CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

CAMINHO PARA A COMPLEXIDADE: O PAPEL DA TOPOLOGIA DE REDE NA DIFUSÃO DE INFORMAÇÃO E DINÂMICA DE MERCADO

FERNANDO ANDRADE DUCHA

Orientador: Arthur Rodrigo Bosco de Magalhães
Centro Federal de Educação Tecnológica de Minas Gerais

Coorientador: Allbens Atman Picardi Faria
Centro Federal de Educação Tecnológica de Minas Gerais

BELO HORIZONTE
OUTUBRO DE 2021

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
COORDENAÇÃO DO PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

"CAMINHO PARA A COMPLEXIDADE: O PAPEL DA TOPOLOGIA DE REDE NA DIFUSÃO DE INFORMAÇÃO E DINÂMICA DE MERCADO".

Tese de Doutorado apresentada por **Fernando Andrade Ducha**, em 21 de outubro de 2021, ao Programa de Pós-Graduação em Modelagem Matemática e Computacional do CEFET-MG, e aprovada pela banca examinadora constituída pelos professores:

Prof. Dr. Arthur Rodrigo Bosco de Magalhães (Orientador)
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Allbens Atman Picardi Faria (Coorientador)
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Marcelo Martins de Oliveira
Universidade Federal de São João del-Rei

Prof. Dr. Leonardo Costa Ribeiro
Universidade Federal de Minas Gerais

Prof. Dr. José Luiz Acebal Fernandes
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Thiago Gomes de Mattos
Centro Federal de Educação Tecnológica de Minas Gerais

Visto e permitida a impressão,

Profª. Drª. Elizabeth Fialho Wanner
Presidenta do Colegiado do Programa de Pós-Graduação em Modelagem Matemática e Computacional

Ducha, Fernando Andrade
D826c Caminho para a complexidade: o papel da topologia de rede na difusão de informação e dinâmica de mercado / Fernando Andrade Ducha. – 2021.
79 f.

Tese de doutorado apresentada ao Programa de Pós-Graduação em Modelagem Matemática e Computacional.
Orientador: Arthur Rodrigo Bosco de Magalhães.
Coordenador: Allbens Atman Picardi Faria.
Tese (doutorado) – Centro Federal de Educação Tecnológica de Minas Gerais.

1. Mercado financeiro – Teses. 2. Análise de séries temporais – Teses. 3. Complexidade computacional – Teses. 4. Fractais – Programas de computador – Teses. 5. Engenharia financeira – Teses. 6. Teoria dos sistemas – Teses. 7. Distribuição (teoria da probabilidade) – Teses. 8. Lei de Zipf – Teses. I. Magalhães, Arthur Rodrigo Bosco de. II. Faria, Allbens Atman Picardi. III. Centro Federal de Educação Tecnológica de Minas Gerais. IV. Título.

CDD 530.13

Elaboração da ficha catalográfica pela bibliotecária Jane Marangon Duarte, CRB 6ª 1592 / Cefet/MG

Resumo

Um novo modelo, em que o fluxo de informação em uma rede social influencia o comportamento de agentes em um mercado, é apresentado. A difusão de informação na rede segue a do modelo baseado na plataforma de *microblogging* Twiter [*Scientific Reports*, Vol.**2**, pp.335], em que um mecanismo de memória permite a simulação de interesses endógenos dos agentes, bem como de atenção limitada. Considera-se que os agentes negociam um único ativo, ofertando ou demandando de acordo com os conteúdos de suas memórias. Tais conteúdos são modelados através de conjuntos de números com distribuição gaussiana, associados à percepção dos agentes quanto ao valor do ativo. Fluxo de informação e dinâmica de preços são analisados por meio de ferramentas estatísticas. Quatro modelos de redes são utilizados: Albert-Barabàsi (AB), Erdös-Rényi (ER), circular regular (CR) e um modelo, aqui desenvolvido, baseado na distribuição de Zipf (Z). Tais modelos são descritos e explorados através de experimentos computacionais. A Lei de Zipf também é investigada, utilizando-se dados da Wikipedia. No caso em que a rede AB é escolhida para o modelo de rede social, são geradas distribuições cumulativas de retornos que decaem como lei de potência, análogas a distribuições empíricas observadas há duas décadas. Quando as redes ER e CR são empregadas, tal decaimento aproxima-se do apresentado por uma gaussiana, compatível com observações recentes. Já para o modelo baseado na distribuição de Zipf, é possível obter, mediante escolha apropriada de parâmetros, cada um dos comportamentos. Correlações temporais não-lineares em séries financeiras frequentemente aparecem sob a forma do fenômeno da clusterização de volatilidade. Tais correlações podem ser associadas ao comportamento multifractal, que, nesta contribuição, é investigado por meio da Análise Multifractal de Flutuação sem Tendência. Nas séries de preços produzidas pelo modelo, observamos correlações mais fortes para a rede AB e para uma das redes Z construídas (a mesma que apresentou caudas pesadas na distribuição de retornos). Resultados análogos, referentes a distribuições e a comportamento multifractal, foram encontrados quando analisamos uma variável relacionada ao fluxo de informação. É digno de nota que, mesmo com entradas gaussianas, o modelo seja capaz de produzir as estatísticas descritas acima, típicas de mercados financeiros, mas que não são trivialmente encontradas em outros contextos. No presente caso, é possível concluir que a topologia da rede de distribuição de informação cumpre um papel fundamental nesse processo.

Palavras-chave: Mercado Financeiro. Expoente Hurst. Multifractalidade. MF-DFA. Redes Complexas. Distribuição de Retornos. Sistemas Complexos. Lei de Zipf.

Abstract

A new model, in which the information flow in a social network influences the behavior of agents in a market, is presented. The dissemination of information on the network follows the model based on the *microblogging* Twiter platform [*Scientific Reports*, Vol.**2**, pp.335], in which a memory mechanism allows the simulation of the agents' endogenous interests, as well as of limited attention. It is considered that agents negotiate a single asset, supplying or demanding according to the contents of their memories. Such contents are modeled through sets of numbers with Gaussian distribution, associated with the agents' perception of the asset's value. Information flow and price dynamics are analyzed using statistical tools. Four network models are used: Albert-Barabàsi (AB), Erdös-Rényi (ER), regular circular (CR) and a model, developed here, based on the Zipf (Z) distribution. Such models are described and explored through computational experiments. Zipf's Law is also investigated, using data from Wikipedia. In the case where the AB network is chosen for the social network model, cumulative return distributions that decay as a power law are generated, analogous to empirical distributions observed two decades ago. When the ER and CR networks are used, such decay is close to that presented by a Gaussian, compatible with recent observations. For the model based on the Zipf distribution, it is possible to obtain, by appropriate choice of parameters, each of the behaviors. Nonlinear temporal correlations in financial series often produce the phenomenon of volatility clustering. Such correlations can be associated with the multifractal behavior, which, in this contribution, is investigated through the *Multifractal Detrended Fluctuation Analysis*. In the price series produced by the model, we observe stronger correlations for the AB network and for one of the constructed Z networks (the same one that presented heavy tails in the distribution of returns). Similar results, referring to distributions and multifractal behavior, were found when we analyzed a variable related to the information flow. It is noteworthy that, even with Gaussian inputs, the model can produce the statistics described above, typical of financial markets but not trivially found in other contexts. In the present case, it is possible to conclude that the topology of the information distribution network plays a fundamental role in this process.

Keywords: Financial Market. Hurst Exponent. Multifractality. MF-DFA. Complex Networks. Return Distributions. Complex Systems. Zipf Law.

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Sumário

1 Introdução

Sistemas complexos são sistemas que, embora formados por várias entidades interconectadas sem a presença de um controlador central, interagem com o mundo externo apresentando, globalmente, comportamentos que não podem ser antecipados a partir do comportamento individual de suas partes. Tais comportamentos são denominados comportamentos emergentes e o estudo desses fenômenos, estudo da complexidade. Sistemas complexos são comumente observados na natureza (MITCHELL, 2009), bem como em sistemas desenvolvidos pelo ser humano. Já na Grécia Antiga, Aristóteles declarava, em sua obra *Metafísica* (ARISTOTLE, 350BCE): *O todo é maior que a simples soma das partes*. No entanto, a forma como os percebemos amadureceu somente no século XX, quando novas ferramentas computacionais mudaram radicalmente nossa habilidade de criar modelos, muitos deles interdisciplinares, englobando áreas da Física, Biologia, Economia e Sociologia. Um exemplo de sistema complexo é o sistema atmosférico da Terra, em que uma multiplicidade de elementos estão conectados levando à emergência dos fenômenos climáticos que observamos. Recentemente, os efeitos que emergem do aquecimento global, tais como enchentes em locais usualmente desérticos na Austrália (JOHNSON et al., 2016) ou secas em regiões onde a água tradicionalmente é abundante (vide o sistema hídrico brasileiro (MONTENEGRO; RAGAB, 2010)), têm ganhado importância, dadas suas consequências dramáticas para muitas populações.

Frequentemente, modelos de sistemas complexos envolvem Teoria de Redes na análise dos comportamentos da natureza e dos seres humanos. Uma rede é formada por nós e pelas arestas que os conectam. A distância entre dois nós é o número mínimo de arestas que devem ser percorridas para ir de um nó ao outro. Um exemplo nesse contexto é o efeito conhecido como mundo pequeno (*small world effect*): em variadas situações, a maior distância entre dois nós da rede será seis (MILGRAM, 1967; TRAVERS; MILGRAM, 1969). Isto já foi detectado em muitos contextos (WATTS; STROGATZ, 1998; BARABÁSI; ALBERT, 1999; LILJEROS et al., 2001). Recentemente, a Internet tem atraído a atenção de pesquisadores de sistemas complexos: o efeito de mundo pequeno tem sido observado na web (BRODER et al., 2000), redes sociais (LILJEROS et al., 2001) e até mesmo na infraestrutura da Internet (VINCIGUERRA; FRENKEN; VALENTE, 2010). Um exemplo de efeito de mundo pequeno em relações humanas não digitais é encontrado em um experimento de Milgran (TRAVERS; MILGRAM, 1969); nele, cartas foram enviadas para várias pessoas pedindo-se que estas colaborassem para que a mesma chegasse a um determinado destinatário em outra cidade, respeitando-se a seguinte regra: as cartas só poderiam ser enviadas para conhecidos. A análise de redes reais é comumente realizada por meio da comparação com redes sintéticas. Exemplos importantes são o modelo de Price e o modelo de Albert-Barabási (NEWMAN, 2018). Ambos produzem redes de mundo

pequeno que apresentam outra importante propriedade: uma distribuição de graus de nodos livre de escala (o grau de um nodo é o número de suas conexões com outros nodos da rede).

O Mercado Financeiro é, com frequência, visto como um sistema complexo (KWAPIEŃ; DROŻDŻ, 2012). A difusão de informação, diante de seu papel na tomada de decisão e na formação de preços, é de grande interesse nesse contexto. A pesquisa nessa área é incentivada não só pela imensa quantidade de dados de origem financeira existentes, mas também pelas investigações quantitativas da estrutura e da dinâmica de redes sociais digitais (KUMAR; NOVAK; TOMKINS, 2010). Modelos baseados em agentes (MBA) têm sido desenvolvidos com a intenção de estudar diferentes aspectos das dinâmicas de mercado (LUX; MARCHESI, 1999; LEBARON; ARTHUR; PALMER, 1999; LEBARON, 2000; CHIARELLA et al., 2003; WEI et al., 2003; LEBARON, 2006; FAN et al., 2009; BAKKER et al., 2010; ATMAN; GONÇALVES, 2012; FENG et al., 2012; WEI; HUANG; HUI, 2013; REKIK; HACHICHA; BOUJELBENE, 2014; STEFAN; ATMAN, 2015; KRAUSE; BÖRRIES; BORNHOLDT, 2015; LIMA, 2017; XAVIER; ATMAN; MAGALHÃES, 2017). Tais modelos são usualmente empregados para tratar problemas com um grande número de variáveis, difíceis de abordar analiticamente. É comum que esses problemas sejam discretos e não-lineares, características frequentes no contexto de Economia e Finanças. Os MBA constituem um meio de aproximar e entender sistemas de interesse utilizando um ou mais tipos de elementos, os agentes, que têm bem definidas suas respostas às diversas entradas possíveis. Tais definições podem depender do tempo. A partir do comportamento interno de cada agente, geram-se relações entre os mesmos. O amálgama entre comportamento dos agentes e comportamento entre agentes cria um ferramental capaz de retornar a quem está empregando o modelo respostas que vão desde o macro, ou seja, o que o conjunto de agentes pode produzir, até o micro, onde, por exemplo, podemos buscar saber se um agente exerce papel preponderante no sistema. Entradas aleatórias podem ser combinadas com dinâmicas endógenas (dinâmicas definidas pelos mecanismos internos de cada agente que constitui o modelo), visando o entendimento do processo pelo qual influências imprevisíveis do mundo externo são capazes de levar o mercado à emergência de comportamentos complexos. Limitação de recursos e tendências comportamentais são ingredientes que modelam tais dinâmicas endógenas. O surgimento de curvas produzidas sinteticamente a partir de um modelo baseado em agentes, curvas estas associadas a estatísticas da mesma natureza das observadas em dados financeiros reais, é uma indicação de que o modelo pode ser útil para que avancemos na compreensão dos mercados reais. Tais estatísticas são, muitas vezes, não triviais, envolvendo, por exemplo, distribuições com caudas pesadas, comportamento fractal e a presença de correlações não lineares, podendo, em certos casos, ser associadas às propriedades emergentes do sistema.

Neste trabalho, apresentamos um novo modelo de mercado, onde a informação sobre o mundo externo é compartilhada em uma rede de agentes. A difusão de informação

segue o modelo introduzido em Weng et al. (2012), que é baseado em observação empírica da plataforma de *microblogging* Twitter; dois importantes pontos a se destacar neste modelo são o mecanismo de memória, que permite aos agentes desenvolverem interesses endógenos, e o conceito de atenção limitada. No presente trabalho, é suposto que os agentes negociam um único ativo no mercado, ofertando e demandando de acordo com o conteúdo de suas memórias. Estes conteúdos correspondem a conjuntos de números reais obtidos aleatoriamente de acordo com uma distribuição gaussiana, escolhida por estar entre as mais utilizadas em modelos estatísticos para variáveis aleatórias contínuas (MAGALHÃES; LIMA, 2002). Os conteúdos nas memórias podem ser entendidos como informações que os agentes possuem: informações que induzem o agente a demandar o ativo correspondem a números positivos e informações que o induzem a ofertar, a números negativos; o módulo do número modela a intensidade da influência da informação sobre o agente. O fluxo de informação, calculado a partir do número médio de mensagens postadas por cada agente, bem como a dinâmica de preço, são estatisticamente analisados. Quatro modelos de redes são utilizados: Albert-Barabàsi (AB) (BARABÁSI; ALBERT, 1999), Erdös-Rényi (ER) (ERDÖS; RÉNYI, 1959), circular regular (CR) (NEWMAN, 2018) e um modelo aqui desenvolvido a partir da distribuição de Zipf. Para AB, distribuições cumulativas de retorno decaem como uma lei de potência, analogamente a resultados empíricos de duas décadas atrás (GOPIKRISHNAN et al., 1999; KWAPIEN et al., 2003). Para ER e CR, o decaimento é próximo ao de uma gaussiana, o que é compatível com algumas observações recentes (KWAPIEN et al., 2003; DROŻDŻ et al., 2007; BOTTA et al., 2015). Para o modelo baseado na distribuição de Zipf, conseguimos obter, mediante a escolha apropriada dos parâmetros que determinam sua construção, cada um dos comportamentos, ou seja, presença ou ausência de caudas pesadas. Correlações não-lineares nas séries de preços, que são essenciais para a clusterização de volatilidade, podem ser associadas a multifractalidade (MANDELBROT, 1997; ASHKENAZY et al., 2003; KALISKY; ASHKENAZY; HAVLIN, 2005; KALISKY; ASHKENAZY; HAVLIN, 2007; KWAPIEŃ; DROŻDŻ, 2012). De fato, o dimensionamento multifractal (*multifractal scaling*) tem sido observado em dados financeiros (CALVET; FISHER, 2002; GU; ZHOU et al., 2010; KWAPIEŃ; DROŻDŻ, 2012; GREEN; HANAN; HEFFERNAN, 2014; BUONOCORE; ASTE; MATTEO, 2016). Para detectar tal comportamento neste trabalho, empregamos a Análise Multifractal de Flutuação sem Tendência (*Multifractal Detrended Fluctuation Analysis* - MFD-FA) (KANTELHARDT et al., 2002) em séries de preços sintéticas e encontramos correlações mais fortes para as redes AB e para uma das redes baseadas na distribuição de Zipf. Resultados análogos foram explicitados para uma medida relacionada ao fluxo de informação. É digno de nota que, mesmo com entradas gaussianas, o modelo seja capaz de produzir respostas com estatísticas não triviais. Para o caso aqui descrito, foi possível concluir que a topologia da rede de distribuição de informação possui um papel fundamental na emergência da dinâmica.

Esta tese está organizada na forma descrita a seguir. No Capítulo 2, alguns conceitos gerais sobre redes são revistos. Nele, definimos o que é uma rede e abordamos medidas relacionadas ao assunto. Um tratamento das redes circular regular e Erdos-Rényi é efetuado, bem como dos modelos de Price e Albert-Barabàsi. Sobre o modelo de Albert-Barabàsi, realizamos um experimento que evidencia a clusterização associada a esse tipo de rede. No Capítulo 3, abordamos a Lei de Zipf e a distribuição de Zipf. Nesse capítulo, descrevemos o modelo de rede baseado na distribuição de Zipf, que desenvolvemos. O MFDFA é descrito no Capítulo 4, onde sua aplicação é exemplificada empregando-se séries de preços simuladas. Os capítulos centrais desta tese são o Capítulo 5 e o Capítulo 6. Seus resultados estão publicados em Ducha, Atman e Magalhães (2021). O Capítulo 5 descreve o modelo de mercado e o modelo de rede social, citados acima. Já o Capítulo 6 traz a análise de correlações não lineares e de distribuições relacionadas às séries temporais concernentes às dinâmicas de informação e de mercado produzidas pelo modelo. No Capítulo 7, apresentamos as conclusões e possíveis trabalhos futuros. Esta tese conta com dois apêndices. No Apêndice A, abordamos métodos de construção de redes. Tratamos, inicialmente, o caso da rede circular regular. A rede Erdos Renyi é explorada em duas versões: o modelo $G(n, M)$ e o modelo $G(p, n)$. Esse apêndice também acomoda uma seção relativa à geração de grafos com distribuição de graus em lei de potência usando o algoritmo de adição de nós, de Albert-Barabàsi. É descrito, também, o modelo de configuração, que pode ser usado para gerar grafos com distribuição de graus definida por uma distribuição de probabilidades parametrizada. No Apêndice B, desenvolvemos um experimento utilizando dados da Wikipedia analisados segundo a Lei de Zipf, empregando ferramentas de tratamento natural de linguagens.

2 Redes

Este capítulo tem como objetivo descrever um breve conteúdo teórico sobre redes. Com isso, explicaremos as diferenças teóricas entre grafos e redes de agentes. Aproveitaremos para introduzir as métricas mais comuns em redes e algumas cujo resultado exigem um pouco mais do contexto em que as redes de agentes estão inseridas. Ainda nesta seção, serão apresentados os modelos de rede utilizados neste trabalho. Cada um deles possui suas peculiaridades topológicas; eles podem ser analizados através de medidas de centralidade (*betweeness*, *closeness*, clusterização) que, por sua vez, também apresentam suas distinções. Muitas vezes vamos nos referir aos agentes da rede como nós, isso acontecerá quando estivermos falando somente sobre a teoria de redes, ou quando não precisarmos descrever ações dos agentes. Deve ficar claro que a expressão *agente* i é equivalente à expressão *nó* i. Os modelos a serem descritos são:

- modelo de rede Circular-Regular (CR);
- modelo de rede de Erdös-Réniy (ER);
- modelo de rede Albert-Barabàsi (AB).

2.1 Grafos e Redes

Tanto redes de uma só camada como grafos direcionados podem ser definidos a partir da reunião de dois conjuntos N e E, constituindo a tupla $G(N, E)$. Neste contexto, $N = \{a_1, a_2, \ldots, a_{n-1}, a_n\}$ é o conjunto de nós (também chamados de nodos ou vértices); já E é o conjunto de arestas, formado por pares ordenados $e_{x,y}$, onde x representa o nó de origem (ou saída) da aresta, e y, o de destino (ou entrada). O que identificamos como um conjunto de nós no caso de grafos será identificado como um conjunto de agentes quando estivermos considerando redes. No caso específico sendo analisado, não serão permitidas arestas duplicadas, ou seja, duas arestas caracterizadas pelo mesmo par ordenado. Também não serão permitidos o que conhecemos como *self-loops*, ou seja, arestas em que a origem seja igual ao destino. Podemos definir o grau de entrada e o grau de saída de cada nó de uma rede: o primeiro é o número de arestas que chegam ao nó; o segundo, o número das que saem do mesmo.

Um grafo conexo é aquele em que é possível ir de qualquer um de seus nós a qualquer outro por um caminho composto por arestas presentes no grafo. Neste caso, o tamanho do grafo será definido como a soma da cardinalidade do conjunto N com a do conjunto E. Um grafo em que o conjunto de nós esteja contido em N e o de arestas em E é um subgrafo de $G(N, E)$. Um grafo que não seja conexo pode ser particionado em subgrafos conexos, chamados de componentes conexas. Neste caso, o tamanho do grafo é o tamanho de sua maior componente conexa. No grafo direcionado, cada nó i tem dois

tipos de vizinhança: a vizinhança de saída, composta pelas arestas que têm como origem o nó i , out_i, e a vizinhança de entrada, formada pelas arestas que têm como destino o nó i, in_i. Observemos que é possível que a conectividade de um grafo dependa do tipo de vizinhança que se analisa. Assim, podemos encontrar um grafo conexo quando analisado pelas vizinhanças de saída e desconexo quando as vizinhanças de entrada são utilizadas. Neste caso, teremos um tamanho do grafo para cada tipo de vizinhança.

Um grafo pode ser simples (não-direcionado): neste caso, os conceitos de origem e destino de uma aresta perdem o sentido, e uma aresta definida por $e_{x,y}$ coincide com a definida por $e_{y,x}$. Assim, há apenas um tipo de vizinhança, e o grau de um nó é simplesmente o número total de nós com o qual ele está diretamente conectado por meio de arestas. A conectividade do grafo, bem como seu tamanho, são estudados através do único tipo de vizinhanca que existe.

Neste trabalho, somente utilizaremos grafos direcionados.

2.1.1 Caracterização de Grafos

A estrutura de um grafo influencia a dinâmica de modelos envolvendo redes de agentes que utilizem tal grafo. Comecemos apresentando três ferramentas simples relacionadas às vizinhanças de saída, que podem ser empregadas na análise de tal estrutura:

- média de out_i: $\langle out_i \rangle = \frac{1}{|N|} \sum_{i=1}^{|N|} out_i$, onde $|N|$ é a cardinalidade do conjunto N (número total de nós);
- densidade de arestas: razão $\frac{|E|}{|N|(|N|-1)}$, onde $|E|$ é a cardinalidade do conjunto E (quantidade total de arestas) e $|N|(|N|-1)$ é o número máximo de arestas que o grafo pode alcançar dada a quantidade de nós $|N|$;
- distribuição de graus advinda de out_i para o conjunto de nós N.

Deve-se deixar claro que estas mensurações são definidas de forma idêntica para as vizinhanças de entrada, podendo apresentar comportamentos distintos para uma mesma rede.

As medidas aqui apresentadas para grafos são tais que representam simplesmente um coeficiente. Porém, quando lidamos com redes de agentes, onde existem mais variáveis para serem levadas em consideração, o entendimento desses coeficientes começa a ganhar interpretação dentro do contexto em que a rede foi modelada. Desta forma, buscam-se relações entre os valores calculados e a dinâmica do modelo, obtendo-se assim uma informação interpretável, e não um simples número desconectado das aplicações.

Descreveremos a seguir três medidas que possuem mais afinidade com redes, pois são frequentemente utilizadas para nos auxiliar a tomar decisões sobre os modelos que utilizamos. São elas:

1. ***Betweenness*:** O conceito de *betweenness* de um nó i é uma medida centralidade que se relaciona com o número total de caminhos mínimos que passam por i. O caminho mínimo do nodo j ao nodo k é um conjunto composto pelo menor número

de arestas que devem ser percorridas para ir de j a k. Dados dois nodos, pode existir mais de um conjunto mínimo de arestas que os une; sendo assim, pode existir mais de um caminho mínimo (evidentemente, todos com a mesma quantidade de arestas). Caminhos mínimos são usualmente calculados através do algoritmo de Dijkstra (NEWMAN, 2018). O cálculo do *Betweenness* para um determinado nodo i é feito da seguinte forma: para cada par de nós j e k diferentes de i, calculam-se todos os caminhos mínimos entre os mesmos; computa-se a quantidade $d_{j,k}(i)$ de tais caminhos mínimos que passam por i; é feita uma normalização dividindo-se $d_{j,k}(i)$ por $d_{j,k}$, a quantidade total de caminhos mínimos entre j e k; os resultados de tal divisão são somados para todos os pares (j, k), com $j \neq i \neq k$. Sendo assim, podemos formular a seguinte equação para o *Betweenness* do nó i:

$$CB_i = \sum_{j \neq i \neq k} \frac{d_{j,k}(i)}{d_{j,k}}. \tag{1}$$

2. ***Closeness*:** A distância $dist(i, j)$ entre dois nós i e j em uma rede é o número de arestas em um caminho mínimo entre eles. A medida de centralidade denominada *closeness* nada mais é do que a soma dos inversos das distâncias de um nó i em relação a todos os outros nós, sendo usualmente normalizada pela quantidade de nós da rede ($|N|$). A definimos, então, como:

$$CC_i = \sum_{j \neq i} \frac{|N|}{dist(i, j)}, \tag{2}$$

onde i é o nó do qual desejamos calcular o valor de *closeness* e o somatório é efetuado para todos os outros nós da rede. Note que CB e CC, apesar de serem métricas diferentes, podem ser conectadas à mensuração de fluxo de informação. Ao calcular o valor de CB para cada nó, ou seja, quantos mínimos caminhos passam por um nó, estamos de forma indireta mensurando qual o papel daquele nó dentro do processo do fluxo de informação, quando levamos em consideração o CB de nós vizinhos ou de nós com valores altos de CB. Em contrapartida, ao calcular o valor de CC de um nó, temos uma indicação de sua importância no fluxo de informação a partir da noção, que esta medida dá, da proximidade do nó em relação ao restante da rede. Deve-se levar em consideração que tais mensurações de CB e de CC não são necessariamente relacionadas com os graus dos nós na rede.
3. **Clusterização local:** É um cálculo feito nó a nó, onde é medido o nível de conectividade da vizinhança de um nó i. Essa mensuração pode ser obtida a partir do descrito a seguir. Consideremos $\mathcal{N}_i$ o conjunto de nós da vizinhança do nó i (observemos que o nó i não pertence a este conjunto). Se $|\mathcal{N}_i|$ é a cardinalidade de $\mathcal{N}_i$, o número máximo de arestas que podem ser definidas entre seus elementos é $|\mathcal{N}_i|(|\mathcal{N}_i| - 1)$.

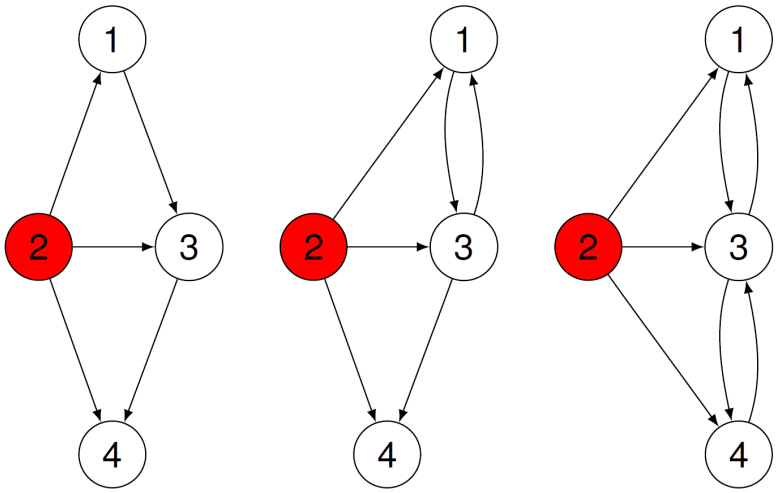


Figura 1 – Nesta figura, mostramos, através da adição de arestas, como o valor da clusteriazação local do nó 2, CL_2, se modifica. No grafo à esquerda, temos duas arestas, $(1,3)$ e $(3,4)$; o número total de arestas possível com este número de vizinhos é 6; logo, $CL_2 = \frac{2}{6} = \frac{1}{3}$. No grafo central, as arestas da vizinhança são $(1,3)$, $(3,1)$ e $(4,3)$, o que nos leva em $CL_2 = \frac{3}{6} = \frac{1}{2}$. No grafo à direita, temos as arestas $(1,3)$, $(3,1)$, $(3,4)$ e $(4,3)$; logo, $CL = \frac{4}{6} = \frac{2}{3}$.

