SENAI CIMATEC

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL E TECNOLOGIA INDUSTRIAL

Doutorado em Modelagem Computacional e Tecnologia Industrial

Tese de doutorado

Entropia da informação em redes semânticas de títulos variáveis no tempo

Apresentada por: Marcelo do Vale Cunha
Orientador: Hernane Borges de Barros Pereira

Outubro de 2020

Marcelo do Vale Cunha

Entropia da informação em redes semânticas de títulos variáveis no tempo

Tese de doutorado apresentada ao Programa de Pós-graduação em Modelagem Computacional e Tecnologia Industrial, Curso de Doutorado em Modelagem Computacional e Tecnologia Industrial do SENAI CIMATEC, como requisito para a obtenção do título de **Doutorado em Modelagem Computacional e Tecnologia Industrial**.

Área de conhecimento: Interdisciplinar

Orientador: Hernane Borges de Barros Pereira
SENAI CIMATEC

Salvador
SENAI CIMATEC
2020

Ficha catalográfica elaborada pela Biblioteca do Centro Universitário SENAI CIMATEC

C972e Cunha, Marcelo do Vale

Entropia da informação em redes semânticas de títulos variáveis no tempo / Marcelo do Vale Cunha. – Salvador, 2020.

92 f.: il. color

Orientador: Prof. Dr. Hernane Borges de Barros Pereira

Tese (Doutorado em modelagem computacional e tecnologia industrial) – Programa de Pós-Graduação, Centro Universitário SENAI CIMATEC, Salvador, 2020.

Inclui referências.

1. Redes semânicas. 2. Grafos Variantes. 3. Entopia da informação. 4. Entropia de Multi Escala. 5. Difusão do conhecimento. I. Universitário SENAI CIMATEC. II. Pereira, Hernane Borges de Barros. III. Titulo.

CDD 620.00113

NDI - 01

Nota sobre o estilo do PPGMCTI

Esta tese de doutorado 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.

Sistema FIEB | SENAI CIMATEC
PELO FUTURO DA INOVAÇÃO

Centro Universitário SENAI CIMATEC

Doutorado em Modelagem Computacional e Tecnologia Industrial

A Banca Examinadora, constituída pelos professores abaixo listados, leu e aprovou a Tese de doutorado, intitulada "**Entropia da informação em redes semânticas de títulos variáveis no tempo**", apresentada no dia 26 de outubro de 2020, como parte dos requisitos necessários para a obtenção do Título de Doutor em Modelagem Computacional e Tecnologia Industrial.

HERNANE BORGES DE BARROS PEREIRA:58646450520
Assinado de forma digital por HERNANE BORGES DE BARROS PEREIRA:58646450520
Dados: 2020.11.09 14:22:42 -03'00'

Orientador: Prof. Dr. Hernane Borges de Barros Pereira
Centro Universitário SENAI CIMATEC

Membro Interno: Prof. Dr. Marcelo Albano Moret Simões Gonçalves
Centro Universitário SENAI CIMATEC

Membro Interno: Prof. Dr. Roberto Luiz Souza Monteiro
Centro Universitário SENAI CIMATEC

Membro Externo: Prof. Dr. José Fernando Mendes
Universidade de Aveiro

Membro Externo: Prof. Dr. José Garcia Vivas Miranda
Universidade Federal da Bahia

Dedico este trabalho a minha tia Lola (In memoriam) e a meus afilhados Luca e Luna.

Agradecimentos

Agradeço primeiramente a Deus pelo dom da vida e por esta grande oportunidade de aprendizado, a meus pais Silvio e Genilda (Painho e Mainha) e às minhas irmãs Danielle e Sinara (Dani e Tai) por todo amor e apoio incondicionais nesta jornada. São minha fortaleza e inspiração. Devo tudo a vocês.

Agradeço à meu orientador Hernane Pereira pela presença constante, orientação acadêmica e pela amizade. Isto me permitiu conviver, observar e aprender de perto as melhores vivências acadêmicas, com ética e bom humor. Tenho grande admiração como pessoa e eterno agradecimento pela sua dedicação em minha formação. Um verdadeiro mestre.

Agradeço aos professores do Programa MCTI, em especial à Marcelo Moret, grande físico que muito aprendo. Agradeço aos colegas do curso, em especial a Carlos César e Jefferson Nascimento pelas parcerias nos trabalhos e amizade. Sou muito grato a todos os colegas do grupo de pesquisa *Fuxicos e Boatos* pelas discussões que vão além das reuniões semanais e pelas parcerias nos trabalhos, em especial a Cleônidas Tavares pela grande ajuda computacional nesta tese.

Agradeço aos membros da Banca que fizeram contribuições fantásticas na qualificação, em especial José Fernando Mendes e José Garcia Miranda que além das contribuições neste trabalho, me inspiram muito como Físico!

Agradeço ao programa de recursos humanos da ANP (PRH55) pela vivência e bolsa nos 2 primeiros anos de doutorado, com excelentes trocas entre os alunos e pesquisadores. Agradeço ao IFBA campus Barreiras e à PRPGI-IFBA por viabilizarem um afastamento de 18 meses e pelo apoio dos colegas e amigos, em especial Anderson Almeida, neste processo.

Agradeço à Marina pela escuta amorosa e apoio em momentos difíceis na reta final. Agradeço aos verdadeiros amigos pelo irrestrito apoio e a todos que incentivaram e acreditaram em minha formação como pesquisador e como pessoa durante esta jornada.

Abraço forte à todos.

Salvador, Brasil — Marcelo do Vale Cunha
26 de Outubro de 2020

Resumo

A difusão do conhecimento em periódicos científicos reflete e também influencia diversos campos de atuação humana. O periódico científico é um sistema formal de comunicação científica que tem sido amplamente estudado nas últimas décadas, principalmente no âmbito da colaboração científica. A teoria de redes fornece possibilidades interessantes de modelagem deste sistema, tais como: redes de coautoria, redes de citações, redes de palavras chaves e redes semânticas de títulos de artigos científicos (RST). A modelagem através de uma RST baseada em cliques foi proposta recentemente e sua análise permite avaliar a comunidade científica que publica, a partir do vocabulário comum utilizado nos títulos dos artigos. Trabalhos recentes empregaram a teoria da informação em redes sociais e complexas e discutem frequentemente a entropia em distribuições de graus. No entanto, há carência de estudos sobre entropia em redes de cliques. Este trabalho visa preencher esta lacuna com uma metodologia que utiliza a entropia da informação em redes de títulos. A modelagem consiste em considerar as palavras dos títulos de publicações em revistas científicas como vértices de cliques de uma rede variante no tempo. O objetivo é propor um conjunto de métodos que permitem acompanhar o surgimento de novas ideias ao longo do tempo, representadas pelo aumento da diversidade de vocabulário dos títulos ou pela robustez e consolidação de ideias e interesses de autores e editores de uma revista em um determinado período. São eles: (I) uso da modelagem *Grafos variantes no tempo* para mostrar a evolução da rede; (II) captura de índices de redes ao longo do tempo; (III) entropia dos valores dos índices de redes ao longo do tempo (IV) cálculo das entropias de Shannon para vértices e arestas em cada janela de tempo; (V) cálculo dos máximos e mínimos para os valores de entropia considerando vínculos em rede de cliques; (VI) comportamento da entropia para uma rede crescente; (VII) uso do índice incidência-fidelidade para encontrar a rede crítica em janelas de tempo e (VIII) identificação dos vértices que se destacam nas redes críticas. Destes métodos, apenas I e II foram desenvolvidos em trabalhos anteriores. O conjunto de dados são os títulos dos trabalhos científicos publicados na Nature e Science ao longo de dez anos. As palavras dos títulos são tratadas manual e computacionalmente segundo regras preestabelecidas para se adequar à modelagem. Os resultados mostram como a diversidade de vocabulário evolui ao longo do tempo, com base nos valores de entropia de vértices e arestas das redes semânticas de cliques. Este estudo além de aumentar o leque de possibilidades de aplicação para redes de cliques, contribui para o estudo da difusão e disseminação do conhecimento científico.

Palavras-chave: Redes Semânticas de Títulos, Grafos variantes no tempo, Entropia da Informação, Entropia de Multi Escala, Redes de Cliques.

Abstract

The diffusion of knowledge in scientific journals reflects and also influences different fields of human activity. The scientific journal is a formal system of scientific communication that has been extensively studied in recent decades, mainly in the scope of scientific collaboration. Network theory provides interesting possibilities for modeling this system, such as: co-authoring networks, citation networks, keyword networks and semantic networks of scientific article titles (STN). Modeling through a clique based STN was recently proposed and its analysis allows to evaluate the scientific community that publishes, based on the common vocabulary used in the titles of the articles. Recent work has employed information theory in social and complex networks and frequently discuss entropy in degree distributions. However, there is a lack of specific work on entropy in network of cliques. This work aims to fill this gap with a methodology that uses the entropy of information in networks of titles. The modeling consists in considering the words of the titles of publications in scientific journals as vertices of cliques of a time-varying network. The objective is to propose a set of methods that allow to follow the emergence of new ideas over time, represented by the increase in the vocabulary diversity of the titles or by the robustness and consolidation of ideas and interests of authors and editors of a magazine in a given period. They are: (I) use of *Time Varying Graphs* modeling to show the evolution of the network; (II) capturing network indexes over time; (III) entropy of the network index values over time (IV) calculation of Shannon entropies for vertices and edges in each time window; (V) calculation of maximum and minimum values for entropy values considering initial conditions in network of cliques; (VI) entropy behavior for a growing network; (VII) use of the incidence-fidelity index to find the critical network in time windows and (VIII) identification of the vertices that stand out in the critical networks. Of these methods, only I and II were developed in previous works. The data set are the titles of scientific papers published in Nature and Science over ten years. The words of the titles are treated manually and computationally according to pre-established rules to suit the modeling.
The results show how the vocabulary diversity evolves over time, based on the entropy values of vertices and edges of the semantic click networks. This study, in addition to increasing the range of application possibilities for network of cliques, contributes to the study of the diffusion and dissemination of scientific knowledge.

Sumário

Lista de Tabelas

Lista de Figuras

Lista de Siglas

PPGMCTI ..	Programa de Pós-graduação em Modelagem Computacional e Tecnologia Industrial
WWW	World Wide Web
MSE	Entropia de multi escala
TVG	Grafos variáveis no tempo
RST	Rede semânticas de títulos
RSTVT	Rede semânticas de títulos variáveis no tempo

Parte I

Introdução

Capítulo Um

Introdução

Ciência das redes tem ganhado um papel de destaque no âmbito de pesquisas que visam entender o comportamento e a estrutura de sistemas que contém entidades que se conectam. Dentre vários exemplos que podemos citar (e.g. redes tecnológicas, redes biológicas, redes sociais, redes organizacionais, redes de informação), destaca-se nesta pesquisa a modelagem por rede semântica de títulos como um sistema de representação do conhecimento proveniente de um periódico científico.

Nas últimas décadas, o periódico científico tem chamado atenção pelos processos de comunicação e memória científica presentes neste sistema. Este canal de difusão do conhecimento tem sido largamente estudado sob diferentes perspectivas: (a) terminologias, funcionalidades e caoktegorização (ZIMAN, 1979; MIRANDA; PEREIRA, 1996; VANZ; STUMPF, 2010; GARVEY, 2014); (b) a influência da escrita na divulgação científica (VOLPATO; FREITAS, 2003; GASTEL; DAY, 2016); (c) as redes de colaboração que fomenta: redes de citações (PRICE, 1965; PRICE, 1986); redes de coautoria (NEWMAN, 2001b; NEWMAN, 2001a; AMBLARD et al., 2011) e (d) rede semântica de títulos de artigos científicos (RST) (FADIGAS et al., 2009; PEREIRA et al., 2011; CUNHA et al., 2013; PEREIRA et al., 2016; GRILO et al., 2017; CUNHA et al., 2020b). Nestes últimos, os processos de comunicação e memória científica são observados a partir de signos linguísticos encontrados em elementos do periódico científico (e.g. título, palavras-chave, etc.), e das análises das propriedades estáticas e dinâmicas das redes semânticas associadas.

Do ponto de vista estático, encontramos alguns trabalhos sobre o estudo topológico de redes semânticas (PEREIRA et al., 2011; CUNHA; MIRANDA; PEREIRA, 2015; NASCIMENTO et al., 2016). O estudo de RSTs, sob análise de índices de teoria de redes, pode ser utilizado para identificar temáticas que se destacam em uma determinada área do conhecimento. Neste sentido, Rodrigues et al. (2017) identificaram temas que se destacam na rede de títulos de artigos na área de Saúde Coletiva e seus respectivos autores, na rede de coautoria associada. Outros trabalhos consideraram a dinâmica de uma RST, como o de Cunha et al. (2013) que investigaram padrões no vocabulário dos títulos de artigos publicados no periódico Nature e suas tendências ao longo do tempo; Pereira et al. (2016) estudaram a evolução da densidade durante a construção de redes semânticas como indicador de diversidade de conceitos de periódicos científicos; e Grilo et al. (2017) propuseram um método que analisa a robustez de uma RST, utilizando estratégias de remoção de vértices, sendo possível identificar uma fração de remoção crítica para a qual a estrutura topológica da rede é mudada.

As redes dos trabalhos supracitados são formadas por cliques. Esta configuração de vértices é comum em sistemas naturais e sociais, onde a rede cresce pela adesão de vértices mutualmente conectados, em por exemplo, redes de atores de filmes (BARABÁSI; ALBERT, 1999), redes de coautoria (NEWMAN, 2001b), redes de conceitos (CALDEIRA et al., 2006), redes de árbitros de futebol (FADIGAS et al., 2020) e redes semânticas (TEIXEIRA et al., 2010; PEREIRA et al., 2011; PEREIRA et al., 2016; GRILO et al., 2017).

Uma rede semântica de cliques é composta por palavras com significado semântico representado pelos vértices, e arestas que representam as conexões entre palavras que aparecem na mesma unidade de significado, ou seja, em uma sentença (frase), um parágrafo ou um título do discurso analisado (PEREIRA et al., 2016; GRILO et al., 2017). Esta proposição é oriunda dotrabalho de Caldeira et al. (2006), marco para rede de palavras formadas por cliques, em que estabelece a sentença como menor unidade de significado de um texto. Outros trabalhos seguiram esta premissa, como o de TEIXEIRA et al. (2010) que identificou um padrão na linguagem humana a partir da rede de discursos orais; e o trabalho de Pereira et al. (2011) que estabelece regras de tratamento para construção de redes semânticas de títulos de periódicos científicos e propõe maneiras de analisar estas redes.

Portanto, em uma rede semântica de cliques cada sentença representa uma clique, em que as palavras são vértices que se conectam com outras da mesma sentença. Esta modelagem pode fornecer respostas interessantes para o estudo da organização da linguagem humana. Nesse sentido, TEIXEIRA et al. (2010) propuseram o índice *incidência-fidelidade* para encontrar uma configuração crítica da rede semântica de um discurso oral. Cunha, Miranda e Pereira (2015) aplicaram este índice em redes de títulos para encontrar a rede crítica. Esta configuração consiste em uma rede ótima, filtrada pelo índice incidência-fidelidade, com o máximo de informação e o mínimo de resíduo textual.

A teoria da informação evoluiu nas últimas décadas e foi aplicada em diferentes campos, como biologia, economia e sistemas quânticos confinados, entre outros (BRILLOUIN, 2013; MOUSAVIAN; KAVOUSI; MASOUDI-NEJAD, 2016; NASCIMENTO; PRUDENTE, 2018; MISHRA; AYYUB, 2019). Nestes trabalhos, entropia é uma medida de informação e é aplicada segundo os preceitos de Shannon (SHANNON, 1948). Mas também pode ser uma medida da complexidade de um sistema, quando é aplicada a uma série temporal associada a dinâmica do sistema, a exemplo dos métodos *Entropia de Multi Escala* (COSTA; GOLDBERGER; PENG, 2005) e o *Entropia da Entropia* (HSU et al., 2017) aplicados a sinais biológicos. Recentemente, alguns autores introduziram a entropia de Shannon para medir as informações contidas na distribuição de graus e distâncias geodésicas em redes observáveis ou em modelos clássicos e redes semânticas para classificar e diferenciar esses sistemas pela heterogeneidade de suas conexões (SOLÉ; VALVERDE, 2004; JI et al., 2008; VIOL et al., 2019).

Apesar do crescente interesse de várias áreas sobre a entropia de Shannon, existe a necessidade de mais estudos desta medida em redes sociais e complexas. Esta tese propõe uma metodologia para o cálculo da entropia em redes de cliques variantes no tempo, respeitando os vínculos associados à formação dessas redes. As possibilidades exploram também a evolução temporal de uma rede com uso de séries temporais.

Os limites para os valores de entropia precisam respeitar os vínculos inerentes ao processo de formação de uma rede de cliques. A entropia é calculada aqui para a configuração inicial das redes de cliques baseadas em títulos de artigos científicos dos periódicos *Nature* e *Science* de 1998 a 2008, mas pode ser generalizado para quaisquer redes de cliques.

O uso de uma série temporal para os valores de entropia e de outros índices de redes ajuda a investigar melhor o comportamento da rede semântica de títulos ao longo do tempo. As redes são modeladaso como um *grafo variante no tempo* (TVG, do inglês *Time-Varying Graphs*), permitindo que uma janela de tempo avance no tempo capturando valores de índices de rede e entropia.

Já é sabido de trabalhos anteriores que este procedimento revela correlação de memória para os vocabulários das SNTs (CUNHA, 2013; CUNHA et al., 2013), através do uso do método DFA nas séries dos índices de rede(Apêndice A). Neste trabalho, o estudo do vocabulário e suas conexões ao longo do tempo para uma RST é ampliado com o uso do método de entropia multiescala nas séries dos índices de rede e nas séries de entropia, calculadas. As séries de entropia são capazes de revelar momentos de maior e menor diversidade de palavras e de pares de palavras. Estas redes são investigadas de acordo com a identificação de vértices (palavras) mais importantes para manter a rede coesa numa configuração denominada de rede crítica. A entropia também é calculada na rede crescente de uma RST para identificar o surgimento de padrões no ganho de informação, a partir do surgimento de palavras novas, ao longo do tempo.

As seções que seguem descrevem a tese, sua importância, o problema e objetivos geral e específicos. Os demais capítulos trazem o referencial teórico, metodologia e resultados esperados.

1.1 Axiomas

Esta pesquisa se baseia em duas premissas referentes a redes semânticas e redes semânticas de títulos:

- Rede semântica representa o conhecimento (TEIXEIRA et al., 2010; PEREIRA et

al., 2016);

- Um título pretende sintetizar em uma frase as principais ideias de um trabalho (CUNHA et al., 2013).

1.2 O problema

Desde os estudos de Price (1965), diversas pesquisas apontam a necessidade de rever métricas para medir a produção e colaboração científica de diferentes áreas, como exemplo os trabalho de Mueller (2005) e Stumpf et al. (2017). Imersos neste cenário, cientistas da atualidade, em geral, desconhecem como sua área se configura nos veículos de comunicação científica e se determinado periódico científico tem potencial considerável para divulgar sua pesquisa ou não.

Com isso, a inexistência de métodos que acompanhem a diversidade do vocabulário e o avanço ou retrocesso de determinadas temáticas ao longo do tempo em periódicos científicos dificultam o aproveitamento do conhecmimento científico difundido. Portanto, faz-se necessário a existência de um método que quantifique a diversidade de vocabulário nos veículos de comunicação científica e que leve em conta as restrições que estes veículos recomendam à quem publica.

1.3 Justificativa

Com a diversidade dos meios de comunicação científica que publica produção qualificada, pesquisadores se deparam com questões importantes, por exemplo:

1. Como é a diversidade de vocabulário em um periódico científico?
2. Qual periódico oferece mais probabilidade de uma dada pesquisa ser bem difundida?
3. Em que épocas de um periódico teve maior proximidade com o vocabulário de um dado artigo?
4. Como uma temática evolui ao longo do tempo em um periódico?

Essas seriam uma das inúmeras questões associadas a eficácia da divulgação científica.

Os trabalhos que investigam o Periódico científico e a divulgação e disseminação da ciência (e.g. Redes semântica de títulos, Rede de palavras chaves, redes de coautoria, redes de

citação) buscam responder perguntas como as supracitadas utilizando a modelagem de Redes.

A modelagem por redes semânticas de títulos de artigos científicos variáveis no tempo aqui descrita oferece pistas para resolver estas questões, uma vez que possibilita investigar: a evolução da rede e seus índices no tempo; a diversidade de palavras e suas conexões a partir da entropia; o vocabulário que se destaca ao longo do tempo em termos de conexões; os efeitos de memória (correlação de longo alcance) para o vocabulário e suas tendências em diferentes escalas de observação.

A necessidade deste trabalho é também incentivada pela carência de estudos que relacionam diversidade de vocabulário em redes semânticas de cliques com entropia de vértices e arestas e com a complexidade (propriedades fractais) nas séries temporais dos índices de rede.

1.4 Hipoteses e suposições

1. A entropia da informação de Shannon é um indicador para medir diversidade de vocabulário de uma rede semântica títulos. O aumento da medida de entropia está associado ao surgimento de novas ideias expressadas nos títulos, enquanto sua diminuição associada à robustez e consolidação de ideias e interesses de autores e editores de uma revista em uma determinado marco temporal;
2. O processo de construção de uma redes de cliques impõe vínculos restritivos para os valores de entropia de seus vértices e arestas;
3. O uso de entropia em séries temporais de índices de redes para rede semântica de títulos captura tendências do vocabulário para um conjunto de títulos, bem como a previsibilidade dos indicadores de redes;
4. A configuração de rede crítica permite uma melhor investigação do vocabulário mais relevante em uma rede semântica e, consequentemente, do interesse de pesquisa.

1.5 Objetivos

1.5.1 Objetivo Geral

Elaborar um conjunto de métodos que (1) auxilie no acompanhamento do interesse de uma dada área do conhecimento a partir da evolução da rede semântica associada à produção

qualificada da área e (2) caracterize esta rede quanto à diversidade de seu vocabulário e conexões entre palavras.

1.5.2 Objetivos específicos

1. Elaborar um método que construa e analise uma rede semântica de títulos variável no tempo;
2. Compreender a dinâmica de uma rede semântica de títulos a partir das séries temporais de seus índices;
3. Elaborar um método que permita mensurar a diversidade de vocabulário de uma rede semântica de títulos, através do conceito de entopia da informação de Shannon;
4. Insvestigar como a entropia é influênciada pelas condições iniciais de uma rede de cliques;
5. Compreender a diversidade de vocabulário para uma rede semântica de títulos crescente no tempo;
6. Avaliar a previsibilidade dos valores de entropia em uma série temporal;
7. Avaliar a importância da existência de *rede crítica* em uma rede semântica de títulos que varia no tempo;
8. Elaborar um método que identifique o principal vocabulário e suas conexões nas redes críticas em momentos de alta e baixa diversidade de vocabulário.

Parte II

Fundamentação Teórica

Capítulo Dois

Redes Variáveis no tempo

A teoria de redes tem sido proeminente em pesquisas que estudam a estrutura e evolução de sistemas que contém elementos ou entidades que se conectam. Neste contexto, uma rede é a abstração desses elementos, chamados de vértices e suas relações são chamadas de arestas ou arcos. Uma rede pode ser estática ou variar no tempo.

2.1 Redes estáticas

Matematicamente, uma rede estática é representada por um grafo $G = (V, \mathcal{E})$, em que $V = \{v_1, v_2, ..., v_n\}$ é o conjunto de nós ou vértices, de n elementos e $\mathcal{E} = \{e_1, e_2, ..., e_m\}$ é o conjunto das arestas, contento m elementos, em que $e_k = \{(i, j)\}$. As arestas de uma rede podem ser ponderadas por algum parâmetro mensurável ou por uma função que represente a repetição de arestas entre o mesmo par de vértice (arestas múltiplas). Redes podem conter arestas dirigidas (arcos). Em outras palavras, uma aresta emana do vértice de origem v_i e incide no vértice destino v_j. Por exemplo, em uma rede de citação, um artigo a que cita um artigo b seria representado pelo arco (a, b). Note que em redes dirigidas $(v_i, v_j) \neq (v_j, v_i)$.

Como exemplo, no contexto do presente trabalho, a Figura 2.1 mostra uma rede de palavras. Estas palavras são oriundas de títulos de artigos científicos com a temática energia da *Nature* entre 1999 e 2008. As arestas conectam palavras pertencentes a um mesmo título, desde que, pelo menos uma extremidade contenha uma palavra relacionada a temática "energia"[1]. Na figura, a rede semântica apresentada refere-se a janela temporal $w_{507,0}$, a ser explicada na Seção 6.3) que significa que os dados foram capturados em um único instantâneo (parâmetro 0) em 507 semanas (parametro 507).

Podemos classificar uma rede de acordo as topologias clássicas:

- Regular, se todos os vértices possuírem o mesmo número de conexões;
- Aleatória, caso sua distribuição de graus siga uma distribuição normal. Este tipo de rede apresenta baixa aglomeração e a distância mínima média entre seus vértices é curta. Acreditava-se que as redes da natureza seguiam esta topologia. O marco para este estudo foi o trabalho de Erdos (1966);

[1]Para o exemplo aqui, qualquer palavra simples ou composta que tenha o termo "energia" em seu nome

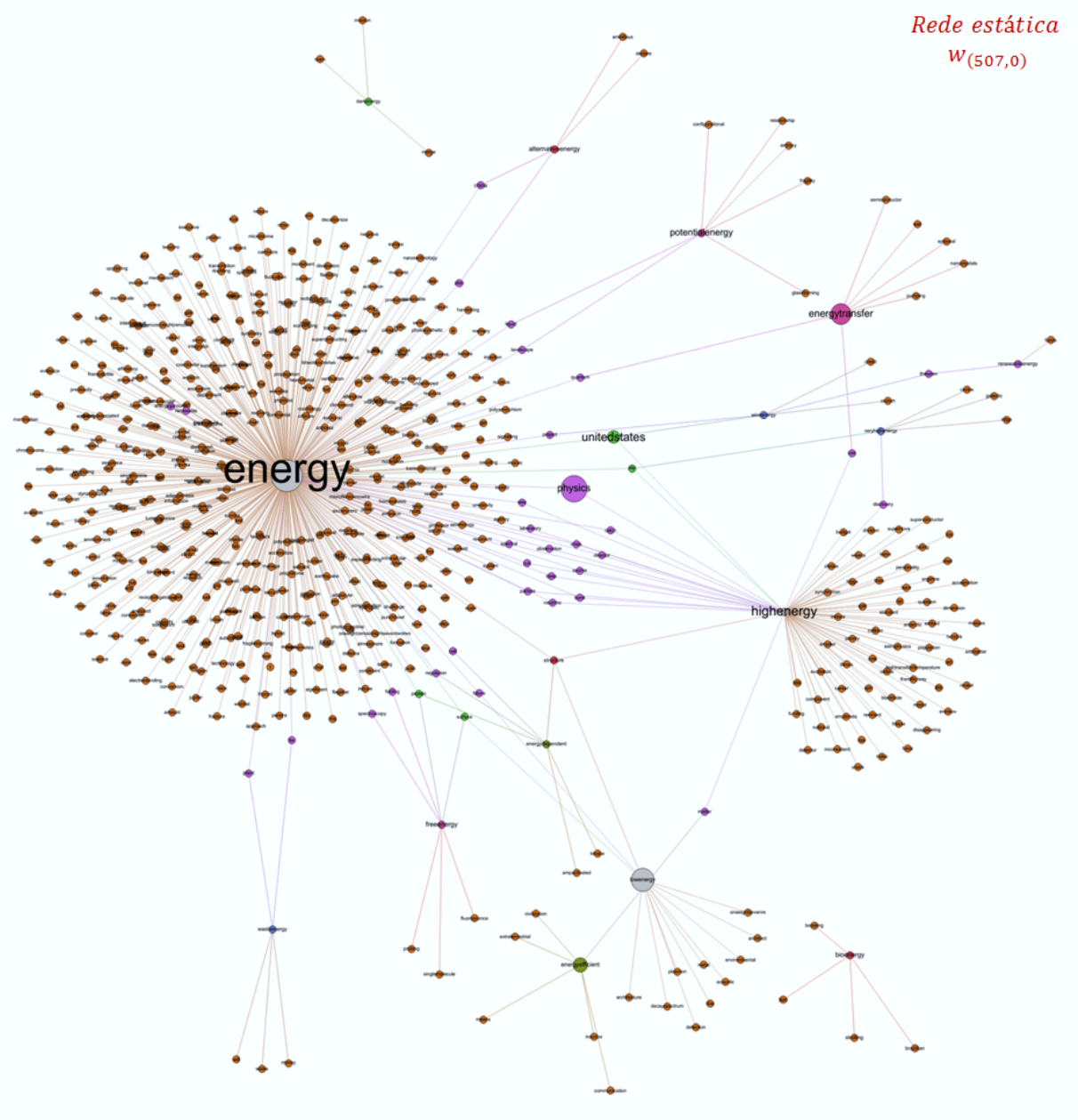


Figura 2.1: Rede da temática "energia" (rede i-"energy"), para a *Nature*, entre 1999 e 2008. As arestas conectam palavras, onde pelo menos uma delas contém o termo "energy". A rede é estática e corresponde a 507 semanas de dados ($w_{507,0}$).

- Mundo pequeno, caso as distâncias entre quaisquer dois vértices da rede sejam curtas, mas com alta aglomeração de arestas em torno dos vértices da rede. Esta rede é eficiente, do ponto de vista da transmissão de informação. Quanto mais atalhos[2] são adicionados entre os vértices, mais a rede se distância da topologia regular e, quanto maior a aglomeração média dos vértices da rede, mais a rede se distância do outro extremo (rede aleatória). A rede mundo pequeno encontra-se entre estes

[2]Um atalho é uma aresta que quando retirada faz com que o menor caminho entre dois vértices que estavam conectados previamente, tenha caminho maior ou igual a 2.

dois extremos. O marco para este estudo foi o trabalho de Milgram (1967) com o experimento das cartas e a observação dos seis graus de separação entre pessoas. Mas o fenômeno "mundo pequeno" só foi formalizado por Watts e Strogatz (1998).

- Livre de escala, caso sua distribuição de graus siga uma lei de potência, na forma $P(k) \sim k^{-\gamma}$. Neste caso, as conexões entre os vértices favorecem a existência de hubs (vértices que concentram muitas conexões). Esta rede é tolerante a falhas (ataques aleatórios), mas não tolerante a ataques coordenados, já que pólos (ou *hubs*) concentram a maior parte das ligações. Se *hubs* são "atacados" (removidos da rede), a estrutura topológica da rede é comprometida. Barabási e Albert (1999) propuseram um modelo de rede livre de escala baseado em duas propriedades: o crescimento contínuo e a adesão preferencial de vértices.

Para revelar a estrutura de uma rede e destacar vértices de alto prestígio é necessário o uso de algumas métricas (BARABÁSI, 2016): Caminho mínimo médio (L); diâmetro (D); densidade (Δ); aglomeração média (C); grau (k_i) de um vértice i; excentricidade de um vértice i ($e(i)$); grau médio ($\langle k \rangle$) e distribuição de graus ($P(k)$).

Muitas vezes torna-se necessário conhecer a dinâmica das conexões entre vértices para melhor analisar a rede. A seção seguinte apresenta uma introdução às redes variáveis no tempo e alguns métodos de análise que são fundamentais para este trabalho.

2.2 Redes temporais

A utilização de grafos com vértices e arestas fixos consegue representar bem as relações entre os vértices envolvidos em uma rede e, em geral, caracterizar bem um sistema. Em diversas redes (sociais e semânticas), autores têm investigado os sistemas como entidades estáticas, estudando suas características em um único instantâneo que abrange toda a evolução do tempo (CUNHA, 2013).

Entretanto, grande parte das redes reais[3] são fortemente influenciadas pela dinâmica de seus vértices (entrada e saída da rede) e mudanças das conexões entre eles. Assim, para um melhor estudo de sistemas deste tipo, é necessário considerar elementos temporais em seus conjuntos de vértices e arestas. Dentre as diversas formas de se estudar os efeitos do tempo em uma rede, existe uma modelagem bem interessante: a dos grafos variantes no tempo, conhecida como *Time-Varying Graphs* (TVG).

