[9]Maiores informações ver em Caldeira (2005).

5.4.1 *Construção das redes*

Nos trabalhos citados no Capítulo 4, nas Seções 4.5.3 e 4.5.4, as redes dos discursos escritos e orais foram construídas através do processo de filtragem, obtido através do índice *incidência-fidelidade*. Por conta deste processo, uma rede filtrada apresentaria apenas pares de vértices, cujas arestas ponderadas, representam valores de *incidência-fidelidade* igual ou maior que um certo valor de corte IF_L[10].

Por exemplo, se o valor de corte determinado pelo pesquisador for $IF_L = 2 \cdot 10^{-4}$, então a rede gerada apresentará arestas cujos valores de IF são maiores ou iguais a $2 \cdot 10^{-4}$. Os pares de vértices (arestas) que não atenderem este critério serão descartados da rede. Quando uma aresta é descartada da rede, não necessariamente os vértices associados a ela também são, já que estes podem estar conectados à outros vértices, formando arestas com IF maiores que o valor de corte. Entretanto, se após a remoção de uma aresta, os vértices associados à ela não estiverem conectados à outros vértices ou se estiverem associados a outras arestas com $IF < IF_L$, estes também serão descartados.

Curiosamente, para os trabalhos supracitados, existe um valor crítico de *incidência-fidelidade* em que a rede, juntamente com o valor de seus índices[11], representa um comportamento típico de mudança de fase, com um ponto crítico bem definido. A rede gerada para esse nível de filtragem é chamada de *rede crítica*.

5.5 *Rede crítica*

Redes críticas foram utilizadas para investigar mecanismos inerentes à linguagem humana, tanto em discursos orais, e.g. estudantes universitários (TEIXEIRA et al., 2010), quanto em discursos escritos, e.g. autores de romances Aguiar (2009). Neste trabalho faremos o mesmo para rede de títulos.

A Figura 5.5 mostra um exemplo de diferentes redes para um mesmo discurso oral, obtido por um indivíduo que narrou aspéctos de sua vida durante 01 hora. Sua narrativa foi gravada, transcrita e tratada manualmente e computacionalmente. Logo depois, foi aplicado o processo de filtragem da sua rede de palavras (TEIXEIRA et al., 2010).

Percebe-se que existe um valor crítico para o ponto de corte da *incidência-fidelidade* nesta rede: $IF_C = 1 \cdot 10^{-3}$. Isso mostra que qualquer aumento no valor de corte da *incidência-fidelidade*, a rede se reduz a poucos pares. Esses pares são de fato os mais fieis no discurso, por serem os mais frequentes no texto e no contexto de palavras em que foram evocados.

[10] Também chamado de valor "limite".
[11] Índices de redes complexas, explicados no Capítulo 3, na Seção 3.4.

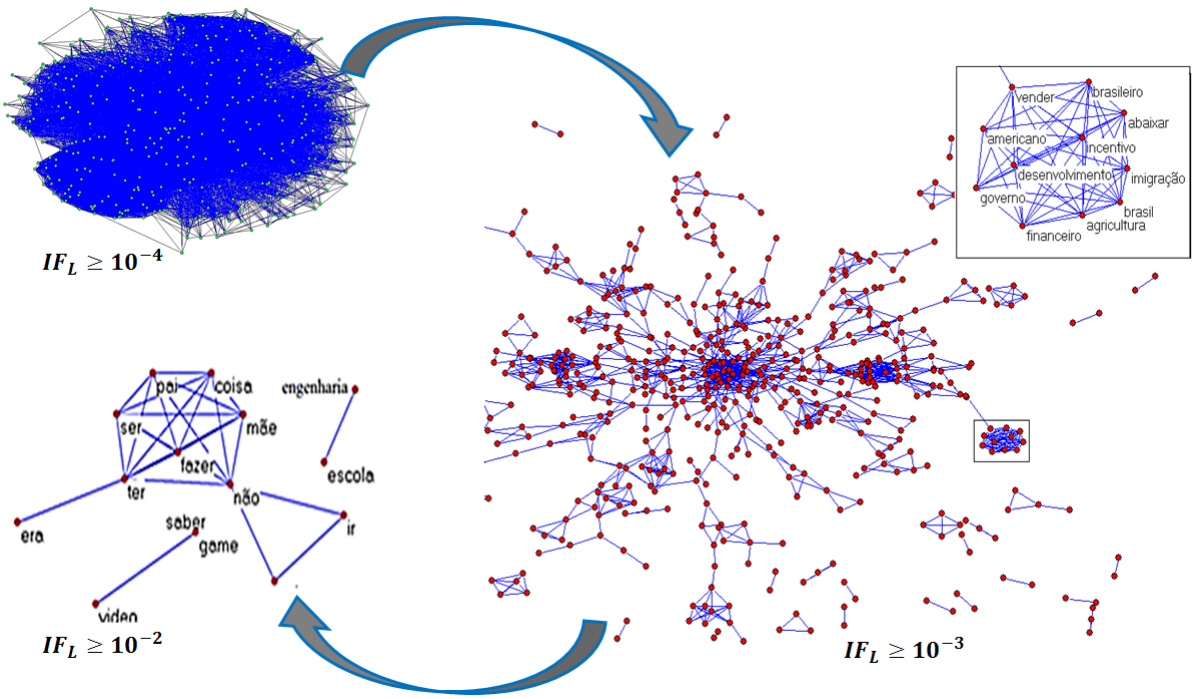


Figura 5.5: Redes de um mesmo discurso oral, para diferentes valores de incidência-fidelidade, inclusive na rede crítica, que na figra apresenta-se na forma de um discurso categorizado. Na figura, é destacado um módulo da rede crítica. Este módulo é composto por palavras de mesma temática. Adaptado de: Teixeira et al. (2010, p.340).

Por outro lado, a rede crítica nos dá mais informações sobre o discurso, já que os pares de palavras de peso insignificante ($IF < IF_C = 1 \cdot 10^{-3}$) foram descartados. Neste valor ($IF_C = 1 \cdot 10^{-3}$), vê-se na rede os pares de palavras mais fieis e também pares de palavras de peso intermediário conectados, em uma estrutura modular. Em geral, cada módulo refere-se a um tema específico do discurso deste indivíduo. A partir da rede crítica é possivel verificar os caminhos, através de palavras, que conectam uma temática específica à outra similar ou até mesmo diferente.

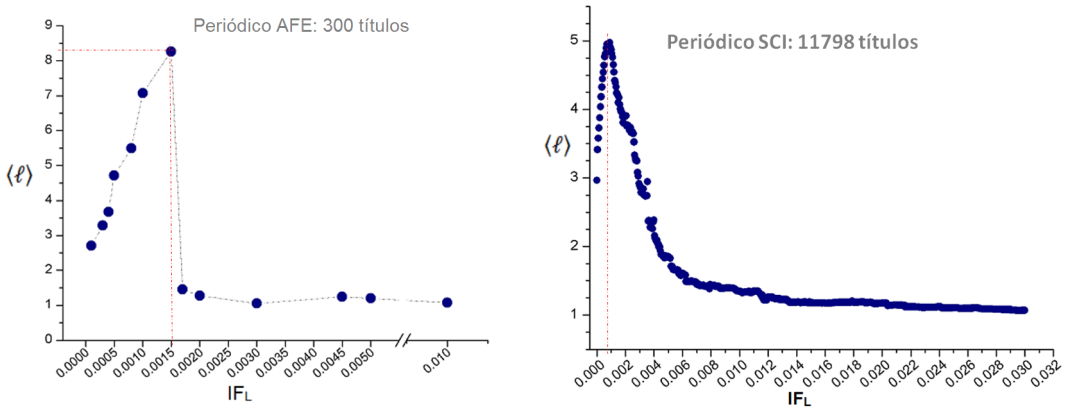
No presente trabalho buscou-se encontrar a existência de uma rede crítica para a rede de palavras baseada em títulos de artigos científicos. A seguir será mostrado o método detalhado para cada abordagem do índice *incidência-fidelidade* e como analisar o problema a partir da ótica proposta por Aguiar (2009) e Teixeira et al. (2010), só que para rede de títulos.

Para encontrar o valor da *incidência-fidelidade* que representa a rede crítica do discurso de um periódico, optou-se por usar o índice *caminho mínimo médio*. A análise consiste em verificar o que acontece com o valor do $\langle \ell \rangle$ a medida que se aumenta o índice *incidência-fidelidade*.

Além desse critério foi verificado o critério da diferença entre vértices e arestas, que

também foi usado nos trabalhos de Aguiar (2009) e Teixeira et al. (2010). Entretanto a análise do $\langle \ell \rangle$ foi mais significativa tanto do ponto de vista quantitativo preciso, quanto do ponto de vista qualitativo. Por esta razão que será usado o critério da $\langle \ell \rangle_{MAXIMO}$ seguido de uma queda brusca do mesmo para determinação da região crítica. Vale ressaltar que o número de *pares não conectados* e o valor do *diâmetro* também são máximos, no ponto crítico, ou no entorno dele.

Como foi visto na Seção 4.5.2, o índice IF atua como um filtro que vai limpando o texto a medida que seu valor cresce. Assim, a medida que se limpa o texto, o valor de $\langle \ell \rangle$ aumenta - já que a rede perde atalhos[12] entre os vértices - até um valor máximo. Um pequeno incremento no valor de IF, faz a rede se quebrar em subredes menores, de forma que a rede toda volta a ter *caminho mínimo médio* baixo, restando apenas pares de palavras importantes, contudo mal conectados, descaracterizando o "discurso da revista"enquanto rede de palavras. Ao observar a Figura 5.6, este raciocínio pode ficar mais claro.



(a) Caminho mínimo médio da rede de títulos da revista AFE em função do índice *incidência-fidelidade limite* proposto por Teixeira et al. (2010).

(b) *Caminho mínimo médio* da rede de títulos da revista Science em função de IF_L proposto por Aguiar (2009).

Figura 5.6: Localização do ponto crítico por duas abordagens metodológicas diferentes do índice *incidência-fidelidade*.

Os gráficos mostram claramente o ponto crítico para as revistas exemplificadas. Este ponto é entendido onde o Caminho Mínimo Médio é máximo. **Antes desse ponto** existe muita informação na rede de palavras, inclusive pares de palavras com pouca relevância para o discurso, que apresentam valores incidência fidelidade muito baixos. A rede exibe um aspecto muito denso (eg. Figura 5.7(a)), com $\langle \ell \rangle$ muito baixo. Assim, ao aumentar o valor de IF_L a rede vai perdendo atalhos (i.e. com a eliminação de pares de palavras, arestas são removidas) e o valor de $\langle \ell \rangle$ vai aumentando, até que **no ponto crítico**, quando $IF_L = IF_C$, o discurso está bem representado e os pares de palavras mais relevantes para

[12] Atalho é um conceito usado em redes de mundo pequeno. Um atalho é uma aresta que quando retirada faz com que o menor caminho entre dois vértices que estavam conectados previamente, tenha caminho maior que 2

quem produziu o discurso estão bem conectados na rede juntamente com pares de menor importância e a rede fica bem conectada. Apesar de exibir topologia de *small world*, tem características de *scall-free*(e.g. Figura 5.7(c)) e assim torna o caminho mínimo alto nessa região. **A partir deste ponto**, um pequeno aumento em IF_L faz o $\langle \ell \rangle$ cair bruscamente. A medida que se aumenta o valor da IF_L o valor do $\langle \ell \rangle$ continua baixo, oscilando muito pouco. Então para valores muito altos de IF_L só resta na rede os pares de palavras mais relevantes para o periódico, pois se repetiram muitas vezes nos títulos de artigos. Entretanto não configura mais um discurso escrito (ver exemplo na figura 5.7(b))

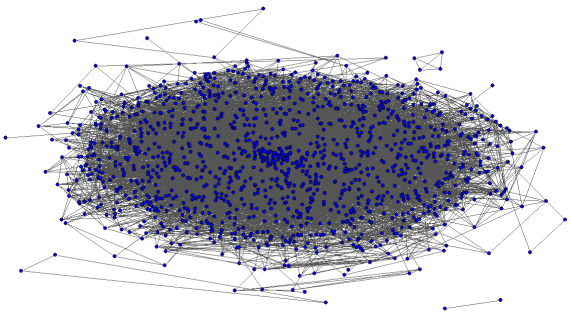
A Figura 5.7 mostra exemplos do aspecto visual de redes de títulos onde foram aplicados índice incidência-fidelidade de Teixeira et al. (2010) (Figuras 5.7(a) e 5.7(c)) e Aguiar (2009) (Figura 5.7(b)).

5.5.1 *Método Utilizando a* incidência-fidelidade *proposto por Teixeira et al. (2010)*

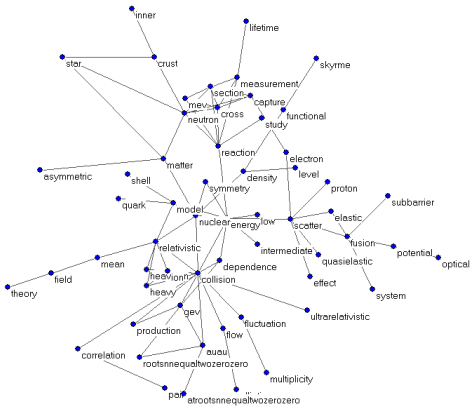
Nesta etapa, as redes foram geradas utilizando o índice *incidência-fidelidade*, proposto por Teixeira (2007), Teixeira et al. (2010). Foi escolhido em torno de 20 valores[13] de incidência fidelidade, que variaram de $1 \cdot 10^{-4}$ até $5 \cdot 10^{-2}$, para cada periódico. Contudo, para esta análise, o número de títulos de cada periódico foi fixado em 300, já que para o método de Teixeira et al. (2010) os discursos escritos não podem diferir muito em quantidade de sentenças. Este número foi escolhido porque a revista *AFE*, de 370 títulos, é a que possui a menor quantidade de títulos da base. A *Nature*, de 35163 títulos, é a revista que possui a maior quantidade de títulos da base. A *Science* vem em 2º lugar, com 11799 títulos.

Os 300 títulos de cada revista foram escolhidos de forma aleatória dentro do universo de títulos de cada uma. Para realizar este procedimento, foi usado o parâmetro "rmsentences"do software livre *guash*, produzido por Monteiro et al. (2010). Para cada periódico foram gerados 100 arquivos de texto contendo 300 títulos cada um. Cada arquivo passou pelo processo de tratamento descrito anteriormente e em seguida foram geradas as redes e calculados os índices de redes para cada valor de *incidência-fidelidade*. Feito isso, foram calculados os valores médios destes índices em cada *incidência-fidelidade*, como mostrado na Tabela 5.2.

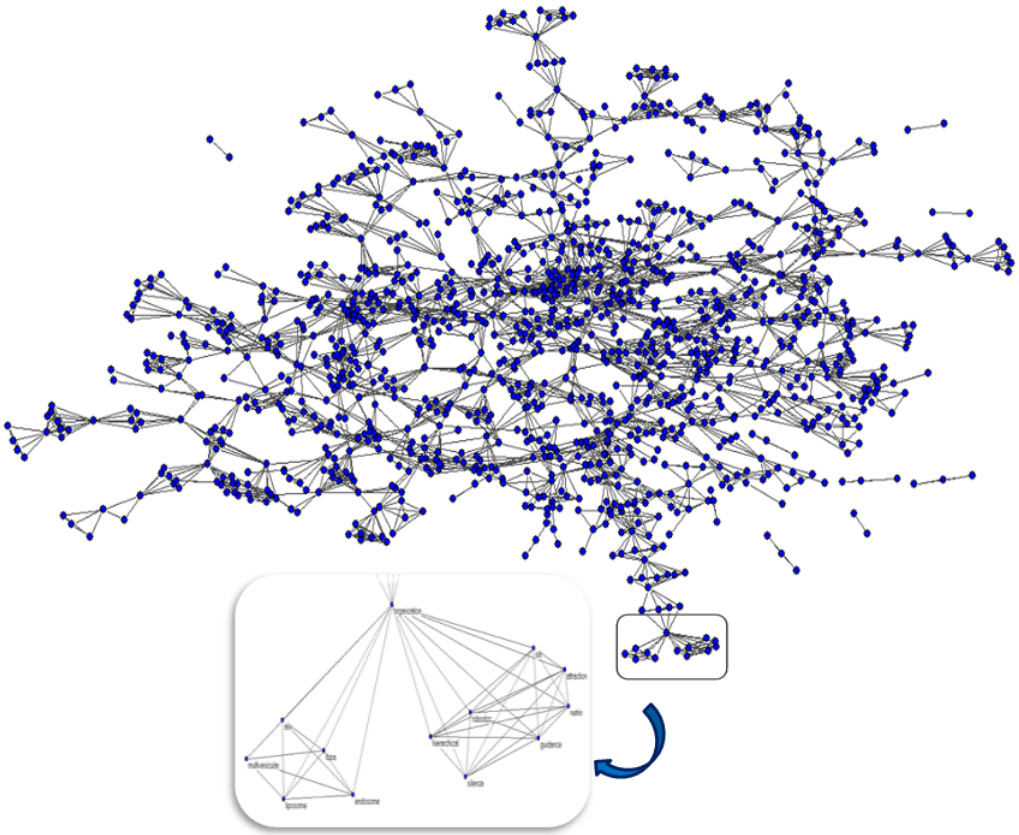
[13] 20 é satisfatório para perceber o fenômeno da rede crítica. Para alguns periódicos foi preciso 22 ou 23 valores, para melhor detalhamento da região crítica. Outros periódicos, com apenas 15 valores já foram suficientes.



(a) Rede formada por 300 títulos da revista CB, para $IF_L > 1 \cdot 10^{-4}$.



(b) Rede formada por 1572 títulos da revista PRC para $IF_L > 1 \cdot 10^{-2}$.



(c) Rede formada por 300 títulos da revista Science para $IF_L > 1 \cdot 10^{-3} = IF_C$.

Figura 5.7: Exemplos de redes de títulos para diferentes valores de *incidência fidelidade*.

5.5.2 *Método Utilizando a Incidência-Fidelidade proposto por Aguiar (2009)*

Nesta etapa da pesquisa, o índice IF proposto por Aguiar (2009) foi aplicado nas redes de títulos. Este índicador não depende da quantidade de sentenças que possui o texto. Assim, para esta aplicação, não foi necessário fixar o número de títulos dos periódicos, ou seja cada periódico da base foi analisado por todos os seus títulos. Isto fez com que se

IF_L	$\bar{n}$	$\bar{m}$	$\bar{D}$	$\overline{\langle C \rangle}$	$\overline{\langle \ell \rangle}$	$\overline{\langle k \rangle}$	$\bar{\Delta}$
0.0001	736.9	10226.6	5.2	0.74	2.80	13.9	0.0189
0.0003	733.1	6439.5	7.3	0.67	3.56	8.8	0.0120
0.0004	730.1	5309.9	8.7	0.65	4.01	7.3	0.0100
0.0005	725.9	4401.4	10.7	0.63	4.62	6.1	0.0084
0.0008	710.4	3180.1	17.4	0.61	6.28	4.5	0.0063
0.001	685.2	2509.1	27.2	0.59	7.31	3.7	0.0054
0.0015	610.4	1696.1	15.5	0.56	2.27	2.8	0.0046
0.0019	461.8	948.6	8.7	0.50	1.25	2.1	0.0045
0.002	453.7	926.5	8.1	0.50	1.22	2.0	0.0045
0.003	424.1	837.7	6.2	0.51	1.10	2.0	0.0047
0.0045	47.9	76.2	5.3	0.18	1.41	1.6	0.0347
0.005	45.4	69.1	4.9	0.18	1.35	1.5	0.0350
0.01	15.9	19.6	2.5	0.07	1.18	1.2	0.0871

IF_L	$\bar{n}$	$\bar{m}$	$\bar{D}$	$\overline{\langle C \rangle}$	$\overline{\langle \ell \rangle}$	$\overline{\langle k \rangle}$	$\bar{\Delta}$
0.0001	1300.4	18893.6	5.9	0.83	2.90	14.5	0.0112
0.0003	1298.4	13909.6	7.5	0.79	3.58	10.7	0.0083
0.0004	1296.7	11937.2	8.7	0.78	4.02	9.2	0.0071
0.0005	1294.8	10439	10.2	0.78	4.53	8.1	0.0062
0.0007	1286.9	8582.9	14.2	0.77	5.76	6.7	0.0052
0.0008	1284.8	8504.5	14.9	0.77	5.94	6.6	0.0052
0.001	1263.7	7231.38	23.5	0.76	7.79	5.7	0.0045
0.0015	1011.1	3438.72	4.9	0.81	1.03	3.4	0.0034
0.002	995.7	3406.96	4.0	0.82	1.02	3.4	0.0034
0.0025	1192.2	5564.9	30.7	0.78	4.33	4.7	0.0039
0.0045	32.9	46.38	2.2	0.20	1.04	1.4	0.0446
0.005	32.0	44.88	2.1	0.20	1.04	1.4	0.0455
0.01	7.5	8.38	1.4	0.09	1.02	1.1	0.2073

Tabela 5.2: Valores médios dos índices das 100 redes de 300 títulos escolhidos aleatoriamente, para cada valor de IF_L, dos periódicos *Probabilistic Engineering Mechanics* (PEM), à esquerda, e *Chemistry & Biology* (CB), à direita.

evitasse dúvidas sobre o valor da *densidade* ou *caminho mínimo médio* na tentativa de diferenciar periódicos.