Tomemos $\mathcal{E}_i$ como o conjunto de arestas ligando os nós de $\mathcal{N}_i$ no grafo sob análise. A cardinalidade de $\mathcal{E}_i$, $|\mathcal{E}_i|$, dá o número de arestas presentes, ligando os nós da vizinhança do nó i. A razão entre o número de arestas existentes entre os elementos de $\mathcal{N}_i$ e o número máximo das que poderiam haver define a clusterização local:

$$CL_i = \frac{|\mathcal{E}_i|}{|\mathcal{N}_i|(|\mathcal{N}_i| - 1)}. \tag{3}$$

Um exemplo de cálculo de clusterização local é dado na Figura 1.

2.2 O Modelo Circular Regular

O modelo de rede circular regular (CR) utilizado neste trabalho é equivalente ao modelo de rede circular encontrado no trabalho de Watts e Strogatz (1998). Este modelo foi escolhido devido à grande simetria entre os nós. Tanto o grau de entrada quanto o de saída de cada nó é o mesmo. Um exemplo de grafo direcionado em que temos 10 nós, cada um deles com grau de entrada e grau de saída iguais a 3 é representado na Figura 2. Embora possamos indexar os nós associando-os a inteiros de 1 a 10, de fato eles não apresentam diferença alguma, e qualquer característica local que seja calculada não apresentará variação com o nó escolhido.

Utilizaremos o modelo de rede circular regular entre os que definirão nosso modelo de difusão de postagens (ver Capítulo 5) com o objetivo de investigar o reflexo da simetria entre os agentes na dinâmica de informação e de preços. ´

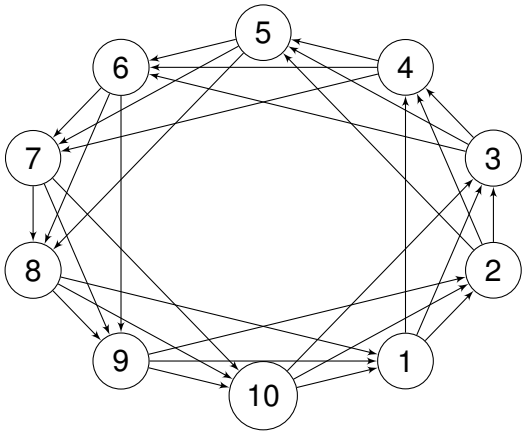


Figura 2 – Representação de uma rede circular regular. A rede possui 10 nós, cada um com 3 arestas entrando e 3 aresta saindo. Embora associemos índices que vão de 1 a 10 a cada nó, suas propriedades são idênticas.

2.3 O Modelo Erdös-Rényi

O modelo Erdös-Rényi (ER), também conhecido como Rede de Poisson, é um dos primeiros modelos de redes aleatórias propostos na literatura (ERDÖS; RÉNYI, 1959). Nos modelos de Erdös-Rényi, desenvolvidos por dois matemáticos húngaros, Paul Erdős and Alfréd Rényi, o objetivo é criar modelos que se comportem de forma aleatória. Dado o tamanho da rede, por meio de um único outro parâmetro busca-se gerar uma rede aleatória cuja construção é realizada através de um processo de Monte Carlo. Tal processo determina em quais nós serão adicionadas novas arestas. A ideia central é que os sorteios sejam sempre efetuados de maneira uniforme no que diz respeito aos diferentes nós, de forma que, durante o processo de construção da rede, não haja mecanismos que gerem tendências que diferenciem as probabilidades de cada nó de receber uma nova aresta.

Erdös e Rényi criaram dois modelos que produzem resultados idênticos no que se refere à distribuição de graus dos nós, mas que se diferenciam na forma de interpretar e de implementar a criação dos grafos. Estes modelos são chamados de $G(n, M)$ e de $G(n, p)$.

O primeiro modelo que abordamos é o $G(n, M)$. Aqui, n é o número de nós e M, o número de arestas que o grafo irá conter. Para empregar este modelo, primeiro gera-se o conjunto de todos os grafos possíveis com n elementos e M arestas, e depois escolhe-se, de forma aleatória com distribuição uniforme, um grafo no conjunto. Observemos que o

número de grafos a serem produzidos nesse processo forma é

$$\binom{\binom{n}{2}}{m}, \tag{4}$$

o que leva a valores enormes mesmo com parâmetros relativamente pequenos. Este modelo é uma alternativa que generaliza o grafo de Poisson: ele dá a possibilidade de gerar a rede sem que a média de graus do grafo que se deseja criar seja o dado central. Por outro lado, essa média pode ser simplesmente obtida por $c = \frac{M}{n}$.

O segundo modelo é o mais comumente utilizado, dentre os dois modelos ER. É conhecido como modelo $G(n, p)$. Nele, n continua sendo o número de nós e p, a probabilidade que cada uma das arestas possíveis tem de existir. Assim, para cada uma das arestas conectando dois nodos do grafo, um processo aleatório é empregado, decidindo que a aresta presente com probabilidade p, e ausente com probabilidade $p - 1$. Desta concepção, temos o resultado direto de que a média de graus do modelo é $c = \binom{n}{2}p$. Apesar de possuir complexidade $O(n^2)$, este modelo é relativamente fácil de aplicar, tendo em vista custo computacional. Ele possui propriedades interessantes (BOLLOBÁS, 2001; ERDÖS; RÉNYI, 1959):

- a rede resultante possuirá distribuição de Poisson de graus;
- se $pn = 1$, a rede será desconexa e a componente principal terá ordem de $n^{\frac{2}{3}}$;
- se $pn = u$, com $u > 1$, onde u é uma constante, o grafo $G(n, p)$ terá a componente principal com o tamanho de uma fração positiva dos vértices e nenhuma outra componente será maior que $O(\log(n))$ vértices;
- se $p < \frac{(1-\epsilon)\ln n}{n}$, onde ϵ é um número real positivo, a rede provavelmente será desconexa (possuirá vértices isolados);
- se $p > \frac{(1+\epsilon)\ln n}{n}$, então provavelmente a rede será conexa.

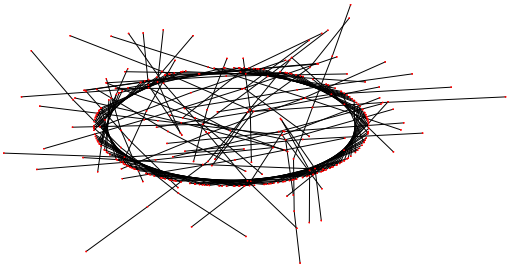
Estas propriedades servem como ferramentas quando o objetivo é lidar com o modelo ER, porque proveem um comportamento esperado para alguns casos do modelo, como, por exemplo, o fato de que o modelo possui um limiar de conectividade, que é $\frac{\ln n}{n}$. Exemplos de como variações no parâmetro p podem afetar a estrutura do grafo ER são apresentados na Figura 3.

A rede de Erdös-Réniy possui distribuição de graus igual a uma distribuição de Poisson:

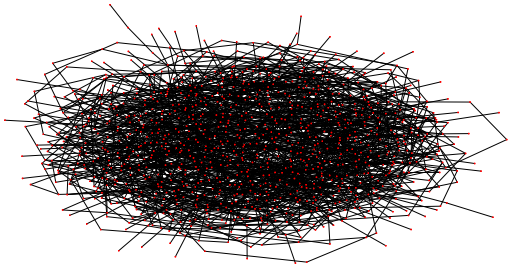
$$f(k; \theta) = Pr(X = k) = \frac{\theta^k e^{-\theta}}{k!}, \tag{5}$$

onde o desvio padrão e a média são iguais a θ, e k é o grau cuja probabilidade é dada. Assim, o modelo ER permite criar grafos conexos quando $p > \frac{\ln n}{n}$, com boa previsibilidade do nó de maior grau e do nó de menor grau.

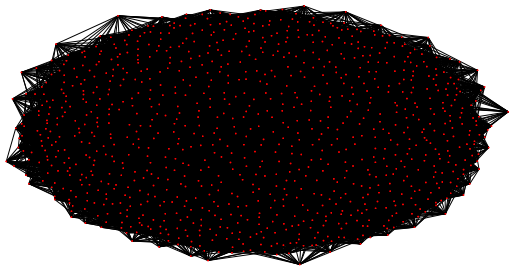
A importância da rede de ER encontra-se na possibilidade de controle da rede resultante, dadas as entradas, tornando-se assim um grafo muito utilizado como base de



(a) $p = 0.0002$ e $n = 1000$



(b) $p = 0.002$ e $n = 1000$



(c) $p = 0.02$ e $n = 1000$

Figura 3 – Exemplos de estados de conectividade, dado o limiar $\frac{\ln n}{n}$. Note que, na Figura 3a, $0.0002 < \frac{\ln 1000}{1000}$, enquanto nas outras figuras isso não acontece, podendo-se perceber o aumento de densidade de arestas nas redes das figuras 3b e 3c.

experimentos (WENG et al., 2012; STEFAN; ATMAN, 2015). Ela é uma rede aleatória não-complexa. É uma das redes empregadas no modelo descrito no Capítulo 5 com o objetivo de estudar como o tipo de rede influencia o aparecimento dos fatos estilizados descritos no Capítulo 6.

2.4 O Modelo Albert-Barabási

O modelo Albert-Barabási (AB) tem como um de seus resultados analíticos a possibilidade de produzir distribuição de graus decaindo em lei de potência com expoente $\lambda = -3$ (BARABÁSI; ALBERT, 1999). Este expoente tornou-se uma particularidade que se apresenta também em outras áreas do conhecimento. Diante disto, encontram-se estudos que demonstraram a importância do modelo (NEWMAN; BARABASI; WATTS, 2011). O modelo AB advém do trabalho pioneiro de Derek J. de Solla Price, criador do modelo de Price (*Price Model*) (PRICE, 1976), que teve como base os conceitos do trabalho de Simon (1955). A equação mestra, que veio a ser chamada de *cumulative advantage*, foi elaborada. Na Subseção 2.4.1, será feita uma breve descrição do que é a *cumulative advantage* (renomeada para *preferential attachtment* em Barabási e Albert (1999); aqui, utilizaremos o equivalente em Português, *anexação preferencial*). Na Subseção 2.4.2, efetuaremos uma apresentação do modelo de Price e continuaremos tratando o modelo AB, explicando as

diferenças.

2.4.1 Anexação Preferencial (*Preferential Attachtment*)

Anexação preferencial é um processo, que acontece em sistemas formados por múltiplos agentes, em que muitos destes buscam atrelar-se a alguns agentes com características que os tornam especiais, os quais denominaremos aqui *agente com influência*. Enquanto isso acontece, uma grande maioria de agentes cumpre papel pouco importante na dinâmica. Anexação preferencial é um processo que permeia um grande número de fenômenos: desde a infra-estrutura gerada pelos seres humanos até a natureza em si. Alguns exemplos são dados a seguir. Certos mamíferos têm necessidade de andar em grupos que tendem a crescer, porém sua população é reduzida por seus predadores. O mesmo serve para cardumes de peixes, onde a unidade do grupo pode permitir que o grupo inteiro tenha informação sobre algo que está por acontecer (SHINCHI et al., 2002). A busca dos seres humanos por trabalho causa, até hoje, fortes movimentos de migração para as grandes cidades industrializadas. O aumento da atividade comercial devido a essa migração recorrente tem, como resultado, a intensificação populacional nos grandes centros urbanos. Isto proporciona uma pulverização populacional em direção às cidades vizinhas, terminando em um movimento de urbanização de cidades consideradas rurais (BUHAUG; URDAL, 2013). Note que o descrito aqui são situações do cotidiano do mundo como um todo. Além disso, não se pode esquecer do papel fundamental desse fenômeno em outros nichos da sociedade humana, como, por exemplo: redes sociais (LILJEROS et al., 2001); a *World Wide Web* (WWW) (BRODER et al., 2000); as rotas de tráfego aéreo (DUNN; WILKINSON, 2016); e o próprio efeito gargalo conhecido como *clogging effect*, causado pelo excesso de carros nas grandes avenidas (NAGATANI, 2002).

Sob a perspectiva financeira, por exemplo, ocorre notável disparidade na concentração de riqueza, muito estudada. Esse desequilíbrio pode ser mais comumente observado em países sub-desenvolvidos ou em países em desenvolvimento, apesar de estar presente também em alguns países de primeiro mundo (CAJUEIRO; TABAK, 2004).

A equação mestra da anexação preferencial e todo o seu desenvolvimento pode ser encontrado em Price (1976) e Newman (2018). O processo em questão pode ser realizado computacionalmente através de um processo de Monte Carlo onde exista um viés a cada sorteio, diante de um determinado critério. Este critério é o elemento que associa um agente ao outro gerando laços entre os dois baseados em alguma característica. Normalmente, em redes utiliza-se o grau do elemento em questão para criar esse viés. Por exemplo, num processo de adição de arestas a um grafo sendo construído por meio da anexação preferencial, podemos definir que a probabilidade de que dois nodos sejam conectados por nova aresta, em um passo específico, seja função dos graus dos nodos. Ressaltemos que o ponto central que define a anexação preferencial na construção de grafos é a presença de um viés no processo de inclusão de novas arestas (isto é uma diferença

fundamental em relação ao modelo Erdös-Rényi). Sendo assim, outras possibilidades podem ser concebidas, para além da consideração dos graus dos nodos. De fato, dado um grafo $G(N, E)$ e considerando uma função genérica do nó i, f_i, que seja válida para todos os nós do conjunto N, pode-se aplicar uma metodologia para gerar a anexação preferencial de acordo com a função f_i, chamada metodologia por adição de nós (veja Subseção A.3.1 para mais detalhes). É importante notar que esta metodologia é costumeiramente utilizada somente com centralidade de graus; a generalização para utilização da mesma usando uma função f_i é uma proposta deste trabalho. O Algoritmo 5, no Apêndice A, descreve esta proposta.

2.4.2 O Modelo de Price

O modelo de Price é definido pela anexação preferencial e originalmente a utiliza para investigar dados empíricos relacionados a citações de obras publicadas (PRICE, 1976; NEWMAN, 2018). Este modelo baseia-se em uma rede direcionada, onde a existência de aresta ocorre quando uma obra A cita uma B; note que a função de citar uma obra é unidirecional por natureza, pois as citações feitas por A são de obras que na linha temporal são anteriores a A, portanto não poderiam citá-la. Assim, quando A cita B, criamos a aresta (A, B), uma vez que a aresta (B, A) não condiz com a lógica imposta pela rede de citações. Se organizamos a rede de Price em termos cronológicos, associando índices crescentes no sentido das mais novas às mais antigas das publicações, poderemos observar que todas as arestas partem dos nós de índices superiores para os de índices inferiores.

A aplicação do modelo de Price leva a uma função de probabilidade que determina, para cada nó, a chance de o mesmo ter o valor de grau q. Para o modelo de rede de Price, a distribuição é descrita por

$$p_q = \sum_{i=1}^{q} \frac{B(i + a, 2 + a/c)}{B(a, 1 + a/c)}, \tag{6}$$

onde:

- p_q é a probabilidade do grau q;
- $B(x, y) = \int_0^1 t^{x-1}(1 - t)^{y-1}dt$ é a função Beta de Euler;
- c é a média dos graus dos nodos do grafo desejada;
- a é uma constante positiva maior que zero.

A Figura 4 apresenta distribuições acumuladas inversas para os graus de redes de Price obtidas através de simulação. Dada uma distribuição acumulada inversa que apresente cauda com decaimento em lei de potência, o coeficiente angular que descreve tal decaimento é uma unidade maior que o coeficiente que descreve o decaimento para a distribuição canônica correspondente (em módulo, uma unidade menor). Para uma rede de Price, o valor do módulo do coeficiente angular relativo ao decaimento da cauda da

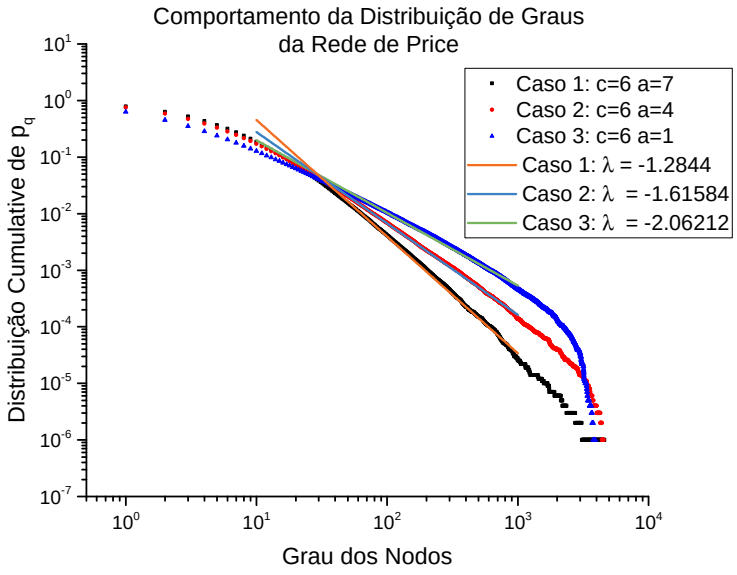


Figura 4 – Distribuição cumulativa inversa referente às probabilidades p_q de que um nodo tenha grau q. É exemplificado como um único parâmetro no modelo de Price pode afetar a morfologia da rede através de sua distribuição de graus. Cada curva é o resultado da média da 100 amostras.

distribuição (canônica) de graus é dado por (NEWMAN, 2018, pp.494)

$$\lambda = 2 + \frac{a}{c}. \tag{7}$$

Uma vez que quanto menor o valor de λ mais pesada é a cauda da distribuição, esta equação indica que tal cauda torna-se cada vez mais leve com o aumento de a e, ao mesmo tempo, cada vez mais pesada com o aumento da média de graus c. Visando a comparação dos coeficientes computados a partir da Figura 4 com os calculados por meio da Equação 7, subtraímos uma unidade de cada coeficiente advindo da Figura 4. Tal comparação é explicitada na Tabela 1, onde observamos razoável proximidade entre os resultados teóricos e os provenientes das simulações.

Caso	Resultado Simulado	Resultado Esperado
Price: $c = 6$, $a = 1$	$\lambda = -2.28$	$\lambda = -2.17$
Price: $c = 6$, $a = 4$	$\lambda = -2.62$	$\lambda = -2.67$
Price: $c = 6$, $a = 7$	$\lambda = -3.06$	$\lambda = -3.17$

Tabela 1 – Coeficientes angulares referentes a diferentes valores do parâmetro a na Equação 6. Comparação de resultados empíricos com os previstos na Equação 7. Cada valor foi obtido através da média de 100 amostras.

2.4.3 Modelo Albert-Barabàsi: Fundamentos

O modelo Albert-Barabàsi (AB) deriva do modelo de Price. Nele, a torna-se igual a c e uma restrição é imposta:

$$q < c \Rightarrow p_q = 0. \tag{8}$$

Em suma, temos a definição de p_q da seguinte forma (NEWMAN, 2018, pp.501):

$$p_q = \begin{cases} \frac{B(q,3)}{B(c,2)} & \text{, para } q \geq c, \\ 0 & \text{, para } q < c. \end{cases} \tag{9}$$

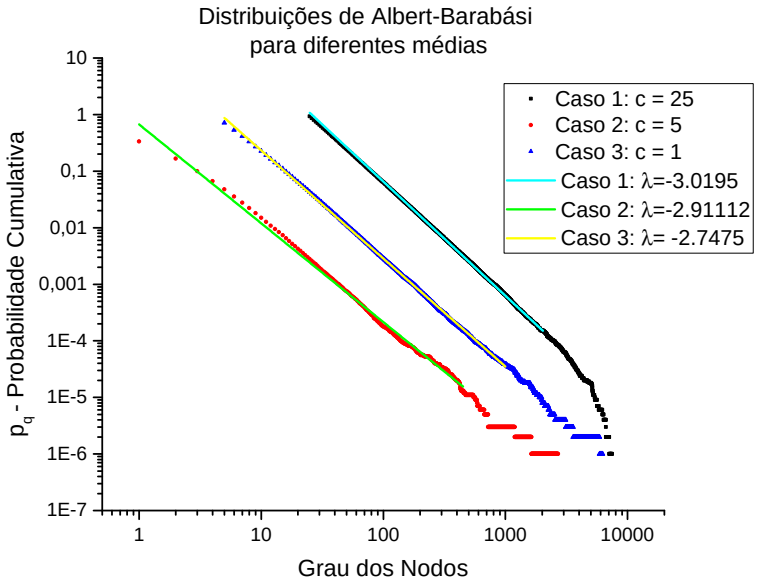


Figura 5 – Distribuição cumulativa inversa referente às probabilidades p_q de que um nodo tenha grau q. Exemplo de como a mudança da média no modelo AB não influencia significativamente a inclinação da distribuição de graus obtida.

Assim, como o modelo de Price apresenta-se como uma lei de potência, o modelo AB também possui a mesma característica. A Figura 5 exemplifica que, independente da média, a relação

$$p_q \sim q^{-3} \tag{10}$$

se manterá verdadeira. Entre diversas propriedades existentes, sabe-se que a rede AB é sempre de mundo pequeno. Outras propriedades particularmente relevantes sobre o comportamento do modelo são:

(a) A média dos tamanhos dos caminhos entre os nós da rede aproxima-se de (COHEN; HAVLIN, 2003)

$$\frac{\ln n}{\ln(\ln n)}. \tag{11}$$

(b) Para o coeficiente de clusterização global, CG, existem dois resultados analíticos (KLEMM; EGUILUZ, 2002),

$$CG \propto N^{-1}(lnN)^2 \tag{12}$$

e

$$CG \propto N^{-0.75}. \tag{13}$$

O cálculo do coeficiente de clusterização global é realizado a partir dos trios de nós conectados no grafo. Consideramos que um trio de nós está conectado se existe um caminho de arestas que os liga. Assim, dado um trio conectado, podemos ter duas ou três arestas associadas ao mesmo. Tal análise é efetuada tomando cada nó do grafo e encontrando todos os trios conectados que ele compõe. Assim, um grafo triangular (formado por três nós e três arestas compondo um triângulo) é computado três vezes na procura por trios conectados, pois é computado a partir de cada nó que o compõe. O coeficiente de clusterização global é definido como a razão entre o número de trios fechados (com três arestas), NF, e o número total de trios, NG:

$$CG = \frac{NF}{NT}. \tag{14}$$

A exploração do resultado dado no item (a) depende da noção de diâmetro e de raio de um grafo. O diâmetro de um grafo é definido como a maior distância encontrada entre dois nodos do grafo. Para entendermos o conceito de raio de um grafo, consideremos primeiro o de excentricidade de um nodo, que é a maior distância deste a outro nodo da rede. O raio de um grafo é a menor excentricidade encontrada entre as de todos os seus nodos. No artigo de Cohen e Havlin (2003), a medida de diâmetro dada é a média das distâncias entre dois nós quaisquer. Desta forma, o diâmetro d tem limites superior e inferior, $r \leq d \leq 2r$. Com relação ao item (b), é interessante ter dois resultados, o que envolve logarítmos, que segue mais de perto a curva numérica, mas não é tão simples, e o que se apresenta como uma lei de potência, no qual é mais fácil encontrar N por meio de um ajuste.

Apesar das relações simples descritas acima, muitas vezes é necessário nos distanciarmos da questão das métricas e avançarmos em questões conceituais. Isto nos permite focalizar a capacidade do modelo AB de se aproximar de diversos casos empíricos reais, cuja topologia da rede consegue descrever com precisão. Alguns exemplos:

- O modelo AB sempre gera redes de mundo pequeno, relacionadas à lei dos seis graus de separação. Isto é uma característica importante porque redes como a Web não precisam de mais de 6 *hops*, o que impacta o número de hubs de roteamento que um pacote precisa passar para chegar em seu destino (VINCIGUERRA; FRENKEN; VALENTE, 2010).

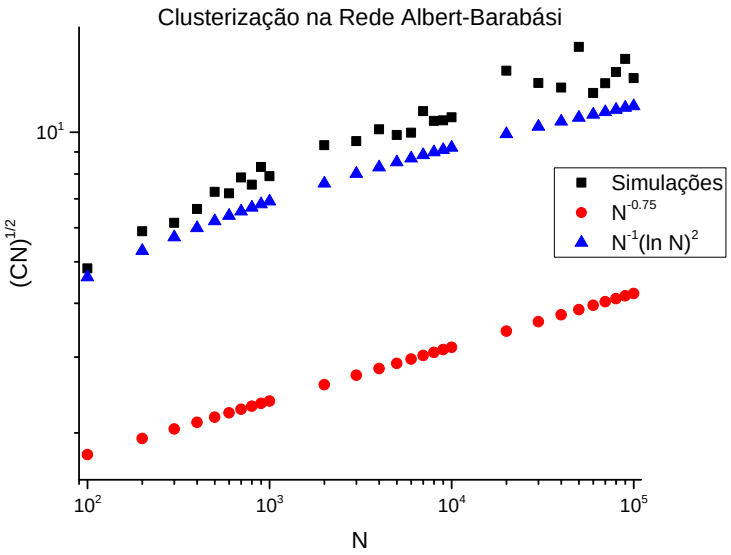


Figura 6 – Exemplo de como os dois modelos teóricos para o coeficiente de clusterização global se comportam. Dados provenientes de simulações também são apresentados. Cada ponto da série de simulações é a média de 100 amostras.

- Levando em consideração o fato de que o modelo AB é uma rede livre de escala, com $\lambda = -3$, e é construído com o uso da anexação preferencial, ele apresenta a capacidade de modelar uma série de sistemas distintos e ao mesmo tempo encontrar desafios para tal modelagem em outros sistemas. Neste contexto, destacamos:
  (a) Um modelo que exemplifica um desses desafios de redes sintéticas, como as do modelo AB, advém do trabalho de Girvan e Newman (2002), que refere-se à formação de comunidades dentro da rede, um fato que nas redes reais é muito comum. Deve-se tomar cuidado para não confundir o conceito de comunidade com o de assortatividade: comunidade corresponde à formação de clusters dentro da rede, de maneira a gerar grupos de nós bem distintos; já para assortatividade o foco é a ligação dos hubs da redes entre si.
  (b) Um caso em que o modelo AB serve de guia para o entendimento da complexidade de um sistema é o de redes neurológicas (BULLMORE; SPORNS, 2009).
  (c) Um modelo para descrever o funcionamento das células através de uma rede complexa de Albert-Barabàsi demonstra a relevância que o modelo AB possui diante de sua aplicações (BARABASI; OLTVAI, 2004).

No modelo AB encontramos diversos atributos, que são de fato as características que demarcam o modelo. A partir desses atributos, inovações são trazidas à tona (BIANCONI; BARABÁSI, 2001), onde o modelo se torna mais próximo do que observamos empiricamente, e novos nós inseridos precisam competir pelos links que possuirão. Entretanto, é importante notar que, apesar de AB ser um modelo de grande importância, este trabalho tem alternativas no que concerne a grafos com distribuição de graus em lei de potência,

como apresentado no próximo capítulo.

3 Modelo do Grafo de Zipf

Se realizarmos a contagem do número de vezes que cada palavra aparece em um determinado texto, podemos classificar tais palavras da seguinte forma: associamos a primeira posição à palavra mais frequente, a segunda posição à segunda palavra mais frequente, e assim por diante. Chamaremos de *rank* da palavra essa posição associada à mesma. O linguista George Kingsley Zipf (1902-1950) observou que, tipicamente, a frequência de uma palavra é, com boa aproximação, inversamente proporcional ao seu rank. A tal relação empírica foi dado o nome de *Lei de Zipf* (ZIPF, 2016). A Lei de Zipf baseia-se na ideia de que o comportamento humano tende a ocorrer pelo caminho de menor esforço. Ela foi investigada em contextos que vão além da linguística, tais como no estudo de distribuições de populações de municípios e de números de empregados de empresas (ZIPF, 2016). Variações da Lei de Zipf original são encontradas quando o rank da palavra é proporcional ao inverso da sua frequência elevada a uma potência diferente de 1. A *Distribuição de Zipf* é uma ferramenta matemática proposta para analisar dados aos quais julgamos que uma Lei de Zipf pode ser associada (NAIR; SANKARAN; BALAKRISHNAN, 2018). Ela é uma distribuição de lei de potência discreta que permite o cálculo do rank de um objeto que pertence a um conjunto que obedeça a uma Lei de Zipf.