[3]Redes são modelos de sistemas "reais" ou teóricos. Neste trabalho, o termo "redes reais" se refere à redes observadas.

2.3 Breve revisão de literatura

Desde meados do século XX, cientistas já se preocupavam em considerar o tempo no estudo de grafos. Em 1958, Bellman propôs um algoritmo que minimiza o tempo de viagem entre duas cidades, de uma malha interligada, em função do tempo de viagem de duas cidades diretamente ligadas. O algoritmo minimiza o erro intuitivo de considerar o tempo de uma rota proporcional a distância entre duas cidades. Afinal, a determinação de um tempo ótimo entre dois pontos depende dos tempos de rotas de pontos intermediários diretamente ligados (BELLMAN, 1958).

Paralelamente, em 1958, Ford e Fulkerson propuseram um algoritmo que determina o fluxo máximo de mercadorias em uma rede de transporte por navios, etiquetando as arestas com o tempo de viagem (JR; FULKERSON, 1958). Em 1966, o algoritmo de Bellman é generalizado por Cooke e Halsey (1966). Os autores, preocupados com inúmeros problemas reais da física e biologia, propuseram um algoritmo que minimiza caminhos considerando o tempo entre dois nós, não mais constante como Bellman considerou, mas sim variáveis no tempo. Halpern e Priess (1974) resolveram a questão do caminho mais curto considerando a indisponibilidade de arestas por alguns períodos de tempo, sugerindo o uso na gestão de logística ferroviária. Halpern melhora seu algoritmo considerando o tempo de formação das arestas e de atraso para o início de uma rota, devido a indisponibilidade do nó (HALPERN, 1977). Décadas depois, em 1990, um trabalho explorou melhor esta questão de atrasos e desligamentos temporários de nós (ORDA; ROM, 1990).

Estes trabalhos, que são de cunho mais teórico, mais tarde foram implementados e melhorados a partir de redes dinâmicas reais, afim de se resolver importantes problemas de planejamento logístico (POWELL; JAILLET; ODONI, 1995). Esses estudos certamente contribuem para a tomada de decisão de gestores, evitando resultados aleatórios que prejudica quem utiliza a rede (e.g. um cliente).

No âmbito das redes sociais, recentemente alguns pesquisadores também tem considerado a influência do tempo nas conexões dos atores, seja na formação de comunidades ou na predições de relacionamentos entre indivíduos. Muitos deles propuzeram novas métricas, considerando o efeito do tempo.

Doreian e Stokman (1997)[4] aplicaram modelos de evolução para estudar o desenvolvimento de estruturas sociais. Barabási et al. (2002) destacaram mecanismos dinâmicos e estruturais presentes em uma rede de coautoria, caracterizando-a topologicamente em instantes de tempo diferentes. Li et al. (2007) estudaram uma rede de colaboração científica a partir de um modelo proposto por eles, que investiga padrões livre de escala nas distribuições dos pesos das arestas ao longo do tempo. Tang et al. (2010) introduzem

[4]Este livro já possui uma edição mais recente: Doreian e Stokman (2013).

conceitos de caminhos temporais e distância em grafos variáveis no tempo e definem o fenômeno *small-world* para um grafo variante no tempo, baseado na condição de alta aglomeração de arestas no tempo e baixa distância média temporal dos nós, em redes de agentes móveis e de sistemas sociais e biológicos.

Em 2012, Nicosia et al. (2012) e Casteigts et al. (2012) formalizaram diversos conceitos e métricas utilizados no estudo das redes dinâmicas, criando assim o conceito de *Time-Varying Graphs* (TVG), que permite modelar redes que possuem arestas e/ou vértices que variam no tempo. Com isso, foi possível integrar a vasta coleção de conceitos, formalismos e resultados encontrados na literatura em um quadro único, que expressa no mesmo formalismo, não só os conceitos comuns a várias áreas, mas também os específicos de cada área (NICOSIA et al., 2012).

Com esta metodologia, Amblard et al. (2011) investigaram as relações de coautoria e citações entre autores de artigos científicos; Silva et al. (2012) analisaram a evolução temporal de sinais cerebrais em redes de neurônios de ratos de comportamento livre; Cunha et al. (2013) investigaram o efeito de mémória em uma rede de títulos variante no tempo; Rad (2016) investigou padrões temporais em redes sociais como o Facebook e o Youtube, utilizando a modelagem *TVG*, com cálculo de centralidades que consideram distâncias temporais entre atores; Chen et al. (2016) discutiram detalhadamente a estrutura de transporte ótima, investigando a relação entre comprimentos temporais e geométricos, em sistemas variantes no tempo, analisando empiricamente o transporte aéreo austríaco; Cunha et al. (2020a) propuseram um método para analisar um *TVG*, a partir de uma janela de tempo deslizante, e criar séries temporais de índices de rede; e Sousa et al. (2020) desenvolveram um modelo denominado *Interação preferencial*, que reproduz uma rede ponderada livre de escala variante no tempo, para sistemas de número fixos de vértices, que pôde ser aplicado na investigação de redes de sinais de eletroencefalograma em indivíduos.

Com a crescente demanda, hoje já existem diversas maneiras de se estudar uma rede variável no tempo. Holme e Saramäki (2012) trouxeram diversas aplicações, sugestões de algoritmos e métricas específicas para redes que variam no tempo. Os autores discutem com mais profundidade no livro organizado por eles, intitulado "*Temporal Networks*" (HOLME, 2014)

2.4 Time-Varying Graphs (TVG)

Considerando a formalização de Casteigts et al. (2012), um grafo variante no tempo pode ser entendido como um grafo estático $G = (V, \mathcal{E})$ acrescido de parâmetros (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,

$$\mathcal{G} = (V, \mathcal{E}, \Upsilon, \varsigma, \Gamma), \tag{2.1}$$

onde:

- $V = \{v_1, v_2, ..., v_n\}$ é o *conjunto de vértices* da rede, que representa os objetos do sistema ou atores da rede, se esta for social;
- $\mathcal{E} = \{(e_1,..., e_m\}$ é o *conjunto de arestas*, sendo $e_k = \{(i, j)\}$; (com $i \neq j$)[5], que representa os relacionamentos entre os objetos do sistema, de acordo com algum critério de conexão;
- O conjunto $\Gamma \subset \mathbb{N}|\Gamma = \{t_1, t_2, t_3, ..., t, ..., t_{(\Psi-1)}, t_\Psi\}$ é o *tempo de vida* do sistema, ou seja, o conjunto discreto de instantes possíveis para existência da rede. O intervalo entre datas extremas é $T = t_\Psi - t_1 + 1$ é o tempo total de existência da rede;
- A função *latência* ς indica quanto tempo necessita para que uma aresta esteja disponível em um instante. Em outras palavras, é o tempo necessário para estabelecer o relacionamento entre dois vértices, em um dado instante t_i;
- A função *presença* $\Upsilon : \mathcal{E} \times \Gamma \rightarrow \{0, 1\}$ garante a existência de uma dada aresta em um dado instante de tempo t.

[5]Podem haver situações em que a aresta tem um único vértice como origem e alvo ($i = j$). Esta conexão é chamada de auto-laço. Há outro caso onde existem arestas "paralelas", ou seja, arestas diferentes para o mesmo par de nós. Arestas nesta situação são denominadas de arestas múltiplas.

Capítulo Três

Redes Semântica de Cliques

3.1 Rede de cliques

Uma clique consiste em um grafo ou subgrafo máximo, ao qual todos os vértices pertencentes à clique se conectam entre si. Uma rede de cliques é uma rede formada pela união de cliques, através dos processos de justaposição de arestas e sobreposição de vertices. A Figura 3.1 mostra como uma rede de cliques é formada a partir desses processos.

As redes de cliques ajustam-se à modelagem de vários sistemas sociais, por exemplo redes de atores de cinema (BARABÁSI; ALBERT, 1999), redes de coautoria (NEWMAN, 2001b), redes de conceitos (CALDEIRA et al., 2006) e redes semânticas (PEREIRA et al., 2011; PEREIRA et al., 2016; GRILO et al., 2017). Alguns trabalhos teóricos trataram destas redes, Derényi, Palla e Vicsek (2005) que abordam a percolação em grafos aleatórios por cliques; e Fadigas e Pereira (2013) que propuseram métricas específicas para estas redes, com exemplos de aplicação em rede semântica de títulos. Fadigas et al. (2020) estudaram redes de cliques constantes, chamadas de *qCliques*, e usaram como objeto a rede social dos árbitros que apitaram as copas do mundo de futebol da FIFA.

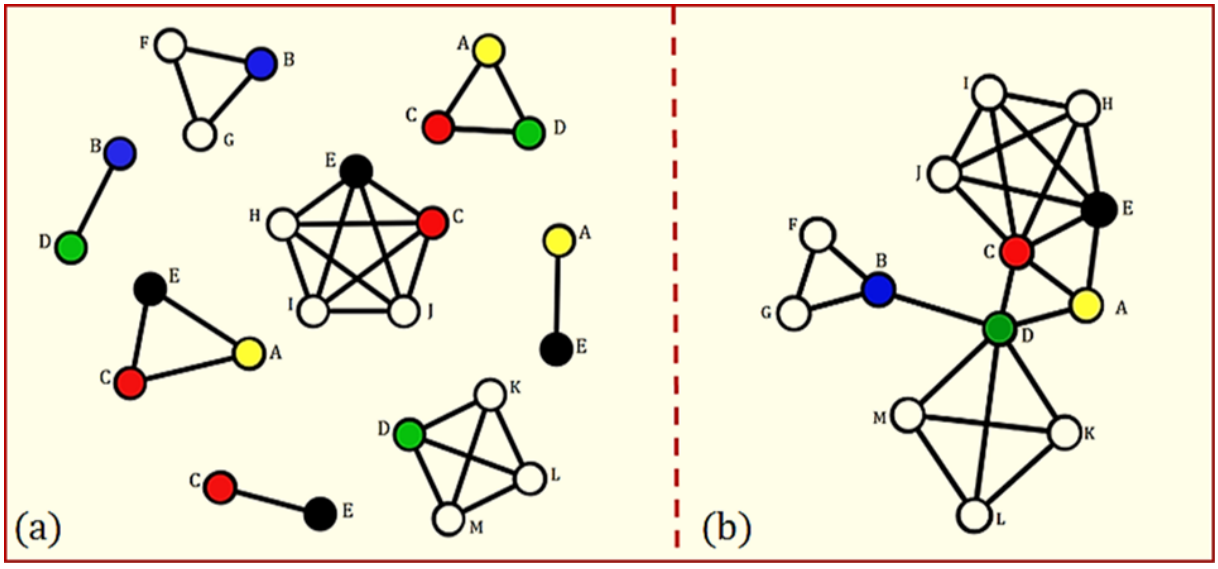


Figura 3.1: Formação da rede de cliques. Em (a), Cliques na configuração original; em (b), cliques unidas a partir de vértices comuns, gerando uma rede de cliques. Na justaposição, as cliques são unidas por apenas um vértice, enquanto na sobreposição a união é feita por pelo menos dois vértices e uma aresta.

3.2 Rede semântica de cliques

Os primeiros trabalhos sobre redes semânticas utilizaram esta modelagem para representar memória. Mais precisamente, a memória declarativa semântica de quem profere o discurso, que é a memória que agrupa tudo que pode ser evocado por palavras. É a partir dela que externalizamos nosso conhecimento sobre o mundo que nos cerca, desde sua história, acontecimentos, ciência, matemática e até sobre nossa própria história pessoal. (STERNBERG, 2000).

De acordo com a definição mais recente de rede semântica (PEREIRA et al., 2016; GRILO et al., 2017) e a premissa de que a sentença é a menor unidade de significado de um texto (CALDEIRA et al., 2006), podemos definir uma rede semântica de cliques como um sistema de representação de conhecimento estabelecido por contexto específico e imbuído de intenção de funcionalidade, onde vértices são elementos (palavras, conceitos ou entidades) com significado semântico e as arestas representam pares destes elementos que aparecem em uma mesma unidade de significado, por exemplo: uma frase de um texto ou discurso; um título de um artigo científico; palavras-chave de um artigo. Estas são as menores unidades de significado[1] do discurso e constituem uma clique na configuração inicial[2] da redes de cliques (CUNHA et al., 2020b).

Assim, a rede é a união dessas unidades menores de significado, ou seja, a união de cliques. Redes semânticas de cliques estão sendo cada vez mais estudadas, e.g., Caldeira et al. (2006) analisam a estrutura de conceitos significativos nos discursos escritos; TEIXEIRA et al. (2010), Neto, Cunha e Pereira (2018) usaram as redes de cliques semânticas para analisar a relação entre as palavras que emergem nos discursos orais, a partir da rede crítica, que é uma configuração encontrada usando o índice de incidência-fidelidade (Seção 3.4). Nesta configuração, a rede exibe mais informações com o mínimo de resíduo textual (TEIXEIRA et al., 2010); Nascimento et al. (2016) analisam uma rede semântica formada pelas palavras-chave de tese de doutorado na área de Ensino de Física, no Brasil, de 1996 a 2006; Andrade et al. (2019) utiliza as medidas de centralidades de grau, proximidade e intermediação para entender a coerência e consistência da proposta de um programa universitário com as ementas das disciplinas.

A próxima seção discute as redes semânticas baseadas em títulos de artigos científicos. Este tipo de rede também é uma rede semântica de cliques.

[1] Em um texto, a menor unidade de significado é a frase. A palavra muda de significado a depender de seus vizinhos na frase. Por isto que a palavra não é a menor unidade de significado de um texto ou discurso e sim a sentença (frase).

[2] A configuração inicial de uma rede de cliques é quando as cliques estão isoladas, antes da união para formar a rede de cliques, como visto em 3.1(a).

3.3 Rede semântica de títulos de artigos científicos (RST)

O conjunto de publicações que compõe uma revista científica faz parte de um sistema formal de comunicação: o periódico científico. 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, que publica, lê e cita artigos desta mesma comunidade (CUNHA, 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. O uso de redes semânticas baseadas em títulos de artigos científicos (RST) contribui para o estudo da colaboração científica dos que publicam em um mesmo periódico.

Uma RST pode ser capaz de apontar evidências dos temas que os autores tratam nos títulos de seus artigos e de como se comporta um periódico em relação às temáticas selecionadas para publicação. Por exemplo, Fadigas e Pereira (2013) afirmam que a quantidade de componentes de uma RST é influenciada pelo vocabulário dos Periódicos e com seu aspecto disciplinar.

Uma RST é uma rede semântica de cliques, em que cada clique representa um título e suas palavras são vértices de cliques. Consequentemente, as arestas representam as conexões entre palavras que pertencem ao mesmo título. Alguns autores propuseram importantes metodologias de estudo sobre RSTs Pereira et al. (2011) foram os pioneiros neste estudo. Os autores propuseram regras para tratamento manual e um método para coleta de dados, construção e análise de redes; Fadigas e Pereira (2013) utilizaram o mesmo conjunto de dados para aplicar índices específicos para redes de cliques, propostos por eles, e caracterizam as redes topologicamente através desses índices. Fadigas e Pereira (2013) estudaram a estrutura topológica de RSTs como um método para analisar a eficiência de difusão da informação, Henrique et al. (2014) usaram RST para comparar os títulos de artigos de periódicos no ensino de matemática em inglês e português; o trabalho Cunha et al. (2013) propôs uma RST variável no tempo e observou um efeito na memória da rede temporal; Cunha, Miranda e Pereira (2015) aplicaram o índice de *incidência-fidelidade* em RSTs em 15 periódicos com alto fator de impacto e encontraram a rede crítica correspondente para cada um; Pereira et al. (2016) estudaram a evolução da densidade durante a construção de redes semânticas como um indicador da diversidade de conceitos de periódicos científicos; e Grilo et al. (2017) propuseram um método que analisa a robustez de uma RST, utilizando estratégias de remoção de vértices, possibilitando identificar uma fração de remoção crítica para a qual a estrutura topológica da rede é alterada.

Na Figura 3.2, 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 são unidas, justapondo a palavra comum. Este procedimento, estendido a todos os títulos de uma revista gera a Rede Semântica de Títulos.

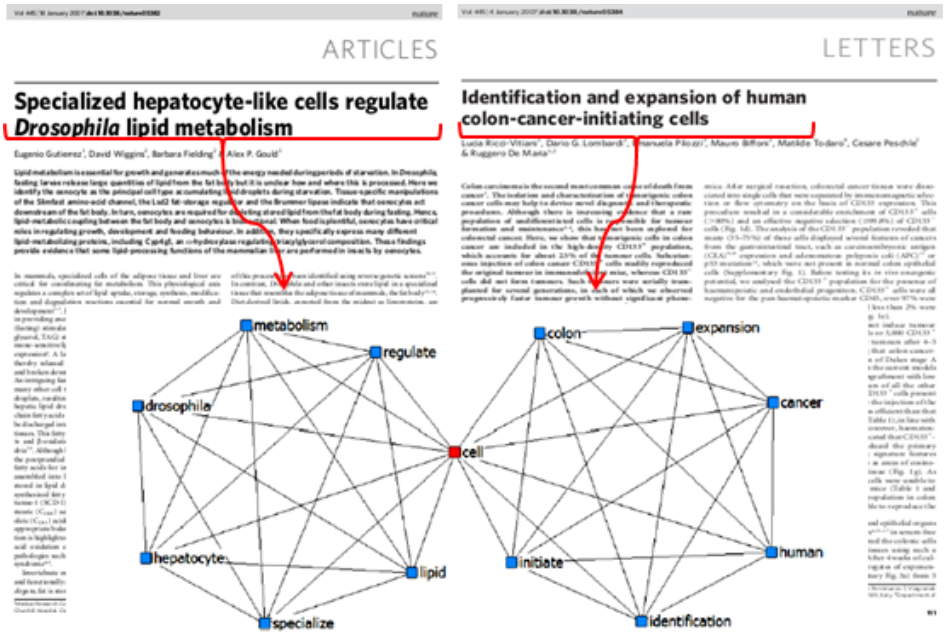


Figura 3.2: Títulos de dois artigos da *Nature* e o processo de construção da rede semântica de títulos (CUNHA, 2013).

Utilizando a modelagem *TVG* aplicada a rede de títulos, propomos neste trabalho, um método de construção e análise de RSTs, detalhado na Seção 6.2 .

3.4 Incidência-fidelidade

Com base na premissa de Caldeira et al. (2006), palavras que ocorrem juntas na mesma frase foram evocadas associativamente na construção da ideia a ser apresentada. TEIXEIRA et al. (2010) acrescentam que, com base nesse critério, pares cuja associação não é muito significativa acabam sendo incluídos na rede e ocultam a estrutura formada pelas associações mais fortes. Desta forma, é necessário filtrar, para que apenas as associações mais relevantes para o discurso sejam consideradas na construção da rede.

Para filtrar uma rede semântica de cliques e encontrar a rede ideal, TEIXEIRA et al. (2010) criaram o índice incidência-fidelidade (IF). Sua aplicação gera uma rede, de configuração crítica, na qual existe o máximo informação com o mínimo de resíduo textual.

Este índice mede o quão "forte" e "fiel" é a relação entre um par de palavras. Considera, para um determinado par de palavras, a frequência de aparecimento no texto (incidência I, Equação 3.1) e a frequência de aparecimento no contexto em que pelo menos uma palavra do par foi evocada (fidelidade F, Equação 3.2) A *incidência-fidelidade* é, portanto, o produto destes dois índices, Equação 3.3.

$$I_{(\alpha,\beta)} = \frac{|C_\alpha \bigcap C_\beta|}{|\bigcup_{i=1}^{n} C_i|} = \frac{|S_{\alpha,\beta}|}{n_q} \tag{3.1}$$

$$F_{(\alpha,\beta)} = \frac{|C_\alpha \bigcap C_\beta|}{|C_\alpha \bigcup C_\beta|} \tag{3.2}$$

$$IF_{(\alpha,\beta)} = I_{(\alpha,\beta)} \times F_{(\alpha,\beta)} = \frac{|S_{\alpha,\beta}|}{n_q} \cdot \frac{|S_{\alpha,\beta}|}{|S_\alpha + S_\beta - S_{\alpha,\beta}|} \tag{3.3}$$

Nas equações, α e β representam as palavras que compõem um par de palavras; C_i o conjunto de sentenças[3] que contêm a palavra i; S_α, S_β e $S_{\alpha,\beta}$ são o número de sentenças que aparecem, respectivamente, a palavra α, a palavra β e o par de palavras (α, β); n_q é o número total de sentenças do texto (número de cliques). Assim, uma vez calculado o IF para todos os pares de palavras, a rede semântica torna-se ponderada nas arestas.

Com isso, o índice IF pode atuar como um filtro para a rede. A filtragem é feita removendo as arestas com o valor $IF < IF_L$, deixando apenas as arestas com $IF \geq IF_L$ na rede. IF_L é o valor mínimo permitido na rede para o índice IF, arbitrado pelo pesquisador.

3.4.1 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; NETO; CUNHA; PEREIRA, 2018), quanto em discursos escritos, e.g. autores romances (AGUIAR, 2009) e em redes de títulos (CUNHA; MIRANDA; PEREIRA, 2015).

Para encontrar o valor de IF_L que representa a rede crítica de um discurso($IF_L = IF_C$), usa-se o índice *caminho mínimo médio*[4](L). A análise consiste em verificar o que acontece com o valor de L a medida que se aumenta o valor de IF_L. Observa-se que, a partir da rede

[3] O termo "sentença" de um texto é o mesmo que "frase" de um texto e o mesmo que clique em uma rede semântica de cliques na configuração original.

[4] Este comportamento também ocorre com outros índices, como o Diâmetro D e a diferença normalizada entre vértices e arestas (TEIXEIRA et al., 2010)

original, a medida que se aumenta o valor de IF_L o, valor de L da rede correspondente aumenta, atinge um valor máximo e em seguida diminui bruscamente.

O índice IF_L atua como um filtro que "limpa" o texto a medida que seu valor cresce. Inicialmente, o valor de L aumenta - já que a rede perde atalhos entre os vértices - até um valor máximo, onde $IF_L = IF_C$, em que IF_C é o valor da *incidência fidelidade crítica*. Neste ponto a rede apresenta algumas propriedades de redes maximizadas, como o caminho mínimo médio (L) e a diferença normalizada entre vértices e arestas [5]. A partir deste ponto, um pequeno incremento no valor de IF_L faz a rede se quebrar e o valor de L cai bruscamente.Segundo TEIXEIRA et al. (2010), esta configuração é denominada *rede crítica* e possui um padrão modular interessante em uma rede que representa bem o discurso, com o máximo de informação e o mínimo de resíduo textual.

[5]Isto é obtido a partir da normalização do número de vértices e do número de arestas pelos seus respectivos valores máximos. O ponto onde o valor absoluto da diferença entre estes valoress é máximo corresponde a um comportamento crítico da rede (TEIXEIRA et al., 2010).

Capítulo Quatro

Teoria da informação

A teoria da informação tem evoluido nas últimas décadas, encontrando aplicação em diferentes campos, tais como as telecomunicações, computação, física pura, genética, ecologia e discussão do processo fundamental de observação científica (BRILLOUIN, 2013).

Em 1928, Ralph Hartley estudou maneiras de medir a informação transmitida por sinais elétricos (HARTLEY, 1928). Em seu trabalho, sugeriu que a informação surgiria da seleção sucessiva de símbolos pertencentes a dado vocabulário, propondo uma formalização matemática para o conceito baseado em logaritmo.

Vinte anos depois, Claude Shannon desenvolveu matematicamente o conceito de informação (SHANNON, 1948) para quantificar seu armazenamento e transmissão em canais com e sem ruído, utilizando também técnicas como códigos de correção de erros. De maneira simplificada, estudou como determinar a espessura mínima de um fio que unisse duas centrais telefônicas de maneira eficiente ou, visto de outra forma, como armazenar e transmitir informação de maneira mais econômica (LIMA et al., 2012).

A informação contida em uma mensagem está associada a quantidade dos possíveis valores que esta mensagem pode ter. Se um sistema possui apenas um estado possível (e.g. rede regular), nenhuma informação é obtida ao ser inspecionado. Quanto mais estados possíveis para um sistema, mais informação ele contém, ou seja, mais poderemos "aprender" pela descoberta de seu estado real. Em outras palavras, a cada exclusão de possibilidade, mais informação se tem do sistema.

Para Shannon, não interessa o significado de uma mensagem, mas sim, que cada mensagem é o resultado de uma escolha entre um conjunto de mensagens possíveis. Isto quer dizer que de um conjunto de mensagens possíveis, apenas uma foi selecionada (SHANNON, 1948).

Para ajudar no entendimento dos aspéctos matemáticos da teoria, trazemos na Seção 4.1 uma revisão da teoria das probabilidades. Para maior aprofundamento ver o trabalho de Papoulis (1990).

4.1 Probabilidades

Em ciência e também no cotidiano, é comum nos depararmos com situações não determinísticas. Mesmo não conhecendo o resultado de um experimento, podemos conhecer os resultados possíveis e assim mensurar a incerteza de cada um dos acontecimentos possíveis. Para obter esta medida, utilizamos o conceito de probabilidade.

Seja $\Omega = \{w_1, ..., w_n\}$ o espaço amostral, ou seja, o conjunto dos resultados possíveis de um experimento aleatório. Os subconjuntos deste espaço são chamados de eventos. Assumindo que todos os resultados sejam equiprováveis, o número de resultados favoráveis a um evento A, dividido pelo número de resultados possíveis, é a probabilidade deste evento, Equação 4.1.

$$P(A) = \frac{|A|}{|\Omega|} \tag{4.1}$$

onde que $|A|$ e $|\Omega|$ são as cardinalidades dos conjuntos A e Ω.

4.1.1 Probabilidade Condicional

Considere A e B como dois eventos de um mesmo espaço amostral, com $P(B) \neq 0$. A probabilidade condicional é definida pela Equação 4.2, ou seja, a probabilidade de A, dado B. Em outras palavras, sabendo que o evento B ocorreu, o evento A ocorre se, e somente se, ocorre a intersecção $A \cap B$.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{4.2}$$

Para quaisquer dois eventos A e B, sabendo que $P(A \cap B) = P(B \cap A)$ e utilizando a equação acima, pode-se deduzir facilmente a Equação 4.3. Esta equação faz parte do *teorema de Bayes*, muito importante em teoria das probabilidades. Ele mostra como alterar as probabilidades *a priori* tendo em conta novas evidências de forma a obter probabilidades *a posteriori*.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{4.3}$$

4.1.2 Variável aleatória

Uma variável aleatória $X|X : \Omega \to \mathbb{R}$ é uma função que associa cada elemento do espaço amostral Ω, seu domínio, a um subconjunto dos reais $X(\omega) \in \mathbb{R}$, seu contradomínio. Assim, o evento $\{w \in \Omega : X(w) = x_i\}$ pode ser representado por $\{X = x_i\}$ e sua probabilidade $P[w \in \Omega : X(w) = x_i] = P[X = x_i]$.

A Probabilidade é então uma função $P(\cdot)$ que associa valores numéricos a um evento A do espaço amostral, e que satisfaz as seguintes condições:

- (1) $P(\Omega) = 1$, $P(\phi) = 0$;
- (2) $0 \leq P(A) \leq 1$;
- (3) $P(A \bigcup B) = P(A) + P(B)$, se e somente se, $A \bigcap B = 0$.

Estes axiomas asseguram que as probabilidades podem ser interpretadas como frequências relativas. Isto é importante neste trabalho para definirmos as probabilidades de vértices e arestas em redes de cliques.

Por fim, uma variável aleatória é a representação numérica real de um evento, por uma variável que assume um valor dentre um conjunto de valores possíveis.

Neste contexto, a incerteza do valor de uma variável será medida pela sua *auto-informação*.

4.2 Auto-informação

Considere uma variável aleatória que pode tomar qualquer um dos valores no conjunto finito $X = \{x_1, x_2, ..., x_n\}$, de elementos estatisticamente independentes e com vetor de probabilidade $p(x) = \{p_1, p_2, ..., p_n\}$, com $p_i \geq 0$; $\sum_{i=1}^{n} p_i$. A *auto-informação* de x_i, I_i, é definida pela Equação 4.4.

$$I_i = -\log_2 p_i \tag{4.4}$$

Na Equação 4.4, a base do logaritmo poderia ser qualquer uma, e ela determina a unidade da *auto-informação*. Neste trabalho utilizaremos a base 2, tornando sua unidade de medida o *bit*. Assim, I_i é, em *bits*, a redução de incerteza provocada pela observação do

símbolo x_i. Quanto mais provável for x_i, menos informação ele agrega à observação da variável x.

Segundo este raciocínio, a informação está associada ao grau de liberdade que se tem da fonte ao selecionar uma mensagem, considerando todo o processo de seleção das possíveis mensagens. Em outras palavras, à "surpresa" do receptor ao receber uma mensagem. Caso este receba um símbolo pouco provável para a variável, então sua surpresa é alta, ou seja, a informação.

A informação é uma medida em termos de decisões. Para a medida dada em bits, por exemplo, o valor da auto-informação corresponde ao número mínimo de perguntas binárias (*sim/não*, 0/1, *ligado/desligado*) que são necessárias para atribuir um objeto ao seu estado correto.

Para estimar quanta informação em média será enviada pela fonte no próximo símbolo, utiliza-se a medida *entropia da informação*.

4.3 Entropia da informação

A informação média $H(x)$ associada aos n símbolos da fonte é dada pela média ponderada das auto-informações de seus símbolos, podenrados por suas probabilidades, Equação 4.5.

$$H(x) = \sum_{i=1}^{n} -p_i \cdot \log_2 p_i \tag{4.5}$$

A entropia então representa a incerteza média de uma variável aleatória. Ao usar a base do logaritmo 2, ela mede o número de *bits* necessários para descrever a variável (COVER; THOMAS, 2012).

Embora o receptor não possa prever qual o próximo símbolo, ele espera obter $H(x)$ bits de informação por símbolo, ou $n \cdot H(x)$ para n elevado, em uma mensagem de n símbolos. A Equação 4.5 possui valor máximo quando os símbolos são equiprováveis ($p_i = \frac{1}{n}$, $\forall i = \{1, 2, ..., n\}$), ou seja, $H(x)_{max} = \sum_{i=1}^{n} -\frac{1}{n} \cdot \log_2 \frac{1}{n} = -\log_2 \frac{1}{n}$ (Equação 4.6).

$$H_{max} = \log_2 n \tag{4.6}$$

A Equação 4.6 mostra que para um dado conjunto de símbolos, a entropia será máxima

se estes forem equiprováveis. Além disso, a equação deixa claro que a entropia também aumenta com o aumento da quantidade de possibilidades para estes estados. Ou seja, dentre quaisquer dois conjuntos com elementos equiprováveis, terá maior entropia o que tiver maior quantidade de elementos (WEAVER, 1953).

Por outro lado, o valor mínimo para a entropia ocorrerá quando a variável tiver apenas um estado possível (incerteza nula), ou seja $H(X) = 0$. Assim, o valor da entropia da informação de uma variável aleatória X está situado entre 0 e $\log_2 n$ (Equação 4.7).

$$0 \leq H(X) \leq \log_2 n \qquad \text{bits} \tag{4.7}$$

4.3.1 *Entropia conjunta*

Sejam X e Y variáveis aleatórias com distribuições $p_i = P(X = w_i)$ e $q_j = P(Y = w_j)$. A distribuição conjunta de (X, Y) é caracterizada pelo vetor de probabilidades $p_{ij} = P(X = w_i, Y = w_j)$. Então, a entropia conjunta do vetor aleatório (X, Y) é dada pela Equação 4.8:

$$H(X,Y) = -\sum_{ij} p_{ij} \times \log_2 p_{ij} \tag{4.8}$$

onde $H(X) = -\sum_{ij} p_{ij} \times \sum_j \log_2 p_{ij}$ e $H(Y) = -\sum_{ij} p_{ij} \times \sum_i \log_2 p_{ij}$

Verifica-se que $p_i = \sum_j p_{ij}$ e $q_j = \sum_i p_{ij}$, onde p_{ij} é a distribuição de (X, Y). Sendo assim, quando X e Y são independentes temos que $p_{ij} = p_i \times q_j$ e a entropia conjunta atinge um máximo (Equação 4.9):

$$H(X,Y) = H(X) + H(Y) \tag{4.9}$$

Ou seja, uma variável não carrega informação sobre a outra e o total da informação no par é a soma das informações em cada variável. Esta é uma condição de máximo. Qualquer situação onde haja dependência entre as variáveis, a entropia conjunta é menor que a soma das entropias de cada uma, conforme sintetizado na Equação 4.10, conhecida como Subaditividade da entropia de *Shannon*.

$$H(X,Y) \leq H(X) + H(Y) \tag{4.10}$$

4.3.2 Entropia condicional e informação mútua

A entropia condicional $H(X|Y)$ mede a incerteza que temos do valor de uma variável X dado que conhecemos o valor de Y. Mas, no caso de independência entre elas, $H(X|Y) = H(X)$, ou seja, o conhecimento de Y nada traz de informação para se conhecer X.