Foram escolhidos 600 pontos igualmente espaçados dentro do intervalo de $0 \leq IF_L < 3 \cdot 10^{-2}$. Com este intervalo é possível gerar não só a rede crítica, como também a rede canônica, i.e. rede sem arestas removidas $IF_L = 0$. A Tabela 5.3 mostra os valores dos índices de redes para alguns valores de IF para o periódico PEM.

IF_N	n	m	D	$\langle C \rangle$	$\langle l \rangle$	$\langle k \rangle$	Δ
0_x10^0	1186	25806	4	0.77	2.46	21.76	0.018
5_x10^{-5}	456	4568	6	0.58	2.62	10.02	0.022
1_x10^{-4}	456	4564	6	0.58	2.62	10.01	0.022
5_x10^{-4}	404	3536	6	0.48	2.87	8.75	0.022
1_x10^{-3}	380	2640	7	0.38	3.09	6.95	0.018
2_x10^{-3}	335	1694	10	0.30	3.36	5.06	0.015
3_x10^{-3}	303	1262	10	0.30	3.09	4.17	0.014
6_x10^{-3}	236	698	8	0.27	2.48	2.96	0.013
9_x10^{-3}	193	486	10	0.28	2.19	2.52	0.013
1_x10^{-2}	167	424	10	0.27	2.25	2.54	0.015
2_x10^{-2}	114	238	6	0.29	1.51	2.09	0.018
3_x10^{-2}	58	122	6	0.15	1.56	2.10	0.037

Tabela 5.3: Valores dos índices, para alguns valores de IF_L, do periódico *Probabilistic Engineering Mechanics* (PEM).

5.5.3 Validação do método

O método computacional de geração das redes semânticas de títulos utilizado nesta pesquisa foi testado com mesma base de dados de Pereira et al. (2011), para os periódicos

no idioma Inglês. Para isto, foram calculados os índices para redes com o uso da menor *Incidência Fidelidade*, a partir do método de Teixeira et al. (2010); e com o uso de $IF = 0$, a partir do método de Aguiar (2009). Em ambas as situações, o *NetPal* gerou as redes e o *NetAll* calculou os índices para as redes *canônicas*[14].

A Figura 5.8 mostra os valores de *densidade* e *caminho mínimo médio* para os 15 periódicos em inglês. Os periódicos utilizados nesta pesquisa são os mesmos utilizados por Pereira et al. (2011), porém com períodos de coletas diferentes em alguns periódicos. Entretanto, para a validação do método foi usado o mesmo período de publicação de todos os periódicos da base de Pereira et al. (2011).

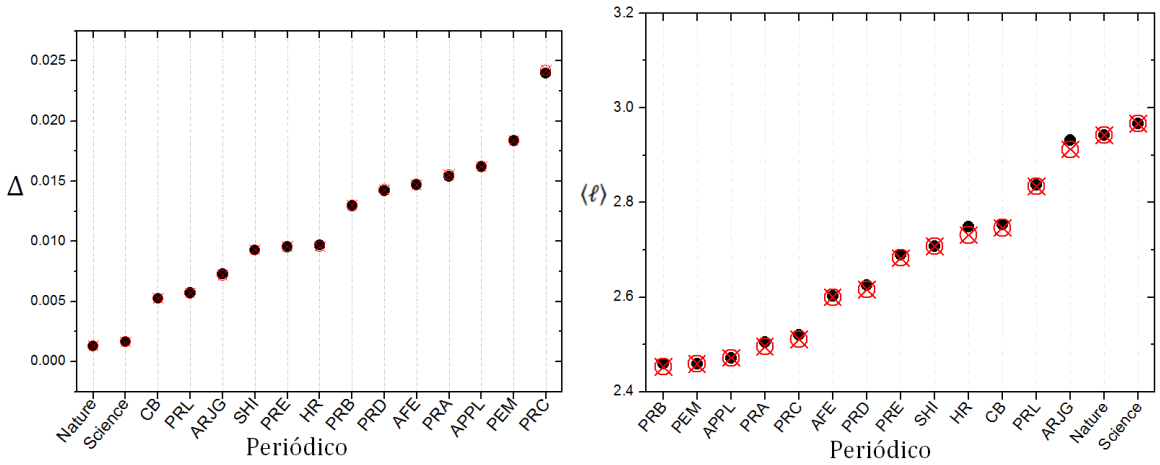


Figura 5.8: Densidade e caminho mínimo médio das redes dos periódicos calculados pelo método computacional proposto - com *IF=0* (cruz vermelha) - e obtidos no trabalho de Pereira et al. (2011) (círculos pretos). A base de dados utilizado no presente trabalho, para este experimento, foi mantida a mesma que o artigo comparado utilizou.

5.6 Grafos que variam no tempo (Time-Varying Graphs - TVG)

Recentemente, alguns trabalhos em redes sociais e complexas têm focado na evolução temporal de redes. Amblard et al. (2011) investiga as relações de coautoria e citações entre autores de artigos científicos. Na biologia, Silva et al. (2012) analisa a evolução temporal de sinais cerebrais em redes de neurônios de ratos de comportamento livre.

Redes sociais estão definidas formalmente por um conjunto V de Vértices (ou atores da rede), que são amarrados por um ou mais tipos de relações (WASSERMAN; FAUST, 1994). Este conjunto de vértices, representa objetos reais (e.g. pessoas, instituições, palavras, neurônios, etc.) e denotaremos por $n = |V|$ a cardinalidade do conjunto V. O conjunto $\mathcal{E}$ das relações entre os elementos de V é chamado conjunto de arestas, que denotaremos

[14]Canônica é a denominação dada à rede completa, sem nenhuma remoção de arestas a partir do *Incidência Fidelidade* (AGUIAR, 2009).

por $m = |\mathcal{E}|$ a cardinalidade deste conjunto. A estrutura de uma rede social de n atores pode ser modelada por um grafo $G = (V, \mathcal{E})$ e sua análise pode ser feita através de índices estatísticos, que dependem unicamente de informações contidas nos dois conjuntos citados acima.

Em redes sociais tradicionais, a utilização de grafos estáticos (com vértices e arestas fixos), no estudo de problemas de conectividade da rede, consegue representar bem as relações entre os objetos envolvidos na rede e, em geral, caracterizar bem uma comunidade. Todavia, em redes dinâmicas, onde existe uma constante mudança dos objetos que compõem a rede e suas interações, fazem-se necessárias várias representações que não possuam vértices e arestas fixos (SANTANA, 2012).

Estudos mais recentes têm usado esta modelagem com a inclusão de, pelo menos, mais um conjunto dentro do grafo original, com elementos que representam o tempo. Isto torna o grafo variável no tempo. Desta forma, segundo Amblard et al. (2011), os vértices de uma rede dinâmica podem aderir, atrair, competir e até cooperar com outros vértices. Eles podem ainda, desaparecer e até afetar a forma e solidez de seu sistema de relacionamentos.

Assim, nos últimos anos, alguns trabalhos têm apresentado diversas formas de estudar redes variáveis no tempo. Casteigts et al. (2011) teve como objetivo unir e formalizar os diversos conceitos e métricas utilizados no estudo das redes dinâmicas, criando assim o conceito de *Grafos que variam no tempo* (Time-Varying Graph ou TVG).

Um TVG pode ser entendido como um grafo estático $G = (V, \mathcal{E})$ acrescido de outros parâmetros que representam funções ou conjuntos temporais: ς (i.e. função de latência), Υ(i.e. função de presença) e Γ(i.e. tempo de vida). Assim um TVG é a quíntupla $\mathcal{G} = (V, \mathcal{E}, \Upsilon, \varsigma, \Gamma)$, onde V e $\mathcal{E}$ representam respectivamente o conjunto de vértices e arestas; a função $\Gamma \subset \mathbb{N}$ representa o tempo de vida do sistema. A função de latência ς indica quanto tempo necessita para que uma aresta esteja disponível em um instante $t \in \Gamma$, em outras palavras, é o tempo necessário para estabelecer o relacionamento entre dois vértices, em um dado instante t. $\Upsilon : \mathcal{E} \times \Gamma \rightarrow \{0, 1\}$ é definido como uma função de presença e garante a existência de uma dada aresta em um dado instante de tempo t.

5.6.1 Aplicação do Método em Rede de Títulos

Com este método é possível investigar a evolução temporal de vértices, arestas e índices de um grafo formado por palavras impressas nos títulos de artigos científicos publicados nos periódicos *Nature*, de 07 de Janeiro de 1999 à 18 de dezembro de 2008, e Science, de 07 de Janeiro de 1999 à 18 de dezembro de 2008.

A escolha destes periódicos se deu pela possibilidade de comparação entre eles, já que publicam semanalmente e possuem artigos publicados na mesma época supracitada. Como também são os periódicos de mais alto impacto em ciência no mundo, este estudo pode contribuir para o estudo da difusão do conhecimento humano. A escolha da época (dez anos de publicações) deveu-se pela possibilidade de comparação com trabalhos de mesma natureza, i.e. Fadigas et al. (2009) e Pereira et al. (2011).

Inicialmente os títulos foram agrupados por semana em 507 arquivos de texto. Posteriormente, para a construção das *janelas temporais*, estes arquivos foram agrupados em grupos de 8 arquivos, ou seja 8 semanas. Consideremos a rede de títulos de artigos publicados nos meses de janeiro e fevereiro de 1999. Todos eles compõe a 1ª janela do TVG, ou seja $t = 1$. Da mesma forma, os títulos das publicações que compreendem o mês de Janeiro, exceto a 1ª semana, todo mês de fevereiro e a 1ª semana de março compõe a 2ª janela do TVG, ou seja $t = 2$. Isso se repete até a ultima janela $t = 507$ que corresponde aos títulos das publicações dos meses de Novembro e Dezembro do ano de 2008. Para melhor adaptar nossa proposta ao referencial teórico sobre TVG proposto por Casteigts et al. (2011), foram consideradas as seguintes condições:

- A função de latência ς é constante para todo o TVG. Não faz sentido quantificá-la, portanto este parâmetro não será levado em consideração em nossa análise;
- O tempo de vida do sistema de nossa amostra é o conjunto $\Gamma = \{t_1, t_2, ..., t_i, t_{i+1}, ..., t_{507}\}$. Em que cada t_i corresponde ao intervalo de tempo de 01 *semana*. $|\Gamma| = 507$ *semanas*;
- O tempo se inicia na 1ª semana de Janeiro de 1999;
- A função de presença pode ser simplificada e melhor entendida com o uso de uma janela temporal, semelhante à utilizada no trabalho de Silva et al. (2012) e aos *footprints* do trabalho de Casteigts et al. (2011). Dessa forma, o TVG em questão será um conjunto de grafos estáticos $\mathcal{G}^{[t_i,t_j]} = (V, \mathcal{E}^{[t_i,t_j]})$. De tal forma que $\forall e \in \mathcal{E}, e \in \mathcal{E}^{[t_i,t_j]} \Leftrightarrow \exists t \in [t_i, t_j], \Upsilon(e, t) = 1$. Ou seja, $\mathcal{G} = \{G^{[t_i,t_j]}, G^{[t_{i+1},t_{j+1}]}, G^{[t_{i+2},t_{j+2}]}, ...\}$;
- Cada *janela* temporal de observação será definida por $\tau_i = [t_i, t_{i+7}]$. Deste modo o TVG pode ser escrito em função dessas janelas: $\mathcal{G} = \{\mathcal{G}^{\tau_1}, \mathcal{G}^{\tau_2}, \mathcal{G}^{\tau_3}, ..., \mathcal{G}^{\tau_{507}}\}$;
- A escolha de uma janela de 8 *semanas* deu-se para suavizar as flutuações dos valores dos índices a cada mudança de janela. Isto, no entanto, não ofusca as tendências dos valores dos índices no TVG como um todo.

Sendo assim, o TVG em estudo é formado por um conjunto de subgrafos estáticos. Cada subgrafo é formado por uma rede de títulos de oito semanas de publicações da revista Nature e existe apenas durante uma unidade de tempo. O tempo do TVG é dado em

semanas. Assim, a cada *semana* t existe uma rede formada por títulos publicados em oito semanas, a partir dela. Ou seja $\mathcal{G}(V_t, \mathcal{E}_t, t) = \mathcal{G}(V_t + ... + V_{t+7}, \mathcal{E}_t + ... + \mathcal{E}_{t+7})$. Por exemplo, o subgrafo grado pelo tempo $t = 51$ corresponde a uma junção das semanas $51, 52, 53, 54, 55, 56, 57$ e 58. Ou seja, começa no final de dezembro de 1999 e termina em meados de fevereiro de 2000. E para o próximo grafo, $t = 52$, o procedimento será análogo.

Assim, a janela temporal de oito semanas avança no tempo, Figura 5.9. E deste modo, é possível analisar o que ocorre com os valores de índices de redes para esta janela ao longo do tempo. Para análises que envolvem todo o TVG em uma única rede, constrói-se a rede colapsada, como na Figura 5.10.

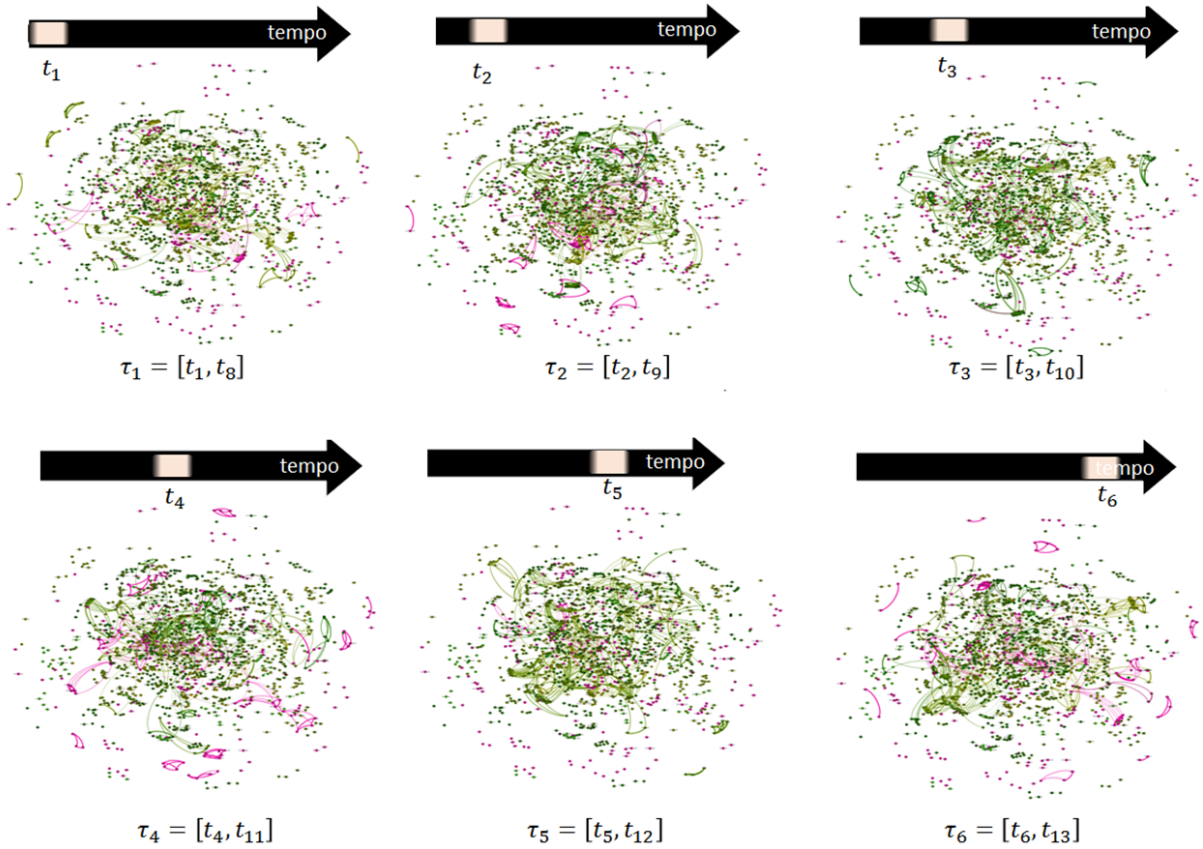


Figura 5.9: Evolução da janela de 8 semanas ao longo do tempo, entre a primeira semana e a decima terceira semana.

5.6.2 Método DFA para redes de títulos

O registro dos valores dos índices de redes, das janelas do TVG, formam séries temporais. Uma série temporal consiste em uuma seqüência de dados obtidos ao longo de um determinado período de tempo (507 semanas para o periódico estudado). A análise das séries temporais feitas neste trabalho buscam identificar correlações nas sequência dos registros

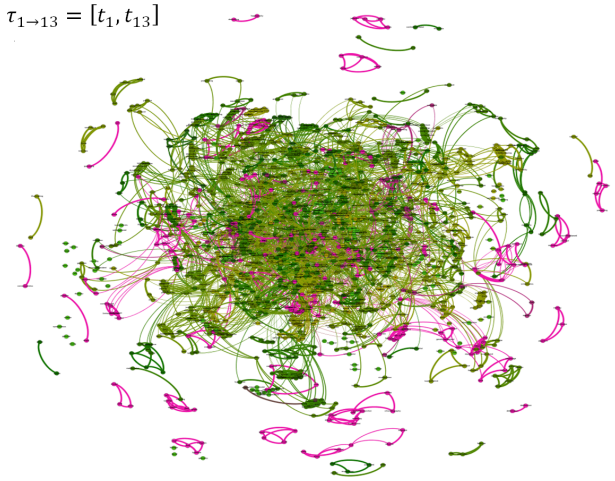


Figura 5.10: Grafo completo de 13 semanas do TVG.

de cada série e a partir disto, verificar se existe um efeito de memória ao longo do tempo nos dados das séries.

O método empregado aqui é o *Detrended Fluctuation Analysis* (*DFA*), proposto por Peng et al. (1994). Este método consiste em medir o quanto uma medida de variação, denotada por F, varia em janelas de tempo de tamanho t. Se para a série temporal a medida de variação se comportar como uma lei de potência, $F \propto t^H$, então a série possui auto-afinidade, ou seja, a sequencia de valores é correlacionada no tempo.

Esta correlação é dita sem memória se $\mathcal{H} = 0,5$. Neste caso, o estado do sistema em um instante não guarda relação para outro estado em instantes posteriores e, portanto, os valores registrados na série temporal não possui memória, como acontece no movimento browniano. Os valores de $\mathcal{H} > 0.5$ representam correlações de longo alcance persistentes. Ou seja, uma tendência positiva no passado é mais provável de continuar positiva e vice-versa. Já para $\mathcal{H} < 0,5$, os dados teriam uma correlação anti-persistente. Isto significa tendências contrárias para o valor de algo, em relação ao que aconteceu no passado.

Este método será útil para verificar, a partir de redes de títulos, se publicações em uma época possuem correlações com publicações de épocas passadas.

5.7 Análise dos Resultados

Existem muitas maneiras de se analisar propriedades emergentes em uma rede semântica, principalmente rede de títulos, que é a rede do "discurso" de um periódico, "escrita" por mais de um autor. Entretanto, devido ao limite de tempo dado para esta pesquisa, a análise foi guiada apenas com o intuito de responder a pergunta norteadora, com foco

nas questões auxiliares, vistas na Seção 1.4, com a utilização dos aspectos metodológicos descritos neste capítulo.

Cada aspecto metodológico para a análise de um determinado resultado é apresentado juntamente com seu respectivo resultado, nos Capítulos 6 e 7 que seguem.