Propomos aqui um novo modelo para a construção de grafos baseado na Lei de Zipf: o *Modelo de Grafo de Zipf*. Os objetivos que motivaram tal proposta são: possibilidade de escolher uma boa estimativa do grau mínimo do grafo, gerando assim um mecanismo que influencia diretamente a média do grafo e o comportamento da componente principal; e liberdade para escolher uma boa estimativa da inclinação da distribuição de graus do grafo. Tais objetivos foram atingidos.[1]

O modelo é construído a partir da função

$$f_{x,\rho,\alpha} = \begin{cases} \frac{x^{-(\rho+1)}}{\zeta_{\rho+1}} + \alpha & \text{, se } x \geq \alpha \\ 0 & \text{, se } x < \alpha \end{cases}, \tag{15}$$

proporcional à probabilidade de uma variável X aleatória assumir o valor x, sendo α e ρ parâmetros reais, e

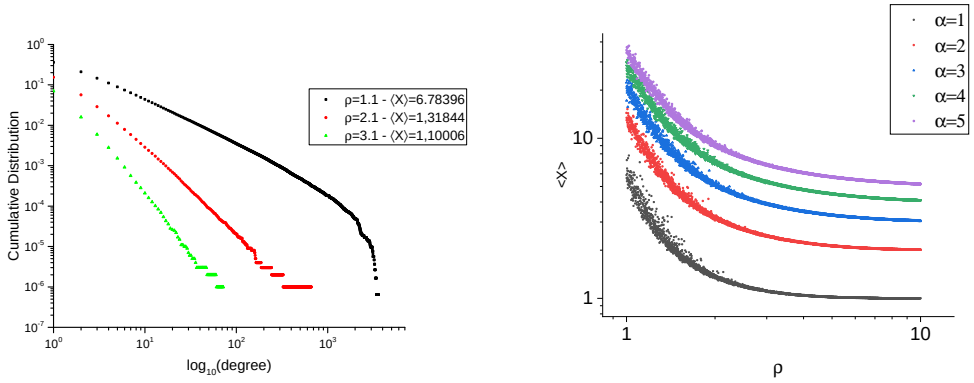
$$\zeta(\rho) = \sum_{i=1}^{\infty} i^{-\rho}, \tag{16}$$

[1] A escolha da palavra estimativa para cada um dos itens se dá pelo fato de que todos os processos propostos, apesar de na média alcançarem o valor desejado, para uma única amostra podem levar a valores díspares que, na realidade, representam um desvio devido ao método de Monte Carlo que é executado para a construção das redes.

a função Zeta de Riemann. A média para $\alpha = 1$ da Equação 15 é dada por:

$$\langle f_{x,\rho,1} \rangle = \frac{\zeta(\rho)}{\zeta(\rho+1)}, \quad \rho > 1. \tag{17}$$

Na Figura 7a, podemos ver com clareza que, por meio do parâmetro ρ da Equação 15, conseguimos controlar a inclinação da distribuição de graus, dada a tarefa de se construir um grafo por esta distribuição. Na Figura 7b, é feita uma análise do modelo da Equação 15 em que queremos saber como se comporta a média de uma distribuição criada pelo modelo. Na Figura 7b, é possível observar uma maior dispersão dos pontos quando os valores estão próximos de $\rho = 1$, para todos os valores de α. Devido ao fato de que $f_{x,\rho,\alpha} \leq f_{x,\rho,\beta}$ quando $\beta > \alpha$, bem como à parte não nula da distribuição $f_{x,\rho,\alpha}$ começar em α, os valores médios calculados crescem com α.



(a) Distribuição inversa acumulada de graus para diferentes valores de ρ. Observe a variação nas médias $\langle X \rangle$ e nas inclinações referentes a cada conjunto de pontos.

(b) Comportamento da média de graus $\langle X \rangle$ de acordo com a mudança do expoente ρ, para cinco valores de α.

Figura 7 – Na Figura 7a, é demonstrada a possibilidade de ajuste da inclinação da distribuição a partir da variação de um único parâmetro do modelo (ver Equação 15). Na Figura 7b, apresentamos um grafo para exemplificar a estabilidade das médias de graus $\langle X \rangle$ geradas pelo modelo (ver Equação 15); cada ponto no gráfico é o resultado médio de 1×10^4 sorteios da distribuição resultante da configuração criada naquele ponto.

A distribuição definida na Equação 15 será utilizada na construção do grafo de Zipf, proposto aqui, de acordo com o descrito a seguir. Consideremos que um grafo com N nós está sendo gerado. Assumamos também que M realizações da variável aleatória X são efetuadas seguindo a distribuição definida na Equação 15, com valores entre α e N. Tais resultados são acumulados em um vetor e o histograma a ele vinculado é produzido. A partir desse histograma, uma distribuição de probabilidades é calculada. Esta é empregada no processo de construção do grafo, associada aos graus dos nós, conforme descrito na Subseção A.3.2 do Apêndice A. Observe que as distribuições de probabilidades encontradas

podem variar de experimento para experimento, uma vez que há componente randômica em sua construção.

Pelo fato de serem os resultados puramente dependentes de sorteios, esse processo assemelha-se a um processo de Monte Carlo gerador de grafos. Um aspecto de destacada relevância é a periodicidade do gerador aleatório sendo utilizado, no presente caso, o Mersenne Twister (MATSUMOTO; NISHIMURA, 1998), cuja periodicidade é de $2^{19937} - 1$, considerado a melhor solução atual.

Podemos perceber que o modelo apresenta grande flexibilidade para a geração de grafos com diferentes estatísticas. Por exemplo, a partir da Figura 7b, concluímos que um ρ menor caracterizará um grafo com maior número médio de arestas, o que torna o grafo mais denso, e um ρ maior, um grafo com menos arestas, menos denso então. Ainda baseados nessa figura, podemos afirmar que tais variações de densidade também são alcançadas mediante a variação de α. Para além da densidade, a Figura 7a indica que outras características dos grafos podem ser buscadas trabalhando-se diretamente com as distribuições, que dependem da escolha de ρ e de α. A comparação de tais estatísticas com a de grafos oriundos de dados empíricos pode ser importante guia na definição da pertinência do método proposto em situações específicas de modelagem. O contexto, evidentemente, vai depender da origem dos dados.

4 Análise Multifractal de Flutuação sem Tendência - MF-DFA

A Análise Multifractal de Flutuação sem Tendência (*Multifractal Detrended Fluctuation Analysis* - MF-DFA) foi criada para calcular correlações temporais de longo alcance que possuam flutuações de diferentes amplitudes (KANTELHARDT et al., 2002). Ela é uma generalização do método (*Detrended Fluctuation Analysis* - DFA).

Consideremos um conjunto de valores x_k ($k = 1, 2, \ldots, T$) com suporte compacto, ou seja, em que a quantidade de elementos nulos é insignificante. Acompanhando o exposto em Kantelhardt et al. (2002), o método aplicado a um tal conjunto pode ser descrito pelos seguintes passos:

- **Passo 1:** Gerar o perfil Y_k a partir da série x_k, conforme definido abaixo:

$$Y_k \equiv \sum_{i=1}^{k} [x_i - \langle x \rangle], \tag{18}$$

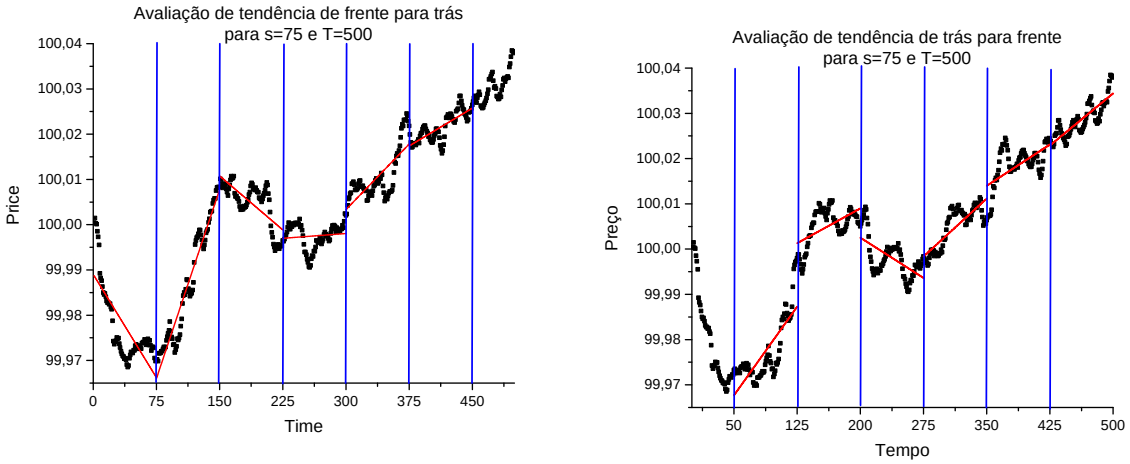
  onde $\langle x \rangle$ é a média tomada sobre todos os valores x_k. Segundo Kantelhardt et al. (2002), o termo da média não é obrigatório na Equação 18, pois no **Passo 3** ela acaba sendo eliminada.
- **Passo 2:** Dado este perfil, deve-se escolher uma escala s e dividi-lo de acordo com esta escala, obtendo-se, assim, T_s segmentos que não se sobrepõem - aqui, T_s é o resultado da divisão inteira de T por s: $T_s = int(T/s)$. Como a divisão na qual obtemos T_s pode não ser exata, alinhamos os segmentos a partir do início da série até o mais próximo possível do final, e, de forma análoga, com o intuito de contemplar os pontos no final da série, é realizado um alinhamento dos segmentos do final para início da série. Ou seja, visando não deixar de lado pontos no início ou no fim da série, esta é percorrida duas vezes; devido a este motivo, utilizamos $2T_s$ segmentos.
- **Passo 3:** Para cada segmento v, onde $v \in \{1, 2, \ldots, 2T_s\}$, é calculada a tendência local por meio do ajuste de um polinômio de grau m; o valor escolhido para m define a ordem do MF-DFA. Tal ajuste é realizado empregando-se o método de mínimos quadrados, utilizando pares ordenados construídos da seguinte forma: as ordenadas são os Y_k; à i-ézima ordenada de cada intervalo é associada abscissa com valor i. Utilizando o polinômio obtido para cada segmento, pontos $y_{v,i}$ são calculados para abscissas $i = 1, 2, \ldots, s$. Determinamos então as funções, análogas a variâncias,

$$F^2_{v,s} = \frac{1}{s} \sum_{i=1}^{s} \{Y_{(v-1)s+i} - y_{v,i}\}^2, \tag{19}$$

para os T_s segmentos definidos no sentido do início para o fim da série ($\upsilon \in \{1, \ldots, T_s\}$), e

$$F^2_{\upsilon,s} = \frac{1}{s}\sum_{i=1}^{s}\{Y_{T-(\upsilon-T_s)s+i} - y_{\upsilon,i}\}^2, \tag{20}$$

para os demais T_s segmentos ($\upsilon \in \{T_s + 1, \ldots, 2T_s\}$), definidos no sentido contrário. A consideração das diferenças entre os pontos da série, $Y_{(\upsilon-1)s+i}$ na Equação 19 e $Y_{T-(\upsilon-T_s)s+i}$ na Equação 20, e as aproximações lineares $y_{\upsilon,i}$ é o que causa a retirada de tendências (que justifica o termo *detrended* no nome do método). A Figura 8 ilustra tal processo para $m = 1$.



(a) (b)

Figura 8 – Exemplos de como as tendências são calculadas, tanto para a Equação 19, quanto para a Equação 20, nas Figuras 8a e 8b, respectivamente. Os pontos representam a série de Y_k. Em cada intervalo, as linhas vermelhas correspondem à aproximação linear, por mínimos quadrados, de cada intervalo υ, que produzem os valores dos $y_{\upsilon,i}$.

- **Passo 4**: Tomar a média sobre todos os segmentos para calcular a função de flutuação de ordem q

$$F_{q,s} \equiv \left\{\frac{1}{2T_s}\sum_{\upsilon=1}^{2T_s}[F^2_{\upsilon,s}]^{\frac{q}{2}}\right\}^{\frac{1}{q}}, \tag{21}$$

onde q pode assumir qualquer valor real. Para $q = 0$, a expressão usada é

$$F_{0,s} \equiv \exp\left\{\frac{1}{4T_s}\sum_{\upsilon=1}^{2T_s}\ln[F^2_{\upsilon,s}]\right\}. \tag{22}$$

Na Equação 21, quando $q = 2$ obtemos o método DFA padrão.

- **Passo 5**: Neste passo, funções de flutuação $F_{q,s}$ são computadas para vários valores de q e de s e o estudo das escalas que relacionam $F_{q,s}$ com um comportamento

exponencial,

$$F_{q,s} \propto s^{h_q}, \tag{23}$$

é realizado. Essa análise de escala pode ser feita através de gráficos log-log da relação apresentada na Equação 23. A Figura 9 exemplifica tais gráficos: os h_q correspondem aos coeficientes angulares obtidos por meio do ajuste de retas às curvas nesta figura, via método de mínimos quadrados. Os h_q são denominados expoentes de Hurst generalizados; para $q = 2$, tal expoente corresponde ao expoente de Hurst canônico.

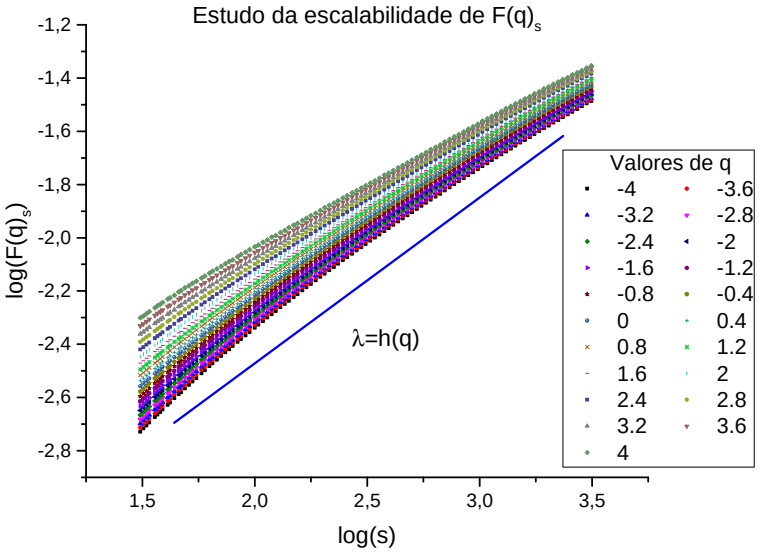


Figura 9 – Exemplo de estudo de escalabilidade para diferentes valores de q. Temos, no eixo das abscissas, as escalas s utilizadas no processo e, no eixo das ordenadas, as funções de flutuação $F_{q,s}$. O intervalo apresentado de valores de s vai de $s = 1.5$ a $s = 3.5$; os valores de q vão de $q = -4$ a $q = 4$. Para cada valor de q, uma curva foi construída. Mediante o ajuste linear de cada curva, os coeficientes de Hurst generalizados h_q são calculados: h_q coincide com o coeficiente angular λ da reta ajustada para o gráfico de $F_{q,s}$ em função de s.

Tendo encontrado um conjunto de expoentes de Hurst generalizados por meio dos passos acima, podemos investigar se existem correlações não lineares na série, e quão fortes essas correlações são. Para tanto, efetua-se uma análise da variação de h_q com q. Essa análise utiliza os resultados obtidos da Figura 9, onde é possível encontrar o coeficiente h_q para cada q. A partir deste ponto, é possível analisar como os expoentes h_q se comportam em relação ao parâmetro q. A Figura 10 exemplifica tal relação.

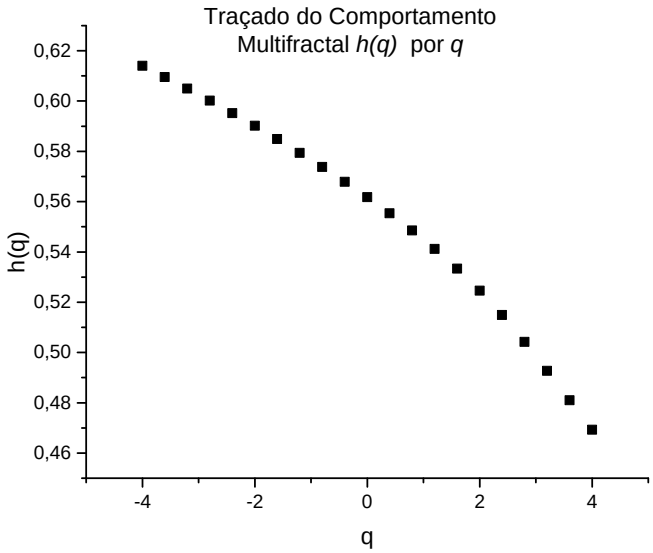


Figura 10 – Relação entre o parâmetro q e seus respectivos expoentes h_q.

Empregando os dados dispostos na Figura 10, torna-se possível construir o espectro de singularidade. O espectro de singularidade é uma forma de investigar a força das correlações não lineares existentes na série analisada. Em tal ferramenta, temos duas grandezas, f_α, que é o espectro de singularidade em si, e α, que é a força da singularidade. Essas duas grandezas podem ser calculadas da seguinte forma:

$$\alpha = h_q + qh'_q, \tag{24}$$

$$f_\alpha = q[\alpha - h_q] + 1. \tag{25}$$

Utilizando os dados da Figura 10 e as equações 24 e 25, é possível montar o espectro de singularidade. Um exemplo é encontrado na Figura 11.

Para uma série monofractal, idealmente h_q não varia com q, e o gráfico do espectro de singularidade colapsa num ponto. A largura do espectro, definida como a diferença entre a maior e a menor abscissa num gráfico como o da Figura 11 ($\Delta\alpha = \alpha_{max} - \alpha_{min}$), é uma testemunha do caráter multifractal da série sob análise. Na prática, uma largura diferente de zero costuma restar mesmo para uma série monofractal, devido à finitude da mesma. Outras causas para uma largura não nula, estas sim evidenciando multifractalidade, são distribuições com caudas pesadas e correlações não-lineares. Uma forma de investigar a presença de correlações não-lineares em uma série é comparar seu espectro de singularidade com o de uma versão embaralhada da mesma. Tal embaralhamento pode ser realizado como descrito a seguir. Consideremos a série x_i, que em sua ordem original gera a série Y_i, de acordo com a Equação 18. Trocando de lugar de maneira aleatória os elementos de x_i, a Equação 18 leva a uma nova série $\overline{Y}_i$, a série embaralhada. Na série embaralhada, possíveis correlações desaparecem. Assim, o uso da série $\overline{Y}_i$ é uma forma de descaracterizar as estruturas que proveem a série Y_i no que diz respeito às correlações não

lineares sendo pesquisadas. Isso significa que, ao calcularmos o espectro de singularidade de $\overline{Y}_i$, tal cálculo será feito sobre uma série mais próxima de um passeio aleatório. Desta forma, se uma das fontes de multifractalidade de uma série for a presença de correlações não-lineares, a largura do espectro de singularidade para a série embaralhada será menor do que para a série original. Uma maneira de detectar correlações não-lineares é, então, comparar tais larguras, da série embaralhada e da série original. Na Figura 11, o espectro de singularidade é apresentado também para a série embaralhada. A significativa redução de sua largura, em relação à da série original, aponta para a presença de correlações não lineares nesta. Na Figura 12, temos espectros de singularidade uma ordem de grandeza mais estreitos do que os apresentados na Figura 11. Embora isso aponte para um comportamento multifractal menos evidenciado, a diferença entre os dois $\Delta\alpha$ (de Y_i e de $\overline{Y}_i$) é uma indicação da presença de correlações não lineares na série original.

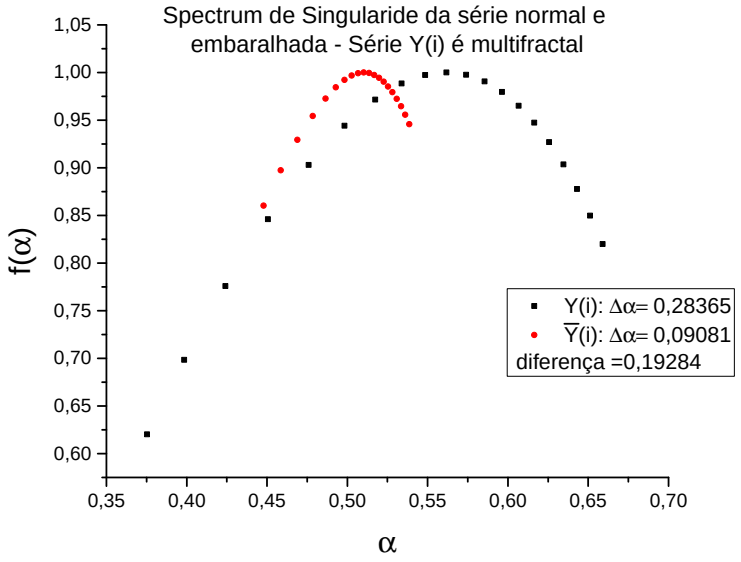


Figura 11 – Exemplo de espectro de singularidade das séries Y_i (original, na cor preta) e $\overline{Y}_i$ (embaralhada, na cor vermelha). Pode-se perceber que no caso da série embaralhada um $\Delta\alpha$ com menor valor é encontrado.

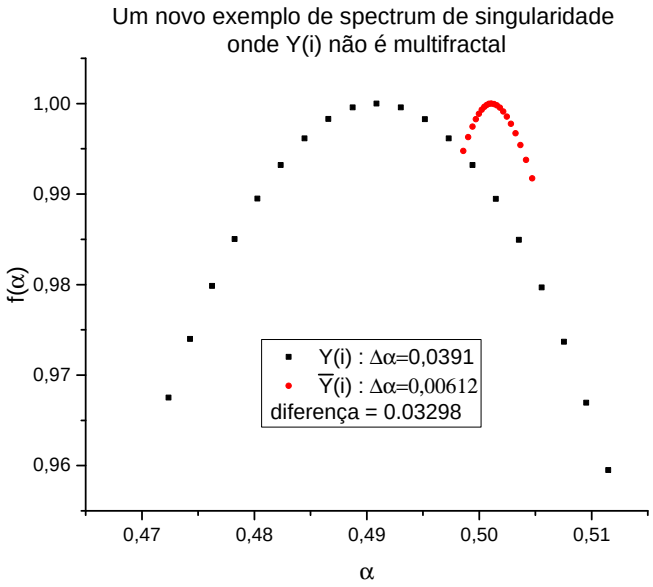


Figura 12 – Exemplo de espectro de singularidade das séries Y_i (original, na cor preta) e $\overline{Y}_i$ (embaralhada, na cor vermelha). Pode-se perceber que no caso da série embaralhada um $\Delta\alpha$ com menor valor é encontrado. Embora os valores de $\Delta\alpha$ nesta figura sejam significativamente menores do que os da Figura 11, a redução observada para a série embaralhada evidencia correlações não-lineares na série original.

5 O Modelo Multi-Agente

O modelo multi-agente delineado neste trabalho é uma adaptação do que é apresentado em Weng et al. (2012). No modelo em Weng et al. (2012), o intuito é simular uma rede social como a do Twitter, e estudar como a informação se espalha e como os *memes* distribuídos nessa rede se comportam. Nesse modelo, existe somente a dinâmica do fluxo de informação. Para a construção do modelo deste trabalho, acoplamos ao modelo em Weng et al. (2012) uma dinâmica de simulação de mercado, onde um grupo de agentes negocia um único ativo financeiro.

5.1 Elementos Estruturais do Modelo

Antes de descrevermos as duas dinâmicas que regem o modelo, serão definidas e comentadas as estruturas que caracterizam os agentes e o meio de comunicação implementado. Tais estruturas, expostas na Figura 13, são o agente e a postagem. Os agentes cumprem o papel de produzir e armazenar informação, bem como de fazê-la fluir em uma rede social, que os conecta. Tal fluxo se realiza por meio de postagens, que são os elementos responsáveis pelo transporte de informação. Com relação à natureza financeira do modelo, esta é centrada no agente, que age no sentido de realizar transações de compra e venda do ativo negociado.

A postagem é composta por um identificador e uma mensagem (ver Figura 13). Duas postagens podem ter mensagens idênticas, mas seus identificadores serão sempre diferentes. Assim, os identificadores distinguem as postagens. As mensagens referem-se às informações oriundas do mundo externo que chegam aos agentes. Cada mensagem é modelada por um número real obtido aleatoriamente de uma distribuição normal padrão. Quando esse número é positivo, a mensagem influencia o investidor no sentido de ir ao mercado demandar o ativo; quando é negativo, ela o influencia no sentido de ir ao mercado ofertar o ativo. Quanto maior for o módulo do número, maior a influência da mensagem sobre a decisão do agente. A forma como tal influência acontece será descrita detalhadamente na Seção 5.3. A escolha da distribuição normal padrão para o modelo da mensagem deve-se ao fato de que tal distribuição é das mais empregadas em modelos estatísticos relativos a variáveis aleatórias contínuas (MAGALHÃES; LIMA, 2002). O modelo pode ser facilmente adaptado para a utilização de outras distribuições.

A estrutura dos agentes também é apresentada na Figura 13. Existem duas filas em cada agente, que servem como repositórios de postagens. Esses repositórios são chamados de janela e de memória, e não têm tamanho máximo definido. Como é descrito em detalhes na Seção 5.2, a janela é o local onde o agente recebe postagens de seus vizinhos; já na memória, ele guarda a informação que o influencia efetivamente, tanto no que

se refere à atuação financeira, como na difusão de informação. De fato, o funcionamento do modelo remete à atenção limitada: nem toda postagem na janela vai para a memória, e somente postagens que estão na memória podem ser encaminhadas para outros agentes. Também somente as postagens na memória influenciam o comportamento financeiro do agente. Cada agente tem um registro associado à quantidade que possui de dinheiro, e outro, que descreve o número de ações que detém do ativo. Tais quantidades determinam o máximo que cada agente pode demandar ou ofertar no mercado.

Como mencionado acima, os agentes estão conectados entre si por uma rede. Diferentes topologias de rede podem ser utilizadas. De fato, os experimentos que realizamos indicam que a emergência de fatos estilizados típicos de séries financeiras depende de tais topologias. As redes complexas que construímos levaram a distribuições de retornos com caudas pesadas, bem como a correlações temporais nas série de preços que foram associadas a comportamento multifractal. Observemos que tal resposta do modelo aconteceu tendo como entrada mensagens, que modelam a influência do mundo exterior, com distribuição gaussiana. Quando redes não complexas foram empregadas, as distribuições de retornos praticamente não se afastaram da gaussianidade, e comportamento multifractal não foi evidenciado.

A comunicação entre agentes, ou seja, o espalhamento de postagens, ocorre de maneira assíncrona: a cada passo assíncrono, um agente é escolhido aleatoriamente e as operações relacionadas à geração, recepção e envio de postagens são executadas para aquele agente. Observemos, assim, que as alterações que ocorrem no sistema em um passo assíncrono são localizadas no agente escolhido e em sua vizinhança imediata. O tipo de rede é fundamental neste processo de difusão de informação. Depois de Γ passos assíncronos, a simulação do mercado financeiro é realizada de maneira síncrona, isto é, considerando todos os agentes simultaneamente, levando a modificações no sistema todo. Na Seção 5.2 descrevemos a dinâmica de informação e na Seção 5.3, a de mercado.

5.2 Dinâmica de Informação

A difusão de informação no modelo segue a dinâmica proposta em Weng et al. (2012), conforme esquematizado na Figura 14. Nesta dinâmica, a cada passo assíncrono um agente é escolhido de maneira aleatória com probabilidade uniforme. Este agente terá probabilidade p_n de criar uma nova postagem. Neste caso, a postagem é guardada na memória do agente escolhido e na janela dos agentes presentes em sua vizinhança. Observemos que uma mesma postagem pode estar presente, simultaneamente, em múltiplos sítios de memória ou de janela dos agentes; assim, uma postagem apresenta, tipicamente, diferentes instâncias. A ação de colocar uma postagem na memória do próprio agente e na janela de sua vizinhança será chamada de postar. No que se refere à dinâmica de informação, a interação entre os agentes ocorre por meio do ato de postar; a alimentação da memória com postagens

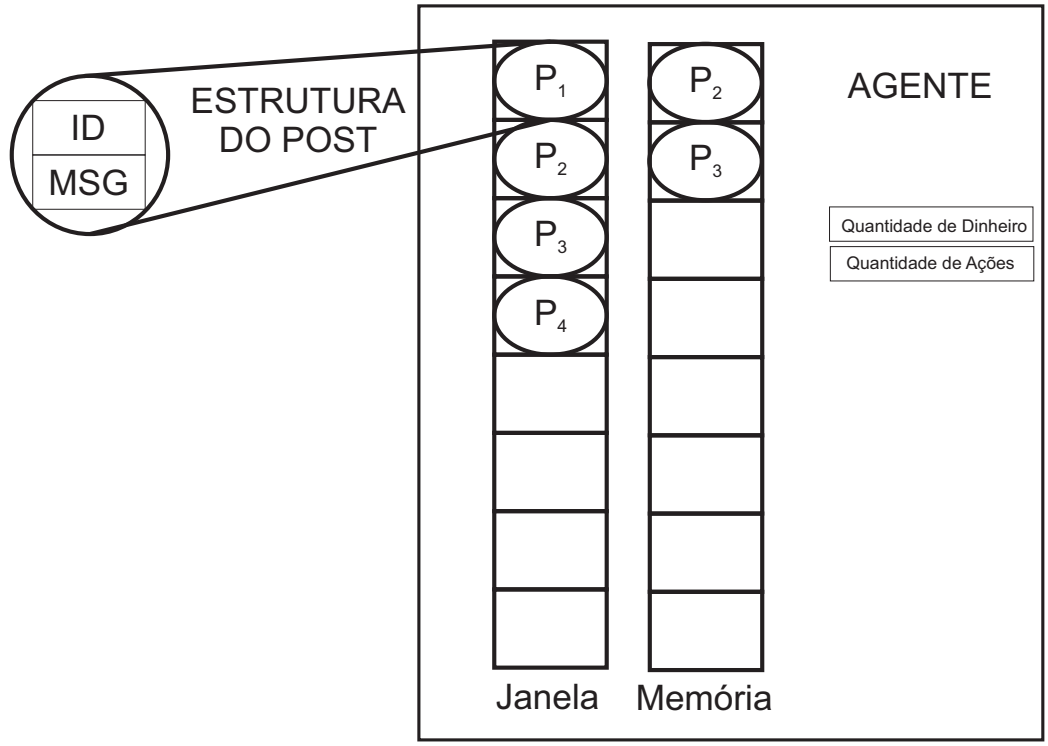


Figura 13 – Estrutura do agente e estrutura da postagem.