A subaditividade (Equação 4.10) mostra que pode existir correlação entre variáveis. Em outras palavras, ao aprendermos sobre uma podemos também ganhar informação sobre a outra. Seguindo este raciocínio, a quantidade de informação sobre X que não está em Y poderia ser encontrada pela Equação 4.11.

$$H(X|Y) = H(X,Y) - H(Y) \tag{4.11}$$

A informação mútua contida no par X e Y, $H(X:Y)$, por outro lado, mede a quantidade de informação que as variáveis possuem em comum. Na soma das informações de cada variável, $H(X) + H(Y)$, a informação em comum é contada duas vezes, enquanto que a informação não comum é contada apenas uma vez. Sendo assim, basta subtrairmos desta soma a informação conjunta, $H(X,Y)$, para obtermos a informação mútua de X e Y, ou seja, a informação comum das duas variáveism, Equação 4.12.

$$H(X:Y) = H(X) + H(Y) - H(X,Y) \tag{4.12}$$

Combinando as Equações 4.12 e 4.11, chegamos a Equação 4.13.

$$H(X:Y) = H(X) - H(X|Y) \tag{4.13}$$

4.4 Entropia da informação em redes semânticas

Alguns trabalhos dedicaram o uso da entropia da informação de Shannon em redes semânticas. De acordo com Solé et al. (2002), as redes sociais e complexas apresentam estruturas heterogêneas que resultam de diferentes mecanismos de evolução. A heterogeneidade da distribuição de graus de uma rede pode ser mensurada a partir do conceito

de Entropia. Neste caso, o sistema é totalmente previsível e ordenado em redes regulares e completamente heterogêneo e desordenado em redes aleatórias (SOLÉ; VALVERDE, 2004). Estes autores buscaram um algoritmo de otimização que aborda simultaneamente heterogeneidade e correlações em redes reais.

Carpi et al. (2011) propuseram o uso de uma métrica baseado na Entropia da Informação para analisar a evolução de redes de mundo pequeno. A escolha desta topologia está associada ao fato de que encontram-se entre as redes regulares (entropia mínima) e as redes aleatórias (entropia máxima).

Rosa (2017) em sua tese de doutorado investigou a robustez de redes semânticas de títulos utilizando diversos indicadores, dentre eles a entropia da informação de Shannon (Equação 4.14), na qual utiliza a distribuição de graus remanescentes $q(k)$, proposta por Solé e Valverde (2004) (Equação 4.15).

$$H = -\Sigma q(k) \log_2(q(k)) \tag{4.14}$$

onde $q(k)$ é a distribuição de graus remanescentes de i (Equação 4.15).

$$q(k) = \frac{(k+1)P_{k+1}}{< k >} \tag{4.15}$$

Os trabalhos de entropia em redes, frequentemente focam nas distribuições de graus dos vértices. Entretanto há carência de trabalhos específicos para redes de cliques, que leve em conta o processo de formação das redes, i.e. as distribuições de frequência de vértices e arestas na configuração inicial (cliques isoladas). Este método é proposto neste trabalho e motivou o trabalho de (CUNHA et al., 2020b).

O próximo capítulo traz aspéctos teóricos sobre o método *MSE (Multi Scale Entropy)*. Este método aplica a entropia de Shannon em séries temporais, o que nos permite uma aplicação paraas séries dos índices das redes deste trabalho.

Capítulo Cinco

Entropia de Multi Escala (MSE)

Para compreender melhor o método de entropia de multi escala, abordaremos conceitos básicos necessários para sua aplicação: (i) séries temporais; (ii) entropia aproximada e (iii) entropia amostral.

5.1 Séries temporais

Uma série temporal, conforme implícita no nome, é uma sequência de valores que representam a evolução temporal de uma determinada variável, oriundas de um processo estocático. Consideraremos nesta abordagem séries discretas. Sendo assim, cabe uma aplicação a analise de redes que variam no tempo a partir das séries temporais dos valores dos ínidices de rede.

Ao analisar uma série temporal, é comum partirmos da premissa de que cada valor depende ou influencia os valores de sua vizinhança. Para validar essa hipótese e investigar a dependência entre valores, vários métodos de regressão linear têm sido desenvolvidos nas últimas décadas. O *Detrended Fluctuation Analysis (DFA)* (PENG et al., 1994) tem sido aplicado em diversas áreas e identifica memória (correlações de longo alcance) nas séries. O Apêndice A traz uma explicação detalhada com uma aplicação para rede de títulos, no trabalho precursor desta tese (CUNHA, 2013); o *Multi Scale Entropy (MSE)*, explicado na Seção 5.4, identifica complexidade nas séries, ao calcular a entropia em escalas diferentes. A título de curiosidade, Ary L. Goldberger e C.-K. Peng participaram da construção destes dois métodos (DFA e MSE).

5.2 Entropia aproximada

Muitos autores têm estudado formas de analisar a complexidade de uma série temporal a partir da entropia de seus dados. A entropia de *Kolmogorov-Sinai* (*entropia KS* ou S_{KS}) avalia a taxa de crescimento de entropia da informação em séries para avaliar sistemas dinâmicos quanto a presença de caos.

Valores de $S_{KS} = 0$ indicam que não há informação sendo criada, ou seja, não há incerteza sobre sua dinâmica e seu movimento deve ser regular; valores finitos de $S_{KS} > 0$ indicam sistemas caóticos, ou seja, que produzem novas informações ao longo de sua evolução,

sendo maior sua complexidade quanto maior for seu valor; para $S_{KS} \to \infty$, o sistema é aleatório, com incerteza máxima.

Devido a dificuldade prática de aplicar a *entropia KS* em séries finitas com poucos pontos (abaixo de 10000), outros métodos de aproximação da *entropia KS* surgiram. Pincus (1991) propôs a *Entropia aproximada* (*ApEn*), que é uma família de entropias, dada pela Equação 5.1:

$$ApEn(m, r, N) = \Phi^m(r) - \Phi^{m+1}(r) \tag{5.1}$$

A equação é aplicada para uma série $\{u_i : 1 \leq i \leq N\}$ de tamanho N. O parâmetro m corresponde ao tamanho da janela de um padrão, pelo vetor $\mathbf{X}_i^m = \{u_i, u_{i+1}, ..., u_{i+m-1}\}$, sendo $i = \{1, ..., N - m + 1\}$, que se repete ao longo da série, com tolerância r para discernimento entre os valores. Esta contagem de padrões pode ser entendida utilizando a Figura 5.1. Assim, temos a Equação 5.2:

$$\Phi^m(r) = (N - m + 1)^{-1} \sum_{i=1}^{N-m+1} \ln C_i^m(r) \tag{5.2}$$

onde, $C_i^m(r) = \dfrac{1}{N - m + 1} \times \{\text{número de vetores } X_j^m \text{ tais que } d[\mathbf{X}_i^m, \mathbf{X}_j^m] \leq r\}$ e $d[\mathbf{X}_i^m, \mathbf{X}_j^m] = \Big(|u(i + k - 1) - u(j + k - 1)|\Big)_{max}$, com $k = 1, ..., m$.

A contagem de padrões é feita iniciando em u_1, a partir de um valor de m, tipicamente é usado $m = 2$. Neste caso, é contado o número de padrões que se repetem, dentro da tolerância $u_i \pm r$, com m e $m + 1$ valores. O procedimento recomeça em u_2 e os números de sequências que combinam cada uma das sequências de modelo de 2 e 3 componentes são novamente contabilizados e adicionados aos valores anteriores (Figura 5.1).

Este processo é então repetido para todos as outras sequências possíveis do modelo, para determinar a relação entre o número total de combinações de modelo de 2 componentes e o número total de correspondências de modelo de 3 componentes.

5.3 *Entropia Amostral*

A entropia aproximada (Equação 5.1) oferece limitações. Na Equação 5.2 o próprio padrão X_i^m é contado, evitando que $C_i^m(r) = 0$, fazendo com que $\ln C_i^m(r)$ sempre exista, mesmo

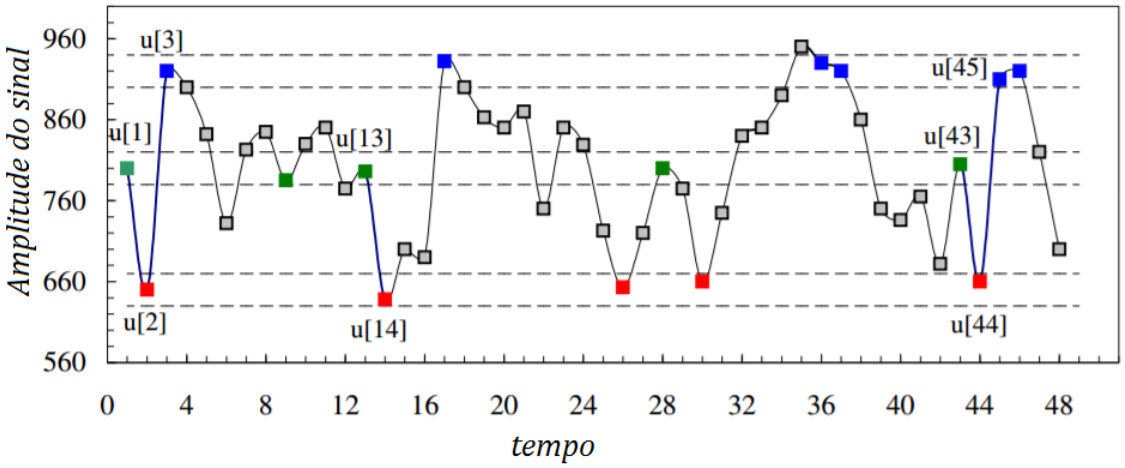


Figura 5.1: Para o caso de $m = 2$ e $r = 20\%$ (tipicamente r está entre 10% e 20% do desvio padrão da amostra. As linhas espaçadas na horizontal indicam a tolerância. Ou seja, todos os pontos que estão no intervalo de tolerância $u(1)$, $u(1) \pm r$, são mostrados em verde, igualmente para intervalos de u(2), em vermelho e de u(3), em azul. Partindo de $u(1)$, existem duas sequências verdes-vermelhas, {u(13), u (14)} e {u(43), u(44)} que correspondem ao padrão {u(1), u(2)}, e apenas uma sequência verde-vermelho-azul que corresponde ao padrão {u(1), u(2), u(3)}. Portanto, neste caso, o número de sequências que correspondem às sequências do modelo de 2 componentes é dois e o número de sequências que correspondem à seqüência do modelo de 3 componentes é 1. Estes cálculos são repetidos para a próxima seqüência de modelo de 2 componentes e 3 componentes , que são {u(2), u(3)} e {u(2), u(3), u(4)}), respectivamente. Fonte: Adaptado de Costa, Goldberger e Peng (2005).

que a contagem de padrões semelhantes tenda a zero. A contagem da própria ocorrência atua como se a série fosse perfeitamente regular, e isto está em desacordo com o conceito de criação de informação.

Para corrigir esta limitação, foi proposto por Richman e Moorman (2000) o método da entropia amostral ou *Sample Entropy* (*SampEn*) que consiste em cálcular o logaritmo natural da contagem de padrões supracitada, sem considerar auto-contagem (Equação 5.3). Esta medida reflete a probabilidade de que as sequências que se combinem entre os dois primeiros pontos de dados também coincidam para o próximo ponto.

$$SampEn(m, r, N) = -\ln \frac{U^{m+1}(r)}{U^m(r)} \tag{5.3}$$

Considerando os mesmos vetores definidos no cálculo de *ApEn*, $U^m(r)$ é dado pela Equação 5.4

$$U^m(r) = \frac{1}{N-m} \sum_{i=1}^{N-m} U_1^m \tag{5.4}$$

Com U_i^m e U_i^{m+1} dados pelas Equações 5.5 e 5.6.

$$U_i^m = \frac{1}{N-m-1} \times \{\text{número de vetores } j \neq i \text{ tais que } d[\mathbf{X}_i^m, \mathbf{X}_j^m] \leq r\} \tag{5.5}$$

$$U_i^{m+1} = \frac{1}{N-m-1} \times \{\text{número de vetores } j \neq i \text{ tais que } d[\mathbf{X}_i^{m+1}, \mathbf{X}_j^{m+1}] \leq r\} \tag{5.6}$$

5.4 Método MSE

O Método *Multi Scale Entropy* possibilita o cálculo da entropia de uma série temporal em múltiplas escalas de tempo. Costa, Goldberger e Peng (2005) propuseram este método para analisar estatisticamente amostras de batimentos cardíacos. Eles descobriram que pessoas saudáveis e pacientes com problemas cardíacos podem ser consistentemente diferenciados por uma medida surpreendentemente simples baseada no conceito de entropia.

Em uma série unidimensional discreta no tempo, $x(i) = \{x(1), x(2), ..., x(N)\}$, que descreve um sistema no tempo, é razoável supor que o estado do sistema em um determinado instante t_i seja parcialmente determinado pela sua história, $\{t_1, t_2, ..., t_{i-1}\}$. No entanto, cada novo estado carrega uma quantidade adicional de novas informações. Essa criação de informação nova, ou seja, a diminuição da incerteza em um receptor, conhecendo o estado atual do sistema e seu histórico, pode ser medida pela sua entropia.

Antes de mensurar entropia, a partir da série original, são construídas séries consecutivas $y = y^{(\tau)}(j)$, em que τ é o fator de escala. Este processo é denominado de "*Coarse Graining*" (COSTA; GOLDBERGER; PENG, 2002a; COSTA; GOLDBERGER; PENG, 2005).

Para $\tau = 1$ temos a série original, de tamanho N (Equação 5.7).

$$y^{(1)}(j) = x(i) = \{x(1), x(2), ..., x(N)\} \tag{5.7}$$

Para $\tau = 2$ temos a série, de tamanho $N/2$, das médias dos dados consecutivos dois a dois sem sobreposição (Equação 5.8).

$$y^{(2)}(j) = \{\frac{x(1)+x(2)}{2}, ..., \frac{x(i)+x(i+1)}{2}, ..., \frac{x(N-1)+x(N)}{2}\} \tag{5.8}$$

Para $\tau = 3$ temos a série, com tamanho $N/3$, das médias de cada grupo de três dados consecutivos sem sobreposição (Equação 5.9).

$$y^{(3)}(j) = \{\frac{x(1)+x(2)+x(3)}{3}, ..., \frac{x(i-1)+x(i)+x(i+1)}{3}, ..., \frac{x(N-2)+x(N-1)+x(N)}{3}\} \tag{5.9}$$

Este processo se repete até um valor limite para τ, definido pelo pesquisador. Resumidamente, cada série $y^{(\tau)(i)}$ pode ser calculada pela Equação 5.10. O método é também ilustrado na Figura 5.2.

$$y^{(\tau)}(j) = \{\frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x_i, 1 \leq j \leq \frac{N}{\tau}\} \tag{5.10}$$

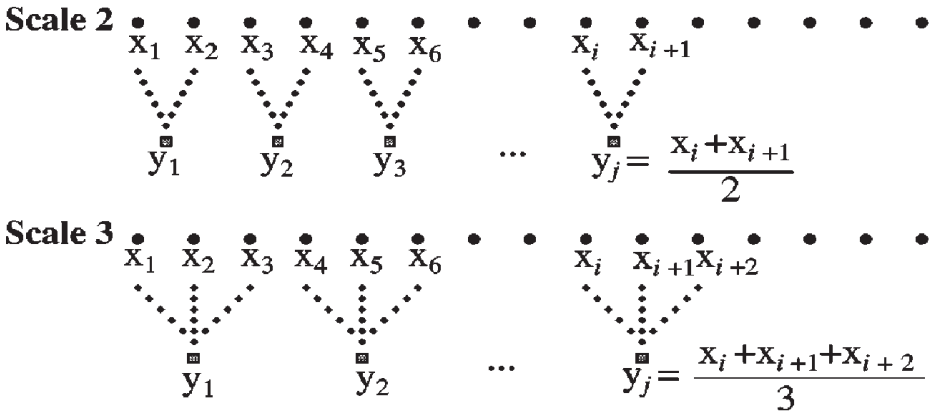


Figura 5.2: Ilustração esquemática do procedimento *Coarse Graining* para as escalas 2 e 3 (COSTA; GOLDBERGER; PENG, 2002a).

Com isso, calcula-se a entropia amostral (SampEn) (Equação 5.3), para cada série temporal $y^{(\tau)}(j)$ e em um gráfico plota-se o valor da Entropia (*Sample Entropy*) em função da escala τ.

Parte III

Procedimentos metodológicos

Capítulo Seis

Materiais e métodos

Este capítulo apresenta os dados, protocolo de coleta e tratamento e discute detalhadamente a construção de métodos novos, o uso conjunto com outros métodos já existentes e algumas aplicações possíveis.

O objetivo deste capítulo e atender os seguintes objetivos operacionais:

1. Construir as redes semânticas de títulos de acordo com as regras tratamento manual e computacional de Pereira et al. (2011);
2. Determinar o melhor tamanho de uma janela temporal deslizante;
3. Construir as séries temporais das redes de títulos variáveis no tempo;
4. Calcular a entropia da informação de Shannon em cada janela de tempo, levando em conta o processo de formação de uma rede de cliques;
5. Calcular os limites dos valores de entropia baseados nos vínculos que existem no processo de formação de uma rede de cliques;
6. Calcular as entropias para as redes crescentes dos periódicos;
7. Aplicar o método MSE nas séries dos índices de redes e nas séries de entropia;
8. Identificar as *redes críticas* nas janelas de tempo, a partir do índice *incidência-fidelidade*;
9. Calcular o grau, excentricidade e intermediação para identificar os vértices mais importantes nas redes críticas de momentos de alta e baixa entropia.

6.1 Coleta e organização dos dados

O conjunto de dados é composto pelos títulos dos artigos publicados nos periódicos *Nature* e *Science*, de 1999 a 2008, que são de circulação internacional e publicam semanalmente trabalhos que frequentemente estão entre os mais impactantes da ciência em todo o mundo, com altos valores de *fator de impacto*[1]. Além disso, a escolha destes periódicos se deu pela possibilidade de comparação entre eles, já que publicam semanalmente e já foram

[1] O impacto de suas publicações pode ser medido, dentre outras métricas, pelo *JIF* (sigla, do inglês Journal Impact Factor).

estudados por outros autores, e.g. Pereira et al. (2011). Tabela 6.1 mostra algumas informações sobre os dados coletados.

Dados	**Nature**	**Science**
Frequência de publicações	Semanal	Semanal
Número de artigos publicados no período	30490	11798
Número de semanas no período	512	514

Tabela 6.1: Dados sobre os periódicos *Nature* e Science no período de 1999 a 2008.

Os títulos coletados são organizados em um ou mais arquivos texto. Cada arquivo contém um ou mais números do periódico. Cada linha corresponde a um título, com a primeira letra (e apenas ela) em caixa alta. O tratamento é dividido em duas etapas,

1. Tratamento manual:

Este procedimento é feito seguindo as regras propostas por Pereira et al. (2011). Eis algumas delas:

1. Colocar todos os títulos no mesmo idioma;
2. Retirar dos títulos os sinais gráficos como ponto, vírgulas, dois pontos, ponto e vírgula, exclamação, etc.;
3. Juntar grupos de palavras em apenas uma, por representar um significado único, e.g. Rio de Janeiro → riodejaneiro, Albert Einstein → *alberteinstein*[2];
4. Escrever os sinais gráficos que representam números por extenso (por exemplo, 2017 → doiszeroumsete).

2. Tratamento Computacional:

O tratamento computacional consiste em utilizar ferramentas apropriadas para classificar as palavras gramaticalmente, modificá-las (se necessário) a fim de reduzir ambiguidades e eliminar as que não possuem significado semântico relevante. Conjecturamos que o sistema a ser analisado considera apenas palavras lexicais, de modo que as palavras gramaticais (por exemplo, preposição, artigo e pronome) não são mais consideradas como elementos do sistema porque não têm significados intrínsecos.

[2]É importante que a mesma palavra seja transformada de maneira igual nos títulos, por exemplo se em um título o pesquisador escolheu tratar Albert Einstein para alberteinstein, deve fazer isto para todos os títulos que esta palavra aparecer, nunca Albert Einstein → *AlbertEinstein*. Isso significa que um vocabulário de controle deve ser elaborado.

Para isto, foi usado o Pacote Unitex[3] e o Ambisin. O primeiro classifica palavras utilizando tabelas léxico-gramaticais, dicionários, gramáticas, entre outros recursos em várias línguas. O Ambisin, criado por Caldeira (2005), permite eliminar ambiguidades, palavras gramaticais, separação entre formas verbais flexionadas e canônicas, além de permitir palavras que não são encontradas no dicionário.

6.2 Construção da rede semânticas de títulos variáveis no tempo

Uma rede semântica de títulos variável no tempo ($RSTVT$) considera as informações temporais contidas em seus títulos, na construção da rede. Neste contexto, sua construção se baseia em um TVG (ver Equação 2.1 da Seção 2.4).

Sendo assim, uma $RSTVT$ é formalizada de acordo com a Equação 6.2 como sendo

$$\mathcal{G} = (V, \mathcal{E}, \Gamma, \Upsilon, \varsigma). \tag{6.1}$$

onde:

- O conjunto de vértices $V = \{v_1, v_2, ..., v_n\}$ é formado pelas palavras tratadas dos títulos no período coletado;
- o conjunto de arestas $\mathcal{E} = \{(e_1,..., e_m\}$ é formado pelos pares de palavras (i, j) que pertencem a um mesmo título;
- o tempo de vida Γ representa o conjunto dos instantes de tempo do período coletado. O valor do tempo total de existência do sistema é $T = t_\Psi - t_1 + 1$, onde t_Ψ é o último instante e t_1 o primeiro instante do TVG;
- a função presença Υ garante a existência de um par de palavras em um dado instante de tempo t. Por, exemplo, se o par de palavras (α, β) =(“gene”,“cell”) acontecem em pelo menos um título em $t = 46$, então $\Upsilon(46)(\alpha, \beta) = 1$.
- a função *latência* ς representaria o tempo de formação de um par de palavras. Sua medida não é relevante para rede de títulos nesta pesquisa, sendo aqui considerada constante e igual a zero;

Como comentado, as redes deste trabalho foram construídas a partir das palavras dos títulos dos periódicos *Nature* e *Science* no período de 1999 a 2008, tratadas manual e

[3]conjunto de programas disponibilizado pela Rede Relex Brasil, disponível em http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/bibliotecas.html

computacionalmente de acordo com Pereira et al. (2011). Para a análise temporal, a rede foi organizada em instântaneos ao longo dos 10 anos de publicações. Estes instantâneos são adquiridos a partir de uma janela de tamanho fixo que se desloca no tempo a um passo constante. A partir disto, os índices de interesse são capturados para análise de suas evoluções no tempo ou para analisar uma determinada época quanto à diversidade de vocabulário e suas conexões. A próxima seção define os parâmetros desta janela temporal deslizante.

6.3 Janela temporal deslizante

A função *janela temporal deslizante* (*time sliding window*) denotada por $w_{\tau,s}$ contribui para a construção e análise de uma rede variável no tempo, pois permite agrupar títulos em janelas de tempo de tamanho τ, separados por um intervalo de tempo de tamanho s entre janelas consecutivas.

Esta função tem por objetivo varrer o TVG permitindo a coleta de informações em diferentes épocas do TVG durante seu tempo de vida. Para $s < \tau$, janelas consecutivas possuirão títulos em comum; para $s \geq \tau$, não haverá sobreposição de títulos.

A Figura 6.1 exemplifica a aplicação desta ferramenta.

Para $s < \tau$, Figura 6.1a, haverá sobreposição de dados para janelas consecutivas. Nesta condição, espera-se que quanto maior for a diferença $\tau - s$, menor será a variação nos valores dos índices de redes para janelas consecutivas; para $s = \tau$, Figura 6.1b, não haverá sobreposição de dados; e para $s > \tau$ não haverá sobreposição e nem lacuna entre as janelas, Figura 6.1c.

Para valores de τ e s sejam constantes, o conjunto de janelas que se encaixa no TVG é uma função de τ, s e T, como mostrado na Equação 6.2. Nesta equação, n_w é o número total de janelas, ou seja, número de redes a serem analisadas.

$$\begin{aligned} w_{\tau,s}(T) =& \{\mathcal{G}^1, \mathcal{G}^2, ..., \mathcal{G}^{n_w-1}, \mathcal{G}^{n_w}\}; \\ \tau, s \in& \mathbb{N}; \\ n_w =& \lfloor \frac{T-(\tau-s)}{s} \rfloor. \end{aligned} \tag{6.2}$$

A Equação 6.2 só vale para $\tau \leq T$ e $s \leq T - \tau$. Para $s = 0$ o TVG possuirá apenas um instantâneo, ou seja, considera apenas os vértices e arestas presentes em $\{t_1, t_2, ..., t_{\tau-1}, t_\tau\}$.

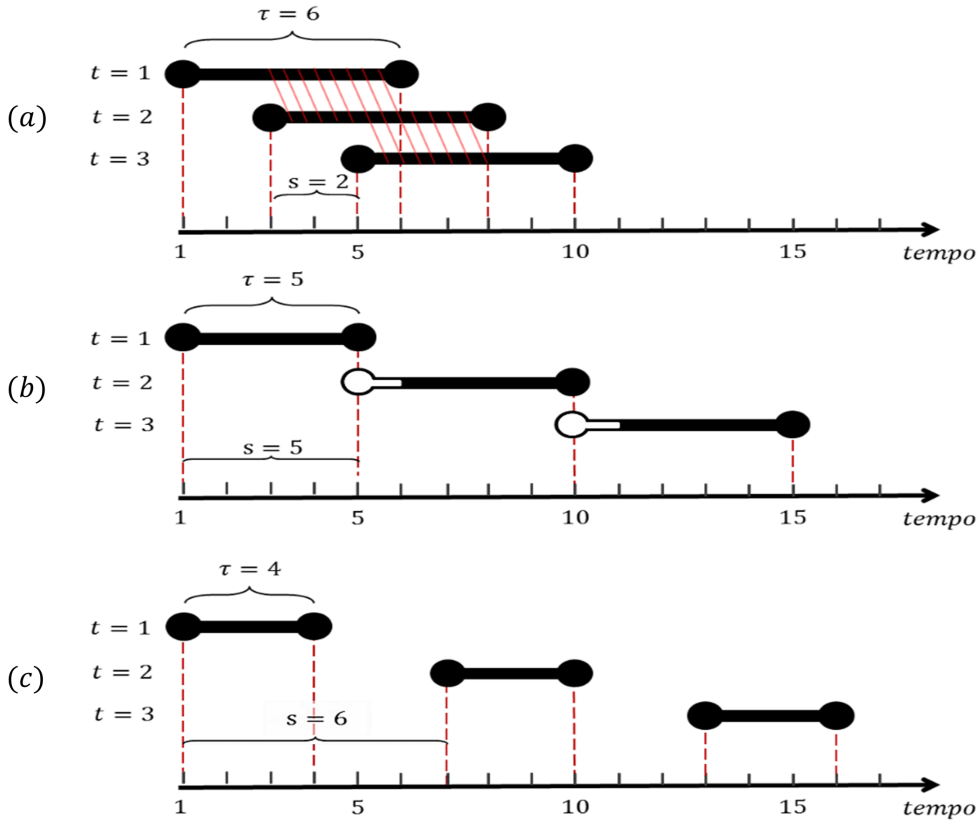


Figura 6.1: Exemplos de janelas deslizantes sobre um TVG. Em (a), $w_{6,2}$, $s < \tau$; em (b) $w_{5,5}$, $s = \tau$; (c) $w_{4,6}$, $s > \tau$. A área rachurada corresponde a sobreposição de dados. Fonte: (CUNHA et al., 2020b).

Para exemplificar, considere o TVG da Figura 6.2, referente à rede descrita na Figura 2.1 (Seção 2.1). A janela escolhida possui tamanho 28 *semanas* (1 semestre) e caminha no tempo anualmente, $s = 52$ *semanas*.

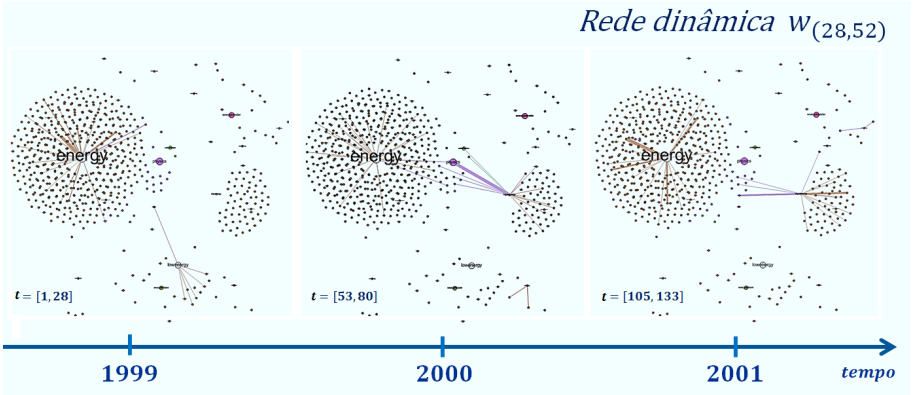


Figura 6.2: Evolução da rede de conceitos associados à temática "Energia" publicadas na *Nature* (em seus títulos) ao logo do tempo com uma janela temporal deslizante $w_{28,52}$. Cada aresta conecta duas palavras que pertencem ao mesmo título e que contenham o termo "energy" no nome.

6.3.1 *Escolha dos parâmetros de w*

A partir da Equação 6.2, vemos que quanto menor os valores de τ e $s > 0$, maior o número de redes (janelas de tempo) a serem analisadas para um dado periodo T. Para o menor valor do passo, $s = 1$. Entretanto, para o menor tamanho da janela, é preciso considerar:

- A rede em uma janela deve ter os números de vértices e arestas capazes de exibir as propriedades desejadas pelo pesquisador. Por exemplo, não faz sentido calcular alguns dos índices de redes (Seção 2.2) para redes com poucos vértices;
- Caso se deseje investigar a evolução da rede pelos seus índices através de uma série temporal é interessante que esta série tenha flutuações suavizadas, de pouco ruído, para identificar tendências. Pequenos valores de s, $s < \tau$ (sobreposição de dados), contribuem para isto. Mas não necessariamente para pequenos valores de τ ($\tau = 1$, por exemplo).

Diante disso, para as análises que envolvem índices de redes, usaremos $s = 1$ e $\tau = 8$. Na Figura 6.3, o índice *Densidade* (Δ) está em função do tempo para janelas de tamanho $\tau = 1$ (1 semana de publicações), $\tau = 4$ (1 mês de publicações) e $\tau = 8$ (1 bimestre de publicações) para *RST* da *Nature*. Observe que para $\tau = 8$ as variações no valor do índice são menos bruscas. A escolha de 8 *semanas* para o tamanho da janela se deu por que ela permite gerarmos o máximo número de redes com tamanhos consideráveis para que seus índices as representem, com a evolução seus valores sem variações bruscas por conta do vocabulário novo inserido a cada janela (observamos na Figura 6.3, que para $\tau = 8$ a série evolui suavemente).