Parte IV

Resultados da Pesquisa

Capítulo Seis

Resultados envolvendo Incidência Fidelidade

Para cada periódico, os resultados mostram que a rede de um dado conjunto de títulos possui configuração crítica. Como foi visto, esta rede contém as associações de palavras mais fiéis em um discurso. Ela apresenta-se de forma categorizada e representa o acesso à memoria[1] do periódico, no que concerne às temáticas de suas publicações.

As análises dos resultados foram motivadas pela busca de propriedades emergentes das interações de palavras em redes de títulos e pela busca de padrões críticos nesse canal da difusão do conhecimento humano: o periódico científico.

A partir da pergunta que norteia esta pesquisa (apresentada na Seção 1.4: *Os periódicos científicos permitem uma classificação a partir dos títulos de seus artigos?*), foi possível definir os objetivos específicos deste trabalho (Seção 1.6). A partir dos objetivos, para guiar a análise dos resultados neste capítulo, faz-se necessário o uso das seguintes perguntas:

1. É possível encontrar uma rede ótima de palavras onde se tem o máximo de informações com o mínimo de ruído, para rede de títulos?;
2. As topologias e índices das redes críticas podem ser capazes de diferenciar periódicos?

Os resultados desta etapa da pesquisa estão divididos em duas seções. A primeira seção apresenta resultados da aplicação do Índice *incidência-fidelidade* proposto por Teixeira et al. (2010) nas redes de títulos da base de dados. A segunda, trata da aplicação do *incidência-fidelidade* ressignificado por Aguiar (2009) nas redes de títulos dos mesmos periódicos.

Na etapa que envolve o índice *incidência-fidelidade* proposto por Teixeira et al. (2010), o número de sentenças dos discursos (aqui visto como títulos de um periódico) precisam ser iguais ou próximos entre si. Assim foi verificado a existência da rede crítica para cada uma das 100 redes de 300 títulos de cada periódico (método descrito na Seção 5.5.1). Na etapa que envolve o índice *incidência-fidelidade* proposto por Aguiar (2009), os discursos não precisam ter o mesmo número de títulos. Assim, na construção da rede crítica de um dado periódico, considera-se todos os títulos de sua base.

[1] De acordo com Sternberg (2011) uma rede semântica pode representar o acesso à memoria declarativa de quem proferiu o discurso.

6.1 *Resultados para o* incidência-fidelidade *de Teixeira et al. (2010)*

Para responder a a questão 1 do início deste capítulo, utilizou-se o procedimento descrito na Seção 5.5.1; Foram feitas 100 retiradas aleatórias de 300 títulos para cada periódico, com objetivo de contemplar o máximo de títulos dentre todos da base de cada periódico.

Para cada periódico da base, em todos os seus 100 arquivos que contém 300 títulos cada, foram calculados alguns índices de redes para vários valores de *incidência-fidelidade limite*. Os valores de *incidência-fidelidade* utilizados foram escolhidos empiricamente de forma à obter um maior detalhamento, que possibilite encontrar a região crítica das redes. A Tabela 6.1 apresenta resultados para dois periódicos, em destaque suas regiões críticas.

IF_L	$\bar{n}$	$\bar{m}$	$\bar{D}$	$\overline{\langle C\rangle}$	$\overline{\langle \ell\rangle}$	$\overline{\langle k\rangle}$	$\bar{\Delta}$
0.0001	736.9	10226.6	5.2	0.74	2.80	13.9	0.0189
0.0003	733.1	6439.5	7.3	0.67	3.56	8.8	0.0120
0.0004	730.1	5309.9	8.7	0.65	4.01	7.3	0.0100
0.0005	725.9	4401.4	10.7	0.63	4.62	6.1	0.0084
0.0008	710.4	3180.1	17.4	0.61	6.28	4.5	0.0063
0.001	685.2	2509.1	27.2	0.59	7.31	3.7	0.0054
0.0015	610.4	1696.1	15.5	0.56	2.27	2.8	0.0046
0.0019	461.8	948.6	8.7	0.50	1.25	2.1	0.0045
0.002	453.7	926.5	8.1	0.50	1.22	2.0	0.0045
0.003	424.1	837.7	6.2	0.51	1.10	2.0	0.0047
0.0045	47.9	76.2	5.3	0.18	1.41	1.6	0.0347
0.005	45.4	69.1	4.9	0.18	1.35	1.5	0.0350
0.01	15.9	19.6	2.5	0.07	1.18	1.2	0.0871

IF_L	$\bar{n}$	$\bar{m}$	$\bar{D}$	$\overline{\langle C\rangle}$	$\overline{\langle \ell\rangle}$	$\overline{\langle k\rangle}$	$\bar{\Delta}$
0.0001	1300.4	18893.6	5.9	0.83	2.90	14.5	0.0112
0.0003	1298.4	13909.6	7.5	0.79	3.58	10.7	0.0083
0.0004	1296.7	11937.2	8.7	0.78	4.02	9.2	0.0071
0.0005	1294.8	10439	10.2	0.78	4.53	8.1	0.0062
0.0007	1286.9	8582.9	14.2	0.77	5.76	6.7	0.0052
0.0008	1284.8	8504.5	14.9	0.77	5.94	6.6	0.0052
0.001	1263.7	7231.38	23.5	0.76	7.79	5.7	0.0045
0.0015	1011.1	3438.72	4.9	0.81	1.03	3.4	0.0034
0.002	995.7	3406.96	4.0	0.82	1.02	3.4	0.0034
0.0025	1192.2	5564.9	30.7	0.78	4.33	4.7	0.0039
0.0045	32.9	46.38	2.2	0.20	1.04	1.4	0.0446
0.005	32.0	44.88	2.1	0.20	1.04	1.4	0.0455
0.01	7.5	8.38	1.4	0.09	1.02	1.1	0.2073

Tabela 6.1: Valores médios dos índices das 100 redes de 300 títulos escolhidos aleatoriamente, para cada valor de IF_L, dos periódicos *Probabilistic Engineering Mechanics* (PEM), à esquerda, e *Chemistry & Biology* (CB), à direita.

A Figura 6.1 mostra os valores de $\overline{\langle \ell\rangle}$ de todos os periódicos em função de IF_L, inclusive a média de seus valores para todos os periódicos. Os dois gráficos da Figura 6.1 revelam o valor da *incidência-fidelidade* crítica para as redes estudadas. Vê-se claramente que as redes críticas dos periódicos analisados são formadas por valores muito próximos de $IF_L = 10^{-3}$.

A partir dos gráficos da Figura 6.1, conclui-se que todos os periódicos apresentam comportamento crítico para $IF_L \cong 10^{-3}$. Os gráficos revelam que a partir da configuração inicial de um dos conjuntos de 300 títulos, ao passo que se aumenta IF_L, arestas são perdidas e atalhos entre as palavras tendem a se perder, já que pares de palavras com baixo valor de IF tem suas arestas eliminadas da rede neste processo. Com isso, o $\langle \ell\rangle$ aumenta até seu valor máximo. Nesta configuração, um pequeno aumento no valor de IF_L acarreta em uma queda brusca nos valores de $\langle \ell\rangle$ da redes.

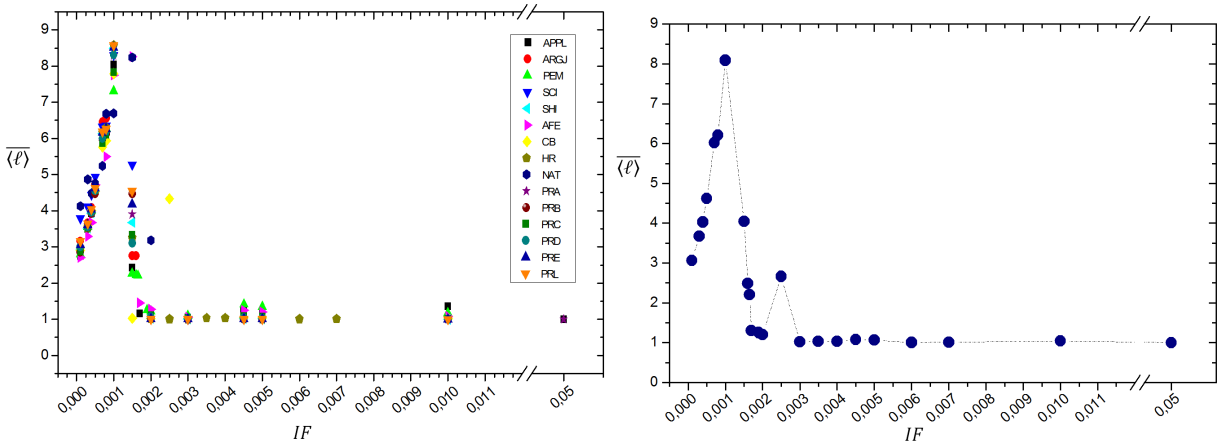


Figura 6.1: Na esquerda: Valores de $\overline{\langle \ell \rangle}$ em função de IF_L para todos os periódicos. Diferente do que possa parecer, este gráfico contém 100 pontos de cada periódico, em cada valor de IF_L. Na direita: Média dos valores de $\overline{\langle \ell \rangle}$ dos periódicos em função de IF_L

Isto acontece com praticamente todos os conjuntos de 300 títulos. Fica evidente com este resultado que todos os periódicos possuem comportamento crítico para o valor de $IF_L = IF_C = 10^{-3}$ e responde a questão 1 do início do capítulo. A Tabela 6.2 mostra o comportamento dos índices de redes na região crítica para todos periódicos estudados. Cada célula da tabela exibe a média dos valores de cada índice nas 100 redes de 300 títulos de cada periódico.

Periódico	IF_C	$\bar{n}$	$\bar{m}$	$\bar{D}$	$\overline{\langle C \rangle}$	$\overline{\langle \ell \rangle}$	$\overline{\langle k \rangle}$	$\bar{\Delta}$
AFE	*0.0015*	*1253 ± 18*	*2838 ± 90*	*43.6 ± 9.1*	*0.71 ± 0.01*	*8.27 ± 1.68*	*4.53 ± 0.10*	*0.0036 ± 0.0001*
SCI	*0.0010*	*1311 ± 24*	*3698 ± 156*	*26.2 ± 3.7*	*0.81 ± 0.01*	*8.36 ± 0.75*	*5.64 ± 0.16*	*0.0043 ± 0.0001*
PRL	*0.0010*	*1139 ± 25*	*2842 ± 136*	*24.8 ± 3.3*	*0.73 ± 0.02*	*8.58 ± 0.56*	*4.99 ± 0.15*	*0.0044 ± 0.0001*
PRB	*0.0010*	*1115 ± 29*	*2742 ± 144*	*23.3 ± 2.6*	*0.70 ± 0.01*	*8.31 ± 0.54*	*4.92 ± 0.16*	*0.0044 ± 0.0001*
PRC	*0.0010*	*955 ± 23*	*2046 ± 104*	*24.8 ± 3.4*	*0.66 ± 0.02*	*7.83 ± 0.57*	*4.28 ± 0.15*	*0.0045 ± 0.0001*
PRE	*0.0010*	*1073 ± 26*	*2587 ± 146*	*25.8 ± 3.3*	*0.70 ± 0.02*	*8.51 ± 0.63*	*4.82 ± 0.17*	*0.0045 ± 0.0001*
CB	*0.0010*	*1264 ± 32*	*3616 ± 189*	*23.5 ± 3.1*	*0.76 ± 0.01*	*7.79 ± 0.55*	*5.72 ± 0.19*	*0.0045 ± 0.0001*
Nat	*0.0015*	*1114 ± 30*	*2814 ± 160*	*32.7 ± 6.0*	*0.80 ± 0.01*	*8.42 ± 1.24*	*5.05 ± 0.17*	*0.0045 ± 0.0001*
SHI	*0.0010*	*1038 ± 23*	*2449 ± 110*	*25.3 ± 3.0*	*0.70 ± 0.01*	*8.50 ± 0.55*	*4.72 ± 0.13*	*0.0045 ± 0.0001*
PRD	*0.0010*	*932 ± 23*	*1995 ± 108*	*25.8 ± 3.5*	*0.65 ± 0.02*	*8.30 ± 0.66*	*4.28 ± 0.16*	*0.0046 ± 0.0001*
PRA	*0.0010*	*1003 ± 25*	*2312 ± 130*	*25.3 ± 3.3*	*0.67 ± 0.02*	*8.30 ± 0.66*	*4.61 ± 0.17*	*0.0046 ± 0.0001*
HR	*0.0010*	*993 ± 23*	*2315 ± 96*	*26.8 ± 3.6*	*0.70 ± 0.02*	*8.58 ± 0.74*	*4.66 ± 0.12*	*0.0047 ± 0.0001*
ARJG	*0.0010*	*944 ± 28*	*2214 ± 141*	*30.6 ± 5.7*	*0.71 ± 0.02*	*8.57 ± 1.05*	*4.69 ± 0.19*	*0.0050 ± 0.0002*
APPL	*0.0010*	*784 ± 21*	*1573 ± 83*	*27.7 ± 4.5*	*0.62 ± 0.02*	*8.04 ± 0.80*	*4.01 ± 0.14*	*0.0051 ± 0.0002*
PEM	*0.0010*	*685 ± 18*	*1255 ± 54*	*27.2 ± 4.2*	*0.59 ± 0.02*	*7.31 ± 0.81*	*3.66 ± 0.11*	*0.0054 ± 0.0002*

Tabela 6.2: Valores médios dos indicadores e seus respectivos desvios para as 100 redes de cada periódico. A segunda coluna fornece os valores de IF_C dos periódicos estudados.

6.1.1 *Discussões para o* incidência-fidelidade *de Teixeira et al. (2010)*

A questão 1 do início do capítulo pôde ser respondida a partir da análise dos valores de $\langle \ell \rangle$ em função dos valores de IF_L das redes. o valor $IF_C = 10^{-3}$ é a *incidênciá-fidelidade limite* para gerar a rede crítica da maioria dos periódicos. Este valor é o o mesmo encontrado nas redes de discursos orais do trabalho de Teixeira et al. (2010). Isto pode indicar que a escolha das publicações pelo periódico tem alguma semelhança com a escolha das frases que um indivíduo faz para expressar suas ideias em um discurso.

A questão 2 do início deste capítulo objetiva diferenciar periódicos a partir das redes críticas. O trabalho de Pereira et al. (2011) mostra que periódicos podem ser diferenciados a partir dos valores de *densidade* (Δ) e *caminho mínimo médio*($\langle \ell \rangle$) de suas redes. Não obstante a Tabela 6.2 mostrar os valores médios destes índices para cada periódico de suas redes críticas, os gráficos da Figura 6.2 revelam que estas médias, na verdade, escondem muita informação.

A Figura 6.2(b) mostra a nuvem de valores de *densidade* para cada periódico e a Figura 6.2(a) mostra o gráfico box-plot[2] para a distribuição deste índice. Percebe-se claramente que existem sobreposições dos valores deste índice para periódicos diferentes.

Para melhor compreensão, considere dois periódicos, e.g. SHI e PRD que tem cerca de $n \simeq 1000$ vértices (Tabela 6.2). O menor incremento no valor de *densidade* $\delta\Delta$ (veja-o como um instrumento de medida), pode ser obtido com incremento de 1 aresta, observe a Equação 6.1:

$$\delta\Delta = 2 \cdot \frac{1}{1000(1000-1)} \simeq 2 \cdot 10^{-6} \tag{6.1}$$

Desta forma, a menor variação de $\delta\Delta$ para uma rede com este número de nós é de $\delta\Delta \simeq 10^{-6}$. Ou seja, os valores de *densidade* de duas revistas só podem ser consideradas diferentes, se estes distarem de pelo menos 10^{-6}. Para o valor médio isto não é problema. Entretanto, a média esconde todos os valores que a geraram. A partir dos máximos e mínimos de *densidade* é possível calcular quantos valores de *densidade* η "cabem" dentro do intervalo $[\Delta_{MINIMO}, \Delta_{MAXIMO}]$, a partir da Equação 6.2.

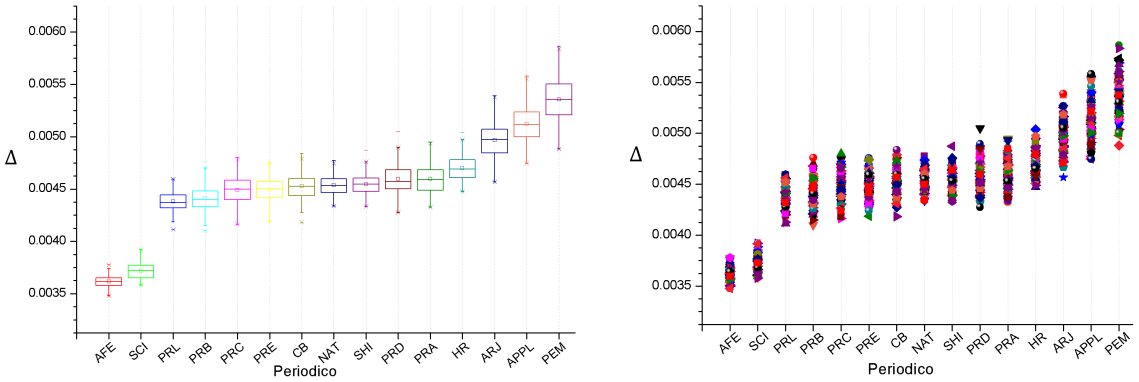
$$\eta = \frac{\Delta_{MAXIMO} - \Delta_{MINIMO}}{\delta\Delta} \tag{6.2}$$

[2] *Boxplot* é uma ferramenta útil para a comparação gráfica de várias amostras. seu gráfico informa importantes aspectos do conjunto de dados através de cinco números: valor mínimo, 1° quartil, 2° quartil, 3° quartil e valor máximo (TUKEY, 1977).

Para os dois periódicos do exemplo, obtém-se os seguintes resultados:

- PRC: $\Delta_{MINIMO} = 0.0042$; $\Delta_{MAXIMO} = 0.0048$ e $\eta = 1271$;
- PRD: $\Delta_{MINIMO} = 0.0042$; $\Delta_{MAXIMO} = 0.0050$ e $\eta = 1544$.

Isto significa que entre os valores máximos e mínimos destes índices "cabem" mais de mil valores diferentes. O gráfico da Figura 6.2 nos mostra que 300 pontos estão distribuídos em intervalos que cabem mais de 1000. Ou seja, é inviável diferenciar um periódico de outro a partir do valor médio de sua *densidade*, pois a faixa de valores para este índice é muito grande, o que gera sobreposição de faixas de valores de vários periódicos.



(a) Média, Máximo, Mínimo e Mediana dos valores de densidade das 100 redes de 300 títulos geradas para cada periódico.

(b) Nuvem de pontos dos valores das densidades das 100 redes de de cada revista.

Figura 6.2: Resultado mostra que é inviável, por este método, diferenciar periódicos a partir da média das densidades das amostras de cada periódico.

Os Gráficos da Figura 6.2 mostram, no que diz respeito ao poder de relacionamento das palavras (*densidade*), que os periódicos PEM e $APPL$ são diferentes dos periódicos AFE e *Science*. Entretanto, não se pode dizer o mesmo de outros periódicos, pois as nuvens de pontos de suas densidades se sobrepõem.

A próxima seção também busca responder as duas questões propostas no ínicio do capítulo, mas a partir do índice *incidência-fidelidade* proposto por Aguiar (2009).

6.2 *Resultados para o* incidência-fidelidade *de Aguiar (2009)*

A Figura 6.3 exibe os pontos críticos para dois periódicos. Todos os outros exibiram comportamento semelhante, exceto um, Figura 6.4.

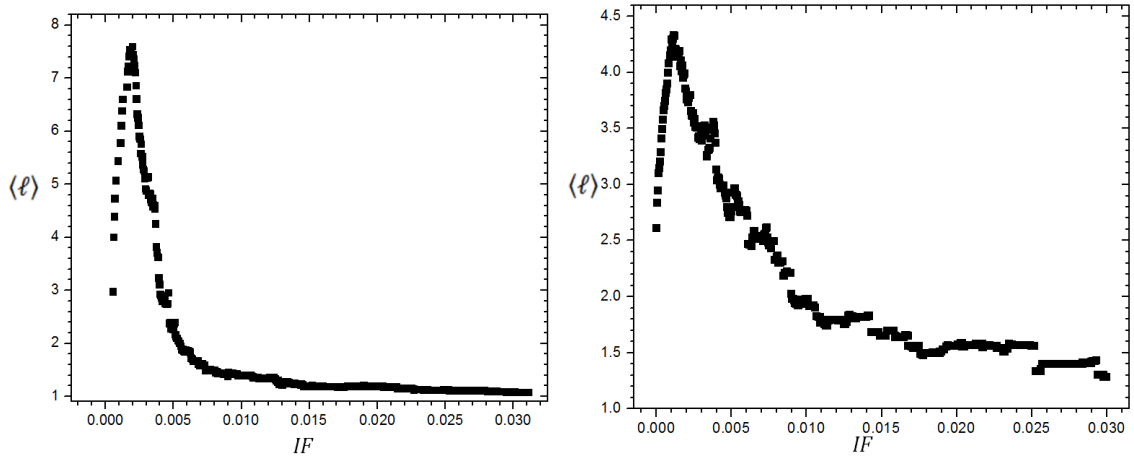


Figura 6.3: *caminho mínimo médio* em função da *incidência-fidelidade limite* para os periódicos Science (à esquerda) e PRC (à direita). O primeiro possui comportamento crítico para $IF_L = 2,5 \cdot 10^{-3}$ e o segundo, $IF_L = 1,1 \cdot 10^{-3}$.