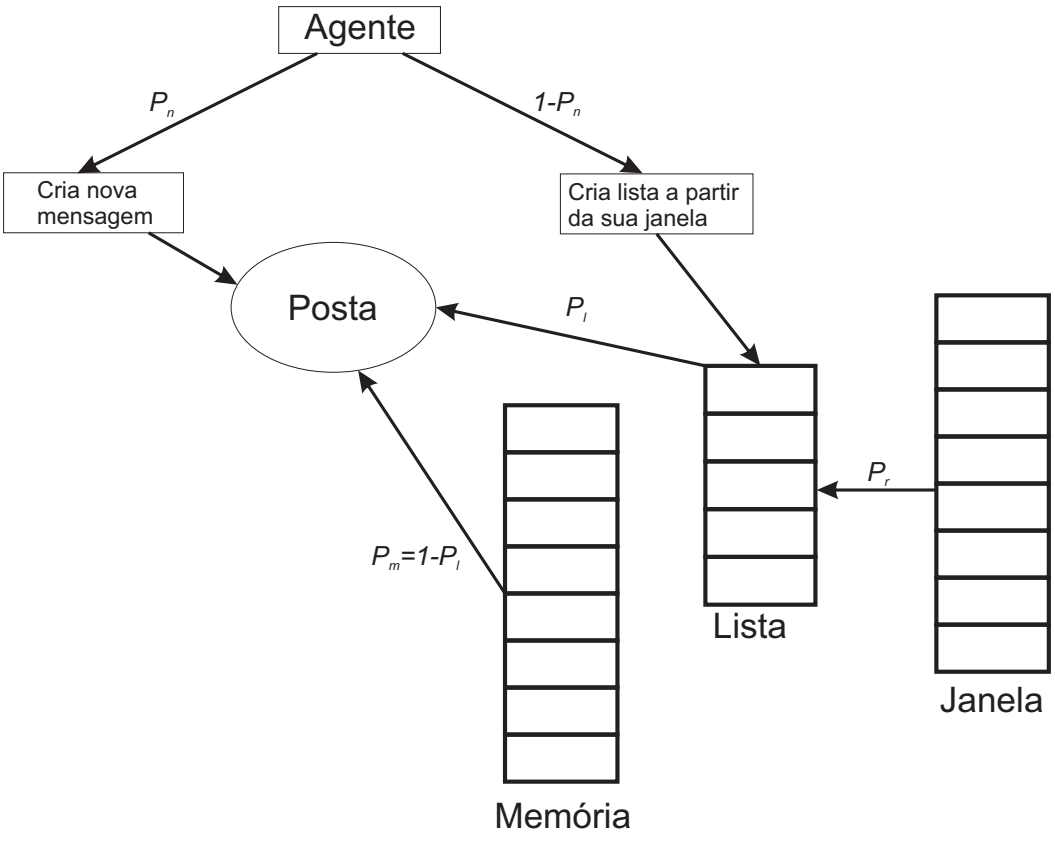


Figura 14 – Fluxograma associado ao processo de difusão de postagens.

também se dá através desse ato. Mesmo que o agente não crie uma nova mensagem, o que ocorre com probabilidade $1 - p_n$, ele irá postar. Neste caso, o primeiro passo do processo é a geração de uma lista que provém das postagens em sua janela; ressaltemos que esta lista é recriada em cada passo assíncrono, ou seja, não há aproveitamento de lista construída em passo anterior. A lista é preenchida da seguinte forma: cada postagem na janela do agente escolhido tem probabilidade p_r de ingressar na lista. Notemos que o tamanho da lista dependerá do valor de p_r e do número de mensagens na janela do agente. Os elementos da lista competirão com os elementos da memória no que se refere ao ato de postar. Para cada elemento da lista, o agente procede da seguinte maneira: com probabilidade p_l, realiza a ação de postar utilizando tal elemento, e, com probabilidade $p_m = 1 - p_l$, uma postagem da memória é escolhida aleatoriamente, com probabilidade uniforme, e é utilizada para ser postada. Vale enfatizar que todos os elementos da lista são considerados nesta etapa; portanto, quando este caminho de compartilhamento de postagens é sorteado, o número de postagens efetuadas no passo síncrono pelo agente escolhido será igual ao número de elementos da lista, que depende, como salientamos, de p_r e do número de postagens na janela.

Observemos que toda vez que um agente posta sua memória é alimentada com a postagem utilizada. Sendo assim, tal memória pode receber postagens originais, criadas pelo agente no passo síncrono em questão, ou postagens presentes na janela ou na memória do agente, num processo de repostagem. A possibilidade de postar mensagens da memória introduz comportamento de auto-interesse: neste caso, a ação de postar replica, nas janelas da vizinhança e na própria memória, as postagens selecionadas na memória do agente. Uma mesma postagem pode ocupar mais de uma posição na memória de um agente. Observemos, por outro lado, que nem todas as postagens da janela chegam à memória, o que introduz no modelo o efeito de atenção limitada. Como veremos na Seção 5.3, somente as postagens na memória do agente influenciam seu comportamento financeiro. Cada instância de postagem possui vida finita: deixa de existir após um certo número (aqui denominado T_w) de passos assíncronos contados a partir do momento de sua inserção, seja na janela, seja na memória. Isto também pode ser interpretado como atenção limitada.

5.3 Dinâmica do Mercado

A cada Γ passos assíncronos, é realizado um passo síncrono, onde o estado financeiro de todos os agentes é atualizado.[1] O tempo t da simulação é definido a partir dos passos síncronos: $t = 1$ corresponde ao primeiro passo síncrono, $t = 2$ ao segundo, e assim por diante. No passo síncrono, os agentes demandam ou ofertam de acordo com

[1] Em modo síncrono, as postagens nas janelas e memórias dos agentes não se alteram: instâncias de postagens presentes antes de um passo síncrono continuarão presentes logo depois do mesmo. O tempo de vida de tais instâncias depende de T_w, que pode ser menor, igual ou maior que Γ.

seu humor e sua condição financeira. A condição financeira do agente i no passo síncrono de tempo t é definida por $M_{i,t}$, que corresponde à quantidade de dinheiro do agente, e por $S_{i,t}$, que corresponde ao número de ações que o agente possui. Neste modelo, $M_{i,t}$ e $S_{i,t}$ podem assumir quaisquer valores reais não-negativos. O humor $H_{i,t}$ do agente i no passo síncrono t é proporcional à soma das mensagens das postagens presentes na memória do agente i no instante t:

$$H_{i,t} = k_i \sum_j m_{i,j,t}, \tag{26}$$

onde a soma é realizada sobre todas as postagens presentes na memória do agente, $m_{i,j,t}$ são a mensagens referentes a tais postagens, e $k_i > 0$ é um parâmetro de controle que determina a intensidade com que o conteúdo das mensagens afeta o humor do agente i. No que segue, consideramos que a sensibilidade de todos os agentes com relação às mensagens é a mesma, o que nos leva num conjunto de valores de k_i iguais entre si: $k_i = k$. Como descrito na Seção 5.1, as mensagens são números aleatórios obtidos a partir de uma distribuição normal padrão. O valor conceitual das mensagens é modelado da seguinte forma: números positivos (negativos) contribuem no sentido do agente demandar (ofertar) o ativo; o módulo do número determina a intensidade dessa contribuição. A resultante de todas as mensagens define o humor. Se $H_{i,t}$ é positivo, o agente demanda no tempo t. Se $0 \leq H_{i,t} \leq 1$, o agente demanda o seguinte número de ações:

$$D_{i,t} = H_{i,t} \frac{M_{i,t}}{P_t}, \tag{27}$$

onde P_t é o preço de uma ação no tempo t. Se $H_{i,t} > 1$, o agente i demanda o equivalente a todo o seu dinheiro:

$$D_{i,t} = \frac{M_{i,t}}{P_t}. \tag{28}$$

De forma análoga, se $H_{i,t}$ for negativo, o agente i ofertará $O_{i,t}$ ações, onde

$$O_{i,t} = \begin{cases} |H_{i,t}| S_{i,t}, & \text{se} \quad -1 \leq H_{i,t} < 0, \\ S_{i,t}, & \text{se} \quad H_{i,t} < -1. \end{cases} \tag{29}$$

A dinâmica do preço é modelada, aqui, considerando-se que a variação relativa do preço $(P_{t+1} - P_t)/P_t$ é proporcional ao excesso de demanda. Assim, o preço P_t evolui da seguinte forma:

$$P_{t+1} = P_t + \beta(D_t - O_t)P_t, \tag{30}$$

onde

$$D_t = \sum_i D_{i,t} \tag{31}$$

é a demanda agregada e

$$O_t = \sum_i O_{i,t} \tag{32}$$

é a oferta agregada, ambas no tempo t. As somas nas definições de D_t e de O_t são realizadas sobre o conjunto de agentes que demandam e sobre o conjunto de agentes que ofertam, no tempo t, respectivamente. Por sua vez, $\beta > 0$ é o parâmetro que define a sensibilidade do preço ao excesso de demanda $D_t - O_t$. Neste modelo, nem toda oferta é completamente integralizada em venda, ocorrendo de maneira análoga no que se refere à demanda. A cada passo das dinâmicas simuladas, o total de ações sendo vendidas é igual ao total de ações sendo compradas. Para que isto seja mantido, a efetivação de compras e vendas a partir das demandas e ofertas acontece da seguinte maneira: se a demanda agregada for maior que a oferta agregada no tempo t ($D_t > O_t$), os agentes que demandam em t comprarão

$$B_{i,t} = D_{i,t}\frac{O_t}{D_t} \tag{33}$$

e aqueles que estiverem ofertando terão toda a sua oferta convertida em venda. Uma correção análoga é realizada quando o total da oferta for maior que o da demanda ($O_t > D_t$): neste caso, cada agente que oferta efetivamente venderá

$$S_{i,t} = O_{i,t}\frac{D_t}{O_t}, \tag{34}$$

ao mesmo tempo em que os que demandam conseguirão comprar o total a que se propõem.

6 Experimentos Realizados

Com o objetivo de testar o modelo, foram criados cinco tipos de configuração para o sistema, definidos pelos tipos escolhidos para os modelos de rede. Cada configuração leva a resultados diferentes naquilo que concerne a fatos estilizados encontrados em estudos lidando com dados reais. Como veremos, a emergência de tais estatísticas a partir de entradas (mensagens) caracterizadas por distribuição gaussiana depende crucialmente do modelo de rede adotado na simulação. O modelo apresenta duas dinâmicas acopladas: difusão de informação e negociação em mercado. Serão analisadas cada dinâmica sozinha e também a relação entre elas.

6.1 Escolha das Redes e dos Parâmetros

Os experimentos foram delineados para mensurar como a topologia que define as vizinhanças dos agentes interfere no volume médio de postagens, bem como na curva de preços do ativo simulado. Para detectar uma real diferença de comportamentos, foram escolhidas cinco redes distintas. Trabalhamos com um modelo de rede regular circular (REG), que é construído da mesma forma que a rede circular da contribuição de Watts e Strogatz (1998), com a única diferença de que aqui a rede é direcionada. Utilizamos também o modelo de Erdös-Rényi (ER) (ERDÖS; RÉNYI, 1959), uma rede aleatória. Este modelo foi inicialmente proposto para redes não direcionadas, mas neste trabalho é utilizado como rede direcionada. A terceira rede é o modelo de Albert-Barabási (AB) (BARABÁSI; ALBERT, 1999), que, por sua vez, foi criado originalmente para redes direcionadas. Também foram construídas duas redes a partir de um modelo baseado na distribuição de Zipf; dois diferentes conjuntos de parâmetros foram utilizados, originando as redes Z_1 e Z_2.

Neste trabalho existem parâmetros fixos, cujos valores são comuns a todos os experimentos: o número de agentes, $N = 10^4$, a probabilidade de gerar novas postagens, $p_n = 0.45$, a probabilidade de postar o que se encontra na própria memória do agente, ou auto-interesse, $p_m = 1 - p_l = 0.40$. A média de graus da rede, $c = 4$, foi escolhida de tal forma que a densidade de arestas, ϵ, fosse da mesma ordem de grandeza que a apresentada em Weng et al. (2012), que serviu de base para nosso modelo de difusão de informação. Aqui, $\epsilon \approx 4 \times 10^4$, em Weng et al. (2012), temos $\epsilon \approx 3 \times 10^4$.

Já os parâmetros, p_r, T_w, Γ, β e k foram escolhidos visando-se encontrar um equilíbrio em que a média do número de postagens por agente por passo síncrono calculada em todas as amostras, Φ, alcance valores na ordem de grandeza entre 0.1 e 1 para todas as redes: $\Phi = 0.3$ para AB, $\Phi = 0.7$ para ER, $\Phi = 0.7$ para REG, $\Phi = 0.5$ para Z_1 e $\Phi = 0.3$ para Z_2. A escolha de um mesmo patamar de Φ ajuda a caracterizar os sistemas, e dizer comparativamente quão semelhantes eles são. É importante notar que p_r controla

Tabela 2 – Valores dos parâmetros não fixos, para cada rede.

MODELO DE REDE	p_r	T_w	Γ	β	k
Albert-Barabási	0.16	0.75	0.5	1×10^{-6}	1×10^{-3}
Erdös-Rényi	0.30	0.75	0.8	1×10^{-6}	2×10^{-3}
Regular	0.40	0.425	0.8	1×10^{-6}	2×10^{-3}
Z_1	0.10	0.75	0.6	1×10^{-6}	1×10^{-3}
Z_2	0.07	0.85	0.7	1×10^{-6}	1×10^{-3}

a chance de um espalhamento maior de uma determinada postagem. O parâmetro T_w controla o tempo de vida da postagem, aumentando ou diminuindo o espalhamento da mesma. Por sua vez, o incremento em Γ tem a propriedade de aumentar a probabilidade de uma postagem ser espalhada com mais intensidade. A constante β não tem influência direta sobre o processo difusivo de informação, servindo como um controlador da magnitude das variações de P_t. O parâmetro k, por sua vez, influencia o valor do humor a ser considerado pelos agentes: modificações em k podem gerar flutuações muito grandes em P_t. A Tabela 2 apresenta os valores destes parâmetros. Todos os experimentos foram realizados utilizando-se $T = 2 \times 10^5$ passos síncronos, correspondendo a $N\Gamma T \approx 10^9$ passos assíncronos.

6.2 Processo de Difusão de Informação

A análise do processo de difusão de informação ocorre, aqui, por meio da procura de distribuições de probabilidades que se afastam da gaussianidade, bem como da observação de flutuações clusterizadas. No contexto da Econofísica, é comum que tais fenômenos sejam estudados a partir de séries de preços ou de retornos, o que será realizado na próxima seção. Para podermos construir um procedimento análogo, referente à dinâmica de informação, é necessário que proponhamos grandezas para serem estatisticamente investigadas de forma equivalente. Com este intuito, começamos definindo ω_j como o número médio de postagens por agente no passo síncrono j. Seguindo os passos do encontrado em Peng et al. (1995), definimos o perfil

$$\Omega_j = \sum_{k=1}^{j} \left(\omega_k - \langle\omega\rangle\right), \tag{35}$$

onde $\langle\omega\rangle$ é a média de ω_j sobre todos os passos síncronos da amostra. Devido à forma da equação acima, denominamos Ω_j de *fluxo de informação integrado*. No presente estudo, Ω_j terá tratamento estatístico análogo ao preço, e $\Delta\Omega_j = \Omega_j - \Omega_{j-1}$, ao retorno. Com o uso destas grandezas, poderemos observar a emergência de distribuições com caudas pesadas e de comportamento multifractal na dinâmica de informação.

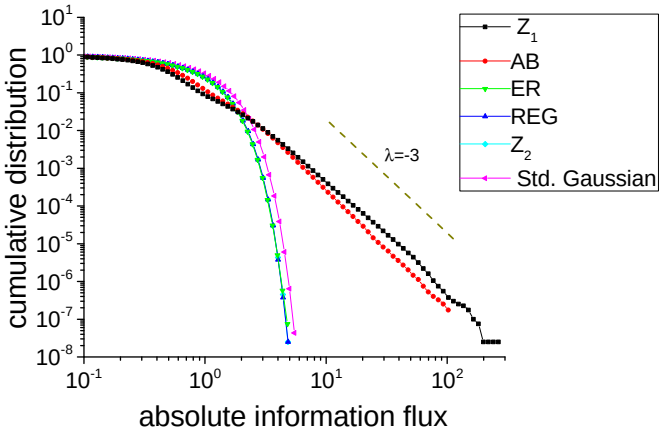


Figura 15 – Distribuições cumulativas de probabilidades de $|\delta\Omega_j|$ para os modelos AB, ER, CR, Z_1 e Z_2. Caudas pesadas são observadas para os casos AB e Z_1. Uma linha reta com coeficiente angular $\gamma = -3$ é colocada como um guia para os olhos.

Na Figura 15, é apresentada a distribuição cumulativa de probabilidades da normalização de $|\Delta\Omega_j|$, que é calculada como

$$|\delta\Omega_j| = \frac{|\Delta\Omega_j| - \langle|\Delta\Omega|\rangle}{\sigma(|\Delta\Omega|)}, \tag{36}$$

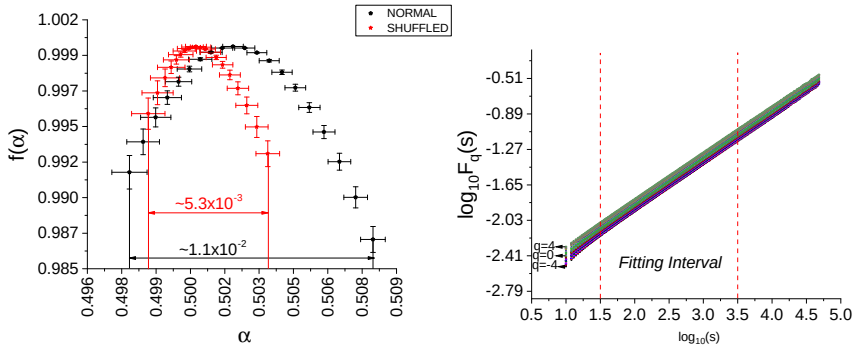
onde $\langle|\Delta\Omega|\rangle$ e $\sigma(|\Delta\Omega|)$ são a média e o desvio padrão de $\Delta\Omega_j$, respectivamente. Note que as distribuições dos experimentos referentes a AB e Z_1 apresentam caudas pesadas. Um linha reta com coeficiente angular $\gamma = -3$ é apresentada como um guia para os olhos. Tal resultado evidencia a importância da topologia da rede para a ocorrência deste fato estilizado no modelo apresentado. Para entendermos melhor essa relação, suponhamos que uma mensagem com probabilidade muito baixa de ocorrer (na cauda da gaussiana que define tais probaiblidades) apareça em um agente com muitas conexões – um hub. Ele distribuirá essa mensagem para muitos outros agentes. Assim, o hub é capaz de fazer com que a frequência de uma mensagem com probabilidade baixa seja significativamente aumentada. Esse mecanismo é impossível em uma rede regular, por exemplo. A não linearidade, representada pela potencialização de algumas mensagens, vem da estrutura da rede.

Foi realizada investigação sobre a existência de correlações não-lineares em séries temporais relacionadas ao processo de difusão de informação. Para tanto, foi utilizado o MFDFA, como descrito em Kantelhardt et al. (2002), nas séries de fluxo de informação integrado Ω_j. Uma vez que a dinâmica do processo de difusão de informação é diretamente conectada à dinâmica de mercado, o aparecimento de regiões (no tempo) em que o número de postagens é singularmente elevado em geral terá consequências na dinâmica de preços. Assim, “bolhas” no fluxo de informação, isto é, intervalos onde o número de postagens

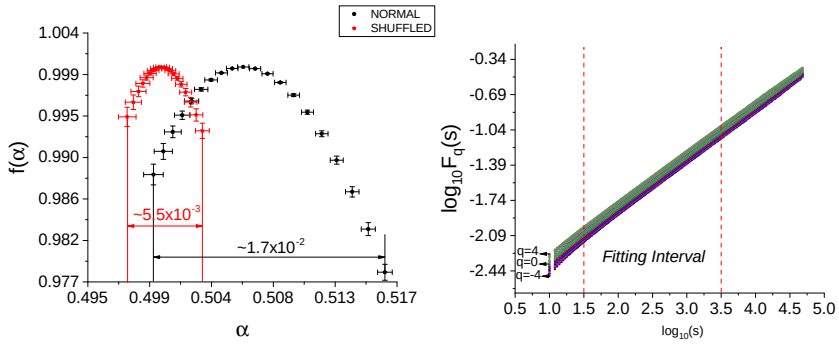
efetuadas assume valores particularmente altos, em princípio podem explicar a ocorrência de clusterização de volatilidade em séries de preços. Tal fenômeno, que pode ser entendido como o concentração de períodos de volatilidade elevada em regiões bem definidas, tem sido observado em séries reais. Como veremos a seguir, o comportamento multifractal, que evidencia a presença de correlações não-lineares, varia consideravelmente com a escolha da rede, o que também joga luz sobre o papel da topologia da mesma para o aparecimento de estatísticas típicas de séries financeiras reais.

Nas figuras 16a, 16c e 16e, é apresentado o espectro de singularidade para as séries de Ω_j de três casos estudados: CR, ER e Z_2. As larguras $\Delta\alpha$ são da ordem de 10^{-2} e 10^{-3}; estes valores são obtidos a partir dos pontos representados pelos pentágonos de cor preta. As figuras 16b, 16d e 16f referem-se ao cálculo dos expoentes de Hurst generalizados empregados na construção dos respectivos espectros de singularidade. Nelas, são indicadas as regiões em que foram calculados tais expoentes. Dada a boa aproximação das curvas por retas, observa-se o comportamento de escala bem caracterizado nessas regiões. Larguras não nulas $\Delta\alpha$ podem ser devidas a três causas: finitude da série analisada, cauda pesada em distribuição associada a seus elementos, e presença de correlações não lineares. Quando consideramos séries embaralhadas,[1] o último fator deixa de existir. Os espectros de singularidade são apresentados também para séries de Ω_j embaralhadas, sendo representados pelas estrelas de cor vermelha. As larguras $\Delta\alpha$ diminuem significativamente quando consideramos séries embaralhadas para CR e ER. Associamos tal diminuição à destruição das correlações presentes nas séries originais; assim, concluímos que tais correlações existem nas séries originais. Como veremos adiante, as larguras calculadas a partir das séries associadas a CR e ER são bem menores do que para as associadas a AB e Z_1, o que indica que as correlações não lineares presentes nas primeiras são menos relevantes do que nas outras. Já para as séries associadas a Z_2, a diminuição da largura do espectro de singularidade não é significativa, o que aponta para a ausência de correlações não lineares neste caso. A largura residual após o embaralhamento pode ser creditada majoritariamente à finitude das séries, principalmente para CR, ER e Z_2, que não apresentam distribuições com caudas pesadas.

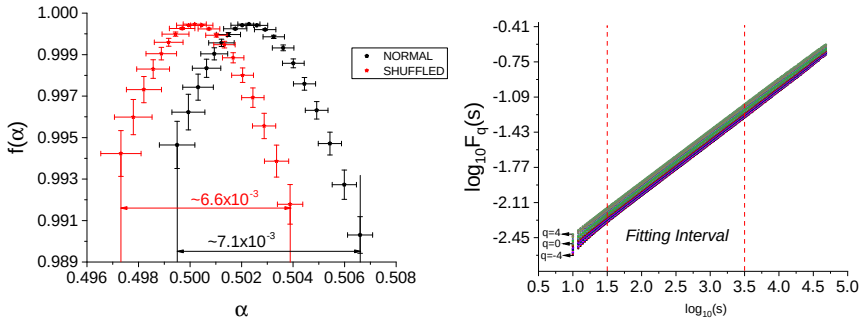
[1] Para embaralhar uma série, primeiro calculamos o conjunto das diferenças entre seus pontos adjacentes. Em seguida, criamos uma série nova da seguinte forma: escolhe-se um valor qualquer para o primeiro ponto; cada ponto seguinte é computado somando-se o valor do ponto mais recentemente obtido com uma das diferenças calculadas, escolhida aleatoriamente no conjunto. Uma vez que uma diferença tenha sido utilizada, ela é extraída do conjunto, de forma que não possa ser empregada novamente. Esse procedimento é repetido até que o conjunto de diferenças calculadas esteja vazio.



(a) CR (b) $F_q(s)$ para CR

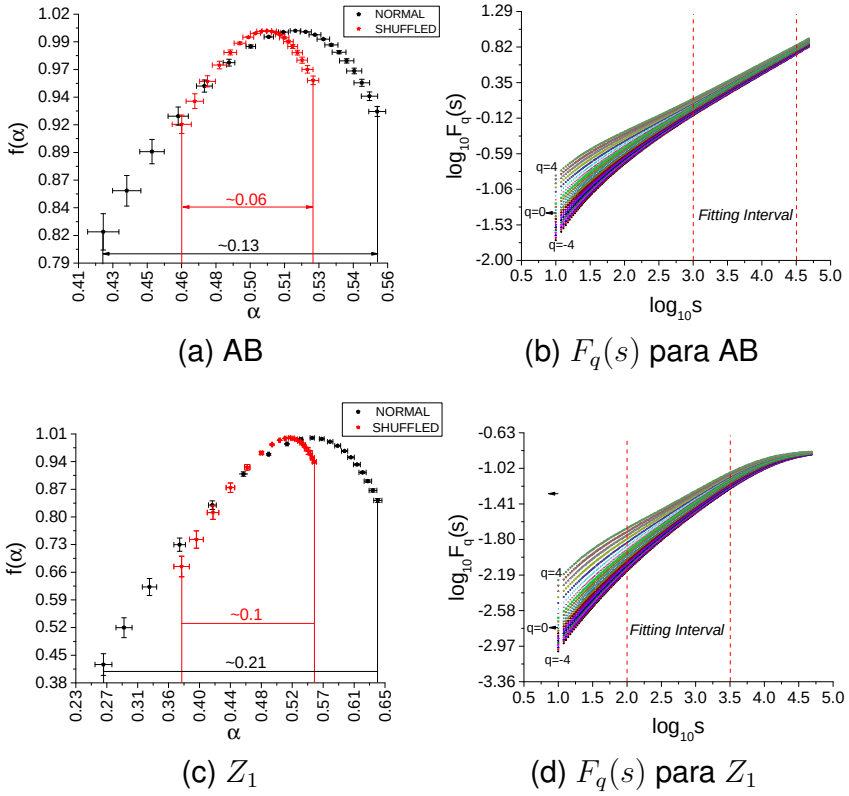


(c) ER (d) $F_q(s)$ para ER



(e) Z_2 (f) $F_q(s)$ para Z_2

Figura 16 – À esquerda, encontram-se espectros de singularidade para séries de Ω_j para os modelos CR, ER e Z_2. Pentágonos pretos correspondem a médias de 2×10^2 amostras de 2×10^5 passos síncronos cada uma, com o método MFDFA aplicado às séries originais. Estrelas vermelhas correspondem ao mesmo número de amostras com o mesmo número de passos, no entanto, para séries embaralhadas. As barras de erro para α e $f(\alpha)$ são apresentadas. No caso das séries originais, as larguras $\Delta\alpha$ para CR e ER são da ordem de 10^{-2}; já para Z_2, são da ordem de 10^{-3}. Após embaralhar as séries, tais larguras caem, todas, apresentando ordem de grandeza de 10^{-3}. Diminuição relevante foi observada para CR e ER, sugerindo a presença de correlações não lineares; a discreta diminuição da largura para Z_2 aponta para correlações pouco significativas neste caso. À direita, funções de flutuação $F_q(s)$ são dadas para diferentes escalas s. As regiões onde os ajustes lineares foram efetuados, visando o cálculo dos expoentes de Hurst generalizados (q variando de -4 a 4), são indicadas.