Já para as análises que envolvem entropia, usaremos $w_{8,1}$ e $w_{1,1}$. Esta última visa evitar sobreposição de dados com o avanço da janela no tempo. A escolha do passo $s = 1$ em todas as situações é para que a observação seja feita no máximo de instantâneos do tempo de vida do TVG.

A *janela deslizante* $w_{8,1}$ ($\tau = 8$ e $s = 1$), ou seja, uma janela de oito semanas que avança no tempo semana a semana é explicada na Figura 6.4. Um exemplo real é dado na Figura 6.5. E deste modo, é possível analisar o que ocorre com os valores de índices de redes para cada grafo estático correspondente a cada janela.

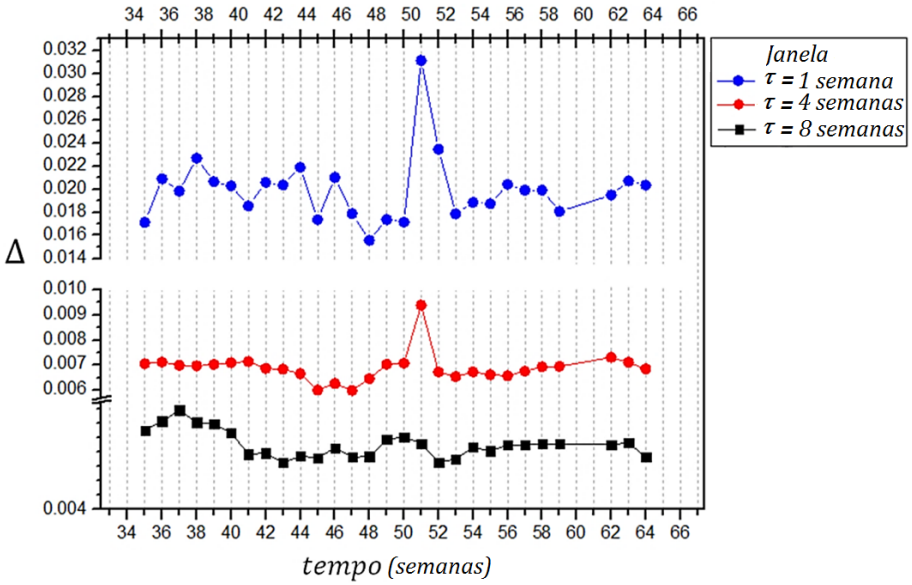


Figura 6.3: Procedimento escolhido para determinar melhor Janela, utilizando a série da *Densidade* (Δ), para janelas de tempo de tamanho 1 semana ($\tau = 1$), 1 mês ($\tau = 4$) e 2 meses ($\tau = 8$).

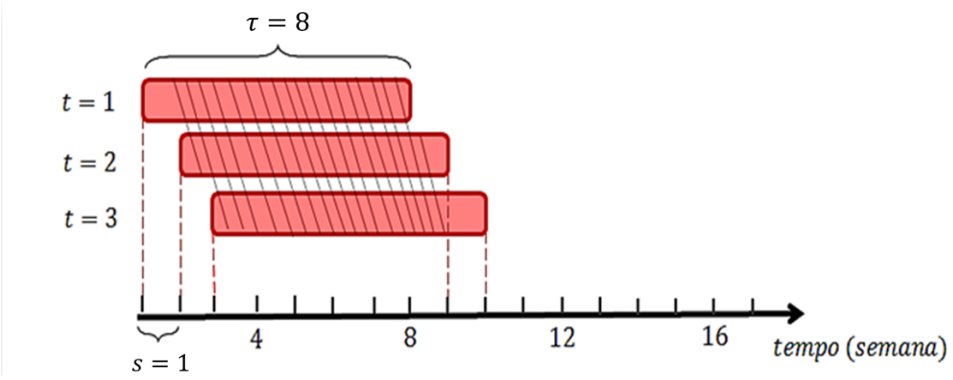


Figura 6.4: Janela de 8 semanas ao longo do tempo que avança semana a semana ($w_{8,1}$).

6.4 Indicadores utilizados

Para analisar a diversidade do vocabulário e suas conexões na rede de cada janela, foram utilizados os índices da Tabela 6.2. Vale lembrar que, para uma rede de cliques, a "configuração incial" são as cliques isoladas, e a "configuração final" são as cliques unidas, formando a rede de cliques, conforme Figura 3.1, da Seção 3.1.

6.5 Cálculo das entropias

Vamos considerar duas variáveis aleatórias no processo de formação de redes de cliques: o vértice e a aresta. As probabilidades de ocorrência de um vértice i e uma aresta $e_k = (i, j)$

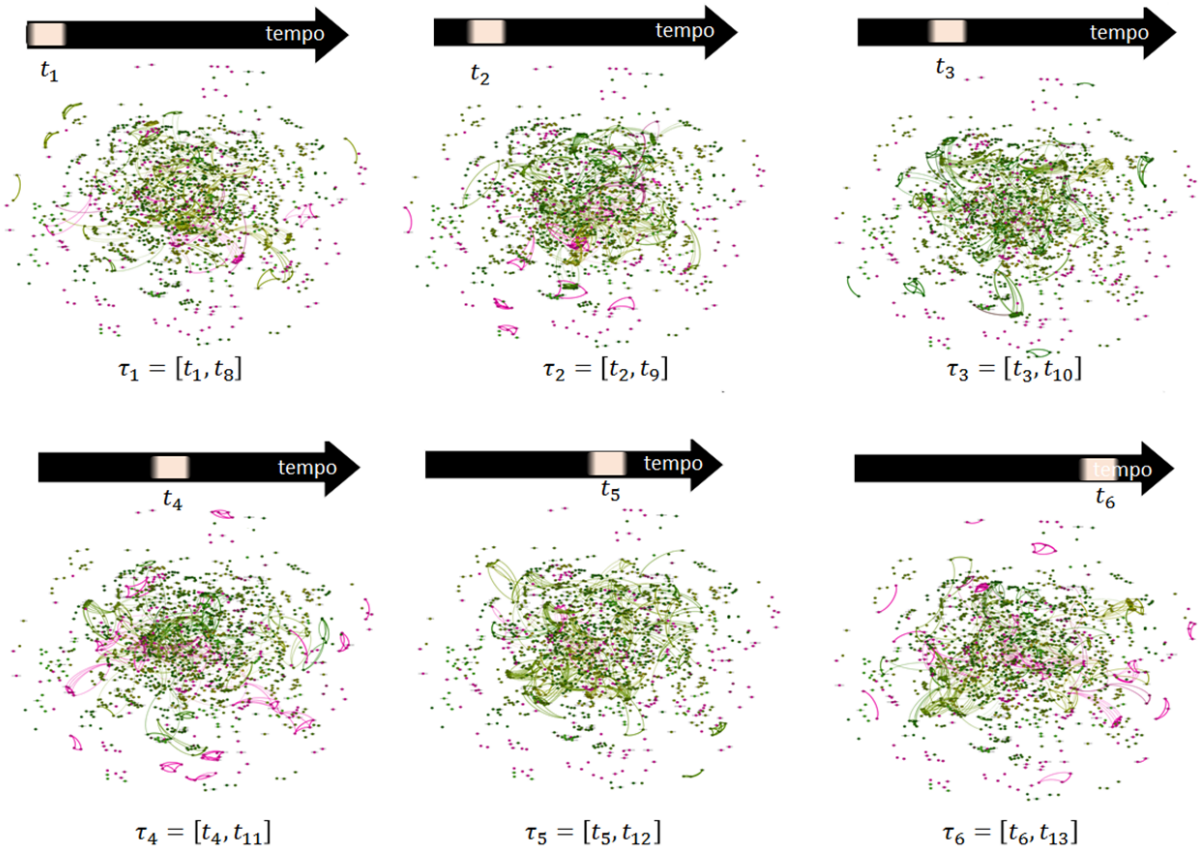


Figura 6.5: Evolução da rede em uma janela $w_{8,1}$ para a *RSTVT* da *Nature* durante 13 *semanas*, ano 1999.

são calculadas, para cada janela de tempo (t) considerada, de acordo com as Equações 6.3 e 6.4. Ou seja, na Equação 6.3 temos a probabilidade de ocorrência de uma palavra i dado um conjunto de palavras n_O e na Equação 6.4 temos a probabilidade de um par de palavras ($e_k = (i, j)$) co-ocorrerem em uma sentença dentro de todos os pares de palavras m_0 do vocabulário da janela no intante de tempo t.

$$p_i(t) = \frac{\#(v_i)(t)}{n_0(t)}, \quad \sum_{i=1}^{n} p_i(t) = 1. \tag{6.3}$$

$$p_{e_k}(t) = \frac{\#(i,j)(t)}{m_0(t)}, \quad \sum_{k=1}^{m} p_{e_k}(t) = 1. \tag{6.4}$$

Índice	Descrição
n_q	Número de cliques na configuração inicial.
n	Número de vértices da rede na configuração final.
m	Número de arestas da rede na configuração final.
n_0 m_0	Número de vértices na configuração inicial., $n_0 \geq n$. Número de arestas na configuração inicial.
$\#(v_i)$	Frequência do vértice i na configuração inicial, ou seja, número de cliques que contém o vértice i na configuração inicial $(1 \leq \#(v_i) \leq n_q)$.
$\#(i,j)$	Frequência da aresta (i,j) na configuração inicial, ou seja, número de cliques que contém as palavras i e j, $1 \leq \#(i,j) \leq n_q$ e $i,j \in \{1,2,...,n-1,n\}$, com $i \neq j$ e $(i,j) = (j,i)$.
q_i	Tamanho do título i. Número de vértices do título i na configuração inicial, $(1 \leq i \leq n_q)$.
q_{min}	Número de vértices da menor clique na configuração inicial, $(1 \leq q_{min} \leq n)$.
q_{max}	Número de vértices da maior clique na configuração inicial, $(1 \leq q_{max} \leq n)$.
$\langle k \rangle$	$\langle k \rangle = \frac{\sum_1^n k_i}{n} = \frac{2m}{n}$, onde $\langle k \rangle$ é o grau médio de uma rede não dirigida e k_i é o grau de um vértice i, que é o número de arestas incidentes sobre o vértice i.
k_i^{hub}	$k_i^{hub} \geq \langle k \rangle + 2\sigma$, são os valores dos graus dos hubs, isto é, vértices de graus muito altos. σ é o desvio padrão da distribuição de graus.
$e(i)$	Excentricidade do vértice i, i.e. $e(i) = max_{j \in V} d(i,j)$, em que $d(i,j)$ é a distância geodésica entre os vértices i e j.

Tabela 6.2: Principais índices de redes de clique usados neste trabalho. "Configuração incial" está relacionada às cliques isoladas, e "configuração final" está relacionada à rede de cliques criada. Os índices são válidos para cada janela de tempo considerada.

As Equações Eq. 6.5 e Eq. 6.6 expressam as entropias de Shannon para essas distribuições, onde $H_v(t)$ e $H_e(t)$ são as entropias de vértices e arestas, respectivamente, em cada instante de tempo t,

$$H_v(t) = -\sum_{i=1}^{n} p_i(t) \cdot \log_2 p_i(t) \quad bits. \tag{6.5}$$

$$H_e(t) = -\sum_{k=1}^{m} p_{e_k}(t) \cdot \log_2 p_{e_k}(t) \quad bits. \tag{6.6}$$

Para melhorar o entendimento sobre o cálculo das entropias, a Figura 6.6 mostra um

exemplo de rede de cliques e seu processo de formação, as probabilidades associadas e os valores de entropia dos vértices e arestas. Na Figura 6.6a temos as cliques na configuração inicial e na Figura 6.6b a rede de cliques construída por processos de justaposição de vértices e sobreposição de arestas (FADIGAS; PEREIRA, 2013). Em 6.6c as probabilidades calculadas de acordo com a configuração inicial, e em 6.6d as entropias da rede para as distribuições de vértices e arestas.

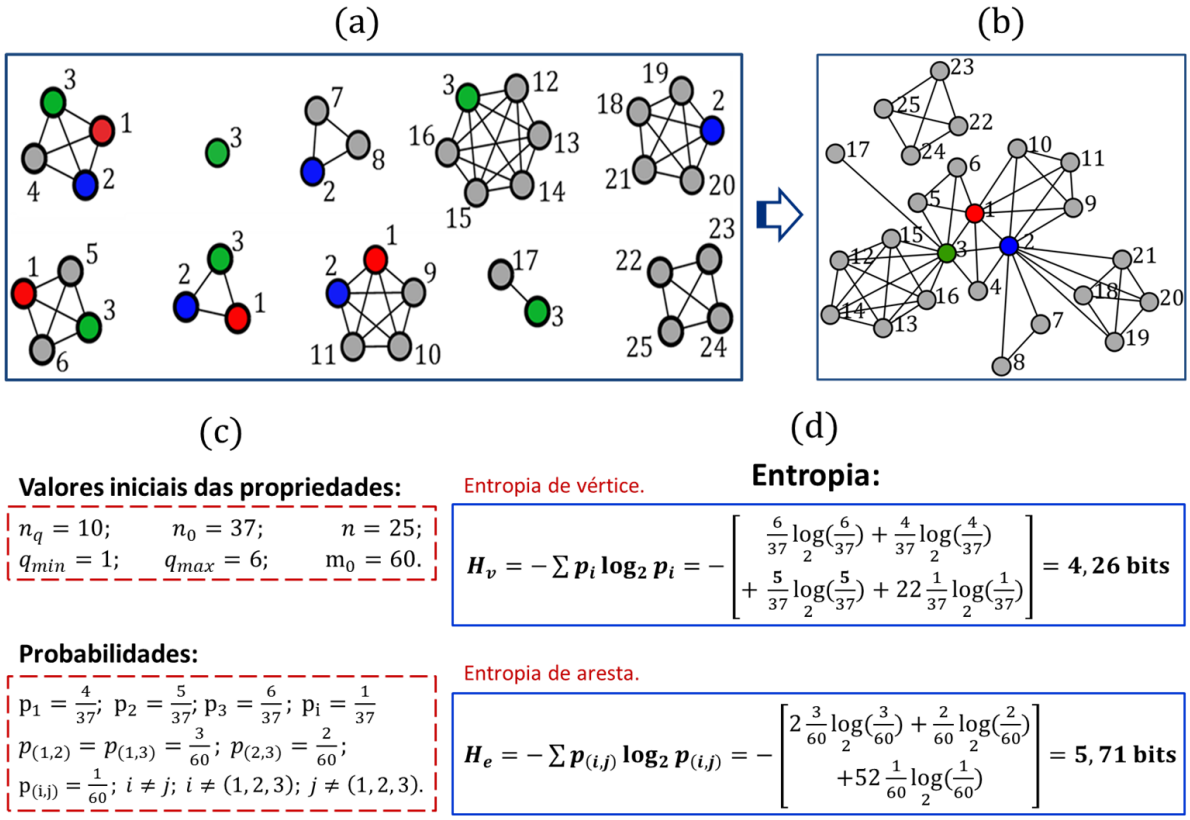


Figura 6.6: Exemplo de uma rede de cliques em seu processo de formação, as probabilidades associadas e os valores de entropia de seus vértices e arestas. (a) as cliques na configuração inicial. (b) a rede de cliques (configuração final) construída por processos de justaposição e sobreposição. (c) valores iniciais de propriedades e probabilidades de vértices e arestas. (d) os valores de entropia de vértices e arestas para a configuração inicial da rede. Fonte: (CUNHA et al., 2020b)

A dinâmica de construção de uma redes de cliques impõe vínculos restritivos para a configuração da rede, de forma a mantê-la como uma rede de cliques. Serão abordados a seguir, os fatores que levam a esses vínculos e como eles afetam a entropia da rede.

6.5.1 Limites para os valores de entropia

É importante destacar os fatores que contribuem para o aumento e redução da entropia de Shannon de um sistema. O valor mínimo de entropia está associado à certeza máxima

de uma variável. Dois fatores contribuem fortemente para isso: (i) mínimo de estados possíveis para a variável e (ii) maior repetição de um ou alguns estados possíveis para a variável. Ou seja, maior probabilidade de x estados, quando $x \ll n$ e n é o total de estados para a variável. Por outro lado, a entropia máxima está associada à certeza mínima da variável, ou seja, ao máximo de estados possíveis para a variável, em que cada estado tem a menor probabilidade possível.

Os limites mostrados na Equação 4.7 não ocorrem para a entropia de uma rede de cliques como foi calculado aqui. Nesta seção, os extremos são calculados baseados nos vínculos a seguir (condições[4] de contorno para a formação de uma rede de cliques):

- número de cliques na configuração inicial n_q;
- número de vértices na configuração inicial n_0;
- número de vértices na configuração final n;
- tamanho da maior clique q_{max};
- menor tamanho de clique $q_{min} \neq 0$.

Para o cálculo dos limites, assumiremos a existência de configurações que maximizam e minimizam a entropia. Para isso, seguiremos 8 passos. Os passos **1** a **2** definem a configuração de uma rede de cliques que possui o máximo de entropia:

Passo 1. Vamos imaginar n_q espaços vazios (chamaremos de "cliques vazias"[5]), e n_0 vértices disponíveis para distribuir nesses espaços, onde $n_0 \geq n$. De n_0 vértices, n é o número de vértices que são necessariamente diferentes.

Passo 2. Os n vértices são distribuídos nas n_q cliques, sem repetição de vértices em cada clique, e sem que o número de vértices por clique q_i exceda o tamanho máximo q_{max} e não seja inferior ao tamanho mínimo de clique q_{min}, ou seja, $q_{min} \leq q_i \leq q_{max}$.

Este momento é chamado de *Configuração 1.* A distribuição é realizada de maneira a não haver repetição de vértices e arestas, usando Equação[6] 6.7. Nesta configuração, teremos todos os vértices e arestas diferentes com o número mínimo de arestas. Na rede final, existem x cliques de tamanho q e y de tamanho $(q+1)$; portanto,

[4] A depender do sitema utilizado, pode não ser necessário utilizar todas estas condições ou ser necessário acrescentar outras.

[5] Não existe uma "clique vazia". Mas, este termo refere-se a um espaço que será, no passo posterior, preenchido por uma clique.

[6] A notação $\lfloor X \rfloor$ representa o menor valor inteiro de X. Sendo assim, y representa o resto da divisão de n por n_q.

$$\begin{aligned} q &= \lfloor \frac{n}{n_q} \rfloor; \\ y &= n - qn_q; \\ x + y &= n_q; \\ xq + y(q+1) &= n. \end{aligned} \tag{6.7}$$

A *Configuração 1* gera a máxima entropia de vértices, $H_{v\ max} = \log_2 n$, porque garante a disposição de todos os vértices sem repetição; e a menor entropia para as arestas da rede, uma vez que garante o menor número de arestas.

Sabemos que a repetição de um estado para a variável (aumento de sua probabilidade) contribui para redução da entropia. Nas redes de clique, esse fenômeno não necessariamente ocorre para arestas porque a repetição de uma aresta implica que ela exista em mais de uma clique. E devido a isso, os dois vértices que a compõem são forçados a serem conectados a todos os outros vértices da clique, o que causa um aumento considerável de arestas, ou seja, mais estados novos são criados e consequentemente um aumento da entropia.

Por exemplo, considere um conjunto de cliques em que não haja repetição de vértices, onde existem duas cliques $Q_1 = \{1, 2, 3, 4, 5\}$ e $Q_2 = \{6, 7, 8, 9, 10\}$. Se acrescentarmos a aresta $e_2 = (1, 3)$ em Q_2 sua probabilidade irá aumentar, entretanto novas arestas serão criadas pela conexão dos vértices $v_1 = 1$ e $v_2 = 3$ com todos os vértices de Q_2 ($v_6 = 6$, $v_7 = 7$, $v_8 = 8$, $v_9 = 9$ e $v_{10} = 10$), o que faz a entropia do sistema aumentar. Portanto, a melhor maneira de garantir a menor entropia de arestas é com o menor número delas, evitando repetição.

Os passos **3** a **6** permitem encontrar a configuração que gera a menor entropia de vértices:

Passo 3. A partir da *Configuração 1*, adiciona-se os restantes $n_0 - n$ vértices repetidos, um por um, com a repetição máxima de vértices para os primeiros vértices adicionados.

Passo 4. Se $n_0 - n \geq n_q - 1$, haverá um vértice repetido em todas as cliques. Após a distribuição, se $(n_0 - n) - (n_q - 1) \geq n_q - 1$, o processo continua, com a escolha de outro vértice para repetir nas cliques.

Passo 5. O processo é repetido até que o número de vértices restantes seja menor que $n_q - 1$, e assim, esses vértices serão distribuídos como um único vértice repetido no número de cliques que podem caber;

Passo 6. O valor $n_q - 1$ é subtraído dos vértices que ainda não foram adicionados até que essa subtração resulte em um número $n' \leq n_q - 1$; portanto, o último vértice é adicionado

repetidamente clique a clique em n' cliques.

A *Configuração 2* aumenta a probabilidade de alguns vértices ao reduzir a entropia ao menor valor possível, respeitando os vínculos do sistema. A Figura 6.7 mostra uma *Configuração 1* e uma *Configuração 2* possíveis para a janela $t = 202$ do *TVG* da *Science*.

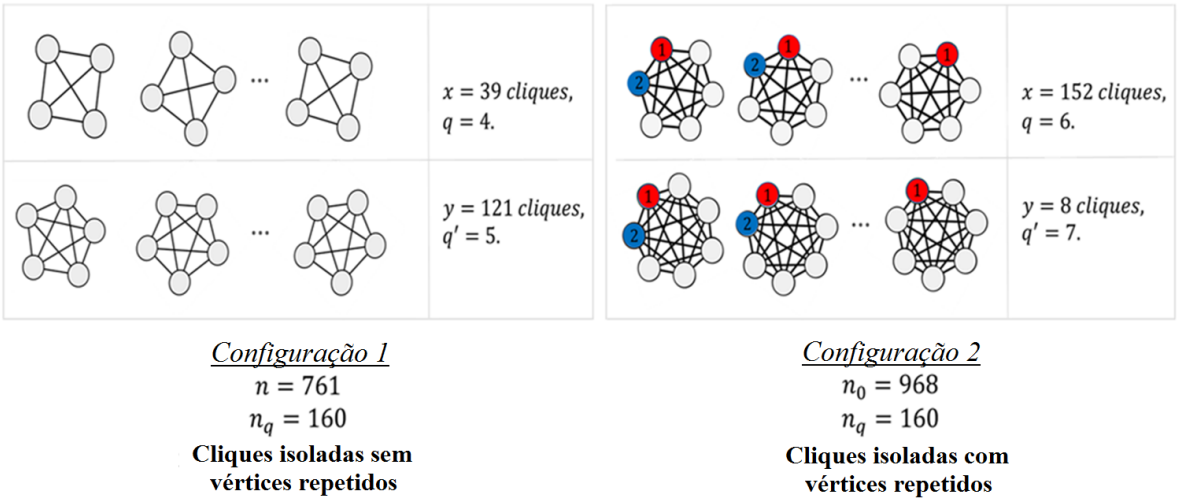


Figura 6.7: Exemplo de um esquema possível baseado na *Configuração 1* e na *Configuração 2* usando dados reais da janela $t = 202$ do *TVG* da rede de títulos da *Science* mostrada neste artigo. Nesta janela, existem $n_q = 160$ títulos com $n = 761$ palavras diferentes de um total de $n_0 = 968$ palavras. A *Configuração 1* minimiza o número de arestas na rede, consequentemente sua entropia, $H_{e\ min} = 10,49\ bits$ e maximiza a entropia de vértices $H_{v\ max} = 9,57\ bits$. A *Configuração 2* minimiza a entropia dos vértices. Portanto, para $t = 202$, *TVG* da *Science*, $H_{v\ min} = 8,34\ bits$. Fonte: (CUNHA et al., 2020a).

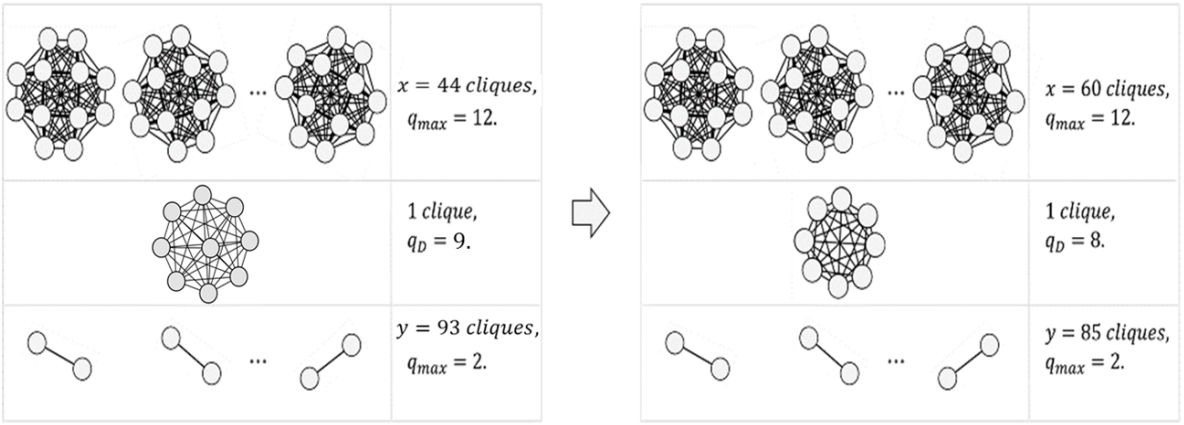
Os passos **7** a **8** permitem encontrar a configuração que gera a maior entropia de arestas. Para isso, o número de arestas diferentes deve ser aumentado o máximo possível, evitando a repetição.

Passo 7 A partir do passo 1, a distribuição de n vértices é feita de forma a obter x cliques de tamanho q_{max} e y cliques de tamanho q_{min}, com a possibilidade de haver uma clique com tamanho q_D, ou seja, $q_{min} < q_D < q_{max}$, chamada *Configuração inicial 3*.

Passo 8 Depois disso, os vértices repetidos $n_0 - n$ são adicionados, um a um nas cliques, evitando a repetição de arestas, de forma a obter o máximo de cliques máximas (*Configuração final 3*).

O aumento no número de cliques máximas eleva o número de arestas distintas e, consequentemente, a sua entropia. A Figura 6.8 mostra o esquema da *Configuração 3* para $t = 188$ do *TVG* da *Science*.

Usando a Figura 6.6 como ponto de partida, a Figura 6.9 resume o processo para calcular os limites máximos e mínimos da entropia de vértices e arestas.



Configuração inicial 3
$n = 723$
$n_q = 146$
Cliques isoladas sem vértices repetidos

Cofiguração final 3
$n_0 = 898$
$n_q = 146$
Cliques isoladas com vértices repetidos

Figura 6.8: Exemplo de um esquema possível baseado na *Configuração 3* usando dados reais da janela $t = 202$ do TVG da rede de títulos da *Science*. Nesta janela, existem $n_q = 146$ títulos com $n = 723$ palavras diferentes de um total de $n_0 = 898$ palavras. A *Configuração 3* maximiza o número de arestas na rede, consequentemente sua entropia, $H_{e\ max} = 11,50\ bits$. Fonte: (CUNHA et al., 2020a).

Com este método, o cálculo das entropias de vértice e aresta de uma rede de cliques e seus limites precisam levar em conta seus vínculos no processo de formação da rede, i.e. distribuição dos vértices e arestas na sua configuração inicial.

Os valores extremos de entropia podem ser reescalonados pelos seus extremos ($H^{'} = \frac{H - H_{min}}{H_{max} - H_{min}}$) para que $0 \leq H^{'} \leq 1$. Por exemplo, na rede da Figura 6.6b temos a 6.6a como sua configuração inicial. Os valores das entropias reais (H), bem como seus valores máximos e mínimos (H_{min} and H_{max}) e os valores de $H^{'}$ para vértices e arestas, calculados nas Figuras 6.6 e 6.9, estão na Tabela 6.3.

Distribuição/entropia	$H^{'}$	H_{min}	H	H_{max}
edges	$H^{'}_e = 0,71$	$H_{e\ min} = 4,32\ bits$	$H_e = 5,71\ bits$	$H_{e\ max} = 6,28\ bits$
vertices	$H^{'}_v = 0,31$	$H_{v\ min} = 4,09\ bits$	$H_v = 4,26\ bits$	$H_{v\ max} = 4,64\ bits$

Tabela 6.3: Entropias para o exemplo contido nas Figuras 6.6 e 6.9. Entropia real das distribuições de vértices e arestas (H), seus valores extremos mínimos e máximos (H_{min} and H_{max}) e a entropia normalizada $H^{'}$ a partir dos extremos.

6.5.2 *Entropia para rede crescente*

Vamos considerar para o mesmo TVG, a janela $w_{t,0}$, ou seja, a janela de tempo de uma semana cresce linearmente com o tempo $\tau = t$, com $t \in N\ 1 \leq t < T$ e não desliza com o

tempo ($s = 0$). A evolução da entropia para este TVG, plotada em uma série temporal, pode fornecer pistas sobre como evolui a diversidade do vocabulário e suas conexões.

6.5.3 Caso $n < n_q$

Para os *TVG*s deste trabalho ($w_{8,1}$), em todas as janelas, o número de vértices (n) é maior que o número de títulos (n_q). Para janelas de tempo maiores, pode acontecer a situação $n < n_q$. Nesse caso, alguns ajustes seriam necessários para o cálculo dos limites. Por exemplo, na *Configuração 1*, se $q = 0$, então $(q + 1) = 1$, $y = n$ e $x = n_q - n$. Isso contradiz uma condição de cotorno, uma vez que $q = 0 < q_{min}$. Assim, uma parcela de $n - n_0$ precisará ser distribuída nas cliques, de forma que cada uma delas possua $q = qmin$.

Sendo assim, para encontrar a entropia máxima de vértices, no caso de $n \leq n_q$, seguiremos os seguintes passos:

1. Considere n_q "cliques vazias" e n_0 vértices disponíveis para distribuir nelas, onde $n \leq n_0$ são necessariamente vértices diferentes;
2. os n vértices serão distribuídos em $n' < n_q$ cliques, sem repetição de vértices em cada clique. Assim, teremos n' cliques com 1 vértice e $n_q - n'$ "cliques vazias";
3. Inserir em uma das clique "vazia" um vértice igual a um dos já adicionados na etapa anterior. Repetir este procedimento para outro vértice, até acabar as "cliques vazias";
4. Este procedimento deixa cada clique com apenas um vértice, com o mínimo de repetição de vértices. Caso o valor de $q_{min} > 1$, a etapa anterior se repete até que todas as cliques tenham tamanho q_{min} (neste trabalho, as redes que $n < n_q$, $q_{min} = 1$).

As Configurações 2 e 3 seguem como descrito na Seção 6.5.1, com a ressalva de substituir a *Configuração 1* pelo resultados dos passos descritos nesta seção.

Desta forma, temos o total de palavras igual a n_q, com $n - (n_q - n)$ vértices com probabilidade igual a $\frac{1}{n_q}$ e $n_q - n$ vértices com probabilidade $\frac{2}{n_q}$. A Equação 6.8 calcula este valor. Note que esta entropia máxima é a maior possível, mas é menor que $\log_2 n$.

$$H_{vmax} = -[(n_q - n) \times \frac{2}{n_q} \log_2 \frac{2}{n_q} + (2n - n_q) \times \frac{1}{n_q} \log_2 \frac{1}{n_q}] \tag{6.8}$$

A Equação 6.8 vale somente para $n_q \geq n$. Note que $H_{max} = \log_2 n$ é o caso particular para $n = n_q$.

Para ilustrar, considere a janela $w_{514,0}$ para a *Nature*. Esta janela compreende todo o vocabulário do período de 1999 a 2008, que corresponde ao tempo de vida do TVG. Nesta janela: $n = 21577$, $n_q = 29580$, com $n_q - n = 8003$. Teremos então um vértice diferente em cada clique, sobrando $n_q - n = 8003$ cliques "vazias", que serão preenchidas por vértices repetidos, de acordo com o passo 4, mas com mínimo de repetição por vértice, ou seja, 8003 vértices diferentes entre si, mas iguais aos que já estavam nas 29580 cliques. Aplicando a Equação 6.8, temos $H_{vmax} = 14,31$ (Menor que $\log_2 n = 14,40$).