A Tabela 6.4 mostra os valores dos índices de redes para os 15 periódicos na região crítica. A Tabela 6.3 apresenta os valores dos mesmos índices para as redes com $IF = 0$, ou seja, construída sem remoção de nenhuma aresta[3]. A rede nesta condição não teve arestas removidas e é denominada *rede canônica*.

PERIÓDICO	*IF*	*n*	*m*	*D*	⟨*C*⟩	⟨*l*⟩	⟨*k*⟩	*γ*	(%) Maior Componente	Δ
PRL	0	7368	108126	5	0.75	2.75	29.35	2.73 ± 0.08	99.9%	0.0040
SHI	0	2098	20442	5	0.76	2.71	19.49	2.52 ± 0.16	100.0%	0.0093
PRE	0	5296	80081	5	0.73	2.66	30.24	2.66 ± 0.18	99.9%	0.0057
PRD	0	4698	73792	5	0.74	2.66	31.41	2.75 ± 0.19	100.0%	0.0067
PRC	0	2736	32887	5	0.77	2.61	24.04	2.33 ± 0.10	100.0%	0.0088
PRB	0	8387	161433	5	0.79	2.59	38.50	2.33 ± 0.10	100.0%	0.0046
PRA	0	4021	59141	5	0.74	2.62	29.42	2.67 ± 0.13	100.0%	0.0073
HR	0	1870	16691	6	0.77	2.73	17.85	2.16 ± 0.10	99.0%	0.0096
PEM	0	1186	12903	4	0.77	2.46	21.76	2.43 ± 0.18	100.0%	0.0184
CB	0	4220	46560	5	0.80	2.75	22.07	2.41 ± 0.11	99.5%	0.0052
ARGJ	0	2155	16633	7	0.76	2.91	15.44	2.18 ± 0.23	100.0%	0.0072
APPL	0	1322	14151	5	0.76	2.47	21.41	2.25 ± 0.17	100.0%	0.0162
AFE	0	1561	17880	5	0.79	2.60	22.91	2.57 ± 0.24	99.8%	0.0147
Science	0	15227	188346	9	0.72	2.97	24.74	2.74 ± 0.12	99.4%	0.0016
Nature	0	23081	172773	5	0.81	2.75	22.06	2.94 ± 0.12	99.5%	0.0006
MÉDIA	**0**	**5682**	**68123**	**5**	**0.76**	**2.68**	**24.71**	**2.51**	**99.8%**	**0.0080**
DESVIO	0	6063	61951	1	0.03	0.14	6.02	0.23	0.3%	0.0051

Tabela 6.3: Índices de redes para os periódicos em suas redes canônicas $IF_L = 0$

As *incidências-fidelidades* críticas foram determinadas como descrito na Metodologia,

[3]Como foi na Seção 4.5.2, esta condição ocorre se I for mínimo ou se F for mínimo.

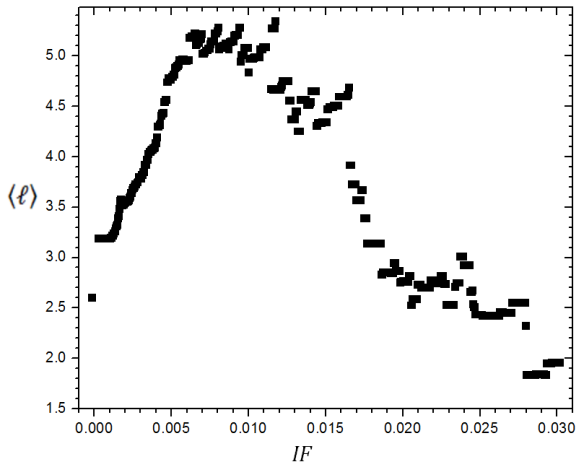


Figura 6.4: *caminho mínimo médio* em função da *incidência-fidelidade limite* para o periódico *AFE*. Na ausência de um ponto crítico bem definido, considerou-se para este periódico o ponto no gráfico onde o valor do *caminho mínimo médio* é máximo.

PERIÓDICO	*IF*	*n*	*m*	*D*	⟨*C*⟩	⟨*l*⟩	⟨*k*⟩	γ	(%) Maior Componente	Δ
PRL	**0.0018**	1544	2470	17	0.46	5.78	3.20	2.84 ± 0.21	63%	0.0021
SHI	**0.0035**	379	566	14	0.28	4.97	2.98	2.04 ± 0.16	84%	0.0079
PRE	**0.0035**	492	963	8	0.48	3.33	3.9	2.02 ± 0.26	96%	0.0038
PRD	**0.0006**	1708	3430	22	0.48	5.76	4.02	2.49 ± 0.11	74%	0.0024
PRC	**0.00115**	637	1552	11	0.43	4.33	4.87	2.0 ± 0.13	89%	0.0077
PRB	**0.0002**	1942	6278	14	0.4	4.54	6.47	2.16 ± 0.08	88%	0.0033
PRA	**0.00065**	968	2119	14	0.36	4.63	4.38	2.33 ± 0.16	85%	0.0045
HR	**0.0079**	313	440	13	0.35	5.07	2.81	2.48 ± 0.21	73%	0.0090
PEM	**0.00205**	335	836	10	0.46	3.83	4.99	2.16 ± 0.24	83%	0.0149
CB	**0.00185**	755	1455	14	0.53	4.86	3.85	2.22 ± 0.15	74%	0.0051
ARGJ	**0.0042**	341	540	16	0.38	5.38	3.17	2.18 ± 0.23	87%	0.0093
APPL	**0.00145**	320	646	11	0.43	4.29	4.04	2.25 ± 0.17	88%	0.0127
AFE	**0.012**	338	488	12	0.39	5.34	2.89	2.57 ± 0.24	69%	0.0086
Science	**0.0025**	1755	1641	23	0.4	7.59	1.9	2.74 ± 0.12	39%	0.0011
Nature	**0.00175**	1596	1256	22	0.48	7.33	1.57	2.94 ± 0.12	19%	0.0010
MÉDIA	**0.00301**	**895**	**1645**	**15**	**0.42**	**4.91**	**4.29**	**2.36**	**74%**	**0.0065**
DESVIO	0.00313	628	1536	5	0.05	1.08	1.91	0.30	21%	0.0042

Tabela 6.4: Índices de redes para os periódicos em suas redes críticas.

Seção 5.5. De posse destes métodos e das Tabelas 6.4 e 6.3, verifica-se que:

a. Todos os periódicos apresentaram um comportamento típico de mudança de fase, com ponto crítico bem definido, com exceção do periódico *AFE*, Figura 6.4;

b. Com exceção do periódico supracitado, os gráficos de $\langle \ell \rangle \times IF$ se adequaram perfeitamente ao método de análise, proposto na Seção 5.5. A Figura 6.3 ilustra dois

exemplos de como as redes críticas foram encontradas. Na figura, os dois periódicos possuem IF_C bem definido;

c. Os valores do expoente γ de todas as redes estão situados dentro do intervalo $2 < \gamma < 3$;

d. Todos os periódicos apresentaram *rede canônica* altamente conectadas, em média a maior componente possui 99.8% da rede. Apenas uma das redes, PRL, manteve-se fortemente conectada na rede crítica, com 99% de seus vértices conectados. A rede crítica da $Nature$ é a menos conectada, apenas cerca de 20% de seus vértices estão conectados;

e. Em geral, os valores de *diâmetro* das *redes canônicas* são menores que nas redes críticas ($D_{canonica} \simeq 5$), independente do número de títulos que possuem. Isto pode indicar algum aspecto universal em redes semânticas de títulos de artigos científicos;

f. Nas *redes canônicas*, os valores do *coeficientes de aglomeração médio* são aproximadamente o dobro dos valores deste mesmo índice nas redes críticas;

g. Os valores de *grau médio* nas redes críticas são muito menores que nas *redes canônicas*.

As Figuras 6.5, 6.7, 6.6, 6.8 e 6.9 exibem exemplos de redes críticas, destacando as *incidências-fidelidades* dos pares de palavras, simbolizadas pela espessura das arestas. A análise destas redes em softwares adequados, como o *Gephi*, pode revelar informações importantes sobre um periódico seleciona suas publicações.

A partir destas visualizações, é possível fazer inferências de cunho semântico ao observar grupos de palavras interagindo na rede. As distâncias entre palavras, em termos de arestas, podem mostrar quais temáticas estão mais ou menos relacionadas nas publicações do periódico. Os valores da *incidência-fidelidade* podem auxiliar pesquisas futuras sobre quais temáticas são mais ou menos relevantes para o periódico, na difusão do conhecimento.

Na próxima seção, serão discutidos os resultados apresentados acima.

6.2.1 *Discussões para o* incidência-fidelidade *de Aguiar (2009)*

Como foi visto, a partir do método proposto, os resultados obtidos respondem à questão 1, do início do capítulo. Para o periódico AFE, entretanto, pode-se perceber que a medida que sua rede é filtrada, ela oferece uma certa resistência em apresentar um ponto crítico. Desse modo, uma faixa de valores de IF é responsável pelos altos valores de $\langle \ell \rangle$, e não um só valor, como nos demais periódicos que apresentam rede crítica.

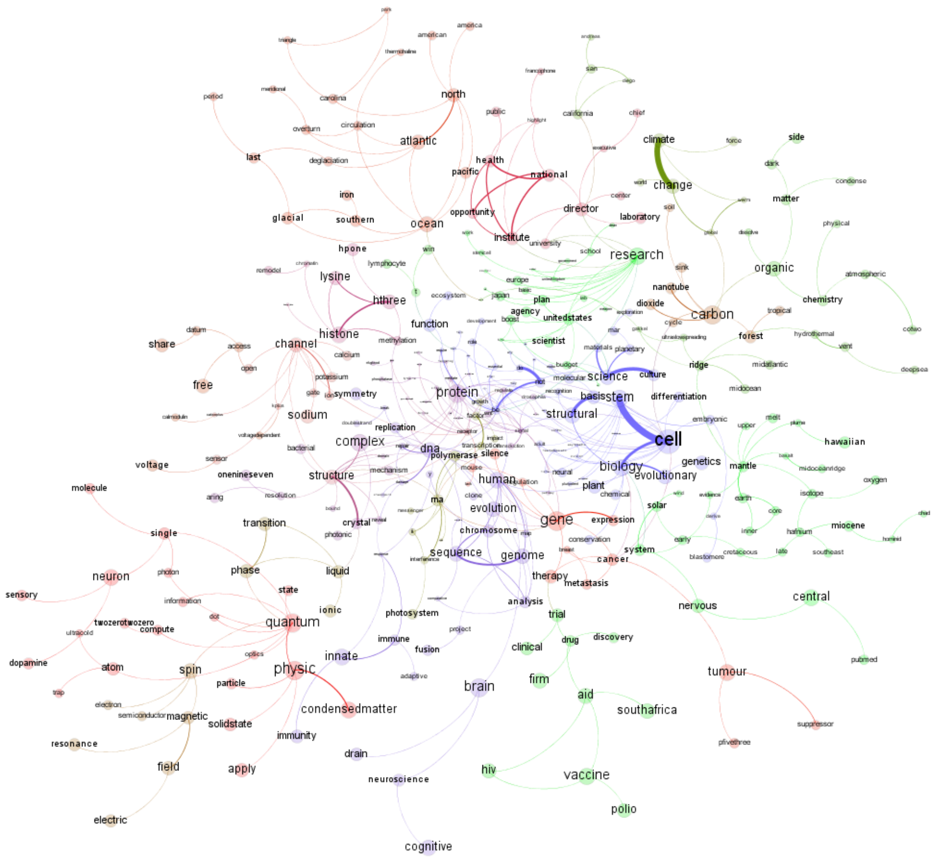


Figura 6.5: Rede crítica da Rede de palavras da revista *Nature* (maior componente). A Espessura das arestas indicam proporcionalmente o valor da *incidência-fidelidade* dos respectivos pares de palavras. A rede possui um diâmetro muito alto em comparação com as demais e cerca de 4 vezes o diâmetro que possui na rede canônica. É possível observar temáticas específicas em determinadas regiões da rede e por qual caminho outra temática surge neste emaranhado de palavras.

Para responder a questão 2 do início do capítulo, é preciso definir sobre quais parâmetros de redes se deseja diferenciar periódicos. As seções que seguem buscam agrupar periódicos a partir de gráficos que envolvam dois índices diferentes e os 15 periódicos.

6.2.1.1 Diâmetro e caminho mínimo médio

É interessante observar que os periódicos na rede *canonica* apresentam em média $D \simeq 5$. Para uma rede que possui *diâmetro* igual a 5, significa que pelo menos duas palavras dela possuem distância geodésica, em termos de arestas, no máximo 5. Ou seja, $\ell_{MAXIMO} \leq 5$ entre quaisquer dois vértices da rede. As redes em questão são formadas por união de cliques. Isto significa, por exemplo, que dois vértices que possuem $\ell = 5$ pertencem uma subrede de 5 cliques de configuração em formato de linha, ou próximo disto[4].

[4]Segundo Fadigas e Pereira (2013) esta configuração maximiza o *diâmetro*.

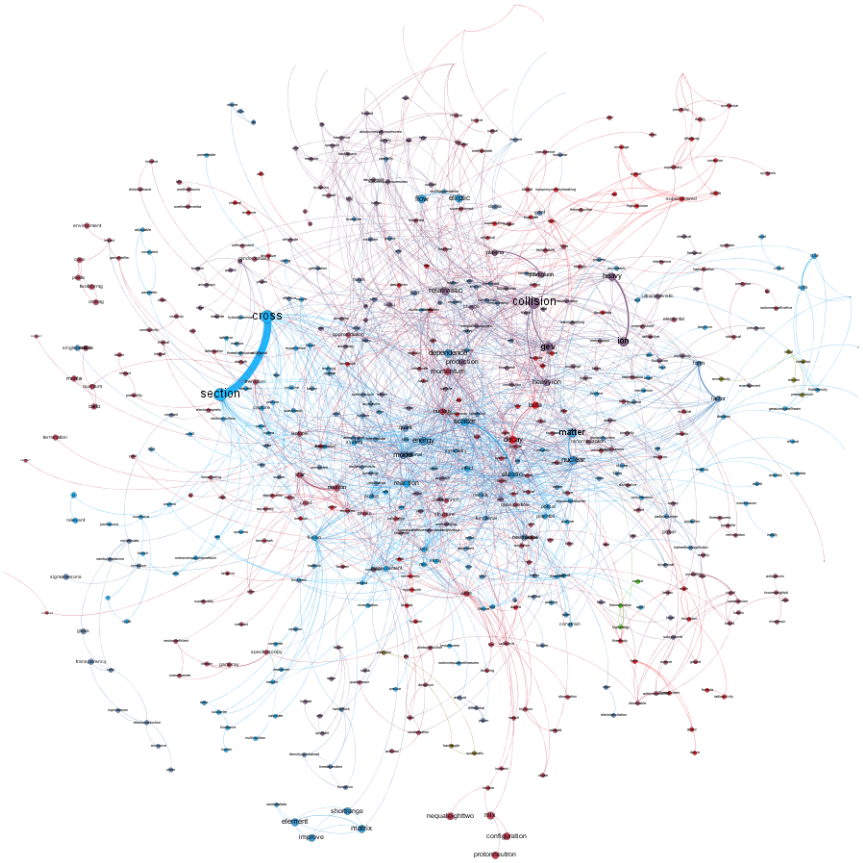


Figura 6.6: Rede crítica da Rede de palavras da revista *PRC*. A Espessura das arestas indicam proporcionalmente o valor da *incidência-fidelidade* dos respectivos pares de palavras. Seu diâmetro é cerca do dobro que quando na rede canônica e metade do diâmetro da rede crítica da Nature, Figura 6.5.

De acordo com a hipótese de Caldeira (2005), inferiu-se na Seção 4.5.5, que um título encerra uma ideia ou menor unidade de significado do conjunto de títulos da revista. Portanto, para as redes canônicas - em que praticamente toda a rede está conectada - dois títulos quaisquer de um dado periódico estão relacionados por no máximo três outros entre eles.

A rede crítica, entretanto, exibe os pares de palavras mais fieis dos conjuntos de títulos. É como se as menores unidades de significado pudessem ser reduzidas a pares de palavras que possuem muita relevância para o discurso. Como foi visto na Tabela 6.4, os *diâmetros* destas redes são muito altos, chegando até aproximadamente 4 vezes o tamanho da rede original (*Nature* e *Science*).

O *caminho mínimo médio* em redes críticas é, em geral, mais alto que em redes canônicas (aproximadamente 2 vezes maior, em média). A eliminação de palavras com IF muito baixo das redes faz o valor de $\langle \ell \rangle$ aumentar até um valor máximo, na rede crítica.

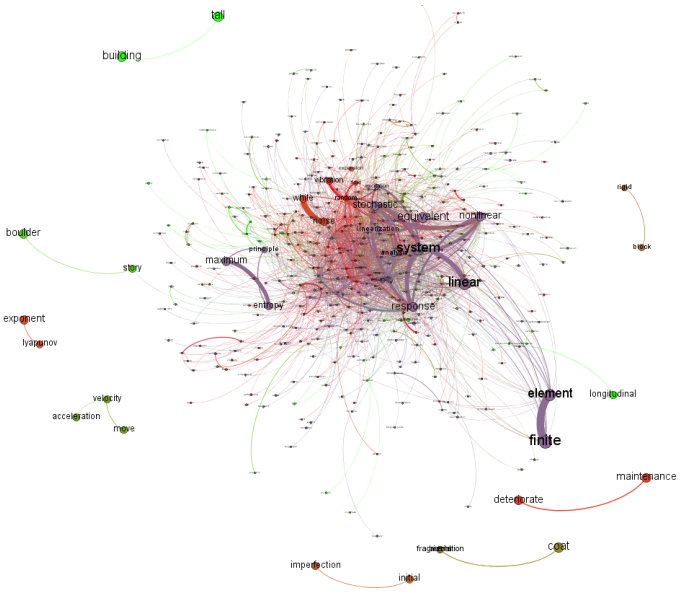


Figura 6.7: Rede crítica da Rede de palavras da revista *PEM*. A Espessura das arestas indicam proporcionalmente o valor da *incidência-fidelidade* dos respectivos pares de palavras. O diâmetro desta rede é aproximadamente o dobro que tinha na rede canônica e já não é mais o menor diâmetro das redes comparadas aqui.

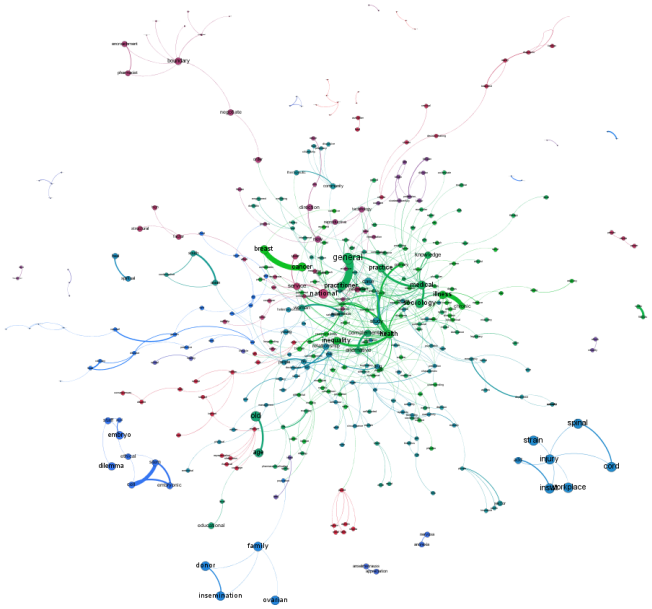


Figura 6.8: Rede crítica da Rede de palavras da revista *SHI*. A Espessura das arestas indicam proporcionalmente o valor da *incidência-fidelidade* dos respectivos pares de palavras.