(a) AB (b) $F_q(s)$ para AB

(c) Z_1 (d) $F_q(s)$ para Z_1

Figura 17 – À esquerda, encontram-se espectros de singularidade para as série de Ω_j para os modelos AB e Z_1. Pentágonos pretos correspondem a médias de 2×10^2 amostras de 2×10^5 passos síncronos cada uma, com o método MFDFA aplicado às séries originais. Estrelas vermelhas correspondem ao mesmo número de amostras com o mesmo número de passos, no entanto, para séries embaralhadas. As barras de erro para α e $f(\alpha)$ são apresentadas. As larguras $\Delta\alpha$ para as séries originais são da ordem de 10^{-1}, sugerindo a existência de correlações não lineares relevantes nas séries. Tal indicação é confirmada pela diminuição significativa de tais larguras quando o espectro é construído a partir das séries embaralhadas. À direita, funções de flutuação $F_q(s)$ são dadas para diferentes escalas s. As regiões onde os ajustes lineares foram efetuados, visando o cálculo dos expoentes de Hurst generalizados (q variando de -4 a 4), são indicadas.

As figuras 17a e 17c mostram o espectro de singularidade para as séries de Ω_j referentes às redes AB e Z_1. Os símbolos pretos correspondem às séries originais e os vermelhos, às séries embaralhadas. Nas figuras 17b e 17d, pode ser observado o comportamento das funções de flutuação $F_q(s)$ utilizadas na computação dos expoentes de Hurst generalizados associados às séries originais. As regiões escolhidas para o ajuste linear que leva a tais expoentes estão indicadas. As larguras dos espectros relativos a AB e Z_1 são uma ordem de grandeza maiores que as larguras associadas a ER, CR e Z_2. Isto indica maior relevância das correlações não lineares presentes nas séries oriundas das redes AB e Z_1. Depois do embaralhamento, há decréscimo significativo tanto para o caso

AB, quanto para Z_1, demonstrando a presença de correlações não-lineares importantes na dinâmica de ambos os experimentos.

6.3 Efeitos sobre a Negociação no Mercado

Resultados análogos foram encontrados nas distribuições cumulativas de probabilidades de retornos e nos espectros de singularidade das séries de preços.

A dinâmica de informação dirige a dinâmica de mercado, que observamos a partir das séries de preços P_t. Estas possibilitam o cálculo de retornos logarítmicos

$$R_t = \log P_t - \log P_{t-1} \tag{37}$$

e de sua versão normalizada

$$r_t = \frac{R_t - \langle R \rangle}{\sigma_R}, \tag{38}$$

onde $\langle R \rangle$ é a média e σ_R o desvio padrão dos R_t, calculados sobre a série inteira. Na Figura 18, encontramos as distribuições cumulativas de probabilidades dos retornos logarítmicos normalizados para os modelos AB, ER, CR, Z_1 e Z_2. Cauda pesada é evidenciada para os casos AB e Z_1. A comparação da cauda encontrada para AB e Z_1 com a linha reta de coeficiente angular $\gamma = -3$ (apresentada como um guia para os olhos) explicita a conformidade com resultados empíricos de duas décadas atrás (GOPIKRISHNAN et al., 1999; KWAPIEN et al., 2003). A ausência de cauda pesada, verificada para as distribuições de retornos relativas a CR, ER e Z_2, é um resultado que espelha períodos mais recentes (DROŻDŻ et al., 2007; BOTTA et al., 2015).

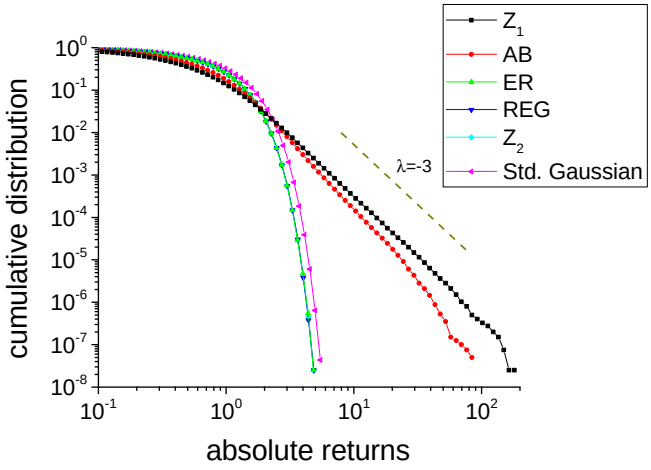
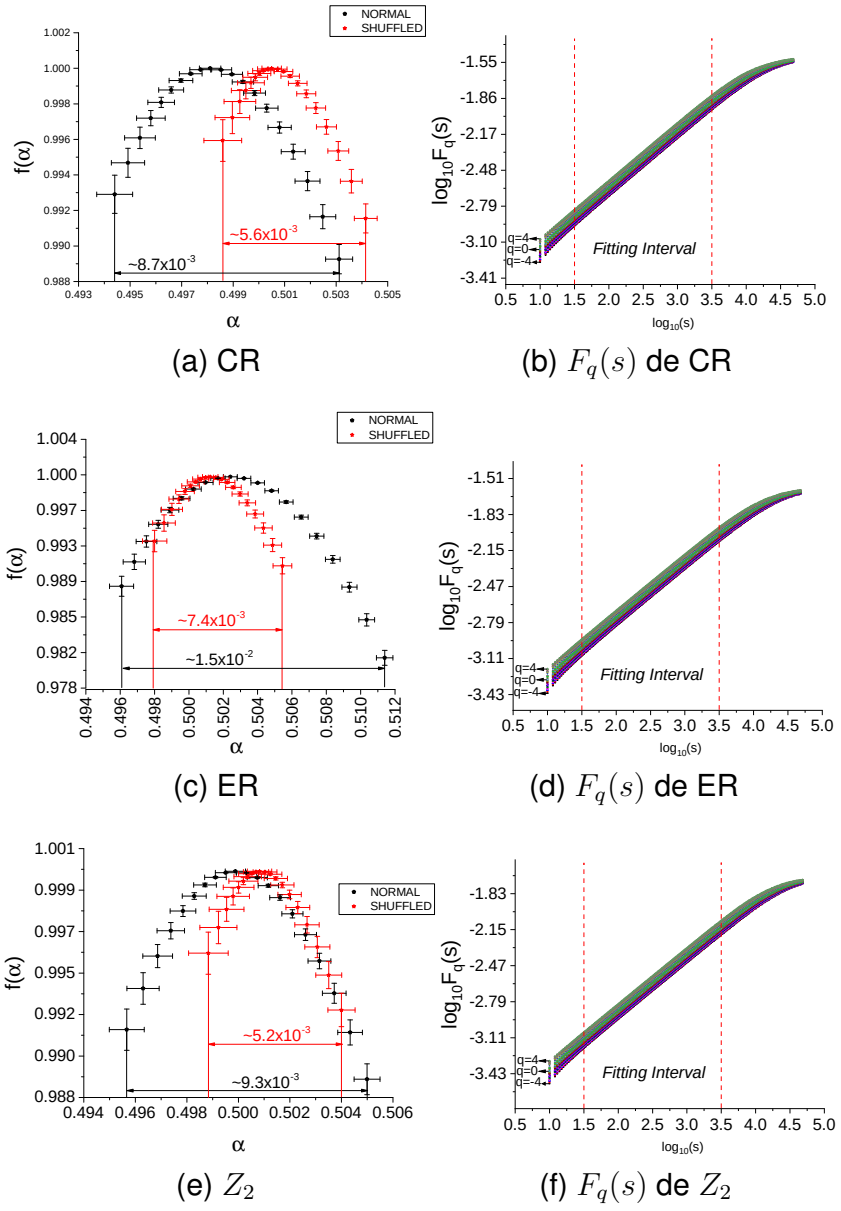


Figura 18 – Distribuições cumulativas de probabilidades dos retornos normalizados para os modelos AB, ER, CR, Z_1 e Z_2. Cauda pesada é observada para os casos AB e Z_1. Uma linha reta com coeficiente angular $\gamma = -3$ é colocada como um guia para os olhos.



(a) CR (b) $F_q(s)$ de CR

(c) ER (d) $F_q(s)$ de ER

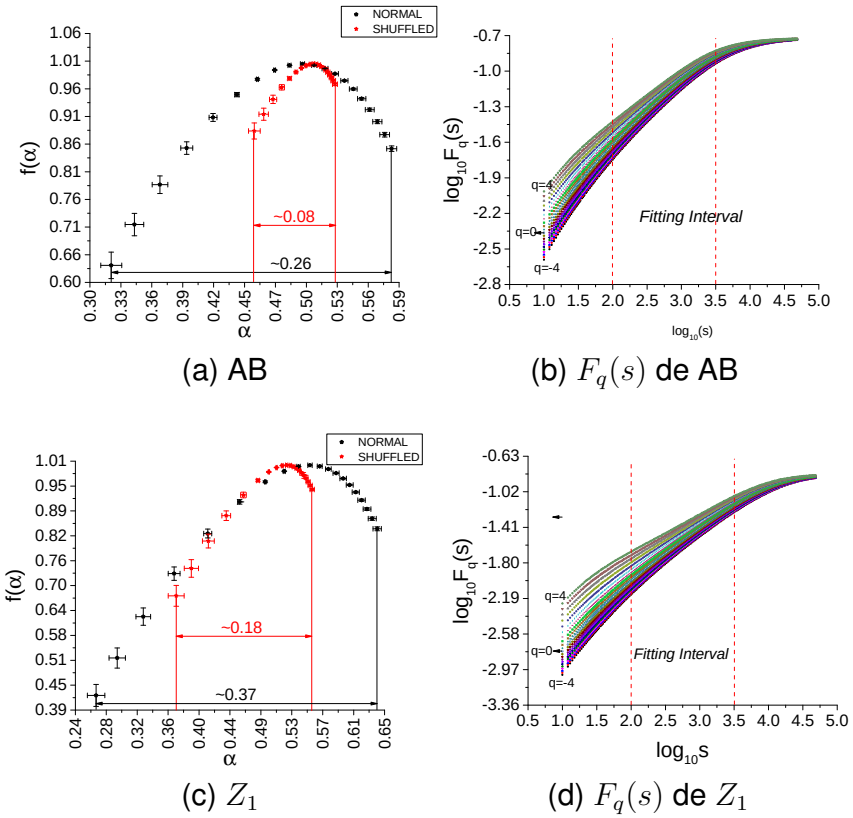
(e) Z_2 (f) $F_q(s)$ de Z_2

Figura 19 – À esquerda, encontram-se espectros de singularidade para as séries de preços P_t concernentes aos modelos CR, ER e Z_2. Os resultados referem-se a médias sobre 2×10^2 amostras de 2×10^5 passos síncronos cada uma, com o método MFDFA aplicado às séries originais. As barras de erro para α e $f(\alpha)$ são apresentadas. Pentágonos pretos correspondem a séries originais e estrelas vermelhas, a séries embaralhadas. As larguras $\Delta\alpha$ para as séries originais são da ordem de 10^{-2} ou 10^{-3}. Para séries embaralhadas, tais larguras caem para aproximadamente a metade do valor referente às séries originais. Isto sugere a existência de correlações não lineares nas séries, com intensidade menor que a dos casos AB e Z_1 (ver figuras 20a e 20c). À direita, funções de flutuação $F_q(s)$ são dadas para diferentes escalas s. As regiões onde os ajustes lineares foram efetuados, visando o cálculo dos expoentes de Hurst generalizados (q variando de -4 a 4), são indicadas.

Foi também realizada a MFDFA para séries de preços sintetizadas pelo do modelo.

As figuras 19 e 20 demonstram que, analogamente ao que foi encontrado na análise da dinâmica da informação, correlações não-lineares (e, possivelmente, clusterização de volatilidade) ocorrem de maneira mais proeminente no caso dos experimentos que utilizam as redes AB e Z_1.

Os resultados para dinâmica de preços confirmam o papel central da topologia da rede no surgimento de estatísticas típicas de mercados financeiros. No presente modelo, diferentes redes levam a diferentes dinâmicas de difusão de informação, que produzem, por sua vez, diferentes dinâmicas de mercado. Redes complexas possibilitam a emergência de estatísticas nas séries de preços e de retornos estruturalmente diferentes das estatísticas que caracterizam as mensagens. Enquanto estas corresponderam, nos experimentos realizados, a números aleatórios com distribuição normal padrão, as séries de preços e de retornos exibiram, quando redes complexas foram utilizadas, caudas pesadas e correlações não-lineares.



(a) AB (b) $F_q(s)$ de AB

(c) Z_1 (d) $F_q(s)$ de Z_1

Figura 20 – À esquerda, encontram-se espectros de singularidade para as séries de preços P_t concernentes aos modelos AB e Z_1. Os resultados referem-se a médias sobre 2×10^2 amostras de 2×10^5 passos síncronos cada uma, com o método MFDFA aplicado às séries originais. As barras de erro para α e $f(\alpha)$ são apresentadas. Pentágonos pretos correspondem a séries originais e estrelas vermelhas, a séries embaralhadas. As larguras $\Delta\alpha$ para as séries originais são da ordem de 10^{-1}. No caso de séries embaralhadas, tais larguras caem, em relação ao valor relativo às séries originais, para aproximadamente um terço para AB e aproximadamente a metade para Z_1. Isto indica a presença de correlações não lineares nas séries, com intensidade maior que a dos casos CR, ER e Z_2 (ver figuras 19a, 19c e 19e). À direita, funções de flutuação $F_q(s)$ são dadas para diferentes escalas s. As regiões onde os ajustes lineares foram efetuados, visando o cálculo dos expoentes de Hurst generalizados (q variando de -4 a 4), são indicadas.

7 Conclusão

A partir deste trabalho, podemos tirar conclusões sob diversas visões científicas. Primeiramente, salientemos, acerca do modelo de distribuição de informação, que conseguimos replicar o modelo de Weng et al. (2012). Nele, o tráfego de pacotes de informação permite a extração de informações que ganham relevância quando for feita diante de medidas relacionadas à rede sendo utilizada. Um dos mais importantes fenômenos que observamos na dinâmica deste modelo é a difusão de informação que acontece através de movimentos que chamaremos de *cascatas de postagens*. Tal fenômeno pode ser reconhecido como um processo difusivo quando acontece de maneira muito intensa para uma determinada postagem. De maneira objetiva, definimos que o efeito cascata ocorre no espaço de tempo em que uma postagem, tendo sido muito repassada, apresenta distância da sua origem até o nodo mais distante equivalente ao diâmetro do grafo. Se esse fenômeno não pudesse ser observado ou replicado, o trabalho não seria viável no que concerne à reprodução de fatos estilizados usualmente observados em séries financeiras. Podemos dizer que uma grande massa de nós ser atingida por uma postagem é um evento raro. Quando ocorre, muitos nós terão a informação relativa a essa postagem em sua janela antes do tempo de vida, T_w, da postagem da origem expirar. Algumas postagens podem difundir-se significativamente, tendo relevância na dinâmica de informação, porém sem atingir o diâmetro do grafo; nestes casos, não falaremos em efeito cascata, pois reservamos a palavra para o caso em que ele alcança o diâmetro do grafo. A frequência com que esse evento ocorre depende das constantes p_r e p_l, bem como de T_w, para determinar a duração do efeito. É importante ressaltar que esses eventos aumentam em muito o custo computacional da simulação, especialmente se ocorrerem muitos concomitantemente.

Percebemos que esse modelo apresenta um conjunto de possibilidades a serem exploradas, porém não esgotadas aqui. Diferentes ajustes dos parâmetros podem levar a dinâmicas bastante diversas. Neste trabalho, usamos as capacidades do modelo em termos de uma metodologia de difusão de informação discreta, postagens, e assim atrelamos às postagens pesos gaussianos, de acordo com a distribuição normal padrão. Os pesos gaussianos são, de fato, uma maneira interessante de atribuir os valores neste caso. Relembremos que o agente tem uma janela e uma memória, e que existem três origens para os itens que estão na memória: uma mensagem criada pelo agente, elementos da janela e elementos repetidos da memória. No entanto, eles correspondem sempre a números que fazem parte de uma distribuição gaussiana e, ao calcular o humor, o somatório de tais variáveis continua sendo gaussiano, com desvio padrão igual ao número de elementos somados. Notemos que em uma rede em que há muitas postagens repassadas e na qual as janelas ficam cheias, dependendo do valor de p_r, os agentes terão memórias com maior número de elementos, gerando um humor que pode se anular, mas, ao mesmo tempo,

pode assumir valores extremos, tanto no sentido negativo como no positivo. É por isso que escolhemos um multiplicador escalar como atenuador das variações da soma da memória, para o cálculo do humor, porque, dependendo dos parâmetros da simulação, as memórias podem ficar mais cheias do que o comum, e este multiplicador deve ser diminuído de acordo com as grandezas.

Com a criação do ferramental do humor, o modelo busca replicar o mercado em termos da precificação e do lapso de tempo entre a chegada da informação e a mudança de preço do ativo. Para isso, os agentes primeiro postam suas ofertas/demandas totais. A atualização de preços é de acordo com tais valores. Porém, nem toda a oferta/demanda tem que ser suprida completamente. De fato, quando um lado é mais volumoso que o outro, a efetivação das transações se faz de acordo com uma constante de proporcionalidade, de forma que os totais de compras e vendas realizadas sejam equivalentes. Desta forma, realiza-se o ciclo síncrono financeiro e então atualizam-se os dados dos agentes no que se refere a quanto cada um conseguiu vender ou comprar.

Determinado o modelo econômico, os modelos de redes agora são agregados ao modelo de informação, são eles: Albert-Barabàsi (AB), Erdös-Rényi (ER), circular regular (CR) e o modelo de Zipf em suas duas versões, Z_1 e Z_2. Os fatos estilizados que conseguimos detectar foram caudas pesadas e correlações não-lineares, que podem sinalizar clusterização de volatilidade. Caudas pesadas foram encontradas na distribuição de retornos e na distribuição de fluxo de informação do modelo AB e Z_1. Este resultado da distribuição de retornos corresponde a um resultado de duas décadas atrás (GOPIKRISHNAN et al., 1999; KWAPIEN et al., 2003), mas o de fluxo de informação é uma mensuração que está de acordo com as medidas realizadas em Weng et al. (2012). As distribuições de retornos de ER e CR são resultados compatíveis com os resultados atuais (DROŻDŻ et al., 2007; BOTTA et al., 2015), já as distribuições de fluxo de informação de ER, CR e Z_2 seguem uma gaussiana. Para AB e Z_1, conseguimos achar fortes correlações não-lineares tanto para as séries de preço como para as séries de fluxo de informação. É importante notar as reduções de uma ordem de grandeza nas larguras dos espectros de singularidade, que ocorreram nas séries de fluxo de informação para ER e CR, das séries embaralhadas em comparação com as séries originais. Nas séries de preços, houve tal redução de uma ordem de grandeza para ER somente. A consideração dessas reduções, tanto para o fluxo de informação como para séries de preços, é importante porque elas representam que uma pequena correlação não-linear existe, que vai além do insignificante.

Explicitamos o caminho para a complexidade no modelo de difusão de informação e de mercado proposto. Tal caminho passa pelo papel fundamental que a rede que caracteriza os contatos dos agentes exerce para a emergência de distribuições com caudas pesadas e de correlações não-lineares a partir de mensagens gaussianas e não correlacionadas. Podemos entender o processo que leva a tais emergências considerando uma mensagem muito rara aparecendo em um agente hiperconectado - um hub. Tal mensagem será

distribuída para muitos outros agentes, elevando significativamente sua frequência na rede, o que impacta o humor médio dos agentes e, desta forma, a dinâmica financeira. Por outro lado, a presença de hubs produz intervalos de hiperatividade na difusão de informação: um hub num estado em que posta um elevado número de mensagens impacta no mesmo sentido parcela significativa da rede. Estes mecanismos não ocorrem numa rede sem hubs. Observamos, assim, a criação de organização a partir de entradas completamente aleatórias, organização esta construída de maneira endógena pelo sistema, por meio da estrutura da rede. A geração de complexidade está associada à topologia da rede, este é o caminho para a complexidade, tema central desta tese.

As abordagens que este trabalho propõe com os experimentos relativos a redes e a Lei de Zipf, além das simulações relacionadas ao modelo de agentes, que foram realizados para que este pudesse ser concretizado, abrem muitos campos para novas ideias. Nossa proposta de uma nova rede, o modelo de Zipf, ainda não está completamente estudada, existe muito a entender da geração dos hubs e da existência de assortatividade ou não. Ainda há simulações para serem feitas que possibilitem uma visão intuitiva das constantes ρ e α, isto ainda há de se modelar. Também seria interessante pesquisar a possibilidade de usar dados empíricos que respeitem uma lei de Zipf e usar a distribuição de probabilidades desses dados no processo de construção das redes empregadas no modelo.

Na pesquisa desenvolvida até o momento, as mensagens foram definidas a partir da distribuição normal padrão. Um estudo relevante como passo futuro é investigar as dinâmicas produzidas pelo modelo empregando outras distribuições na caracterização das mensagens, visando testar a robustez das conclusões obtidas com relação ao tipo de entrada no modelo. Um exemplo nesse sentido seria verificar se mensagens com distribuições uniformes levam ao surgimento de fatos estilizados nas mesmas redes em que tais estatísticas apareceram ao utilizarmos a distribuição gaussiana. Por outro lado, a investigação da situação em que já as entradas (mensagens) estejam associadas a distribuições com caudas pesadas, visando a comparação dos resultados obtidos para redes complexas e não-complexas, também pode fornecer um caminho interessante para pesquisa futura. A pergunta, neste caso, refere-se a que tipos de fatos estilizados serão encontrados nas respostas que o sistema fornece.

Outro ponto possível a ser pesquisado é conseguir achar um expoente de difusão de informação e encontrar uma maneira de medir essa difusão com ideias atreladas à Equação do Calor $\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2}$. Considerando que as redes, muitas vezes, não têm estrutura semelhante à do espaço euclidiano, a Equação do Calor não pode ser diretamente aplicada, este é mais um problema aberto. A justificativa de usar ideias oriundas da Equação Calor é que ela está profundamente arraigada em diversos campos da Física. Conseguir engendrar um processo não físico a um processo físico como esse é extremamente interessante e deve responder muitas perguntas em diferentes áreas.

Em notas finais, é interessante simular um livro de ofertas para definição do preço do

ativo, que, então, não seria mais regimentado somente pela equação de preços. Também poderia-se considerar que as ações não pudessem ser fracionadas. Um protótipo neste sentido está sendo desenvolvido. Nessa nova abordagem, um modelo de rede de difusão de informação também estaria presente. Diferentes redes podem ser consideradas, de forma que uma investigação do caminho para a complexidade como a que realizamos aqui também poderá ser levada adiante.

Referências

ARISTOTLE. **Metaphysics**. [S.l.]: The Internet Classics Archive, 350BCE. Citado na página 1.

ASHKENAZY, Y. et al. Magnitude and sign scaling in power-law correlated time series. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 323, p. 19–41, 2003. Citado na página 3.

ATMAN, A.; GONÇALVES, B. A. Influence of the investor's behavior on the complexity of the stock market. **Brazilian Journal of Physics**, Springer, v. 42, n. 1-2, p. 137–145, 2012. Citado na página 2.

BAKKER, L. et al. A social network model of investment behaviour in the stock market. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 389, n. 6, p. 1223–1229, 2010. Citado na página 2.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **science**, American Association for the Advancement of Science, v. 286, n. 5439, p. 509–512, 1999. Citado 9 vezes nas páginas xi, 1, 3, 11, 34, 62, 63, 64 e 65.

BARABASI, A.-L.; OLTVAI, Z. N. Network biology: understanding the cell's functional organization. **Nature reviews genetics**, Nature Publishing Group, v. 5, n. 2, p. 101, 2004. Citado na página 17.

BIANCONI, G.; BARABÁSI, A.-L. Competition and multiscaling in evolving networks. **EPL (Europhysics Letters)**, IOP Publishing, v. 54, n. 4, p. 436, 2001. Citado na página 17.

BOLLOBÁS, B. Random graphs. 2001. **Cambridge Stud. Adv. Math**, 2001. Citado 2 vezes nas páginas 10 e 61.

BOTTA, F. et al. Quantifying stock return distributions in financial markets. **PloS one**, Public Library of Science, v. 10, n. 9, p. e0135600, 2015. Citado 3 vezes nas páginas 3, 40 e 45.

BRODER, A. et al. Graph structure in the web. **Computer networks**, Elsevier, v. 33, n. 1-6, p. 309–320, 2000. Citado 2 vezes nas páginas 1 e 12.

BUHAUG, H.; URDAL, H. An urbanization bomb? population growth and social disorder in cities. **Global Environmental Change**, v. 23, n. 1, p. 1–10, 2013. Citado na página 12.

BULLMORE, E.; SPORNS, O. Complex brain networks: graph theoretical analysis of structural and functional systems. **Nature reviews neuroscience**, Nature Publishing Group, v. 10, n. 3, p. 186, 2009. Citado na página 17.

BUONOCORE, R. J.; ASTE, T.; MATTEO, T. D. Measuring multiscaling in financial time-series. **Chaos, Solitons & Fractals**, Elsevier, v. 88, p. 38–47, 2016. Citado na página 3.

CAJUEIRO, D. O.; TABAK, B. M. The hurst exponent over time: testing the assertion that emerging markets are becoming more efficient. **Physica A: Statistical Mechanics and its Applications**, v. 336, n. 3, p. 521 – 537, 2004. ISSN 0378-4371. Disponível em:

⟨http://www.sciencedirect.com/science/article/pii/S0378437103011828⟩. Citado na página 12.

CALVET, L.; FISHER, A. Multifractality in asset returns: theory and evidence. **Review of Economics and Statistics**, MIT Press, v. 84, n. 3, p. 381–406, 2002. Citado na página 3.

CHIARELLA, C. et al. Asset price dynamics among heterogeneous interacting agents. **Computational Economics**, Springer, v. 22, n. 2-3, p. 213–223, 2003. Citado na página 2.

COHEN, R.; HAVLIN, S. Scale-free networks are ultrasmall. **Physical review letters**, APS, v. 90, n. 5, p. 058701, 2003. Citado 2 vezes nas páginas 15 e 16.

DROŻDŻ, S. et al. Stock market return distributions: From past to present. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 383, n. 1, p. 59–64, 2007. Citado 3 vezes nas páginas 3, 40 e 45.

DUCHA, F.; ATMAN, A.; MAGALHÃES, A. B. de. Information flux in complex networks: Path to stylized facts. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 566, p. 125638, 2021. Citado na página 4.

DUNN, S.; WILKINSON, S. M. Increasing the resilience of air traffic networks using a network graph theory approach. **Transportation Research Part E: Logistics and Transportation Review**, v. 90, p. 39–50, 2016. ISSN 1366-5545. Risk Management of Logistics Systems. Citado na página 12.

EGBERT, J.; BIBER, D. Incorporating text dispersion into keyword analyses. **Corpora**, v. 14, n. 1, p. 77–104, 2019. Citado na página 77.

ERDÖS, P.; RÉNYI, A. On random graphs, i. **Publicationes Mathematicae (Debrecen)**, v. 6, p. 290–297, 1959. Citado 5 vezes nas páginas 3, 9, 10, 34 e 56.

FAN, Y. et al. The effect of investor psychology on the complexity of stock market: An analysis based on cellular automaton model. **Computers & Industrial Engineering**, Elsevier, v. 56, n. 1, p. 63–69, 2009. Citado na página 2.

FENG, L. et al. Linking agent-based models and stochastic models of financial markets. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 109, n. 22, p. 8388–8393, 2012. Citado na página 2.

GIRVAN, M.; NEWMAN, M. E. Community structure in social and biological networks. **Proceedings of the national academy of sciences**, National Acad Sciences, v. 99, n. 12, p. 7821–7826, 2002. Citado na página 17.

GOPIKRISHNAN, P. et al. Scaling of the distribution of fluctuations of financial market indices. **Physical Review E**, APS, v. 60, n. 5, p. 5305, 1999. Citado 3 vezes nas páginas 3, 40 e 45.

GREEN, E.; HANAN, W.; HEFFERNAN, D. The origins of multifractality in financial time series and the effect of extreme events. **The European Physical Journal B**, Springer, v. 87, n. 6, p. 129, 2014. Citado na página 3.

GRIES, S. T. Dispersions and adjusted frequencies in corpora. **International journal of corpus linguistics**, John Benjamins, v. 13, n. 4, p. 403–437, 2008. Citado 3 vezes nas páginas x, 74 e 75.

GU, G.-F.; ZHOU, W.-X. et al. Detrending moving average algorithm for multifractals. **Physical Review E**, APS, v. 82, n. 1, p. 011136, 2010. Citado na página 3.