6.6 Método MSE

O método de entropia multi escala (MSE) será aplicado nas séries temporais conforme descrito na Seção 5.4. As séries utilizadas são (i) as dos índices de rede para os TVGs da Science e da *Nature* para $w_{8,1}$ e (ii) as dos valores de entropia de vértice, para $w_{8,1}$ e $w_{1,1}$. Os parâmetros para o MSE utilizados foram $r = 15\%$ do desvio padrão e $m = 2$.

O objetivo da aplicação deste método é investigar se existe algum padrão de previsvibilidade destas medidas ao mudar a escala de observação. Além disso de poder comparar resultados do MSE nos índices de rede com o método DFA em rede semática de títulos variáveis no tempo (CUNHA et al., 2013), e entre as séries de entropia compará-las quanto à previsibilidade de seus valores ao mudar o tamanho da escala de observação.

As curvas MSE são usadas para comparar a complexidade relativa de séries temporais. Neste sentido: (i) Se para a maioria das escalas a entropia é maior para uma série temporal, implica que a sua complexidade é superior; (ii) Se uma curva MSE for monótona e decrescente, a série temporal apenas contém informação relevante nas primeiras escalas temporais.

Os padrões teóricos para ruído branco e ruído $1/f$ são descritos em Costa, Goldberger e Peng (2005). Eles também são úteis para usar como referência para os resultados daqui. A Figura 6.6 mostra a análise MSE para estes dois ruídos.

A complexidade está associada à capacidade de um sistema se ajustar a um ambiente de constante mudança, o que requer funcionalidade multiescalar integrativa (e.g. sistemas vivos). Em contraste, em condições de funcionamento livre, uma diminuição sustentada na complexidade reflete uma capacidade reduzida do sistema de funcionar em certos regimes dinâmicos, possivelmente devido ao desacoplamento ou degradação dos mecanismos de controle (COSTA; GOLDBERGER; PENG, 2005).

Neste sentido, a série temporal de ruído $1/f$ é mais complexa do que a série temporal de ruído branco, devido a presença de correlações de longo alcance em suas séries, que não ocorre em séries temporais de ruído branco. Não obstante sua entropia poder ser maior na primeira escala, o padrão complexo fica evidente pela constância ou pequena oscilação do valor de entropia com o aumento da escala.

6.7 O uso da rede crítica

A Figura 6.7 mostra a componente gigante[7] da rede do TVG da *Nature* em $t = 224$. Esta configuração se deu com $IF_L \geq 1,1 \cdot 10^{-3}$, valor próximo do encontrado em TEIXEIRA et al. (2010) para discursos orais.

Esta rede foi obtida conforme explicado na Seção 3.4. Esta configuração de rede pode trazer pistas sobre o vocabulário que se destaca na referida janela, já que pertencem ao grupo de palavras responsáveis pela sua robustez.

Para encontrá-la, utilizamos o índice IF como um filtro a partir de um valor arbitrário IF_L. O valor de IF_L é aumentado a partir do valor mínimo de IF, eliminando da rede arestas com valores de $IF \leq IF_L$ e vértices isolados. A Figura 6.7 mostra a configuração da rede da *Science* em $(t = 384)$ para diferentes valores de IF_L.

Este fenômeno ocorreu com todas as redes da base. Sendo assim, podemos imaginar uma rede semântica baseada em cliques composta por três grupos de conexões de acordo com os valores de IF de seus pares de palavras: (i) as de "fraca conexão" (IF baixo); (ii) as de "forte conexão" (IF alto) e (iii) as de "conexão moderada"(IF intermediário). A Figura 6.7(b) mostra os picos para estes três grupos de IF. Através do gráfico em 6.7(a), pode-se identificar o ponto correspondente à rede crítica.

$IF = IF_C$ não necessariamente corresponde a maioria dos pares de palavras,i.e. o aumento de IF_L a partir deste ponto não necessariamente provoca a maior queda de arestas na rede (neste exemplo, isto acontece logo após $IF_L = 6,54 \times 10^{-3}$). Entretanto, $IF = IF_C$ é o ponto onde possui as arestas responsáveis pela conexão entre comunidades de vértices que possuem os pares de mais altos valores de IF.

[7] A componente gigante de uma rede é a maior componente conectada, desde que tenha pelo menos $50\% + 1$ dos vértices da rede.

6.7.1 *Vértices que se destacam na rede crítica*

Existem diversas maneiras de identificar vértices importantes em uma rede. Duas são apresentadas nesta seção:

- vértices *Hubs*: vértices com alto valor de grau;
- vértices de centro e de periferia: vértices que possuem o menor e o maior valor de excentricidade, respectivamente.

Os vértices que possuem os maiores graus na rede crítica não necessariamente são os mesmos na rede geral (de IF_L mínimo). Para identificar os vértices *hubs*, utilizaremos a Equação 6.9 (SILVA et al., 2012; SANTOS; PEREIRA; CUNHA, 2018).

$$k_i^{hub} \geq \langle k \rangle + 2\sigma; \tag{6.9}$$

Na Equação 6.9, k_i^{hub} é o valor do grau de um vértice i considerado *hub*, isto é, vértice de grau muito alto, em que σ corresponde ao desvio padrão da distribuição de graus. Caso seja de interesse do pesquisador que menos palavras sejam consideradas *hubs*, pode-se aumentar mais um desvio padrão na Equação 6.9 para termos os *hubs* a partir da Equação 6.10.

$$k_i^{hub} \geq \langle k \rangle + 3\sigma; \tag{6.10}$$

O grau de uma palavra i em uma rede de títulos representa a influência do (i) número de palavras que juntos com i pertencem a um mesmo título, i.e. quanto mais palavras i são necessárias para compor um título que i faz parte; e do (ii) número de títulos que a palavra i aparece. Palavras que contem alto grau, em uma rede que foi filtrada as conexões mais fracas, merecem destaque.

A excêntricidade de um vértice representa o quanto ele está distante dos demais, ou seja, $e(i) = max_{j \in V} d(i,j)$, em que $d(i,j)$ é a distância geodésica entre os vértices i e j. Utilizando este conceito, deriva-se o *diâmetro* $D(G) = max_{i,j \in V} d(i,j)$ e raio $r(G) = min_{v \in V} e(i)$ (TAKES; KOSTERS, 2013).

Os vértices que possuem as maiores excentricidades compõe a periferia da rede e os vértices que possuem a menor excentricidade compõe o centro da rede. Estes vértices possuem grande importância na rede de títulos. As palavras que compõe o centro da rede, por

exemplo, são as mais próximas de todas as outras da rede. E como a rede crítica exclui arestas "fracas" da rede, palavras de centro possuem papel fundamental na composição do vocabulário de uma RST em uma dado instante.

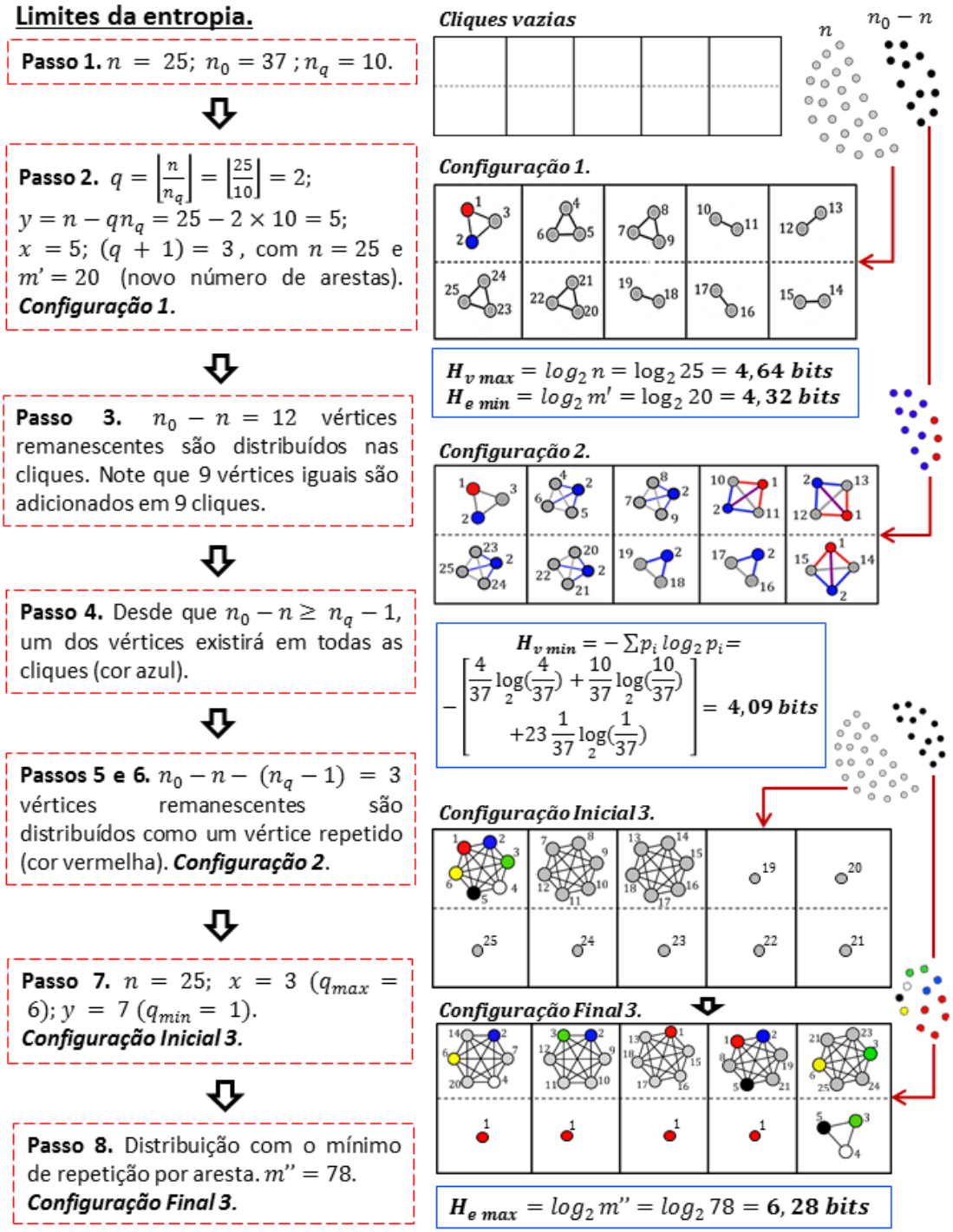


Figura 6.9: Passo a passo do cálculo dos limites máximos e mínimos para os valores de entropia dos vértices e arestas da rede de cliques na Figura 6.6. Fonte: (CUNHA et al., 2020b).

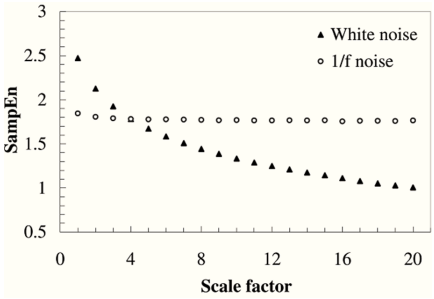


Figura 6.10: MSE para séries de ruído branco e ruído $1/f$. Os símbolos representam os valores médios de entropia para 30 séries temporais e barras de erro do desvio S_D, com uma média de $0,05$ para ruído branco e $0,02$ para ruído $1/f$. Fonte: Adaptado de Costa, Goldberger e Peng (2002b).

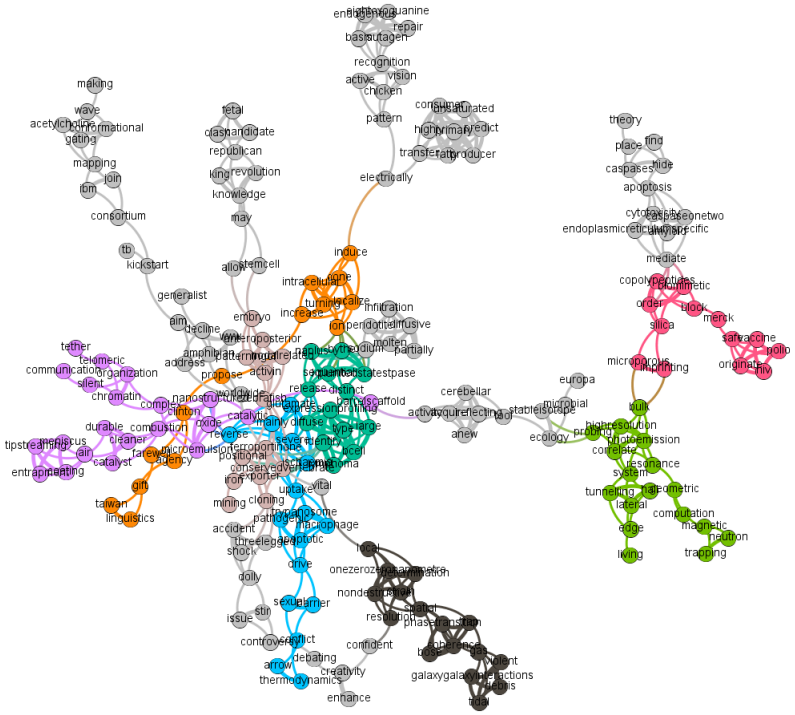


Figura 6.11: Componente gigante (64% da rede total) da rede crítica para o periódico *Nature*, no TVG deste trabalho, para o instante $t = 224$.

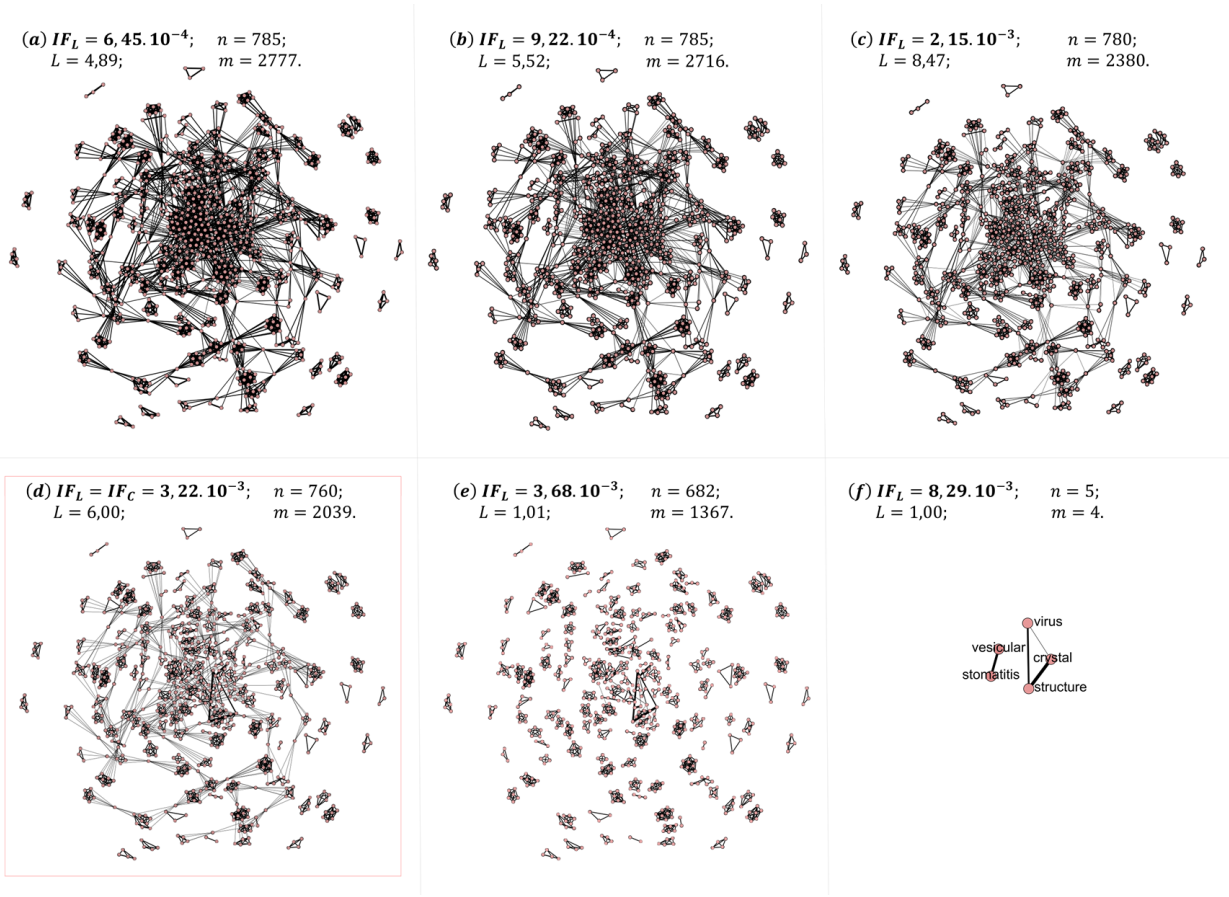


Figura 6.12: Atuação do índice IF na rede da *Science* em $t = 384$ (instante de maior valor de entropia de arestas normalizada para o TVG da *Science*). Cada rede contém arestas que possuem $IF \geq IF_L$. Os estágios mostram a rede para valores de IF_L: (a)mínimo, i.e. considerando todas as arestas da rede; (b) baixo, pouco maior que seu valor minímo; (c) imediatamente inferior ao valor crítico; (d) crítico; (e) imediatamente após o valor crítico; (f) elevado; as arestas estão ponderadas pelo IF. Observe que as arestas que possuem $IF = IF_C$ conectam diferentes grupos de palavras que possuem valores altos de IF.

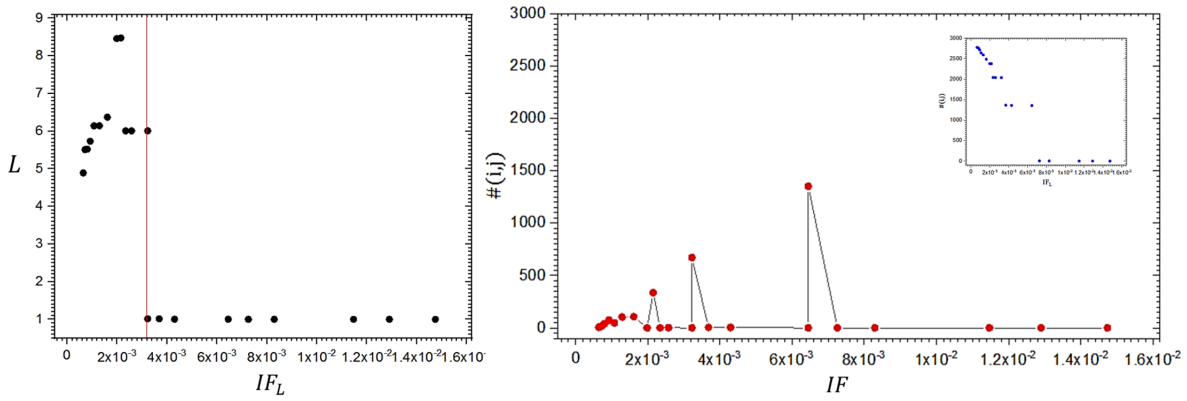


Figura 6.13: Encontrando a rede crítica para $t = 384$ da *Science*. Em (a) a evolução do caminho mínimo médio (L) em função da incidência fidelidade limite (IF_L), com ponto crítico em $IF_C = IF_L = 3,22 \times 10^{-3}$. Em (b) a distribuição dos pares de palavras $\#(i,j)$ por valor de IF. Em (b-inset) a quantidade de pares de palavras da rede em função de IF_L, tal como foi obtido por TEIXEIRA et al. (2010).

Parte IV

Resultados

Capítulo Sete

Resultados e Discussões

Para guiar a leitura, é recomendado a leitura dos objetivos específicos na Seção 1.5.2. Alguns desses objetivos, por serem propostas metodológicas, estão resolvidos no Capítulo 6:

Obj. 1 Elaborar um método que construa e analise uma rede semântica de títulos variável no tempo;

Obj. 3 Elaborar um método que permita mensurar a diversidade de vocabulário de uma rede semântica de títulos, através do conceito de entopia da informação de Shannon;

Obj. 8 Elaborar um método que identifique o principal vocabulário e suas conexões nas redes críticas em momentos de alta e baixa diversidade de vocabulário.

Para os outros objetivos específicos, as próximas seções propõem e discutem suas resoluções.

7.1 Entropias e seus limites teóricos para as redes semânticas de títulos

As Equações 6.3, 6.4, 6.5 e 6.6 (Seção 6.5) foram aplicadas às janelas de tempo para as redes de títulos para os dois periódicos. A Figura 7.1 mostra o comportamento da entropia de vértices e arestas para janela de 8 semanas que avança ao logo do tempo semana a semana, como também os valores máximos e mínimos para as entropias de vértice e aresta, seguindo os passos da Seção 6.5.1.

As entropias reais foram reescalonadas $H^{'} = (\frac{H - H_{min}}{H_{max} - H_{min}})$ para que $0 \leq H^{'} \leq 1$ (Figura 7.2). O reescalonamento retira o efeito do tamanho e permite uma melhor comparação entre os TVGs, quanto a diversidade de vocabulário dos títulos usados para construir as redes semâticas dos periódicos. Na Figura 7.2, as linhas horizontais dividem os valores em quatro regiões para $H^{'} = (0,00 - 0,25; 0,25 - 0,50; 0,50 - 0,75; 0,75 - 1,00)$ que podem ser úteis para classificar periódicos quanto à diversidade do vocabulário e suas conexões em suas redes de títulos.

Os valores da entropia do vértice são mais altos e variam substancialmente menos que os

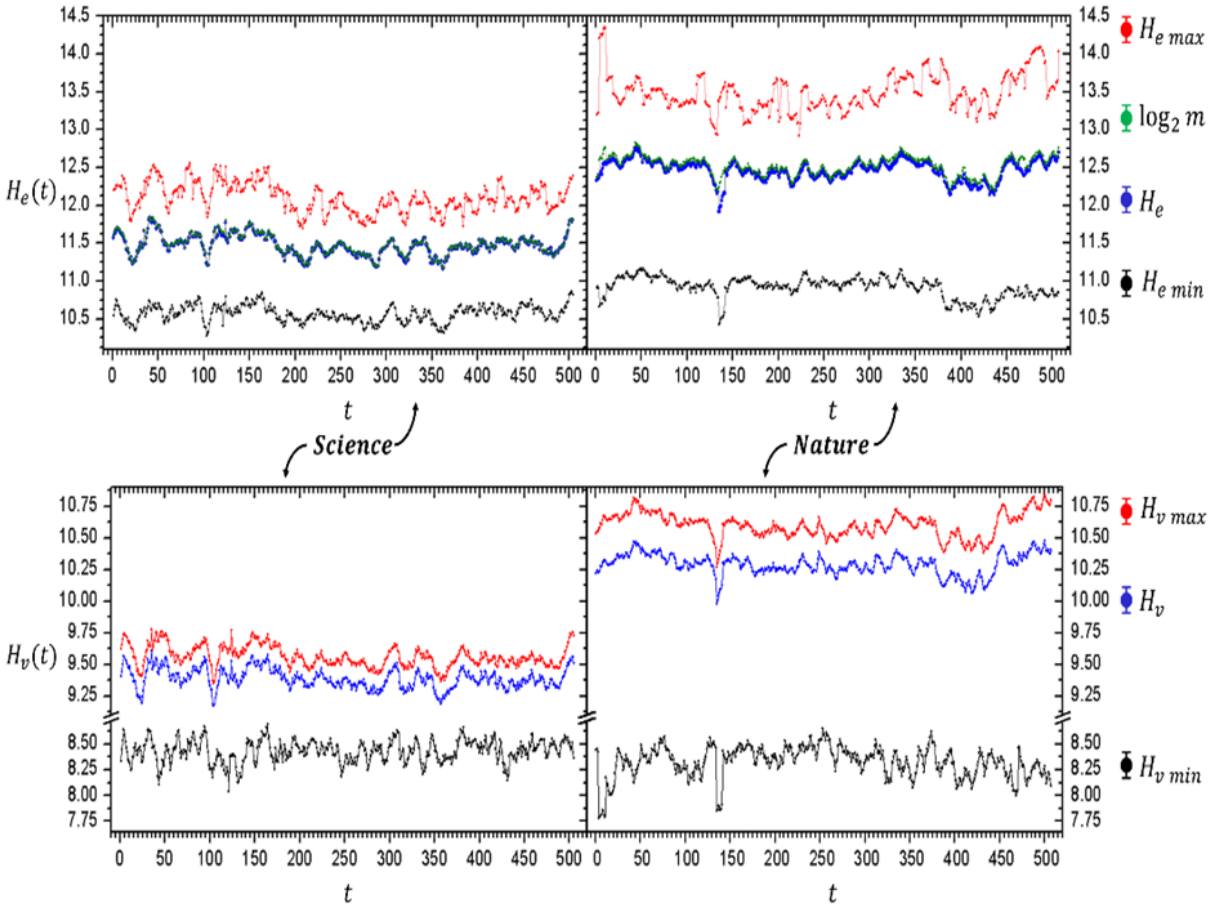


Figura 7.1: Entropia de vértices (H_v) e entropia de arestas (H_e) para os TVGs dos periódicos em função do tempo t dado em semanas, com janela deslizante $(w_8, 1)$. As séries iniciam em 5 de Janeiro de 1999 para Science e 7 de Janeiro de 1999 para a *Nature*. Embora não seja necessariamente um valor máximo, $H_e = \log_2 m$ está no gráfico para mostrar o quão semelhantes e fortemente correlacionadas são as entropias das arestas com esses valores. Isto mostra que as janelas possuem pouca sobreposição de arestas, com potencial para mais. Fonte: (CUNHA et al., 2020a).

valores da entropia das arestas. Essa descoberta mostra que as janelas têm sobreposição mínima de arestas, o que nos leva a supor que elas possuem mais potencial para novos relacionamentos entre o vocabulário existente do que a adição de novo vocabulário.

Os momentos em que a entropia diminui do máximo podem indicar tendências para o vocabulário da revista ao longo do tempo, como ilustra a Figura 7.3, que mostra um pequeno trecho das séries de entropia de vértices. no intervalo $321 < t < 330$ *semanas* para a *Nature*, por exemplo, enquanto a Entropia máxima aumenta, devido a inclusão de novos vértices ($H_{v\ max} = \log_2 n$), a entropia real se mantém constante. Isto se deve a uma "resistência" em diversificar o vocabulário, o que não acontece para a *Science* no mesmo período.

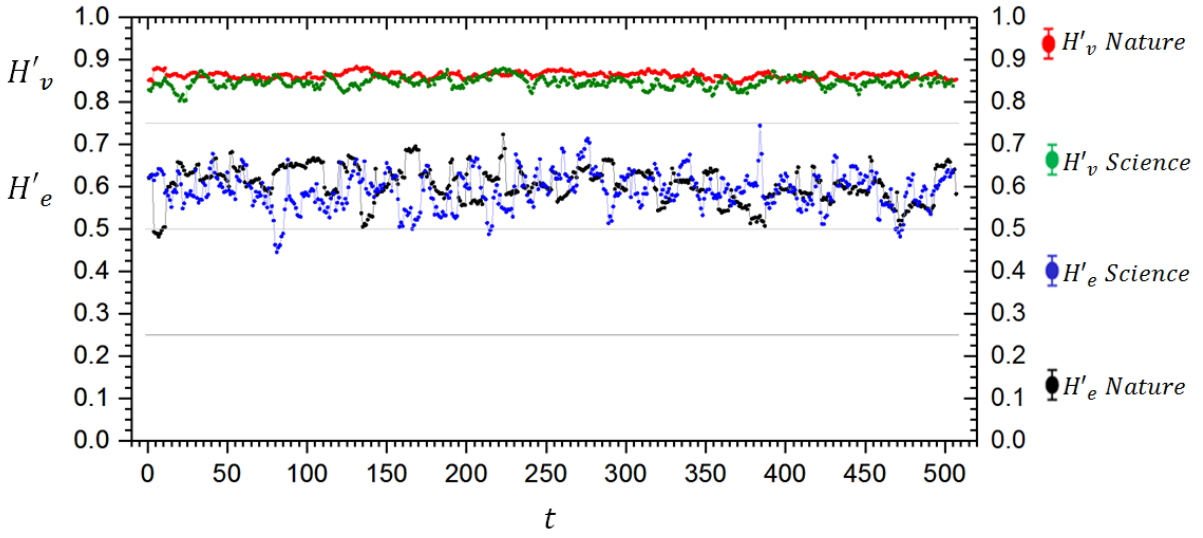


Figura 7.2: Entropias reescalonadas pelos extremos máximos e mínimos dos TVGs da *Nature* e *Science*, para $w_{8,1}(valoresdeentropiaentre0e1.)$. Fonte: (CUNHA et al., 2020a).

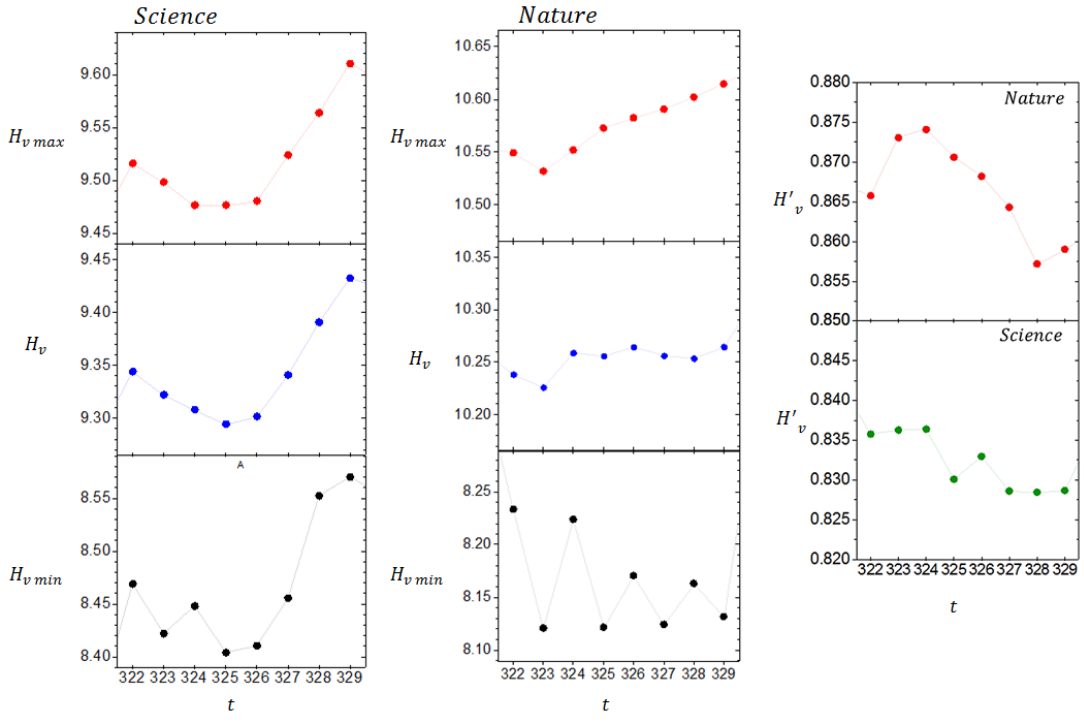
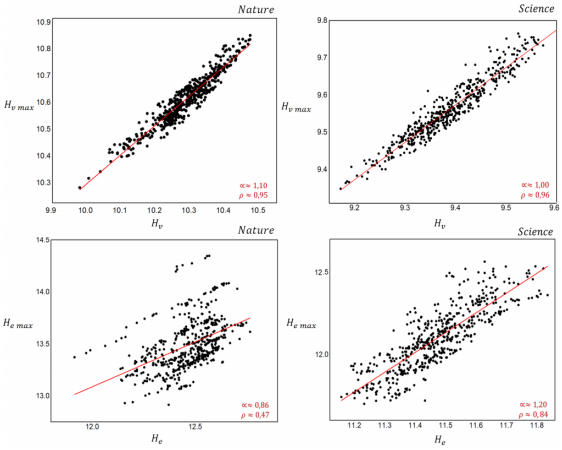


Figura 7.3: Excertos dos gráficos da Figuras 7.1 e 7.2 para $321 < t < 330$.