Na Seção 6.2.2 será apresentado um ranking dos valores de $\langle \ell \rangle$ na rede crítica.

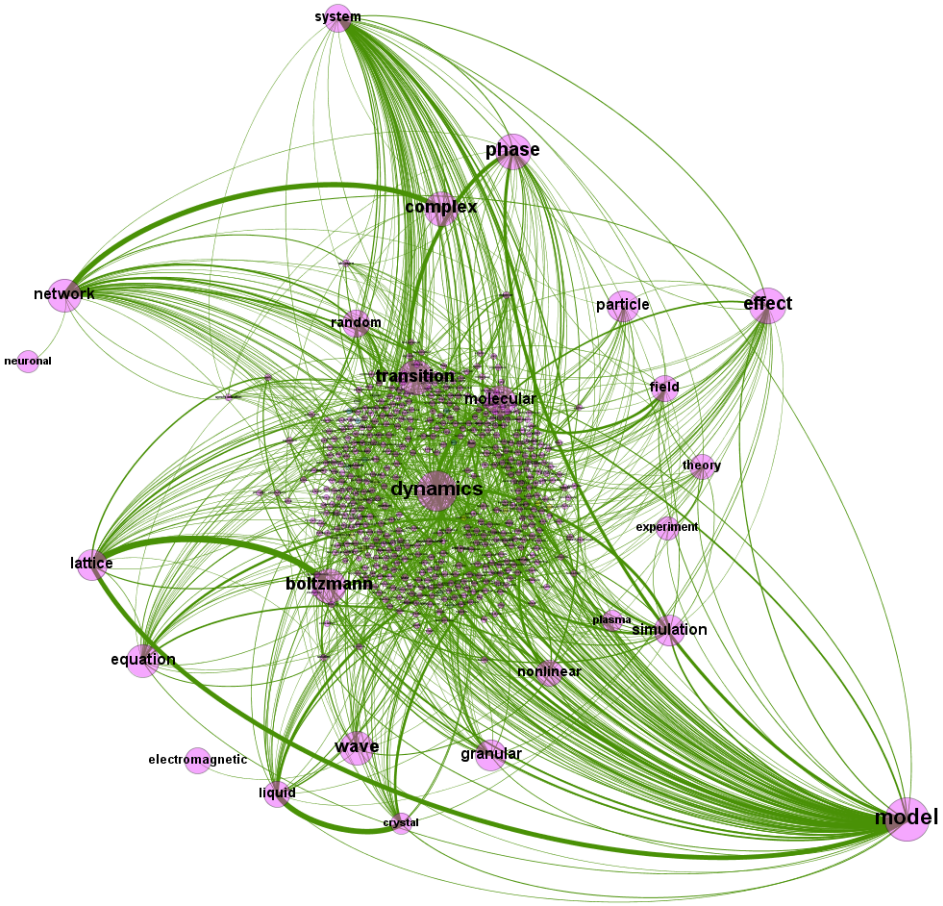


Figura 6.9: Rede crítica da Rede de palavras da revista *PRE*. A Espessura das arestas indicam proporcionalmente o valor da *incidência-fidelidade* dos respectivos pares de palavras.

6.2.1.2 Coeficiente de aglomeração médio

A partir da Tabela 6.3 e da definição de Watts e Strogatz (1998) para redes de mundo pequeno, pode-se inferir que as redes *canônicas* exibem o fenômeno *small-world*, por apresentarem coeficiente de aglomeração altos ($\langle C \rangle = 0.42 \pm 0.05$), *caminhos mínimos médios* pequenos ($\langle \ell \rangle = 2.68 \pm 0.14$), além de alta conectividade entre os vértices (maior componente representa toda a rede) comparados com redes aleatórias similares.

Entretanto, as redes críticas não possuem alta aglomeração e portanto não podem ser consideradas redes de *mundo pequeno*, com exceção de PRE e PEM. Isto significa que a grande quantidade de atalhos entre palavras de diferentes títulos são provocadas por pares de palavras de baixa *incidência-fidelidade* para o conjunto de títulos.

6.2.1.3 Porcentagem do maior componente

A partir da Tabela 6.3, pode-se perceber que quanto mais conectada é uma *rede crítica*, menor é o valor do seu *diâmetro*. A Figura 6.10 ilustra este fenômeno.

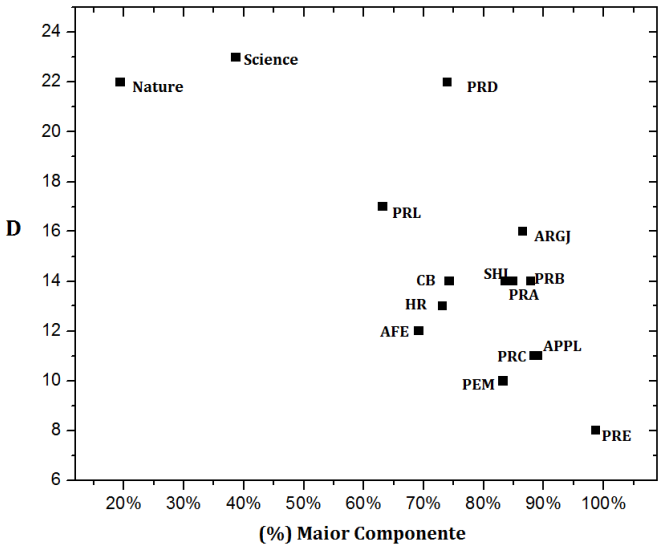


Figura 6.10: Porcentagem do maior componente em função do diâmetro das redes críticas. Dados da Tabela 6.4.

Este resultado pode nos dar indícios sobre a interdisciplinariedade de um periódico. Perceba que *Nature* e *Science* com baixa conectividade e alto *diâmetro* no topo, à esquerda, enquanto que PRC, $APPL$ e PEM com alta conectividade e baixo *diâmetro* estão no canto inferior direito, ou seja, grupos diametralmente opostos no gráfico.

6.2.1.4 Expoente γ e densidade

A Tabela 6.4 nos informa que $2 < \gamma < 3$, ou seja, as redes possuem uma topologia Livre de Escala (BARABASI; ALBERT; JEONG, 1999). A Figura 6.11 mostra os periódicos, em suas redes críticas, dispostos em um diagrama que relaciona *densidade* $\times$ expoente γ. A Figura 6.12, por sua vez, mostra dois periódicos em diagramas que relacionam $IF \times \gamma$.

A partir da análise dos gráficos da Figura 6.12[5], pode-se verificar que os periódicos apresentam densidades mínimas no ponto crítico. Este fenômeno pode ser explicado pela queda brusca de vértices quando IF_L torna-se maior que IF_C. Ou seja, a rede crítica se apresenta com o máximo de palavras fieis ao discurso, com o mínimo de conexões entre elas[6], por isto a densidade é mínima nesta configuração.

[5] Graficos semelhantes foram encontrados em todos os periódicos, exceto para AFE, que não possui um ponto de mínimo definido, mas sim uma larga região de densidades baixas.

[6] De maneira similar, o trabalho de Aguiar (2009) utilizou o critério da "diferença normalizada entre arestas e

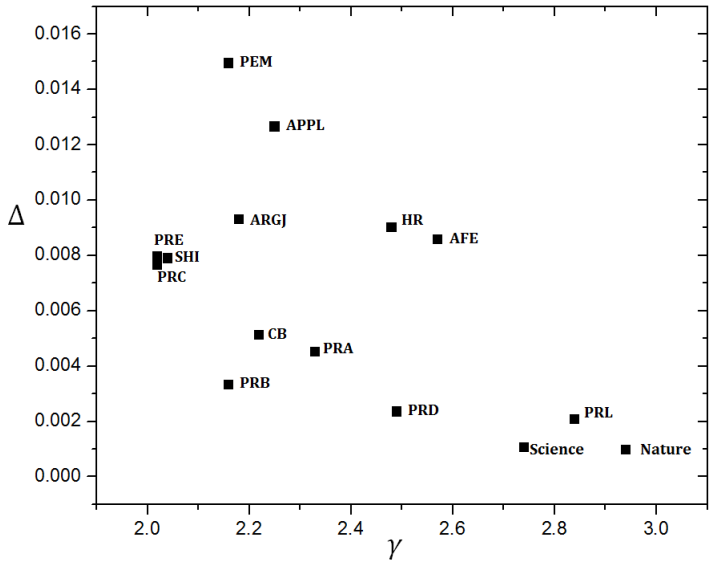


Figura 6.11: Densidade em função do expoente γ da distribuição de graus das redes críticas. Dados da Tabela 6.4.

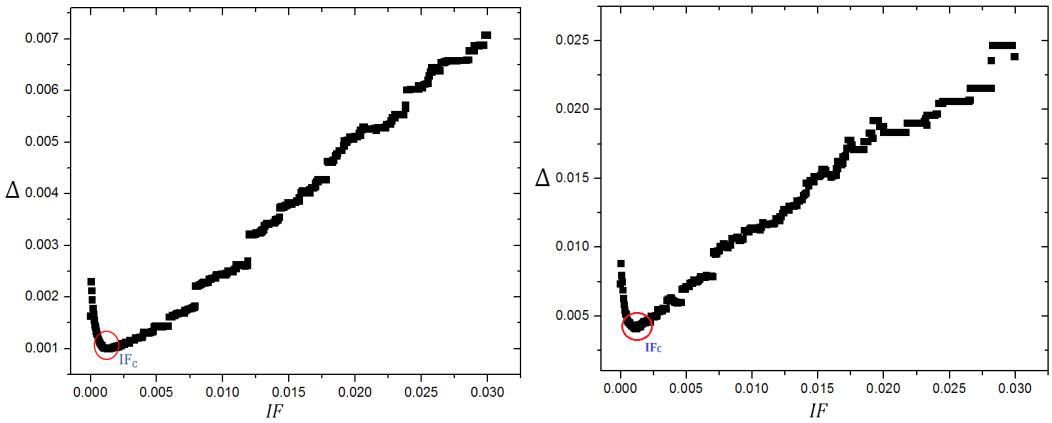


Figura 6.12: Densidade em função de IF_L para os periódicos *Science* (à esquerda) e *PRA* (à direita) da distribuição de graus das redes críticas. Dados da Tabela 6.4.

6.2.1.5 Grau médio

Nas redes críticas os valores de *grau médio* são muito menores que na rede canônica. Devido a remoção de arestas, as palavras conectam-se menos nesta configuração. A evidência de que estas redes são livres de escala mostra que existem alguns Hubs nestas redes. Para um trabalho futuro, uma análise qualitativa e visual poderia mostrar a relação destes Hubs com módulos destas redes. Por enquanto, pode-se analisar aqui a relação entre a *variação do grau médio* $\delta\langle k\rangle$ (Equação 6.3) com a *variação da densidade* $\delta\Delta$ (Equação 6.4) das redes críticas em relação às configurações iniciais (i.e. redes canônicas).

vértices" para determinar as incidências fidelidades críticas de redes de discursos escritos.

$$\delta\langle k\rangle = |\frac{\langle k\rangle_{critica} - \langle k_{canonica}\rangle}{\langle k_{canonica}\rangle}| \tag{6.3}$$

$$\delta\Delta = |\frac{\Delta_{critica} - \Delta_{canonica}}{\Delta_{canonica}}| \tag{6.4}$$

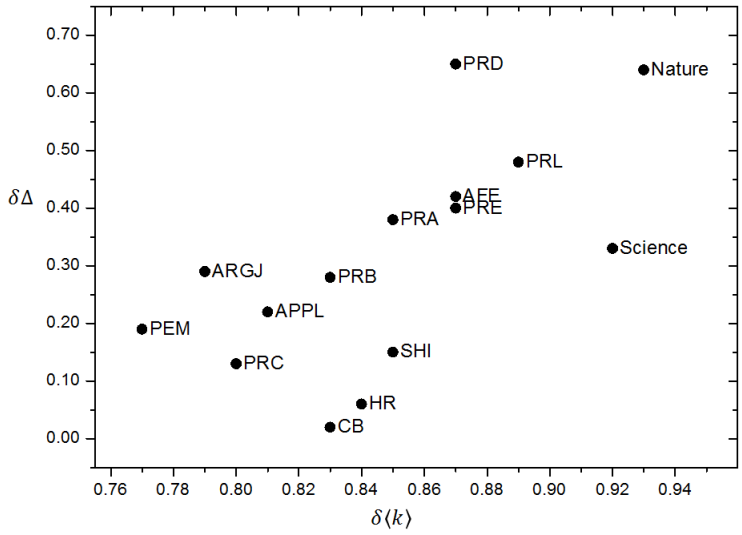


Figura 6.13: *Variação da densidade* em função da *variação do grau médio* das *redes críticas* em relação às *redes canônicas*.

6.2.2 Comparação com trabalhos anteriores

A Figura 6.14 mostra os valores das densidades e caminhos mínimos médios para as redes críticas, comparando-os com Pereira et al. (2011). Vale ressaltar que o período de publicações da coleta deste trabalho é maior que o trabalho comparado para algumas revistas.

A Tabela 6.5 mostra valores médios de alguns dos índices usados aqui e quais seus respectivos valores em discursos orais e escritos, inclusive em redes críticas.

6.3 Comentários finais

As duas questões propostas no início do capítulo foram respondidas de maneira satisfatória. Entretanto, devido a sua natureza interdisciplinar, o problema norteador, proposto na Introdução (Seção 1.4), está longe de ser resolvido, mas este trabalho contribui utilizando a modelagem de redes complexas.

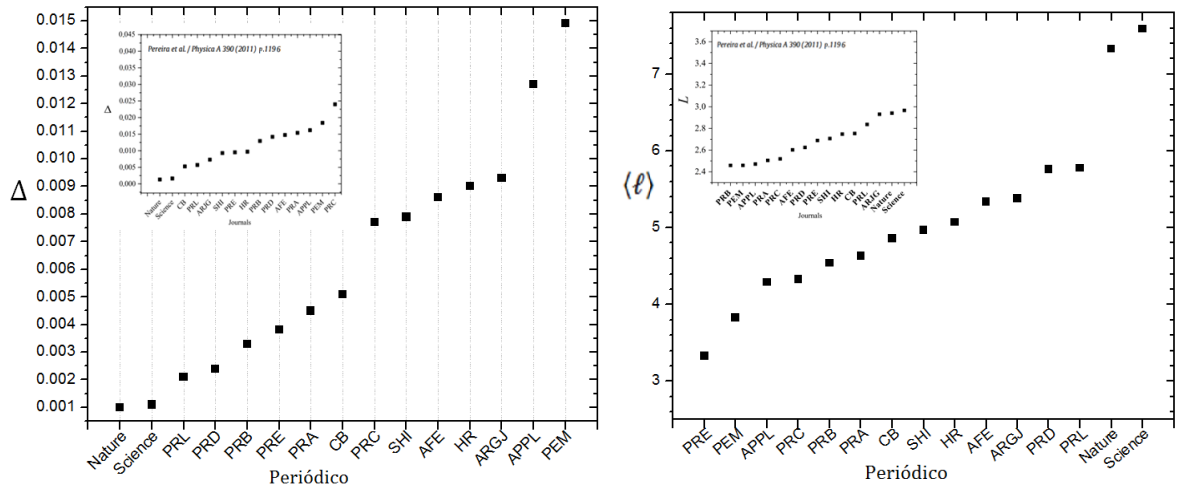


Figura 6.14: Ranking dos valores de δ e de $\langle \ell \rangle$ para os periódicos em suas redes críticas. Os gráficos menores representam os mesmos índices para o trabalho de Pereira et al. (2011). Dados da Tabela 6.4.

Rede sem Filtro	Textos Escritos 312 amostras (Caldeira 2005)	Textos Orais 12 amostras (Teixeira 2007)	Textos Escritos 50 amostras (Aguiar 2009)	Títulos 15 amostras (Este Trabalho 2013)
$< C >$	0.77	0.80	0.74	0.76 ± 0.03
$\langle \ell \rangle$	2.3	2.1	2.3	2.68 ± 0.14
D	5	4	5	5 ± 1
γ	1.6	1.7	1.8	2.5

Rede Crítica	(Teixeira et al 2007; 2010) $IF_C = 10^{-3}$	(Aguiar 2009) $IF_C = 4,4.10^{-4}$	(Este Trabalho 2013) $2.10^{-4} \leq IF_C \leq 1,2.10^{-2}$
$< C >$	0.58	0.20	0.42 ± 0.05
$\langle \ell \rangle$	3.8	4.0	4.91 ± 1.08
D	16	14	15 ± 5
γ	2.59	1.7	2.36

Tabela 6.5: Alguns índices de redes deste trabalho e de trabalhos anteriores. Na esquerda, os valores de índices em redes que não sofreram o processo da *incidência-fidelidade*. Na direita, encontra-se trabalhos em que foi considerado o fenômeno da *rede critica*. As duas últimas colunas se referem ao índice proposto por Aguiar (2009) e a primeira se refere ao índice proposto por Teixeira et al. (2010).

O uso do índice *incidência-fidelidade* em redes de títulos permitiu filtrar redes (a partir das frequências de pares de palavras nos títulos das revistas), até esta chegar em uma configuração crítica, o que responde a questão 1 proposta no iníncio do capítulo. Por outro lado, as respostas da questão 2 mostram que as topologias das redes críticas podem indicar padrões nas conexões das palavras das redes e as análises dos índices podem fornecer sugestões de diferenciar periódicos.

A Seção 6.2 demonstrou que é possível encontrar rede crítica para a maioria das redes de títulos. Esta rede apresenta o máximo de informações com o mínimo de resíduos e pode indicar algum padrão intrínseco da linguagem humana. Por exemplo, a maneira como um grupo de cientistas escolhem palavras para compor um ou mais títulos de seus trabalhos para submissão em um periódico e também como estes trabalhos são escolhidos pelo cientistas avaliadores do periódico podem ser responsáveis pelo surgimento do fenômeno da rede crítica.

Para mostrar que isso pode ser intrinseco da linguagem humana, foi feito o seguinte experimento: Para um certo conjunto de títulos que apresentou o fenômeno da rede crítica, o tamanho de cada título é mantido (número de palavras em cada título) e as palavras deste cojunto são redistribuidas aleatoriamente nos títulos. Isto produz um texto embaralhado.

A rede de cliques, gerada a partir do texto embaralhado, não apresenta o fenômeno da rede crítica. Os gráficos da Figura 6.15 mostram como varia *caminho mínimo médio* em função de *incidência-fidelidade limite* as redes de títulos com palavras embaralhadas para os periódicos *Science* e *Chemistry and Biology*. Observe que não há um ponto crítico, como foi observado nestes mesmos periódicos sem o processo da embaralhamento.

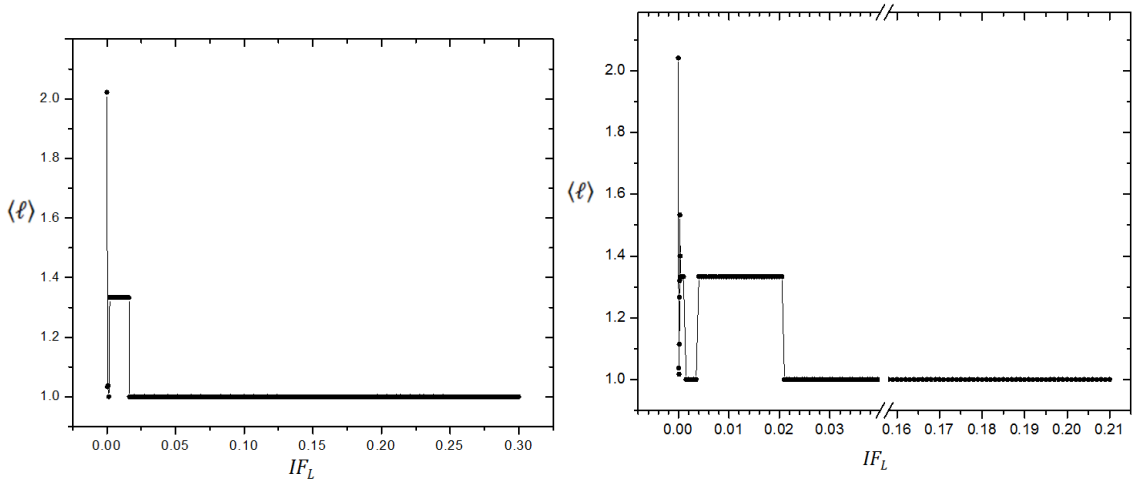


Figura 6.15: Não existem pontos críticos para redes de títulos com palavras embaralhadas. À esquerda o periódico *Chemistry and Biology*. À direita, o periódico *Science*.