HASAN, M. R. et al. A highly sensitive gold-coated photonic crystal fiber biosensor based on surface plasmon resonance. **Photonics**, v. 4, n. 1, 2017. Citado na página 54.

ISLAM, R. et al. Bend-insensitive and low-loss porous core spiral terahertz fiber. **IEEE Photonics Technology Letters**, v. 27, n. 21, p. 2242–2245, 2015. Citado na página 54.

JOHNSON, F. et al. Natural hazards in australia: floods. **Climatic Change**, Springer, v. 139, n. 1, p. 21–35, 2016. Citado na página 1.

KALISKY, T.; ASHKENAZY, Y.; HAVLIN, S. Volatility of linear and nonlinear time series. **Physical Review E**, APS, v. 72, n. 1, p. 011913, 2005. Citado na página 3.

KALISKY, T.; ASHKENAZY, Y.; HAVLIN, S. Volatility of fractal and multifractal time series. **Israel Journal of Earth Sciences**, v. 56, n. 1, 2007. Citado na página 3.

KANTELHARDT, J. W. et al. Multifractal detrended fluctuation analysis of nonstationary time series. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 316, n. 1-4, p. 87–114, 2002. Citado 3 vezes nas páginas 3, 22 e 36.

KLEMM, K.; EGUILUZ, V. M. Growing scale-free networks with small-world behavior. **Physical Review E**, APS, v. 65, n. 5, p. 057102, 2002. Citado na página 16.

KNUTH, D. E. **The Art of Computer Programming, Volume 4, Fascicle 2: Generating All Tuples and Permutations (Art of Computer Programming)**. [S.l.]: Addison-Wesley Professional, 2005. ISBN 0201853930. Citado 2 vezes nas páginas xi e 57.

KRAUSE, S. M.; BÖRRIES, S.; BORNHOLDT, S. Econophysics of adaptive power markets: When a market does not dampen fluctuations but amplifies them. **Physical Review E**, APS, v. 92, n. 1, p. 012815, 2015. Citado na página 2.

KUMAR, R.; NOVAK, J.; TOMKINS, A. Structure and evolution of online social networks. In: **Link mining: models, algorithms, and applications**. [S.l.]: Springer, 2010. p. 337–357. Citado na página 2.

KWAPIEŃ, J.; DROŻDŻ, S. Physical approach to complex systems. **Physics Reports**, Elsevier, v. 515, n. 3-4, p. 115–226, 2012. Citado 2 vezes nas páginas 2 e 3.

KWAPIEN, J. et al. Are the contemporary financial fluctuations sooner converging to normal? **Acta Physica Polonica B**, v. 34, p. 4293, 2003. Citado 3 vezes nas páginas 3, 40 e 45.

LEBARON, B. Agent-based computational finance: Suggested readings and early research. **Journal of Economic Dynamics and Control**, Elsevier, v. 24, n. 5-7, p. 679–702, 2000. Citado na página 2.

LEBARON, B. Agent-based computational finance. **Handbook of computational economics**, Elsevier, v. 2, p. 1187–1233, 2006. Citado na página 2.

LEBARON, B.; ARTHUR, W. B.; PALMER, R. Time series properties of an artificial stock market. **Journal of Economic Dynamics and control**, Elsevier, v. 23, n. 9-10, p. 1487–1516, 1999. Citado na página 2.

LILJEROS, F. et al. The web of human sexual contacts. **Nature**, Nature Publishing Group, v. 411, n. 6840, p. 907, 2001. Citado 2 vezes nas páginas 1 e 12.

LIMA, L. Modeling of the financial market using the two-dimensional anisotropic ising model. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 482, p. 544–551, 2017. Citado na página 2.

LOU, J.; CHENG, T.; LI, S. High sensitivity photonic crystal fiber sensor based on dual-core coupling with circular lattice. **Optical Fiber Technology**, Elsevier, v. 48, p. 110–116, 2019. Citado na página 54.

LUX, T.; MARCHESI, M. Scaling and criticality in a stochastic multi-agent model of a financial market. **Nature**, Nature Publishing Group, v. 397, n. 6719, p. 498, 1999. Citado na página 2.

MAGALHÃES, M. N.; LIMA, A. C. P. D. **Noções de probabilidade e estatística**. [S.l.]: Editora da Universidade de São Paulo, 2002. v. 5. Citado 2 vezes nas páginas 3 e 28.

MANDELBROT, B. B. The variation of certain speculative prices. In: **Fractals and scaling in finance**. [S.l.]: Springer, 1997. p. 371–418. Citado na página 3.

MATSUMOTO, M.; NISHIMURA, T. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. **ACM Transactions on Modeling and Computer Simulation (TOMACS)**, ACM New York, NY, USA, v. 8, n. 1, p. 3–30, 1998. Citado 2 vezes nas páginas 21 e 59.

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

MITCHELL, M. **Complexity: A Guided Tour**. [S.l.]: OUP USA, 2009. ISBN 9780195124415. Citado na página 1.

MONTENEGRO, A.; RAGAB, R. Hydrological response of a brazilian semi-arid catchment to different land use and climate change scenarios: a modelling study. **Hydrological Processes**, Wiley Online Library, v. 24, n. 19, p. 2705–2723, 2010. Citado na página 1.

NAGATANI, T. The physics of traffic jams. **Reports on progress in physics**, IOP Publishing, v. 65, n. 9, p. 1331, 2002. Citado na página 12.

NAIR, N. U.; SANKARAN, P.; BALAKRISHNAN, N. Chapter 3 - discrete lifetime models. In: NAIR, N. U.; SANKARAN, P.; BALAKRISHNAN, N. (Ed.). **Reliability Modelling and Analysis in Discrete Time**. Boston: Academic Press, 2018. p. 107 – 173. ISBN 978-0-12-801913-9. Disponível em: ⟨http://www.sciencedirect.com/science/article/pii/B9780128019139000038⟩. Citado 2 vezes nas páginas 19 e 72.

NEWMAN, M. **Networks: an introduction**. 2nd edição. ed. [S.l.]: Oxford university press, 2018. Citado 11 vezes nas páginas 1, 3, 7, 12, 13, 14, 15, 54, 56, 66 e 70.

NEWMAN, M.; BARABASI, A.-L.; WATTS, D. J. **The structure and dynamics of networks**. [S.l.]: Princeton University Press, 2011. v. 12. Citado na página 11.

NEWMAN, M. E.; STROGATZ, S. H.; WATTS, D. J. Random graphs with arbitrary degree distributions and their applications. **Physical review E**, APS, v. 64, n. 2, p. 026118, 2001. Citado 2 vezes nas páginas 66 e 69.

PENG, C.-K. et al. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. **Chaos: An Interdisciplinary Journal of Nonlinear Science**, AIP, v. 5, n. 1, p. 82–87, 1995. Citado na página 35.

PRICE, D. d. S. A general theory of bibliometric and other cumulative advantage processes. **Journal of the Association for Information Science and Technology**, Wiley Online Library, v. 27, n. 5, p. 292–306, 1976. Citado 3 vezes nas páginas 11, 12 e 13.

REKIK, Y. M.; HACHICHA, W.; BOUJELBENE, Y. Agent-based modeling and investor's behavior explanation of asset price dynamics on artificial financial markets. **Procedia Economics and Finance**, Elsevier, v. 13, p. 30–46, 2014. Citado na página 2.

SHINCHI, T. et al. Fractal evaluations of fish school movements in simulations and real observations. **Artificial Life and Robotics**, Springer, v. 6, n. 1, p. 36–43, 2002. Citado na página 12.

SIMON, H. A. On a class of skew distribution functions. **Biometrika**, Oxford University Press, Biometrika Trust, v. 42, n. 3/4, p. 425–440, 1955. ISSN 00063444. Disponível em: ⟨http://www.jstor.org/stable/2333389⟩. Citado na página 11.

STEFAN, F.; ATMAN, A. Is there any connection between the network morphology and the fluctuations of the stock market index? **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 419, p. 630–641, 2015. Citado 2 vezes nas páginas 2 e 11.

TRAVERS, J.; MILGRAM, S. An experimental study of the small world problem. **Sociometry**, v. 32, n. 4, p. 425–443, 1969. Citado na página 1.

VINCIGUERRA, S.; FRENKEN, K.; VALENTE, M. The geography of internet infrastructure: an evolutionary simulation approach based on preferential attachment. **Urban Studies**, SAGE Publications Sage UK: London, England, v. 47, n. 9, p. 1969–1984, 2010. Citado 2 vezes nas páginas 1 e 16.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of "small-world" networks. **nature**, Nature Publishing Group, v. 393, n. 6684, p. 440, 1998. Citado 5 vezes nas páginas 1, 8, 34, 54 e 55.

WEI, J.; HUANG, J.; HUI, P. An agent-based model of stock markets incorporating momentum investors. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 392, n. 12, p. 2728–2735, 2013. Citado na página 2.

WEI, Y.-m. et al. The cellular automaton model of investment behavior in the stock market. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 325, n. 3-4, p. 507–516, 2003. Citado na página 2.

WENG, L. et al. Competition among memes in a world with limited attention. **Scientific reports**, Nature Publishing Group, v. 2, p. 335, 2012. Citado 7 vezes nas páginas 3, 11, 28, 29, 34, 44 e 45.

WIKIPEDIA.ORG. **Despejo do Wikipedia em Ingles**. 2021. Disponível em: ⟨http://dumps.wikimedia.your.org/enwiki/20210701/⟩. Citado 3 vezes nas páginas ix, 73 e 74.

XAVIER, P. O.; ATMAN, A.; MAGALHÃES, A. B. de. Equation-based model for the stock market. **Physical Review E**, APS, v. 96, n. 3, p. 032305, 2017. Citado na página 2.

ZIPF, G. **Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology**. [S.l.]: Ravenio Books, 2016. Citado 2 vezes nas páginas 19 e 72.

A Métodos de Criação de Redes

A.1 Modelo de Rede Circular Regular

O método de construção da rede circular regular (*circular lattice*) advém de um método proposto em Watts e Strogatz (1998). Por serem estruturas muito versáteis, redes circulares não são raras na ciência, como podemos exemplificar na área da óptica (HASAN et al., 2017), da engenharia elétrica (ISLAM et al., 2015) e em telecomunicações (LOU; CHENG; LI, 2019). Na área de redes aleatórias, até onde abrange nosso conhecimento, uma rede circular foi utilizada pela primeira vez no trabalho de Watts e Strogatz (WATTS; STROGATZ, 1998).

Em uma rede, a distância entre dois nós é a quantidade mínima de arestas que devem ser percorridas para que se vá de um nó a outro. A cada nó pode ser associada uma distância média A_i, que é a média das distâncias do mesmo a todos os outros nós da rede. Tal resultado é conhecido como *tamanho do caminho médio* (*average path length*), aqui denotado por $\mathcal{L}$ (NEWMAN, 2018). Na Tabela 3, as distâncias entre os nós da rede da Figura 21 são explicitadas, o que permite o cálculo de $\mathcal{L}$ ($\mathcal{L} = 2$ para esta rede). A

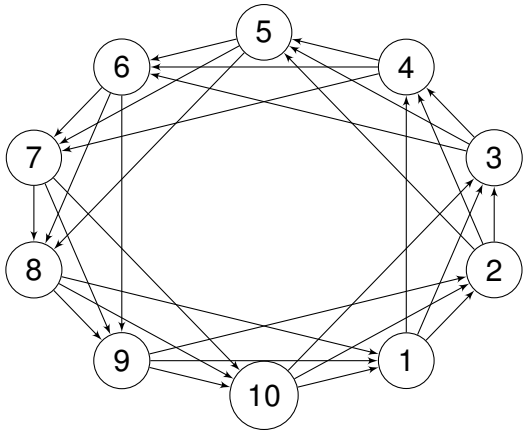


Figura 21 – Exemplo de rede circular regular com 10 nodos. Em cada nodo, há 3 arestas saindo e 3 arestas entrando.

Tabela 3 também nos permite computar o diâmetro da rede circular regular representada na Figura 21, que nada mais é do que a maior distância entre dois nodos A_i e A_j pertecentes à mesma: é fácil verificar que tal diâmetro é 3. Para entendermos o conceito de raio de um grafo, comecemos considerando, para cada nodo, a maior distância entre ele e outro nodo da rede, que é chamada de excentricidade. O raio é o menor valor de excentricidade dentre todos os encontrados na rede. A partir da Tabela 3, vemos que para todos os nodos a excentricidade vale 3, logo o raio é 3.

–	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
A_1	0	3	3	3	2	2	2	1	1	1
A_2	1	0	3	3	3	2	2	2	1	1
A_3	1	1	0	3	3	3	2	2	2	1
A_4	1	1	1	0	3	3	3	2	2	2
A_5	2	1	1	1	0	3	3	3	2	2
A_6	2	2	1	1	1	0	3	3	3	2
A_7	2	2	2	1	1	1	0	3	3	3
A_8	3	2	2	2	1	1	1	0	3	3
A_9	3	3	2	2	2	1	1	1	0	3
A_{10}	3	3	3	2	2	2	1	1	1	0

Tabela 3 – Distâncias mínimas de nó para nó na rede circular regular da Figura 21. A distância média associada a cada nó é 2. Dada a simetria da rede, este é o valor do tamanho do caminho médio da rede: $\mathcal{L} = 2$. É fácil constatar que tanto o diâmetro quanto o raio valem 3 para a rede analisada.

Para o desenvolvimento do modelo multi-agente descrito no Capítulo 5, consideramos vários tipos de rede. O algoritmo da rede circular regular apresenta flexibilidade no que se refere à escolha do grau dos nós. O diâmetro, no caso do modelo multi-agente que desenvolvemos, é relacionado ao número máximo de passos gastos para uma mensagem chegar, partindo de um nodo, a qualquer outro. Ele tem impacto importante na dinâmica de informação do modelo. O tamanho do caminho médio também tem grande relevância para a difusão de informação no modelo. A facilidade com que tais características da rede podem ser estabelecidas a partir do número de nós e de arestas entrando e saindo de cada nó é um fator que torna uma rede circular regular interessante na composição do modelo.

A construção da rede circular regular é descrita no Algoritmo 1. Este é um algoritmo iterativo, que adiciona arestas da direita para esquerda, no entanto respeitando uma condição de contorno que é imposta pelo número de elementos. O número de graus de cada nodo pode mudar o algoritmo no sentido de que, se for par, a adição das arestas deve seguir uma alternância entre os nodos, e não sequencialmente, como seria de se esperar. Mais detalhes sobre o porquê desta implementação podem ser encontrados em (WATTS; STROGATZ, 1998).

A.2 Modelo Erdös-Rényi

Uma vez que existem dois modelos Erdös-Rényi, o $G(n, M)$ e o $G(n, p)$, vamos separar esta seção em duas, dedicando uma subseção para tratar de cada um deles separadamente. Devido às suas peculiaridades, fica mais interessante abordarmos de maneiras separadas como esses modelos se comportam, como são construídos e onde nos deparamos com as margens das definições e com o que ainda está aberto.

Algoritmo 1: Algoritmo de criação da rede circular regular, tanto para valores de k pares, quanto ímpares. O valor k determina qual o grau que cada nodo terá.

Entrada: $G(N, E)$, a rede em que os nós serão iniciados, o número de nós é n, e k, o grau de regularidade.
Saída: $G(N, E)$ com n nós.

```
1  function CreateCircRegular(G(N,E),n,k)
2      AddNodes (G,n);
3      if k%2 then
4          for i = 0, i < c.n do
5              for j = 1, j <= c.k/2 + i%2 do
6                  d = (i + j)%n;;
7                  AddEdge (i, d);              // Adiciona a aresta e_{i,d} a E.
8                  j = j + 1;
9              end
10             i = i + 1
11         end
12     end
13     else
14         for i = 0, i < c.n do
15             for j = 1, j <= c.k/2 do
16                 d = (i + j)%n;;
17                 AddEdge (i, d);              // Adiciona a aresta e_{i,d} a E.
18                 j = j + 1;
19             end
20             i = i + 1;
21         end
22     end
```

A.2.1 Erdös-Rényi $G(n, M)$

O modelo Erdös-Rényi (ER), descrito na Seção 2.3, foi uma das primeiras propostas de redes aleatórias (ERDÖS; RÉNYI, 1959). Primeiramente tomamos a definição de rede dada na Seção 2.1. Consideremos os dois modelos existentes, o modelo $G(n, M)$, onde escolhe-se o número de nodos n e o número de arestas desejadas M para o grafo G, e o modelo $G(n, p)$, onde é escolhido o tamanho do grafo n e a probabilidade com a qual os nós irão se ligar p, no grafo G. Na Seção 2.3, foram descritas muitas características sobre os grafos ER em geral; vamos reportar, aqui, pontos relevantes, alguns deles, citados na seção mencionada. Os dois grafos dos modelos de Erdõs-Rényi são também conhecidos como grafos de Poisson. São assim chamados porque, erroneamente, pensou-se que um grafo aleatório teria distribuição uniforme. No entanto, isso não ocorre, conforme bem descrito em Newman (2018, pp.347), demonstrando que a distribuição do modelo $G(n, p)$ é uma distribuição de Poisson. Aqui, iremos investigar se o mesmo ocorre para o modelo $G(n, M)$, como é dito comumente na literatura.

Algoritmo 2: Algoritmo de Knuth (KNUTH, 2005) para enumerar todas a combinações de uma combinação qualquer $\binom{n}{k}$ (deve-se levar em consideração a questão da viabilidade de se enumerar e guardar todas essas combinações em $\mathcal{A}$ na memória de um computador).

Entrada: n e k do combinatorial $\binom{n}{k}$ e uma estrutura de dados que é um conjunto de arestas $\mathcal{A}$

Saída: $\mathcal{A}$ com todas as combinações enumeradas.

```
1  function EnumCombinations(n,k)
2      i, j = 1, c[k + 3], x;
3      for i = 1, ..., k do
4          c[i] = i;
5      end
6      c[k + 1] = n + 1;
7      c[k + 2] = 0;
8      j = k;
9      visit:// Este um label ponto de desvio de fluxo de código.;
10     PushArray(A,e_{c[0],c[1]})// Copia c e armazena em A;
11     if j > 0 then
12         x = j + 1;
13         goto incr// Desvia o fluxo de código para o label incr.;
14     end
15     if c[1] + 1 < c[2] then
16         c[1]+ = 1;
17         goto visit// Desvia o fluxo de código para o label visit.;
18     end
19     j = 2;
20     domore:// Este um label ponto de desvio de fluxo de código.;
21     c[j − 1] = j − 1;
22     x = c[j] + 1;
23     if x == c[j + 1] then
24         j + +;
25         goto domore// Desvia o fluxo de código para o label
             domore.;
26     end
27     if j > k then
28         return A;
29     end
30     incr:// Este um label ponto de desvio de fluxo de código.;
31     c[j] = x;
32     j − −;
33     goto visit// Desvia o fluxo de código para o label visit.;
```

Comecemos com o algoritmo de geração de combinações de Knuth (KNUTH, 2005). No Algoritmo 2, apresentamos o método para construção do conjunto de combinações $\mathcal{A}$ a partir dos parâmetros n e $k = 2$, de forma que cada elemento de $\mathcal{A}$ possa ser associado a uma aresta $e_{a,b}$. Então, como queremos gerar todas as combinações de grafos com M

arestas, isto nos leva a combinar $\binom{\mathcal{A}}{M}$ possibilidades. Esta última combinação, no entanto, gera números astronômicos. Portanto, tomamos a decisão de seguir por um algoritmo que tem a seguinte lógica: criamos um grafo vazio de arestas $G(N, E)$, sorteamos iterativamente arestas de $\mathcal{A}$ e cada aresta sorteada é adicionada a E e, enquanto $|E| \ngeq M$ e todos os nodos não tiverem sido sorteados, o algoritmo não para (ver Algoritmo 3).

Algoritmo 3: Construção do grafo $g(n, M)$ através de iterações sucessivas. Arestas vão sendo adicionadas ao grafo até quando a condição na linha 14 for verdadeira para terminar.

Entrada: Grafo $G(N, E)$, com $E = \{\}$, o vetor resultante de $\binom{N}{2} = \mathcal{A}$ e o número M de arestas. Vamos considerar, durante o algoritmo, $|\mathcal{A}|$ como a cardinalidade deste conjunto.

Saída: Grafo $G(N, E)$ com aproximadamente M arestas.

```
1  function ConstructGNM(G(N,E),A,M)
2      Nl = {}// Lista de nodos inicializada como vazia;
3      while True do
4          l =rand-i(|A|)// Seleciona um índice aleatório entre
             [0, |A| − 1];
5          p =getFrom(i,A)// Pega o i-ésimo elemento de A;
6          if AddEdge(E,e_{p0,p1}) then // Adiciona aresta entre o par
             ordenado p.
7              if p0 ∉ Nl then
8                  pushNode(p0,Nl);
9              end
10             if p1 ∉ Nl then
11                 pushNode(p1,Nl);
12             end
13         end
14         if |E| ≥ M AND |Nl| ≥ |N| then ;        // Condição de parada do
             algoritmo.
15
16             return G(N,E)
17         end
18     end
```

O Algoritmo 3 é uma solução para o problema do modelo $G(n, M)$, ela é rápida, simples e não impede que sorteios não sejam utilizados. Uma alternativa para essa implementação, que seria um pouco mais restritiva em termos de aleatoriedade, seria se, quando o conjunto E chegasse a M arestas, fossem aceitos sorteios que só contenham nodos que não foram descobertos ainda. A nossa abordagem pode ser, por assim dizer, gulosa, no sentido do número de arestas que são inseridas no grafo $G(N, E)$. No entanto, se formos comparar com outras abordagens, que façam um sorteio por iteração e ainda assim queiram não adicionar uma aresta por iteração, vale observar como a função `AddEdge(`e`,`E`)` retornará verdadeiro se todos os sorteios de `rand-i(`$|\mathcal{A}|$`)` forem distintos. Mas isso não é sempre verdade, então a função `AddEdge(`e`,`E`)` só retornará falso quando `rand-i(`$|\mathcal{A}|$`)`

repetir um sorteio. Observemos que a probabilidade de sorteios idênticos segue igual, dada uma execução do já citado Algoritmo 3.

Além das questões levadas em consideração até o momento, não podemos deixar de falar da eficiência do gerador aleatório. Como já dito neste trabalho, usamos o Mersenne Twister (MATSUMOTO; NISHIMURA, 1998), que tem o período $2^{19937} - 1$ e retorna inteiros de 32 de bits. Não é preciso elaborar muito a análise, basta supor que o algoritmo seja capaz de retornar os inteiros como uma sequencia de números que não se repete, completamente aleatória. Seria uma longa série, mas, no fim, após o período do algoritmo, seria só o início de uma nova série. Além do mais, a repetição não excessiva dos números dentro de uma série pseudo-aleatória a torna mais real em termos de uma série completamente aleatória real. Alguns eventos tendem a se repetir, seja por sorte, seja por insistência, mas um gerador aleatório, seja com baixo período ou alto, deve ter *dentro dele* a capacidade de repetir números sorteados no passado curto de vez em quando.

Para termos um indício da dificuldade dessa repetição acontecer, fizemos um teste de 10000 N e 20000 E, um grafo bem esparso. Queríamos somente testar as qualidades estatísticas da distribuição de graus. Fizemos mil amostras de grafos, com o valor de $\|\mathcal{A}\| = 49995000$, este era o limite L que o gerador aleatório tinha para fazer seus sorteios $L = [0, 49995000 - 1]$. Porém, para sabermos qual a ordem de grandeza da probabilidade de se escolher um elemento usando esses dados, podemos fazer o seguinte cálculo:

$$\log_{10}\left(\frac{1}{49995000}\right) = -7.6989. \tag{39}$$

Observemos que a ordem de grandeza da probabilidade de se sortear uma aresta de $\mathcal{A}$ é $10^{-7.6989}$. No experimento, fizemos 1000 grafos como amostras e, na média, temos 4.8838×10^{-4} arestas por amostra, o que acarreta uma probabilidade final de $4.8838 \times 10^{-4} \times 10^{-7.6989} = 9.769 \times 10^{-4}$. Desta maneira, com o número obtido da probabilidade final, podemos afirmar que a probabilidade de sortear duas arestas em uma mesma amostra é da grandeza de 100 milionésimos. Tal probabilidade é ínfima diante das grandezas do próprio problema. Caso a repetição ocorra, a aresta não é adicionada e um novo sorteio é realizado no próximo laço, e o algoritmo continua normalmente.

Para este modelo, realizamos alguns testes. Um, envolvendo uma quantidade máxima de arestas dada pelo combinatorial $\binom{10000}{2}$, é refletido na Figura 22a; a rede é extremamente esparsa, mas consegue replicar a distribuição de Poisson. Três, envolvendo a quantidade máxima de arestas dada por $\binom{3969}{2}$, com os seguintes valores de M, $M = 2000$, $M = 20000$ e $M = 200000$, também foram efetuados. É possível perceber que nesta parte do estudo as redes se mantêm esparsas (observe as figuras 22b e 22c). O mais interessante é a análise da rede que não é esparsa (Figura 22d). Percebemos que a curva, apesar de ter a forma de uma curva de Poisson, por muito pouco não ficou sendo um processo gaussiano. Isto ocorre por causa da propriedade advinda do modelo $G(n, p)$ (ver Seção 2.3),

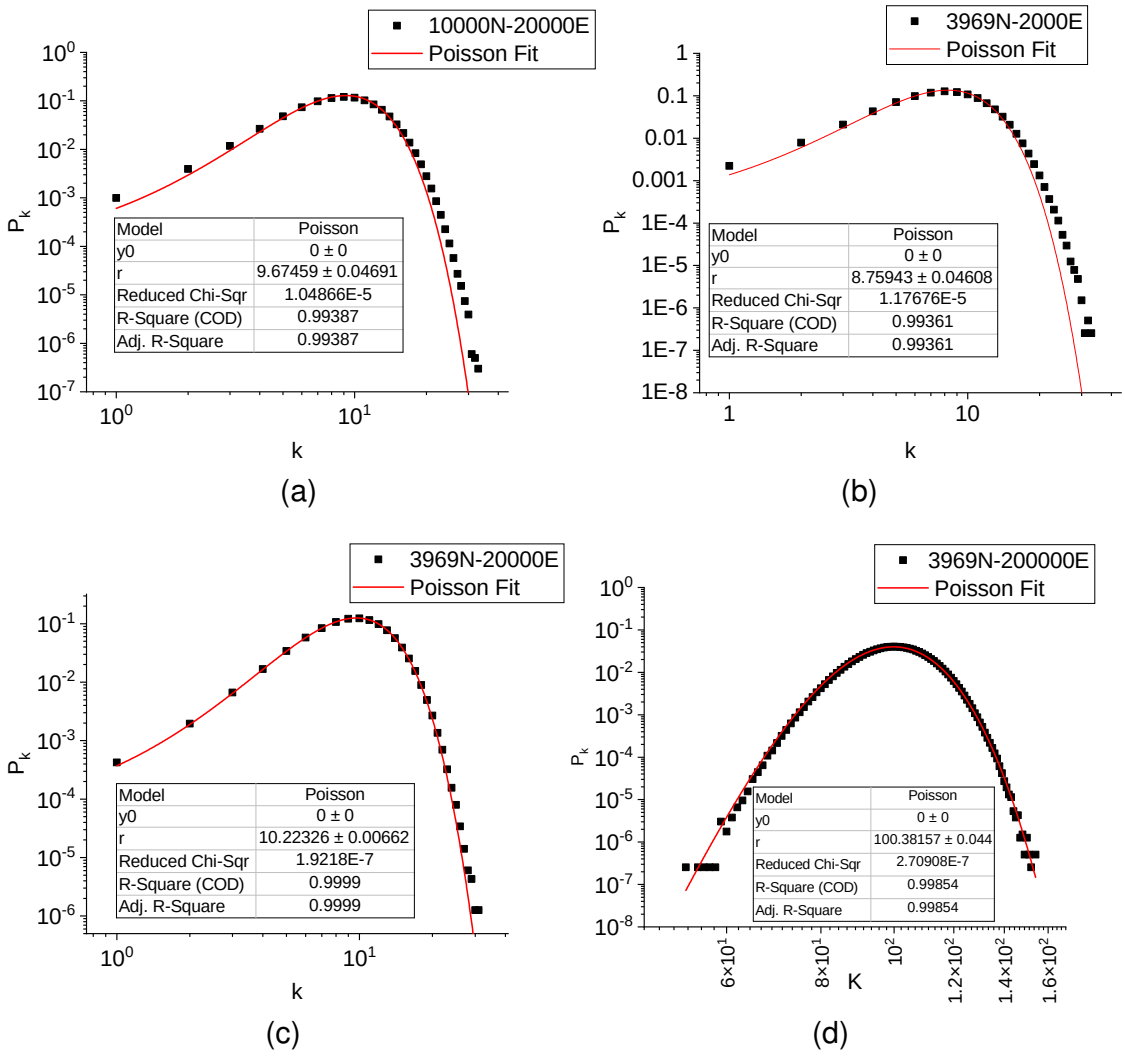


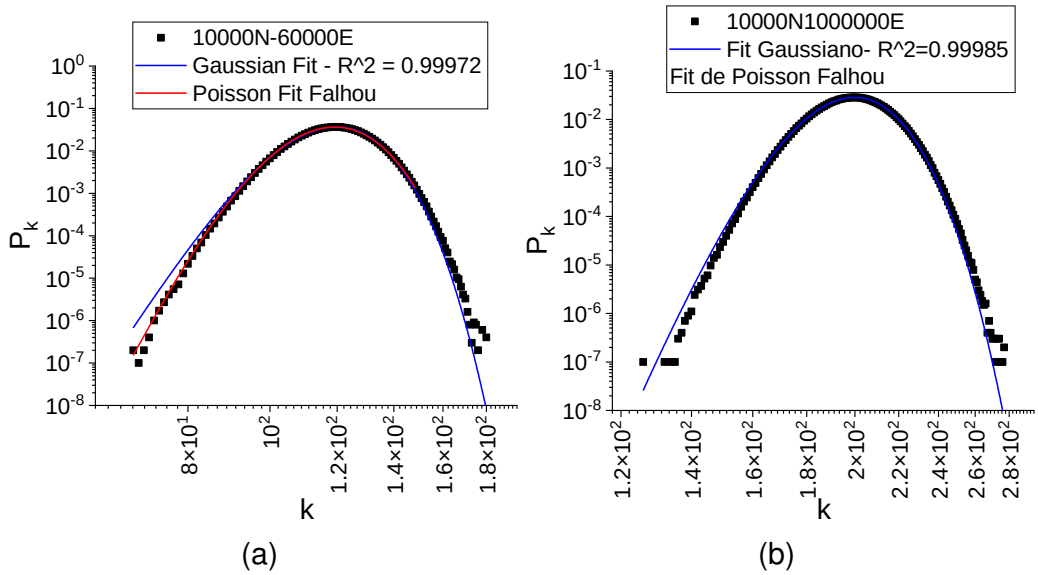
Figura 22 – Resultados dos experimentos feitos com o algoritmo $G(n, M)$, cada figura é o resultado da média de mil amostras.

que também é aplicável aqui:

$$p > \frac{(1+\epsilon)\ln|N|}{|N|}. \tag{40}$$

Tal propriedade diz que, quando p for maior que a expressão dada na Equação 40, o grafo será conexo. Pois bem, $p = \frac{|N|}{\langle|E|\rangle}$, onde o denominador nada mais é que a média de graus por nodo, considerando que nossos grafos são direcionados; então $\langle|E|\rangle = \frac{|E|}{|N|}$. Agora percebamos que este valor possui duas vertentes, a de uma constante onde você pode fixar um valor p para o grafo em questão, e a segunda vertente, dinâmica, na qual, de acordo com as modificações feitas no grafo, p mudará. Então ele passa a não ser uma constante, ele se modifica de acordo com a evolução do grafo.