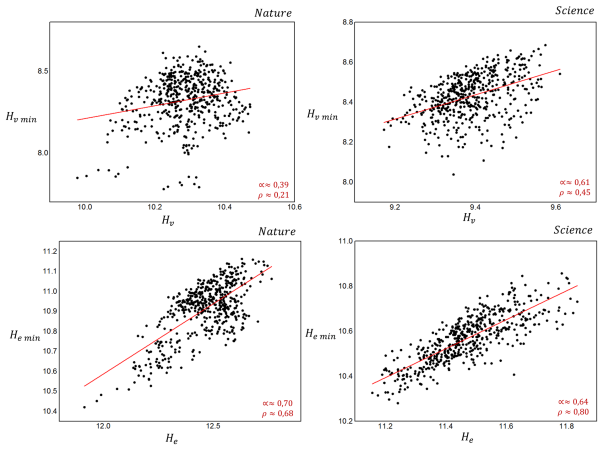
Observe que, respeitando os vínculos das cliques em cada janela, a Entropia mínima é a situação ideal para a menor diversificação de vocabulário (muitas palavras se repetem), a máxima entropia é a situação ideal para máxima diversificação de vocabulário e, a entropia real é intermediária, podendo estar mais próxima de um destes dois extremos ideias. O mesmo raciocínio vale para entropia de arestas, que se refere à diversificação das conexões do vocabulário da janela.

Além disso, em vários intervalos, H_v e H_e apresentam tendências de crescimento opostas. Sabemos que um aumento de H_e implica na geração de novas arestas, o que é possível devido ao incremento de vértices repetidos nos cliques, causando a diminuição de H_v. Em alguns períodos, uma tendência de crescimento oposta foi observada entre os Periódicos para a entropia de arestas: um Periódico alcançou um alto valor de entropia e o outro, um baixo valor de entropia.

Não obstante as medidas de entropia serem sensíveis ao tamanho da amostra, foi utilizado aqui todo o conjunto de dados do período coletado. Isso permite uma comparação adequada dos dois periódicos, mesmo com valores de entropia próximos. A Figura 7.4 mostra como as entropias reais de vértices e arestas estão correlacionadas com seus valores máximos (Subfigura 7.4(a)) e com seus valores mínimos (Subfigura 7.4(b)).



(a) Valores das entropias máximas, em função dos valores reais.



(b) Valores das entropias mínimas em função dos valores reais.

Figura 7.4: Valores das entropias mínimas e máximas, em função dos valores reais H_v e H_e para os dois periódicos. A linha mostra o ajuste linear para os pontos e evidencia a diferença entre a correlação das entropias de vértices e de arestas, onde α é o coeficiente de ajuste linear e ρ é o coeficiente de correlação de Pearson. Fonte: (CUNHA et al., 2020b).

Observamos uma forte correlação nas entropias de vértices com seus valores máximos, o que nos permite concluir que $H_v \simeq \log_2 n$ em qualquer janela de tempo. Já a entropia de arestas correlacionam melhor com seus mínimos, para os dois periódicos. Isto sugere que ao longo do tempo o vocabulário das revistas mantiveram alta diversificação para $w_{8,1}$. Por isto que as entropias de vértices obtiveram valores altos para entropia normalizada. O que poderia não ocorrer para o caso de redes em janelas de tempo maiores, que é explorado na próxima seção.

7.2 *Entropias para as RSTs crescentes.*

Uma vez que a medida de entropia apresentada aqui se refere à diversidade de vocabulário, vamos ver como cada rede semântica de títulos se comporta ao longo do tempo, considerando suas redes crescentes. A janela utilizada será $w_{t,0}$, ou seja, não desliza no tempo ($s = 0$) e seu tamanho aumenta com o passar do tempo, semama a semana ($\tau = t$).

A Figura 7.5 mostra a entropia de vértices, conforme o método proposto aqui (Seção 4.4) para as redes crescentes e suas respectivos valores máximos, já que são bem correlacionadas. Para cada periódico $t = 1$ representa todos os títulos das publicações da primeira semana[1], $t = 2$ são os títulos da primeira semana acrescidos aos títulos da segunda semana. Então, o instante t representa o acúmulo de todos os títulos da semana t mais os das semanas que o antecederam.

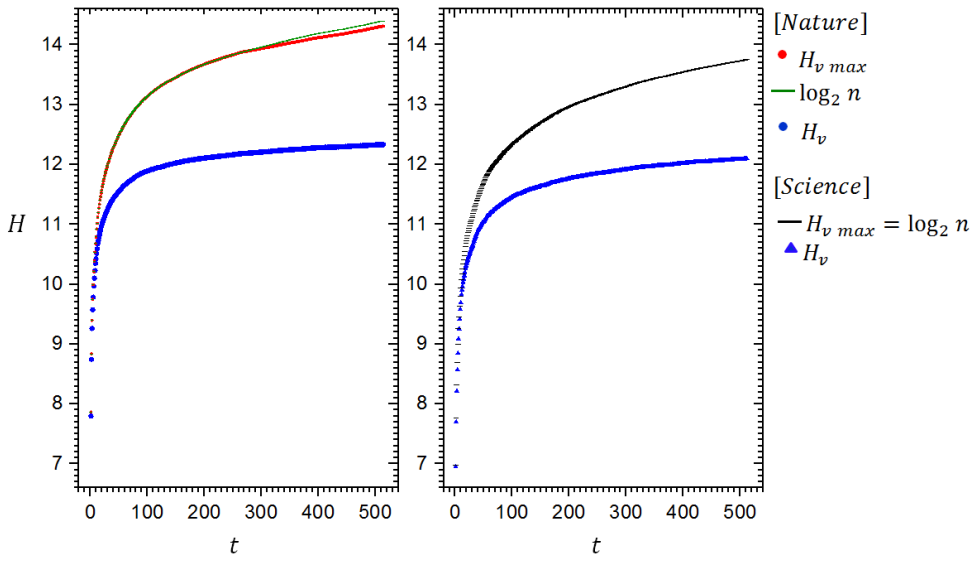


Figura 7.5: Evolução da entropia para a rede crescente da *Nature* e *Science* e suas respectivas entropias máximas. O tempo t representa a quantidade de revistas publicadas pelos periódicos, que como sabemos possuem frequência semanal de publicação. Para instantes onde $H_{vmax} < \log_2 n$, $n < n_q$.

Notamos que existe um aumento brusco no valor da entropia nas primeiras semanas até aproximadamente 1 ano de publicações, quando depois disto a taxa de variação é menor. A entropia real varia semelhante à entropia máxima, mas notamos que alguns pontos ocorre uma resistência para um aumento semelhante.

A Figura 7.6 mostra como evolui a razão $\frac{H_v}{H_{v\ max}}$ para as redes, permitindo a comparação das duas revistas sem o efeito do tamanho das redes.

Observamos pela Figura 7.5 que a *Nature* diversifica mais seu vocabulário com o passar

[1]Lembramos que uma semana aqui representa um número da revista, já que elas publicam semanalmente.

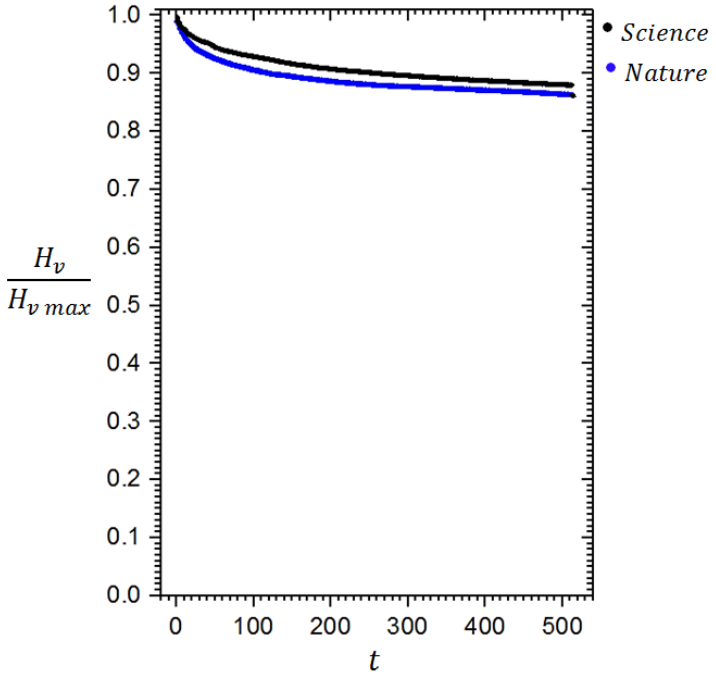


Figura 7.6: Evolução de $\frac{H_v}{H_{v\ max}}$ para as redes semânticas de títulos crescentes da *Nature* e *Science* entre 1998 a 2008.

do tempo. Mas, ao retirarmos o efeito do tamanho (Figura 7.6) vemos que a *Science* diversifica mais, já que a entropia de seus vértices estão mais próximas de seus valores máximos.

Para os TVGs deste trabalho, apenas na *Nature* acontece $n < n_q$, ou seja, o número de vértices menor que o número de títulos em uma dada janela temporal. Conforme veremos a seguir, este fenômeno acontece para janelas de tamanho $\tau \geq 270$. Nesta situação a *Configuração 1* (Seção 6.5.1) não será válida, já que ao distribuir n vértices nas cliques, cliques "vazias" precisarão ser preenchidas com vértices repetidos.

Nos instantes em que $H_{v\ max} < \log_2 n$, $n < n_q$. O instante crítico t_c em que $n = n_q$ tem um significado importante. É a partir dele que o vocabulário da revista "colapsa". Este termo é para ilustrar o fenômeno associado às redes de títulos crescentes, devido a seus vínculos: para $t > t_c$ não é possível com o vocabulário existente todas as cliques serem preenchidas sem haver repetição de vértices em cliques diferentes[2].

Para este instante crítico, $\frac{n}{n_q} = 1$. Ao buscarmos a configuração em que todos os vértices sejam diferentes, com o mínimo de arestas (*Configuração 1*, Seção 6.5.1), calculamos $q = \lfloor \frac{n}{n_q} \rfloor$ e $y = n - q.n_q$. Matematicamente, y é o resto da divisão $\frac{n}{n_q}$ e representa a quantidade de cliques que terão $q + 1$ vértices cada, quando a outra parcela das cliques

[2]Lembre-se que na mesma clique não pode ter vértices repetidos, mas um determinado vértice pode aparecer em mais de uma clique.

terão q vértices.

Obviamente, se $q = 0$ a *Configuração 1* torna-se impossível, fato que ocorre quando $n < n_q$, sendo necessário incluir (dos $n_0 - n$ vértices remanescentes) vértices repetidos para não deixar cliques vazias. A Figura 7.7 mostra como encontrar este instante crítico, com $q = \lfloor \frac{n}{n_q} \rfloor$ e $\frac{n}{n_q}$ em função do tempo. A Figura 7.8 mostra o valor de $y = n - q.n_q$ em função do tempo. Conforme foi visto em 6.5.1, os parâmetros y e q são importantes para deterinar a *Configuração 1*, que maximiniza a entropia de vértices.

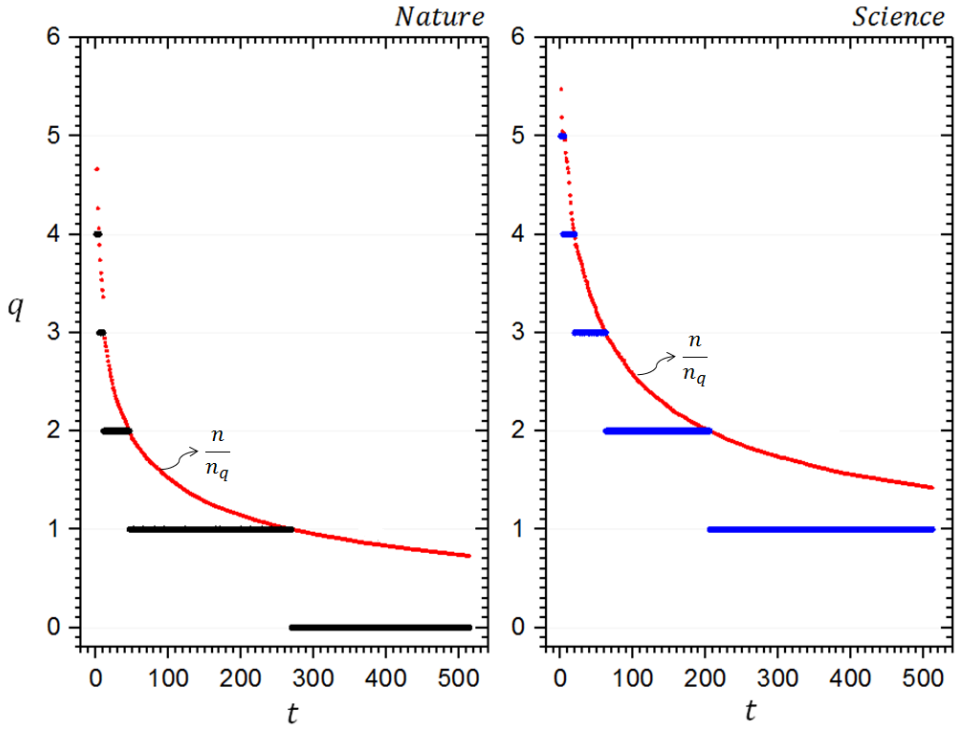


Figura 7.7: Razão inteira de n por n_q em função do tempo (azul e preto) e razão real q por n_q (em vermelho) para as redes crescentes da *Nature* e *Science* entre 1998 a 2008. Observe que o vocabulário da *Nature* "colapsa" quando $t = 269$, já que $q = 1$.

Vemos pelos gráficos das Figuras 7.7 e 7.8 que para *Nature* $q = 1$ em $47 \leq t \leq 269$. A partir daí, para $t \geq 270$ o vocabulário "colapsa". Ou seja, é necessário a repetição de vértices, sendo impossível a entropia ser $H_v = log_2 n$, mas sim algum valor menor que isto. Para a *Science*, este instante não foi encontrado para o tempo de vida do TVG. Mas, vemos indícios nos gráficos através das semelhanças com a *Nature* (inclusive na Figura 7.8) que este ponto aconteceria caso a base de dados da *Science* fosse maior, independente do intervalo.

A faixa de valores que este fenômeno ocorre e o valor do instante crítico certamente é um bom indicador para comparar periódicos, quanto a diversidade do vocabulário de seus títulos, através de suas redes semânticas de títulos crescentes.

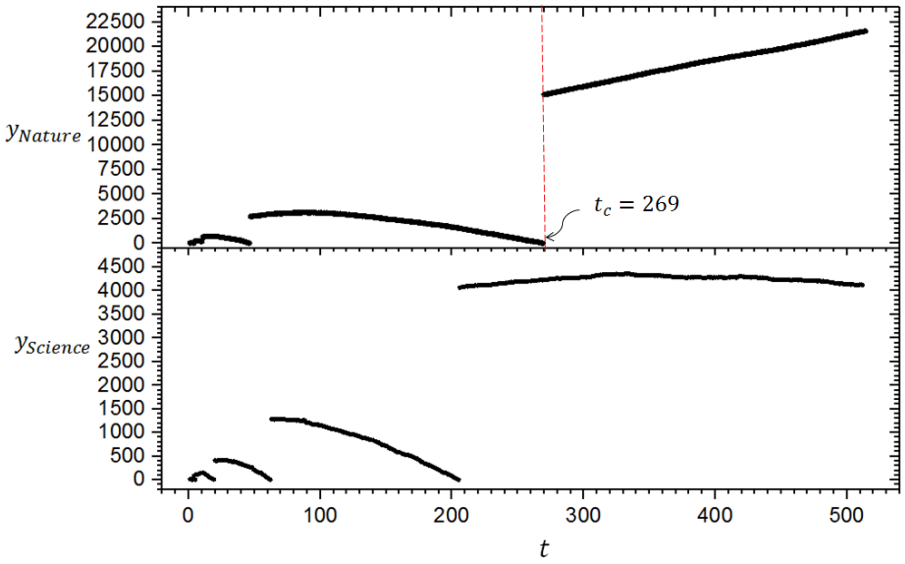


Figura 7.8: Valor $y = n - q.n_q$ para as redes crescentes da *Nature* e *Science* entre 1998 a 2008. Observe que a partir do instante crítico ($t_c = 269$ para a *Nature*) o valor de y cresce linearmente com o tempo.

A Figura 7.9 mostra a evolução da entropia considerando diferentes instantes para o início do TVG. O objetivo é identificar o mesmo padrão de evolução considerando a rede crescente.

Percebe-se que independente de quando no tempo a rede começa a crescer, ela segue o mesmo padrão de aumento de entropia, o que sugere um padrão para este periódico que se repete em qualquer época. A partir deste conjunto de dados, a Figura 7.10 mostra a distribuição da evolução da entropia para a rede crescente correspondente a 29 semanas (aproximadamente 1 semestre) de cada ano entre 1998 e 2008.

A partir da Figura 7.10 observamos que a distribuição dos valores é pouco dispersa em volta da média, o que indica que uma rede semântica de títulos crescente se comporta de maneira muito semelhante quanto a diversificação de seu vocabulário, qualquer que seja a época que ela comece.

7.3 Séries temporais para os índices de redes dos TVGs

A Figura 7.11 exibe as séries temporais de índices de redes para as janelas $w_{8,1}$ dos TVGS da *Science* e *Nature* entre 1998 e 2008.

A evolução dos índices ajuda a identificar tendências e comparar os periódicos, umas vez que as janelas de observação têm mesmo tamanho e passo e a frequência de publicações é a mesma. Por exemplo, em 2008, a densidade decai de forma mais expressiva para o

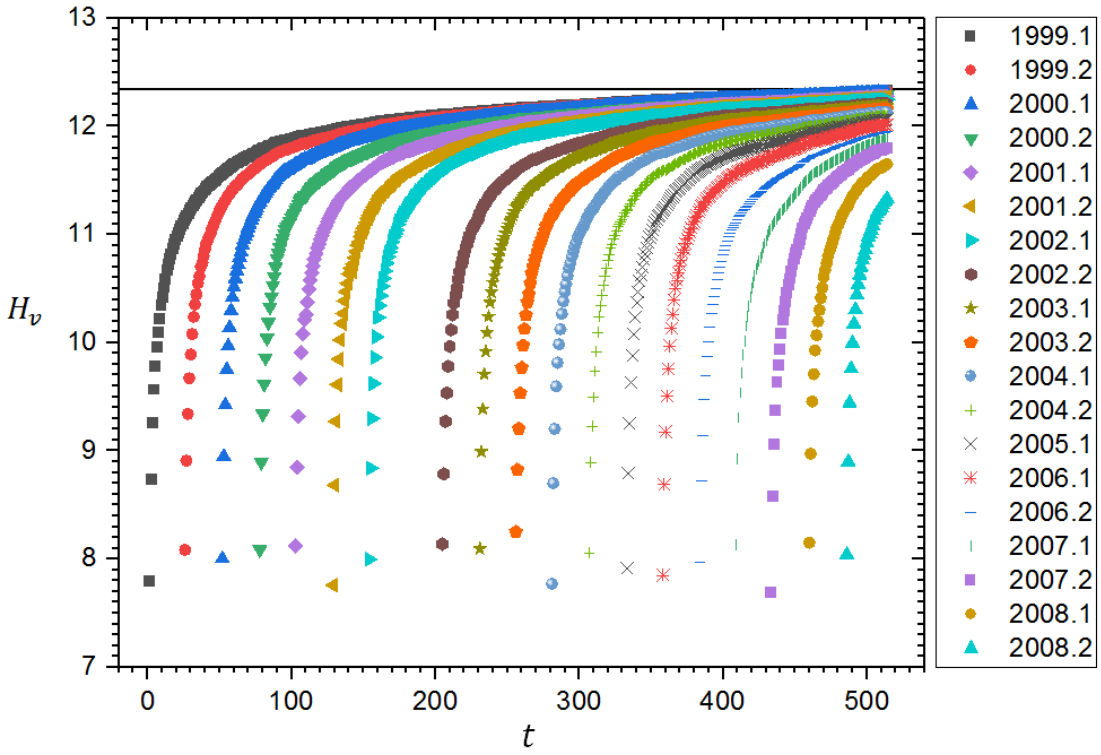


Figura 7.9: Entropia para rede crescente para diferentes instantes de inícios para o periódico *Nature*. Por exemplo, o conjunto de pontos que representa 2006.1 representa a evolução da entropia para a rede crescente que se inicia na primeira semana do primeiro semestre de 2006.

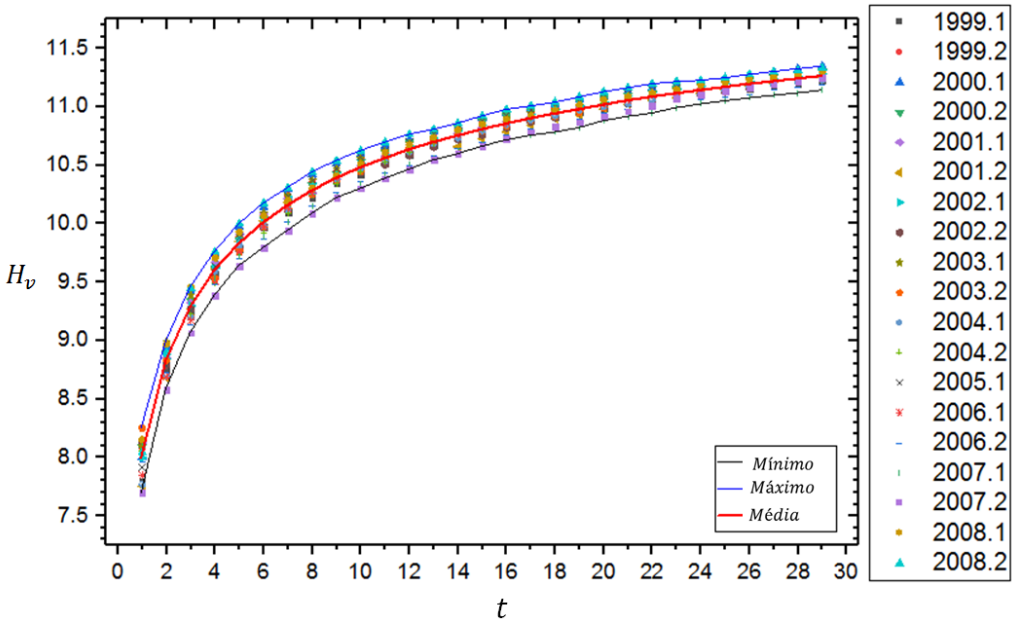


Figura 7.10: Entropia para rede crescente de 29 semanas a partir do ínicio do primeiro e segundo semestre de cada ano para o periódico *Nature*. As linhas de ajustes representam os valores médios, mínimos e máximos da distribuição.

periódico *Nature*, devido a um aumento no vocabulário (n), ao passo que a *Science* no mesmo período tem um aumento considerável no número de arestas.

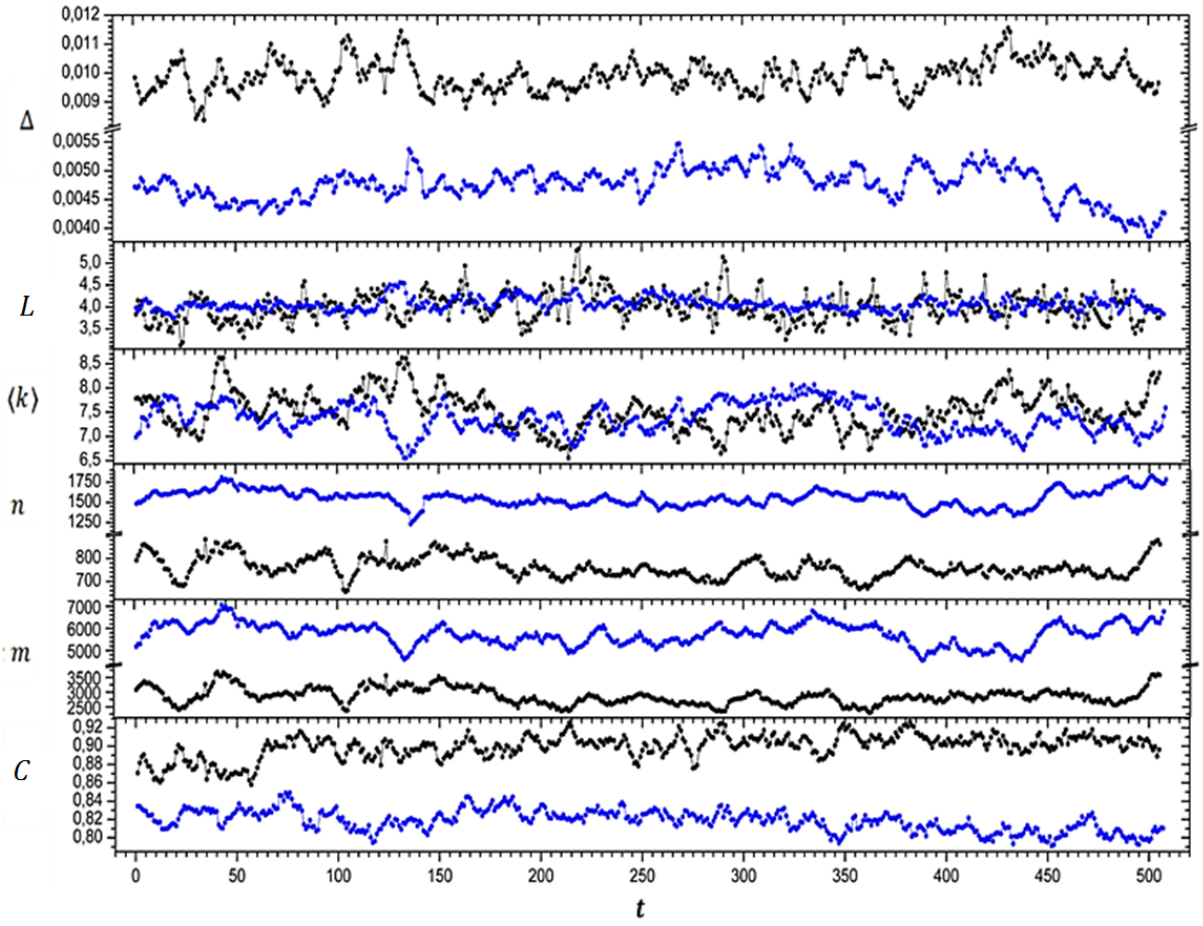


Figura 7.11: Evolução dos índices das janelas temporais entre 1999 e 2008 para a revista *Nature* (em zaul) e Science (em preto). Cada ponto no gráfico representa a rede de uma janela de oito semanas de publicações. Fonte: Cunha (2013) com acréscimo do Periódico *Science*.

O trabalho Cunha et al. (2013) faz uma aplicação interessante para o TVG do Periódico *Nature*, nas mesmas condições, utilizando o método DFA (Apêndice A) para identificar mémoria nas séries. Vale destacar aqui alguns resultados deste trabalho, que usaremos na próxima seção.

- Os índices tem correlação persistente para o intervalo $4 \leq \Delta t \leq 21$ *semanas*. Ou seja, uma variação positiva "hoje" tende a ser positiva após 4 semanas e isto se mantém até 21 semanas depois;
- O número de vértices n obteve a correlação mais forte, seguido pelo número de arestas m, pelo grau médio $\langle k \rangle$ e por último o coeficiente de aglomeração médio C;
- O Caminho mínimo médio tem correlação sem memória, portanto, segue um passeio aleatório para o mesmo período.

O próximo passo é verificar se essa correlação se mantém para outras escalas de observação,

ao calcular a entropia nestas séries, a partir do método MSE, para complementar os trabalhos supracitados.

7.4 Método MSE

7.4.1 Séries de índices de rede

Redes semânticas de títulos possuem memória. O método DFA, Apêndice A, foi aplicado nos títulos do Periódico *Nature* e encontrou correlações de longo alcance para as séries dos índices de redes nas janelas do TVG (CUNHA, 2013).

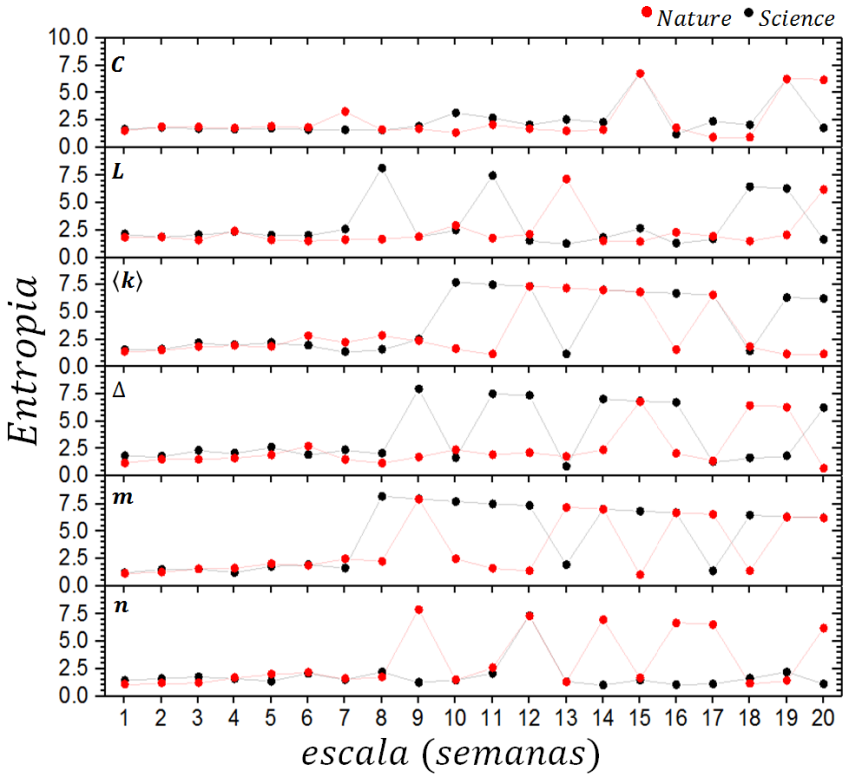
Ao aplicar o método MSE (Seção 5.4) para as séries dos índices de rede deveremos encontrar para a primeira escala de tempo resultado semelhante ao do trabalho supracitado. Além disso, é possível observarmos em mais escalas de tempo o comportamento dos índices. Por exemplo, algum índice da rede semântica de títulos que exibe algum grau de previsibilidade entre uma semana e outra (escala igual a 1), pode se apresentar diferente quando se considera de um mês pra outrro (escala igual a 4), por exemplo.

Isto significa que mesmo para escalas de tempo maiores, sempre há ganho de informação ao inspecionar sua medida, semelhante quando em escalas menores. Em geral, redes semânticas de cliques permitem o surgimento de comunidades fortes, já que os vértices são sempre bem aglomerados. E, mesmo que se varie a escala, esse comportamento de estrutura modular continua. Na Figura 7.12(a) observamos o valor da entropia amostral para 20 escalas diferentes utilizando o método MSE.

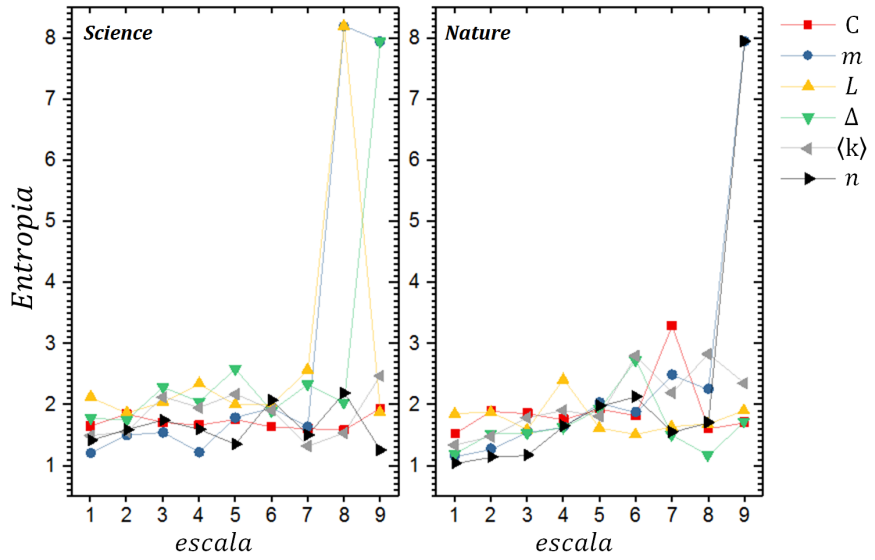
Na Figura 7.4.1, o caminho mínimo médio L inicia (escala=1) com maior valor de entropia, o que sugere uma maior imprevisibilidade de suas tendências. Na mesma escala, n possui a menor entropia (maior previsibilidade), o que corrobora com Cunha et al. (2013), que com o método *DFA* observou n sendo o mais correlacionado (aqui com menor entropia na primeira escala, Figura 7.12(b)) e L sem mémoria (aqui com maior entropia na primeira escala, ou seja, o mais imprevisível). Mas este comportamento se inverte para escala de tamanho 5 na *Nature*. Ou seja, para esta escala a variação no vocabulário é mais incerta que o caminho entre duas palavras.