A Tabela 6.16 mostra os valores dos índices de redes para alguns valores de IF_L. Pode-se observar que a primeira filtragem ($IF_L = 5 \cdot 10^{-5}$) faz a rede se desmontar, mesmo com IF_L muito baixo. A rede nessa configuração, apesar de conter 499 vértices, apresenta poucas arestas, configurando-se em apenas pares de palavras isolados (Figura 6.17).

Neste processo de filtragem a rede não apresentou o fenômeno da perda de atalhos, comum quando se aumenta IF_L, a partir da rede canônica. Dessa forma, o valor de $\langle \ell \rangle$ caiu bruscamente, ao invés de aumentar para depois ter a queda brusca. Fica evidente com este experimento que a o fenômeno da rede crítica pode ser algo intrínseco à rede de títulos.

A Seção 6.2 apresentou também os valores de índices de redes para redes críticas dos 15 periódicos. Estas redes são diferentes das redes contruídas a partir do *incidência-fidelidade* proposto por Teixeira et al. (2010). Entretanto, com este método é possível diferenciar periódicos a partir da rede crítica, já que a quantidade de títulos não precisa ser igual para

IF	n	m	$\langle \ell \rangle$	$\langle C \rangle$	$\langle k \rangle$	Δ
0.00000	15039	261405	2.04	0.34	34.77	0.0023
0.00005	499	263	1.04	0.01	1.05	0.0021
0.00010	221	116	1.02	0.02	1.05	0.0047
0.00015	28	18	1.11	0.10	1.28	0.0476
0.00020	12	10	1.27	0.24	1.67	0.1515
0.00025	10	9	1.32	0.29	1.80	0.2000
0.00030	6	7	1.53	0.478	2.33	0.4667
0.00035	5	6	1.40	0.60	2.40	0.6000
0.00040	4	4	1.33	0.58	2.00	0.6667
0.00100	4	4	1.33	0.58	2.00	0.6667
0.00150	3	3	1.00	1.00	2.00	1.0000
0.00350	3	3	1.00	1.00	2.00	1.0000
0.00350	3	3	1.00	1.00	2.00	1.0000

Figura 6.16: Índices de redes para cada valor de IF_L para rede de títulos embaralhados da *Science*.

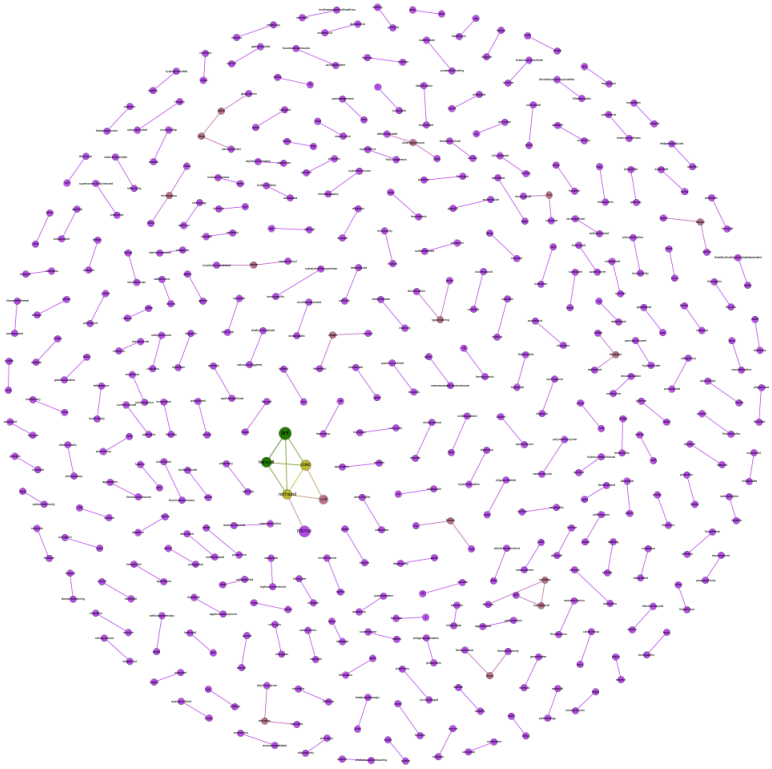


Figura 6.17: Rede para $IF_L = 5 \cdot 10^{-5}$ para rede de títulos embaralhados do periódico *Science*.

todos os periódicos. Ainda na Seção 6.2, foram apresentadas várias formas de diferenciar periódicos a partir da rede crítica, seus índices de redes e gráficos que os relacionam.

Por outro lado, a Seção 6.1 utilizou o índice *incidência-fidelidade* proposto por Teixeira et al. (2010) para demonstrar que existe rede crítica para um conjunto qualquer de 300 títulos de um periódico. Além disso, o valor da *incidência-fidelidade crítica* da grande

maioria das redes destes grupos tem o mesmo valor que a de discursos orais do trabalho de Teixeira et al. (2010), ou seja, $IF_C = 10^{-3}$. Entretanto, os índices de redes na configuração crítica para um dado periódico, variam muito a depender do grupo de 300 títulos que se observa, o que dificulta qualquer tentativa de diferenciar periódicos.

Cada grupo de 300 títulos de um dado periódico pode conter títulos de diferentes épocas. Por exemplo, escolhendo-se ao acaso 1 dentre os 100 grupos de 300 títulos do periódico *Chemistry and Biology*, existe uma probabilidade extremamente baixa de se obter todos os títulos pertencentes a uma mesma época[7], já que existem $\binom{1643}{300}$ grupos diferentes de 300 títulos. Entretanto, para o periódico AFE esta probabilidade não é tão baixa, pois existem $\binom{370}{300}$ grupos diferentes de 300 títulos. Os gráficos da Figura 6.2 evidenciam esta diferença a partir da faixa de valores de *densidade* de cada um destes periódicos.

Este resultado pode indicar dependência temporal entre os títulos dos periódicos. Se não fosse assim, qualquer combinação de 300 títulos de um dado periódico iria ter valores de Δ muito próximos, por exemplo. Em outras palavras, o surgimento de publicações em uma revista pode estar relacionado com publicações anteriores na mesma revista

O Capítulo 7 apresenta resultados de comparações entre grupos de títulos do periódico *Nature*, só que de épocas diferentes. Nesta abordagem, explicada na Seção 5.6.1, cada grupo de títulos pertencem à mesma janela de observação, que corresponde à 8 semanas. E esta janela varia no tempo, com isso pode-se perceber tendências no padrão de relacionamentos das palavras dos títulos da revista ao longo do tempo.

[7] O periódico *Chemistry and Biology* publica cerca de 100 artigos por ano. É pouco provável que um dos grupos de 300 títulos tenha exatamente três anos seguidos de publicações.

Capítulo Sete

Resultados envolvendo TVG

Como foi explicado na Seção 5.6.1, a rede da revista Nature pôde ser dividida em uma sequência de janelas, onde cada uma representa uma época composta por títulos referentes a 8 semanas de publicações. A análise desta sequência de redes mostra a evolução no tempo para o padrão de conexão das palavras publicadas na revista, a fim de diferenciar épocas distintas do mesmo periódicos a partir de suas redes.

A diferença desta abordagem para a abordagem proposta na Seção 5.5.1 - onde se analisa um periódico a partir de 300 títulos aleatórios - é que na abordagem com TVG, cada grupo contém títulos pertencentes a uma mesma época. A partir dos resultados vistos no capítulo anterior, Seção 6.1, pôde-se perceber que grupos distintos de títulos retirados de maneira aleatória de um periódico podem gerar redes completamente diferentes. Isto é um forte indício de que o surgimento de publicações em uma dada época depende de publicações anteriores na mesma revista, no que concerne às ideias expressas nos títulos dos seus artigos.

Dessa forma, cabe aqui, neste capítulo, responder às seguintes questões (adaptadas dos objetivos específicos - Introdução, Seção 1.6):

1. As redes de diferentes épocas do periódico em questão exibem propriedades topológicas que permitam diferenciá-las? É possível verificar tendências no padrão de conexões destas redes, ao longo do tempo?;
2. Pode-se admitir que a configuração de uma rede de títulos de uma certa época tem correlação com as configurações de redes de títulos em épocas passadas?;
3. Qual a influência do vocabulário da revista nos comportamentos observados das questões anteriores?.

7.1 Respostas para a questão 1

A resposta da questão 1, proposta no início deste capítulo, foi obtida a partir da comparação dos valores dos índices de redes nas janelas do TVG. Estes índices são denominados *atemporais*[1], que são os índices clássicos de redes e os índices específicos para redes

[1]Como a modelagem do problema se deu com o TVG como um conjunto de grafos estáticos, os índices de redes complexas são satisfatórios para esta análise, não sendo relevante aqui o uso de índices *temporais* em TVG (e.g jornada).

de cliques.

A Figura 7.1 exibe os valores dos indices clássicos (eixo das ordenadas), em função do tempo, dado em semanas (eixo das abscissas). Cada ponto no gráfico representa a rede de uma janela $\tau_i = [t_i, t_{i+7}]$ de oito semanas de publicações, que se inicia a partir da semana i, correspondente abscissa do ponto. Vale lembrar que as primeiras janela do TVG ($\tau_1 = [t_1, t_8]$) se inicia em $t = 1$, 1º de janeiro de 1999. A partir destas séries temporais é possível verificar tendências de comportamento da revista em determinadas épocas.

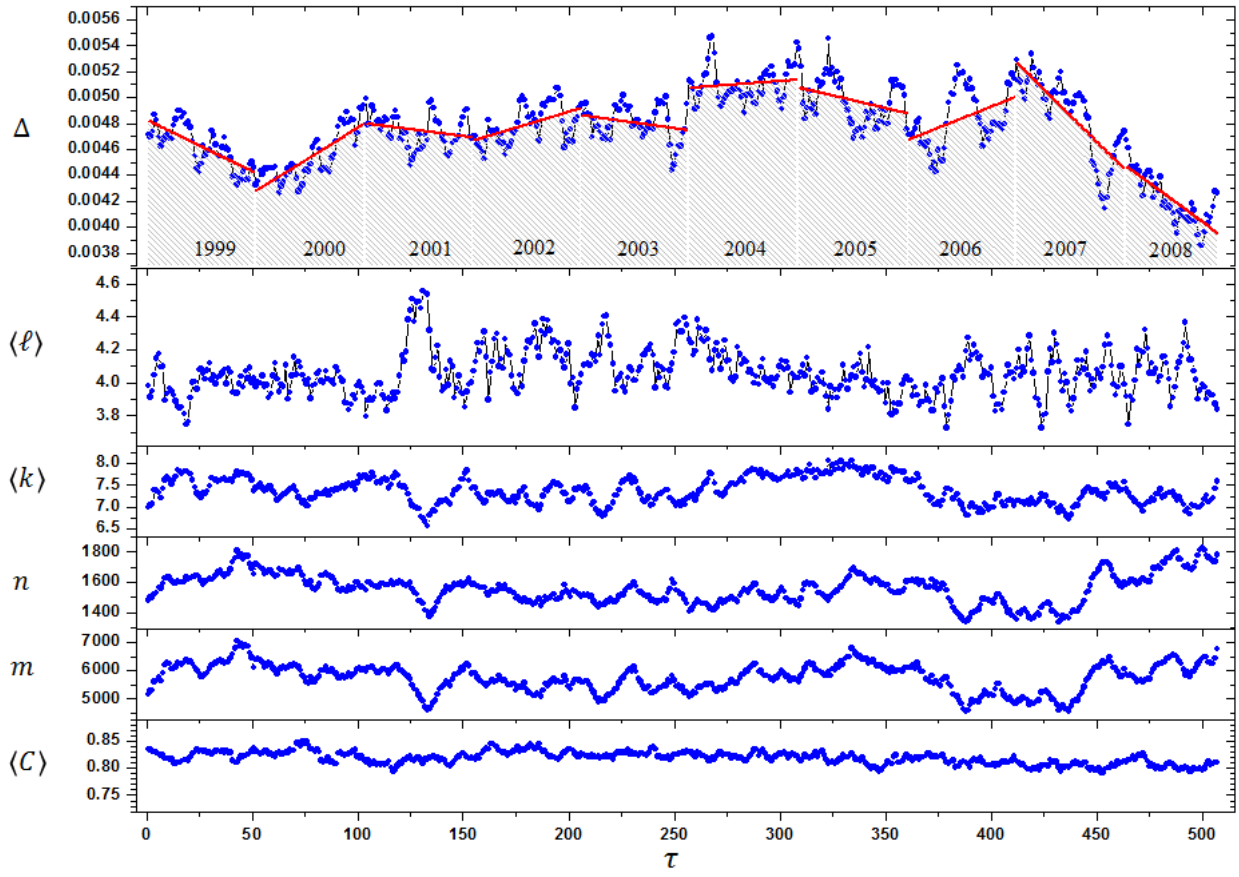


Figura 7.1: Evolução dos índices das janelas temporais entre 1999 e 2008 para a revista Nature. As linhas retas no gráfico de Δ representam o melhor ajuste linear para janelas de 1 ano. Fonte: Cunha et al. (2013).

7.1.1 Discussões que envolvem Índices Clássicos

A *densidade* Δ, é relativamente baixa (comparada a outras redes semânticas de cliques) para redes com base em títulos de trabalhos científicos (FADIGAS; PEREIRA, 2013). Entretanto, é possível verificar tendências de crescimento, decrescimento e constância em torno dos valores de densidade ao longo do tempo, nas linhas de ajuste linear, no gráfico

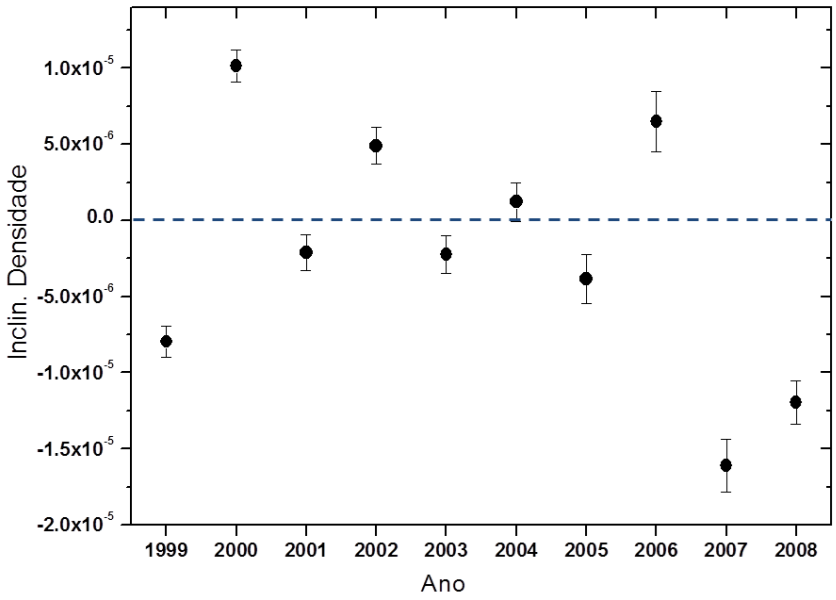


Figura 7.2: Valores das inclinações do gráfico de Δ por ano.

da Figura 7.1.

Os valores das inclinações das retas de ajuste podem ser obtidos no gráfico da Figura 7.2. Percebe-se, de acordo com estas inclinações, que em média:

- Em 2007 e 2008 a densidade evolui semelhante à 1999. Ou seja, a rede tende a ser mais esparsa;
- Em 2000 e 2006 ocorre o contrário, a rede tende a ser mais densa;
- De 2001 à 2005 a rede tende a manter a densidade de suas relações entre seus vértices, já que suas inclinações são próximas de zero.

Pode-se reescrever a expressão da densidade de uma rede não dirigida para em função do tempo $t \in \Gamma$ (Eq. 7.1), assim:

$$\Delta_t = \frac{m_t}{\frac{n_t(n_t-1)}{2}} = \frac{2\langle k\rangle_t}{(n_t-1)} \simeq \frac{2\langle k\rangle_t}{n_t} \tag{7.1}$$

Sabe-se que para todos os subgrafos do TVG em questão $n \gg 1$, por isso foi usado a aproximação $n_t - 1 \simeq n_t$. De acordo com a Equação 7.1, para épocas em que as redes

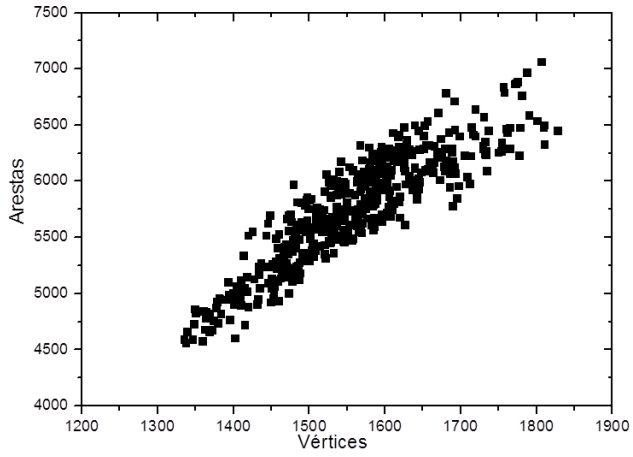


Figura 7.3: O número de vértices cresce em média proporcionalmente ao número de arestas. Fonte Cunha et al. (2013).

tiveram o mesmo valor de grau médio $\langle k \rangle_t$, a densidade é apenas efeito do tamanho da rede. Neste caso, quanto maior o número de vértices n_t menor será o valor da densidade Δ_t.

Para exemplificar, considere as janelas τ_{257}, τ_{372} e τ_{419}. Como se pode ver no gráfico da Figura 7.1 e na Tabela 7.2, as redes para estes instantes possuem o mesmo valor de grau médio, ou seja, para estes três momentos da história de publicação da Nature, os "relacionamentos"entre os vértices (o grau médio) em média se mantiveram, mas o "poder de relacionamento"das redes (a densidade) inicialmente diminuiu e em seguida aumentou. Isto se deve ao fato de que inicialmente a diversidade de palavras da janela (número de vértices) aumentou e em seguida diminuiu. De maneira análoga podemos comparar a janela τ_{18} com a janela τ_{445}. Percebe-se que o poder de relacionamento destas redes é mesmo. Houve uma redução no vocabulário e no "relacionamento médio"entre as palavras dos títulos (Tabela 7.2).

O comportamento dos gráficos de vértices e arestas no tempo (Figura 7.1) são similares, quase que sobrepostos, respeitando as diferenças entre escalas. Isto sugere que, em algumas épocas, n e m são em média proporcionais, ou seja:

$$m_t \propto n_t \tag{7.2}$$

Este resultado nos mostra que em vários períodos de tempo, maiores que τ_i, a proporção em entre vértices e arestas, ou seja, o grau médio $\langle k \rangle_t$ das redes com o passar do tempo, em média, se mantém constante. Este mesmo resultado pode ser entendido observando o gráfico da Figura 7.3.

Neste caso, pode-se rearrumar a Equação 7.1 da densidade para:

$$\Delta_t \cong \frac{2\langle \overline{k} \rangle_{\tau_k}}{n_t} \tag{7.3}$$

Ou seja, já que o grau médio das redes das janelas variaram muito pouco, para alguns períodos de tempo $\tau_k = [t_k, t_z] > \tau_i$, a densidade Δ_t das janelas τ_i é, em média, função exclusiva do numero de vértices da rede de cada janela, sendo $(t_k, t_z) \in \Gamma$.

7.2 Respostas para a questão 2

Para responder a questão 2, proposta no início do capítulo, foi realizado o teste de normalidade proposto por Shapiro e Wilk (1965) e calculado o *expoente de Hurst* para as séries dos índices de redes em diferentes épocas, a partir do *método DFA* (*Detrended Fluctuation Analysis*) proposto por Peng et al. (1994). O primeiro procedimento busca verificar a existência de normalidade nas distribuições dos valores dos índices. O segundo, verifica se existe correlação temporal entre os índices de redes ao longo do tempo.

A Figura 7.4 apresenta as distribuições de valores de índices clássicos[2], bem como as curvas de gauss de melhor ajuste.

Percebe-se pela Figura 7.4 que as curvas se ajustam bem à curvas de Gauss.Entretanto, a Tabela 7.1 revela que a distribuição dos índices não segue uma distribuição normal, já que $(p\text{-}valores) < \alpha$.