Tangendo o problema de grafos cujas distribuições de graus se aproximam de uma gaussiana, podemos explicar a partir de $p = \frac{|N|^2}{|E|}$. Note que, se o número de arestas subir, o valor p diminuirá, o que gera um relação de proporção inversa entre p e $|E|$. Isso faz com



(a) (b)

Figura 23 – Demonstração da tendência gaussiana com o crescimento da média de graus, advindo da média de mil amostras.

que, se o grafo crescer, inserir arestas torne-se um processo cada vez mais *difícil*. Refletindo mais cuidadosamente sobre o assunto, observemos que, num processo estocástico como o Algoritmo 3, a cada aresta adicionada deixamos o processo ser executado um número imenso de vezes, crescendo um grafo que possua número máximo de nodos condizentes (quando dizemos *condizentes*, estamos nos referindo a um grafo cujo número de arestas, $|N|(|N|-1)$, seja bem superior a esse crescimento proposto). O que vai acontecer é que todos os nodos obterão valores de grau alto, e, com essas grandezas de valores para os nós, com o deslocamento para a direita, é difícil distinguir a distribuição de Poisson existente de uma distribuição Gaussiana, que fica mais evidenciada. Isso é ilustrado na Figura 23.

A.2.2 Erdös-Renyi $G(n,p)$

Este modelo de Erdös-Rényi é o mais famoso dos modelos de redes aleatórias. Uma curiosidade: foi este modelo que permitiu a existência do número de Erdös, que é a distância, num grafo de coautoria de publicações, até o autor Paul Erdös. Este modelo foi muito estudado sob diversos vieses, a obra de Bollobás (2001) demonstra isso veementemente. Como é um modelo cujas informações mais importantes já foram dadas na Seção 2.3, vamos, aqui, mostrar quatro exemplos onde variamos a probabilidade em quatro ordens de grandeza, calculando a distribuição de graus para cada um deles (ver Figura 24). Exibiremos, aqui, a maneira mais simples de se implemetar $G(n,p)$. No Algoritmo 4, pode-se perceber uma otimização, pois ele consegue gerar todos os sorteios de um grafo direcionado sem fazer a varredura quadrática, que seria $\sum_{i=0}^{|N|}\sum_{j=0}^{|N|}$. Nesse algoritmo, nós realizamos uma leitura triangular que, apesar de ser considerada $O(n^2)$, na realidade apresenta-se como uma pequena redução de cálculos, que em alguns casos é bem vinda, dependendo da

Algoritmo 4: Algoritmo de implementação do $G(n,p)$ direcionado. A cada iteração são feitos dois sorteios e por consequência dois testes, um considerando a aresta em um sentido e o outro considerando a aresta em outro sentido.

Entrada: Grafo $G(N,E)$, com $E = \{\}$, probabilidade p
Saída: Grafo $G(N,E)$ com média de arestas $p|N|$.

```
1  function ConstructGnp(G(N,E),p)
2      for i = 0,...,|N| − 1 do
3          for j = i + 1,...,|N| − 1 do
4              a0=0to1rand();
5              if a0 < p then
6                  AddEdge(E,e_{i,j});
7              end
8              a1=0to1rand();
9              if a1 < p then
10                 AddEdge(E,e_{j,i});
11             end
12         end
13     end
```

escala da rede que se deseja montar. Esse algoritmo tem a mesma característica que o modelo $G(n,M)$ na questão da escalabilidade. Quando fazemos grafos muitos densos, eles tendem a ter distribuição de graus gaussiana. É importante notar, no entanto, que os grafos gerados por esse modelo aproximam-se muito mais de uma distribuição de Poisson do que os produzidos pelo Algoritmo 3 (ver Figura 24).

A.3 Métodos de Criação de Grafos em Lei de Potência

A.3.1 Método de Adição de Nós (Node Addition Method)

Um dos métodos mais antigos, criado para ser uma modelo simples, é o modelo de adição de nós (BARABÁSI; ALBERT, 1999). Nele, através da adição de nós e com o ganho de links com essa dinâmica incremental de um nó por vez, aqueles que possuírem maior grau possuem maior probabilidade de receber os links dos novos nodos que estão sendo inseridos. A metodologia desta abordagem começa com uma rede $G(N,E)$, conforme definida na Seção 2.1, com m nodos iniciais completamente conectados. Vamos, por exemplo, supor que $m = 3$. Dada esta rede G, ela teria a lista de graus $D = \{2,2,2\}$. Quando adicionamos um quarto nó, temos uma lista de graus com algo do tipo $D = \{2,3,2,1\}$. Notemos que o total de graus anterior era 6 e agora é 8. Nesta nova configuração que se formou, podemos considerar que a chance de um nó conseguir um novo link está ligada à razão entre o grau de cada nó e a soma de todos os graus, $\frac{d_i}{n}$, onde d_i é o grau do nodo i, dado pelo i-ézimo elemento em D, e n é a soma dos elementos de D. Se assim

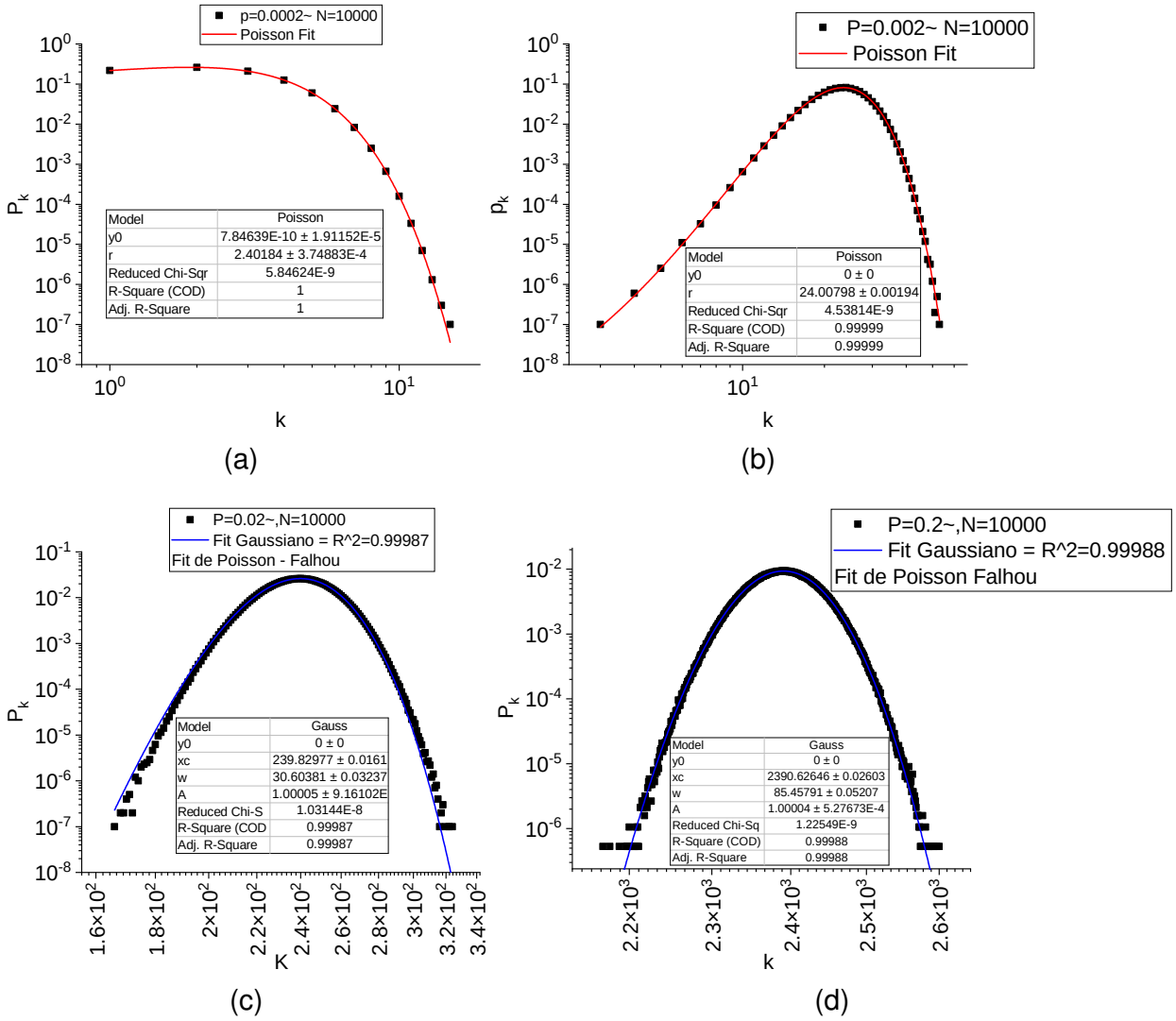


Figura 24 – Modelo $G(n,p)$ para $N = 10000$. Na Figura 24a, $p \approx 0.0002$, na Figura 24b, $p \approx 0.002$. Note que a densidade vai aumentando e naturalmente a média vai se movendo para a direita. O efeito da *gaussianização* ocorre aqui também. Isso pode ser observado nas figuras com probabilidades nos valores $p \approx 0.02$, 24c, e com $p \approx 0.2$, 24d.

o fizermos, teremos o conjunto de probabilidades $P_d = \{0.25, 0.375, 0.25, 0.125\}$. A partir de P_d, os nós, ao invés de terem probabilidades uniformes de sorteio, como descrito na Seção A.2, passam a ter probabilidades baseadas na quantidade de graus que possuem, o que caracteriza o *preferential attachment*.

Em 1999, os autores de Barabási e Albert (1999) tinham consciência da importância do trabalho que faziam, mas fica a questão de se eles sabiam do impacto que o trabalho causaria no mundo inteiro. O Artigo Barabási e Albert (1999) é hoje um dos trabalhos mais citados do mundo. O algoritmo desse processo não tem definição tão rígida a ponto de todas as implementações precisarem de ser iguais, mas, como vamos apontar, há dois pontos em que elas não podem se diferenciar. O primeiro: o parâmetro m, que representa

o número inicial de nós, define a maneira como a distribuição de nós irá se desenvolver, como se fosse uma impressão digital deixada no rastro da execução do algoritmo. O segundo: a probabilidade do novo nó inserido no grafo a um nó i pertencente ao grafo $G(N, E)$ deve depender da probabilidade demonstrada na linha linha 14 do Algoritmo 5. Pode-se considerar quase uma diretiva do Algoritmo 5 que o mesmo é de estados temporais assíncronos. Isso significa que os estados do algoritmo modificam a rede a cada nó inserido, mas cada inserção não depende das inserções anteriores e nem dependerá das inserções posteriores. Para comprovar isso, basta notar que, apesar das inserções modificarem a rede em termos de estarem adicionando arestas a essa rede, essas modificações não geram nenhum tipo de condicionamento para que as próximas inserções ocorram. Os autores do artigo Barabási e Albert (1999) relevam o fato de que o modelo não precisa ter n definido para que o mesmo possa funcionar. Se $n \to \infty$, podemos executar o algoritmo e tirar cópias do estado da rede $G(N, E)$ nos instantes desejados, ou seja, podemos executar o algoritmo e, se quisermos, podemos obter o estado da rede a cada potência 10^{p_i}, com os expoentes p_i no conjunto $p = \{3, 4, 5, 6, 7, \dots\}$.

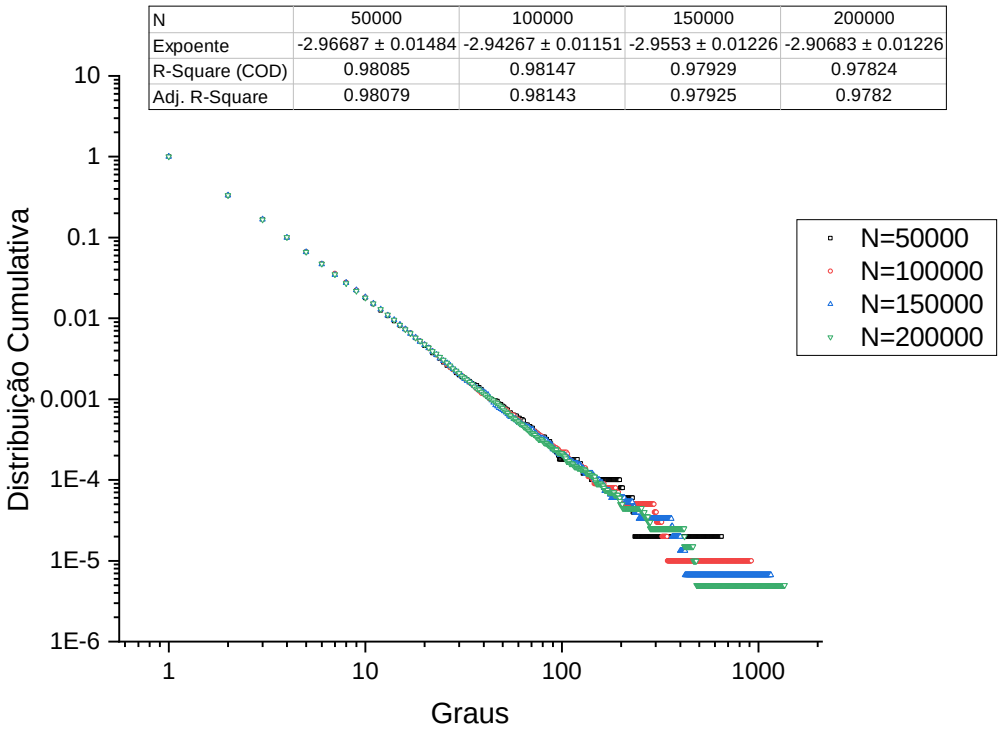


Figura 25 – Distribuição de graus do grafo $G(N, E)$ para quatro tamanhos do sistema diferentes. Note que o modelo de Barabási-Albert, quando há *preferential attachment*, fica invariante a mudanças de tamanho. É apresentada a distribuição acumulada inversa, relativa aos graus dos nodos, em escala $log \times log$, para facilitar a visualização do comportamento em questão.

O Algoritmo 5 é base para duas análises importantes sobre o *preferential attachment*. Neste trabalho, replicamos uma delas (ver discussão em torno da Figura 25). Note que o Algoritmo 5 é invariante em relação ao tamanho da rede, isto porque, como dito antes, ele

Algoritmo 5: Algoritmo que exemplifica o modelo de Barabási e Albert (1999), onde as entradas são um rede $G(N, E)$, a quantidade m de nós iniciais ($m > 0$) e a quantidade de nós a serem inseridos, n. Este algoritmo constrói uma rede cujo valor da inclinação da lei de potência relativa à distribuição dos graus dos nodos é constante, independente do valor inicial de m.

Entrada: $G(N, E)$, a rede em que os nós serão iniciados, $m > 0$, o número de nós inicial e o número de nós a serem adicionados, n.
Saída: $G(N, E)$ com n nós

```
1  function PowerLawByNodeAddition(G(N,E),m,n)
2      if |N| == 0 then
3          GenerateComplete(G,m)// G(N,E) completo com m nós.;
4      end
5      ε = |E|// ε é a quantidade de graus gerados pelas arestas;
6      while |N| < n do
7          N = N ∪ {i + 1};
8          i = |N| − 1;
9          j = i;
10         while i == j do
11             ψ=0to1rand();
12             α=randInt(|N| − 2);
13             H=Degree(α);
14             p = H/ε // Probabilidade para conectar com i.;
15             if p > ψ then
16                 if AddEdge(E,α,i) then //  Caso α ≠ i , e_{α,i} ∉ E seja
                    Verdadeiro tem-se E = E ∪ {e_{α,i}} e retorna Verdadeiro,
                    se não retorna Falso;
17                     i = i + 1;
18                     ε = ε + 2;
19                 end
20             end
21         end
22     end
23     return G(N,E);
```

é um algoritmo em que o parâmetro n não precisa estar definido de antemão. A limitação relacionada à determinação do parâmetro n é de âmbito puramente computacional, no entanto, matematicamente, pode-se gerar uma equação de recursão onde não precisamos colocar um ponto de parada:

$$G_0(E_0, N_0) \rightarrow N = \{n_0\}, E_0 = \{\}, \tag{41}$$

$$G_1(N_1, E_1) = G_0(N_0 \cup \{n_1\}, E_0 \cup \{e\}), \tag{42}$$

$$\dots$$

$$G_i(N_i, E_i) = G_{i-1}(N_{i-1} \cup \{n_i\}, E_{i-1} \cup \{e\}). \tag{43}$$

Nas equações 42 e 43, é importante notar que o conjunto E sempre ganha uma nova aresta e o conjunto N sempre ganha um novo elemento, aumentando em uma unidade o tamanho do sistema. Os operadores de união denotados por $\cup$ somente serão concretizados se os nós escolhidos na aresta e respeitarem a lei de *preferential attachment*.

O Algoritmo 5 possui independência da entrada. Por este motivo, irá sempre gerar grafos com a mesma distribuição de graus em lei de potência com o expoente de decaimento $p_k \approx k^{-3}$, ver Figura 25.

A.3.2 Modelo de Configuração (*Configuration Model*)

O Modelo de configuração é um método de construção de grafos considerado atualmente uma das ferramentas em redes complexas mais completas em termos de construção de redes. Em Newman (2018), temos uma abordagem sobre o assunto de Modelo de Configuração que pode ser considerada um texto guia para leitores de todos os níveis. Primeiramente é importante diferenciar dois parâmetros que em alguns casos podem se sobrepor: *distribuição de graus* e *sequência de graus*. Note que a distribuição de graus está vinculada a características estatísticas de excentricidade de grau dos nós em uma rede; já sequência de graus, seja qual for a qualidade que se atribua a ela, sempre será uma sequência finita de números inteiros positivos. Vale notar que, se a sequência for infinita, no nosso contexto não podemos considerá-la sequência, uma vez que seu último termo é desconhecido.

O método de construção de grafos que descrevemos aqui apareceu na literatura com mais destaque em Newman, Strogatz e Watts (2001), onde os autores, com muita didática, explicam como geraram redes de milhões de nós com a tecnologia da época.[1] Nesse artigo, os autores abordam o problema de gerar redes que tenham como produto uma distribuição de graus determinada arbitrariamente seguindo as características de uma dada distribuição probabilística $\mathcal{D}$. O fator especial do modelo de configuração vem exatamente dessa definição abrangente, porque na estatística a maioria das distribuições surgiu com o fim de descrever um fenômeno existente que é frequente ou na vida das pessoas ou nos estudos científicos em si. Apesar de haver diversos tipos de distribuição, existem dois tipos de função que servem para definir a informação probabilística das distribuições, uma delas é a função de probabilidade, quando $\mathcal{D}$ é discreto, e a outra é a função de densidade de probabilidade, quando $\mathcal{D}$ é contínuo.

Neste trabalho, vamos tratar somente o caso da função de probabilidade discreta f, o único tipo que usaremos aqui. Tal função recebe inteiros e retorna uma probabilidade daquele inteiro ocorrer. Para construir uma distribuição d de parte do intervalo da distribuição discreta $\mathcal{D}$, é importante delimitarmos um intervalo $[i, j]$ onde teremos d_k como o valor da probabilidade do inteiro k ($i \leq k \leq j$) ocorrer em $\mathcal{D}$ no intervalo $[i, j]$. Desta forma,

[1] Os autores, doutores em Física, comprovam que muitos problemas computacionais ainda hoje estão mais próximos da Física do que da Ciência da Computação.

quando tal distribuição for utilizada na construção de redes, ao escolhermos $[i, j]$, estamos definindo o grau mínimo e máximo que a rede pode ter como i e j, respectivamente. Assim, criaremos um vetor com elementos correspondentes a d_k para todos os valores de k em $[i, j]$, $\mathcal{V} = \{d_i, d_{i+1}, \ldots, d_{j-1}, d_j\}$, com $j - i + 1$ componentes, atribuindo d_i para cada posição existente do vetor: $\mathcal{V}_i = d_i$. Importante lembrar que em princípio desconhecemos a distribuição, e que alguns valores de d_i no intervalo definido podem assumir valores $\pm\infty$ ou nan (não é um número). Quando isso acontece, não é possível seguir com o processo.

Um algoritmo para o preenchimento deste vetor funcionaria da forma como está descrita no Algoritmo 6. As funções de probabilidade são maneiras de extrair dados de uma

Algoritmo 6: Criação do um vetor de elementos normalizado e acumulado $\mathcal{V}$ retirado de uma amostragem da distribuição $\mathcal{D}$, d no intervalo $[i, j]$ onde i e j respeitam os limites de $\mathcal{D}$. Com este vetor de elementos $\mathcal{V}$, é possível fazer operações de sorteio não uniforme para uma distribuição qualquer $\mathcal{D}$, mesmo que esta seja linear.

Entrada: Distribuição intervalar d no intervalo $[i, j]$.
Saída : Vetor $\mathcal{V}$ de probabilidades acumuladas pronto para ser usado pelo Modelo de Configuração.

```
1  function CreatingVectorV(d,[i, j])
3      for k = i, ..., j,l = 0, ..., j − i do
5          v = d_k;
7          if IsNan (v) OR IsInf (v) then
9              V_l=0;
11
12         end
13         else
15             V_l = v;
16         end
17     end
19     V = V / (Σ_k^{j−i} V_k) // Divisão de cada elemento de V pela soma total.;
21     for k = 0, ..., j − i do
23         V_k = V_{k−1} + V_k;
24     end
26     return V;
```

distribuição. Como d_i é uma probabilidade, temos que,

$$d_i = f_{\mathcal{D},i}, \qquad i \in \mathcal{I}, \tag{44}$$

onde $f_{\mathcal{D},i}$ é a probabilidade de ocorrência de i na distribuição de probabilidade $\mathcal{D}$, e $\mathcal{I}$ é o intervalo finito ao qual i pertence. O intervalo deve ser finito porque não podemos representar intervalos infinitos no computador. Para dar um exemplo de como é o resultado

final do vetor $\mathcal{V}$, usemos a distribuição exponencial,

$$p_{x,\lambda} = \int_{x}^{\infty} \lambda \exp^{-\lambda x}, \tag{45}$$

onde a função exponencial é definida no limite $[0, \infty)$ e λ é conhecida como a taxa de decaimento. Para nosso exemplo, vamos usar $\lambda = 0.5$, assim:

$$P_{x,0.5} = \int_{x}^{\inf} 0.5 \exp^{-0.5x}. \tag{46}$$

Tomando o intervalo $[i, j] = [1, 10]$, temos os resultados na Tabela 4. Logo, a partir da coluna

x	$\mathcal{V}_x = P_{x,0.5}$	$\frac{\mathcal{V}}{\sum_k^{10} \mathcal{V}_k}$	$\mathcal{V}$ Acumulado
1	0.30327	0.39614	0.39614
2	0.18394	0.24027	0.63641
3	0.11157	0.14573	0.78214
4	0.06767	0.08839	0.87053
5	0.04104	0.05361	0.92414
6	0.02489	0.03252	0.95666
7	0.0151	0.01972	0.97638
8	0.00916	0.01196	0.98834
9	0.00555	0.00726	0.9956
10	0.00337	0.0044	1

Tabela 4 – Etapas do processo de construção do vetor de elementos $\mathcal{V}$.

quatro da Tabela 4 os sorteios de números inteiros do intervalo escolhido passam a ser feitos através de sorteios uniformes independentes e identicamente distribuídos (iid) no intervalo $[0, 1)$, relativos à variável aleatória u_t. Note que temos exatos 10 intervalos definidos a partir da coluna 4 da Tabela 4 e que estão sendo explicitados na Tabela 5. É importante observar que entre cada intervalo existe um valor x pertencente ao intervalo anteriormente escolhido, que era de $[1, 10]$. Por exemplo, de acordo com a Tabela 5, observamos que $x = 1$ corresponde ao intervalo $[0, 0.39]$, $x = 2$ ao intervalo $[0.39, 0.64]$, e assim por diante. Para

–	x	$\mathcal{V}_1$	x	$\mathcal{V}_2$	x	$\mathcal{V}_3$	x	$\mathcal{V}_4$	x	$\mathcal{V}_5$	x	$\mathcal{V}_6$	x	$\mathcal{V}_7$	x	$\mathcal{V}_8$	x	$\mathcal{V}_9$	x	$\mathcal{V}_{10}$
0	1	0.39	2	0.64	3	0.78	4	0.87	5	0.92	6	0.95	7	0.97	8	0.988	9	0.995	10	1

Tabela 5 – Intervalos em que os sorteios irão ocorrer, perceba que cada x está entre dois valores.

que possamos sortear estes valores x de maneira não uniforme existem muitas abordagens, no entanto, neste trabalho escolhemos a abordagem do Algoritmo 7.

Algoritmo 7: Algoritmo de sorteio de números inteiros não uniforme.

Entrada: O vetor $\mathcal{V}$ acumulado, um sorteio u_t e o tamanho do vetor de elementos $\mathcal{V}$, n.
Saída : Número inteiro distribuído de maneira exponencial.

```
1  function SorteioAleatorioNU(V,u_t,n)
2      aux = 0;
3      i = 1;
4      while i ≤ n do
5          aux = aux + V_i;
6          if aux ≥ u_t then
7              return i;
8          end
9          i = i + 1;
10     end
```

A.3.2.1 O Cerne do Modelo de Configuração

Primeiramente, vamos utilizar as definições de redes já feitas na Seção 2.1 e considerar que queremos um grafo direcionado $G(N, E)$ de tamanho L que, por enquanto, tem zero arestas no seu conjunto, $E = \{\}$, e $N = \{a_1, a_2, \ldots, a_{L-1}, a_L\}$. Vamos usar uma distribuição $\mathcal{D}$ no intervalo $[1, L]$ e construir o vetor $\mathcal{V}$ de acordo com o Algoritmo 6. Depois desse passo, estaremos fazendo as seguintes considerações: temos um vetor de elementos $\mathcal{V}$; o vetor só se refere à Distribuição $\mathcal{D}$ no intervalo dado, $[1, L]$. O algoritmo segue fazendo uma sequência de sorteios cujos resultados s_i são utilizados na composição de pares ordenados (a_i, s_i), onde os a_i correspondem aos nós da rede. Como foi explicado nas tabelas 4 e 5, s_i só assume valores inteiros e, neste caso, define o grau que o nó a_i terá. Logo, o par ordenado (a_i, s_i) significa que o i-ésimo elemento da rede, a_i, vai receber o grau s_i. Em Newman, Strogatz e Watts (2001), este processo é chamado de criação de *stubs*, e a construção da rede no algoritmo dos autores corresponde a resolver o problema de conectar todos os stubs até que cada nó tenha atingido o grau desejado. Aqui, vamos demonstrar uma técnica inspirada na proposta feita em Newman, Strogatz e Watts (2001), onde a solução desse problema torna-se trivial. A partir do conjunto de pares ordenados de índices e seus correspondentes graus de nós, $\mathcal{S} = \{(a_1, s_1), (a_2, s_2), \ldots, (a_{L-1}, s_{L-1}), (a_L, s_L)\}$, criamos o conjunto S onde cada elemento a_i aparece s_i vezes, conforme o Algoritmo 8. Usando o conjunto S, coletamos aleatoriamente seus elementos para a geração das arestas, definindo, desta forma, cada e_{S_i,S_j}, se $S_i \neq S_j$ e $e_{S_i,S_j} \notin E$; então, $E \leftarrow E \cup \{e_{S_i,S_j}\}$ (ver Algoritmo 9).