7.4.2 Séries de entropia

A Figura 7.13 mostra o método MSE aplicado para as séries de entropia normalizada de vértices para os TVGs da *Nature* e *Science* que foram construídos com janelas semanais



(a) Método MSE para os índices de rede.



(b) Método MSE para os índices de rede em um mesmo gráfico, para escalas baixas, em semanas.

Figura 7.12: Método MSE para os índices de rede dos TVGs da *Nature* e *Science*, com $m = 2$ e $r = 0.15$.

$(w_{1,1})$ e bimensais $(w_{8,1})$, que avança semana a semana.

Observamos que com o aumento do tamanho da escala, as janelas semanais $(w_{1,1})$ tendem

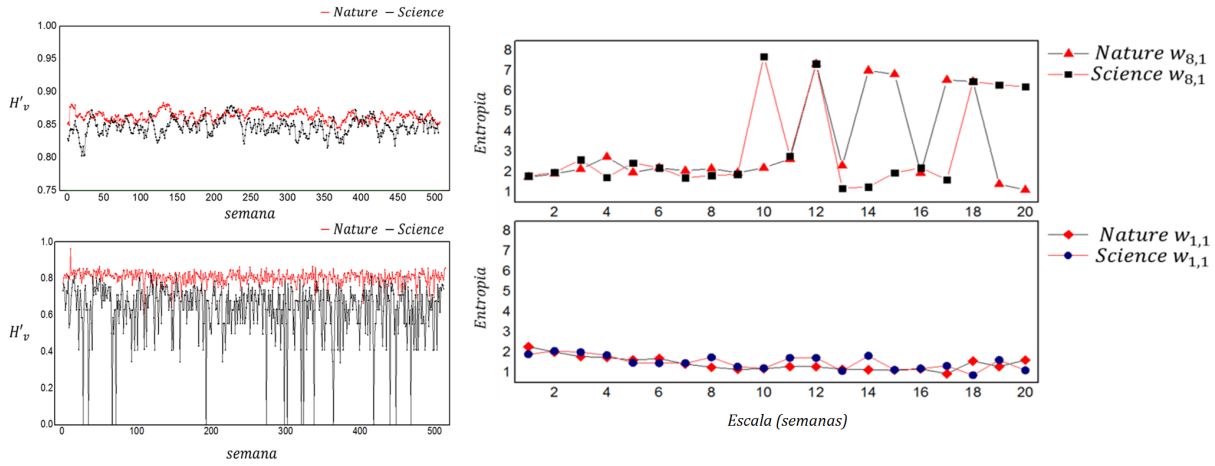


Figura 7.13: À esquerda, séries de entropia de vértices para janelas semanais ($w_{1,1}$) e bimensais ($w_{8,1}$) para os TVGs da *Nature* e *Science*. À direita, método MSE aplicado para estas séries.

a diminuir o valor da entropia, o que significa que a variabilidade dos valores da série não depende da escala, efeito também encontrado em séries heterocedásticas e no extremo de séries sem mémoria, como o ruído branco. Já para a janela bimensal ($w_{8,1}$) este comportamento acontece para escalas de tamanho até 10 (Nature). Após este valor, a entropia aumenta bruscamente, o que sugere surgimento de novas estruturas (informação) para escalas maiores.

Sendo assim, a janela bimensal sustenta uma tendência complexa para até escala de tamanho 9 ou 10 (entropia aproximadamente constante), quando que a janela mensal no mesmo período diminui sua complexidade. A janela bimensal possui mais palavras e seu avanço no tempo ocorre com sobreposição de dados ($\tau > s$, ver Seção 6.3). Isto colabora para correlações de longo alcance para os valores de entropia.

A título de curiosidade, ao aplicar o método DFA nestas séries, temos para $w_{8,1}$ o expoente $\alpha \approx 1,0$ (caso especial que acontece para o ruído $1/f$) e para $w_{1,1}$ o expoente $\alpha \approx 0,5$ (caso que acontece para o ruído branco)(PENG et al., 1995). A Figura 7.14 mostra este resultado. No Apêndice A, o método DFA é explicado com exemplo e aplicações no contexto deste trabalho.

De certa forma, este resultado reforça a importância de se usar uma janela de tamanho 8 *semanas* para estudar as redes de títulos em detrimento de uma janela de 1 *semana*, que se mostrou mais descorrelacionada no tempo e com queda do ganho de informação com aumento da escala de observação. Entretanto, a oscilação no valor da entropia nas janelas $w_{8,1}$ a partir da escala de tamanho 10 precisa ser melhor investigada em um trabalho futuro.

A próxima seção apresenta os resultados para a aplicação do índice IF para determinação

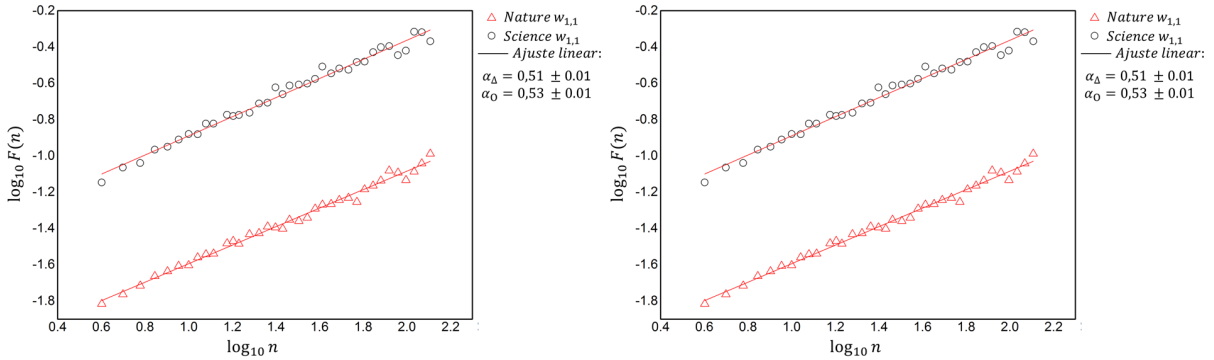


Figura 7.14: Método DFA aplicado nas séries séries semanais ($w_{1,1}$) e bimensais ($w_{8,1}$) de Entropia normalizada de vértices da *Nature* e *Science*. A Função $F(n)$ é ajustada em escala logartmica. Para $\alpha \approx 1,0$ temos séries que se assemelham ao ruído $1/f$ e para $\alpha \approx 0,5$ temos séries que se assemelham ao ruído branco.

de rede crítica e com isso,identificarmos os vértices principais em determinados momentos.

7.5 Incidência Fidelidade e rede crítica

Sabemos que a rede crítica é formada por arestas ponderadas pelo IF, que são importantes por manterem a rede coesa e conectar módulos (grupos de palavras altamente conectadas). Sendo assim, vale a pena investigar as redes críticas de janelas que obtiveram maiores e menores entropias para a distribuição de suas arestas. Usaremos esta configuração para identificar vértices importantes na rede, a partir de suas excentricidades e altos graus.

Para cada janela de tempo, foi encontrado a rede crítica utilizando o índice de incidência-fidelidade (IF, Seção 3.4). Estas redes nos permitem identificar os vértices (palavras) mais relevantes considerando suas conexões. As Figuras 7.15 e 7.16 mostram as redes críticas para a *Nature* em $t = 223$ (maior $H_e^{'}$) e $t = 7$ (a menor $H_e^{'}$), destacando os vértices considerados *hubs*. O grau de um vértice *hub* é dado por[3] $k_i^{hub} \geq \langle k \rangle + 2\sigma$.

Assim como TEIXEIRA et al. (2010), as redes estudadas neste trabalho apresentaram rede crítica para $IF_C = IF_L \sim 10^{-3}$. Na rede crítica oriunda do maior valor de entropia normalizada de arestas (Figura 7.15), os vértices *hubs* são fracamente conectados uns com os outros, indicando alta diversidade de vocabulário. Por outro lado, na rede crítica oriunda do menor valor de entropia normalizada de arestas (Figura 7.16, os vértices *hubs* são fracamente conectados uns com os outros, indicando robustez e recorrência de vocabulário.).

A Tabela 7.5 mostra os vértices que se destacam nas redes críticas oriundas dos instantes

[3]Caso a rede seja muito grande, pode-se aumentar o limiar para $\geq \langle k \rangle + 3\sigma$.

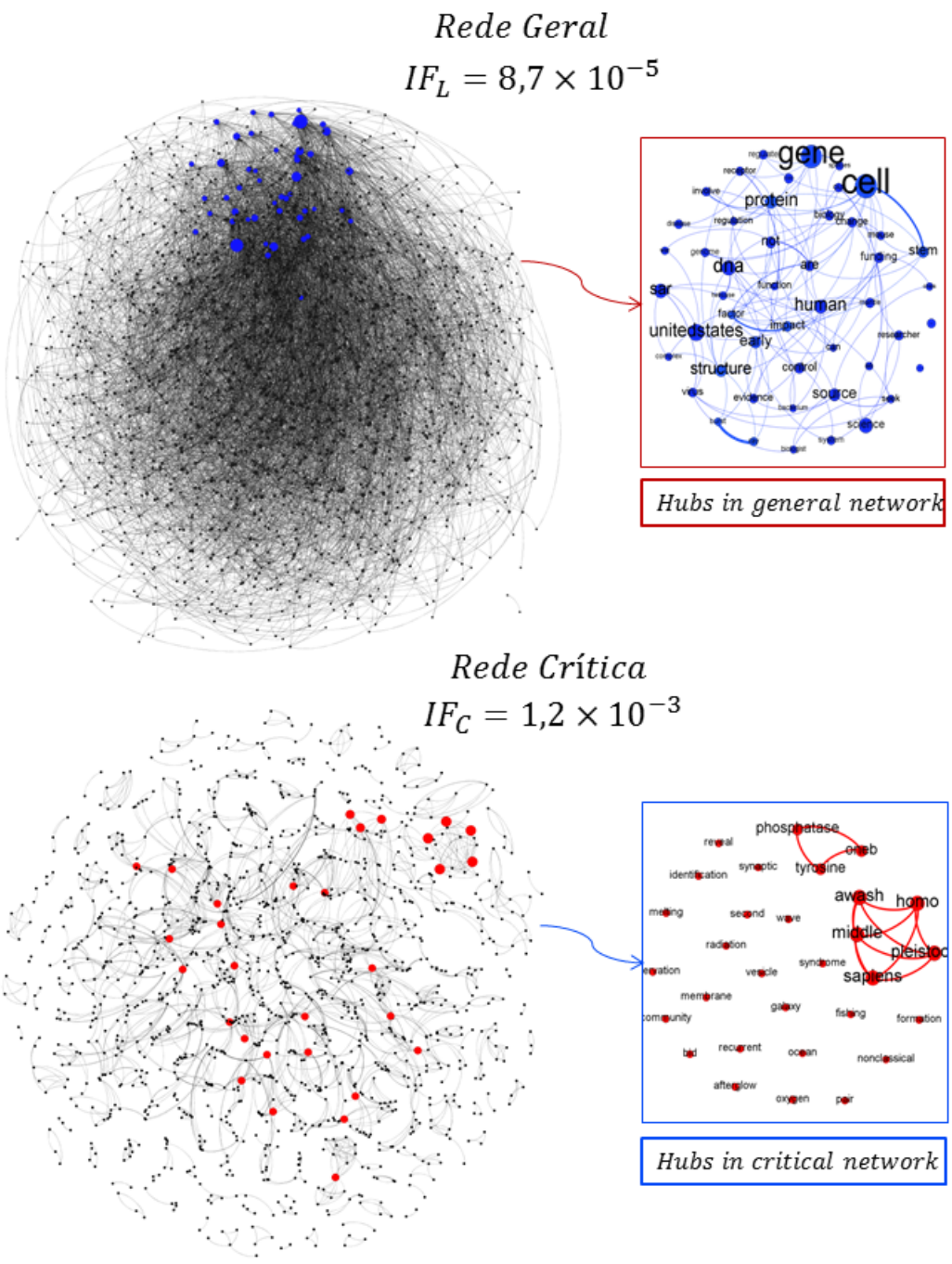


Figura 7.15: Redes geral e crítica para a *Nature* em $t = 223$, com a maior entropia $H_e^{'}$. *Hubs* ($k_i^{hub} \geq \langle k \rangle + 2\sigma$) são mostrados para a rede geral (destaque ao lado superior) e para a rede crítica (destaque ao lado inferior).

de maior e menor entropia normalizada de arestas para os TVGs de $w_{8,1}$. Os grupos de vértices estão agrupados por "*hubs*" - vértices de alto grau, que são os vértices mais conectados; "centro" - vértices de menor valor de excentricidade, que são os vértices de maior proximidade, em média, com todos os outros vértices da rede; "periferia" - vértices de mais alto grau de excentricidade, que são os vértices mais distantes, em média, de todos os outros da rede. A Figura 7.17 mostra as redes para estas configurações, destacando os vértices *hubs*, centro e periferia das redes.

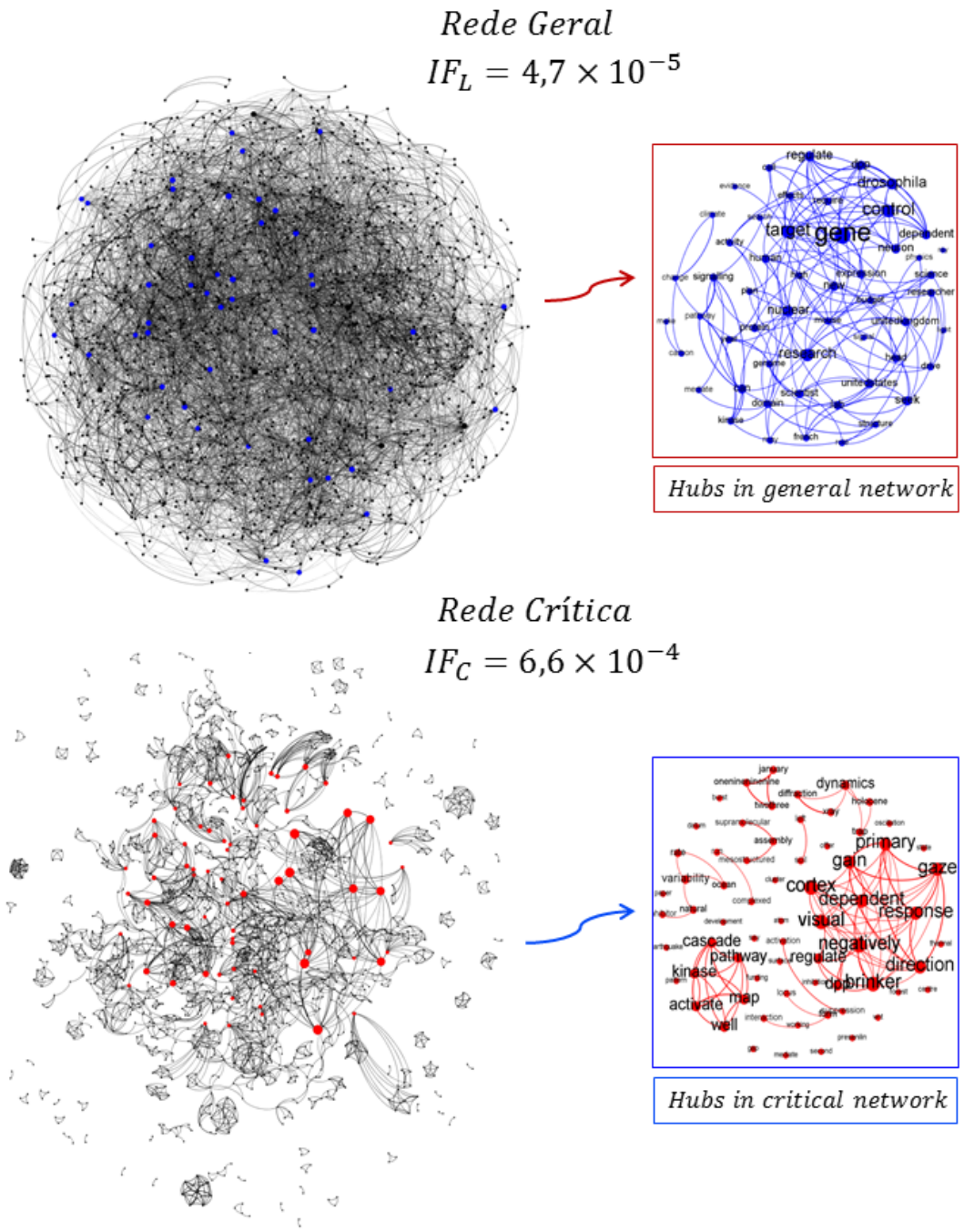


Figura 7.16: Redes geral e crítica para a *Nature* em $t = 7$, com a menor entropia $H_e^{'}$. Hubs ($k_i^{hub} \geq \langle k \rangle + 2\sigma$) são mostrados para a rede geral (destaque ao lado superior) e para a rede crítica (quadro inferior).

	(a) $t = 7$ Menor $H_e^{'}$ Crítica *Nature*	**(b)** $t = 81$ Menor $H_e^{'}$ Crítica *Science*	**(c)** $t = 223$ Maior $H_e^{'}$ Crítica *Nature*	**(d)** $t = 384$ Maior $H_e^{'}$ Crítica *Science*
centro	Second; say; gap; calcium; lymphocyte; messenger; cyclic; adpribose; power; export; store build; twofive; lord.	Marine.	Canadian; official.	Coastal.
Hubs	Kinase; pathway; dynamics; well; brinker; negatively; form; complexed; mediate; assembly; activate; map; cascade; inhibitor; activation; regulate; thermal; mesostructured; interaction.	quantum; target; fault; Escherichia; coli.	formation; syndrome; nonclassical.	cell; plant; protein; structure; mouse; human; virus; promote; complex; signal; dna; site; genetic; induce; kinase; crystal; disease.
periferia	welcome; wages; house; company; flat; earthers; creationism; matter; wave; fourwave.	Sequence; ribozyme; emergence; cell; lifespan; superoxide; dismutasecatalase; mimetics; oligodendrocyte; precursor; reprogram; become; multipotential; cns.	genetic; why; shall; politician; wage; osteoclast.	large; wind; great; plains; medieval; warm; period; early; pleistocene; glacial; integrate; summer; insolation; forcing; middle; paleolithic; bead; Israel; algeria; multiple; phosphorylation; reproducibility; rod; singlephoton; extinction; conservation; priority; anaphase; inactivation; spindle.

Tabela 7.1: Alguns vértices que se destacam nas redes de maior e menor entropia de arestas. Em (a), rede $IF_C = 6,63.10^{-4}$; em (b), $IF_C = 3,37.10^{-3}$; em (c) $IF_C = 1,22.10^{-3}$; em (d) $IF_C = 3,22.10^{-3}$. Para esta aplicação, os *hubs* foram obtidos com três desvios padrões acima do grau médio, ou seja $k_i^{hub} \geq \langle k \rangle + 3\sigma$.

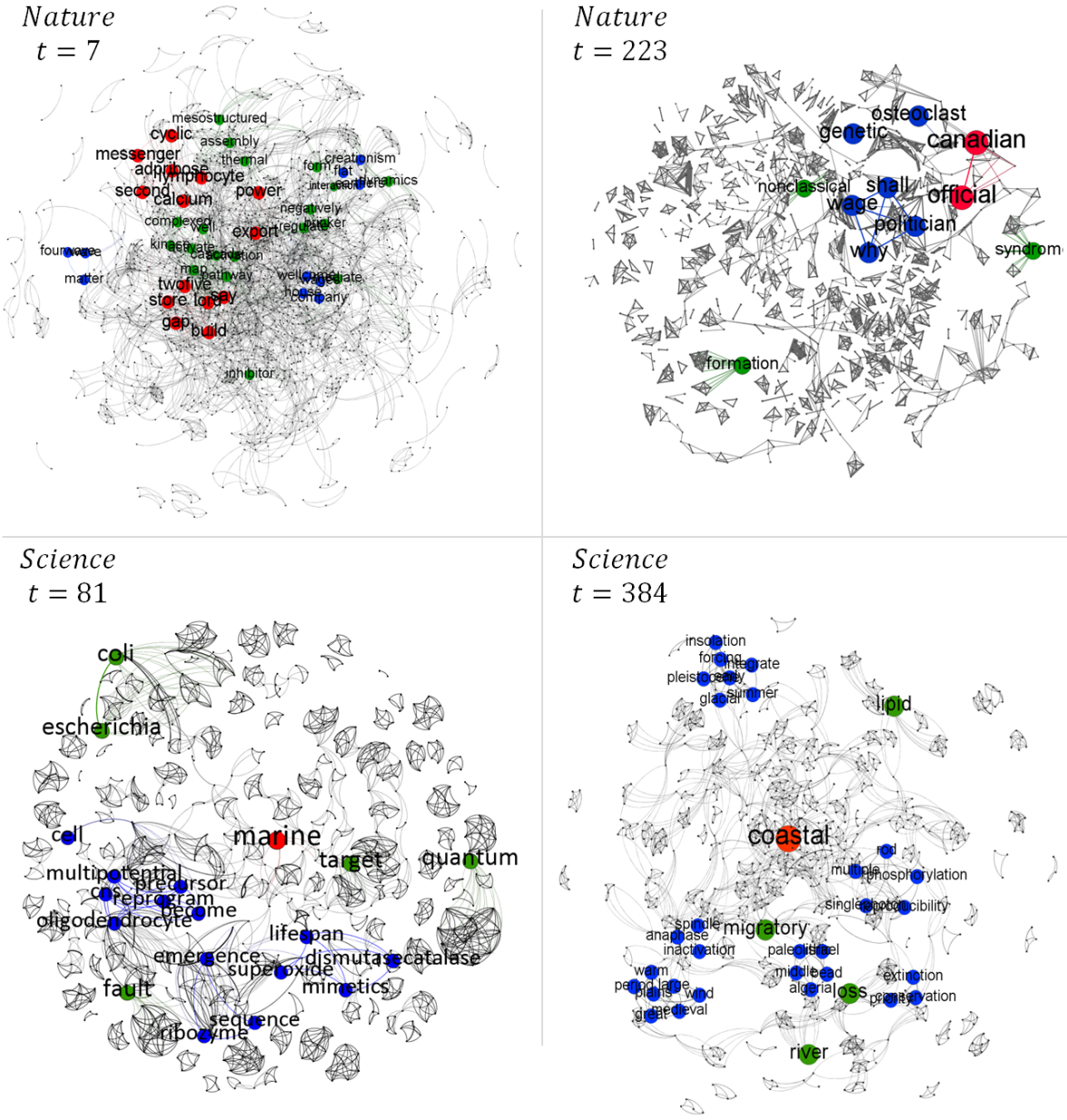


Figura 7.17: Redes em suas configurações críticas, para os intantes de maior e menor entropia normalizada de arestas. Em vermelho, os vértices de centro; em azul, vértices de periferia; em verde, vértices *hubs* ($k_i^{hub} \geq \langle k \rangle + 3\sigma$).

Capítulo Oito

Conclusão

No presente trabalho adaptamos a entropia de Shannon para colaborar com o estudo de uma rede semântica de cliques, tanto pela previsibilidade nas séries temporais dos índices de rede (método MSE), quanto pela diversidade do vocabulário e dos pares de palavras (método Entropia em redes de cliques).

O uso do método MSE nas séries dos índices de rede complementa estudo anterior que aponta alta correlação persistente para o vocabulário da *Nature* e correlação sem memória para o caminho mínimo médio. Neste trabalho vemos que a entropia amostral na primeira escala para os dois periódicos é menor para vértices e arestas e maior para o caminho mínimo médio. Para escalas maiores, vemos que o padrão de previsibilidade se mantém com baixa flutuação até escalas de tamanho 8 e 9, mas para o coeficiente de aglomeração este padrão se estende até a escala de tamanho 14, caracterizando complexidade. A entropia está associada a previsibilidade dos valores dos índices no tempo.

O método MSE nas séries de entropia de vértice nos revelou a diferença entre as janelas de 8 semanas e de 1 semana. A janela de tamanho 8 se mostra correlacionada e está mais próxima de uma estrura complexa (ruído $1/f$). Em contraste, a diminuição na complexidade das janela de tamnho 1 reflete uma capacidade reduzida do sistema de se ajustar a mudanças, caracterizando a não correlação, típica de séries de ruído branco. Isto reforça a escolha da janela deslizante $w_{(8,1)}$ para captura dos índices de rede em uma semântica rede de títulos variável no tempo.

A medida de entropia de vértices e arestas proposta neste trabalho avalia o grau de diversidade do vocabulário (entropia de vértices) e de suas conexões (entropia de aresta) em uma rede semântica de cliques. Ao reescalonar seus valores, retiramos o efeito do tamanho e as redes diferentes podem ser melhor comparadas. Entretanto, as configurações extremas de valores teóricos máximos e mínimos para entropia precisam respeitar os vínculos que existem no processo de formação de uma rede de cliques.

A evolução dos valores de entropia para uma janela de tempo de tamanho fixo nos mostra os momentos em que os valores reais da diversidade se aproximam mais da configuração ideal de entropia máxima ou mínima. Os valores extremos para entropia de arestas são úteis para extrair informações de vértices importantes da rede, a partir de índices que focam ems arestas. Aqui as redes foram filtradas para sua configuração crítica, através do incidência-fidelidade.

Os valores de entropia para rede crescente mostram que a diversidade do vocabulário tem aumento expressivo para os primeiros 6 meses (devido a considerável entrada de palavras novas na rede) e depois evolui lentamente (devido a recorrência de vocabulário a partir daí) semelhante a função $\log_2 n$, que é a entropia máxima, com exceção para a *Nature* $t \geq 269$ em que o valor é ligeiramente menor do que isto. Esta correção foi necessária porque neste intervalo, o número de títulos é maior que o vocabulário ($n < n_q$), sendo necessário obrigatoriamente palavras se repetirem, tornando $H_{v\ max} < \log_2 n$. Este fenômeno acontece para $w_{t,0}$, quando $t > t_c$. Este instante t_c foi chamado de instante crítico e, a partir dele, o vocabulário "colapsa".

Os valores de entropia calculados aqui não requereram o uso de um modelo nulo (isto é, rede aleatória) para comparação. O processo de construção das Configurações 1, 2 e 3 já é aleatorizado. Também é importante enfatizar que uma rede de cliques possui um cluster alto e isso significa que não existe uma rede aleatória correspondente, pois em redes aleatórias o coeficiente de cluster tende a zero ($C \to 0$) (WATTS; STROGATZ, 1998).

Há uma forte correlação entre os valores de entropia e seus respectivos valores máximos, especialmente para entropia de vértices. É razoável dizer que é equivalente calcular a entropia máxima para estimar a entropia, em mesmo nos casos em que $n < n_q$. As Figuras 7.1 e 7.2 mostram que os periódicos têm uma diversidade maior de palavras que pares de palavras. Isto sugere que ao longo do tempo o vocabulário das revistas mantiveram alta diversificação para $w_{8,1}$.

A mensuração da diversidade de vocabulário e a diversidade de conexões entre palavras em uma rede semântica de títulos de artigos científicos nos permitem acompanhar (i) o surgimento de novas ideias ao longo do tempo, representadas pelo aumento da diversidade de vocabulário dos títulos ou (ii) pela robustez e consolidação de ideias e interesses de autores e editores de uma revista em um determinado período de tempo.

A observação de ideias (representada pelas redes semânticas de títulos) com base nas medidas de entropia propostas nos ajuda a entender o "caminho" da produção científica (ou seja, ideias e interesses de autores e editores). Ao aplicar o índice IF, percebemos que na rede crítica é possível identificar os principais temas e como eles estão vinculados por meio de seu vocabulário (especificamente, maior diversidade do vocabulário para redes com alta entropia e robustez e recorrência do vocabulário para redes com baixa entropia).

O uso da modelagem por redes semânticas vem sendo usado cada vez mais em sistemas de representação do conhecimento. No contexto da psicologia cognitiva, redes semânticas representam uma faceta da memória declarativa de quem profere o discurso. A medida de entropia como diversidade de vocabulário, proposta neste trabalho, colabora com a ava-

liação de uma propriedade emergente do discurso coletivo nos Periódicos científicos. Este discurso traz elementos da memória declarativa semântica dos indivíduos responsáveis pela construção dos títulos dos artigos, i.e. autores e colaboradores que constroem os artigos e também dos avaliadores e editores de um Periódico que filtram os mais adequados para publicação científica. Sendo assim, este é um estudo de um fenômeno social, humano e global.

Para trabalhos futuros, o método de entropia de vértices e arestas em redes de cliques pode (a) ser combinado com o surgimento de comunidades nas redes; (b) correlacionado com outros indicadores específicos para esse tipo de rede (por exemplo, diâmetro de referência e fragmentação (FADIGAS; PEREIRA, 2013)); (b) adaptado para sistemas modelados em redes semânticas de linha e círculo; (c) aplicado nas RST de outros periódicos científicos e (d) aplicado à Repositórios de ciência que não possui filtros (editores, avaliadores, etc.).

Esta tese abre caminhos para se pensar em como uma rede de cliques pode evoluir, mais especificamente uma rede semântica de títulos, levando em conta os vínculos presentes em seu processo de formação (por exemplo, a maior probabilidade de um vértice é $p_{imax} = \frac{n_q}{n_0}$). Qual sentido termodinâmico existe na evolução destas redes? Estes conceitos certamente poderão ser estendidos para outros sistemas que também possuem vínculos e que sejam possíveis de se calcular suas entropias, principalemnte os que possam ser comparados em termos de entropia termodinâmica.

Parte V

Apêndice

Apêndice A

Detrended Fluctuation Analysis (DFA)

Proposto por Peng et al. (1994) para estudar sequências de genes no DNA, este método consiste em encontrar e quantificar correlações de longo alcance em uma série temporal não estacionária.

A.1 O método DFA.

Vamos ilustrar o método com as séries temporais da Figura A.1, como exemplo. Na Figura, temos os valores das entropias de vértices calculadas para as redes semânticas de títulos da *Nature* e *Science*, para uma janela de observação de 1 semana ($w_{1,1}$).