Teste de Normalidade	Δ	$\langle k \rangle$	n	m	D	$\langle C \rangle$	$\langle \ell \rangle$
W	0.99	0.98	0.99	0.99	0.88	0.99	0.98
$p-valor$	0.00006	0.00003	0.00387	0.00020	0.00000	0.00499	0.00000
Método DFA	Δ	$\langle k \rangle$	n	m	D	$\langle C \rangle$	$\langle \ell \rangle$
$\mathcal{H}$	0.661	0.640	0.783	0.767	-	0.550	0.499
erro	0.013	0.006	0.007	0.006	-	0.008	0.016
R^2(ajuste)	0.994	0.999	0.999	0.999		0.998	0.985

Tabela 7.1: Teste de Normalidade: Valores da estatística de Shapiro-Wilk, com o nível de confiança de $\alpha = 0.05$. *Método DFA*: *expoente de Hurst* $\mathcal{H}$ das séries temporais dos índices de redes, erros associados e R^2 dos ajustes.

Para os resultados do *método DFA* no TVG, os *expoentes de Hurst* na Tabela 7.1 nos permite identificar que, com exceção do *Diâmetro* (não foi identificada auto-afinidade em

[2]Optou-se por não apresentar a distribuição do Diâmetro ($\langle D \rangle = 10 \pm 1$), já que não foi possível um bom ajuste para curva de Gauss em seus pontos ($R^2 = 0.45$).

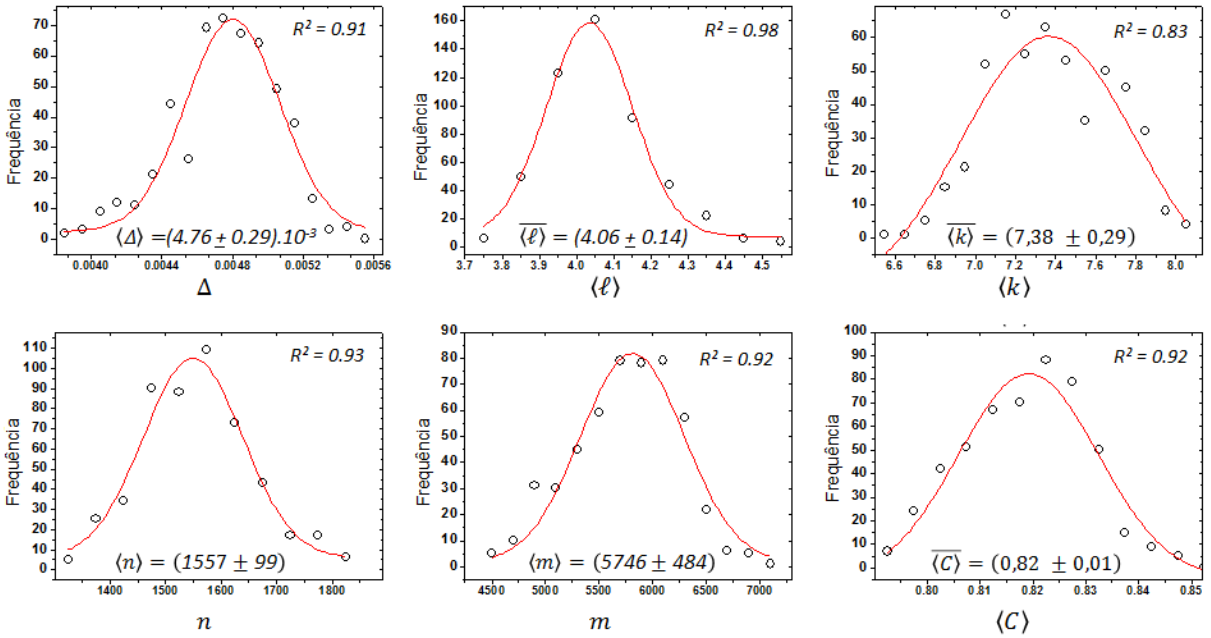


Figura 7.4: Distribuição de frequências para os índices durante o intervalo de 1999 a 2008. As linhas representam curvas de Gauss de melhor ajuste. Em cada gráfico o valor médio do respectivo índice, sua incerteza e o valor de R^2 para o ajuste da curva.

sua série temporal), todos os índices de redes têm um padrão de auto-afinidade, com correlação temporal persistente.

7.3 *Discussões sobre o* expoente de Hurst

O gráfico da Figura 7.5 revela o período comum para as correlações das séries, $t = [4, 21]$, que é o intervalo de 4 a 21 semanas[3].

O único índice que possui inclinação $\mathcal{H} \simeq 0.5$ é o *caminho mínimo médio* ($\langle \ell \rangle$). Isto significa, que para o intervalo de tempo considerado, apesar de correlacionado, sua série temporal não possui "memória"e segue uma caminhada aleatória. Entretanto, as outras quantidades possuem memória para este mesmo intervalo de tempo. Os índices n e m destacam-se pelos altos valores de $\mathcal{H}$.

Sabe-se que n representa o número de vértices, ou seja, o número de palavras diferentes na rede (vocabulário da rede). E m representa a quantidade de relacionamentos que as palavras deste vocabulário fazem. Assim, para uma dada época (ou janela do TVG), se o vocabulário aumentou, existe uma forte tendência dele continuar aumentando de 4 a 21 semanas depois.

[3] A partir do gráfico da Figura 7.5, $log(4) = 1,2$ e $log(21) = 1,3$.

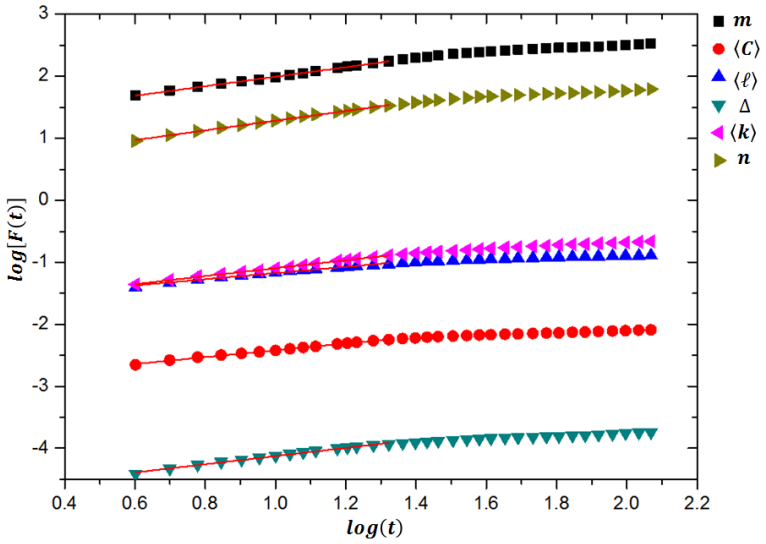


Figura 7.5: Valores do logaritmo da função F_{DFA} em função do logaritmo do tempo, dado em semanas. Fonte: Cunha et al. (2013).

7.4 Resposta para a questão 3

A Tabela 7.2 mostra os valores dos índices[4] de redes de cliques para algumas janelas do TVG e o confronto deles com alguns índices clássicos.

Os índices para rede de cliques são capazes de caracterizar, quanto às topologias, as redes de diferentes épocas do TVG. Dessa forma, analisar os índices de redes das janelas do TVG ajuda a responder a questão 1, proposta no início deste capítulo. Além disso, eles dão informações de como se comporta o vocabulário das janelas ao longo do tempo (o que atende à questão 3). A análise dos valores da Tabela 7.2 permite-nos observar que:

- A linha que apresenta (n_0/n_q) mostra que, em média, aproximadamente 50% dos vértices das janelas apresentadas constituem palavras que se repetem;
- Para todo o TVG $D_{ref445} \leq D_{ref} \leq D_{ref106}$. Ou seja os extremos possuem o maior e o menor[5] D_{ref}, respectivamente. Ou seja, a rede da janela τ_{445} possui o maior diâmetro D por quantidade de títulos n_q. A janela τ_{25} é praticamente um bimestre a frente da janela τ_{18} no ano de 1999. Ainda assim, houve uma redução brusca no valor da densidade da janela. Isto se torna claro ao observar que existe um leve aumento no número de vértices e uma redução no valor do grau médio;
- A janela τ_{257}, apesar de ter um dos maiores valores de Δ, possui um valor pequeno para $v(\Delta)$e para $v(\langle k \rangle)$. Ou seja, em média, o relacionamento dos vértices e o poder

[4] A maior funcionalidade destes índices é ver o quão uma rede de cliques real difere de sua configuração inicial (i.e. estado inicial das cliques isoladas)

[5] Caso a rede de cliques fosse minimamente conectada, teria um layout tipo *Estrela*

Semana	*18*	*25*	*106*	*249*	*257*	*268*	*339*	*372*	*419*	*445*	*499*
Δ	0.0049	0.0044	0.0048	0.0044	0.0051	0.0055	0.0047	0.0045	0.0053	0.0049	0.0039
Δ_{q0}	0.0020	0.0021	0.0021	0.0022	0.0027	0.0026	0.0021	0.0022	0.0024	0.0023	0.0017
$\upsilon(\Delta)$	1.4541	1.1060	1.2963	1.0277	0.8809	1.1390	1.2238	1.0303	1.2280	1.1429	1.3159
$\langle k \rangle$	7.82	7.24	7.65	7.18	7.20	7.76	7.75	7.20	7.20	7.42	6.98
$\langle k_{q0} \rangle$	5.11	5.05	5.14	5.09	5.31	5.38	5.26	5.17	4.94	5.16	4.68
$\upsilon(\langle k \rangle)$	0.53	0.43	0.49	0.41	0.36	0.44	0.47	0.39	0.46	0.44	0.49
n_0	2558	2416	2466	2349	1958	2095	2489	2332	2061	2278	2810
n	1606	1637	1598	1619	1405	1420	1652	1617	1350	1523	1811
n_q	481	459	480	458	368	391	477	465	429	460	594
n_0/n	1.59	1.48	1.54	1.45	1.39	1.48	1.51	1.44	1.53	1.49	1.55
n_0/n_q	5.3	5.3	5.1	5.1	5.3	5.4	5.2	5.0	4.8	5.0	4.7
D	10	10	8	10	11	11	11	11	10	14	10
D_{ref}	0.29	0.30	0.25	0.30	0.33	0.32	0.31	0.31	0.30	0.36	0.28
$\langle C \rangle$	0.82	0.83	0.81	0.82	0.83	0.82	0.82	0.82	0.80	0.81	0.80
$\langle C \rangle_{rd}$	0.006	0.005	0.004	0.003	0.005	0.006	0.005	0.005	0.005	0.006	0.004
$\langle \ell \rangle$	3.81	4.08	3.91	4.11	4.26	4.08	4.06	3.83	4.13	4.16	3.95
$\langle \ell \rangle_{rd}$	3.99	3.99	3.84	3.98	3.91	3.73	3.85	3.95	3.89	3.82	4.05
$\%Cp_{maior}$	90.9%	90.3%	91.3%	89.9%	88.6%	94.3%	92.5%	85.4%	90.9%	91.4%	89.1%

Tabela 7.2: Índices de redes complexas e índices de rede de cliques para algumas janelas do TVG. Fonte: Cunha et al. (2013).

de relacionamento do estado inicial das cliques isoladas já eram, relativamente, altos antes das cliques se juntarem;

- As janelas τ_{372} e τ_{419} possuem o mesmo valor de $\langle k \rangle$. O aumento de Δ, de uma janela para outra, se deu pela redução de n. Isto reforça o que já foi discutido: Para este TVG o aumento no vocabulário dessas janelas de tempo, na maioria das vezes, torna a rede mais esparsa, já que as palavras, em média, relacionam-se da mesma forma;

- A janela τ_{499} possui o menor valor de Δ das redes que formam o TVG. Janelas próximas à ela representam a tendência mais atual[6] da revista, i.e. um vocabulário cada vez maior a medida que o tempo passa;

- De acordo com os resultados da Tabela 7.2, podemos inferir que todas[7] as redes exibem o fenômeno *small world* para o modelo de Watts e Strogatz (1998).

[6] Entende-se por "mais atual", nessa base de dados, as ultimas janelas deste TVG, ou seja, publicações no final do ano de 2008

[7] Para isso as redes precisam estar conectadas ou que tenham um componente com a maioria dos vértices da rede. Na tabela, $\%Cp_{Maior}$ representa a porcentagem do maior componente da rede

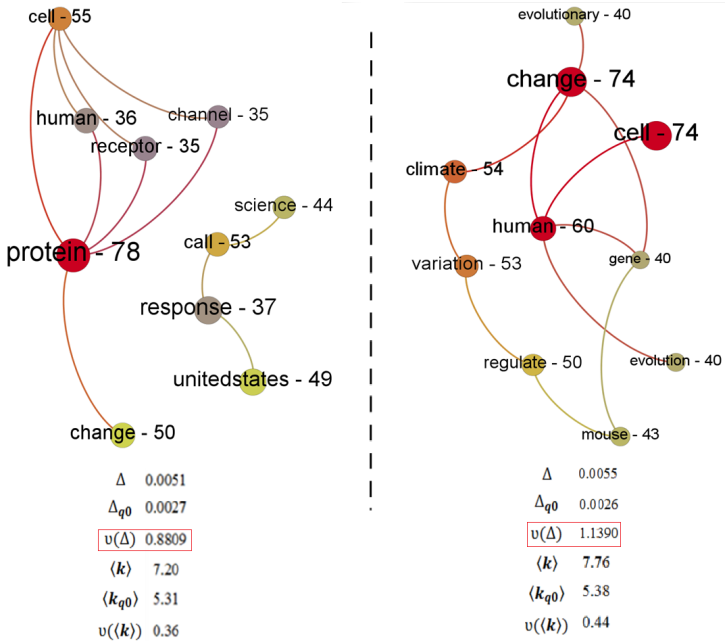


Figura 7.6: Cada Rede possui 10 palavras de maior grau na janela correspondente. Abaixo das redes, encontra-se os valores dos índices de redes de cliques. Em destaque, o índice $\upsilon(\Delta)$, que representa o adensamento em uma rede de cliques.

Para melhor responder à questão 3, foi feito um ranking das palavras de maior grau[8] em épocas diferentes.

As figuras 7.6 e 7.7 representam as redes de algumas janelas do tvg filtradas pelas palavras de maiores graus. As figuras destacam o índice $\upsilon(\Delta)$. Este indicador representa o fator de adensamento de uma rede de clique, ou seja, compara o quanto muda a densidade de uma rede de cliques real, para sua configuração inicial quando as cliques estão isoladas.

De acordo com a Figura 7.7 para estes três instantes de tempo, a partir de $t = 419$, a rede diminui seu adensamento e em seguida aumenta para um valor maior que o inicial. Parece que este indicador está relacionado com as palavras de maior frequência nas janelas. Afinal, quando $\upsilon(\Delta)$ cai, a rede das palavras de maior grau fica menos densa.

Ou seja, pode ser possível a partir dos valores de $\upsilon(\Delta)$ em alguma janela prevermos como as temáticas importantes da revista nesta época (janela) se relacionam. Por exemplo, as palavras "cell" e "science" na janela $t = 419$ se relacionam através do vértice comum "biology". Cada uma dessas três palavras funciona como se fosse um vértice central de uma rede de clique tipo estrela (oculta na figura) que representa uma dada temática.

Na Janela $t = 445$, estas duas palavras não se conectam mais, o que indica um certo "afastamento" entre elas nesta época. E isto está e acordo com o valor de $\upsilon(\Delta)$, que

[8] O grau de um vértice(palavra) mede o numero de relacionamentos(arestas) feitos por ele.

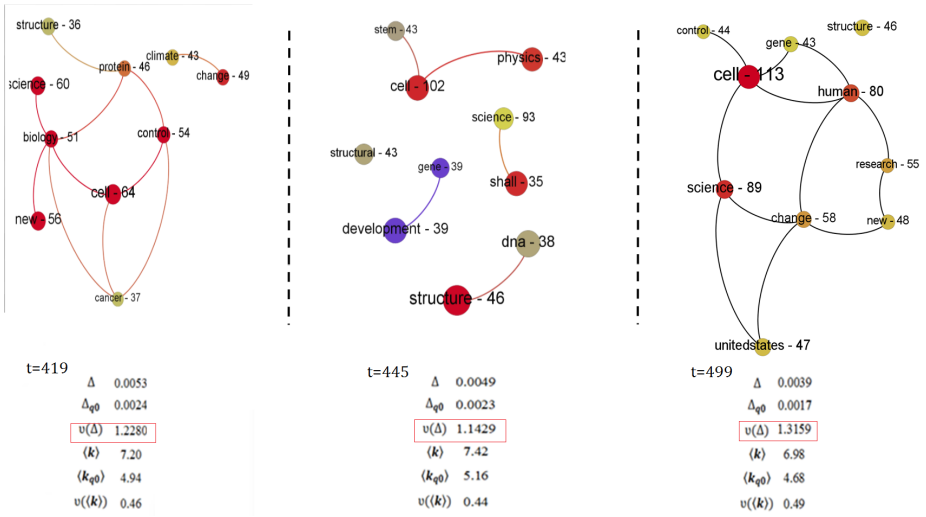


Figura 7.7: Cada Rede possui 10 palavras de maior grau na janela correspondente. Abaixo das redes, encontra-se os valores dos índices de redes de cliques. Em destaque, o índice $\upsilon(\Delta)$, que representa o adensamento em uma rede de cliques.

também caiu. Na Janela $t = 499$, o valor de $\upsilon(\Delta)$ aumenta e a rede de palavras de maior grau volta a ser bem conectada.

7.5 *Comentários Finais*

As questões propostas no início do capítulo estão longe de serem respondidas de maneira precisa, devido a natureza interdisciplinar do problema (questão norteadora, proposta na Introdução, Seção 1.4). Este trabalho sugeriu alguns caminhos para as respostas (e.g. redes, séries temporais) e possíveis interpretações dos resultados encontrados.

Os valores dos índices da janela do *TVG*, em geral, são muito próximos de seus valores médios, com baixo desvio padrão. Entretanto, os testes de normalidade apontam que as distribuições não são normais. É possível que o teste tenha rejeitado a hipótese de normalidade por conta do tamanho da amostra. De acordo com o teste proposto por Shapiro e Wilk (1965), amostras grandes (como neste trabalho - 507) tornam o teste mais preciso e por isso rejeitou a normalidade dos índices das janelas.

Testes adequados para grandes amostras poderiam dar melhores resultados sobre este *TVG*. Assim, caso um teste assim acusasse como normal as distribuições dos índices, seria possível inferir que um conjunto de títulos de qualquer época representa todo o TVG, já que todas as redes possuem índices muito próximos da média.

Ainda não se sabe ao certo o que significa "não ter memória" para o *caminho mínimo*

médio. Mas algumas interpretações podem ser feitas. Por exemplo, considere duas palavras na rede de uma janela do TVG que pertencam à dois cliques (i.e. títulos) que distam entre si um certo número de arestas (e.g. $\ell = 5$, Figura 7.8 - à esquerda), estes dois títulos contém ideias que não interagem diretamente na rede, mas sim indiretamente, através de outras palavras intermediárias que conectam estes dois títulos (caminho verde, Figura 7.8 - à esquerda). Mas, de acordo com o resultado apresentado neste capítulo, o índice ℓ está associado à um passeio aleatório e portanto, a janela imediatamente posterior pode ter estes títulos interagindo de maneira completamente diferente, como na Figura 7.8, à direita.

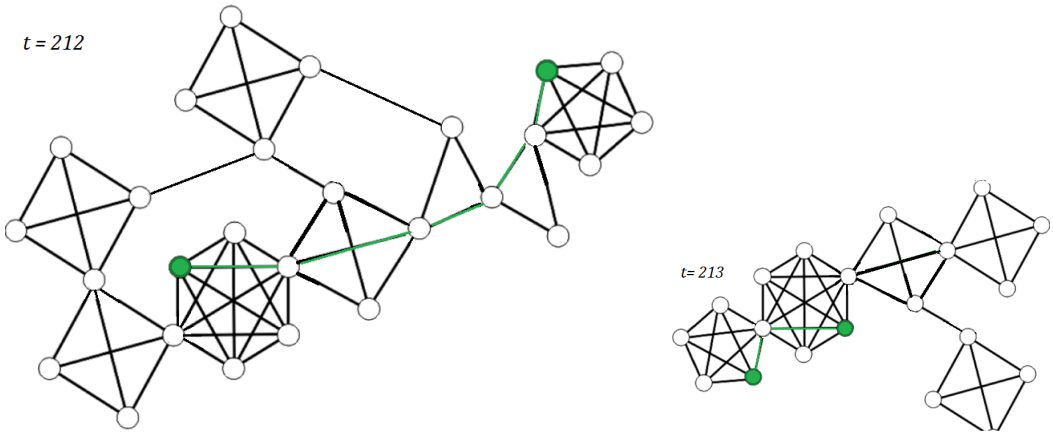


Figura 7.8: À esquerda, dois títulos que interagem indiretamente por meio de cliques intermediários. À direita, um instante de tempo depois, os mesmos títulos interagindo diretamente. Este exemplo evidência para rede de títulos a falta de memória para o índice $\langle \ell \rangle$ na série temporal.