A partir do arcabouço criado pelos algoritmos 6, 7, 8 e 9, chegamos ao algoritmo que finalmente preencherá o conjunto E e conseguirá dar para quase todos os nós o número de arestas de saída (*outgoing*) determinadas na tuplas do conjunto $\mathcal{S}$. Nossa abordagem resolve o problema de casamento dos *stubs* de maneira trivial, mas não garante exatidão no resultado. Este método está descrito no Algoritmo 10.

Algoritmo 8: Algoritmo para criação de série de índices de nós do grafo $G(N, E)$, S.

Entrada: Vetor de stubs $\mathcal{S} = \{(a_1, s_1), (a_2, s_2), \ldots, (a_{L-1}, s_{L-1}), (a_L, s_L)\}$
Saída : Vetor com sequência de índices de nós a_i, S

```
1 function MakeIndexSequence(S)
2     n = |S|// Obtendo o número de pares ordenados;
3     foreach par ordenado p ∈ S do
4         for i = 0, ..., p[1] do // p[1] acessando o grau do par ordenado p
5             append (S, p[0])// p[0], índice do nó no par ordenado p
6         end
7     end
8     return S
```

Algoritmo 9: Algoritmo que conecta dois nós do grafo de maneira aleatória. Se a conexão entre ambos ainda não tiver sido feita, relembrando que $e_{i,j} \neq e_{j,i}$, cria-se a aresta em questão e adiciona-se a mesma a E. Remove-se o elemento na posição i, porque aquele nó tem uma aresta que sai (*outgoing*) a menos.

Entrada: Grafo $G(N, E)$ e sequência de índices dos nós pertencentes a G,S.
Saída: Sequência de índices dos nós pertencentes a G,S.

```
1  function connectNodePair(G(N,E),S) // |S| cardinalidade de S.
2      i = 0to1rand{} * |S|;
       // 0to1rand() Sorteia um valor no intervalo [0,1).;
3      t=GetFrom(i,S);
4      j = 0to1rand{} * |S|;
       // GetFrom(a,b) retorna b_a caso a esteja nos limites de b. ;
5      h=GetFrom(j,S);
6      if t ≠ h AND e_{t,h} ∉ S then
7          E = E ∪ {e_{t,h}};
8          RemovePos(i,S)// Remove posição i de S;
9      end
10     return S;
```

O algoritmo em sua forma original, descrito em Newman (2018), faz os *stubs* serem pontos de entrada de cada aresta, ou seja, quando gero *stubs* de um nó, está sendo construída a vizinhança de entrada (*incoming*) de cada nó. A decisão de mudar essa arquitetura do algoritmo, de gerar a vizinhança de saída e não a de entrada, advém da natureza do problema com que se lidou neste trabalho: buscou-se controlar os parâmetros da distribuição da vizinhança dos agentes, que define quem eles seguem, ou seja, a vizinhança de saída. A estruturação das vizinhanças de entrada de cada agente pode ser importante num passo futuro, mas não é o caso agora. Ainda em Newman (2018), são demonstradas características interessantes de grafos montados a partir do Modelo de Configuração; o autor faz uma análise em que uma distribuição qualquer é levada em conta, trazendo uma fonte rica de informações sobre grafos construídos por este processo.

Algoritmo 10: Método de construção de grafos pelo Modelo de Configuração. É importante perceber que os passos para chegar nesse ponto também fazem parte do processo.

Entrada: Grafo sem arestas $G(N, E)$, sequência de índices S, tolerância ϵ para tentativas falhas de adição de arestas.

Saída: Grafo com densidade ou valor aproximado da densidade desejada $G(N, E)$.

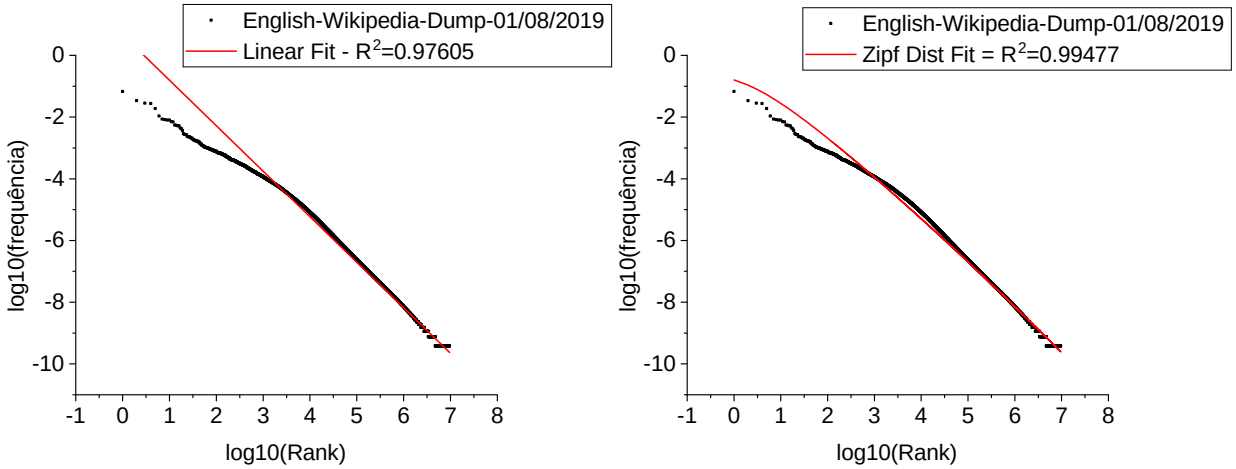
```
1  function ProduceGraph(G(N,E),S,ε)
2      ed = |S|/(N*(N-1)) // densidade esperada;
3      sum = |S|;
4      lsum = 0;
5      ad = 0// densidade atual;
6      while ad < ed AND sum > 0 do
7          lsum = sum;
8          S=connectNodePair(G(N,E),S);
9          sum = |S|// Se houve inserção de aresta sum = lsum − 1;
10         if lsum − sum == 0 then // Se não houve inserção de aresta.
11             ε = ε − 1;
12             if ε == 0 then
13                 return G(N,E);
14             end
15         end
16         ad = |E|/(N(N-1))// |E| cardinalidade do conjunto de arestas.;
17     end
18     return G(N,E);
```

B Wikipedia e Lei de Zipf

O trabalho de George Kingsley Zipf (G. K. Zipf, Linguista, 1902-1950) baseia-se na teoria da *Lei do Menor Esforço*. Em Zipf (2016), o autor discorre sobre esta teoria em diversos aspectos, principalmente em Linguística. Zipf procurou estudar esse efeito também usando tipos de dados que vão além de frequências de palavras, como a distribuição populacional de municípios em um determinado estado ou a distribuição do número de empregados nas empresas. Deve-se lembrar que, devido aos avanços tecnológicos dos últimos anos aplicados às linhas de produção e também ao fato da data do estudo ser da década de 1940, as regras podem não ser mais aplicáveis. De toda forma, em todos esses casos foi possível observar uma lei de potência denominada *Lei de Zipf*.

É importante notar que esta lei é um objeto matemático empírico, onde não existe um lei de universalidade para todos os sistemas, somente para os quais pode ser observada. Para entendermos como surge a Lei de Zipf, consideremos um exemplo. Tomemos inicialmente a totalidade dos artigos da Wikipedia em um certo momento. Calculemos as frequências de todas as palavras que aparecem em cada artigo. Uma palavra, juntamente com as suas frequências em cada artigo, forma o que é denominado um objeto, a ser analizado segundo a Lei de Zipf. Associa-se a um objeto uma frequência total, que é a soma das suas frequências em todos os artigos. O objeto de rank 1 é o que tem maior frequência total. O de rank 2, aquele com a segunda maior frequência; analogamente, define-se o rank dos demais objetos. Segundo a Lei de Zipf, a frequência total de um objeto decai, em função do rank, de acordo com uma lei de potência com coeficiente -1. Pode-se considerar a existência de uma Lei de Zipf anômala, cujo coeficiente de decaimento é diferente de -1. Nestes casos, características relacionadas a menor esforço podem ser perdidas.

Ao lidarmos com fenômenos humanos ou naturais, é interessante que tenhamos uma ferramenta que consiga confirmar a natureza da Lei de Zipf do que está sendo observado. Uma das propostas é a utilização da *Distribuição de Zipf*, que é bem descrita em Nair, Sankaran e Balakrishnan (2018)[pp. 132]. A Distribuição de Zipf é uma distribuição de lei de potência discreta que é intimamente ligada à lei de Zipf. Por meio dela, é possível calcular o rank de um objeto que pertence a um conjunto que, em princípio, obedece a uma Lei de Zipf. Para dissertarmos sobre a Distribuição de Zipf, voltemos ao nosso exemplo relacionado à Wikipedia. Tomemos um objeto com suas frequências (cada uma relacionada a cada artigo). Para cada valor de frequência, contabilizemos o número de artigos em que a palavra em questão apresenta aquela frequência. Utilizando tais dados, constroi-se um histograma a partir do qual podem ser calculadas probabilidades, que serão usadas para realizar um



(a) Fit usando Equação Linear. (b) Fit usando Equação 49.

Figura 26 – Exemplo de investigação da Lei de Zipf obtido através de um experimento com um despejo (*dump*) da Wikipedia, (WIKIPEDIA.ORG, 2021), ***enwiki-20190801-pages-articles.xml.bz2***. Os pontos pretos correspondem aos dados empíricos, frequências de palavras em função do *rank* em escala log-log, e as retas vermelhas, aos ajustes efetuados a partir dos dados. Na Figura 26a, é apresentado um ajuste linear no gráfico em escala log-log; o coeficiente de decaimento em lei de potência encontrado foi -1.48, com coeficiente de determinação $R^2 = 0.996$. Tal coeficiente não corresponde ao valor -1 esperado segundo a Lei de Zipf. Na Figura 26b, o ajuste é realizado com o uso da Equação 49, proposta como alternativa ao ajuste linear, por permitir o cálculo de estatísticas como a média e o desvio padrão. Os parâmetros encontrados: $\rho = -1.16$ e $\theta = -0.566$, sendo $R^2 = 0.99477$.

ajuste de acordo com a equação

$$q_x = \frac{x^{-1-\rho}}{\zeta(\rho+1)} \begin{cases} x > 0 \\ x \in \mathcal{Z} \\ \rho > 1 \end{cases}, \tag{47}$$

onde q_x é proporcional à probabilidade de encontrarmos uma palavra, no corpus, com a fequência x (torna-se a probabilidade, após normalização). Neste ajuste, estima-se o valor de ρ; se este conjunto obedecer à distribuição de Zipf, ρ terá o mesmo rank do objeto. Aqui, $\zeta(\rho)$ é a função Zeta de Riemann,

$$\zeta(\rho) = \sum_{i=1}^{\infty} i^{-\rho}. \tag{48}$$

Assim, podemos dar um exemplo da verificação dessa condição da seguinte forma: dado o objeto O_z da Lei de Zipf z com rank o, a frequência dos elementos do objeto O_z produzirá um rank o que, somado a mais um, $o + 1$, será igual ao rank encontrado na

Equação 47, ou, simplesmente, $o + 1 = \rho + 1$. Isso significa que, quando encontramos a constante ρ de um conjunto de dados (que define um objeto), existe uma grande probabilidade de que este conjunto de dados faça parte de uma lei de Zipf com rank exatamente igual a ρ.

Na Figura 26, é exposto um exemplo de aplicação da Lei de Zipf. O experimento foi realizado com um despejo (*dump*) da Wikipedia, (WIKIPEDIA.ORG, 2021), ***enwiki-20190801-pages-articles.xml.bz2***, do ano de 2019. Apesar dessa versão não ser mais acessível, não parece haver motivos para que versões mais novas levem a resultados qualitativamente diferentes. Na Figura 26a, o procedimento canônico para investigação da Lei de Zipf é seguido: um ajuste linear foi realizado, explicitando um decaimento em Lei de potência com coeficiente -1.48, sendo o coeficiente de determinação $R^2 = 0.996$. Na Figura 26b, o ajuste é efetuado de acordo com a distribuição

$$p_x = \frac{x^{-\rho}}{\zeta(\rho)} + \theta \quad x \geq 0, \tag{49}$$

inspirada na Equação 47, mas utilizada em um contexto diferente do descrito acima, referente àquela equação. A Equação 49 é proposta aqui como uma alternativa ao modelo de decaimento em lei de potência. Para os dados utilizados na construção da Figura 26, foram encontrados $\rho = -1.16$ e $\theta = -0.566$. A qualidade do ajuste, quantificada por $R^2 = 0.99477$, mostrou-se bastante satisfatória. A quantidade de parâmetros a serem estimados no ajuste aqui proposto é a mesma do referente a lei de potência, uma vez que este, sendo realizado através do ajuste de uma reta, produz um coeficiente angular e um coeficiente linear. A vantagem do ajuste pela Equação 49 sobre o ajuste por lei de potência é a possibilidade de cálculo de estatísticas como média e desvio padrão, que não são passíveis de definição em uma distribuição livre de escala.

Na Tabela 6, abordamos o tema da dispersão das palavras. Para o corpus com que estamos trabalhando, a técnica apresentada por Gries (2008) é a que possui o perfil mais adequado, porque se adequa a corpora com definições de partes de diferentes tamanhos. No caso de técnicas de cálculo de dispersão, as unidades de divisão de um corpus podem se diferenciar em diversos níveis: capítulos, seções, páginas, verbetes e, no caso do Wikipedia, artigos. Um aspecto importante neste contexto é a normalização textual: a busca de uma divisão do corpus que não gere tendências de acordo com o impacto (raridade) de cada palavra. Outras maneiras de cálculo de dispersão podem ser encontradas tanto para divisões iguais como para diferentes. Também relevante é a técnica chamada IDF (frequência inversa do documento ou, em inglês, *inverse document frequency*), muito utilizada na área de processamento de linguagens naturais.

Para definir o problema de contagem e dispersão de palavras, consideraremos dois conjuntos: $\mathcal{P}$, de palavras, cujo conteúdo inicialmente é vazio, e o conjunto $\mathcal{A}$, que representa o corpus, os dados conhecidos, $\mathcal{A} = \{a_1, a_2, \ldots, a_j, \ldots, a_{n-1}, a_n\}$, onde cada

Palavra	Frequência(F)	Dispersão(D)	$F \times D$
the	1.76905×10^8	0.14584	2.58007×10^7
of	8.96166×10^7	0.19478	1.74555×10^7
in	7.18088×10^7	0.18265	1.31156×10^7
dominated	71858	0.94378	67818
angle	69525	0.97538	67814
otto	69525	0.97499	67786
wished	26895	0.97404	26197
fold	26564	0.98617	26197
oaks	26427	0.9912	26194
dictator	12174	0.99054	12059
rods	12131	0.99383	12056
sorts	12190	0.98873	12053

Tabela 6 – Esta tabela trata de doze palavras espaçadas em intervalos de quatro mil passos, ou seja, as três primeiras palavras, as três palavras a partir de quatro mil, as três a partir de oito mil, seguindo-se analogamente até doze mil. Na primeira coluna, temos o nome da palavra, na segunda coluna temos a frequência não normalizada (F) da palavra, na terceira coluna temos a taxa de dispersão D (GRIES, 2008), e na quarta coluna temos o produto $F \times D$, que foi sugerida em Gries (2008) como uma nova forma de rankeamento.

a_j representa um artigo da Wikipedia ($j = 1, 2, ..., n$). Logo no principio no processo de contagem das palavras, vamos preenchendo o conjunto $\mathcal{P}$ de tal forma que definiremos seus elementos p_i como vetores coluna de tamanho n, onde cada posição j deste vetor possuirá a contagem da palavra i no artigo j, $p_{i,j}$. Definiremos $|p_i|$ como a contagem total relativa à palavra de índice i e $|a_j|$ como a quantidade de palavras do artigo j. Desta forma, definimos

$$f_{i,j} = \frac{p_{i,j}}{|p_i|} \tag{50}$$

como a frequência relativa da palavra i no artigo j, sendo o total de palavras

$$T = \sum_{k=1}^{n} |a_k| \tag{51}$$

e

$$\epsilon_j = \frac{|a_j|}{T}, \tag{52}$$

o valor esperado do corpus. A dispersão das palavras é calculada da seguinte forma:

$$D_{p_i} = \sum_{j=1}^{n} \frac{|f_{i,j} - \epsilon_j|}{2}. \tag{53}$$

Dois algoritmos, descritos abaixo, demonstram como a contagem de palavras é feita e como a dispersão é calculada.

Algoritmo 11: Técnica de contagem de palavras.

Entrada : Conjunto $\mathcal{P}$, o corpus $\mathcal{A}$, e n, o número de partes (artigos, páginas ou qualquer tipo de granulometria em A que se deseja usar, no nosso caso, artigos.)
Saída : Conjunto P preenchido.

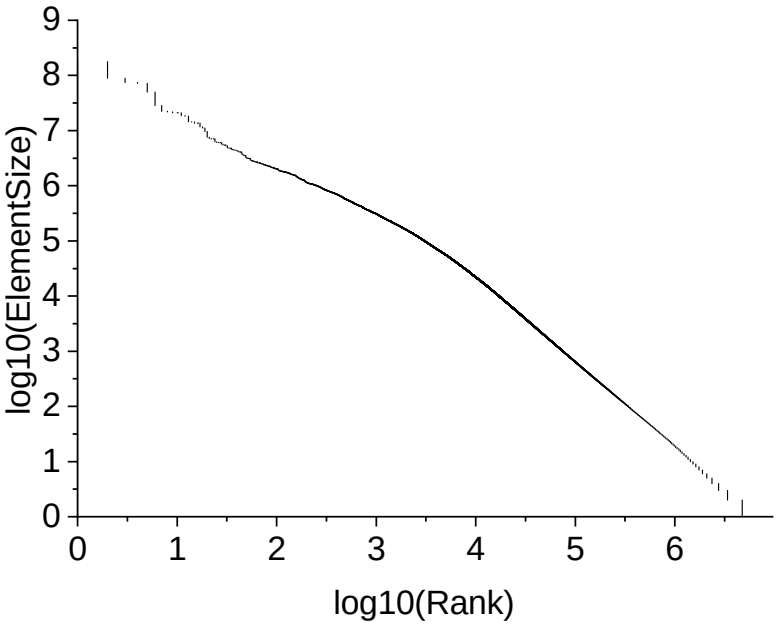
for $j \leftarrow 1$ **to** n **do**
  H_j=`WordHist` $(j,\mathcal{A})$`// Faz histograma do j-ésimo artigo de` $\mathcal{A}$.;
  for $p_i \in H_j$ **do**
    if $p_i \notin \mathcal{P}$ **then**
      $\mathcal{P} \leftarrow p_i$;
    end
    $p_{i,j} \leftarrow$ `Count`(p_i, H_j)`// Conta quantas palavras` p_i `em` H_j.;
    `Put`$(\mathcal{P}, p_i, p_{i,j})$`// Put(),Põe` $p_{i,j}$ `no vetor de` p_i.;
  end
end
return $\mathcal{P}$

Algoritmo 12: Cálculo da dispersão de palavras contidas no corpus, palavra por palavra, depois por cada vez em que houve presença da palavra p_i em um artigo j, $p_{i,j}$, e também quando não há.

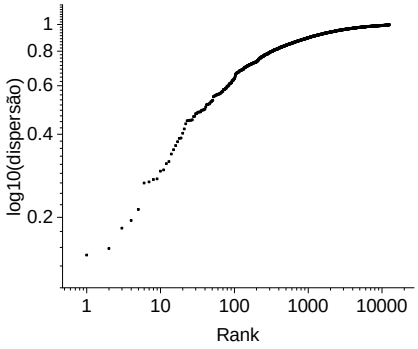
Entrada : Conjunto $\mathcal{P}$ preenchido pelo algoritmo 11
Saída : Densidade de todas as palavras, $\mathcal{D}$.

$\overline{\mathcal{A}}$=`SomaPalavras` $(\mathcal{A})$`// Soma as palavras de cada artigo` j.;
$T = \sum_k^n \overline{\mathcal{A}}_k$`// Soma todas a somas e obtem-se o total de palavras`.;
$\langle T \rangle = \frac{T}{n}$;
$\mathcal{D} \leftarrow \{\}$;
for $p_i \in \mathcal{P}$ **do**
  $d_i \leftarrow 0$;
  for $p_{i,j} \in p_i$ **do**
    $f \leftarrow \frac{p_{i,j}}{|p_i|}$`// Frequência relativa de` $p_{i,j}$`, Equação 53.`;
    $\epsilon_j \leftarrow \frac{|a_j|}{T}$`// Valor esperado do artigo` j`, Equação 53.`;
    $d_i \leftarrow d_i + \frac{|f-\epsilon_j|}{2}$;
  end
  for $p_{i,j} \notin p_i$ **do**
    $d_i = d_i + \frac{\epsilon_j}{2}$;
  end
  $\mathcal{D} \cup \{d_i\}$;
end
return $\mathcal{D}$

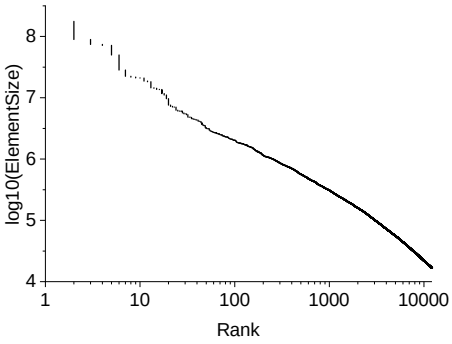
Os algoritmos 11 e 12 explicam tecnicamente os fatores que vamos discutir doravante. Consideramos que o rank r_i de uma palavra i é a posição em ordem decrescente da soma



(a) Gráfico de barras móveis do corpus em estudo.



(b) Dispersão das primeiras 12.000 palavras.



(c) Gráfico de barras móveis das primeiras 12.000 palavras.

Figura 27 – Mostramos na Figura 27a um gráfico de barras onde o tamanho de cada barra representa a distância das frequências de palavras adjacentes. Os gráficos 27b e 27c representam a dispersão das 12.000 primeiras palavras e o gráfico de barras das mesmas, respectivamente.

do vetor da palavra i em relação à soma das outras palavras cujas frequências são dadas em $\mathcal{P}$. Como dito anteriormente, r_i se relaciona diretamente com o expoente da distribuição de Zipf da palavra i, caso aplicável, ou seja, Lei de Zipf com inclinação próxima a -1. No entanto, em trabalho mais recente (EGBERT; BIBER, 2019), a dispersão tem sido alvo de estudos, por ser uma informação adicional sobre a formação do corpus, complementando, desta maneira, informações sobre a morfologia do corpus, antes que qualquer outro tipo de análise léxica e ou semântica sejam feitas. Deve ficar claro que uma baixa taxa de dispersão significa que a palavra tem alta probabilidade de aparecer na maioria dos artigos, e uma alta taxa de dispersão significa que a palavra tem alta probabilidade de aparecer, de maneira condensada, em poucos artigos. Deve ficar claro que detectar a dispersão das palavras

informa sobre o comportamento da morfologia do corpus em si, mas pode potencializar análises textuais de outros níveis quando agregados.

Na Figura 27a, podemos notar diferentes flutuações nas larguras das barras, que explicitam a diferença entre pontos adjacentes. Na região onde tais flutuações estão localizadas, elas mostram o desvio de uma Lei de Zipf canônica. Note as flutuações no tamanho das barras na faixa de ordenadas de 4×10^8 até 5×10^6. Por outro lado, há uma seção da série em que o tamanho dessas barras se torna padronizado, de 10^5 até 10. Essa seção está localizada numa faixa de frequências de ordens de grandeza muito pequenas em comparação com as faixas onde as anomalias ocorrem. Por outro lado, correspondem a uma região significativa do gráfico, envolvendo grande concentração de pontos. Isto sugere, mas não faz necessário, o comportamento da série anômala como observado na Figura 26. Fazendo uma comparação simples entre as figuras 27b e 27c, a aparente proporção inversa entre os dois gráficos não leva em consideração uma normalização de escalas. Esta pode ser feita através da razão Q, uma relação definida originalmente neste trabalho:

$$Q = \frac{D}{F}, \tag{54}$$

onde D é a dispersão, definida na Equação 53, e F a frequência, relativas a cada palavra. Normalizando Q, obtemos a grandeza Q',

$$Q'_i = \frac{Q_i - \langle Q \rangle}{\sigma(Q)}, \tag{55}$$

onde:

- Q_i é o i-ésimo valor de Q, referente à palavra de índice i;
- $\langle Q \rangle$ é o valor médio de Q;
- $\sigma(Q)$ é o desvio padrão de de Q.

Dessa forma, ao construírmos o gráfico de $Q' \times R$, onde R é o rank, temos a Figura 28.

Na Figura 28, encontramos que, para o corpus estudado em questão (***enwiki-20190801-pages-articles.xml.bz2***), para $R = 12000$ o inverso da frequência das palavras multiplicado pela dispersão das mesmas encontra uma relação exponencial em relação ao rank das palavras. Isto gera uma nova forma de relacionar as estatísticas de corpus, levando em consideração o parâmetro de dispersão, D. Esta relação deve ser investigada em outros contextos.

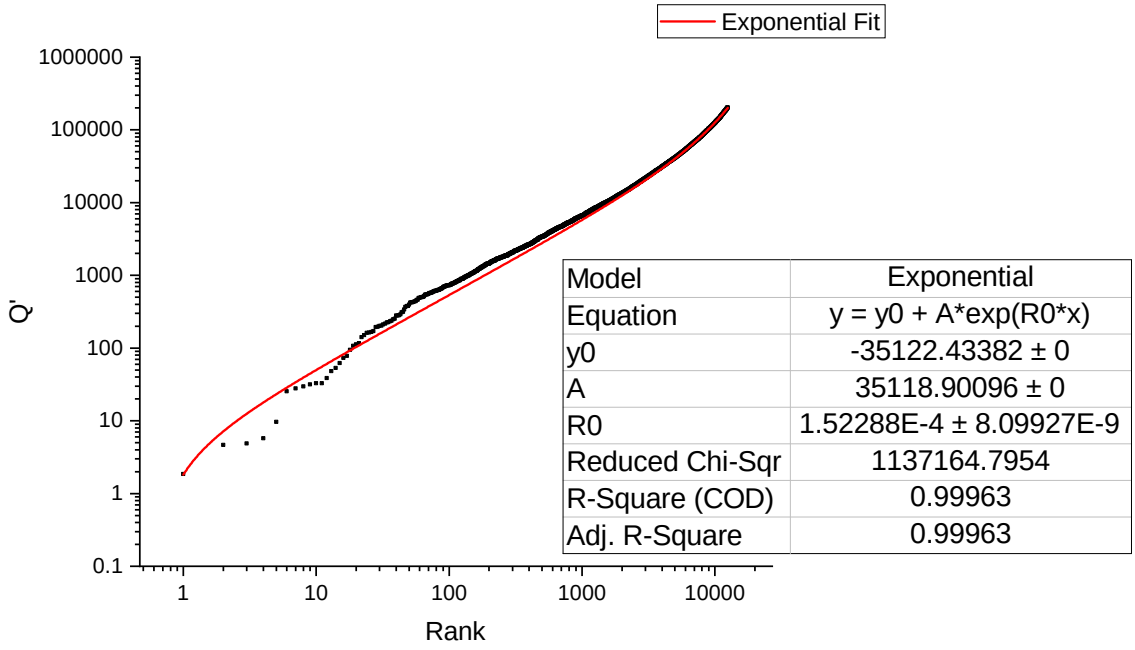


Figura 28 – A relação apresentada na Equação 54 tem como resultado uma relação exponencial entre Q, calculado até o rank 12000 e o rank R, onde essa exponencial é descrita da seguinte forma: $y = y0 + A * e^{R0*x}$, sendo $y0$ a constante de deslocamento vertical, A o valor inicial e $R0$ a razão de crescimento ou decrescimento da função, dependendo de seu sinal.