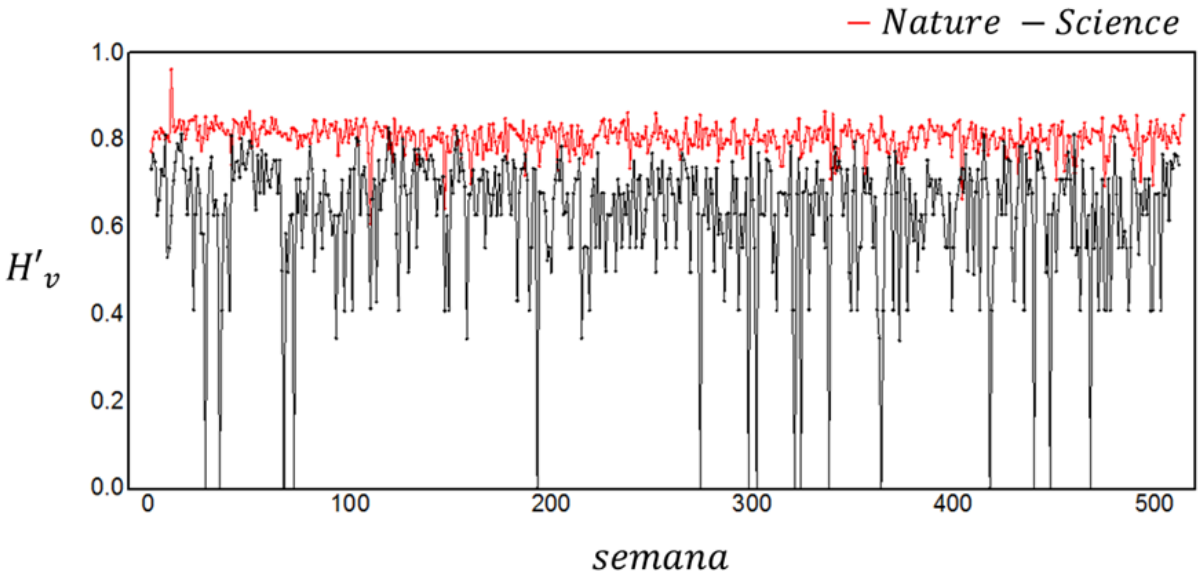


Figura A.1: Séries temporais das entropias de vértices das redes semânticas de títulos da Nature e Science no período 1999 a 2008, em uma janela temporal de 1 semana que avança semana a semana ($w_{1,1}$). A normalização é dada por $H^{'} = \frac{H - H_{min}}{H_{max} - H_{min}}$. Os valores de entropia (real, máximo e mínimo) são explicados na Seção 6.5.

Para realizar uma análise DFA, quatro passo devem ser seguidos (PENG et al., 1994):

1) A partir da média $\overline{y}$ (Equação A.1) da série original $y(i) = \{y(1), y(2), ..., y(N)\}$, obtemos a série integrada $Y(k)$, equação A.2.

$$\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y(i) \tag{A.1}$$

$$Y(k) = \sum_{i=1}^{k} (y(i) - \overline{y}) \tag{A.2}$$

2) A série integrada é então dividida em N/t caixas (intervalos de tempo) de tamanho t. Em cada caixa, o conjunto de pontos é ajustado pela melhor reta $Y_n(k)$ (método dos mínimos quadrados), sendo que $n = \{1, 2, ..., N/t - 1\}$ é a numeração de cada caixa, considerando a parte inteira de N/t. A Figura A.2 mostra as séries integradas da Figura A.1.

3) Em cada janela é feito um ajuste polinomial. A Figura A.2 mostra duas séries integradas, com exemplo de ajuste linear para um dado tamanhos de janela.

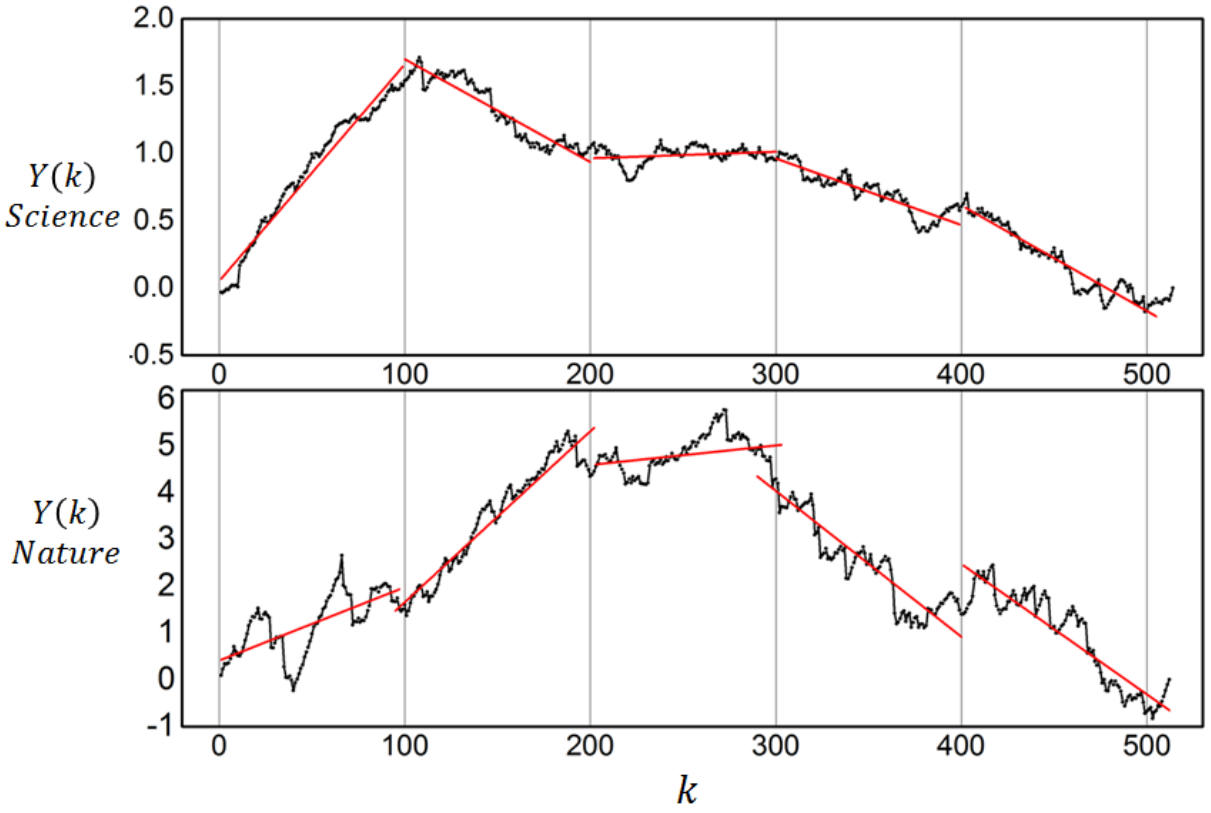


Figura A.2: Séries integradas, obtidas pela retirada das médias ao acumular os valores das séries originais. Os pontilhados verticais corresponde ao tamanho da caixa $n = 100$. As retas sólidas representam as tendências em cada caixa, através do método dos mínimos quadrados.

4) Posteriormente, para retirar as tendências da série, calcula-se a função flutuação (Equação A.3) em cada janela:

$$F(n) = \sqrt{\frac{1}{n}\sum_{k=1,}^{n}[Y(k) - Y_n(k)]^2} \tag{A.3}$$

Feito isto, se $F(n)$ se comportar como uma lei de potência, $F \sim n^{\alpha}$, então a série possui auto-afinidade, ou seja, a sequência de valores é correlacionada no tempo. Temos algumas opções para o valor de α (PENG et al., 1995):

Esta correlação é dita sem memória se $\alpha = 0,5$. 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 seguem um "passeio aleatório", como acontece no movimento browniano ou ruido branco. Este fenômeno também é alcançado quando se embaralha a ordem dos valores da série original.

Para valores de $0.5 < \alpha \leq 1,0$, as correlações são persistentes. Ou seja, uma tendência positiva no passado é mais provável de continuar positiva a longo alcance. Já para $0 \leq \alpha < 0,5$, os dados teriam uma correlação anti-persistente. Isto significa tendências opostas para o valor do sinal da série, ou seja, uma tendência positiva no passado tem alta probabilidade de tornar-se negativa no futuro.

O caso especial $\alpha = 1$, corresponde ao ruído $1/f$. Para $\alpha > 1$, exite correlação, porém deixa de ter a forma de uma lei de potência; $\alpha = 1,5$ corresponde ao ruído marrom, que é a integração do ruído branco.

A Figura A.3 mostra a função de correlação $F(n)$ e o valor de α para duas regiões em cada gráfico.

Observamos que para janelas de tamanho até $n = 10$, os valores de entropia de vértice possuem correlação persistente ($\alpha > 0,5$).Para janelas maiores temos passeio aleatório ($\alpha \simeq 0,5$).

A.2 Aplicação: índices de rede para os TVGs da Nature e Science.

Este método foi aplicado nas séries dos índices de redes, exibidos na Figura 7.11. Com exceção do índice *Diâmetro*, as séries se mostraram auto-afins. Os ajustes são exibidos na Figura A.3 e os parâmetros do ajuste encontram-se na Tabela A.1.

O gráfico da Figura A.4 revela o período comum para as correlações das séries, $t = [4, 21]$,

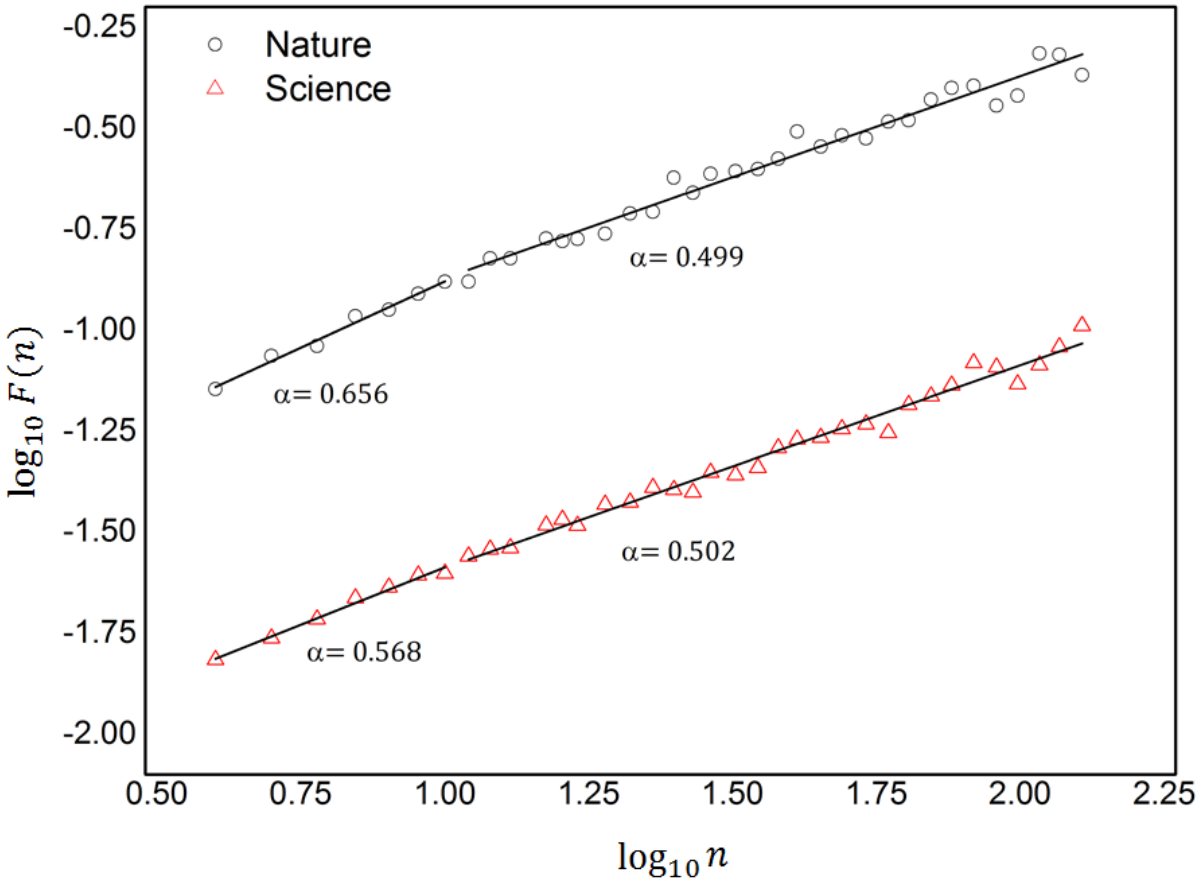


Figura A.3: Função $F(n)$, ajustada em escala logartmica. Para o intervalo $4 \leq n \leq 10$ a correlação é persistente nas duas séries (*Nature*: $\alpha = 0,656 \pm 0,033$ e *Science*: $\alpha = 0,568 \pm 0,028$). Para valores de $n \geq 10$ temos passeio aleatório (Nature: $\alpha = 0,499 \pm 0,16$ e Science: $\alpha = 0,502 \pm 0,013$).

Método DFA							
Nature	***m***	***n***	***L***	$\langle k \rangle$	***D***	***C***	Δ
α	0,767	0,783	0,499	0,640	–	0,550	0,661
erro	0,006	0,007	0,016	0,006	–	0,0076	0,013
$R^2(ajuste)$	0,999	0,999	0,985	0,999	–	0,998	0,994
Science	***m***	***n***	***L***	$\langle k \rangle$	***D***	***C***	Δ
α	0,560	0,561	0,446	0,623	–	0,589	0,624
erro	0,013	0,009	0,014	0,007	–	0,014	0,005
$R^2(ajuste)$	0,994	0,997	0,987	0,999	–	0,993	0,999

Tabela A.1: Parametros do Método DFA para as séries auto-afins: Expoente α, erros associados e R^2 dos ajustes. Fonte: Adaptado de Cunha et al. (2013).

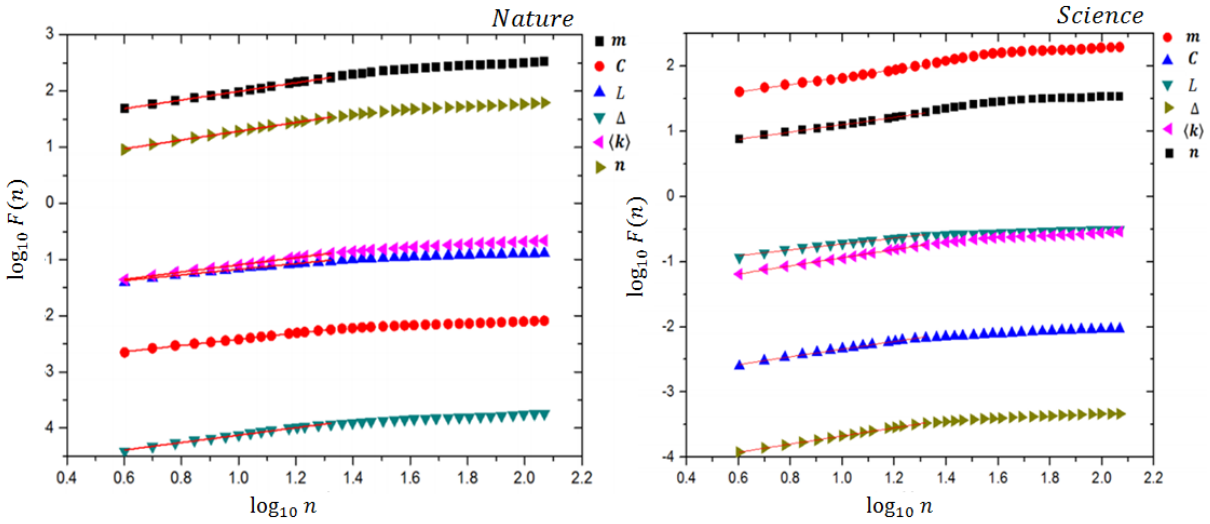


Figura A.4: Função $F(n)$, ajustada em escala logarítmica para séries dos índices para os TVGs da *Nature* e *Science* para janela de 8 semanas ($w_{8,1}$). Para o intervalo $4 \leq n \leq 21$ a correlação é persistente nas séries, exceto para o caminho mínimo médio que possui passeio aleatório ($\alpha \approx 0,5$). Fonte: Adaptado de Cunha et al. (2013).

que é o intervalo de 4 a 21 semanas (a partir do gráfico da Figura A.4, $log(4) = 1,2$ e $log(21) = 1,3$.).

O único índice que possui inclinação $\mathcal{H} \simeq 0.5$ é o *caminho mínimo médio* (L). Isto significa que, para o intervalo de tempo considerado, sua série temporal não possui "memória"e segue uma caminhada aleatória. Entretanto, as outras quantidades possuem memória para este intervalo de tempo.

Os índices n e m destacam-se pelos altos valores de α. Sabe-se que n representa o número de palavras diferentes (vocabulário dos títulos). E m representa os relacionamentos das palavras deste vocabulário. 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. Vemos que as correlações persistentes são mais fracas, em geral, para a Science.

Referências Bibliográficas

AGUIAR, M. S. *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. 3.4.1

AMBLARD, F. et al. On the temporal analysis of scientific network evolution. In: *CASoN*. [S.l.: s.n.], 2011. p. 169–174. 1, 2.3

ANDRADE, J. C. et al. Interdisciplinaridade e teoria de redes: rede semântica de cliques baseada em ementas e rede de componentes curriculares. *iSys-Revista Brasileira de Sistemas de Informação*, v. 12, n. 3, p. 24–52, 2019. 3.2

BARABÁSI, A.-L. *Network science*. [S.l.]: Cambridge University Press, 2016. 2.1

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. *Science*, v. 286, p. 509–512, 1999. 1, 2.1, 3.1

BARABÁSI, A.-L. et al. Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, Elsevier, v. 311, n. 3, p. 590–614, 2002. 2.3

BELLMAN, R. On a routing problem. *Quarterly of applied mathematics*, v. 16, n. 1, p. 87–90, 1958. 2.3

BRILLOUIN, L. *Science and information theory*. [S.l.]: Courier Corporation, 2013. 1, 4

CALDEIRA, S. *Caracterização de Rede de Signos Linguísticos: um modelo baseado no aparelho psíquico de Freud*. Tese (Doutorado) — Mestrado Interdisciplinar em Modelagem Computacional. Fundação Visconde de Cairu, Salvador, Brasil, 2005. 6.1

CALDEIRA, S. M. et al. The network of concepts in written texts. *The European Physical Journal B-Condensed Matter and Complex Systems*, Springer, v. 49, n. 4, p. 523–529, 2006. 1, 3.1, 3.2, 3.4

CARPI, L. C. et al. Analyzing complex networks evolution through information theory quantifiers. *Physics Letters A*, Elsevier, v. 375, n. 4, p. 801–804, 2011. 4.4

CASTEIGTS, A. et al. Time-varying graphs and dynamic networks. *International Journal of Parallel, Emergent and Distributed Systems*, Taylor & Francis, v. 27, n. 5, p. 387–408, 2012. 2.3, 2.4

CHEN, Q. et al. Optimal transport in time-varying small-world networks. *Phys. Rev. E*, American Physical Society, v. 93, p. 032321, Mar 2016. 2.3

COOKE, K. L.; HALSEY, E. The shortest route through a network with time-dependent internodal transit times. *Journal of mathematical analysis and applications*, Elsevier, v. 14, n. 3, p. 493–498, 1966. 2.3

COSTA, M.; GOLDBERGER, A.; PENG, C.-K. Multiscale entropy to distinguish physiologic and synthetic rr time series. In: IEEE. *Computers in Cardiology, 2002*. [S.l.], 2002. p. 137–140. 5.4, 5.2

COSTA, M.; GOLDBERGER, A. L.; PENG, C.-K. Multiscale entropy analysis of complex physiologic time series. *Physical review letters*, APS, v. 89, n. 6, p. 068102, 2002. 6.10

COSTA, M.; GOLDBERGER, A. L.; PENG, C.-K. Multiscale entropy analysis of biological signals. *Physical review E*, APS, v. 71, n. 2, p. 021906, 2005. 1, 5.1, 5.4, 6.6, 6.6

COVER, T. M.; THOMAS, J. A. *Elements of information theory*. [S.l.]: John Wiley & Sons, 2012. 4.3

CUNHA, M. do V. *Redes semânticas baseadas em títulos de artigos científicos.* Dissertação (Mestrado em modelagem computacional e tecnologia industrial) — Faculdade de Tecnologia Senai CIMATEC, Salvador, 2013. 1, 2.2, 3.3, 3.2, 5.1, 7.11, 7.4.1

CUNHA, M. V.; MIRANDA, J. G. V.; PEREIRA, H. B. B. Incidência fidelidade aplicada a rede semântica de tıtulos. In: *IV Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*. [S.l.: s.n.], 2015. 1, 3.3, 3.4.1

CUNHA, M. V. et al. Redes de títulos de artigos científicos variáveis no tempo. In: *Anais do II Brazilian Workshop on Social Network Analysis and Mining*. Porto Alegre, RS, Brasil: SBC, 2013. p. 194–205. 1, 1.1, 2.3, 3.3, 6.6, 7.3, 7.4.1, A.1, A.4

CUNHA, M. V. et al. Shannon entropy in time-varying clique networks. In: SPRINGER. *VIII International Conference on Complex Networks and Their Applications*. [S.l.], 2020. p. 507–518. 2.3, 6.7, 6.8, 7.1, 7.2

CUNHA, M. V. et al. Shannon entropy in time-varying semantic networks of titles of scientific paper. *Applied Network Science*, Springer Nature, v. 5, n. 1–17, p. 128, 2020. 1, 3.2, 4.4, 6.1, 6.6, 6.9, 7.4

DERÉNYI, I.; PALLA, G.; VICSEK, T. Clique percolation in random networks. *Physical review letters*, APS, v. 94, n. 16, p. 160202, 2005. 3.1

DOREIAN, P.; STOKMAN, F. *Evolution of Social Networks*. [S.l.]: Taylor & Francis, 2013. (Routledge Contemporary Human Geography). ISBN 9781136647321. 4

DOREIAN, P.; STOKMAN, F. N. *Evolution of social networks*. [S.l.]: Psychology Press, 1997. v. 1. 2.3

ERDOS, P. On cliques in graphs. *ISRAEL JOURNAL OF MATHEMATICS*, v. 4, p. 233–234, 1966. 2.1

FADIGAS, I. S. et al. Fifa world cup referees' networks: a constant-size clique approach. *Social Network Analysis and Mining*, Springer, v. 10, n. 1, p. 1–12, 2020. 1, 3.1

FADIGAS, I. S. et al. Análise de redes semânticas baseada em títulos de artigos de periódicos científicos: o caso dos periódicos de divulgação em educação matemática. *Educação Matemática Pesquisa*, Pontificia Universidade Catolica de Sao Paulo PUC-SP, Programa de Estudos Pós-Graduados em Educação Matemática, v. 11, n. 1, 2009. 1

FADIGAS, I. S.; PEREIRA, H. B. B. A network approach based on cliques. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 392, n. 10, p. 2576–2587, 2013. 3.1, 3.3, 6.5, 8

GARVEY, W. D. *Communication: the essence of science*: facilitating information exchange among librarians, scientists, engineers and students. [S.l.]: Elsevier, 2014. 1

GASTEL, B.; DAY, R. A. *How to write and publish a scientific paper*. [S.l.]: ABC-CLIO, 2016. 1

GRILO, M. et al. Robustness in semantic networks based on cliques. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 472, p. 94–102, 2017. 1, 3.1, 3.2, 3.3

HALPERN, J. Shortest route with time dependent length of edges and limited delay possibilities in nodes. *Mathematical Methods of Operations Research*, Springer, v. 21, n. 3, p. 117–124, 1977. 2.3

HALPERN, J.; PRIESS, I. Shortest path with time constraints on movement and parking. *Networks*, Wiley Online Library, v. 4, n. 3, p. 241–253, 1974. 2.3

HARTLEY, R. V. Transmission of information. *Bell Labs Technical Journal*, Wiley Online Library, v. 7, n. 3, p. 535–563, 1928. 4

HENRIQUE, T. et al. Mathematics education semantic networks. *Social Network Analysis and Mining*, Springer, v. 4, n. 1, p. 200, 2014. 3.3

HOLME, P. *Temporal networks*. [S.l.]: Springer, 2014. 2.3

HOLME, P.; SARAMÄKI, J. Temporal networks. *Physics reports*, Elsevier, v. 519, n. 3, p. 97–125, 2012. 2.3

HSU, C. F. et al. Entropy of entropy: Measurement of dynamical complexity for biological systems. *Entropy*, Multidisciplinary Digital Publishing Institute, v. 19, n. 10, p. 550, 2017. 1

JI, L. et al. Network entropy based on topology configuration and its computation to random networks. *Chinese Physics Letters*, IOP Publishing, v. 25, n. 11, p. 4177, 2008. 1

JR, L. R. F. .; FULKERSON, D. R. Constructing maximal dynamic flows from static flows. *Operations research*, INFORMS, v. 6, n. 3, p. 419–433, 1958. 2.3

LI, M. et al. Evolving model of weighted networks inspired by scientific collaboration networks. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 375, n. 1, p. 355–364, 2007. 2.3

LIMA, B. N. et al. Entropia: introdução à teoria matemática da (des) informação. 2012. 4

MILGRAM, S. The small world problem. *Psychology today*, New York, v. 2, n. 1, p. 60–67, 1967. 2.1

MIRANDA, D. B. de; PEREIRA, M. d. N. F. O periódico científico como veículo de comunicação: uma revisão de literatura. *Ciência da informação*, v. 25, n. 3, 1996. 1

MISHRA, S.; AYYUB, B. M. Shannon entropy for quantifying uncertainty and risk in economic disparity. *Risk Analysis*, Wiley Online Library, v. 39, n. 10, p. 2160–2181, 2019. 1

MOUSAVIAN, Z.; KAVOUSI, K.; MASOUDI-NEJAD, A. Information theory in systems biology. part i: Gene regulatory and metabolic networks. In: ELSEVIER. *Seminars in cell & developmental biology*. [S.l.], 2016. v. 51, p. 3–13. 1

MUELLER, S. P. M. A publicação da ciência: áreas científicas e seus canais preferenciais. 2005. 1.2

NASCIMENTO, J. O. d. et al. *Redes Sociais e Complexas: um modelo computacional para a investigação da pós-graduação Brasileira em Ensino de Física*. 2016. 110–114 p. 1, 3.2

NASCIMENTO, W. S.; PRUDENTE, F. V. Shannon entropy: A study of confined hydrogenic-like atoms. *Chemical Physics Letters*, Elsevier, v. 691, p. 401–407, 2018. 1

NETO, J. L. d. A. L.; CUNHA, M. V.; PEREIRA, H. B. B. Redes semânticas de discursos orais de membros de grupos de ajuda mútua= semantic networks of oral discourses of members of mutual aid groups. *Obra digital: revista de comunicación*, Universitat de Vic, n. 14, p. 51–66, 2018. 3.2, 3.4.1

NEWMAN, M. E. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 98, n. 2, p. 404–409, 2001. 1

NEWMAN, M. E. J. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical review E*, APS, v. 64, n. 1, p. 016132, 2001. 1, 3.1

NICOSIA, V. et al. Components in time-varying graphs. *Chaos: An interdisciplinary journal of nonlinear science*, AIP, v. 22, n. 2, p. 023101, 2012. 2.3

ORDA, A.; ROM, R. Shortest-path and minimum-delay algorithms in networks with time-dependent edge-length. *Journal of the ACM (JACM)*, ACM, v. 37, n. 3, p. 607–625, 1990. 2.3

PAPOULIS, A. *Probability & statistics*. [S.l.]: Prentice-Hall Englewood Cliffs, 1990. v. 2. 4

PENG, C.-K. et al. Mosaic organization of dna nucleotides. *Phys. Rev. E*, American Physical Society, v. 49, n. 2, p. 1685–1689, 1994. 5.1, A, A.1

PENG, C.-K. et al. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos: an interdisciplinary journal of nonlinear science*, American Institute of Physics, v. 5, n. 1, p. 82–87, 1995. 7.4.2, A.1

PEREIRA, H. B. B. et al. Density: A measure of the diversity of concepts addressed in semantic networks. *Physica A: Statistical Mechanics and its Applications*, v. 441, p. 81 – 84, 2016. ISSN 0378-4371. 1, 1.1, 3.1, 3.2, 3.3

PEREIRA, H. B. B. et al. Semantic networks based on titles of scientific papers. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 390, n. 6, p. 1192–1197, 2011. 1, 3.1, 3.3, 1, 6.1, 6.1, 6.2

PINCUS, S. M. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 88, n. 6, p. 2297–2301, 1991. 5.2

POWELL, W. B.; JAILLET, P.; ODONI, A. Stochastic and dynamic networks and routing. *Handbooks in operations research and management science*, Elsevier, v. 8, p. 141–295, 1995. 2.3

PRICE, D. J. *Little science, big science... and beyond*. [S.l.]: Columbia University Press New York, 1986. 1

PRICE, D. J. D. S. Networks of scientific papers. *Science*, JSTOR, p. 510–515, 1965. 1, 1.2

RAD, A. A. *Social Network Analysis and Time Varying Graphs*. Tese (Doutorado) — Université d'Ottawa/University of Ottawa, 2016. 2.3

RICHMAN, J. S.; MOORMAN, J. R. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, Am Physiological Soc, v. 278, n. 6, p. H2039–H2049, 2000. 5.3

RODRIGUES, A. Á. A. d. O. et al. Um método para analisar a temática de periódicos na saúde coletiva. *Revista Eletrônica de Comunicação, Informação & Inovação em Saúde*, v. 11, n. 1, 2017. 1

ROSA, M. G. Modelo empírico para analisar a robustez de redes semânticas. Faculdade de Educação, 2017. 4.4

SANTOS, C. C. R.; PEREIRA, H. B. B.; CUNHA, M. V. do V. Identificando hubs na rede marítima da cabotagem brasileira utilizando time-varying graphs. *Revista Mundi Engenharia, Tecnologia e Gestão (ISSN: 2525-4782)*, v. 3, n. 2, 2018. 6.7.1

SHANNON, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, v. 27, n. 4, p. 623–656, Oct 1948. ISSN 0005-8580. 1, 4

SILVA, B. et al. Statistical characterization of an ensemble of functional neural networks. *European Physical Journal B*, v. 392, p. 85–358, 2012. 2.3, 6.7.1

SOLÉ, R. V. et al. Selection, tinkering, and emergence in complex networks. *Complexity*, Wiley Online Library, v. 8, n. 1, p. 20–33, 2002. 4.4

SOLÉ, R. V.; VALVERDE, S. Information theory of complex networks: on evolution and architectural constraints. In: *Complex networks*. [S.l.]: Springer, 2004. p. 189–207. 1, 4.4

SOUSA, R. et al. Preferential interaction networks: A dynamic model for brain synchronization networks. *Physica A: Statistical Mechanics and its Applications*, Elsevier, p. 124259, 2020. 2.3

STERNBERG, R. J. *Psicologia cognitiva*. [S.l.]: Piccin, 2000. 3.2

STUMPF, I. R. C. et al. Scientific output indicators and collaboration in southern brazil. *Revista Interamericana de Bibliotecología*, Universidad de Antioquia, v. 40, n. 1, 2017. 1.2

TAKES, F. W.; KOSTERS, W. A. Computing the eccentricity distribution of large graphs. *Algorithms*, Multidisciplinary Digital Publishing Institute, v. 6, n. 1, p. 100–118, 2013. 6.7.1

TANG, J. et al. Small-world behavior in time-varying graphs. *Physical Review E*, APS, v. 81, n. 5, p. 055101, 2010. 2.3

TEIXEIRA, G. M. et al. Complex semantic networks. *International Journal of Modern Physics C*, World Scientific, v. 21, n. 03, p. 333–347, 2010. 1, 1.1, 3.2, 3.4, 3.4.1, 4, 5, 6.7, 6.13, 7.5

VANZ, S. A. de S.; STUMPF, I. R. C. Colaboração científica: revisão teórico conceitual. *Perspectivas em Ciência da Informação*, SciELO Brasil, v. 15, n. 2, p. 42–55, 2010. 1

VIOL, A. et al. Characterizing complex networks using entropy-degree diagrams: unveiling changes in functional brain connectivity induced by ayahuasca. *Entropy*, Multidisciplinary Digital Publishing Institute, v. 21, n. 2, p. 128, 2019. 1

VOLPATO, G. L.; FREITAS, E. G. d. Desafios na publicação científica. *Pesquisa Odontologica Brasileira*, scielo, v. 17, p. 49 – 56, 05 2003. ISSN 1517-7491. 1

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of'small-world'networks. *Nature*, v. 393, n. 6684, p. 409–10, 1998. 2.1, 8

WEAVER, W. Recent contributions to the mathematical theory of communication. *ETC: A Review of General Semantics*, JSTOR, p. 261–281, 1953. 4.3

ZIMAN, J. Comunidade e comunicação. *Conhecimento público. São Paulo: Itatiaia/EDUSP*, p. 115–130, 1979. 1

Entropia da informação em redes semânticas de títulos variáveis no tempo

Marcelo do Vale Cunha

Salvador, Outubro de 2020.