A abordagem de rede de cliques proposta por Fadigas e Pereira (2013) é adequada para a base de dados. Sabemos que uma rede de títulos é formada por junção de cliques através de palavras comuns. Os títulos dos artigos publicados em uma revista são apenas grupos de palavras isoladas que interagem entre si e representam bem a ideia dos trabalhos publicados.

Com os resultados obtidos neste capítulo, estudos futuros poderão estabelecer relações significativas entre os vocabulários das janelas. Por exemplo, pode ser feita uma análise sobre a frequência de palavras de alto grau nas janelas, como foi iniciado na Seção 7.4 deste capítulo, ao buscar respostas para a questão 3. Respostas mais precisas para essa questão poderão revelar algo sobre a capacidade das palavras de alto grau nas janelas interagirem com grupos de palavras diferentes, indicando interações entre áreas do conhecimento humano, que podem ser visualizadas nas redes com algum software de visualização, como o Gephi ou Pajek.

Parte V

Conclusão

Capítulo Oito

Considerações Finais

8.1 Conclusões

As redes complexas oferecem muitas possibilidades de análise de fenômenos naturais, e.g. a linguagem. Neste sentido, a comunicação científica é fruto de uma complexa rede de pesquisadores que interagem de diversas formas. Vimos aqui, que a partir do vocabulário comum dos cientistas nos títulos de suas publicações, é possível modelar uma maneira de visualizar localmente e dinamicamente como funciona o sistema formal de comunicação científica, i.e. o periódico científico.

Conforme vimos, as redes dos periódicos possuem um ponto crítico bem definido, assim como em redes de discursos escritos e de discursos orais, de trabalhos anteriores a este. Isto nos leva a supor que existem muitas semelhanças entre discursos de pessoas comuns e de conjuntos de publicações em um dado periódico. Este conjunto precisa passar rigorosamente por um comitê avaliador, composto por mais de uma pessoa, diferentemente das redes de discursos orais e escritos mencionados, que são oriundos de uma só pessoa.

Com a rede crítica foi possível indicar algumas maneiras de diferenciar periódicos, cruzando os índices de suas redes críticas. A partir deste olhar, pesquisadores que estudam colaboração e cooperação científica poderão contribuir com interpretações mais precisas para diferenciar periódicos com o uso de redes semânticas. Assim, com a verificação do fenômeno da rede crítica em redes de títulos, bem como a existência de correlações persistentes nas séries temporais do TVG, novas possibilidades se abrem para o estudo da colaboração científica. Como se sabe, ao longo do tempo, os periódicos tornaram-se estruturas de representação do conhecimento científico da humanidade.

A abordagem TVG foi fundamental para se perceber tendências no comportamento do relacionamento das palavras. O método *DFA* foi útil para perceber que existe uma forte correlação no número de vértices e arestas no intervalo $4 \leq \tau \leq 21$. Isto significa, por exemplo, que para um dado tempo $t \in \Gamma$, se o vocabulário da Nature aumentou nos últimos 2 meses, então existe uma alta probabilidade dele continuar aumentando a partir do próximo mês, e essa tendência se mantém fortemente correlacionada até aproximadamente 4 meses depois.

A *densidade* nas janelas ao longo do tempo, em média, tende a diminuir. Isto não ocorre pela redução das arestas entre seus vértices, mas pelo aumento médio no número de vértices dessas janelas. Ou seja, o vocabulário da revista em uma janela de observação de

8 semanas aumenta, a medida que o tempo passa.

A abordagem de rede de cliques pôde mostrar o quanto mudam os indicadores, a partir da junção das cliques, a partir de uma configuração inicial em que estão isoladas. Os altos coeficientes de aglomeração e a constatação do fenômeno *small-word* nos leva a supor que é alta a probabilidade de que duas palavras ligadas a uma outra, estejam, elas mesmas, ligadas entre si nas janelas do TVG.

A partir dos resultados desta pesquisa, trabalhos futuros poderão investigar: (a) o que representa a memória persistente nas séries temporais aqui apresentadas, explicar tendências nos padrões de relacionamentos das palavras utilizadas pelos pesquisadores que publicam em um dado periódico; (b) investigar padrões modulares nas redes críticas aqui apresentadas, a fim de observar quais temáticas são mais ou menos abordadas em publicações de um dado periódico.

8.2 Atividades Futuras de Pesquisa

Os *comentários finais* nos capítulos 7 e 6 (Seções 7.5 e 6.3) revelam a importância dos resultados obtidos. Trabalhos futuros poderão aproveitar as lacunas contidas nestas seções e ampliar o estudo da colaboração científica. Como sugestão, utilizando a mesma modelagem desta pesquisa:

- Incluir outras bases para o método TVG (está sendo feito para o periódico *Science*);
- Buscar redes críticas em janelas do TVG;
- Ampliar o tempo de vida do TVG;
- Fazer um estudo sobre a função de latência doTVG (está sendo feito com a base de dados desta pesquisa);
- Analisar os resultados levando em conta as relações semânticas e de frequência de aparição das palavras dos títulos em épocas distintas;
- Catalogar métodos de redes semânticas para ajudar futuros pesquisadores em usar os métodos adequados às suas demandas;
- Verificar mais evidências que comprovem que a rede crítica é um mecanismo instrínseco da linguagem humana;
- Desenvolver e aplicar testes de normalidades para distribuições com muitos elementos.

Dessa forma, presente trabalho funciona como um “vértice” de uma grande “rede dirigida” que representa o estudo de redes complexas, redes semânticas e redes de informação. Este vértice emana vetores de informação importantes para o fluxo de atividades que estão por vir, para enriquecer ainda mais o estudo da difusão do conhecimento humano.

Parte VI

Apêndice

Apêndice A

Limitações da Pesquisa

Inicialmente o trabalho consistia em usar o índice *incidência-fidelidade* - proposto por Teixeira et al. (2010) - para encontrar redes críticas para redes de títulos - estudadas por Pereira et al. (2011), baseadas nas mesmas metodologias descritas por estes trabalhos. Entretanto, as redes dos periódicos possuíam tamanhos consideravelmente diferentes, o que impossibilitou a comparação de suas redes críticas.

Foi decidido, então, a fixação de 300 títulos para cada periódico, devido à revista *AFE* (371 títulos) - que possui a menor quantidade de títulos da base de dados. Estes títulos foram retirados aleatoriamente do conjunto total de títulos de cada periódico. Inicialmente foi proposto 1000 retiradas aleatórias para cada periódico a fim de abranger o máximo de possibilidades de combinações de grupos com 300 títulos.

Entretanto, para a primeira revista escolhida (HR) este procedimento demorou aproximadamente 3,5 dias para que o o conjunto de softwares utilizados, em funcionamento ininterrupto, gerasse as redes e calculasse os índices de redes para cada valor de *incidência-fidelidade* considerado. Ficou evidente que o procedimento ao ser estendido aos demais periódicos iria demandar muito mais tempo. Isto se explica por conta do aparato computacional disponível (CPU de 1,7 GHZ e 3Gb de memória RAM), bem como por limitações nos softwares advindos dos trabalhos em que esta pesquisa baseia a sua metodologia.

Felizmente, os valores médios e seus respectivos desvios, para os índices das 1000 redes de 300 títulos da revista *HR* na rede crítica e demais incidências fidelidades tiveram discrepância desprezível comparados com o mesmo experimento, na mesma revista, só que para 100 retiradas aleatórias. A partir de então, todos os periódicos passaram pelo mesmo processo, com exceção da Nature e Science que tiveram 180 retiradas - por conta da enorme quantidade de títulos de suas bases.

Não obstante este procedimento revelar um importante resultado: todas as redes possuem um comportamento crítico para o mesmo valor de *incidência-fidelidade*[1] obtido em discursos orais, com a análise dos primeiros resultados, percebeu-se que existe uma larga faixa de valores de um mesmo índice na rede crítica, mostrando que sua média esconde muita informação, visando assim não contribuir para diferenciar um periódico de outro.

Para contornar esta limitação foi experimentado o uso do índice *incidência-fidelidade* em

[1] Este índice foi chamado inicialmente e *força fidelidade*(TEIXEIRA, 2007).

sua forma reescalonada[2], em que se permite apenas valores situados entre 0 e 1. Isto elimina o efeito do tamanho do texto na comparação de redes críticas. Assim todos os periódicos puderam ser analisados em seus tamanhos originais (i.e. com a mesma quantidade de títulos em que se apresentam na base de dados desta pesquisa).

Diante da limitação anterior, surgiu um questionamento: Por que larga faixa de valores para um dado índice na rede crítica? A hipótese que surgiu foi que o surgimento de um título em um periódico depende de títulos que o antecederam em um curto período de tempo. Quando o conjunto de 300 títulos é composto por títulos alatórios, a sequência temporal que demonstra o surgimento deles é provavelmente quebrada. Ou seja, é prudente verificar os valores dos índices de redes no caso de os 100 grupos de 300 títulos fossem oriundos de datas diferentes do periódico. Então, seria possível verificar se existe dependência temporal no surgimento deles, e até caracterizar as redes do periódico em diferentes épocas e compará-las.

Dessa forma, a investigação subsequente contemplou métodos de TVG[3] aplicado ao periódico *Nature*, com janelas temporais de 8 semanas de publicação. Anteriormente foi feito o mesmo procedimento com uma janela menor (i.e. 4 semanas de publicações), entretanto os resultados mostraram flutuações bruscas dos valores dos índices ao longo do tempo. Dessa forma, a escolha do tamanho atual para a janela foi feita por que as flutuações dos valores dos índices são mais suaves, podendo indicar tendências evolutivas nestes valores ao longo do tempo. O prazo para a entrega desta dissertação limita que o mesmo método possa ser empregado e analisado em outras revistas da base. Sugere-se isto para trabalhos futuros.

[2] O Índice foi inicialmente chamado de *força fidelidade normalizada*(AGUIAR, 2009).

[3] TVG significa *Time-Varying Graphs* e trata de redes em que a existência de suas arestas e vértices depende do instante de tempo que o sistema se encontra (CASTEIGTS et al., 2011).

Referências

AGUIAR, M. S. F. *Redes de palavras em textos escritos: Uma análise da linguagem verbal utilizando redes complexas*. Dissertação (Programa de Pós-Graduação em Físca) — Universidade Federal da Bahia, Salvador, 2009.

AGUIAR, M. S. F. de et al. Análiselise da resiliência de redes de textos escritos. In: *XXV Encontro de Físicos do Norte e Nordeste, Livro de Resumos*. Natal, RN: [s.n.], 2007.

ALBUQUERQUE, F. J. B.; PIMENTEL, C. E. Uma aproximação semântica aos conceitos de urbano, rural e cooperativa. *Psicologia: Teoria e Pesquisa*, SciELO Brasil, v. 20, n. 2, p. 175–182, 2004.

AMBLARD, F.; CASTEIGTS, A.; FLOCCHINI, P.; QUATTROCIOCCHI, W.; SANTORO, N. On the temporal analysis of scientific network evolution. In: *CASoN*. [S.l.: s.n.], 2011. p. 169–174.

ANDRADE, M. T. T. *A Colaboração em Comunidades Científicas: das Redes de Coparticipação à Difusão do Conhecimento*. Tese (Tese (Doutorado Multiinstitucional e Multidisciplinar de Difusão do Conhecimento)) — Universidade Federal da Bahia, Faculdade de Educação, UFBA, Salvador (BA), 2013.

ANGELIS, A. F. D. *Tutorial Redes Complexas*. Instituto de Física de São Carlos - Universidade de São Paulo: FAPESP, 2005.

BAK, P.; TANG, C.; WIESENFELD, K. Self-organized criticality. *Physical Review A*, v. 38, n. 1, p. 364–374, 1988.

BARABASI, A. L. The architecture of complexity, from network structure to human dynamics. *Ieee Control Systems Magazine*, p. 33–42, 2007.

BARABASI, A.-L.; ALBERT, R. Emergence of scaling in random networks. *Science*, v. 286, p. 509–512, 1999.

BARABASI, A.-L.; ALBERT, R.; JEONG, H. Mean-field theory for scale-free random networks. *Physica A*, v. 272, p. 173–187, 1999.

BARABáSI, A. L. *Linked: the new Science of networks*. Cambridge, MA: Perseus, 2002.

BARABáSI A. L., B. E. Scall-free networks. *Scienticf American*, p. 50–59, 2003.

CALDEIRA, S. *Caracterização da rede de signos linguísticos: Um modelo baseado no aparelho psíquico de Freud*. Dissertação (Mestrado Interdisciplinar em Modelagem Computacional) — Fundação Visconde de Cairu, Salvador, 2005.

CASTEIGTS, A.; FLOCCHINI, P.; QUATTROCIOCCHI, W.; SANTORO, N. Time-varying graphs and dynamic networks. In: *ADHOC-NOW*. [S.l.: s.n.], 2011. p. 346–359.

CUNHA, M. do V.; ROSA, M. G.; FADIGAS, I. de S.; MIRANDA, J. G. V.; PEREIRA, H. B. de B. Redes de títulos de artigos científicos variáveis no tempo. In: *BraSNAM - II Brazilian Workshop on Social Network Analysis and Mining*. [S.l.: s.n.], 2013. p. 1744–1755.

ERDOS, P. On cliques in graphs. *ISRAEL JOURNAL OF MATHEMATICS*, v. 4, p. 233–234, 1966.

ERDOS, P.; RENYI, A. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, v. 5, p. 17–61, 1960.

EULER, L. Solutio problematis ad geometriam situs pertinentis. *Graph Theory 1736-1936*, Oxford University Press, USA, 1736.

FADIGAS, I.; HENRIQUE, T.; PEREIRA, H.; SENNA, V.; MORET, M. Análise de redes semânticas baseada em títulos de artigos de periódicos científicos: o caso dos periódicos de divulgação em educação matemática. *Educação Matemática Pesquisa*, São Paulo, v. 11, n. 1, p. 167–193, 2009.

FADIGAS, I.; PEREIRA, H. A network approach based on cliques. *Physica A: Statistical Mechanics and its Applications*, v. 392, n. 10, p. 2576 – 2587, 2013.

FADIGAS, I. S. *Produção e difusão do conhecimento em educação matemática sob perspectiva das redes sociais e complexas*. Tese (Tese (Doutorado Multiinstitucional e Multidisciplinar de Difusão do Conhecimento)) — Universidade Federal da Bahia, Faculdade de Educação, UFBA, Salvador (BA), 2011.

FADIGAS, I. S.; CUNHA, M. do V.; ROSA, M. G.; PEREIRA, H. B. de B. Análise de redes de coautoria por meio de redes semânticas uniformes. In: *BraSNAM - II Brazilian Workshop on Social Network Analysis and Mining*. Maceió: [s.n.], 2013. p. 1553–1564.

FEOFILOFF, P.; KOHAYAKAWA, Y.; WAKABAYASHI, Y. Uma introdução suscinta à teoria dos grafos. In: *II Bienal da SBM, Salvador - Ba, 2004*. USP, SP: Sociedade Brasileira de Matemática, 2007. v. 2, p. 1–61. Disponível em: <http://www.ime.usp.br/~pf/teoriadosgrafos/>.

GALVãO, V. et al. Modularity map of the network of human cell differentiation. *Proceedings of the National Academy of Sciences*, v. 107, n. 13, p. 5750–5755, 2010.

GARVEY, W. D. *Communication: The Essence of Science*. [S.l.]: Oxford, NY: Pergamon Press, Inc., 1979. ISBN 0-08-022254-4.

MANDELBROT, B. B. *The fractal geometry of nature*. [S.l.]: Macmillan, 1983.

MARTINS, N. S. *Introdução à estilística: a expressividade na língua portuguesa*. [S.l.]: EdUSP, 2008. ISBN 9788531410123.

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

MIRANDA, D.; PEREIRA, M. O periódico científico como veículo de comunicação: uma revisão de literatura. *Ibict*, Ci. Inf., Brasília, v. 25, n. 3, p. 375–382, 1996.

MONTEIRO, R. L. S. AND PEREIRA, H. B. B. Roberto Luiz Souza Monteiro, Hernane Borges de Barros Pereira, Marcelo A. Moret e Inácio S. Fadigas. *SCNTOOLS - Social and Complex Network Tools*. 2010. BR n. RS 11102-5, rPI 2083 de 07/12/2010.

MONTEMURRO, M.; ZANETTE, D. Entropic analysis of the role of words in literary texts. *Adv. Complex. Syst.*, v. 1, n. 5, p. 7–17, 2002.

NASCIMENTO, C. Helano Aquino do. *Aplicação de Redes Complexas no Estudo de Redes Elétricas*. Dissertação (Mestrado Interdisciplinar em Modelagem Computacional e Tecnologia Industrial) — SENAI CIMATEC, Salvador, 2012.

NELSON, D. L.; MCEVOY, C. L.; SCHREIBER, T. A. The university of south florida word association, rhyme and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, Springer, 1998.

NEWMAN, J. R. Leonhard euler and the koenigsberg bridges. *Scientific American*, v. 45, 1953.

NEWMAN, M. E. Scientific collaboration networks. i. network construction and fundamental results. *Physical review E*, APS, v. 64, n. 1, p. 016131, 2001.

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

NUSSENSVEIG, H. M. *Complexidade e Caos*. [S.l.]: UFRJ, 2008.

OLIVEIRA-FILHO, J. de. *Aparelho Psíquico de Freud: uma busca pela caracterização de seus caminhos*. Dissertação (Mestrado em Modelagem Computacional e Tecnologia Industrial) — SENAI Cimatec, Salvador, 2012.

PAUMIER, S. *Unitex Manual de Utilização. Rede Relex Brasil*. 2002.

PENG, C.-K. et al. Mosaic organization of dna nucleotides. *Phys. Rev. E*, American Physical Society, v. 49, n. 2, p. 1685–1689, 1994.

PEREIRA, H.; FADIGAS, I.; SENNA, V.; MORET, M. Semantic networks based on titles of scientific papers. *Physica A: Statistical Mechanics and its Applications*, v. 390, n. 6, p. 1192–1197, 2011.

SANTANA, A. *Caracterização da jornada máxima em redes dinâmicas*. Dissertação (Programa de Pós-Graduação em Matemática) — Universidade Federal da Bahia, Salvador, 2012.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, v. 52, n. 3/4, p. 591–611, 1965.

SILVA, B. et al. Statistical characterization of an ensemble of functional neural networks. *European Physical Journal B*, v. 392, p. 85–358, 2012.

STERNBERG, R. *Psicologia cognitiva*. Porto Alegre, RJ: Artes Médicas Sul, 2011.

TEIXEIRA, G. et al. Complex semantics networks. *International Journal of Modern Physics C*, v. 21, n. 3, p. 333–347, 2010.

TEIXEIRA, G. M. *Redes semânticas baseadas em discursos orais: Uma proposta metodológica baseada na psicologia cognitiva utilizando redes complexas*. Dissertação (Mestrado Interdisciplinar em Modelagem Computacional) — Fundação Visconde de Cairu, Salvador, 2007.

TUKEY, J. W. *Exploratory Data Analysis*. [S.l.]: Addison-Wesley, 1977.

VANZ, S.; STUMPF, I. Colaboração científica: revisão teórico-conceitual. *Perspectivas em Ciência da Informação*, scielo, v. 15, p. 42–55, 2010.

WASSERMAN, S.; FAUST, K. *Social Network Analsis*. [S.l.]: Cambridge: Cambridge University Press., 1994.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of'small-world'networks. *Nature*, v. 393, n. 6684, p. 409–10, 1998.

ZIMAN, J. Comunidade e comunicação. In: _____. [S.l.]: Conhencimento público. São Paulo: EDUSP, 1979. p. 115–138.

ZIPF, G. K. Human behavior and the principle of least effort. addison-wesley press, 1949.

ZIPF, G. K. The principle of least effort: an introducion to human ecology. Hafner Publishing Company, 1972.

Redes semânticas baseadas em títulos de artigos científicos

Marcelo do Vale Cunha

Salvador, Novembro de 2013.